NOVA IMS
Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

A segment analysis to understand the behavior of greenhouse gases

Carolina Oliveira de Araújo

Dissertation presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

A segment analysis to understand the behavior of greenhouse gases

by

Carolina Oliveira de Araújo

Dissertation presented as partial requirement for obtaining the master's degree in Data Science and Advanced Analytics

Advisor: Roberto Henriques

July 2021

ACKNOWLEDGEMENTS

I wish to express the gratitude to my thesis advisor, Professor Roberto Henriques, whose guidance was vital when doing this research. He constantly allowed this essay to be my own creation but pushed me in the right path whenever I was most lost.

Likewise, I must acknowledge the support of my family and friends for their encouragement in pursuing this final step of my master's degree.

ABSTRACT

"If something is important enough, even if the odds are against you,
you should still do it."

Elon Musk

For years I admired this statement, and it is safe to express this applied to my master thesis. This research is focused on understanding which countries characteristics influences the production of greenhouse gases throughout 1990 to 2017. The data used to produce the project consists on economic and environment related data. The methodology applied is based on *Knowledge Discovery in Databases Process*, with an emphasize on the step, Data Mining, where the approach used was unsupervised learning.

Only after I started exploring research about time series clustering for greenhouse gases, it was when I realized probably this would not be the best approach for the main question. Nevertheless, I was determined to validate if this was not the case since, I firmly believe all approaches should be tested in order to understand which one is better suited to answer a research question.

KEYWORDS

Greenhouse Gases, Emissions, Time Series, Clustering

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

BRICS	Brazil, Russia, India, China and South Africa
CCS	Carbon Capture Storage
CH_4	Methane
CO_2	Carbon Dioxide
C_2F_6	Hexafluoroethane
CRISP-DM	Cross Industry Standard Process for Data Mining
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DR	Dimensionality Reduction
EIT	Economies in Transaction
EM	Expectation-Maximization
FA	Factor Analysis
F-gases	Fluorinated greenhouse gases
GMST	Global Mean Surface Temperature
GWP	Global Warming Potential
GHG	Greenhouse Gases Emissions
IPCC	Intergovernmental Panel on Climate Change
IQR	Interquartile Range
KDD	Knowledge Discovery in Databases
KMO	Kaiser-Meyer-Olkin
KNN	K-Nearest Neighbors
Kt	Kiloton
LULUC	Land Use, Land-Use Change
LULUCF	Land Use, Land-Use Change and Forestry
NASA	National Aeronautics and Space Administration
NF_3	Nitrogen Trifluoride

N_2O	Nitrous Oxide
OECD	Organisation for Economic Co-operation and Development
PC	Principal Component
PCA	Principal Component Analysis
SEMMA	Sample, Explore, Modify, Model and Assess
SF_6	Sulfur hexafluoride
SOM	Self-Organizing Map
TPES	Total Primary Energy Supply
UK	United Kingdom
UNFCCC	United Nations Framework Convention on Climate Change
USA	United States of America

1. INTRODUCTION

The world is transforming right in front of our eyes, but we keep neglecting due to our self-indulgence and reluctance against change. It seems nowadays every little thing we see in the news is about climate change, even though some reports may apparent that it is not, it can be an indirect effect.

Let's think about the pandemic we are going through, deforestation is killing the home to several animals, forces them to leave their habitat and look for other ones, and in the process contact with different animals they were not used to, creating new environments for new bacteria (*Coronavirus and Climate Change – C-CHANGE | Harvard T.H. Chan School of Public Health, 2021*). Another case is Madagascar being the first country in the world starving exclusively by climate change (*Madagascar Famine is First Caused Entirely by Climate Change | Time, 2021*).

1.1. BACKGROUND AND PROBLEM IDENTIFICATION

Climate change and everything concerned to it, is acknowledged as the biggest concern of our time (IPCC, 2014) and it can happen due to natural reasons or human intervention. It is known the theme *climate change* has been present in our planet since millions ago as it has been the reason for dinosaurs extinction (*Climate Change Killed The Dinosaurs. 'Drastic Global Winter' After Asteroid Strike, Say Scientists, 2020*) however this matter was the work of mother nature. Since the past century until the present, human intervention is the main factor for these changes due to a tremendous booster, the Industrial Revolution, and it is maintained by the unsustainable consumption of population.

Greenhouse gases are what drives climate change and these chemical substances lead to the warmth of the atmosphere (Anderson et al., 2016). Such powerful phenomenon has repercussions throughout time, and it is interesting to understand what could have affected this issue and how the world as behaved since then, for better or for worst.

1.2. MOTIVATION

Even though climate change is such an interesting subject with several information available, the proposed research mainly focused on greenhouse gases (GHGs). It was thought that due to the fact that GHGs are at the core of the reason for climate change, it would be appropriate to study this question.

There are countless studies about GHGs, research goes from the impacts of greenhouse gases on society up to investigations on how countries and/or industries implement approaches in order to mitigate the damaging outcomes (Arioli et al., 2020; Flämig et al., 2019; Homma et al., 2012; Kaveh et al., 2020; Mohan, 2018; Röck et al., 2020; Serra et al., 2019). Nonetheless, it was not discovered much information concerning segmentation on GHGs (Homma et al., 2012; Mohan, 2018).

1.3. Research objectives

Due to the lack of information regarding GHGs segmentation, there was a call for this dissertation to focus on this matter and an attempt to answer the following question: *Which, if any, years had an impact on the GHGs emissions evolution through 1990-2017?*

In order to achieve the desired outcome, one needs to understand there are several methods for working with time series data. One of the purposes of the study is to corroborate if performing a segmentation, based on which countries and years have similar behavior, is the best practice to answer the main question. It is expected with the approach implemented that it can be easily understood what main characteristics can be observed in each clustering and thus, realize what type of evolution the planet has been through.

1.4. Document structure

Adding to this chapter, the presented research is structured as follows:

- Chapter 2 is where it is introduced a literature review concerning the subject greenhouse gases and the best practices to implement on the given data;
- Chapter 3 is the description of the methodology used to answer the main question;
- Chapter 4 is presented the segmentation approach used and the results derivates of that;
- Chapter 5 is the conclusion of the study and it references limitations during the process of the investigation and future work.

2. LITERATURE REVIEW

2.1. GREENHOUSE GASES

2.1.1. Climate change

Climate change, as stated by United Nations Framework Convention on Climate Change (UNFCCC), is considered a phenomenon where variations in climate's properties happen due to natural causes or due to direct and/or indirect human activity for long periods of time (United Nations, 1992). Intergovernmental Panel on Climate Change (IPCC), an organization affiliated to UNFCC, reckons that climate change can be recognized, using proper statistical tools, as fluctuations in the average and/or oscillations of climate's composition for a prolonged time (IPCC, 2018). It also underlines that climate change can happen as a result of natural internal transformations, for instance, variations of the solar cycles and volcanic eruptions or due to human activity, such as, anthropogenic variations in the atmosphere or land use (IPCC, 2018).

The term anthropogenic is what it is called for the outcomes from human activity such as the burn of fossil fuels, deforestation, land use and land-use changes (LULUC), livestock production, fertilization, waste management and industrial processes (IPCC, 2018). IPCC affirms that these consequences on climate has been the main reason for global warming since mid-20th century (IPCC, 2018).

2.1.2. Greenhouse gases

The greenhouse effect is a phenomenon that occurs when solar energy crosses the atmosphere, warming the planet's surface and re-emitting as infrared energy by greenhouse gases (GHGs). Earth needs these natural gases to keep the ecosystem warmth or, otherwise, living beings would be living in a planet where the average temperature would be around -21°C (Anderson et al., 2016).

Intergovernmental Panel on Climate Change (IPCC) expresses that the main GHGs that contribute to the planet's heat are water vapor (H_2O), carbon dioxide (CO_2), nitrous oxide (N_2O), methane (CH_4) and ozone (O_3) and that these gases happen due to both natural and anthropogenic causes (IPCC, 2018). There are also other toxic GHGs that are caused by human activity, such as substances containing chlorine and bromine and halocarbons - halons, methyl chloride, methyl bromide, chlorofluorocarbons (CFCs), hydrochlorofluorocarbons (HCFCs), hydrofluorocarbons (HFCs), sulfur hexafluoride (SF_6), and perfluorocarbons (PFCs) (IPCC, 2018).

A recent study developed by the University of Michigan (2019) states that are ten GHGs that most contribute for global warming (Center for Sustainable Systems, 2019). The research has shown that H_2O, CO_2, CH_4 and N_2O happen due to natural reasons whereas perfluorocarbons (CF_6, C_2F_6), hydrofluorocarbons (CHF_3, CF_3CH_2F, CH_3CHF_2) and sulfur hexafluoride (SF_6) occur as a result of industrial activities (Center for Sustainable Systems, 2019). H_2O is the GHG that represents the major share in the atmosphere and its absorption mainly relies on temperature and other meteorological features and not directly due to human actions (Center for Sustainable Systems, 2019). As opposed, CO_2 is the anthropogenic GHG that most contributes for global warming caused by human activities (Center for Sustainable Systems, 2019).

The Center for Climate and Energy Solutions realized a research where distinguishes two important concepts concerning the evaluation of the strength of atmospheric gases on the greenhouse effect (Center for Climate and Energy Solutions, 2016):

- **Global Warming Potential (GWP).** It is an estimation of the radiative effect of each unit of GHG over a specific period, comparative to the radiative effect of CO_2 (Center for Climate and Energy Solutions, 2016). A high GWP for a certain amount of GHG warms the atmosphere more than the same amount of CO_2 (Center for Climate and Energy Solutions, 2016).
- **Atmospheric lifetime.** It is an estimation for how long a GHG prolongs in the atmosphere before natural procedures eradicate it (Center for Climate and Energy Solutions, 2016). When comparing two GHGs with equal GWP, the GHG with higher lifetime does more harm to the atmosphere (Center for Climate and Energy Solutions, 2016).

The GHGs emissions have a specific measurement to evaluate, *kt (kiloton)*, that can only be compared to CO_2 (kt CO_2 equivalent) since is the gas that most contributes to global warming (Center for Sustainable Systems, 2019; Environmental Indicators for Agriculture, 2001). This measure is estimated based upon a GHG global warm potential, usually, for an atmospheric lifetime of 100 years (Environmental Indicators for Agriculture, 2001).

The same study also assures that the main GHGs are CO_2, CH_4, N_2O, SF_6, chlorofluorocarbon-12 (CCl_2F_2), hydrofluorocarbon-23 (CHF_3) and Nitrogen Trifluoride (NF_3). It affirms that gases like CO_2 occur by natural and anthropogenic processes while fluorinated gases are outcomes of manmade activities (Center for Climate and Energy Solutions, 2016). These fluorinated gases or F-gases contribute for a global warming 23.000 times worse than CO_2 and HFCs are the most problematic for the atmosphere in the short run, nevertheless, PFCs and SF_6 can endure thousands of years (Center for Climate and Energy Solutions, 2016).

2.1.2.1. GHGs drivers

According with IPCC, the human activities that contribute more to the increase of anthropogenic GHG emissions are the constantly growth of population size and its lifestyle, economy, energy use, land use and technology (IPCC, 2014). The Journal of Cleaner Production believes that there are three indicators to measure the impact of the increase of anthropogenic GHGs: energy intensity, economic growth and carbon factor (Zheng et al., 2019). **Energy intensity** is the most used indicator to measure the level of energy efficiency of a country. According with the study, the energy intensity of a country rises throughout the industrialization stage and since the post-industrialization stage has been decreasing due to the fact of manufacturing is being substituted by services (Zheng et al., 2019). The carbon intensity of a country's energy is measured by the **carbon factor**, where GHGs emissions per unit are relative to the total primary energy supply (Zheng et al., 2019).

Rosa and Dietz (2012) did a more focused investigation of anthropogenic GHGs drivers and how they affect the atmosphere, and these are considered the ones:

- **Population:** the continuous growth of the population is a major risk for the environment and its pace of growth it is also worrying (Rosa & Dietz, 2012). Besides these, the growth of the number of households is far more crucial for the overall anthropogenic GHGs (Rosa & Dietz, 2012);
- **Consumption:** the crushing consumption patterns of the population due to cultural differences and the technology used to produce what it is demanded are reasons for the growth of anthropogenic emissions (Rosa & Dietz, 2012);
- **Urbanization:** the fact that more than half of the population live in urban cities, with tendency to grow, the consumption of energy has high levels for anthropogenic emissions in the atmosphere (Rosa & Dietz, 2012);
- **Trade:** it is believed that the liberalization of trade worldwide jeopardizes climate change since developing countries do not have much control over environmental taxes, therefore, companies use these for their advantage and jeopardize the atmosphere (Rosa & Dietz, 2012);
- **Institutions:** international agreements have been made to reduce anthropogenic GHG emissions (Rosa & Dietz, 2012). At a national level, it has been verified that democratic countries are more willingly to implement more environmental policies than oppressed countries (Rosa & Dietz, 2012);
- **Values, beliefs, norms, trust and worldviews:** these are major drivers that influence every day human activities which is of a big influence on climate change (Rosa & Dietz, 2012). Since this is such a subjective variable it is hard to find a consistent evaluation of the impact that these actions have on the environment, nevertheless, it cannot be denied that are of a major influence in everyone choices (Rosa & Dietz, 2012).

2.1.2.2. GHGs players

As mentioned in the last section, everyone is responsible for the continuous growth of anthropogenic GHGs but there are groups of countries that have major economic impacts in the world (Zheng et al., 2019). This is the case of the G7, which is represented by the seven most developed economies: Canada, France, Germany, Italy, Japan, United Kingdom (UK) and United States of America (USA). Together they correspond 60% of the global net wealth and almost 50% of the global GDP. Emerging economies also have a big influence on GHGs emissions, which is the case of BRICS countries, an acronym for Brazil, Russia, India, China and South Africa (Zheng et al., 2019).

In 2017, the drivers of G7 and BRICS GHGs emissions accounted for 60% excluding land use, land-use changes and forestry (LULUCF). China is considered the largest emitter of GHGs emissions, 24% of the global emissions, due to the vast consumption of coal and largest solar technology manufacturer (Zheng et al., 2019).

2.1.3. Consequences of climate change

For the reasons above, there are costs for these actions that can affect both positively and negatively the climate. IPCC (2018) classifies global warming as the rise of the global mean surface temperature (GMST) and it is considered the most significant consequence of anthropogenic GHGs emissions. National Aeronautics and Space Administration (NASA) shows there are other significant consequences for the climate (NASA, 2018):

- **Global temperature rise.** Since 1850, when record-keeping of the global temperature has started, Earth's mean surface temperature has increased 0.9 °C, mainly, due to the rise of carbon dioxide and other human activities in the atmosphere (NASA, 2018). The last decade was considered the most worrying since it was recorded with the highest temperatures, where 2016 was the warmest year ever been registered (NASA, 2018);
- **Oceans warming.** This rise of the temperature has been heating the oceans since 1969 (NASA, 2018);
- **Shrinking ice sheets.** Ice sheets in Greenland and Antarctica have been losing hundreds of billions of tons of ice between 1993 and 2016. In the past decade, the disappearance of Antarctica's ice-covered area has tripled (NASA, 2018);
- **Glacial retreat.** Glaciers are diminishing not only in Greenland and Antarctica but also in the Alps, Himalayas, Andes, Rockies, Alaska and Africa (NASA, 2018);
- **Decreased snow cover.** In the northern hemisphere, snow cover has been decreasing in the last 50 years and the snow is dissolving even earlier (NASA, 2018);
- **Sea level rise.** In the past century, global sea level increased around 20 centimeters, with every year rising a small portion (NASA, 2018);
- **Declining Arctic sea ice.** The amount and depth of the Artic sea ice has been decreasing in the past decades (NASA, 2018);
- **Extreme events.** Since 1950, high temperature records have been growing as opposed to low temperature records, that have been diminishing (NASA, 2018);
- **Ocean acidification.** The reason for the continuous rise of carbon dioxide is because of Industrial Revolution, in 1760. This contributed to the growth of ocean's acidity about 30%, where the amount of carbon dioxide absorbed every year increases 2 billion tons (NASA, 2018).

2.1.4. International environmental agreements

Due to the worrying prospects for the planet, it was necessary for countries to act and to start compromising in order to reduce the global temperature rise. UNFCCC was implemented as an international environment agreement in March 1994, where 196 countries and European Union signed the deal (IPCC, 2018). The purpose of this institution is to stabilize GHGs emissions in the atmosphere at a level that must not jeopardize the climate (IPCC, 2018). There are two treaties that give provisions for the countries who signed (IPCC, 2018):

- **Kyoto Protocol**

It was written in 1997 in Kyoto, Japan, but only in February 2005 it was fully implemented for its 192 parties (IPCC, 2018). The goal of the treaty had two phases, where the first one was for countries to commit themselves in decreasing anthropogenic GHGs emissions no less than 5% below 1990 levels, where these results should be achieved between 2008 and 2012 (IPCC, 2018). The second phase was decided in December 2012, known as the Doha Amendment, where a new group of countries agreed in diminishing anthropogenic GHGs emissions no less than 18% below 1990 levels between 2013 and 2020 (IPCC, 2018). The Doha Amendment did not receive enough approval to put into action since May 2018 (IPCC, 2018).

- **Paris Agreement**

It was implemented in November 2016 and 196 countries agreed on the intentions but later, in May 2018, it only had 195 signatures and was approved by 177 countries (IPCC, 2018). One of the objectives of the treaty was to support countries taking responsibility and action of the effects of climate change (IPCC, 2018). Another important aim was to decrease the global average temperature to 2°C lower than pre-industrial levels and to implement actions to limit the rise of the temperature to 1.5°C higher than pre-industrial levels (IPCC, 2018). Pre-industrial levels are a reference for GHGs emissions before the Industrial Revolution (in 1760), and IPCC uses the timeframe 1850-1900 to approximate to pre-industrial levels of emissions, since record-keeping only started in 1850 (IPCC, 2018).

2.2. TIME SERIES CLUSTERING

2.2.1. The problem of time series data mining

The introduction of time series data mining comes from the claim that individuals need to picture the shape of data since it is easier to recognize resemblances among patterns (Esling & Agon, 2012).

Various studies concluded that due to the unique behavior of time series data, information regarding this subject still is insufficient and this is considered one of the 10 problems in data mining (Qiang & Xindong, 2006). Researchers focused mainly in developing time series data mining methods (Fu, 2011) instead of improving problems like high data dimensionality, indexing issues and similarity measure (Esling & Agon, 2012; Fu, 2011). Esling and Agon (2012) state that the major obstacles to solve this matter are data representation, similarity measure and indexing mechanisms (Esling & Agon, 2012).

The study realized by Fu (2011), divided the various approaches for **data reduction** into two sections, time series dimension reduction directly in the time domain and dimension reduction in the transformation domain (Fu, 2011).

When it comes to dimensionality reduction in the **time domain**, **sampling** is the easiest approach in doing so (Åström, 1969). However, the distortion of the shape can be compromised so adopting the average of each segment can be another way to do it (Fu, 2011) or with **linear interpolation** (Keogh, 1997; Keogh & Smyth, 1997; Smyth & Keogh, 1997).

As for data reduction in the **transformation domain**, approaches used are principal component analysis (PCA) (Yang & Shahabi, 2005; Yoon et al., 2005) and self-organizing maps (SOM) (Hammer et al., 2005).

Similarity measure is crucial for time series analysis (Fu, 2011). When it comes to conventional databases, similarity measure has an exact match whereas in time series databases, similarity measure has an approximate distance (Fu, 2011). Using Euclidean distance in large data sets proves to be adequate to **measure similarity** since there is a big probability of finding an exact match in the data (Esling & Agon, 2012).

3. METHODOLOGY

The choice of methodology to use for the research was based on a study realized by Azevedo and Santos, where they discuss three different methodologies (Azevedo & Santos, 2008): **KDD** (Knowledge Discovery in Databases) (Fayyad et al., 1996), **SEMMA** (Sample, Explore, Modify, Model and Assess) and **CRISP-DM** (Cross Industry Standard Process for Data Mining). The research concluded that both SEMMA and CRISP-DM can be seen as an application of the KDD process (Azevedo & Santos, 2008). When comparing KDD and SEMMA, the study clearly identifies the five steps of SEMMA to be very similar to the five steps of KDD (Azevedo & Santos, 2008). Regarding to a comparison between KDD and CRISP-DM, the only differences that can be seen are in business understanding and deployment stages of CRISP-DM since it is more business oriented (Azevedo & Santos, 2008). The other steps of the process are considered the same as the equivalent steps in KDD (Azevedo & Santos, 2008). To this extent, the following chapter consists of the usage of KDD Process to answer the main question.

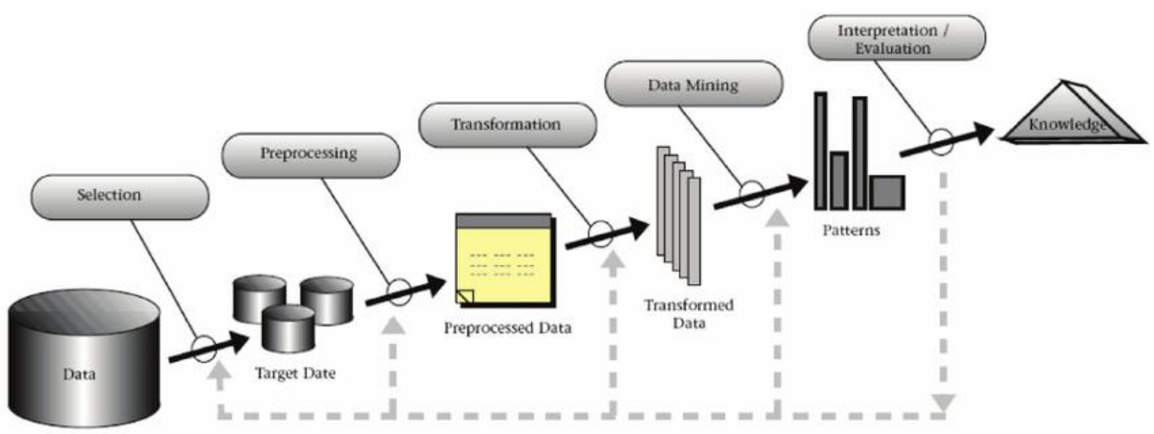


Figure 1 - KDD Process.

The method consists in five phases (Fayyad et al., 1996; Table 1):

1. **Data Selection:** data used from UNFCCC, OECD and World Bank.
2. **Data Preprocessing:** outlier detection and missing values imputation.
3. **Data Transformation:** data normalization for both continuous and categorical data and it was performed Principal Component Analysis, Factor Analysis and Self-Organizing Map as dimensionality reduction techniques.
4. **Data Mining:** the approach used was unsupervised learning - Clustering.
5. **Data Interpretation:** interpretation and evaluation of the results given in the data mining step (Fayyad et al., 1996) through contingency tables.

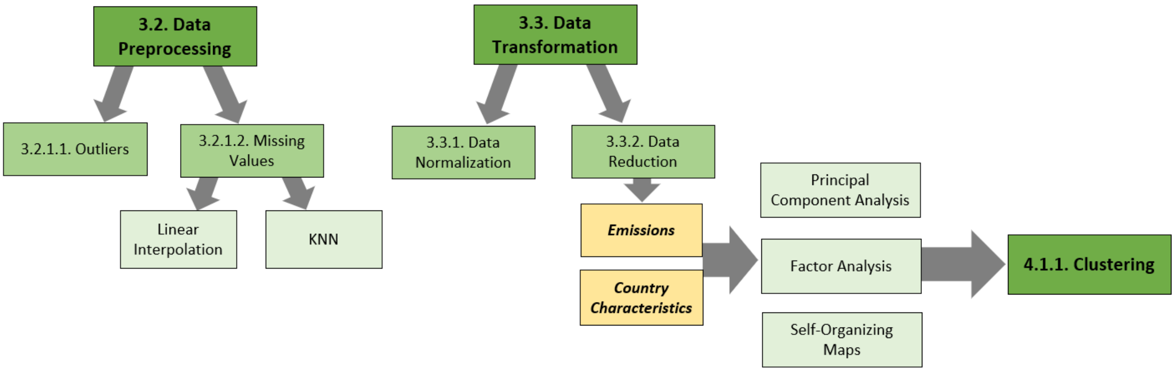


Figure 2 - Research process.

Regarding the Data Preprocessing, the practice used for outlier treatment was the interquartile range (IQR) method where, quartile 1 is set as 25% quantile and quartile 3 is set as 75% quantile. It was considered as outliers, values that have a factor of 1.5 of the IQR below the 25th percentile or above the 75th percentile. In order to fill missing data the chosen approach was linear interpolation with a maximum of 5 null values to fill and the methodology used to fill these numbers were both backward and forward fill.

As for the step Data Transformation, it was created a new feature, *nº years*, where it was converted the variable *Year* to the number of years passed until 2020. Due to the need of variable comparison the data was shaped through *StandardScaler* and *dummies* transformation, where continuous variables were molded with the first one and categorical variables were converged into the latter. For *dummy* variables, the values could only be 1 or 0, where 1 refers to the presence of the feature and 0 refers to the absence of the feature.

The data set was divided into two new ones, *Country Characteristics* and *Emissions*. As for *Country Characteristics*, belong variables that can characterize a country's effort to sustainability, whereas *Emissions*, belong variables that measures a country's emissions related to *kt* CO_2 *equivalent*.

For both data it was produced data reduction approaches. *Country Characteristics* was used PCA with 4 components while *Emissions* was used SOM with a network weights 15x15, where *Periodic Boundary Conditions* was set as true, a training network for 10.000 epochs and a learning rate of 0.1.

The approach used for Data Mining step was *unsupervised learning* – Clustering. Even though in the approach above it was performed *Emissions* with dimensionality reduction, for the chosen partition method, K-means with 2 clusters, it was performed without data reduction. As far comes *Country Characteristics*, the selected clustering technique was K-means with 4 clusters. These outcomes can be seen in the following chapter (Chapter 4).

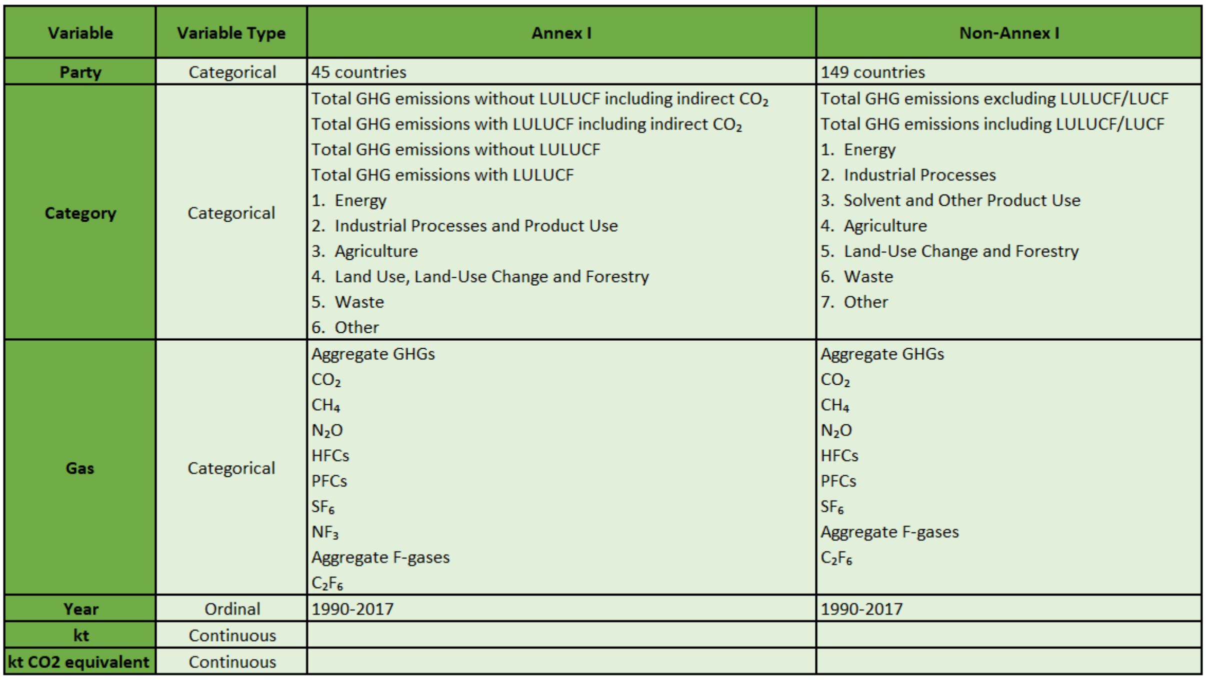
3.1. Data Selection

Data from UNFCCC, World Bank and OECD were considered the most suitable for the investigation due to countries' information about GHGs emissions and environmental related subjects, economy and population (Organisaton for Economic Co-operation and Development, 2015; World Bank, 2018). The time series established for the project was data from 1990 to 2017.

UNFCCC differentiates countries into three types: Annex I, Annex II and Non-Annex I (UNFCCC, 1992). Countries that belong to Annex I are industrialized countries and economies in transaction (EIT Parties), Annex II only has industrialized parties from OECD countries whereas developing countries belong to Non-Annex II (UNFCCC, 1992).

Working with Annex II was not considered useful for the research since Annex I had data from Annex II and more countries. In this degree, only Annex I and Non-Annex I were used in the project. In the table below, there is a more clarified description of both data sets.

Table 1 - Data description.



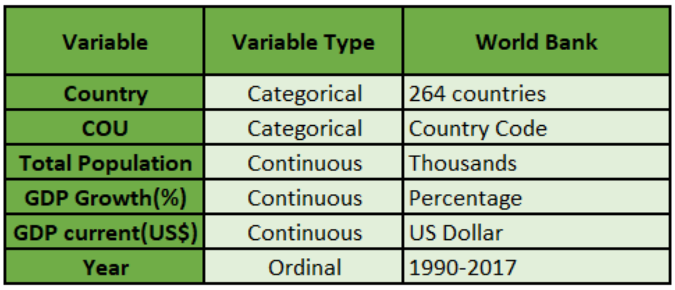
Variable	Variable Type	Annex I	Non-Annex I
Party	Categorical	45 countries	149 countries
Category	Categorical	Total GHG emissions without LULUCF including indirect CO_2 Total GHG emissions with LULUCF including indirect CO_2 Total GHG emissions without LULUCF Total GHG emissions with LULUCF 1. Energy 2. Industrial Processes and Product Use 3. Agriculture 4. Land Use, Land-Use Change and Forestry 5. Waste 6. Other	Total GHG emissions excluding LULUCF/LUCF Total GHG emissions including LULUCF/LUCF 1. Energy 2. Industrial Processes 3. Solvent and Other Product Use 4. Agriculture 5. Land-Use Change and Forestry 6. Waste 7. Other
Gas	Categorical	Aggregate GHGs CO_2 CH_4 N_2O HFCs PFCs SF_6 NF_3 Aggregate F-gases C_2F_6	Aggregate GHGs CO_2 CH_4 N_2O HFCs PFCs SF_6 Aggregate F-gases C_2F_6
Year	Ordinal	1990-2017	1990-2017
kt	Continuous		
kt CO2 equivalent	Continuous		

The variable *Party* is the one referring to the number of countries used in the research, for Annex I there are 45 countries and for Non-Annex I there are 149 countries. In this first step, *Category* is considered a categorical variable but later it will be assumed as a continuous variable due to its description passing to features of the data. Regarding to feature *Year*, the data is treated as ordinal because it is possible to compare how many years have passed until the present. The timeline being studied is 27 years, from 1990 to 2017. The variable *Gas* is categorical due to its description of what type of greenhouse gas is being analyzed (Table 1).

For every value of *Category* and *Gas*, there is an amount of *kt* that each GHG gas emits into the atmosphere and *kt CO_2 equivalent* is the measure to use when comparing the volume of emissions of all GHGs with *CO_2*, since is the worst gas issued in the atmosphere (Table 1). For instance, a global warming potential for methane over 100 years is 28, thus 1 million metric tons of CH_4 is equivalent to 28 million metric tons of carbon dioxide (Greenhouse Gas Protocol, 2015).

World Bank and OECD data description can be seen below.

Table 2 - World Bank data.

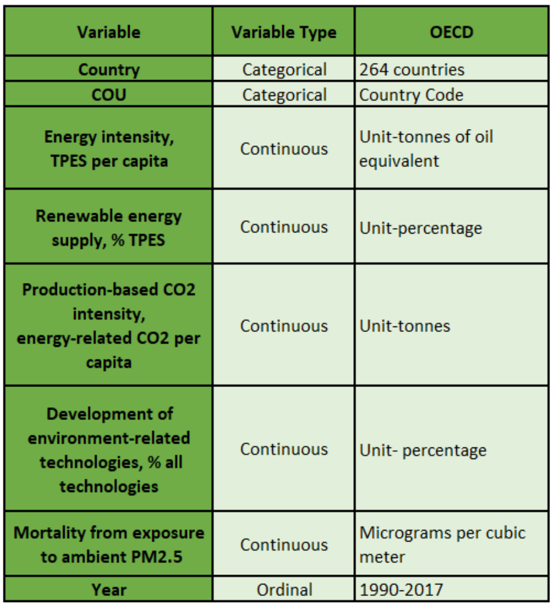


Variable	Variable Type	World Bank
Country	Categorical	264 countries
COU	Categorical	Country Code
Total Population	Continuous	Thousands
GDP Growth(%)	Continuous	Percentage
GDP current(US$)	Continuous	US Dollar
Year	Ordinal	1990-2017

From World Bank it was collected general information about *Total Population, GDP Growth and GDP current (US$)* for 264 countries for 27 years (Table 2). This new evidence it was taken into consideration in order to understand if a country's economic indicators had some kind of influence in the behavior of greenhouse gases emissions.

Country and *COU* are categorical, *Year* is, once again, an ordinal variable and the three new variables are all continuous (Table 2).

Table 3 - OECD data.



Variable	Variable Type	OECD
Country	Categorical	264 countries
COU	Categorical	Country Code
Energy intensity, TPES per capita	Continuous	Unit-tonnes of oil equivalent
Renewable energy supply, % TPES	Continuous	Unit-percentage
Production-based CO2 intensity, energy-related CO2 per capita	Continuous	Unit-tonnes
Development of environment-related technologies, % all technologies	Continuous	Unit- percentage
Mortality from exposure to ambient PM2.5	Continuous	Micrograms per cubic meter
Year	Ordinal	1990-2017

As for OECD, there was an attempt to match the same countries of World Bank with different information for the same countries in OECD. The new variables for analysis are *Energy intensity, TPES per capita* measured in tons of oil equivalent, *Renewable energy supply, % TPES* and *Development of environment-related technologies, % all technologies* are classified as unit percent (Table 3). *Production-based CO_2 intensity, energy-related CO_2 per capita* and *Mortality from exposure to ambient PM2.5* are other features for analysis, and they are measured in tons and micrograms per cubic meter, respectively (Table 3). PM2.5 is a small particle that can be inhaled into the deepest part of the lung and is measured *per* 1 000 000 inhabitants.

As for the final data set to use, some values had to be dropped since they had too many null values and its interesting analysis was not sufficient to carry on with this data. From this step and forward *kt* was not considered a good measure to compare different GHGs emissions and the *kt CO_2 equivalent* was used instead. The values from *Category* were transformed as features since it was assumed to lead to an interesting segmentation analysis.

3.2. Data Preprocessing

3.2.1. Data Cleaning

The information used in the investigation was well structured and, as so, there was almost no need to clean the data.

Regarding the variable *Gas*, its emissions were referred to both aggregated and individual gases. Aggregated values such as Aggregate GHGs, Aggregate F-gases, HFCs and PCFs were removed due to the fact the project aims at a more specific analysis, data about each GHG gas was more helpful.

There is also an interesting note to consider. Assuming that, if at least one country for all 27 years has one feature with no data, these countries would also be dropped. The reason behind this is, when predicting missing values, it is essential to have at least one value to fill the rest of the time series for each country.

These changes led to a final data dimension 8676 rows and 20 columns.

3.2.1.1. Outliers

Outliers' visualization was important to comprehend how features were distributed and if there were any strange values that needed to be removed. For purposes of results comparison, boxplots with and without outliers for each feature were performed.

It was verified for most features, outlier removal led to better outcomes for data distribution (Annex 1). For some cases made no significative difference removing outliers, this is the case for GDP (Annex 2), PM2.5 Mortality (Annex 3), Production CO_2 intensity (Annex 4) and Waste (Annex 5).

3.2.1.2. Missing values

This step is of a big importance for this project for the reason that every year of the time series was demanded to have at least one value for each feature of each country, as was mentioned before. It was performed three methods for handling missing values: no values, linear interpolation and K-Nearest Neighbors (KNN), all this performed in data with and without outliers. When removing all null values, it did not give a viable data set to work with since, for both data sets, it only remained one third of the data. Due to this, the method was discarded and predicting missing values was the best path to have feasible data.

Linear interpolation was an interesting approach since it was a combination of backward fill and forward fill. The reason behind this is, if there is a single value for each feature for a certain country and it's time series, use this value to fill both backward and forward years. It was also performed the most used missing values prediction algorithm, KNN.

In order to have the final data that is going to be used in Data Transformation, an evaluation of the methods used for predicting missing values with and without outliers was crucial.

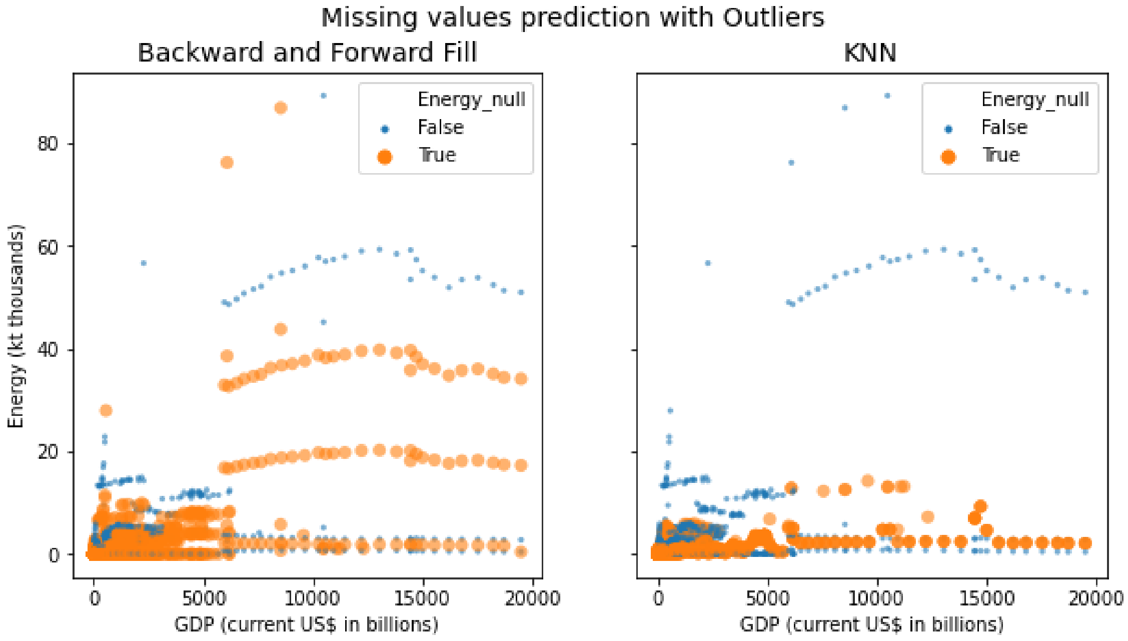


Figure 3 - Missing values with outliers.

The graph above represents the prediction for data with outliers for linear interpolation and KNN. The blue color represents the real data values whereas the orange color represents the predicted values. It can be understood that linear interpolation is closer to the real values and KNN was not a good prediction algorithm for the research (Figure 3).

As for missing values prediction without outliers, the results can be seen below.

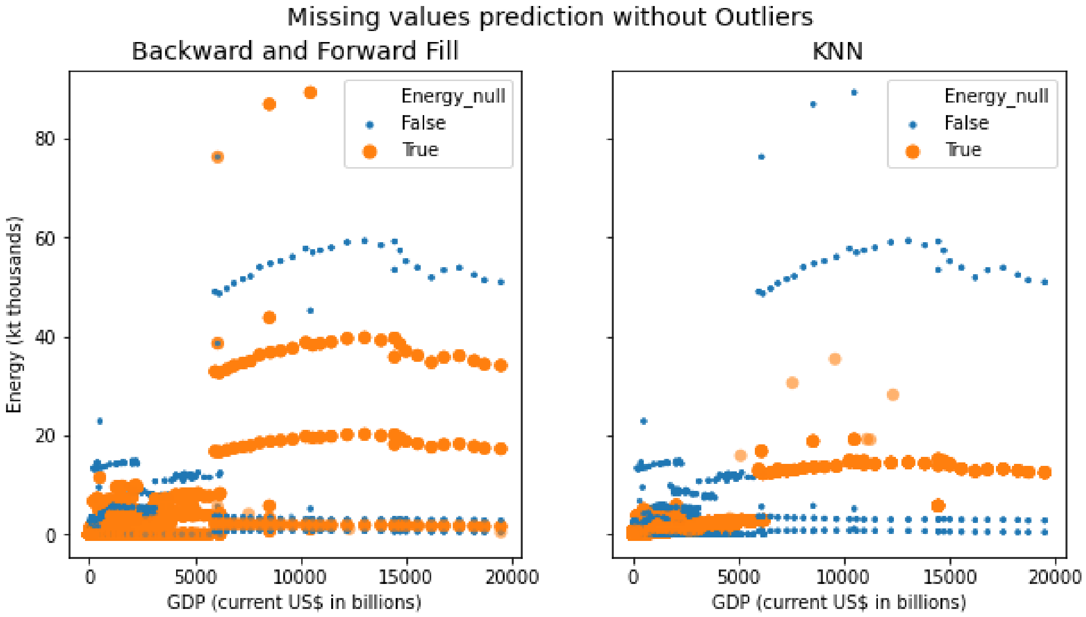


Figure 4 - Missing values without outliers.

It can be easily seen that linear interpolation also led to better results for data without outliers, unlike KNN. When comparing outcomes with and without outliers for linear interpolation, there are some values on data without outliers that were predicted the same, thus making this data set the chosen one to continue the investigation (Figure 4).

3.3. Data Transformation

For this research, Data Transformation mainly consisted of data normalization and data reduction techniques. It was introduced a new feature, *Nº years*, resulting in the number of years that have passed since 1990 in order to understand if this variable would be interesting to evaluate. There was an effort to add new variables in the data set but it led to very correlated variables or not meaningful ones. Even though it makes sense to remove Total Population and GDP for GDP *per capita*, this approach was not done assuming these features would be of a good input for understanding the final data.

3.3.1. Data normalization

Data normalization was necessary since all variables had different measures and was not possible to compare them, so the data was transformed into comparable variables belonging to a range between 1 and -1.

When visualizing a correlation matrix, it was understood there were a lot of variables either directly or inversely correlated to each other. Features that had a correlation higher than 0 were directly correlated and features which had a correlation lower than 0 were inversely correlated. The map can be seen below.

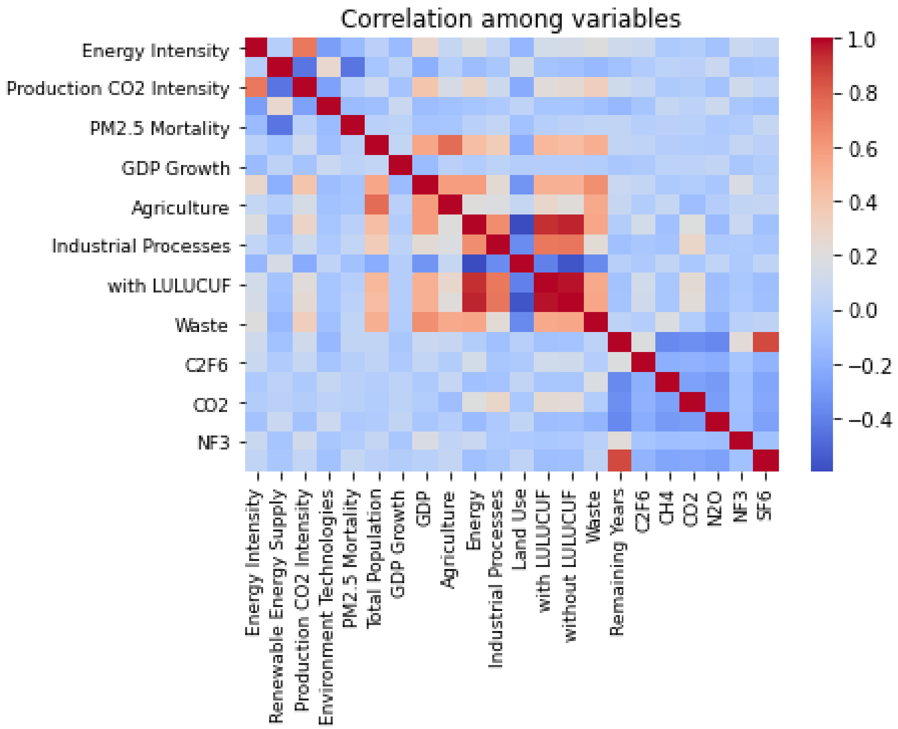


Figure 5 - Variables correlation.

Variables with positive correlation are the ones nearest to 1 whereas features with negative correlation are the ones closest to -1. A positive correlation means that both variables move in the same direction and negative correlation are variables that move in opposite directions (Figure 5).

The following are the features with **positive correlation** (Figure 5):

- Energy intensity with Production CO_2 Intensity;
- Energy with total GHGs emissions with and without LULUCF;
- Industrial Processes with Energy and LULUCFs;
- Total GHGs emissions with LULUCF with total GHGs emissions without LULUCF;
- Total Population with LULUCFS, Agriculture and Waste;

- GDP with LULUCFS, Energy and Waste.

The following are the features with **negative correlation** (Figure 5):

- Renewable Energy Supply with Production CO_2 Intensity and PM2.5 Mortality;
- Energy with Land Use;
- Land Use with total GHGs emissions without LULUCF and Waste.

3.3.2. Data Reduction

After understanding the heatmap, it was realized that the data set should be divided into two distinguished data sets, *Country Characteristics* and *Emissions*. The reason for this was to have a number of correlated features in order to have a transformed number of uncorrelated variables for Data Reduction (Hui, 2019).

Emissions contained *Agriculture, Energy, Industrial Processes, Land Use, Total GHGs emissions with* and *without LULUCF, Waste* and *Nº Years*. As for *Country Characteristics*, the data set incorporated *Energy Intensity, Renewable Energy Supply, Production CO_2 Intensity, Environment Technologies, PM2.5 Mortality, Total Population, GDP Growth* and *GDP.* In short, *Emissions* had all values related to GHGs emissions and *Country Characteristics* had values associated to a country's development regarding features that are important for the research.

3.3.2.1. Principal Component Analysis

The use of Principal Components for data reduction was implemented since it would allow to understand which features with high correlation reached to uncorrelated components among them (Loukas, 2020). Principal Component Analysis (PCA) can only be employed on continuous features (Walker, 2019) which led to a discard for this analysis of CO_2, N_2O, CH_4, NF_3, SF_6 and C_2F_6 on *Emissions* since they are categorical.

The choice of number of components for *Emissions* and *Country Characteristics* was based on each PCA cumulative explained variance and a target of at least 85% of the data explained by the first components (Annex 15 and Annex 16). In this way, a PCA with 4 components and with 3 components was performed on *Country Characteristics* and *Emissions*, respectively.

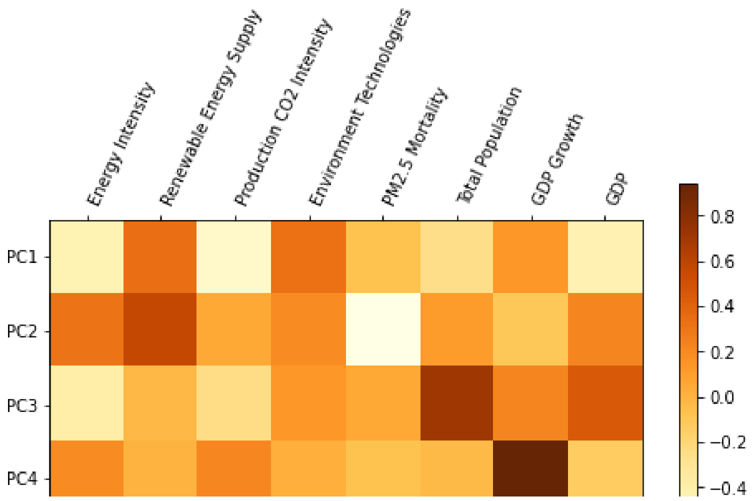


Figure 6 - Principal Components for Country Characteristics.

Regarding *Country Characteristics*, the first Principal Component (PC) consisted in data from *Renewable Energy Supply, Environment Technologies* and *GDP Growth*. PC2 has *Energy Intensity* and *Renewable Energy Supply* whereas PC3 has *Total Population* and *GDP*. There is only one variable that belongs to PC4, *GDP Growth* (Figure 6).

It seems the first component considers how producing renewable energy and implementation of environment technologies might be related to a growth on a GDP country while the second, is a component in which renewable energy might have a big share in energy intensity. The third component can be understood as GDP *per capita* and the last one, it only stood out GDP growth (Figure 6).

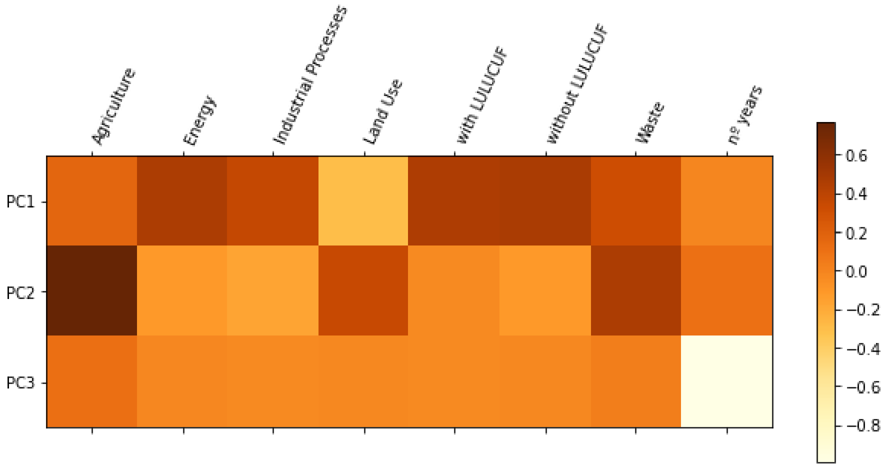


Figure 7 - Principal Components for Emissions.

The first PC for *Emissions* is composed with *Energy, Industrial Processes, LULUCFs* and *Waste*, the second has *Agriculture* and PC3 did not give conclusive results. PC1 seems to be a component more related to the secondary sector and PC2 connected to the primary sector (Figure 7).

It is visible for both *Country Characteristics* and *Emissions*, how Principal Components places data with high variability into much more uncorrelated features (Figure 8; Figure 9).

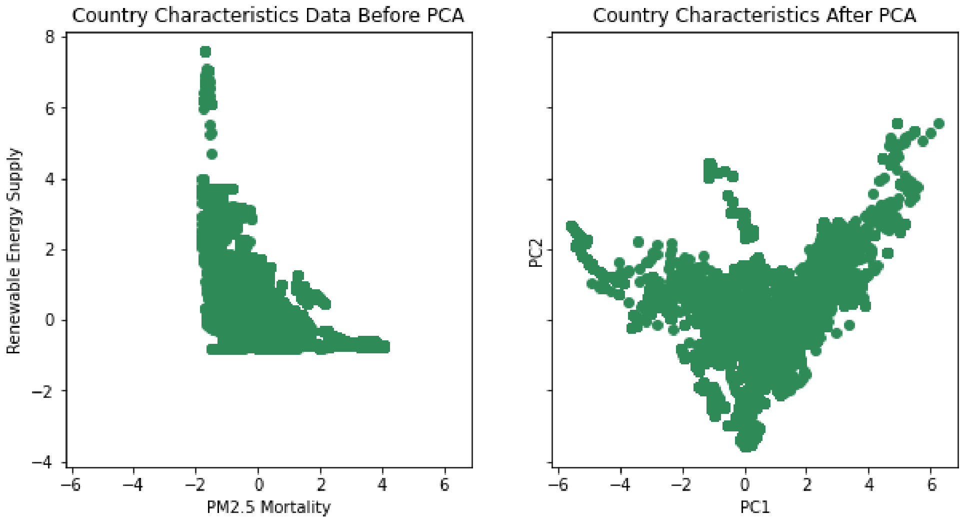


Figure 8 - Country Characteristics data before and after PCA.

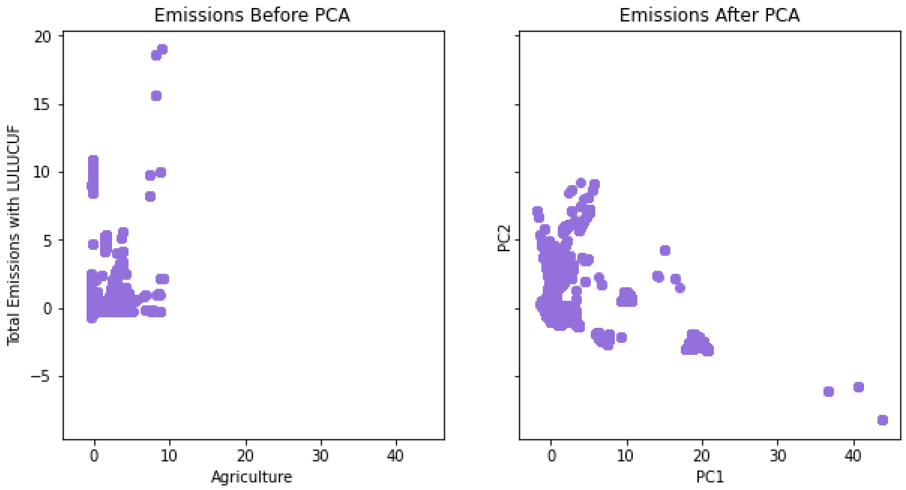


Figure 9 - Emissions data before and after PCA.

3.3.2.2. Factor Analysis

A Factor Analysis (FA) was also tested since it can handle categorical variables (UCLA Statistical Consulting Group, 2016). In order to do so, it was combined two tests to understand if the data was feasible for a FA, **Kaiser-Meyer-Olkin (KMO)** and **Bartlett's test.** If KMO had a result above 0.50 and Bartlett’s test with a significance level below 0.05, the data could be used for FA (IBM Corporation, 2018).

Country Characteristics outcomes were 0.47 for KMO and 0.0 for Bartlett’s test whereas for *Emissions*, KMO generated 0.59 but for Bartlett’s test it did not give any results. This led to conclude that neither of the data was viable to perform a Factor Analysis.

3.3.2.3. Self-Organizing Map

Even though Self-Organizing Map (SOM) is most used as a Data Mining technique, there was no information available on how to do it with Python technology but instead use it as data reduction. The use of SOM on both *Country Characteristics* and *Emissions* was an interesting idea due to the fact it was possible to put the feature *Year* as a target for dimensionality reduction.

Country Characteristics

It was implemented a network of 9x9 dimensions since the data contained 9 features. The network was trained with *Periodic Boundary Conditions* activated, for 10.000 epochs and with a learning rate of 0.1.

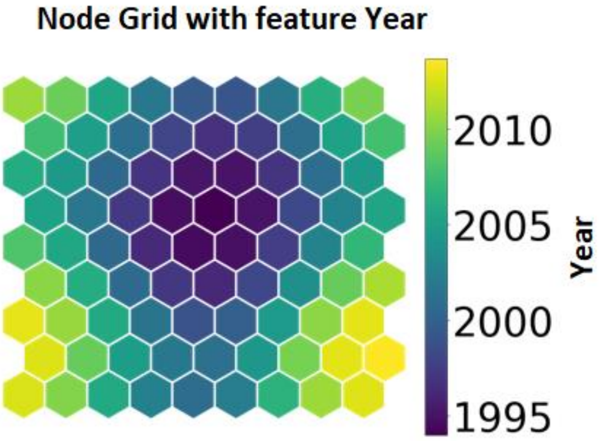


Figure 10 - Self-Organizing Map for Country Characteristics.

Taking into consideration that the darker parts of the outcome represent a cluster, while the lighter ones represent the division from clusters, it was concluded the SOM performance gave one cluster (Figure 10). Below, it is demonstrated how data was distributed before and after SOM, and it can be understood how the data reduction technique was well distributed after SOM (Figure 11).

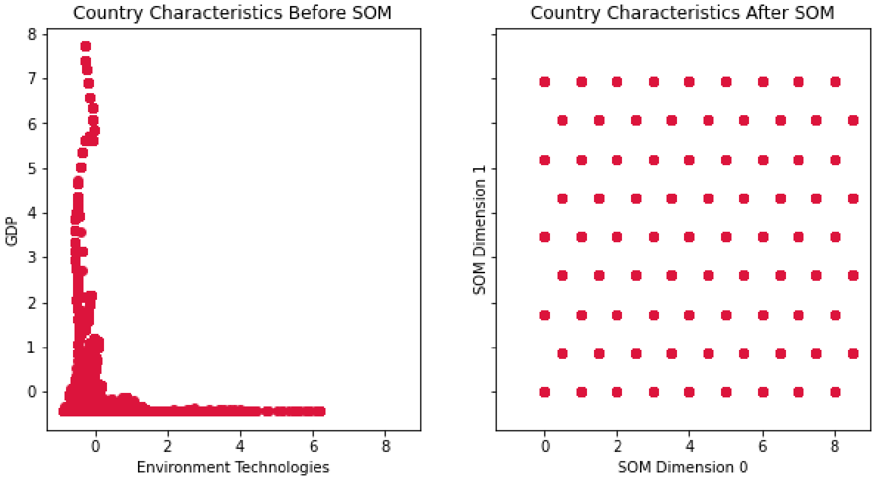


Figure 11 - Country Characteristics data before and after SOM.

Emissions

It was performed a network of 15x15 dimensions since the data had 15 features. The network was trained with *Periodic Boundary Conditions* activated, for 10.000 epochs and with a learning rate of 0.1.

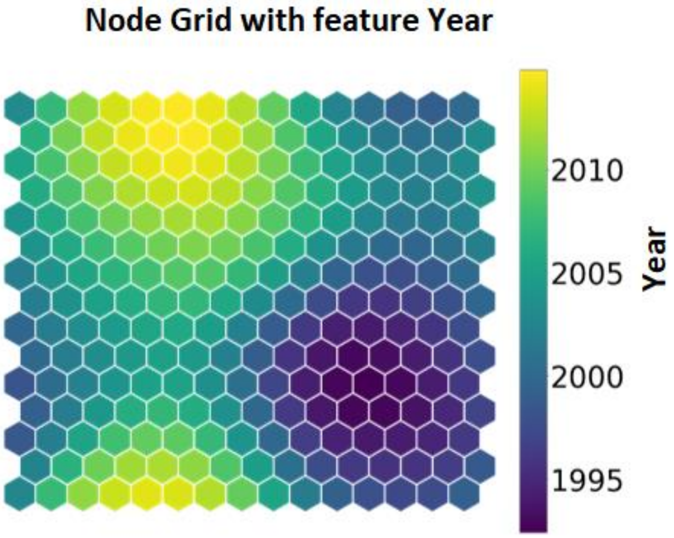


Figure 12 - Emissions Self-Organizing Map.

As assumed previously, the lighter parts of SOM results are clusters partition so, it seems to have clustered three groups for *Emissions* (Figure 12). Below, it is demonstrated how data was distributed before and after SOM, once again, data after SOM was more dispersed (Figure 13).

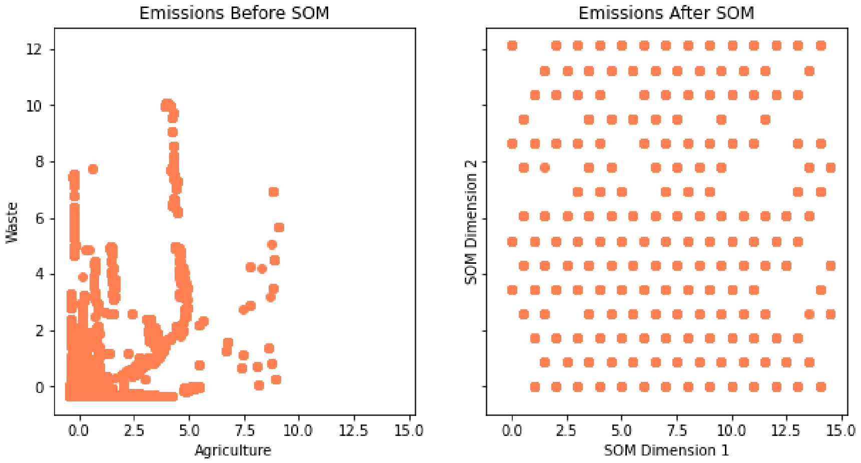


Figure 13 - Emissions data before and after SOM.

Even though using SOM as a dimensionality reduction technique distributed better the data than before, comparing with PCA results it gives different insights. When it comes for *Country Characteristics*, PCA with 4 components gave valuable and interesting information instead of a not so conclusive SOM data. As for *Emissions*, the results given by PCA were not good to carry on with the analysis since using SOM was more interesting.

After understanding which Data Reduction technique gave better results it was concluded PCA worked better for *Country Characteristics* and SOM enhanced *Emissions* data.

4. RESULTS AND DISCUSSION

4.1. DATA MINING

4.1.1. Clustering

Taking into consideration this is the most important step of the research, the chosen *clustering* method would lead to a more insightful analysis, the expectations for good outcomes were high. Thus, requiring different clustering techniques to compare. In order to do so, it was implemented the five most common clustering algorithms: **K-means, Hierarchical, DBSCAN, Mean-Shift** and **Expectation-Maximization (EM)**.

4.1.1.1. Emissions

The poorest algorithm performance was DBSCAN since it gave 3.412 clusters (Annex 17). This was not considered a reliable cluster number to do a proper analysis.

Mean-Shift with 4 clusters was also an unsuccessful algorithm, it can be understood that data mostly belongs to cluster 2 and 3 (Annex 18). Cluster 1 is bad at grouping similar values and also, data is not well partitioned as Cluster 3 shows some overlapping with other clusters.

The selected clusters numbers for hierarchical clustering was achieved by representing a dendrogram (Annex 19) thus, making 2 clusters as the best outcome. Nevertheless, this partitioning procedure gave bad results since most data seems to belong to cluster 1 and, cluster 0 suffers overlapping of the first (Annex 20).

As for the set-up of EM clusters number, it was established (for all data sets) as the same clusters number as K-means since EM is a more general technique than K-means. When comparing both K-means with 3 clusters and EM with 3 clusters, EM seems to group the data into more homogenous groups whereas in K-means, cluster 2 overlapped the others (Annex 21 and Annex 22). Data distribution with K-means mostly belonged to cluster 1 and with EM data most fitted in cluster 1 (Annex 21 and Annex 22).

EM seems to group the data into more homogenous groups whereas K-means partitioning, cluster 2 overlapped the others, as observable in Land Use and Agriculture (Annex 21 and Annex 22). So, EM was the chosen algorithm for *Emissions*.

The chosen algorithm for *Emissions* was Expectation-Maximization but this was not still sufficient to carry on with *Emissions* analysis. Clustering *Emissions* without dimensionality reduction was crucial in order to understand if it gave better results than data reduction with SOM.

4.1.1.2. Emissions without dimensionality reduction

Since *Emissions* with no dimensionality reduction (DR) possessed *dummy* variables, besides performing five clustering algorithms as mentioned it was also implemented **K-modes** for the categorical variables. As a result of K-modes being a variation of K-means, it was established as the same clusters number as K-means.

DBSCAN and Mean-Shift were the worst performers, giving 191 and 43 clusters, respectively (Annex 23 and Annex 24). These were an unacceptable number of clusters due to not grouping homogenous values. EM with 2 clusters (Annex 25) was also a bad outcome as it was seen that almost all data belonged to cluster 1.

K-means with 2 clusters and Hierarchical with 2 clusters (Annex 26, Annex 27, Annex 28), gave the exact same results thus, leading to Hierarchical exclusion since K-means provides a better performance and easy implementation. For both algorithms, most data resided in cluster 1 (Annex 26 and Annex 28).

Comparing K-means with K-modes with 2 clusters, it was visible that K-means grouped better more homogenous data whereas cluster 1 in K-modes, overlapped some data belonging to cluster 0, as it can be seen with variables Land Use and Agriculture (Annex 29).

In order to understand which results, from *Emissions* with SOM and *Emissions* without dimensionality reduction, were better, a comparison was made between the winning algorithms for both approaches. It was concluded K-means, for *Emissions* without dimensionality reduction, gave the best group of similar values, as understood with variables Energy and Agriculture (Annex 22 and Annex 25).

4.1.1.3. Country Characteristics

Once again, DBSCAN was a bad algorithm for the research due to presented 92 clusters which was not good for data understanding (Annex 30). Hierarchical with 2 clusters (Annex 31 and Annex 32) and Mean-Shift with 5 clusters were also poor performers, as the data was not grouped by similar values (Annex 33).

Between EM with 4 clusters (Annex 34) and K-means with 4 clusters (Annex 35) was a tight decision since it gave identical outcomes but there were some features, such as Energy Intensity and Renewable Energy Supply, where K-means had a better distribution for each cluster.

4.1.1.4. Conclusion

When comparing both results for *Emissions*, it is easily understood that clustering without dimensionality reduction, for its segmentation gave better results than *Emissions* performed with SOM (Figure 14; Figure 15; Figure 16; Figure 17; Figure 18; Figure 19; Figure 20;Figure 21). Thus, the chosen final clustering for *Emissions* is partitioning without dimensionality reduction, K-means with 2 clusters.

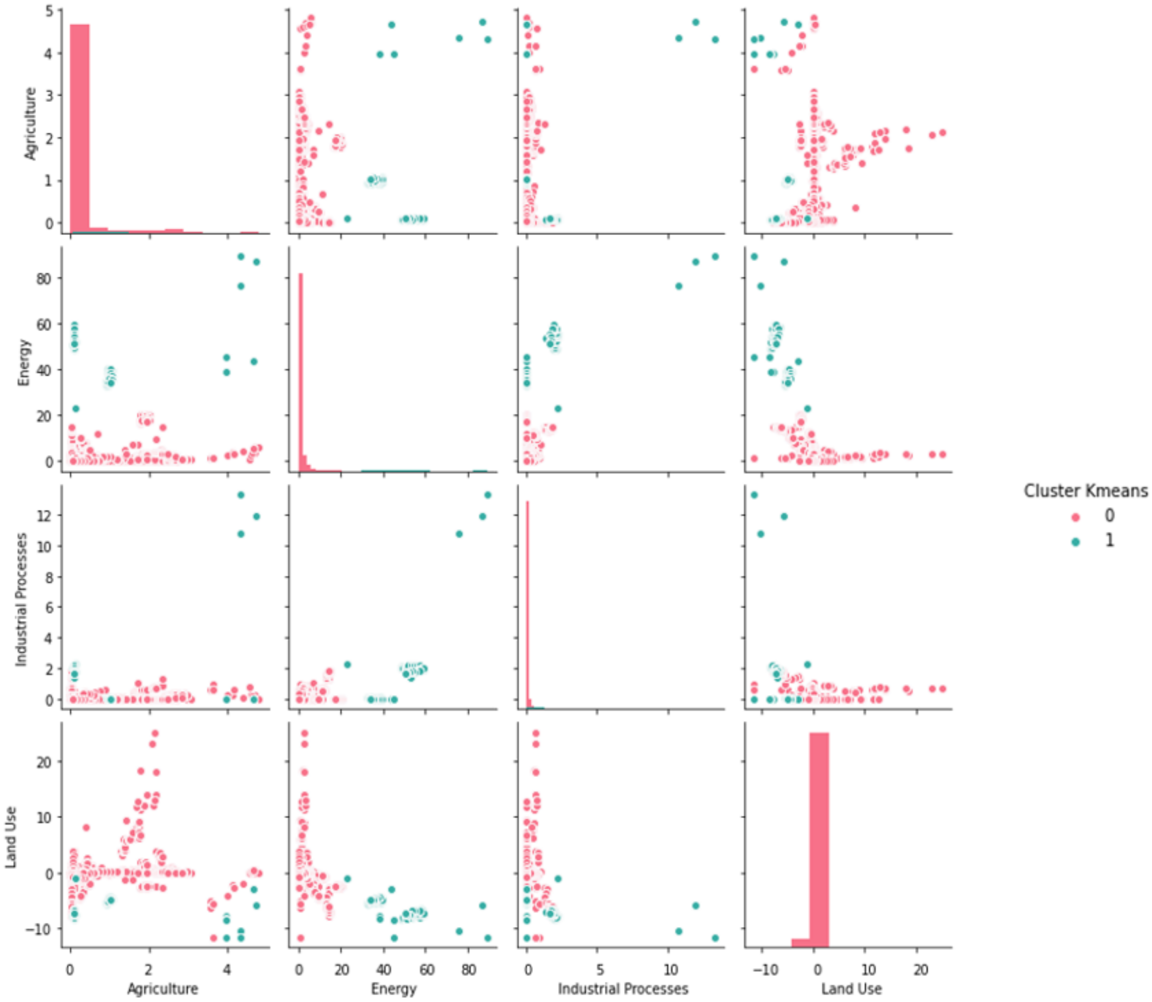


Figure 14 - Emissions clustering without dimensionality reduction (Part I).

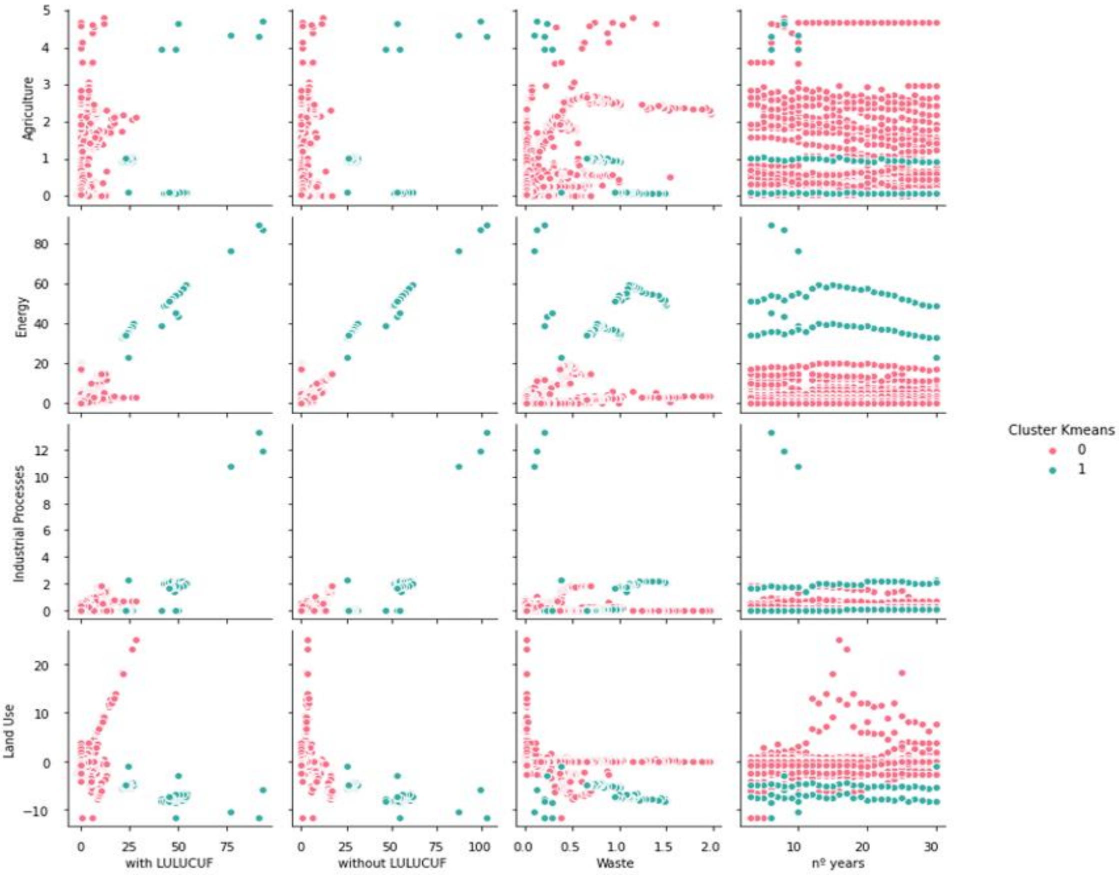


Figure 15 - Emissions clustering without dimensionality reduction (Part II).

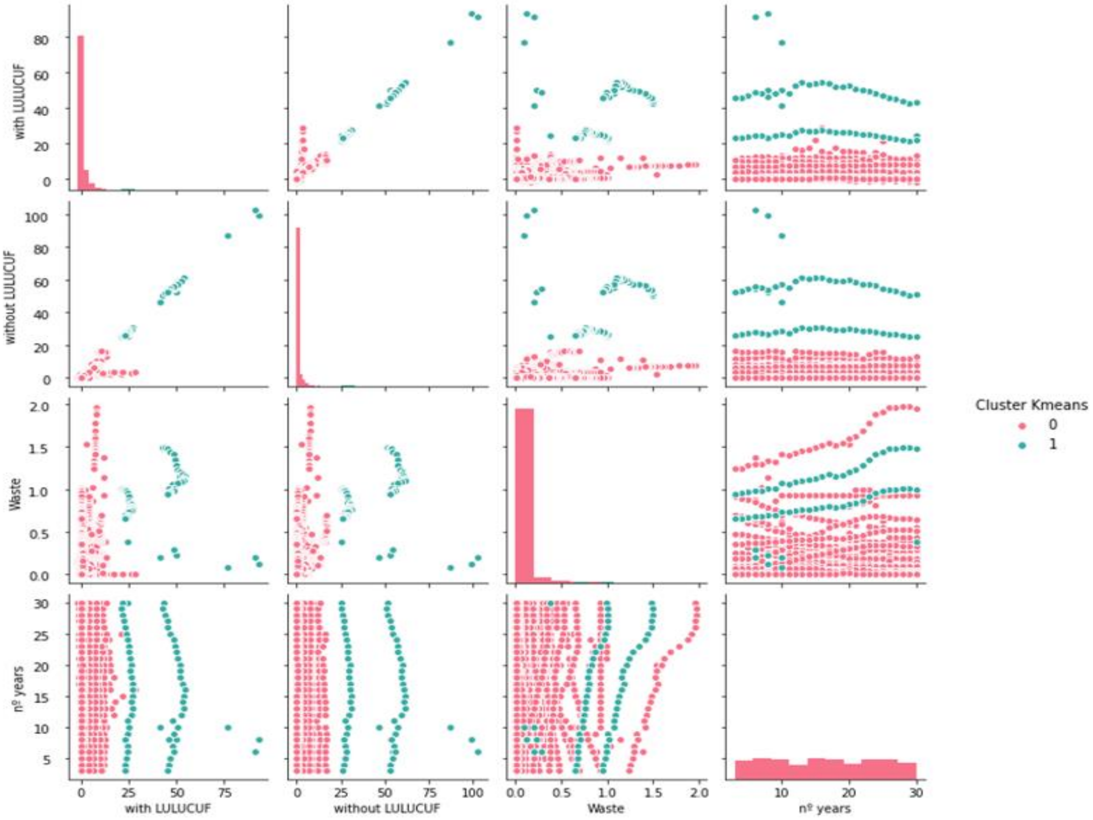


Figure 16 - Emissions clustering without dimensionality reduction (Part II).

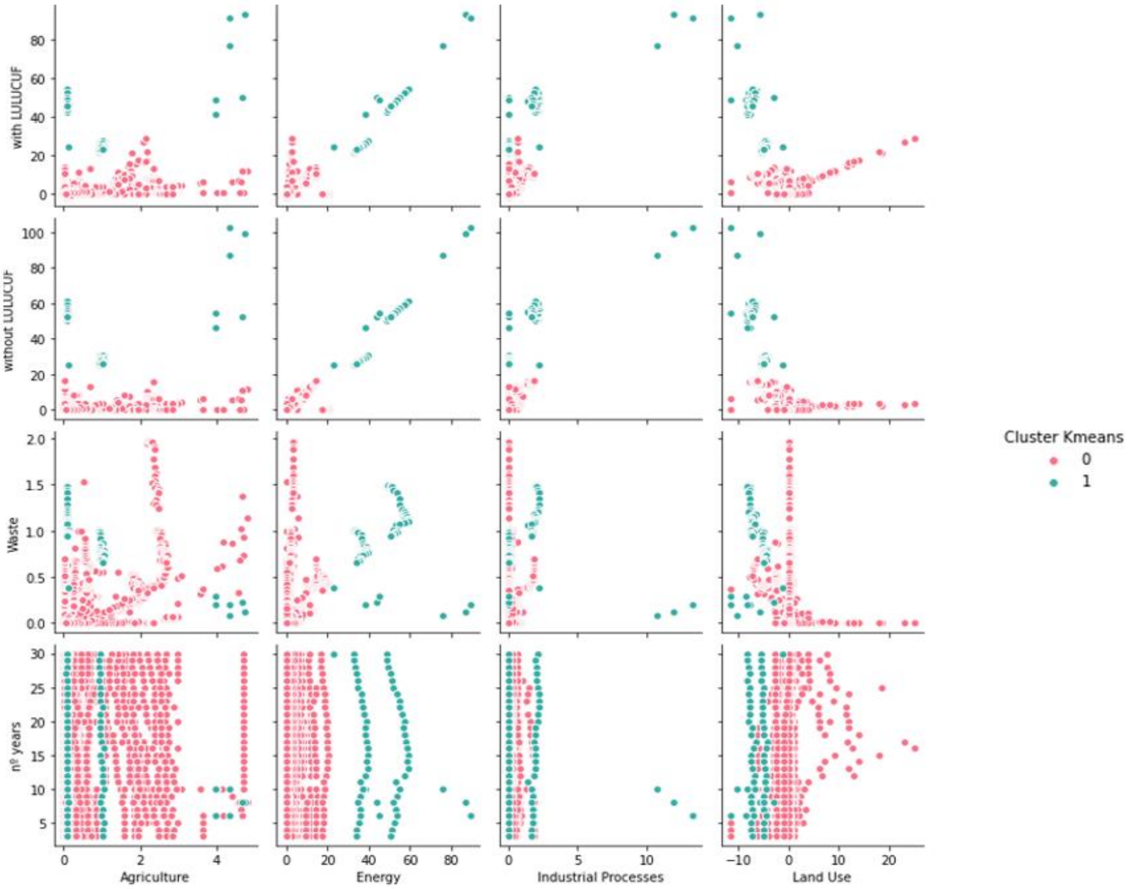


Figure 17 - Emissions clustering without dimensionality reduction (Part IV).

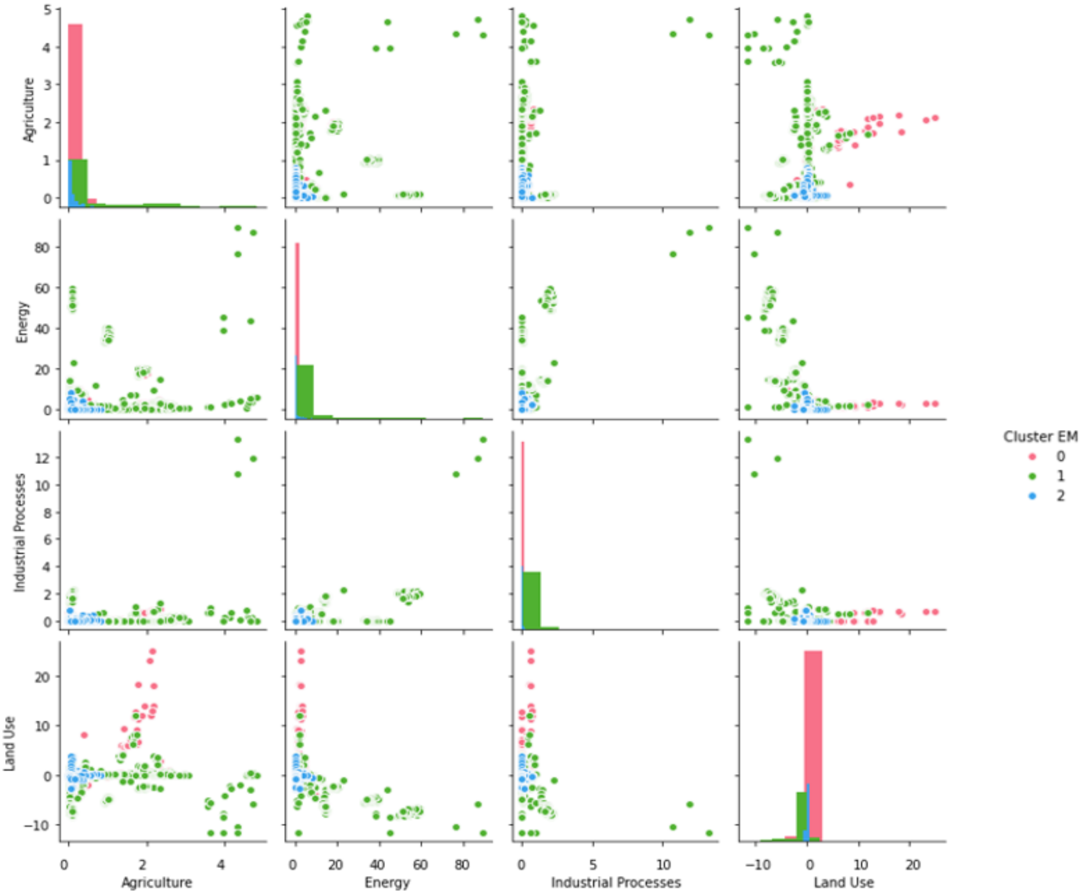


Figure 18 - Emissions clustering with dimensionality reduction (Part I).

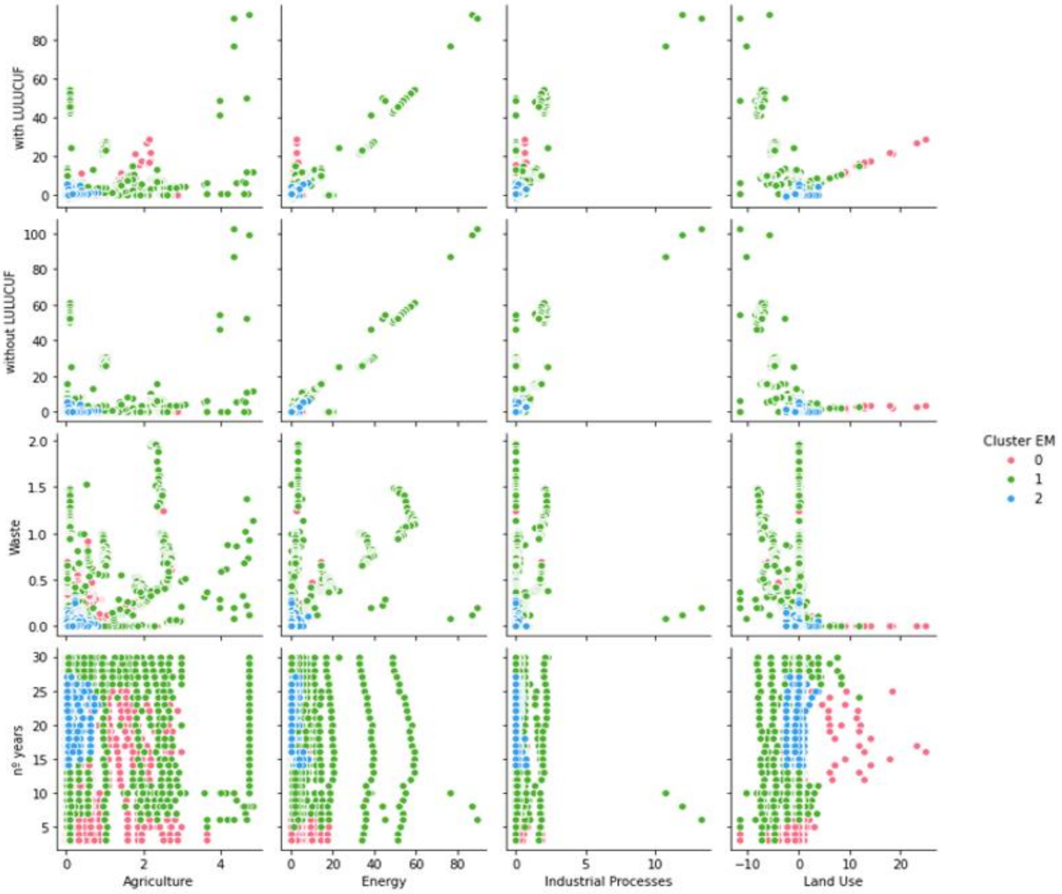


Figure 19 - Emissions clustering with dimensionality reduction (Part II).

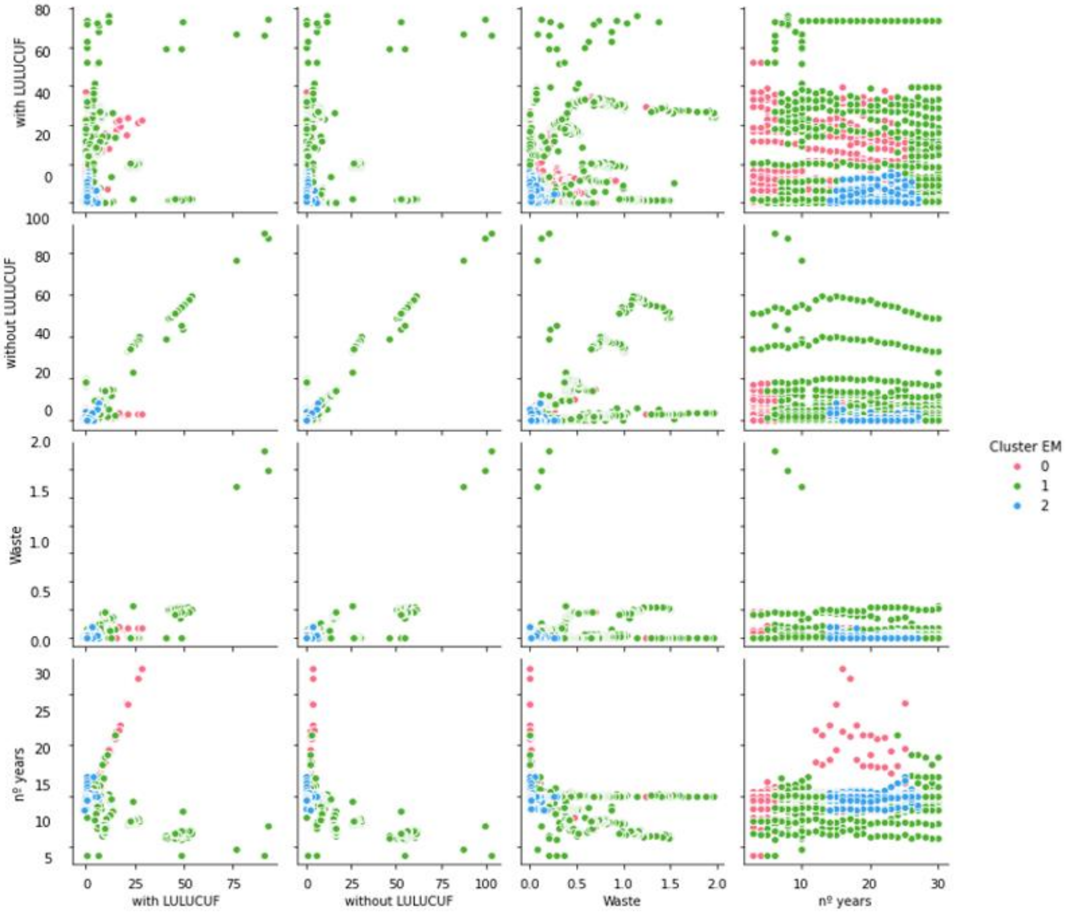


Figure 20 - Emissions clustering with dimensionality reduction (Part III).

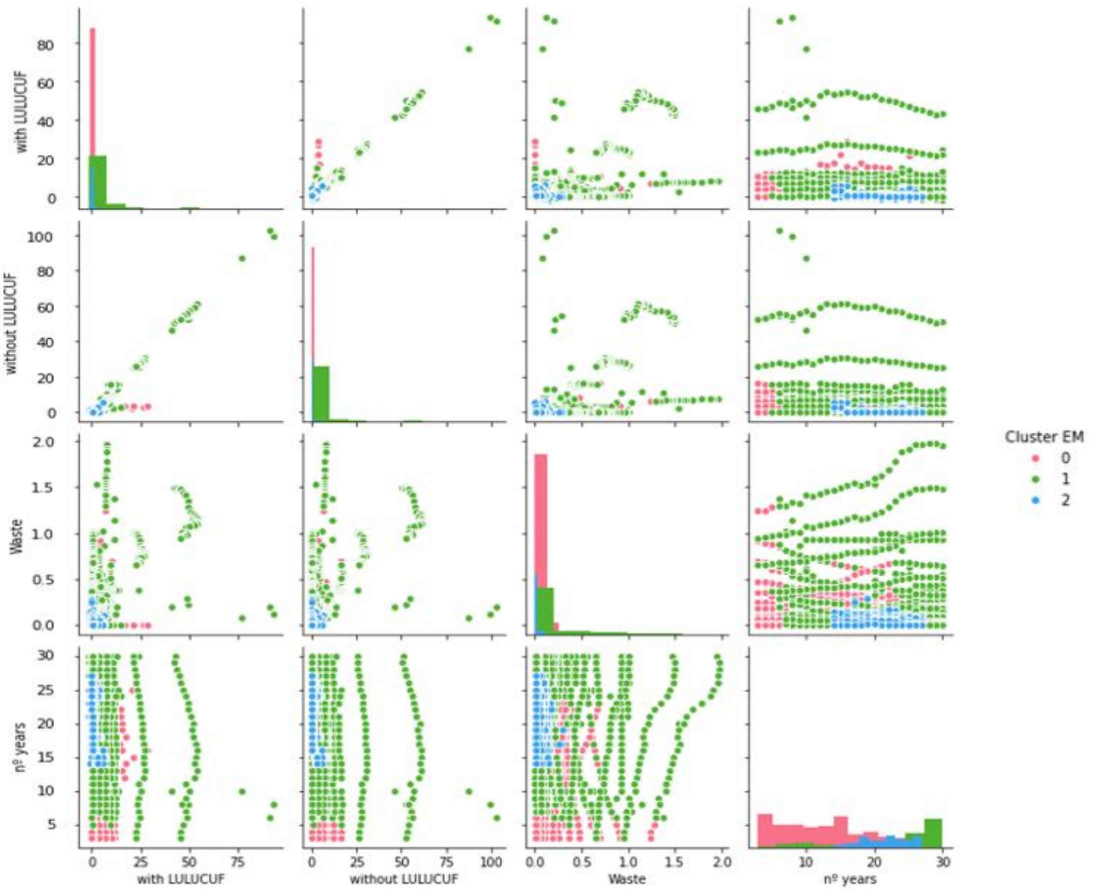


Figure 21 - Emissions clustering with dimensionality reduction (Part IV).

Regarding *Country Characteristics*, the selected algorithm is also K-means with 4 clusters (Figure 22; Figure 23; Figure 24; Figure 25).

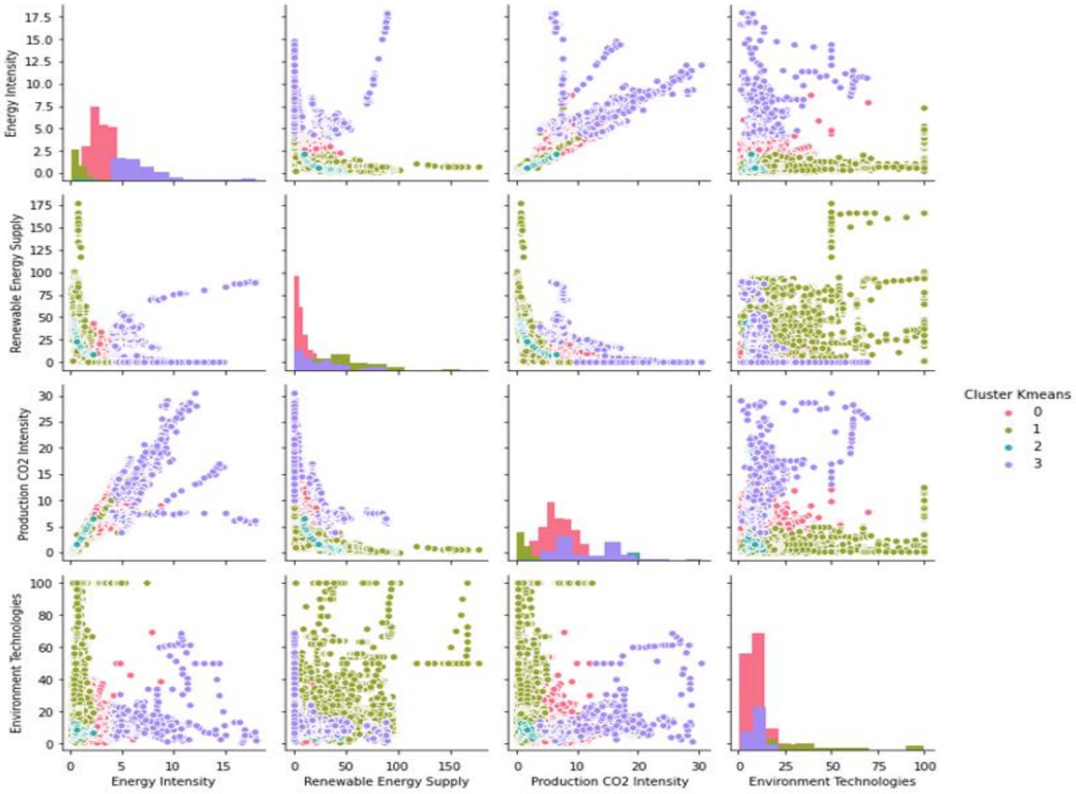


Figure 22 - Country Characteristics clustering (Part I).

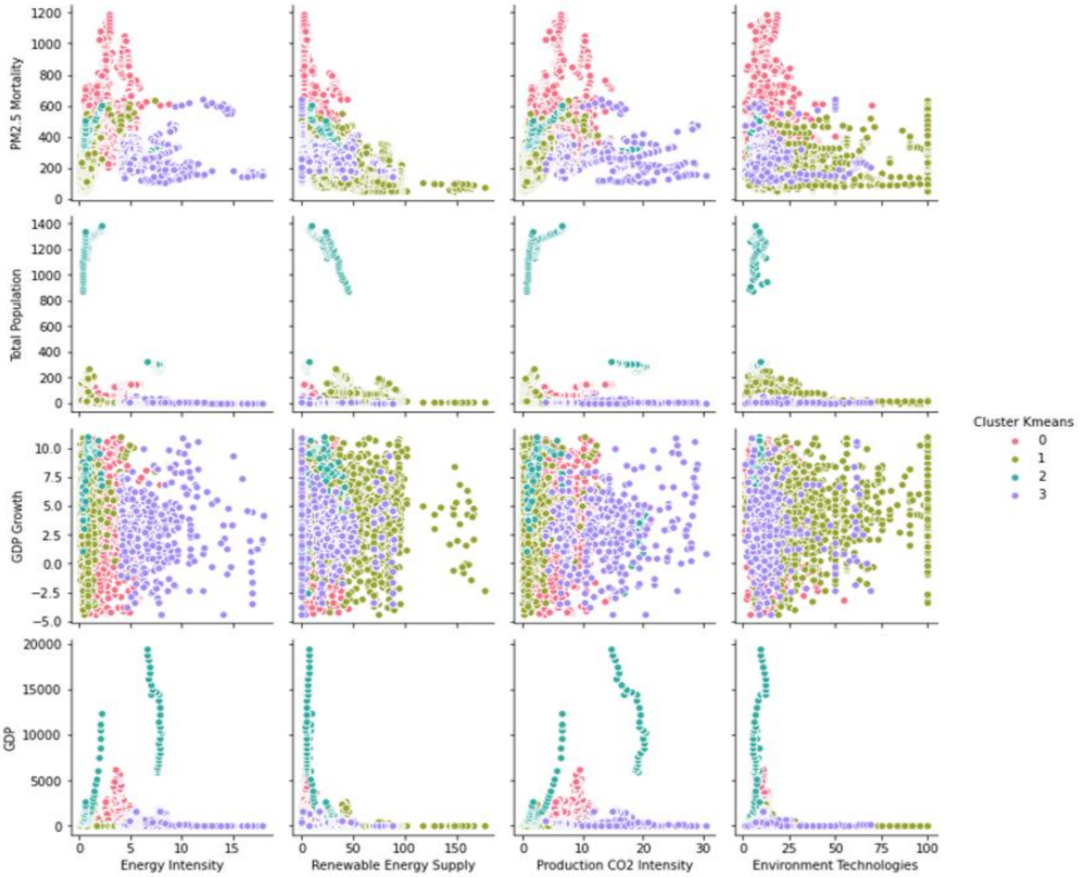


Figure 23 - Country Characteristics clustering (Part II).

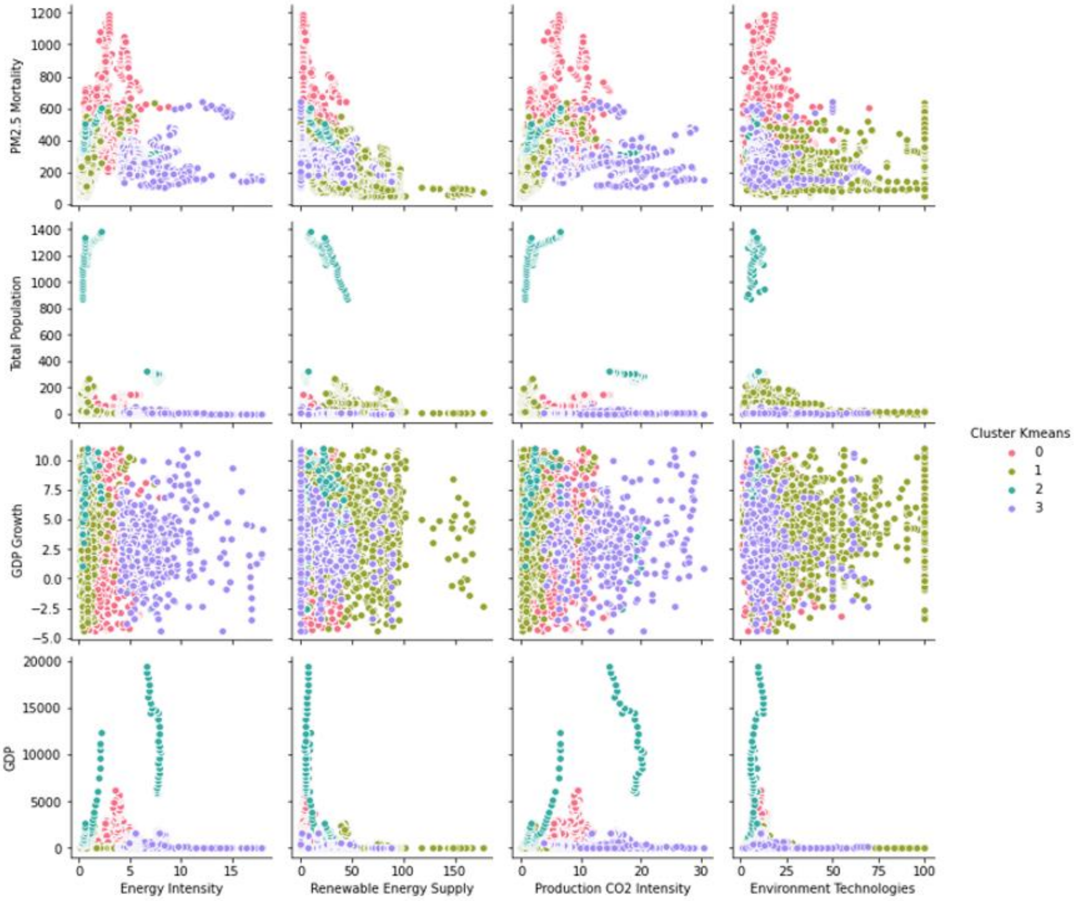


Figure 24 - Country Characteristics clustering (Part II).

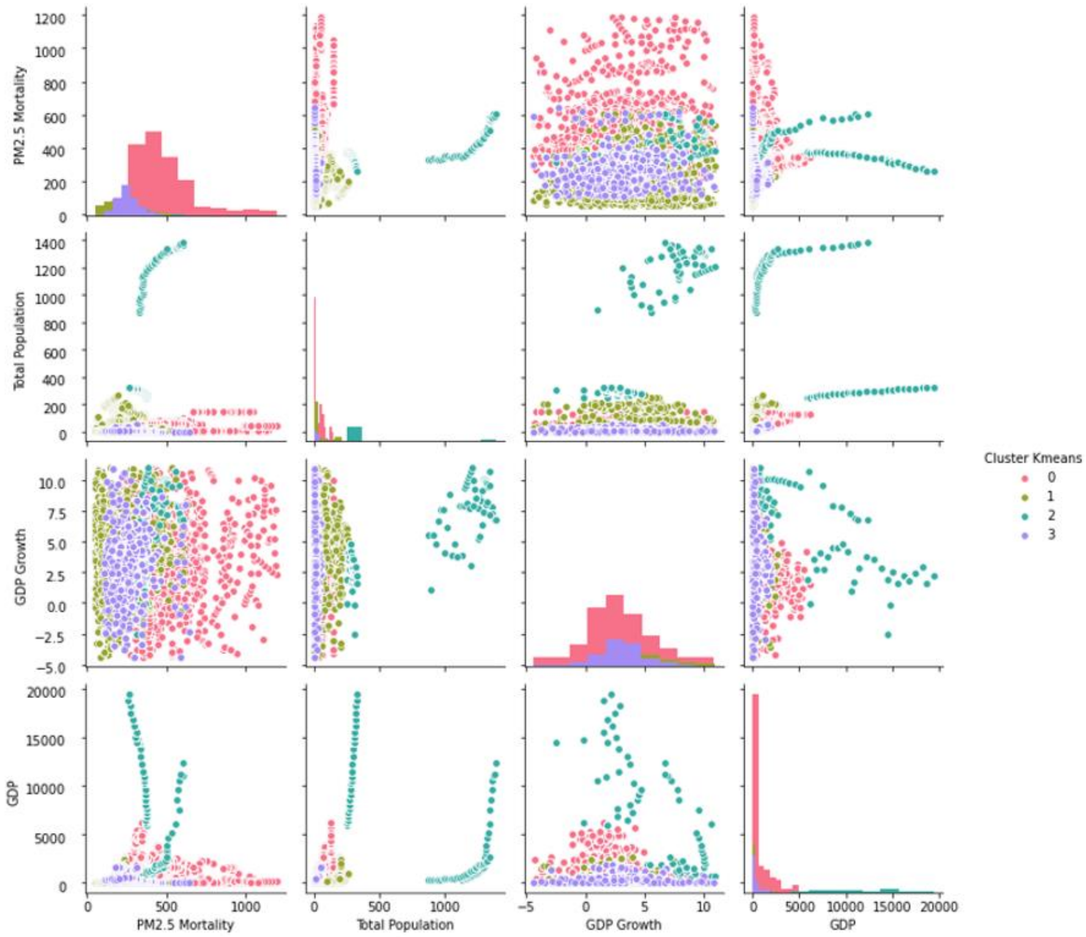


Figure 25 - Country Characteristics clustering (Part IV).

4.2. DATA INTERPRETATION

Data Interpretation is one of the most important steps of the KDD process since it is where is gained valuable insights for the study. It will be explained the cluster conclusions reached for both data sets, *Emissions* and *Country Characteristics*, separately, and also for the combination of both data done through a contingency table.

Regarding *Emissions*, it will be analyzed GHGs emissions (kt CO_2 equivalent) by category (Energy, Industrial Processes, Land Use, Waste GHGs with LULUCF and GHGs without LULUCUF) across 27 years. As for *Country Characteristics*, for the same time series, it will be evaluated Energy Intensity, Renewable Energy, Production-based CO2 Intensity, Development of Environment-related Technologies and Mortality from Exposure to Ambient PM2.5 in its respective unit measures.

Finally, it will be considered four years of the time series (1990-2017) to represent clusters in the world map for *Emissions* and *Country Characteristics*. The chosen years selected to analyze are:

- 1990: beginning of the time series;
- 1995: UNFCCC (1994);

- 2006: Kyoto Protocol (2005);
- 2017: Paris Agreement (2016) and final year of the time series.

The final three years suffer a lag of one year from its implementation, for purposes of a better analysis.

4.2.1. Emissions

After carefully understood the behavior of GHGs emissions through time, it was concluded that Energy, GHGs with LULUCF and without LULUCF are the categories that most contribute to the rise of GHG emissions (Figure 26). Agriculture, Industrial Processes and Waste are categories that almost do not produce GHGs emissions and Land Use do not produce harmful emissions (Figure 26).

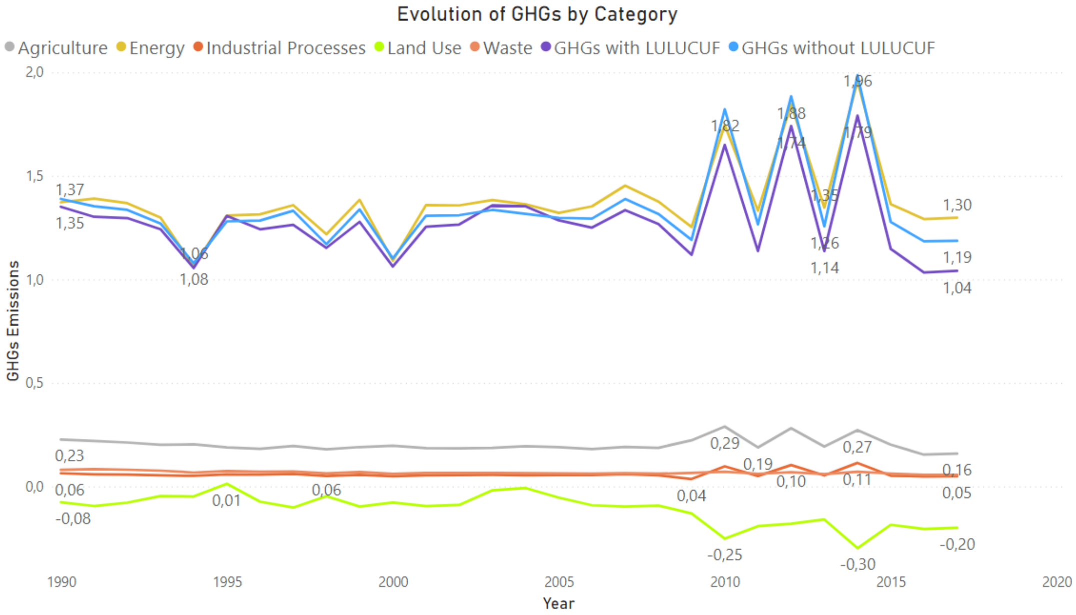


Figure 26 - Evolution of GHGs by Category.

GHGs emissions were constant from 1990 to 2009 but, since 2009 until 2014 they were very volatile, reaching the peak of emissions in 2014 with 1,96 kt CO_2 equivalent for Energy and GHGs without LULUCF and, 1,79 kt CO_2 equivalent (Figure 26). As expected, CO_2 is the gas that most contributes to the rise of emissions, as it can be seen, its values for Energy and GHGs with and without LULUCF are above the average of the total emissions (Figure 27). CO_2 is not the only one contributing for these high values as CH_4 and C_2F_6 do their fair share of molding the GHGs emissions (Annex 36, Annex 37).

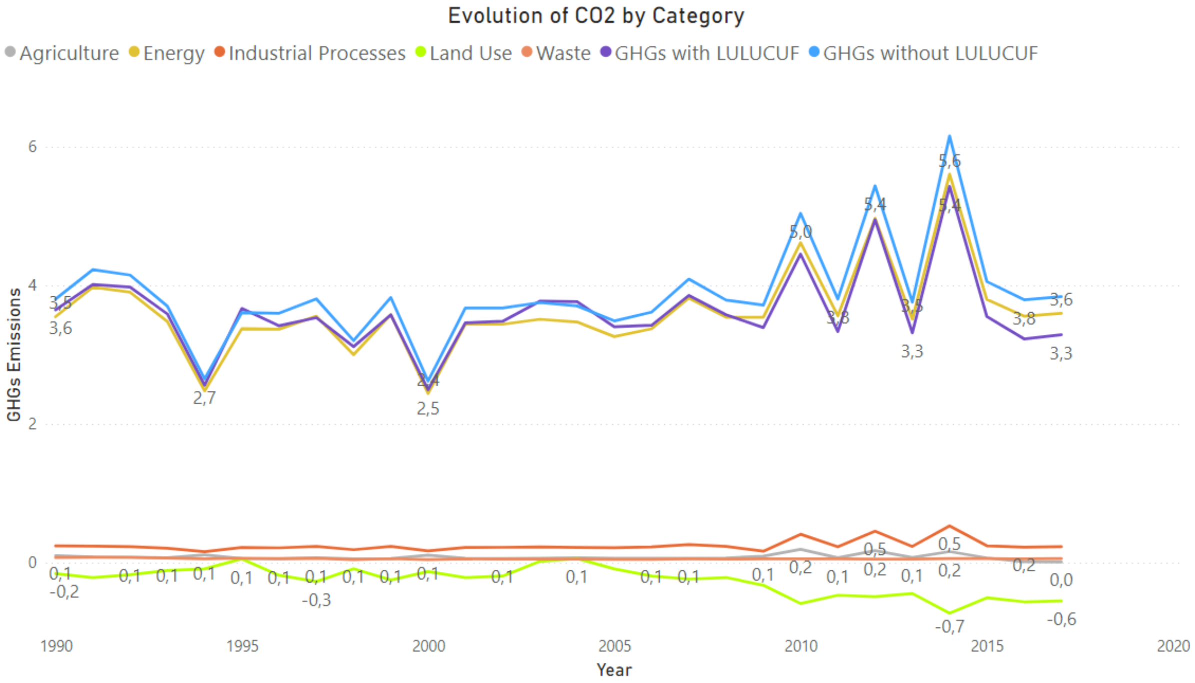


Figure 27 - Evolution of CO2 by category.

There were two clusters belonging to *Emissions*, for **cluster 0** there were 31.753 observations and for **cluster 1** only 371 occurrences. **Cluster 0** represents countries that emit all GHGs but mostly are shaped by C_2F_6 and CO_2 (Annex 38 and Annex 39) and **Cluster 1** are countries only by C_2F_6 and CO_2. Even though **cluster 0** is the one with most observation, it is clearly stated **cluster 1** are the observations that shape *Emissions* (Figure 28; Figure 29).

As it can be seen, from the graphs below, the difference between clusters is in the amount of emissions produced by each category. **Cluster 0** has low emissions values for Energy, Industrial Processes, Waste and GHGs with and without LULUCF but for Land Use, emissions are higher than for **cluster 1** (Figure 28; Figure 29). As for **cluster 1** it is clearly stated that emissions for Energy, Industrial Processes, Waste and GHGs with and without LULUCF but for Land Use are much higher than for **cluster 0** (Figure 28; Figure 29).

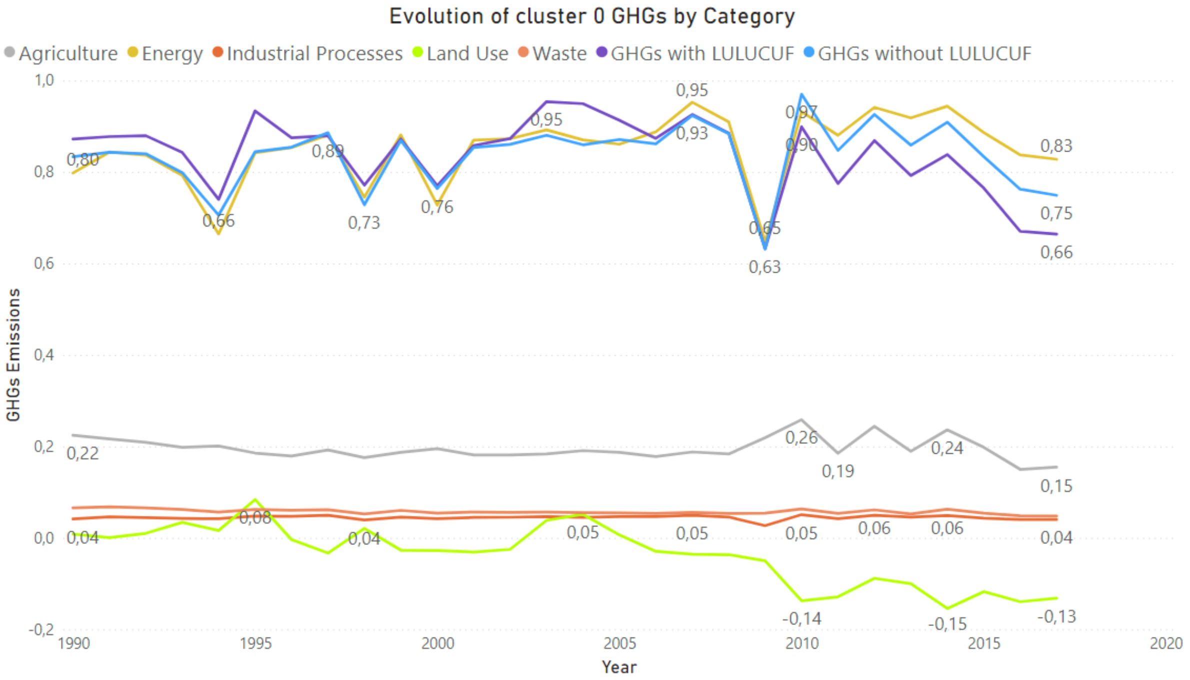


Figure 28 - Evolution of cluster 0 GHGs by category.

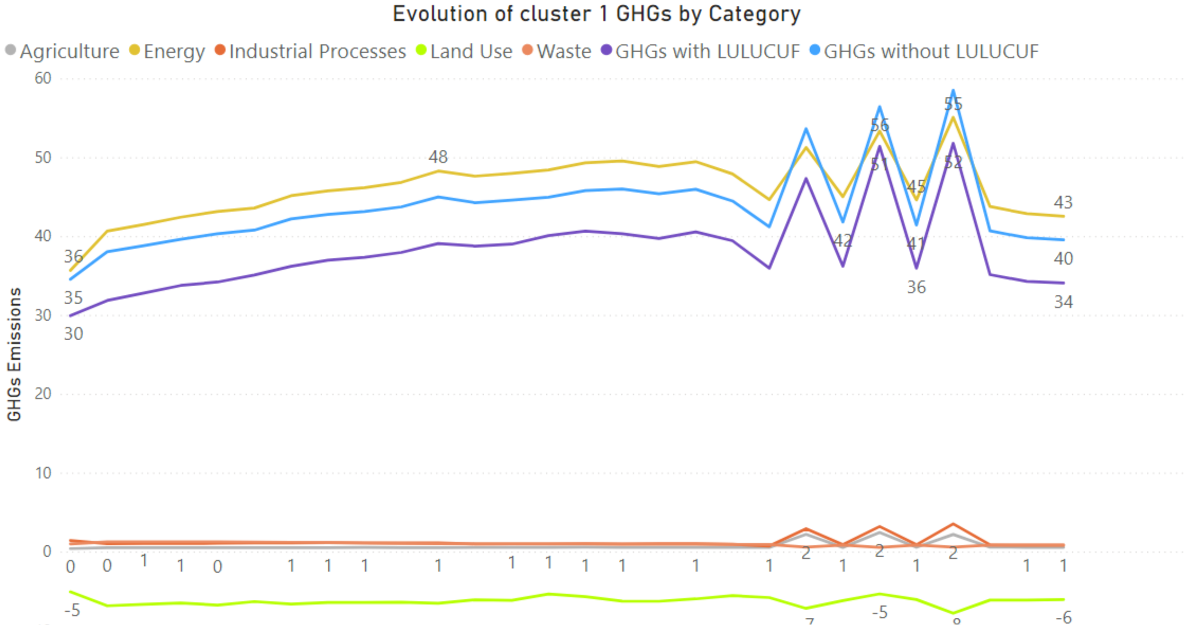


Figure 29 - Evolution of cluster 1 GHGs by category.

Regarding the analysis of *Emissions* clusters for countries, all years used in analysis (1990, 1995, 2006 and 2017), it was verified that Russia, in 1990, produced high values of C_2F_6 and CO_2 in the atmosphere due to the fact of belonging to **cluster 1** but, over time, its emissions diminished and then shifted to **cluster 0** (Annex 40, Annex 41, Annex 42, Annex 43, Annex 44, Annex 45, Annex 46 and Annex 47). United States of America is the constant country belonging to **cluster 1** in all four years in analysis, meaning its C_2F_6 and CO_2 remained with high values in the time series in analysis.

4.2.2. Country Characteristics

Country Characteristics data are defined by variables such as **Energy intensity, Renewable Energy Supply, Production-based CO_2 intensity, Development of environment-related technologies** and **Mortality from exposure to ambient PM2.5**.

In general, all of *Country Characteristics* features were very volatile throughout the time series but there was something in the year 2000 that made Energy intensity, Mortality from exposure to ambient PM2.5, Production-based CO_2 intensity and GDP at one of their lowest values and, Renewable Energy Supply, Development of environment-related technologies, Total population and GDP growth at their highest (Annex 48, Annex 49, Annex 50, Annex 51, Annex 52, Annex 53, Annex 54 and Annex 55). In 2009, both Total population and GDP growth, had really low values due to the economic crisis (Annex 54 and Annex 55). It can be verified that GDP ever since 2000 has been growing exponentially (Figure 35). Since 2011, Mortality from exposure to ambient PM2.5 and Production-based CO_2 intensity values have been decreasing (Figure 32; Figure 33), and since 2012, Energy intensity and Development of environment-related technologies have been increasing and decreasing, respectively (Figure 30; Figure 34). As for Renewable Energy Supply, since 2007 it has been increasing significantly (Figure 31).

As mentioned before, it was considered four clusters for *Country Characteristics*, **cluster 0** had 20.821 observations, **cluster 1** had 4.160 cases, for **cluster 2** and **cluster 3**, there were 1.197 and 5.946 scenarios, respectively.

Regarding clusters belonging to **cluster 0**, they can be defined as countries with very high Mortality from exposure to ambient PM2.5 (Figure 33) and low values for the rest of the features. **Cluster 1**, can be considered as countries that bet on environment policies since the values for Renewable Energy Supply and Development of environment-related technologies are soaring (Figure 31; Figure 34) whereas **cluster 2** are countries that still are very dependent on traditional energy since values for Energy intensity, Production-based CO_2 intensity and Mortality from exposure to ambient PM2.5 are very high, they are also rich economies and very populated (Figure 30; Figure 32; Figure 33). As for **cluster 3**, it can be understood as countries which are very dependent on energy in general since, values for Energy intensity, Renewable Energy Supply and Production-based CO_2 intensity are above average (Figure 30; Figure 31; Figure 32).

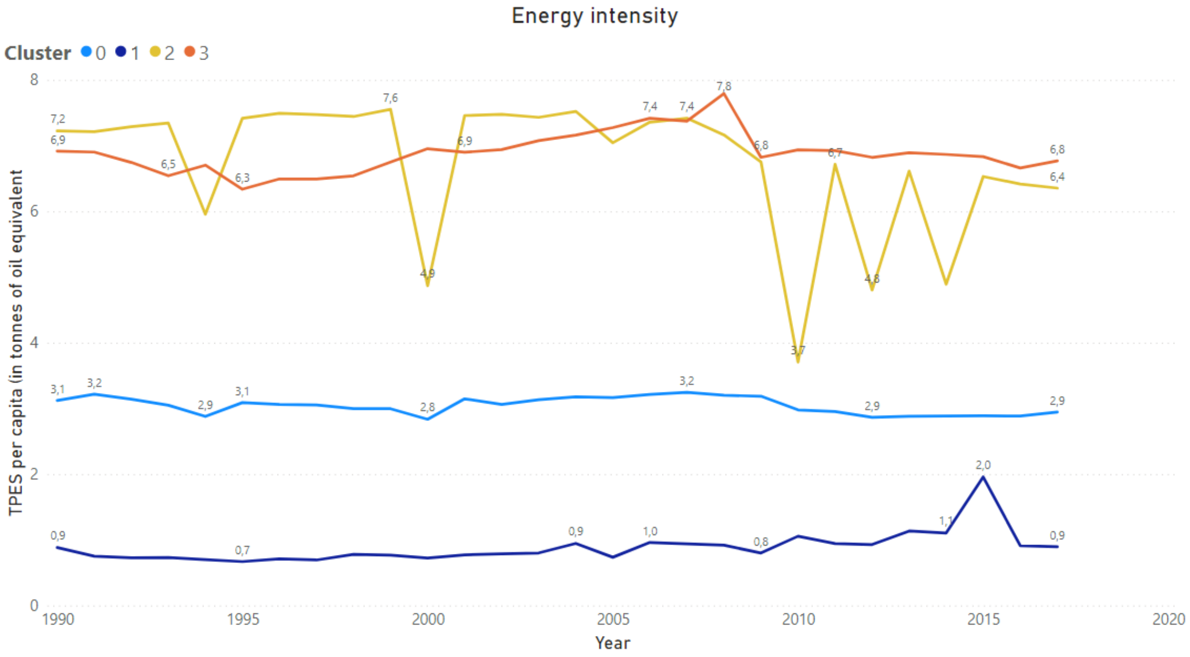


Figure 30 - Country Characteristics clusters in Energy intensity.

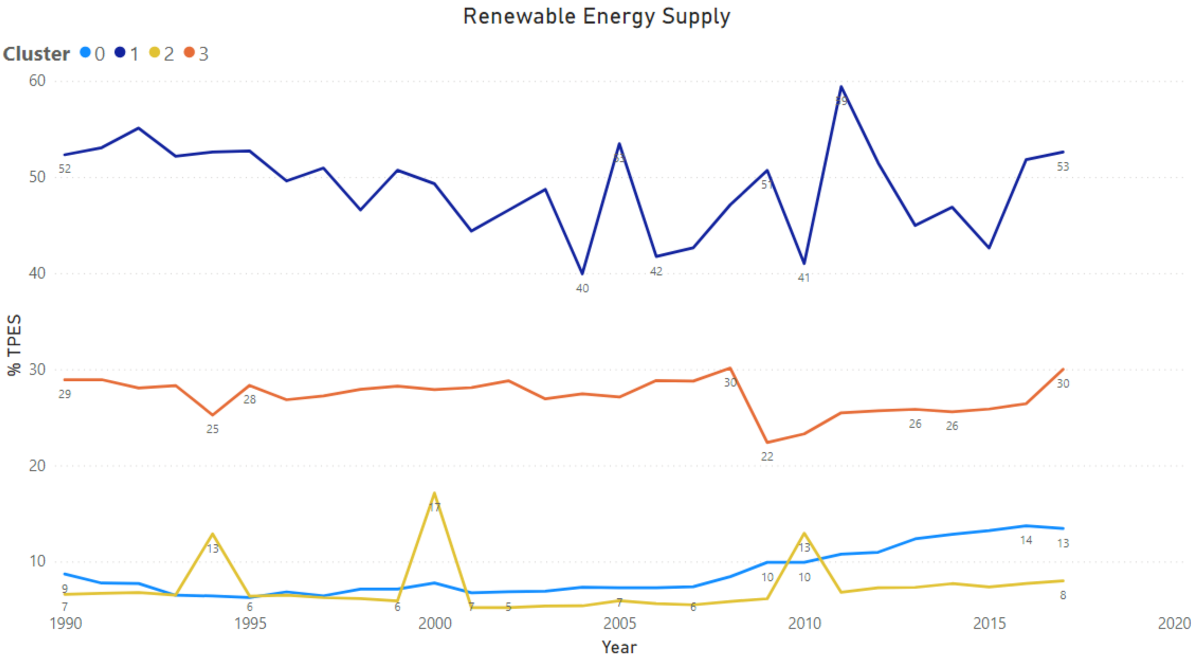


Figure 31 - Country Characteristics clusters in Renewable Energy Supply.

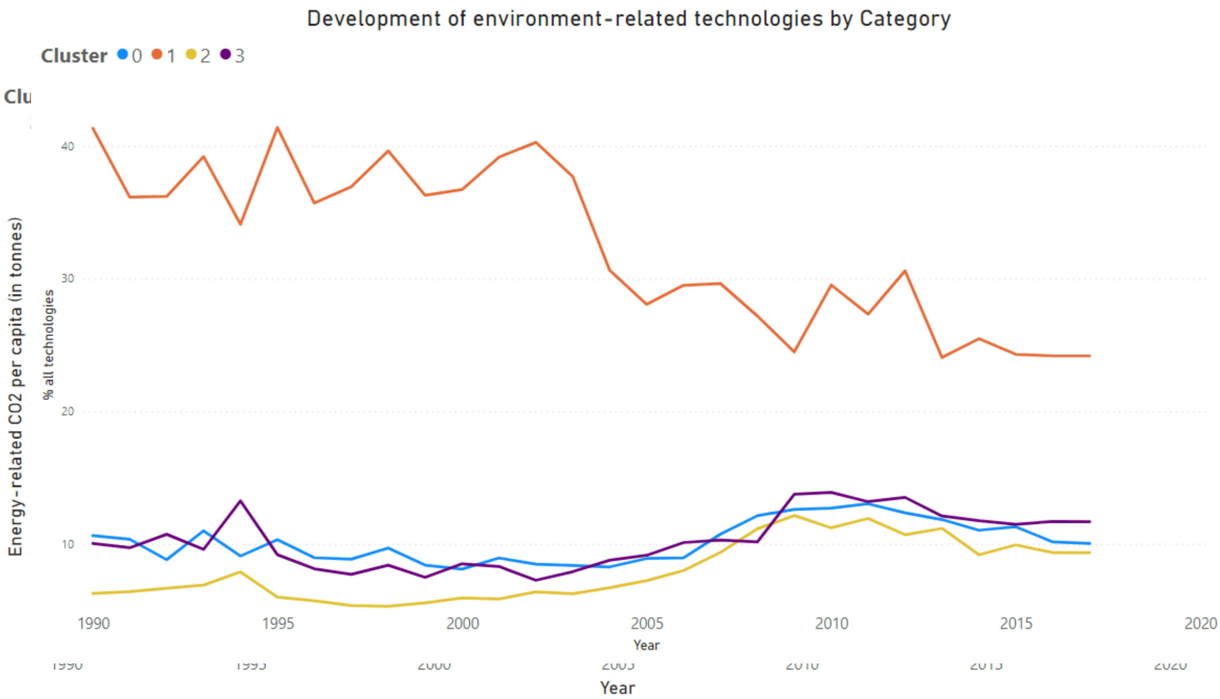


Figure 32 - Country Characteristics clusters in Production-based CO2 intensity.

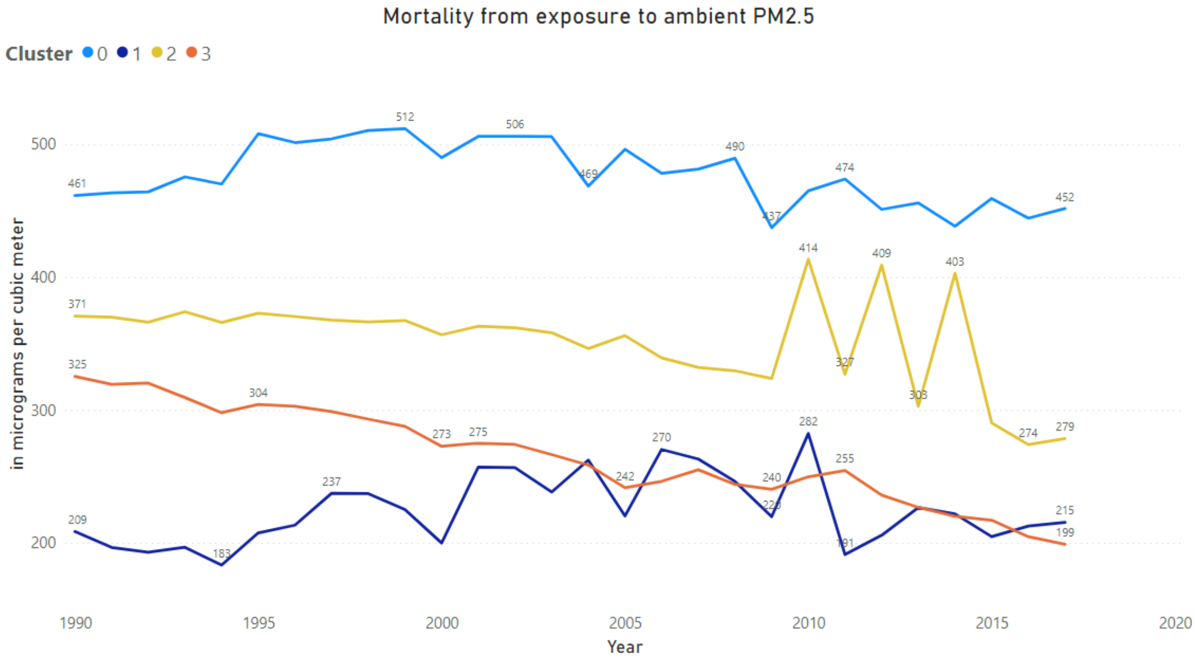


Figure 33 - Country Characteristics clusters in Mortality from exposure to ambient PM2.5.

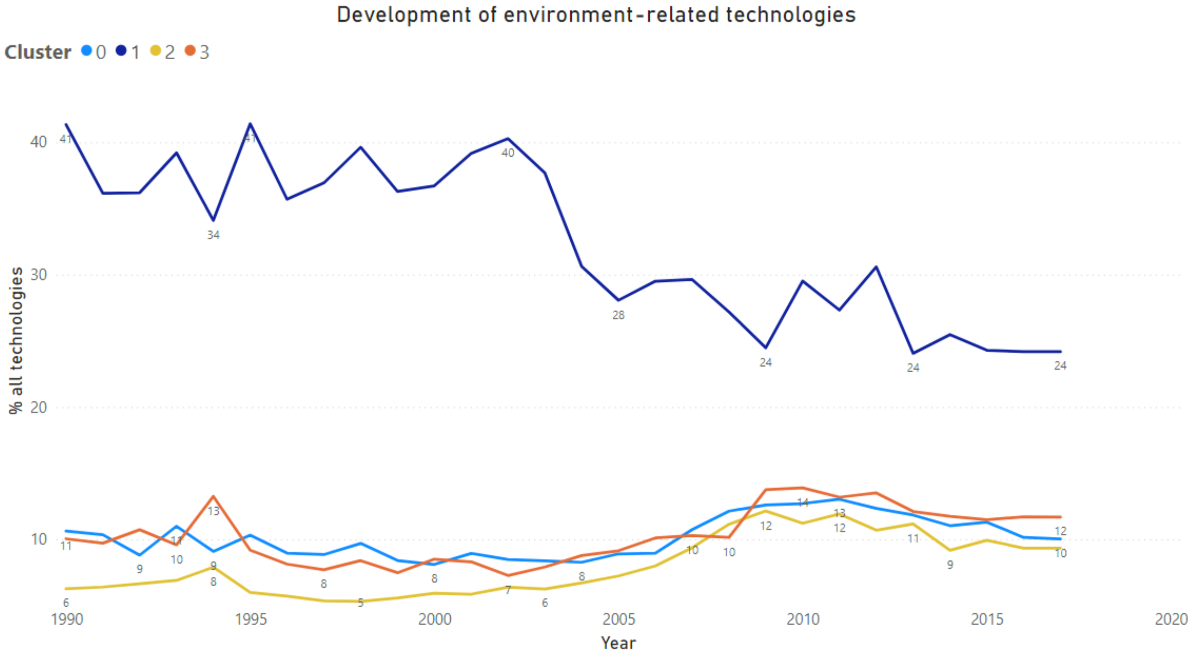


Figure 34 - Country Characteristics clusters in Development of environment-related technologies.

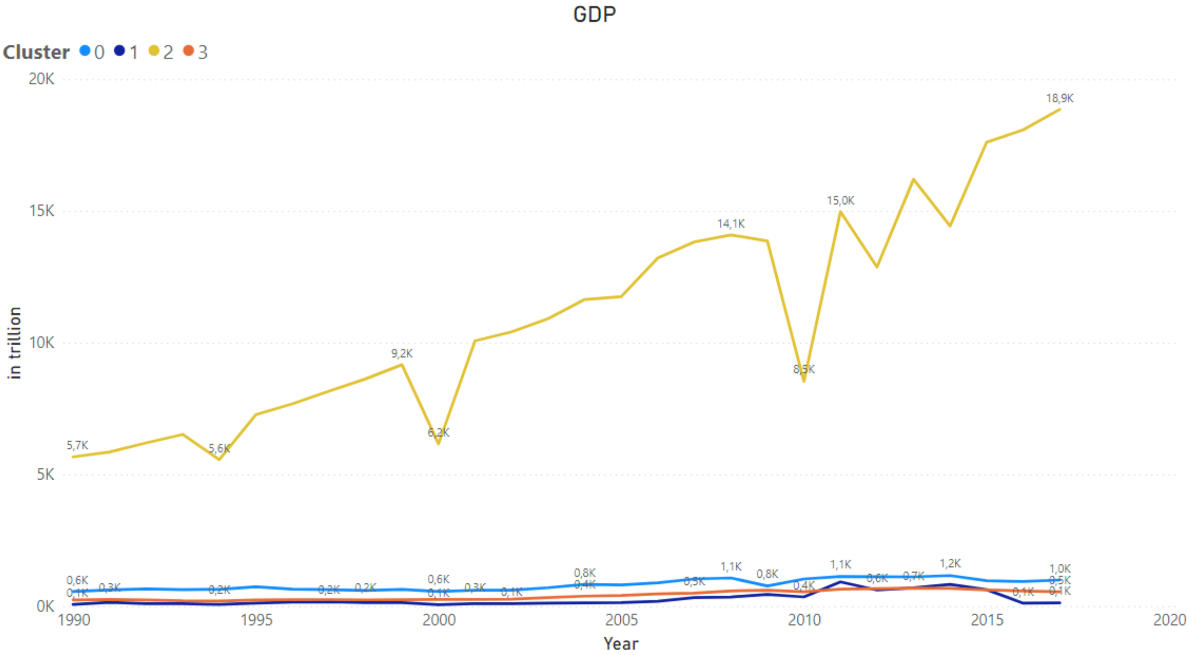


Figure 35 - Country Characteristics clusters in GDP.

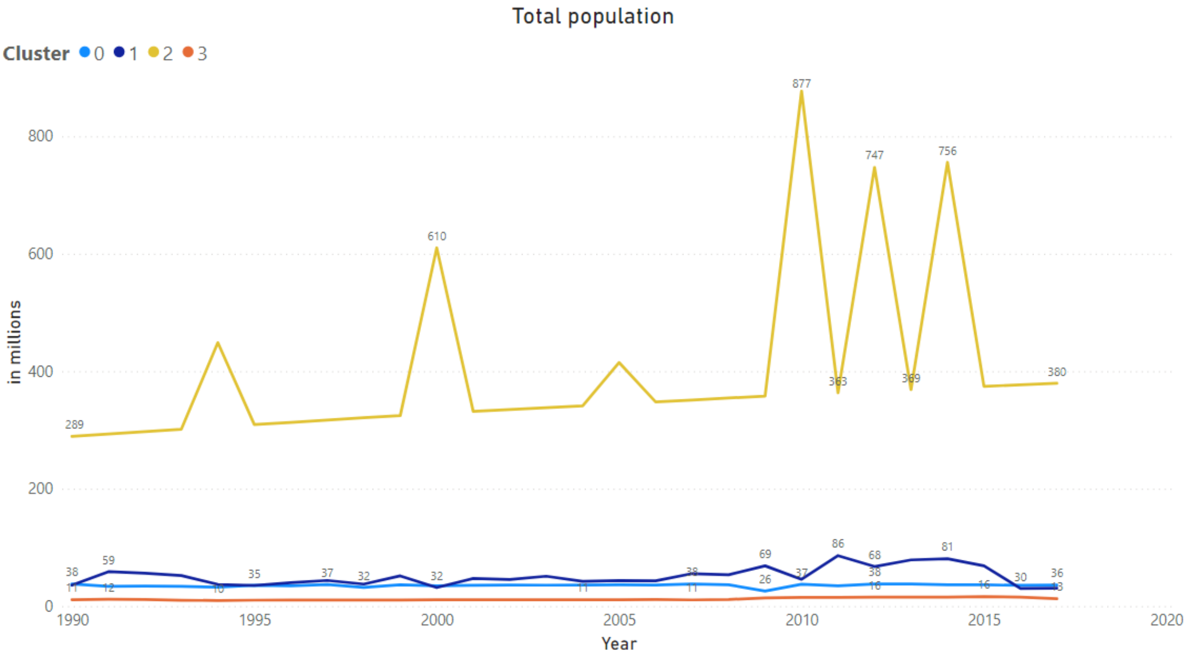


Figure 36 - Country Characteristics clusters in Total population.

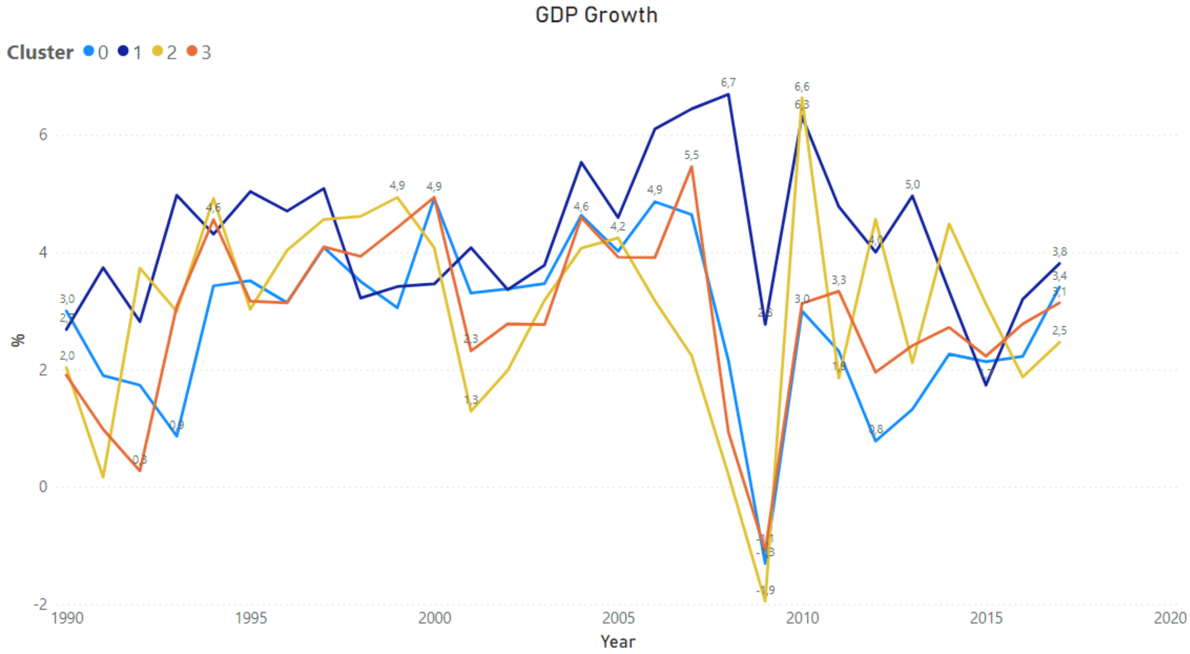


Figure 37 - Country Characteristics clusters in GDP growth.

Before UNFCCC implementation, in 1990, most of countries belonged to **cluster 0**, at least one in each continent fits in this cluster, and **cluster 1** can be understood as developing economies. United States of America, India and China, the three most polluting countries, are the only ones associated to **cluster 2** whereas Canada, Iceland, Norway, Finland, Kuwait, United Arab Emirates and Australia are allied to **cluster 3** (Figure 38).

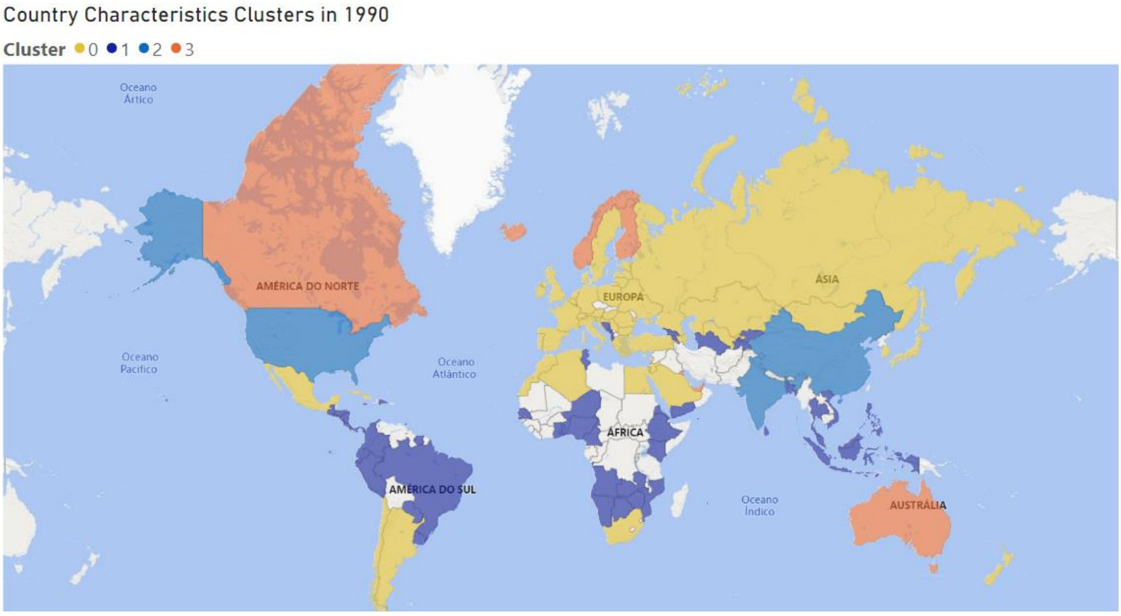


Figure 38 - Country Characteristics clusters in 1990.

The year 1995 is marked by a shift of Tunisia and Malaysia from cluster 1 to **cluster 0** and Mongolia, Jordan, Algeria and Chile performed the reverse, in 1990 belonged to cluster 0 but in 1995 moved to **cluster 1**. Regarding **cluster 2**, it remained with the same three countries as before and **cluster 3** got new countries that shifted from cluster 0 (Figure 39).

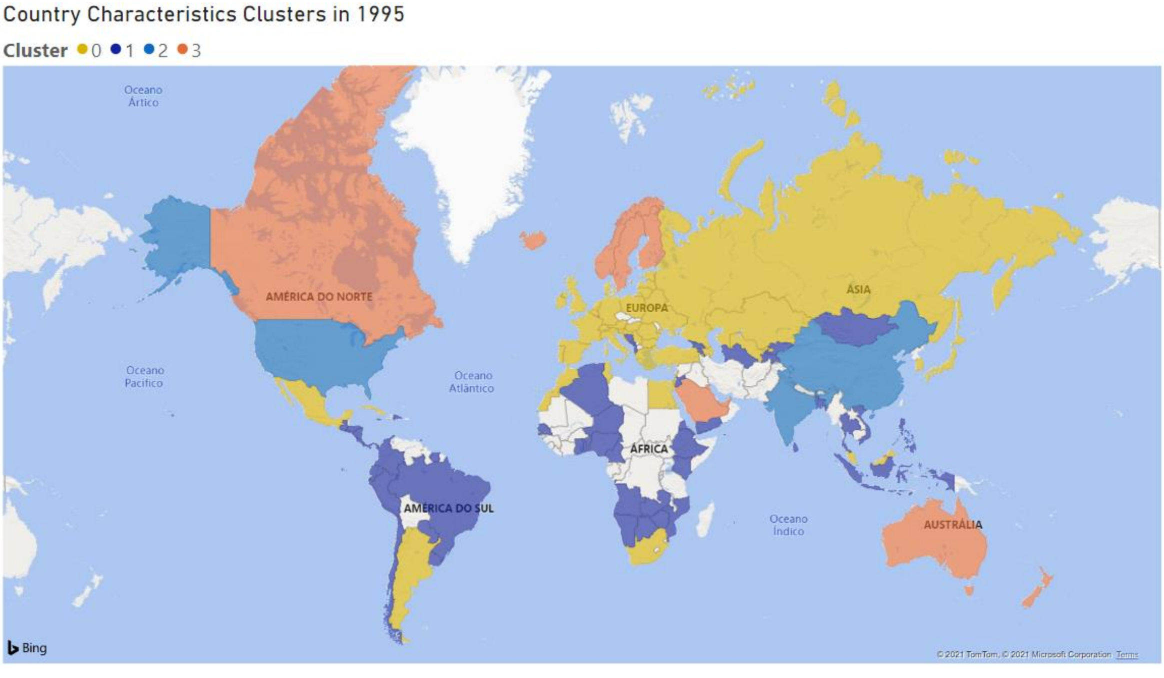


Figure 39- Country Characteristics clusters in 1995.

After the Kyoto Protocol, 2006 is defined as the year Uzbekistan changed from cluster 0 to **cluster 1** and everything stayed the same as the previous year (Figure 40).

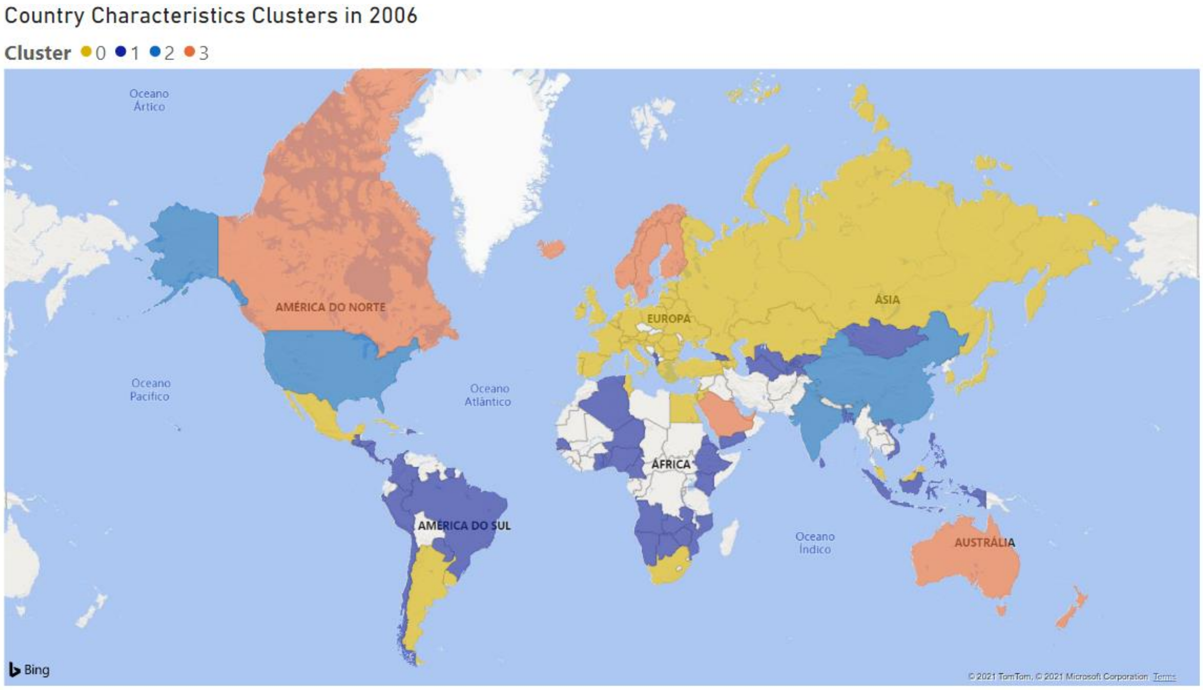


Figure 40- Country Characteristics clusters in 2006.

Finally, in the year 2017, is registered for Thailand having valuable information which contributes for belonging to **cluster 0**. Chile, Namibia, Algeria, Montenegro, Uzbekistan, Turkmenistan, Mongolia and Georgia also fit in this cluster whereas Uruguay is in **cluster 1** and Korea, Estonia and Kazakhstan moved to **cluster 3** (Figure 41).

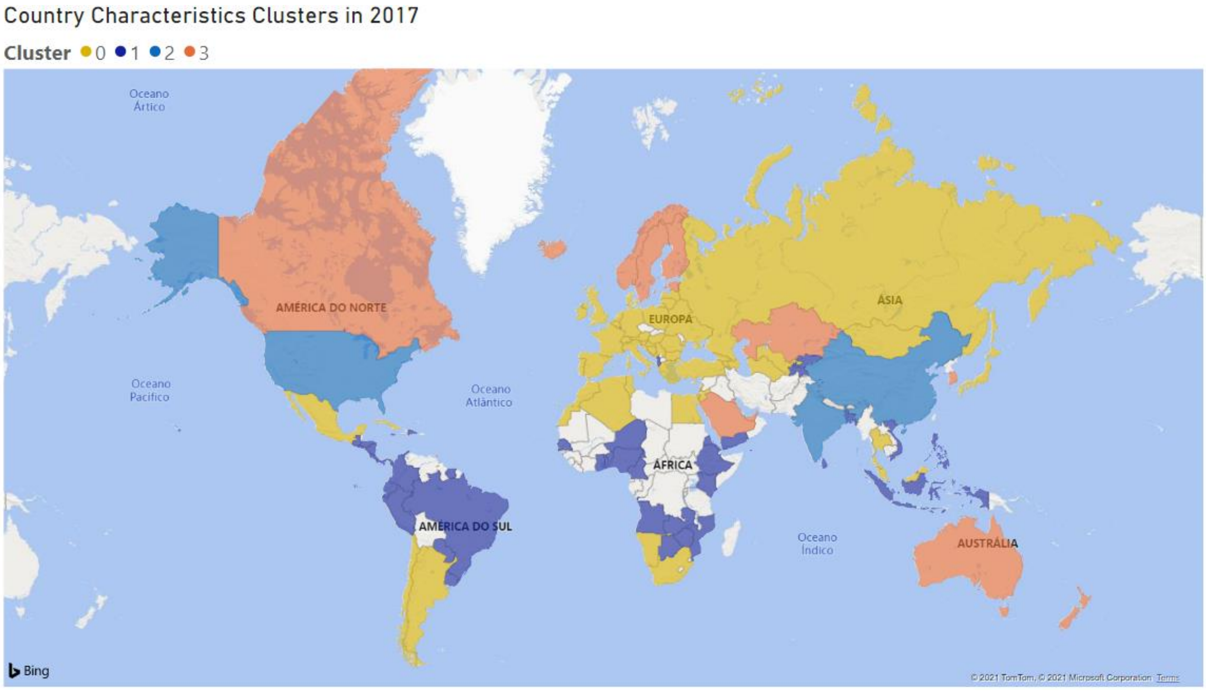


Figure 41- Country Characteristics clusters in 2017.

4.2.3. Emissions and Country Characteristics

In order to understand better the behavior of GHGs emissions it was performed a **contingency table** between clusters from both *Emissions* and *Country Characteristics* clustering in which gave the following observations:

Table 4 - Contigency table observations.

	Country Characteristics Clustering				
Emissions Clustering		0	1	2	3
	0	20.816	4.160	831	5.946
	1	5	0	366	0

In the table below, it is represented the average cluster values for each feature when combining *Emissions* (E) and *Country Characteristics* (CC) (Table 5).

Table 5 - Average values of clusters from contingency table.

Clusters	Energy Intensity	Renewable Energy Supply	Production CO2 Intensity	Environment Technologies	PM2.5 Mortality	Total Population	GDP Growth	GDP	Agriculture	Energy	Industrial Processes	Land Use	with LULUCF	without LULUCF	Waste	Nº Years	Year
E0 CC0	3,041	8,722	6,857	10,126	478,257	35,548	2,921	817,59	10 000	87 200	4 700	-12 900	79 400	88 900	4 700	17	2003
E1 CC0	5,96	3,01	14,664	14,14	710,684	148,292	-2,999	516,814	10 100	2 289 400	225 800	-105 100	2 420 300	2 525 500	37 900	30	1990
E0 CC1	0,845	48,862	1,42	33,809	223,27	47,009	4,282	209,776	36 000	18 900	3 200	-54 300	94 800	43 000	3 400	19	2001
E0 CC2	6,266	8,69	15,308	8,175	357,581	466,522	3,408	10462,406	251 200	554 100	12 400	-63 400	330 600	330 200	67 600	17	2003
E1 CC2	7,05	5,558	17,243	8,111	355,286	377,175	2,955	11804,637	83 700	4 687 000	139 100	-635 400	3 912 600	4 483 700	93 600	16	2004
E0 CC3	6,889	27,092	11,871	10,476	260,395	12,537	3,006	431,829	8 500	58 600	3 100	-4 000	57 200	60 700	2 100	16	2004

To begin with, it is important to do an analysis of the outcomes for each feature, separately from cluster analysis. Since *Energy Intensity* is in TPES (Total Primary Energy Supply) per capita, and primary energy can be understood as energy with no human transformation whether renewable or non-renewable, it is difficult to understand if this feature has good results or not for the main research question. Maybe this feature had to be combined with another variable such as *Renewable Energy Supply* or *Production CO_2 Intensity*. As for *Renewable Energy Supply*, considering it is in percentage of TPES, it does not seem to affect *Energy Intensity* but when compared to *Production CO_2 Intensity (per capita)* it can be easily understood that in fact it is. This leads to a conclusion that the *Energy Intensity* supply represented is the non-renewable since it is the one who “feeds” the human usage of fossil fuels.

It can be verified *Land Use*, all its average cluster values, are negative which means this feature is a **Carbon Capture and Storage (CCS)** whereas the other categories are still producing harming GHGs emissions (Environmental Technology, 2017)**.**

Regarding clusters characterization, it is going to be described the average object belonging to each cluster combination:

- **E0CC0:** is defined mostly by the year 2003, where the average countries have positive GDP but not too high compared with other clusters. It has small population and low GHGs emissions. This makes sense since the continuous growth of population leads to higher emissions. Regarding to renewable energy supply, CO_2 production and environment technologies, its values are considered lower than other clusters. PM2.5 mortality is high, but it is not CO_2 intensity what influences, it seems it is emissions provided from Energy, Total GHGs emissions with and without LULUCF. Some countries identified are Russia, Japan, Mexico, South Korea, South Africa, Morocco, Argentina and almost every country in Europe (Annex 56).

- **E0CC1:** is defined mostly by the year 2001 and it is considered an average GDP country. The combination of low CO_2 per capita, low population, low CO_2 intensity, high renewable energy production, high investment in environment technologies affects the low PM2.5 mortality. It can be confirmed that developing countries belong here (Annex 57).

- **E0CC2:** is defined mostly by the year 2003 and characterized by rich countries. Even though they are countries with monetary power to invest in environment technologies and renewable energy, these values are too low for what they could be. The fact that are countries highly populated and high GHGs emissions, makes PM2.5 mortality relatively high. It is not a surprise that China, United States of America (USA) and India are the countries characterized at this juncture (Annex 56).

- **E0CC3:** is defined mostly by the year 2004 and an average GDP country. This cluster’s combination behaves differently from the others in the sense that they have a high renewable energy production but low environment technologies, possibly exports their production. Their CO_2 intensity is high due to GHGs emissions from Energy and emissions with and without LULUCF. Countries characterized with small population and relatively low GHGs emissions production, influences the low PM2.5 mortality. Countries related to this clustering are Canada, Iceland, Sweden, Finland, Norway, Saudi Arabia, Australia and New Zealand (Annex 58).

- **E1CC0:** is defined only by the year 1990 and the country behaves similar to **E0CC2** with the exception it is not a rich country and is an economy in recession. Its economy is based on industries that affects Energy, Industrial Processes and emissions with and without LULUCF. There is only one country that belongs to this combination, Russia (Annex 60).
- **E1CC2:** is defined mostly by the year 2004 and is also similar to **E0CC2** but countries in this combination are distinguished for very high GHGs emissions per category. United States of America is the only country associated to this combination (Annex 60), nonetheless, is the one with highest **CCS**.

4.2.4. Conclusion

When doing this research, it was expected CO_2 to be the greenhouse gas that most contributes to the rise of emissions, and it was verified. CH_4 and C_2F_6 are also gases that produce very high chemical substances in the atmosphere. The growth of these emissions is explained mostly by the Energy industry. Since GHGs with and without LULUCF are the aggregation of all GHGs emissions by category, it can be retained that Land Use is a Carbon Capture Storage still, it does not affect that many emissions. GHGs with land use, land-use changes and forestry have minor values when comparing with GHGs emissions without LULUCF. It is noteworthy to say production-based CO_2 intensity has been decreasing since 2007 but the energy sector, in terms of total primary energy supply (TPES) *per capita*, since 2000, has been increasing, due in part to the rise of population. This constant growth of energy can be explained by the shift from production-based CO_2 intensity to renewable energy production. The investment made in the development of environment related technologies is not accompanied by the constant growth of GDP, this could mean, even though the world is getting richer, these efforts are not being applied in the reduction of GHGs emissions. Since 2011 production-based CO_2 intensity is declining and, this change has direct effect in the reduction of mortality exposure to ambient PM2.5.

In short, *Emissions* clusters are defined by which type of greenhouse gases each country most produce:

- **Cluster 0:** countries with all greenhouse gases in analysis;
- **Cluster 1:** countries shaped only C_2F_6 and CO_2.

As for *Country Characteristics*:

- **Cluster 0:** countries with high mortality exposure to ambient PM2.5;
- **Cluster 1:** countries that invest in environment related policies;

- **Cluster 2:** countries very dependent on traditional energy, that are rich and have high population;
- **Cluster 3:** countries very dependent energy, in both traditional energy and renewable energy.

For the combination of both *Emissions* and *Country Characteristics*, taking into account the chosen variables for analysis, two groups of clusters that have low values for renewable energy supply and at the same time high values for mortality exposure to ambient PM2.5, which makes sense since betting on renewable energy reduces this type of death. There are also four groups where production-based CO_2 intensity and development of environment related technologies have opposite values, highlighting the fact that CO_2 intensity have high values and the other feature has low values. In general, throughout a time series of 27 years, there are not many changes stated regarding GHGs emissions, but this conclusion is biased, as it would be necessary more information to verify this.

5. CONCLUSIONS

This chapter summarizes the main conclusions retrieved in the presented research and identifies its main limitations.

In the process of completing this study, it was possible to understand the behavior of GHGs emissions and important traits of the countries throughout 1990-2017, in each cluster. There was an attempt in understanding if the characteristics of countries, in some extent, influenced the rise of anthropogenic greenhouse gases emissions.

As the literature review demonstrates, there were a lot of variables to take in consideration when trying to understand what influences GHGs emissions. There were direct contributors for this growth such as energy intensity and some indirect, more social indicators, that were high influencers, but it was hard to measure with quantitative data. This was the case of consumption habits and beliefs of each person. The closest approach in order to best reach these insights was, with information about each country.

The KDD process was the booster to connect the information reviewed for the main research question with the given outcomes. The approach was important for the research since in the Data Mining step, *clustering*, was the method used to obtain insights.

The product of using a clustering technique, gave results different from what was expected. There were chosen certain years for the analysis of GHGs emissions and some country indicators, separately, to realize if in fact the measures applied to reduce emissions are changing during this time. The results are somewhat inconclusive since there are countries shifting from cluster to cluster, but it is noteworthy to say that three of the most pollutants countries remained with this legacy during the 27 years in the analysis. It is worth mentioning when joining data from both emissions and countries characteristics, the given average years belonging to each cluster were not influenced by sustainable measures. It was curious how, even though, both analysis gave different results years, the same was not applied for the rest of the variables, once again it was demonstrated there is still much to do in order to mitigate the consequences of anthropogenic greenhouse gases emissions.

5.1. LIMITATIONS AND FUTURE WORK

There were some limitations during the research. First and foremost, the difficulties in finding data easy to understand from someone who's background is not an expert in climate change or similar fields, there was data found but in order to treat it, was necessary more inside knowledge. Since the variables were constrained, it was difficult to create new and interesting variables.

Another limitation was the methodology used for the investigation, using unsupervised learning for such an ambitious question led to inconclusive results. Even though it was possible to conclude there were not big changes in the decrease of GHGs emissions this would have been possible to reach without performing clustering.

In order to achieve a more insightful outcome, instead of using unsupervised learning, applying supervised learning is considered a better practice.

6. BIBLIOGRAPHY

Anderson, T. R., Hawkins, E., & Jones, P. D. (2016). CO2, the greenhouse effect and global warming: from the pioneering work of Arrhenius and Callendar to today's Earth System Models. *Endeavour*, *40*(3), 178–187. https://doi.org/10.1016/j.endeavour.2016.07.002

Arioli, M. S., D'Agosto, M. de A., Amaral, F. G., & Cybis, H. B. B. (2020). The evolution of city-scale GHG emissions inventory methods: A systematic review. In *Environmental Impact Assessment Review* (Vol. 80, p. 106316). Elsevier Inc. https://doi.org/10.1016/j.eiar.2019.106316

Åström, K. J. (1969). On the choice of sampling rates in parametric identification of time series. *Information Sciences*, *1*(3), 273–278. https://doi.org/10.1016/S0020-0255(69)80013-7

Center for Climate and Energy Solutions. (2016). *Main Greenhouse Gases*. https://www.c2es.org/content/main-greenhouse-gases/

Center for Sustainable Systems. (2019). *Greenhouse Gases Factsheet*. http://css.umich.edu/factsheets/greenhouse-gases-factsheet

Climate Change Killed The Dinosaurs. 'Drastic Global Winter' After Asteroid Strike, Say Scientists. (n.d.). Retrieved July 27, 2021, from https://www.forbes.com/sites/jamiecartereurope/2020/06/29/climate-change-killed-the-dinosaurs-a-drastic-global-winter-after-asteroid-strike-say-scientists/?sh=5856591a2e34

Coronavirus and Climate Change – C-CHANGE | Harvard T.H. Chan School of Public Health. (n.d.). Retrieved July 27, 2021, from https://www.hsph.harvard.edu/c-change/subtopics/coronavirus-and-climate-change/

Environmental Indicators for Agriculture. (2001). *OECD Glossary of Statistical Terms - Carbon dioxide equivalent Definition*. Glossary of Statistical Terms. https://stats.oecd.org/glossary/detail.asp?ID=6323%0Ahttps://stats.oecd.org/glossary/detail.asp?ID=285

Environmental Technology. (2017). *What Are Negative Emissions?* Environmental Technology. https://www.envirotech-online.com/news/environmental-laboratory/7/breaking-news/what-are-negative-emissions/44272

Esling, P., & Agon, C. (2012). Time-series data mining. *ACM Computing Surveys*, *45*(1). https://doi.org/10.1145/2379776.2379788

Flämig, H., Lunkeit, S., Rosenberger, K., & Wolff, J. (2019). Enlarging the scale of BEVs through environmental zoning to reduce GHG emissions: A case study for the city of Hamburg. *Research in Transportation Business and Management*, 100418. https://doi.org/10.1016/j.rtbm.2019.100418

Fu, T. C. (2011). A review on time series data mining. In *Engineering Applications of Artificial Intelligence* (Vol. 24, Issue 1, pp. 164–181). Elsevier Ltd. https://doi.org/10.1016/j.engappai.2010.09.007

Greenhouse Gas Protocol. (2015). Global Warming Potential Values. In *Greenhouse Gas Protocol* (Vol. 2014, Issue 1995). www.ipcc.ch

Hammer, B., Micheli, A., Neubauer, N., Sperduti, A., & Strickert, M. (2005). *(PDF) Self organizing maps for time series*.

https://www.researchgate.net/publication/215386291_Self_organizing_maps_for_time_series #read

Homma, T., Akimoto, K., & Tomoda, T. (2012). Quantitative evaluation of time-series GHG emissions by sector and region using consumption-based accounting. *Energy Policy*, *51*, 816–827. https://doi.org/10.1016/j.enpol.2012.09.031

Hui, J. (2019). *Machine Learning — Singular Value Decomposition (SVD) & Principal Component Analysis (PCA) Matrix diagonalization*. https://medium.com/@jonathan_hui/machine-learning-singular-value-decomposition-svd-principal-component-analysis-pca-1d45e885e491

IBM Corporation. (2018). *KMO and Bartlett's Test*. IBM Knowledge Center. https://www.ibm.com/support/knowledgecenter/en/SSLVMB_24.0.0/spss/tutorials/fac_telco_kmo_01.html%0Ahttps://www.ibm.com/support/knowledgecenter/SSLVMB_subs/statistics_casestudies_project_ddita/spss/tutorials/fac_telco_kmo_01.html%0Ahttps://www.ibm.com/sup

IPCC, 2014: *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. IPCC, Geneva, Switzerland, 151 pp.

IPCC, 2018: Global Warming of 1.5°C. *An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty* [Masson-Delmotte, V., P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R. Matthews, Y. Chen, X. Zhou, M.I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield (eds.)]. In Press.

Kaveh, M., Amiri Chayjan, R., Taghinezhad, E., Rasooli Sharabiani, V., & Motevali, A. (2020). Evaluation of specific energy consumption and GHG emissions for different drying methods (Case study: Pistacia Atlantica). *Journal of Cleaner Production*, *259*, 120963. https://doi.org/10.1016/j.jclepro.2020.120963

Keogh, E. (1997). Fast similarity search in the presence of longitudinal scaling in time series databases. *Proceedings of the International Conference on Tools with Artificial Intelligence*, 578–584. https://doi.org/10.1109/tai.1997.632306

Keogh, E., & Smyth, P. (1997). *A Probabilistic Approach to Fast Pattern Matching in Time Series Databases*. www.aaai.org

Loukas, S. (2020). *PCA clearly explained —When, Why, How to use it and feature importance: A guide in Python | by Serafeim Loukas | Towards Data Science*. https://towardsdatascience.com/pca-clearly-explained-how-when-why-to-use-it-and-feature-importance-a-guide-in-python-7c274582c37e

Madagascar Famine is First Caused Entirely by Climate Change | Time. (n.d.). Retrieved July 27, 2021, from https://time.com/6081919/famine-climate-change-madagascar/

Main Greenhouse Gases | Center for Climate and Energy Solutions. (n.d.). Retrieved March 25, 2020, from https://www.c2es.org/content/main-greenhouse-gases/

Mohan, R. R. (2018). Time series GHG emission estimates for residential, commercial, agriculture and fisheries sectors in India. *Atmospheric Environment*, *178*, 73–79.

https://doi.org/10.1016/j.atmosenv.2018.01.029

NASA. (2018). *Facts – Climate Change: Vital Signs of the Planet*. https://climate.nasa.gov/evidence/

Organisaton for Economic Co-operation and Development. (2015). *OECD Data*. OECD Website. https://data.oecd.org/

Qiang, Y., & Xindong, W. (2006). 10 Challenging problems in data mining research. *International Journal of Information Technology and Decision Making*, *5*(4), 597–604. https://doi.org/10.1142/S0219622006002258

Röck, M., Saade, M. R. M., Balouktsi, M., Rasmussen, F. N., Birgisdottir, H., Frischknecht, R., Habert, G., Lützkendorf, T., & Passer, A. (2020). Embodied GHG emissions of buildings – The hidden challenge for effective climate change mitigation. *Applied Energy*, *258*, 114107. https://doi.org/10.1016/j.apenergy.2019.114107

Rosa, E. A., & Dietz, T. (2012). Human drivers of national greenhouse-gas emissions. In *Nature Climate Change* (Vol. 2, Issue 8, pp. 581–586). Nature Publishing Group. https://doi.org/10.1038/nclimate1506

Serra, R., Niknia, I., Paré, D., Titus, B., Gagnon, B., & Laganière, J. (2019). From conventional to renewable natural gas: can we expect GHG savings in the near term? *Biomass and Bioenergy*, *131*, 105396. https://doi.org/10.1016/j.biombioe.2019.105396

Smyth, P., & Keogh, E. (1997). *Clustering and Mode Classiication of Engineering Time Series Data*.

UCLA Statistical Consulting Group. (2016). *Stata FAQ How can I perform a factor analysis with categorical (or categorical and continuous) variables ?* https://stats.idre.ucla.edu/stata/faq/how-can-i-perform-a-factor-analysis-with-categorical-or-categorical-and-continuous-variables/

UNFCCC. (1992). *Parties and Observers*. https://unfccc.int/parties-observers

United Nations. (1992). *United Nations Framework Convention on Climate Change*.

Walker, B. (2019). *PCA Is Not Feature Selection. What it actually does and when you can… | by Brandon Walker | Towards Data Science*. https://towardsdatascience.com/pca-is-not-feature-selection-3344fb764ae6

World Bank. (2018). *World Bank Open Data | Data*. World Bank Database. https://datos.bancomundial.org/%0Ahttps://data.worldbank.org/%0Ahttps://data.worldbank.org/%0Ahttps://data.worldbank.org/%0Ahttps://data.worldbank.org/%0Ahttps://data.worldbank.org/%0Ahttps://data.worldbank.org/%0Ahttps://data.worldbank.org/%0Ahttps://dat

Yang, K., & Shahabi, C. (2005). On the stationarity of multivariate time series for correlation-based data analysis. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 805–808. https://doi.org/10.1109/ICDM.2005.109

Yoon, H., Yang, K., & Shahabi, C. (2005). Feature subset selection and feature ranking for multivariate time series. *IEEE Transactions on Knowledge and Data Engineering*, *17*(9), 1186–1198. https://doi.org/10.1109/TKDE.2005.144

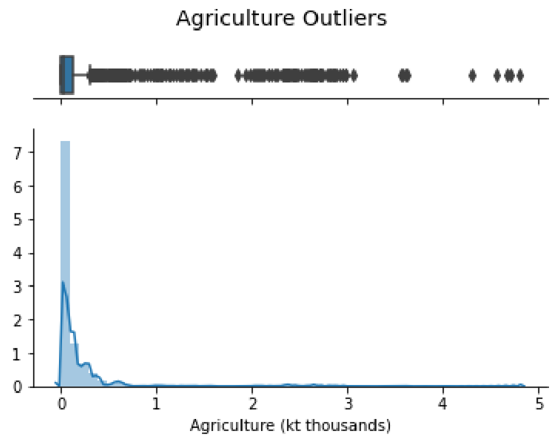
Zheng, X., Streimikiene, D., Balezentis, T., Mardani, A., Cavallaro, F., & Liao, H. (2019). *A review of greenhouse gas emission profiles, dynamics, and climate change mitigation efforts across the key climate change players*. https://doi.org/10.1016/j.jclepro.2019.06.140

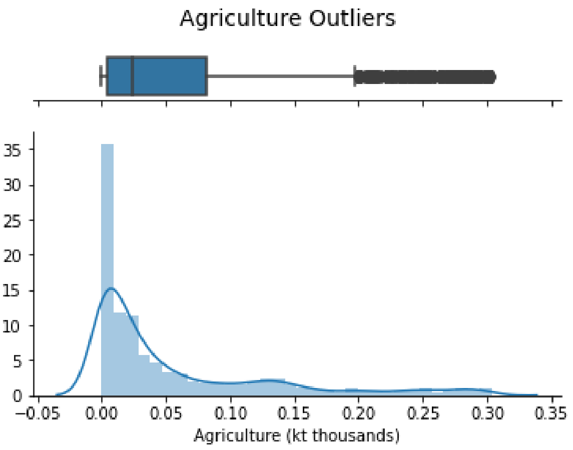
7. ANNEXES

Annex 1 – Outliers Agriculture – Before & After (Boxplot)

Before

After

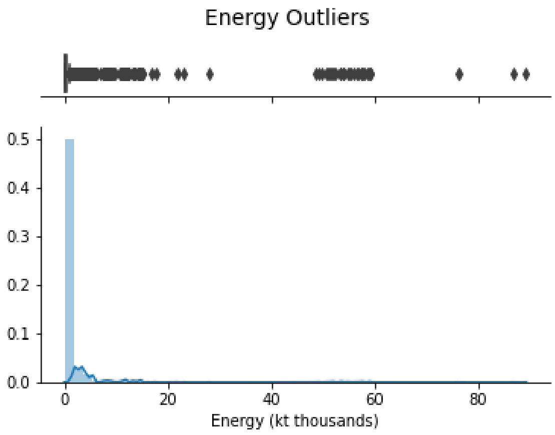


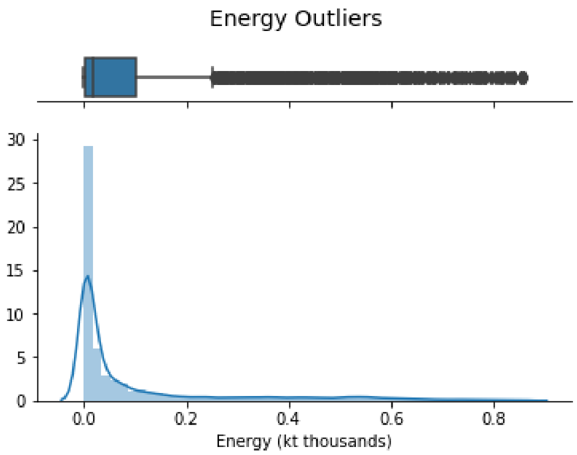


Annex 2- Outliers Energy Before & After (Boxplot)

Before

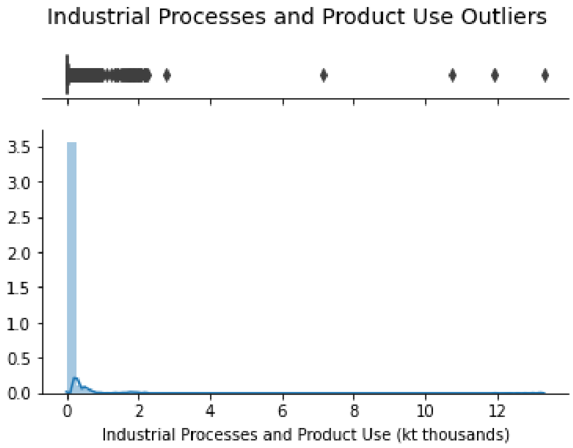
After



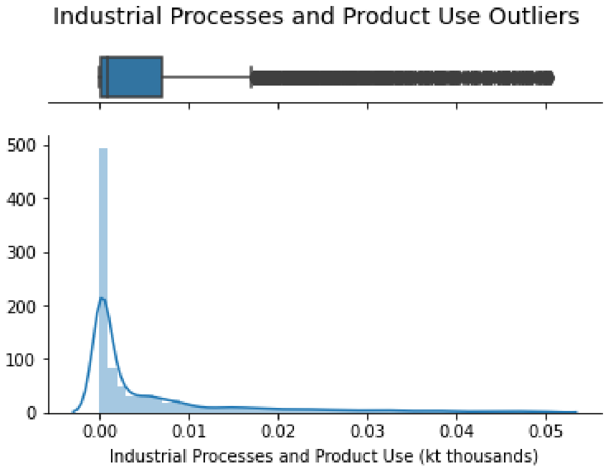


Annex 3- Outliers Industrial processes and product use Before & After (Boxplot)

Before

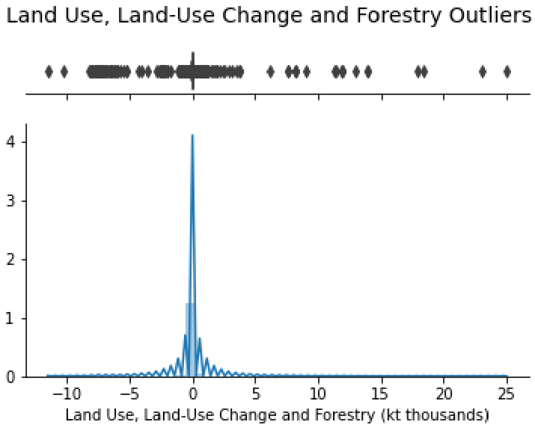


After

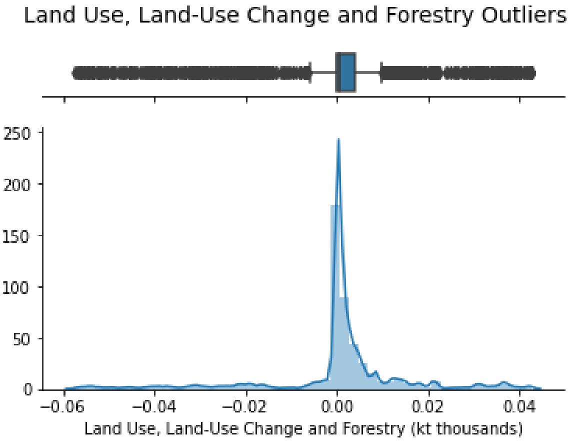


Annex 4- Outliers Land Use, Land Use Change and Forestry Before & After (Boxplot)

Before



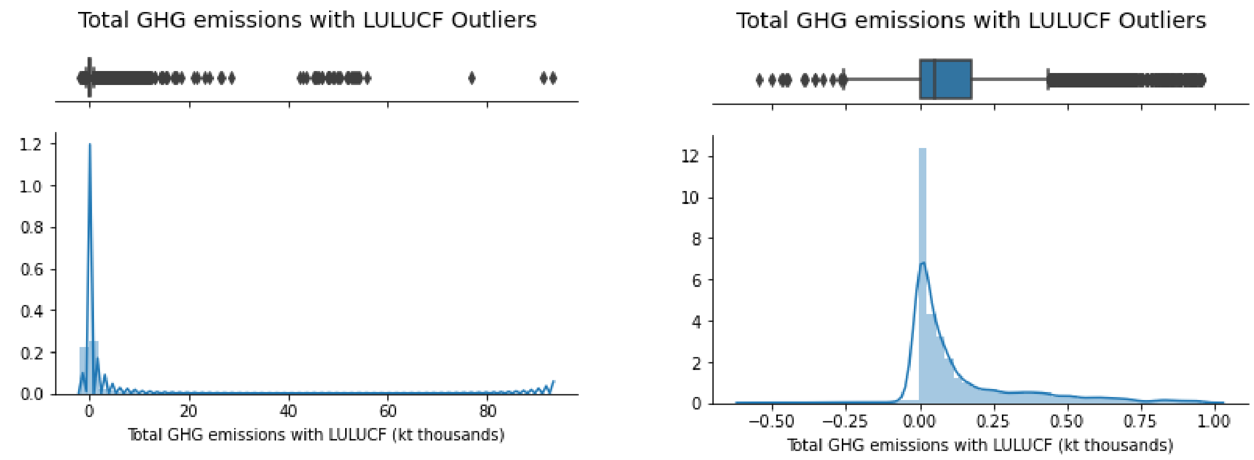
After



Annex 5- Outliers Total GHGs emissions with LULUCF Before & After (Boxplot)

Before

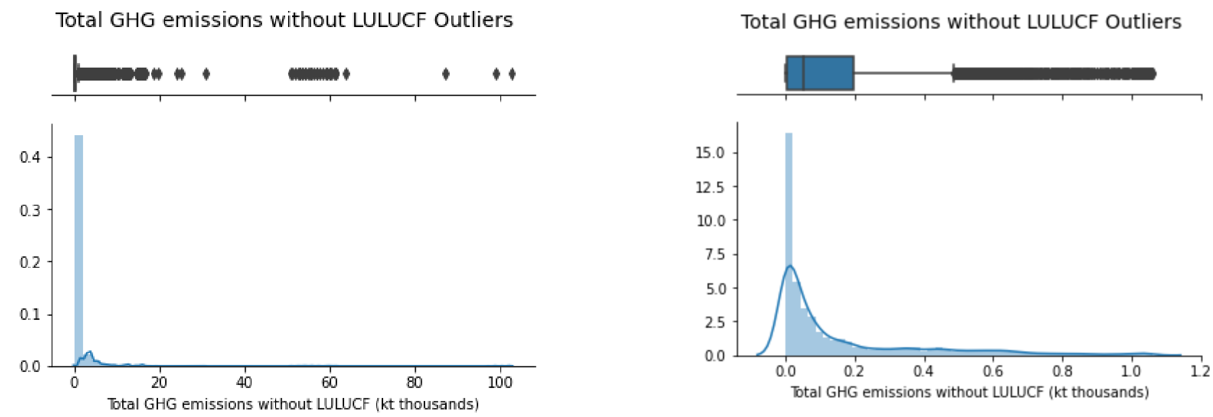
After



Annex 6- Outliers Total GHGs emissions without LULUCF Before & After (Boxplot)

Before

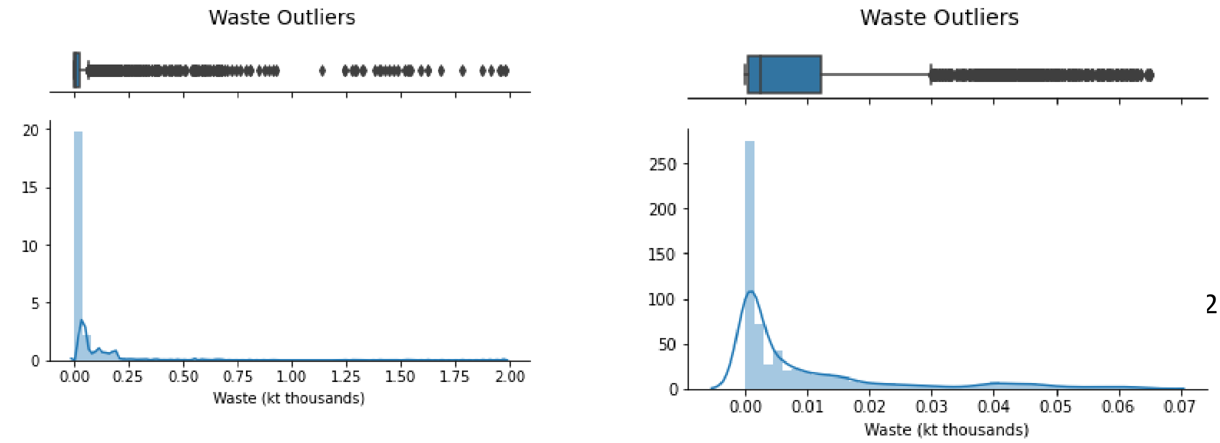
After



Annex 7- Outliers Waste Before & After (Boxplot)

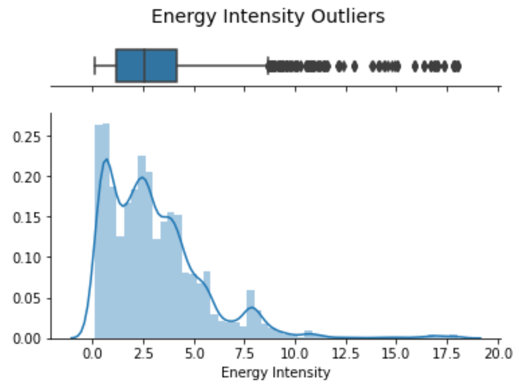
Before

after

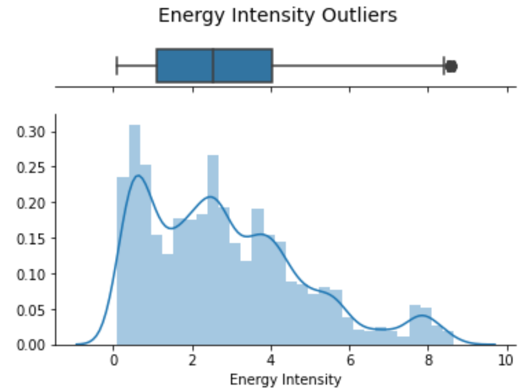


Annex 8- Outliers Energy intensity Before & After (Boxplot)

Before

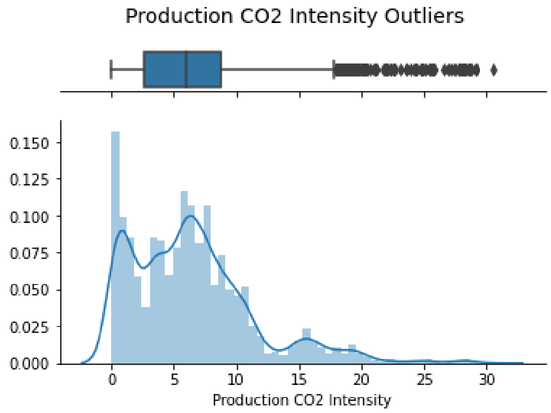


After

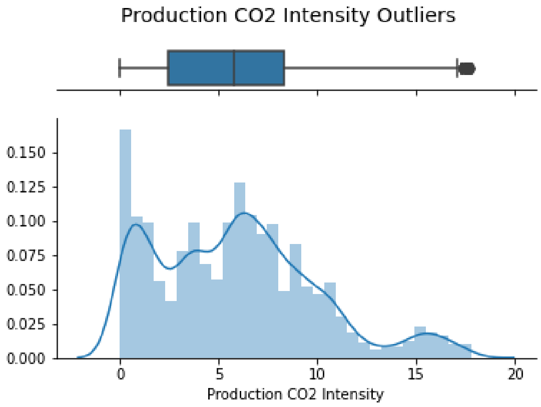


Annex 9- Outliers Production CO_2 Before & After (Boxplot)

Before

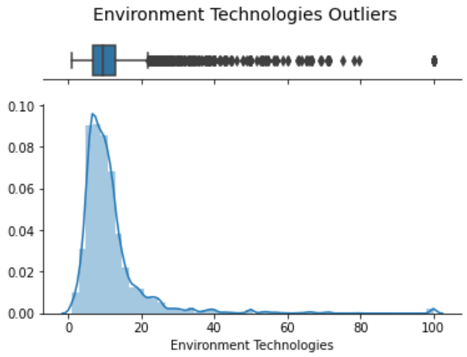


After

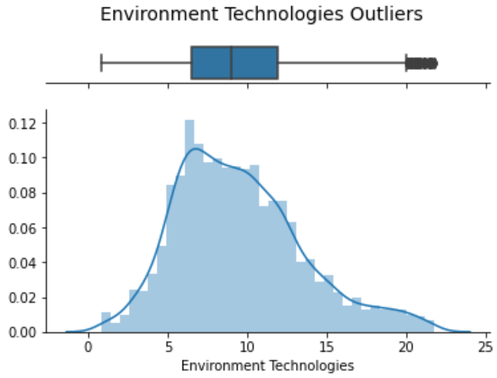


Annex 10- Outliers Environment Technologies Before & After (Boxplot)

Before

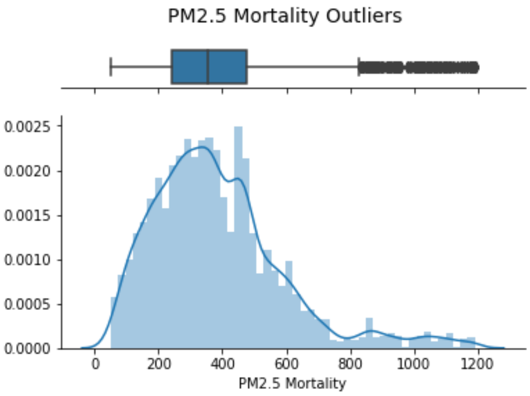


After

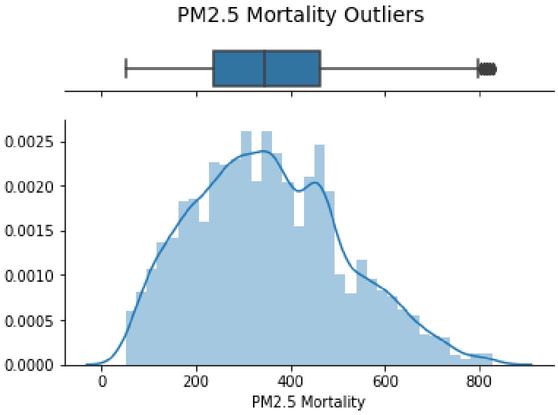


Annex 11- Outliers PM2.5 Mortality Before & After (Boxplot)

Before

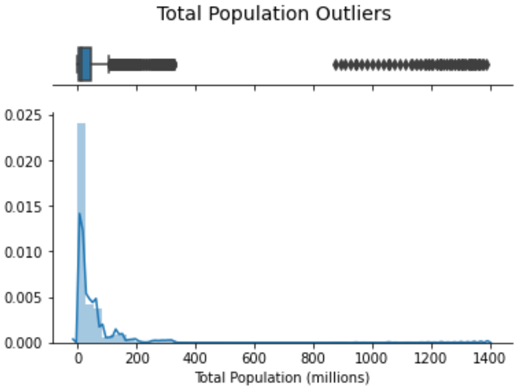


After

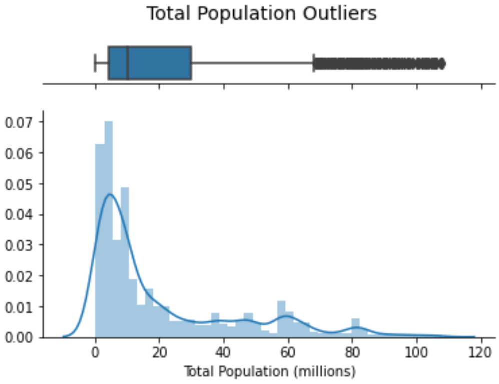


Annex 12- Outliers Total Population Before & After (Boxplot)

Before

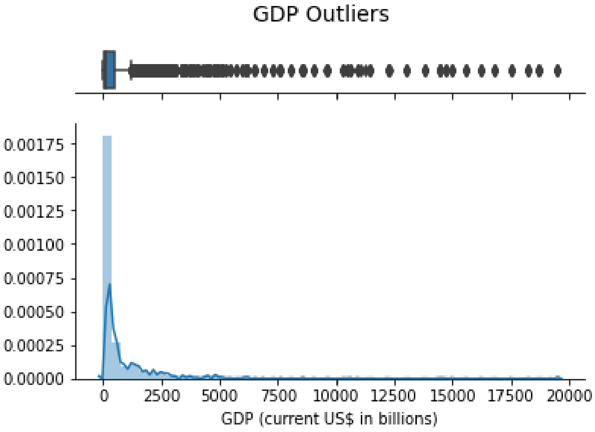


After

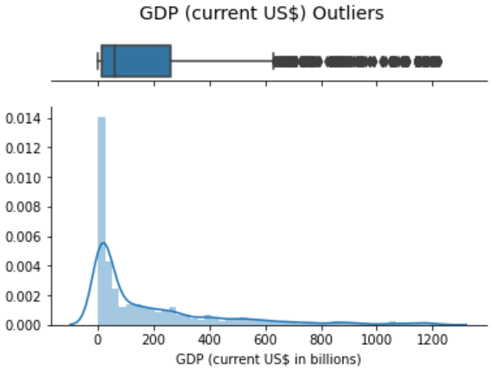


Annex 13- Outliers GDP Before & After (Boxplot)

Before



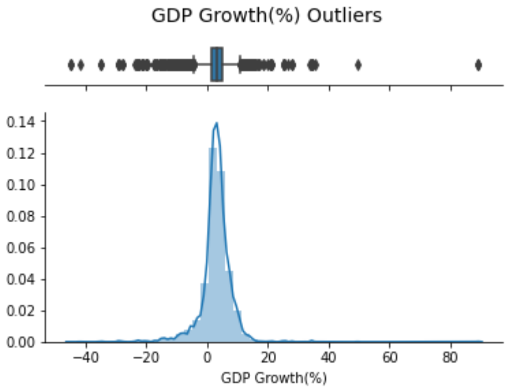
After

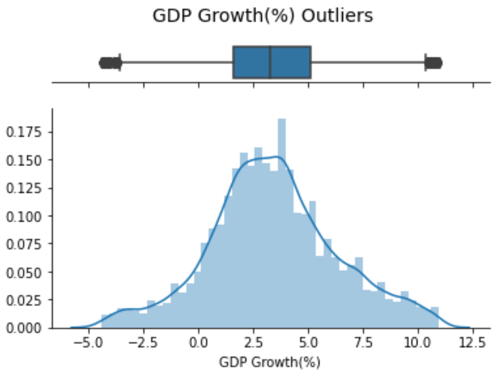


Annex 14- Outliers GDP Growth Before & After (Boxplot)

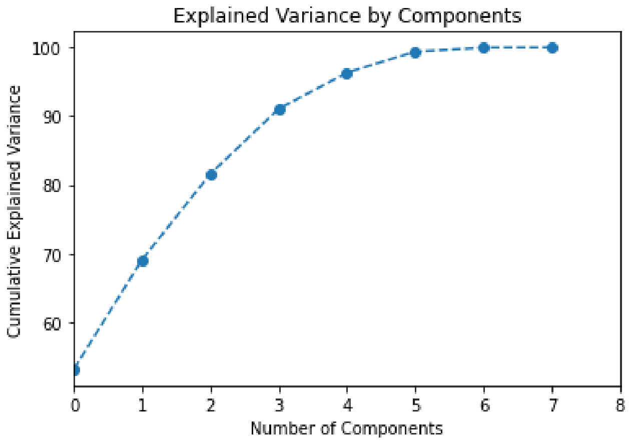
Before

After

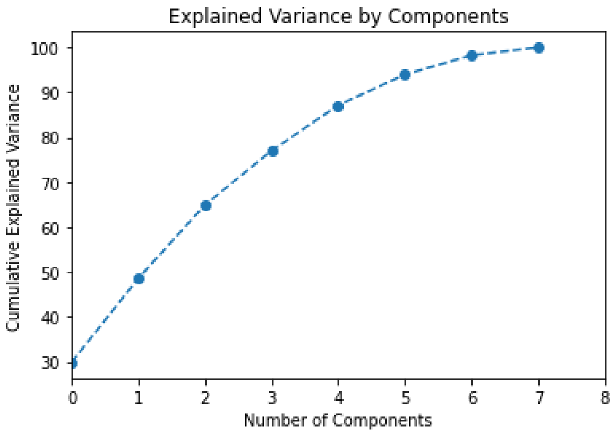




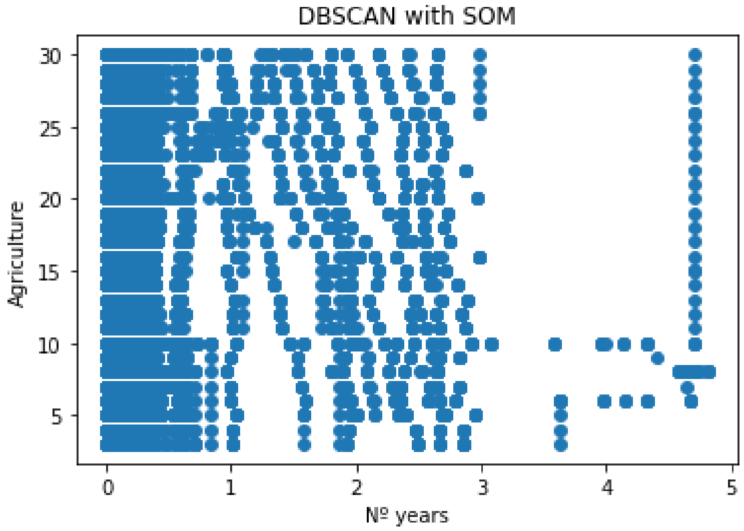
Annex 15- Cumulative explained variance for Emissions



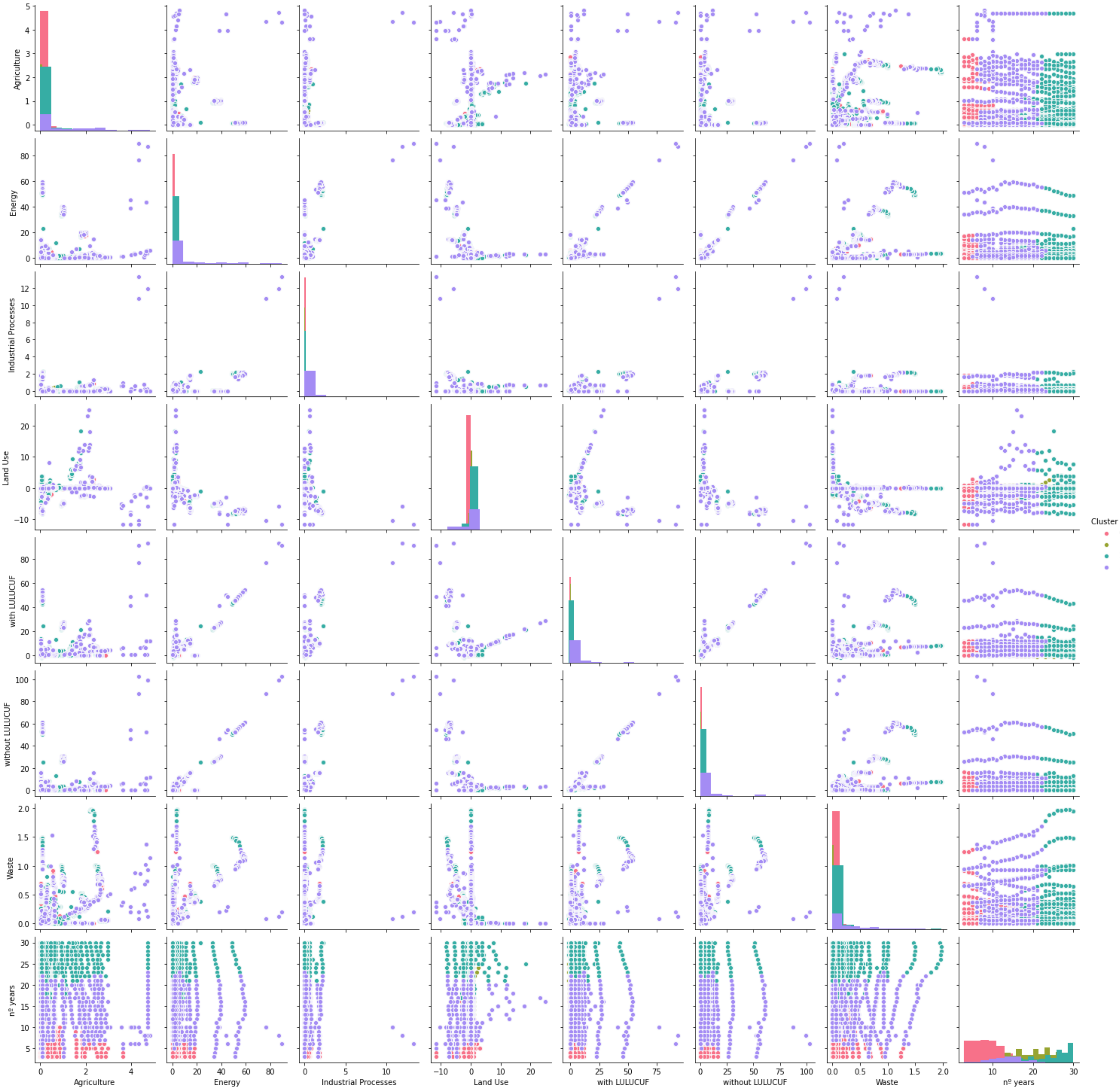
Annex 16- Cumulative explained variance for Country Characteristics



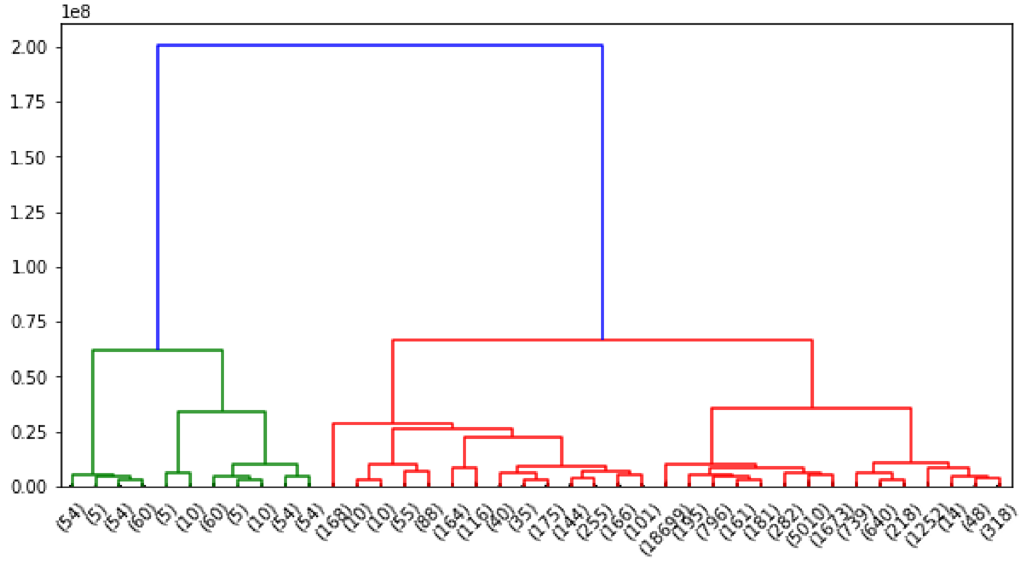
Annex 17- Emissions clustering: DBSCAN



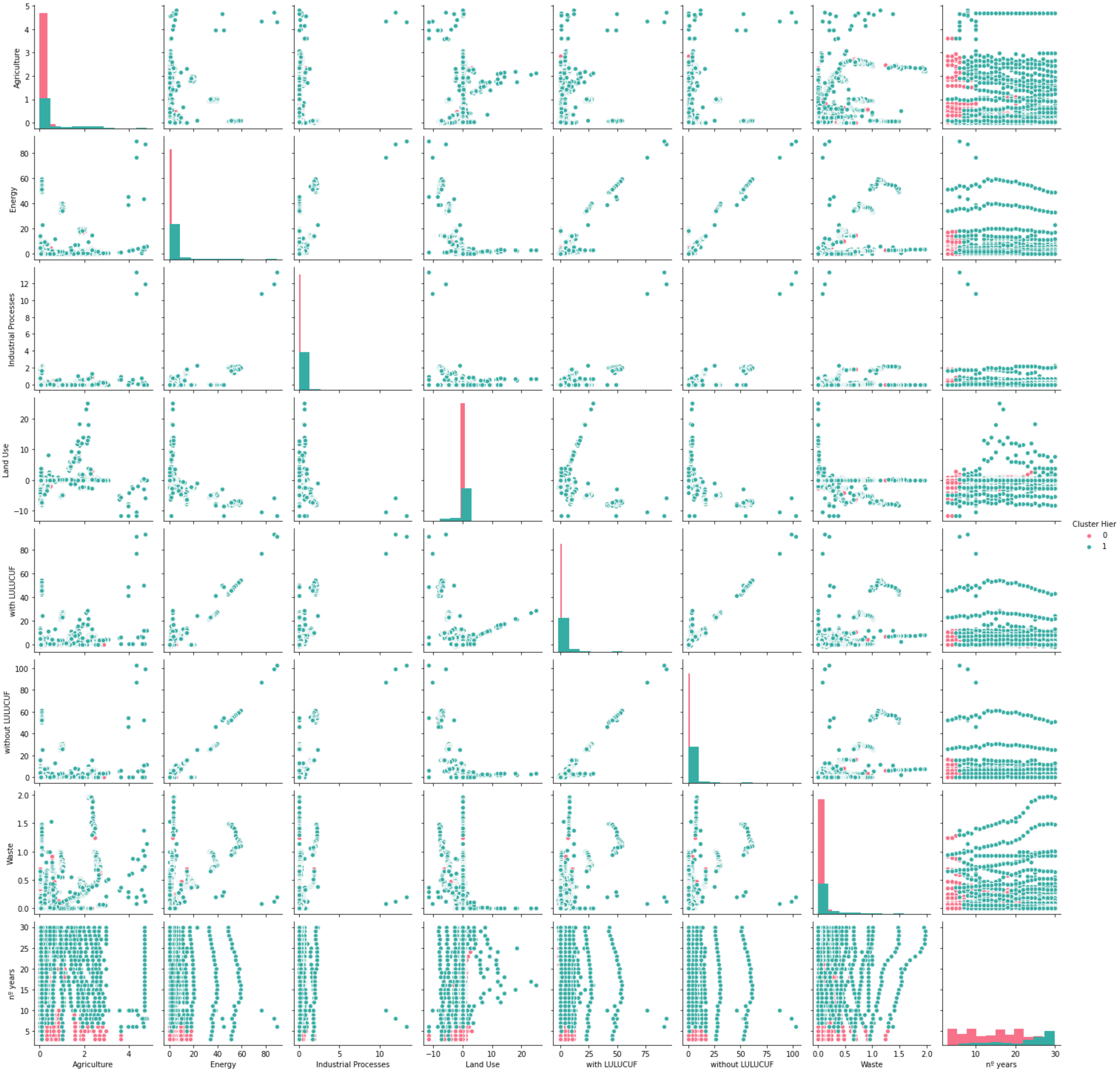
Annex 18- Emissions clustering: Mean-Shift



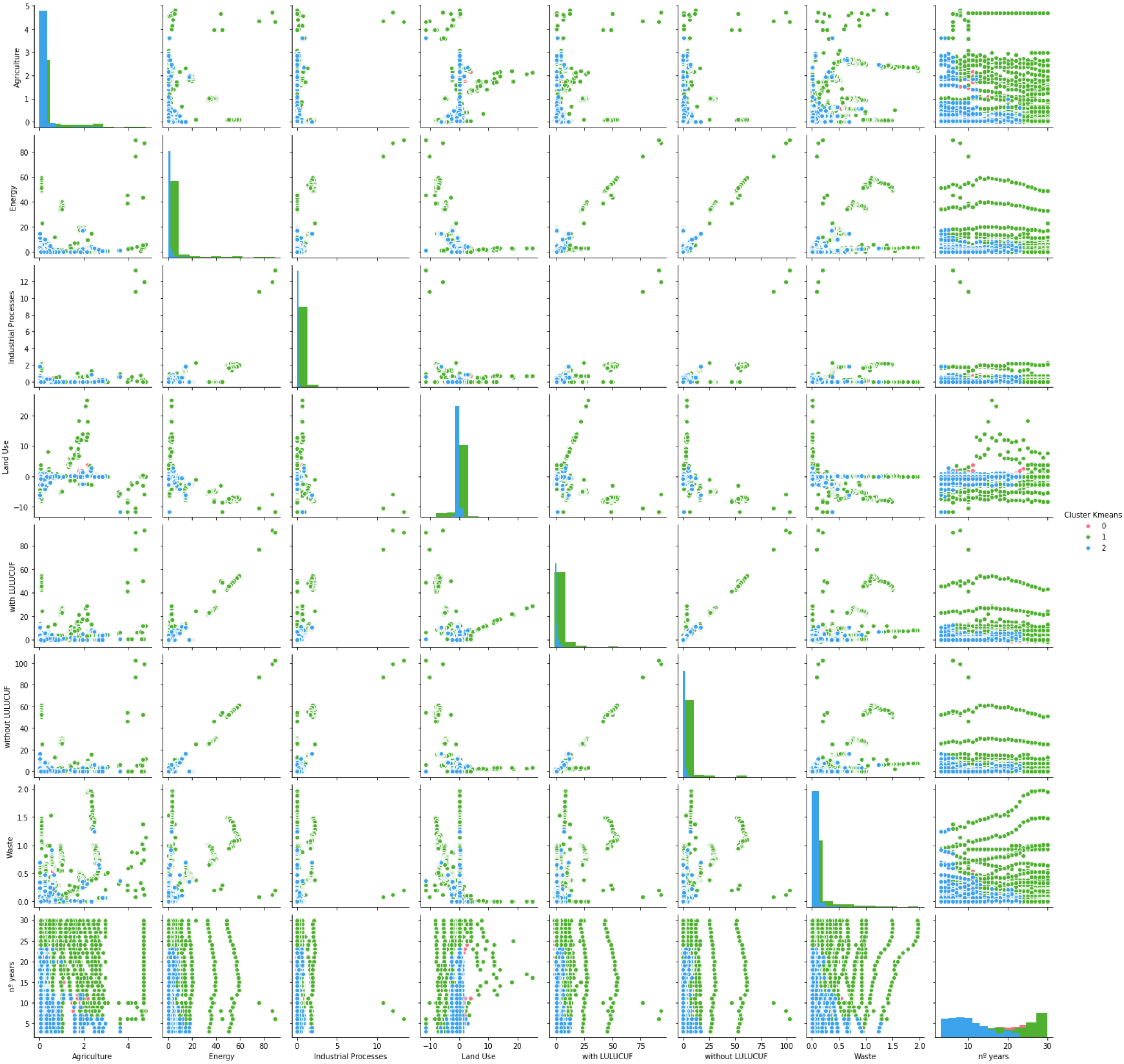
Annex 19- Emissions clustering: dendrogram



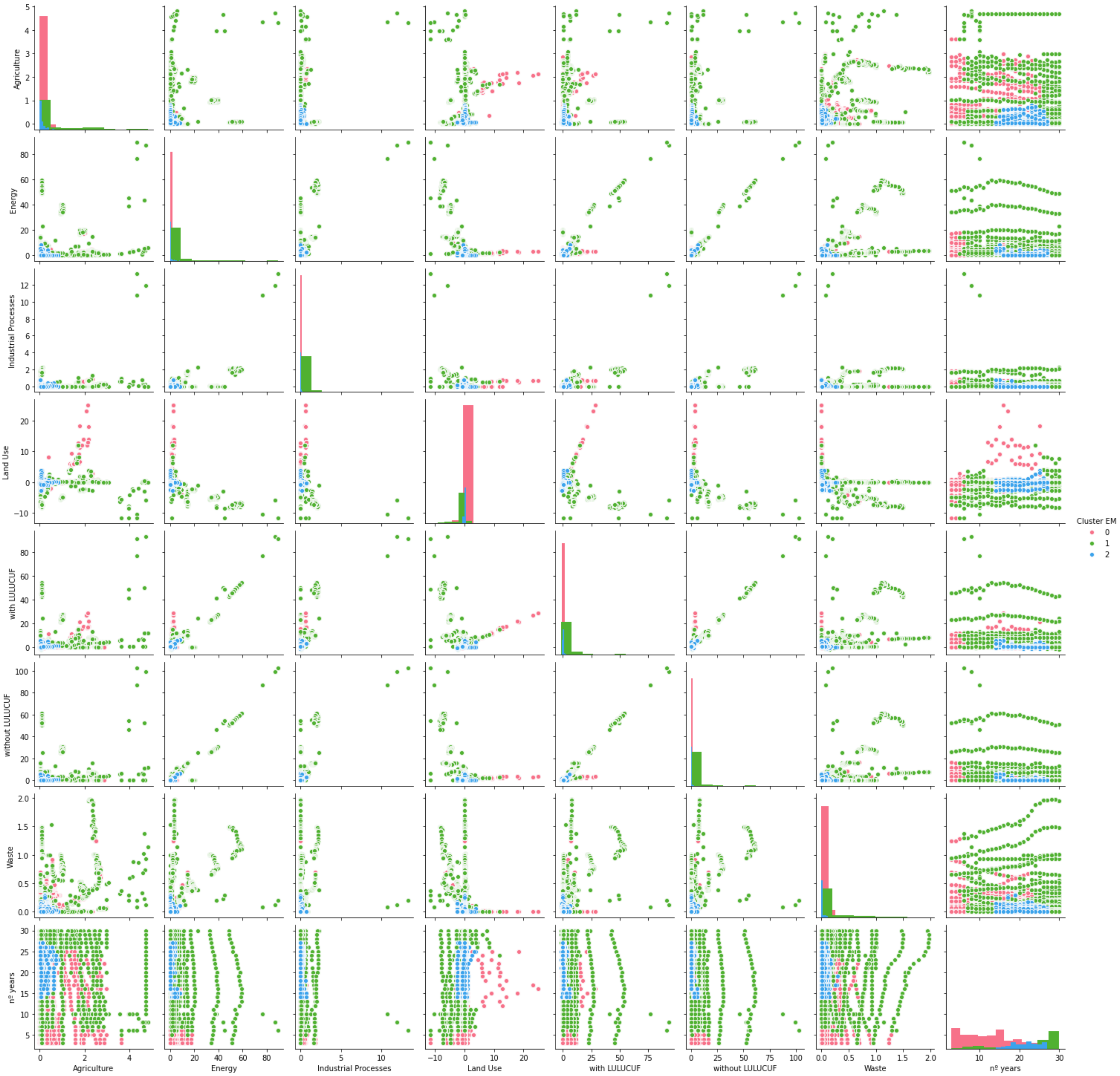
Annex 20- Emissions clustering: Hierarchical



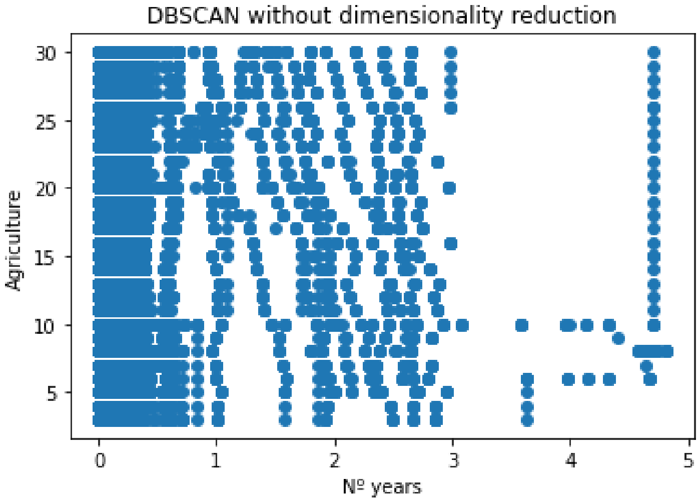
Annex 21- Emissions clustering: K-Means



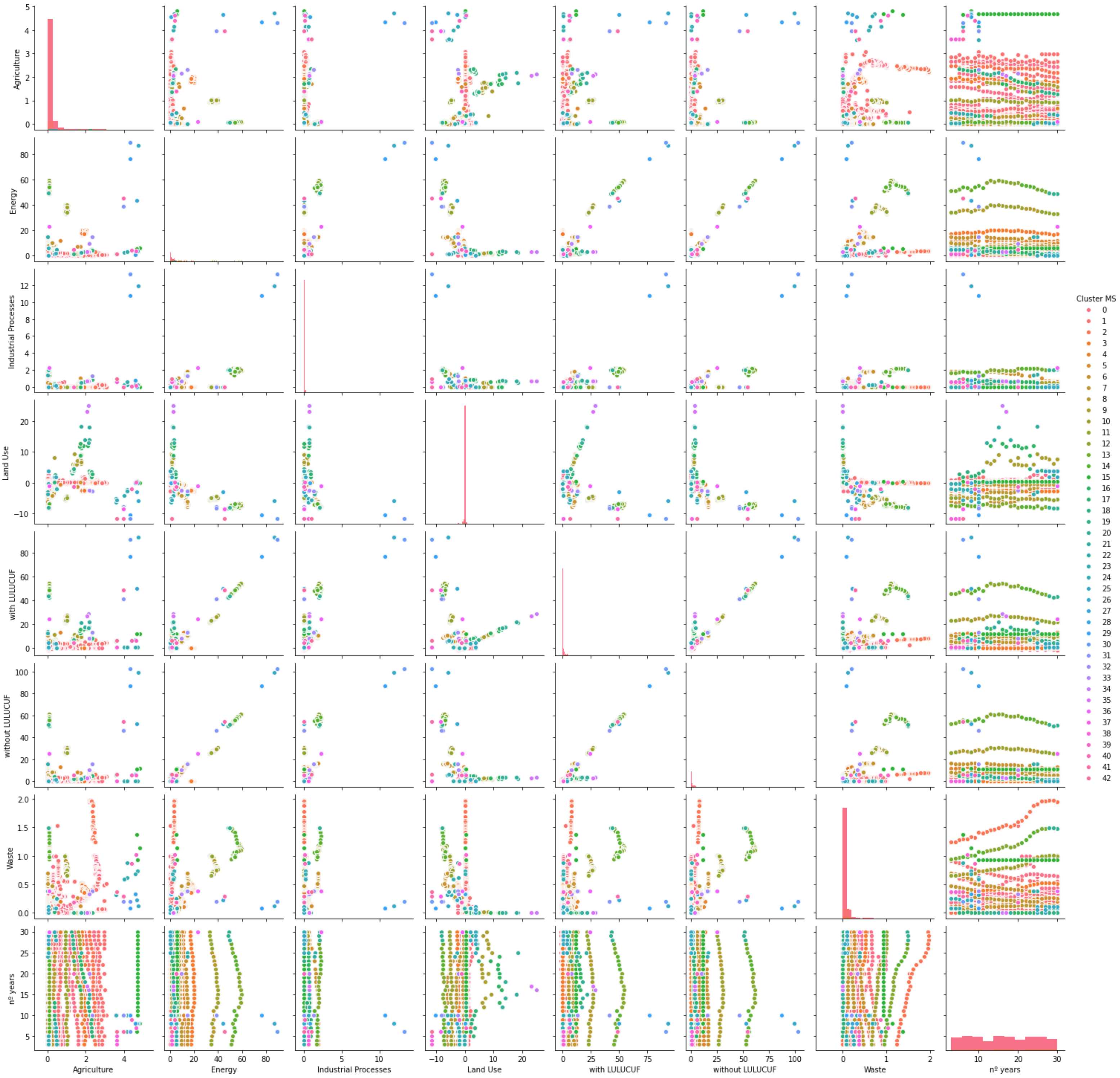
Annex 22- Emissions clustering: Expectation-Maximization



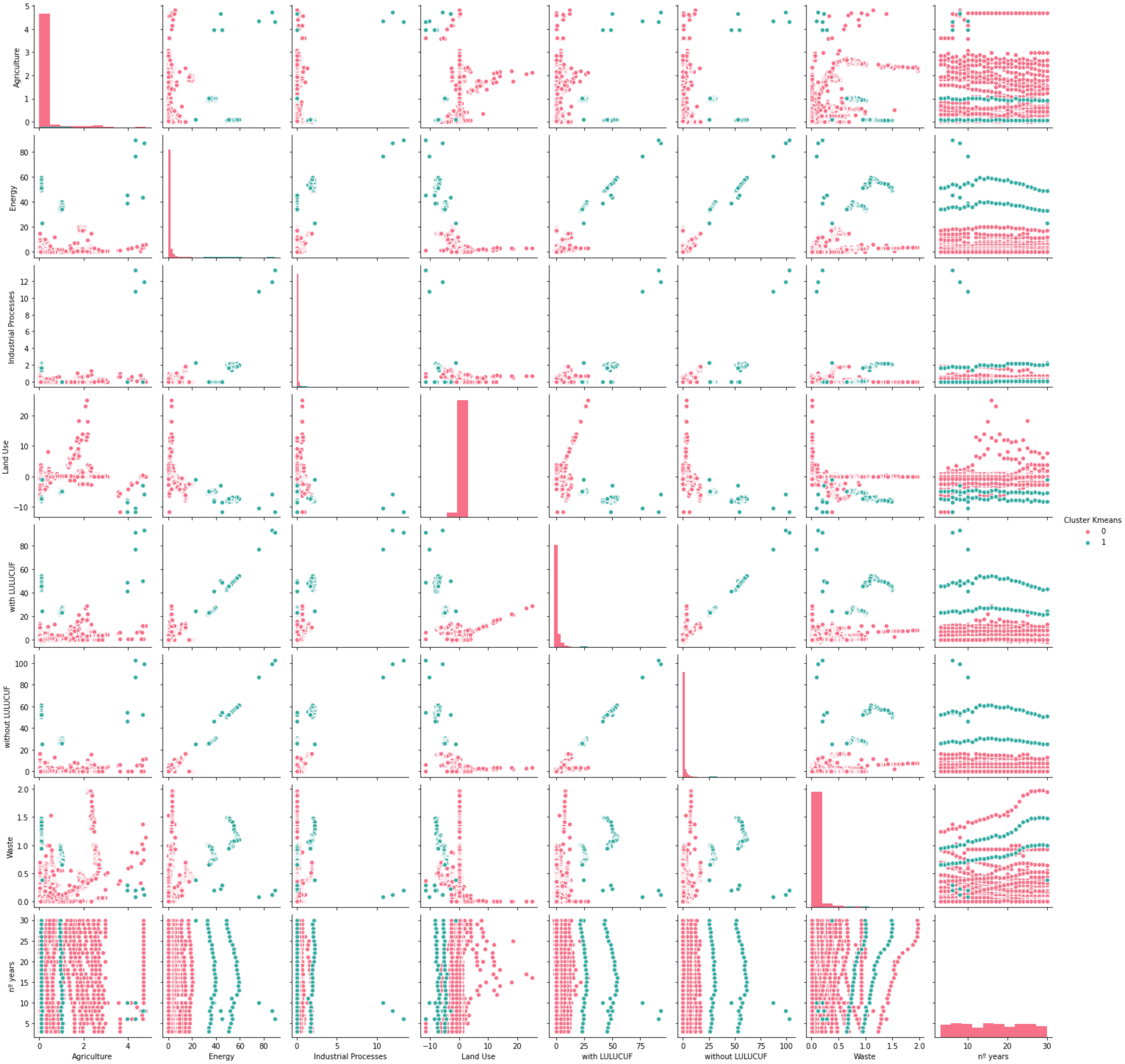
Annex 23- Emissions without dimensionality reduction clustering: DBSCAN



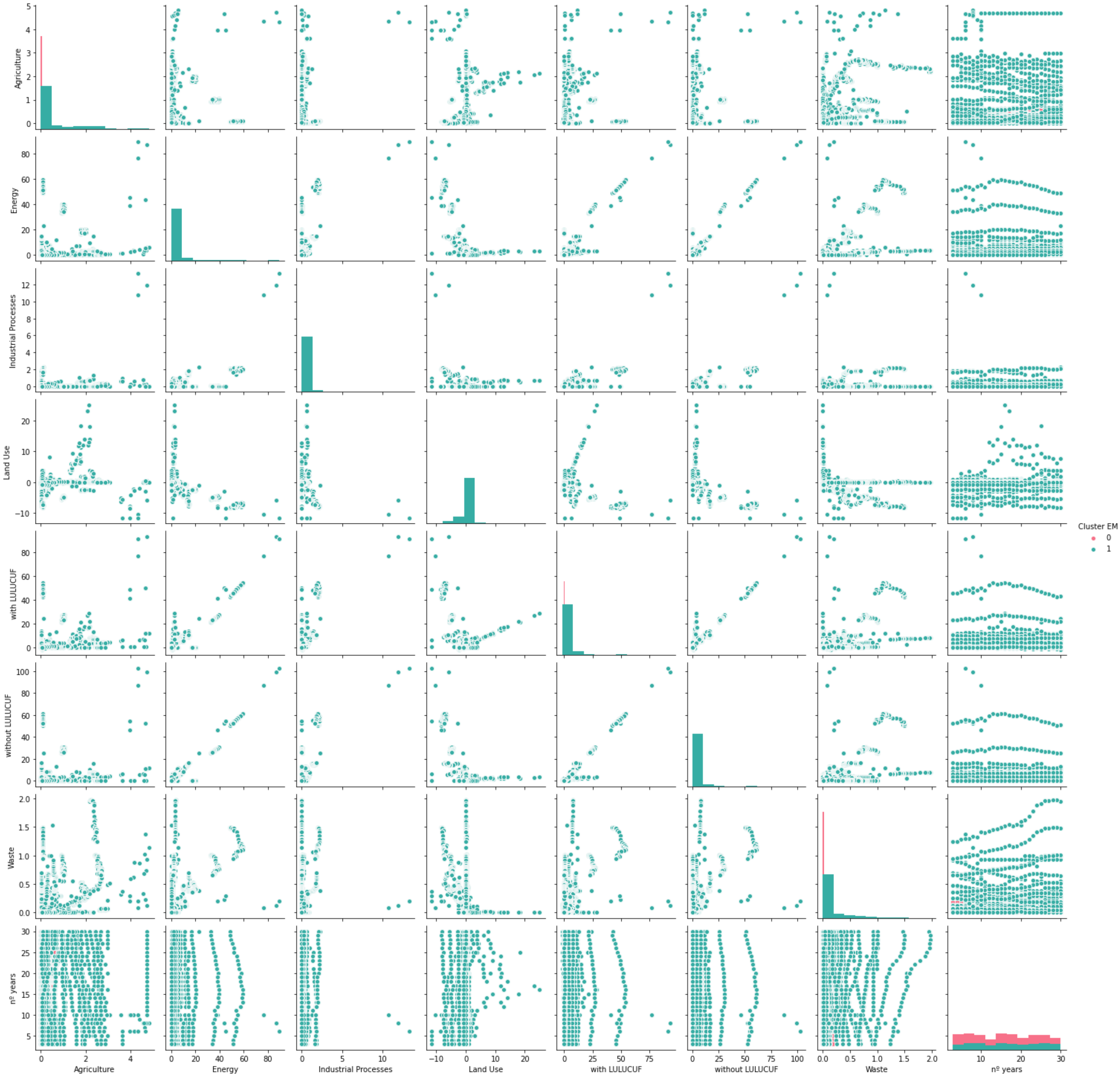
Annex 24- Emissions without dimensionality reduction clustering: Mean-Shift



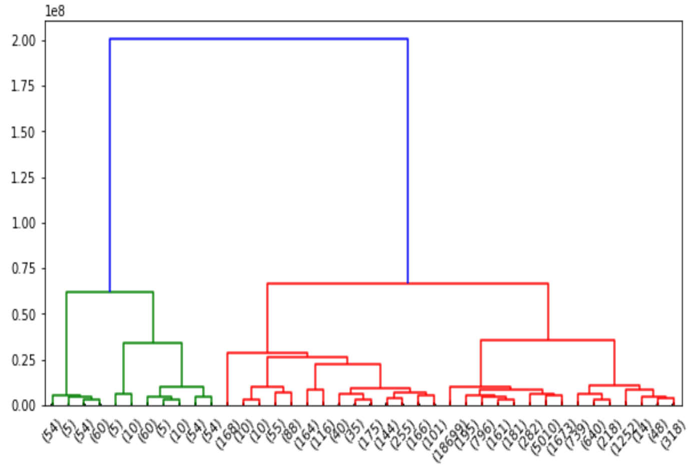
Annex 25- Emissions without dimensionality reduction clustering: K-means



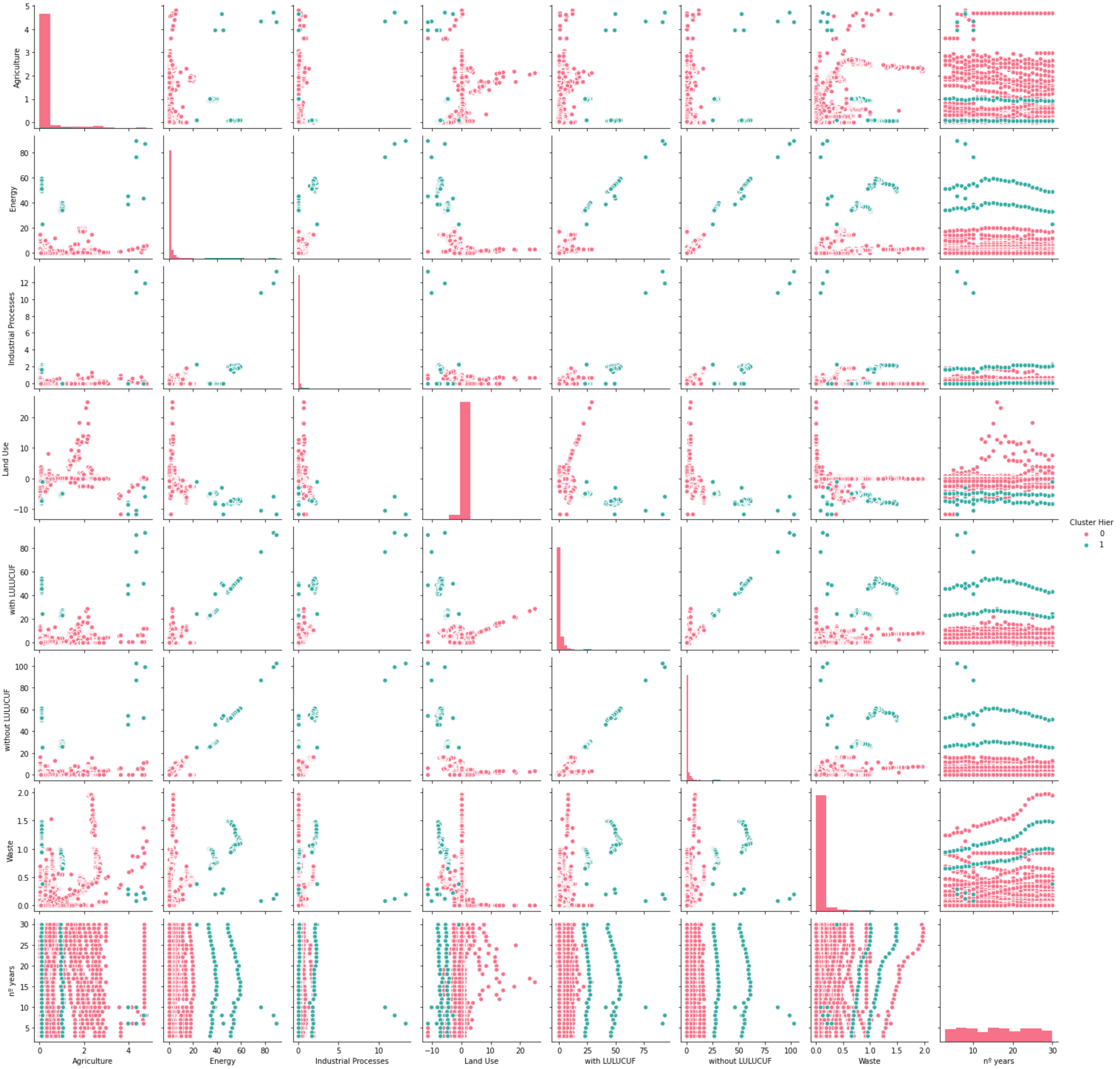
Annex 26- Emissions without dimensionality reduction clustering: Expectation-Maximization



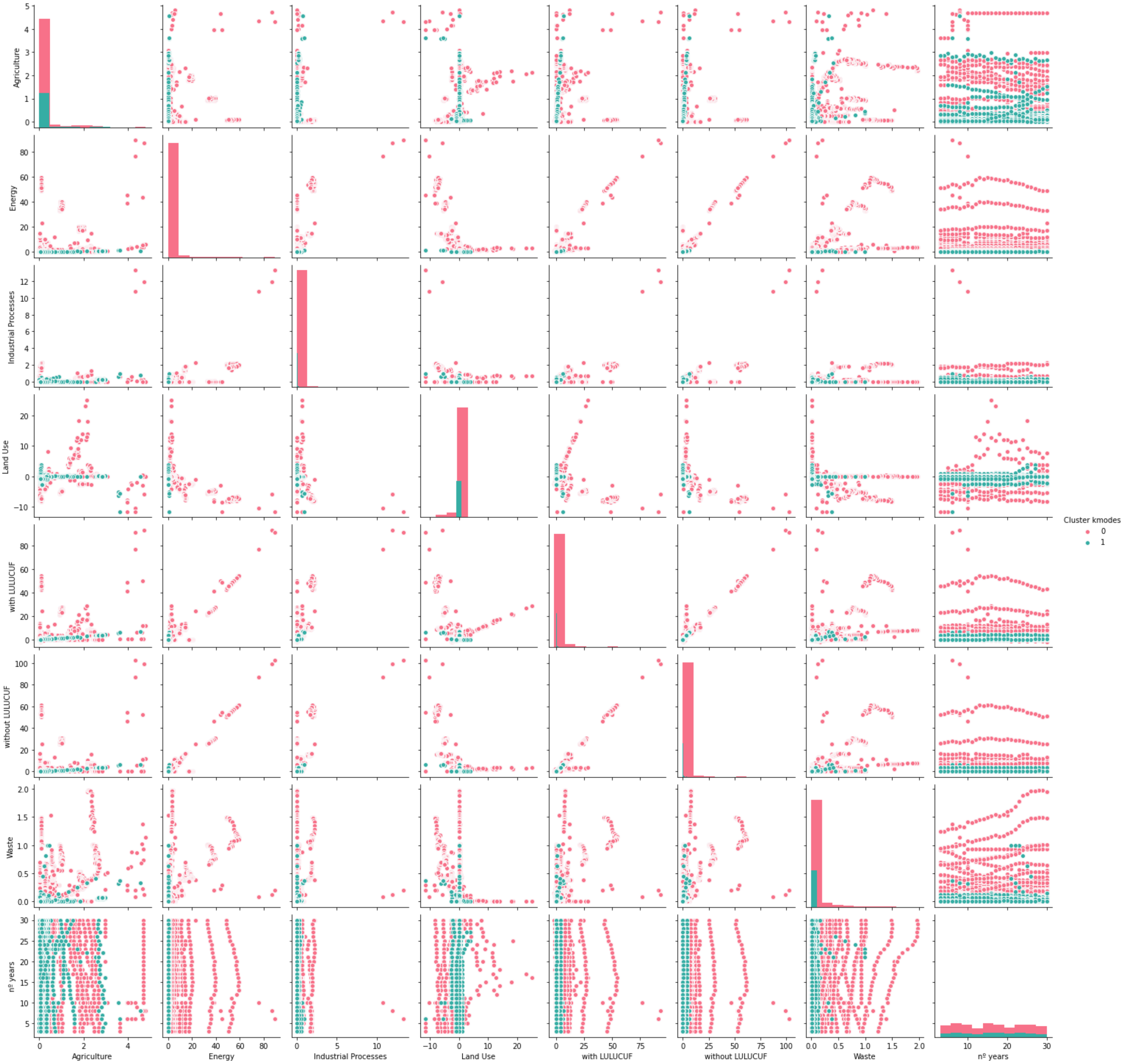
Annex 27- Emissions without dimensionality reduction clustering: dendrogram



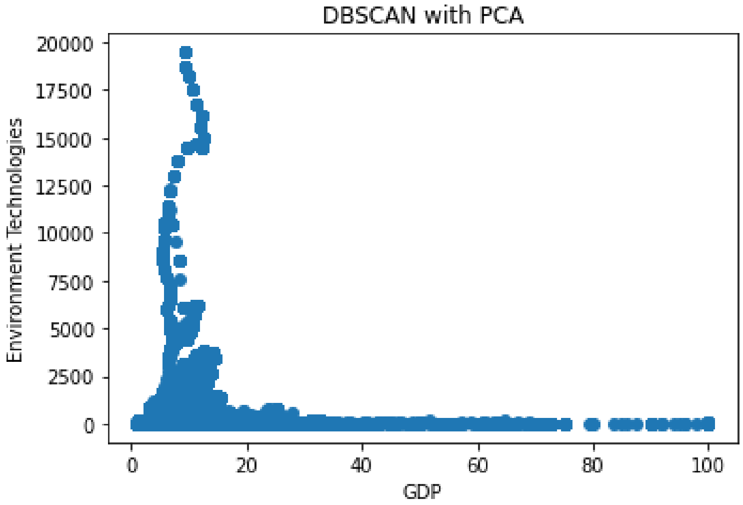
Annex 28- Emissions without dimensionality reduction clustering: Hierarchical



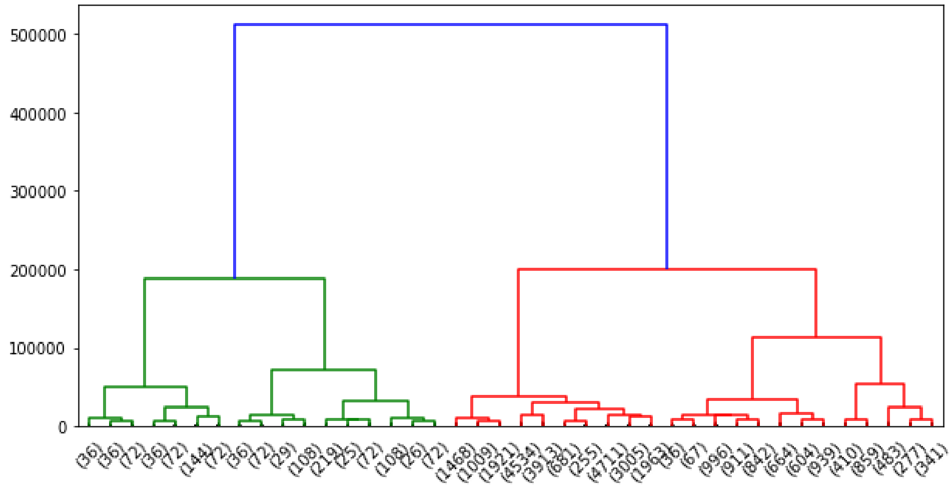
Annex 29- Emissions without dimensionality reduction clustering: K-Modes



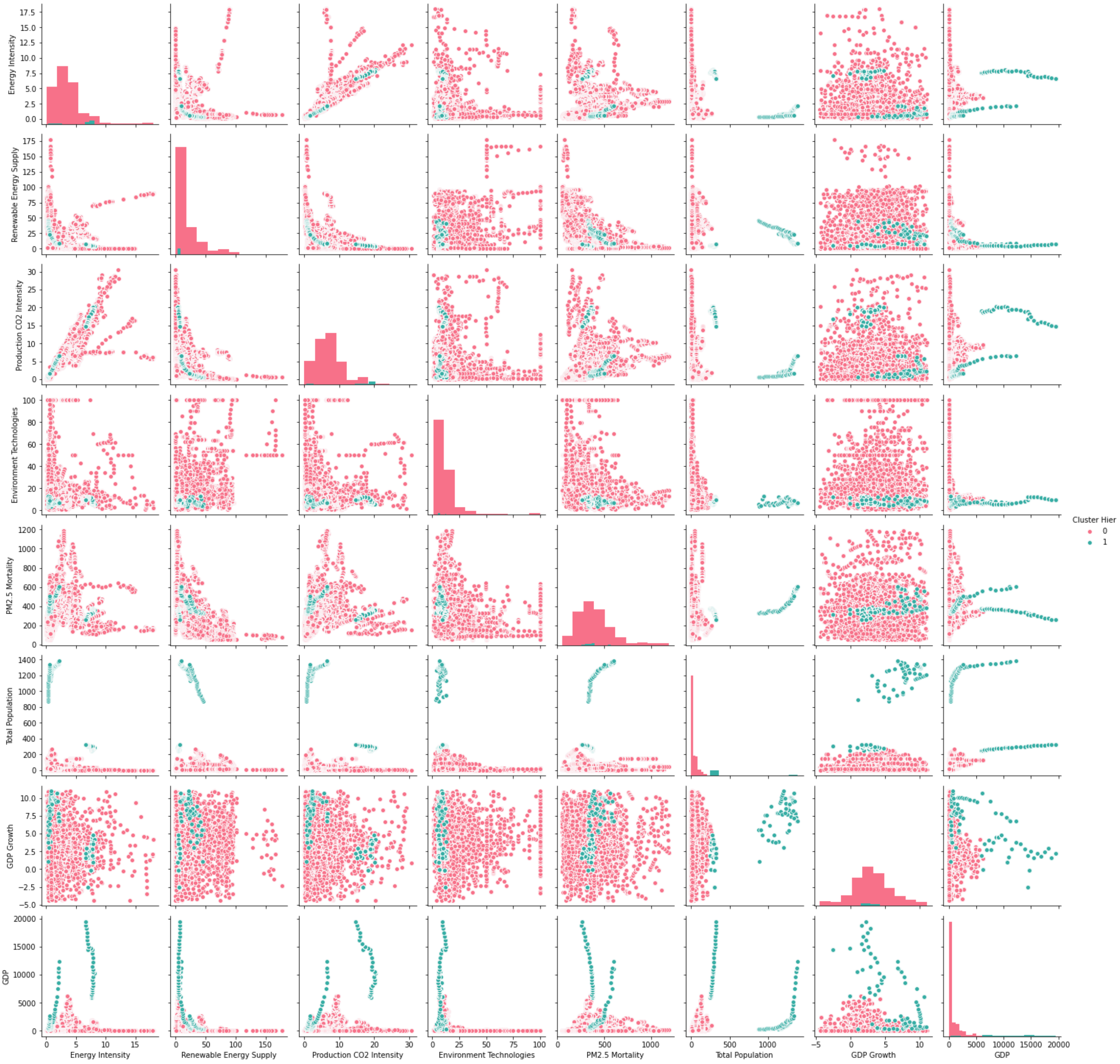
Annex 30- Country Characteristics clustering: DBSCAN



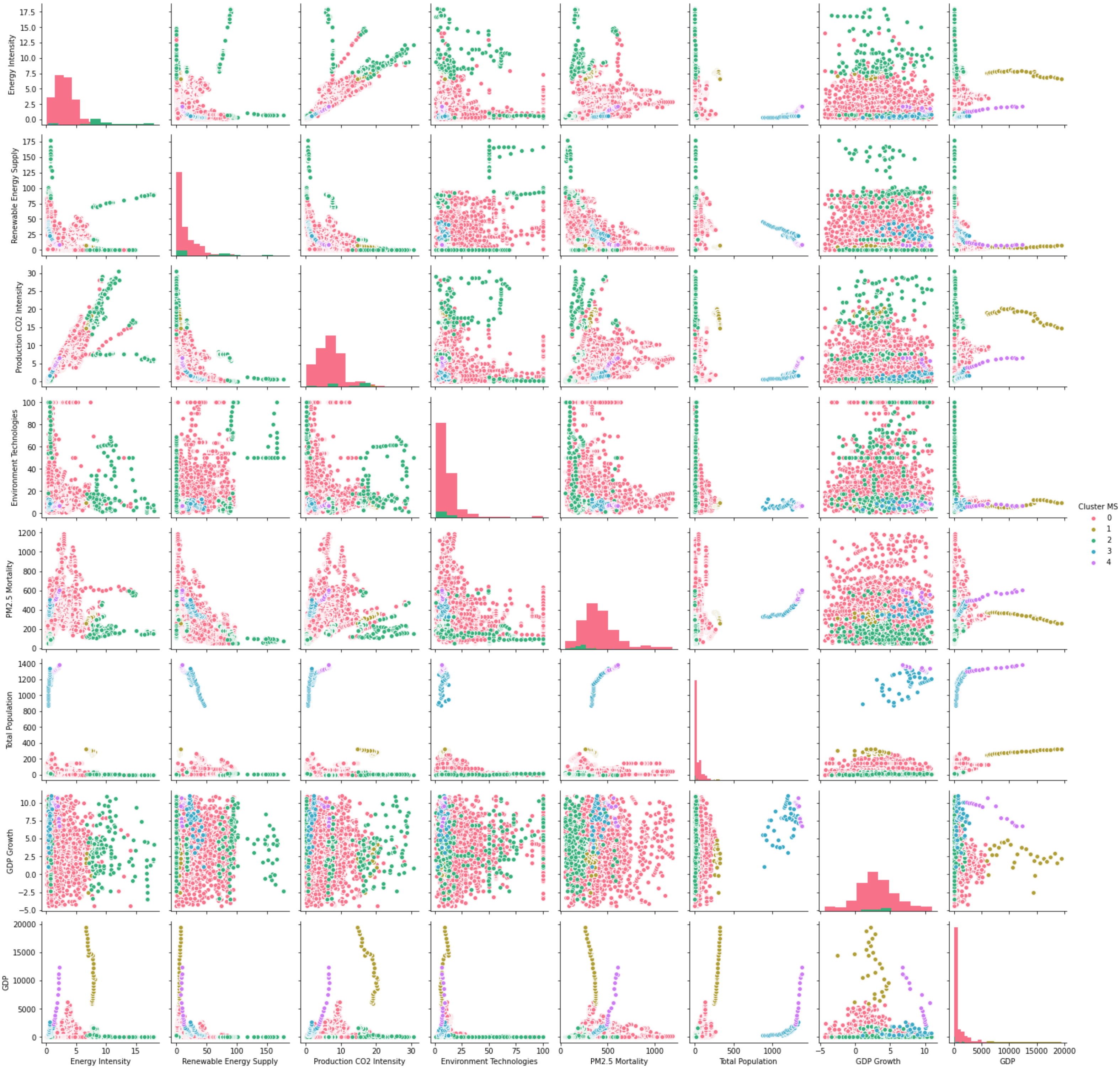
Annex 31- Country Characteristics clustering: dendrogram



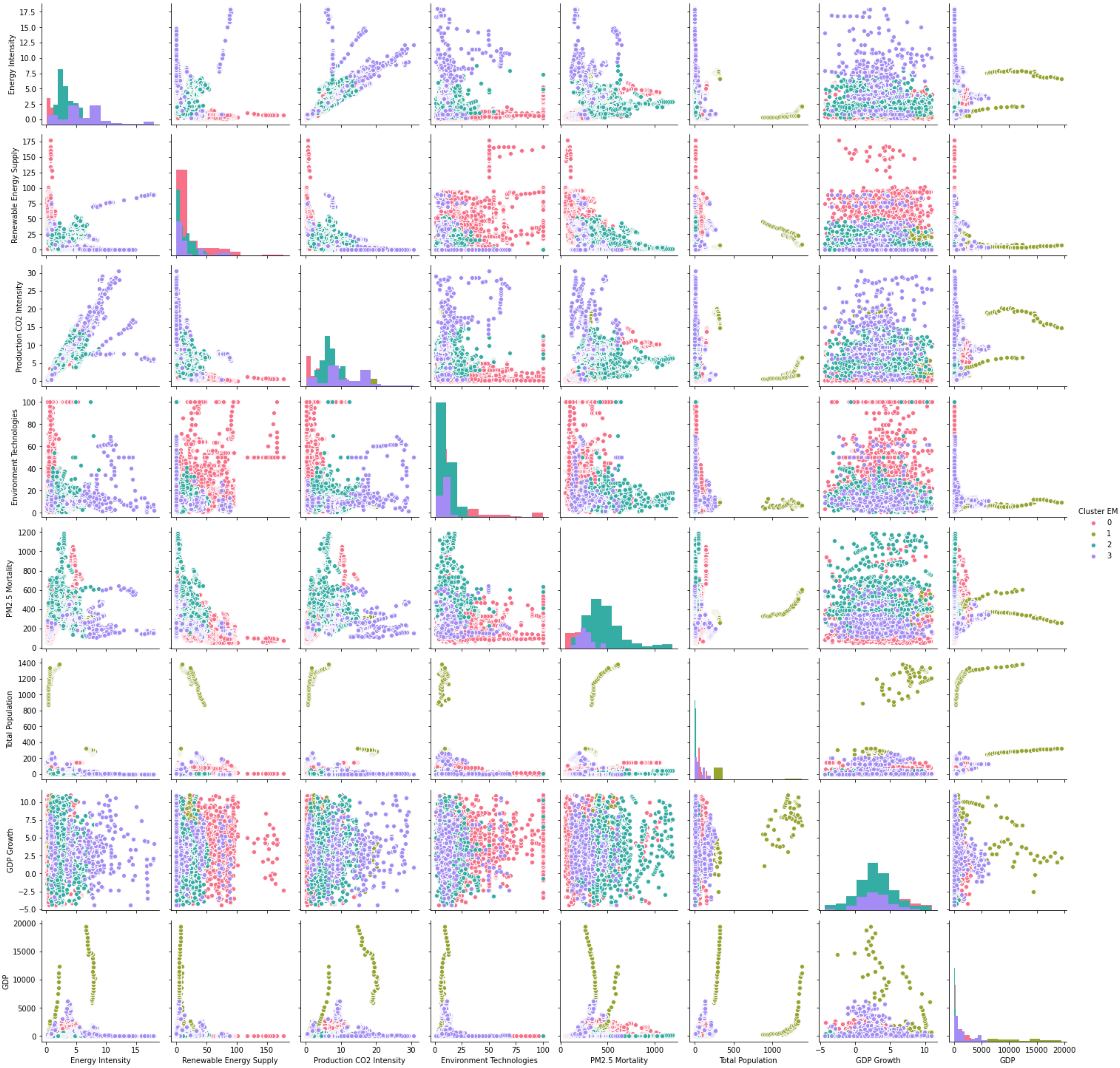
Annex 32- Country Characteristics clustering: Hierarchical



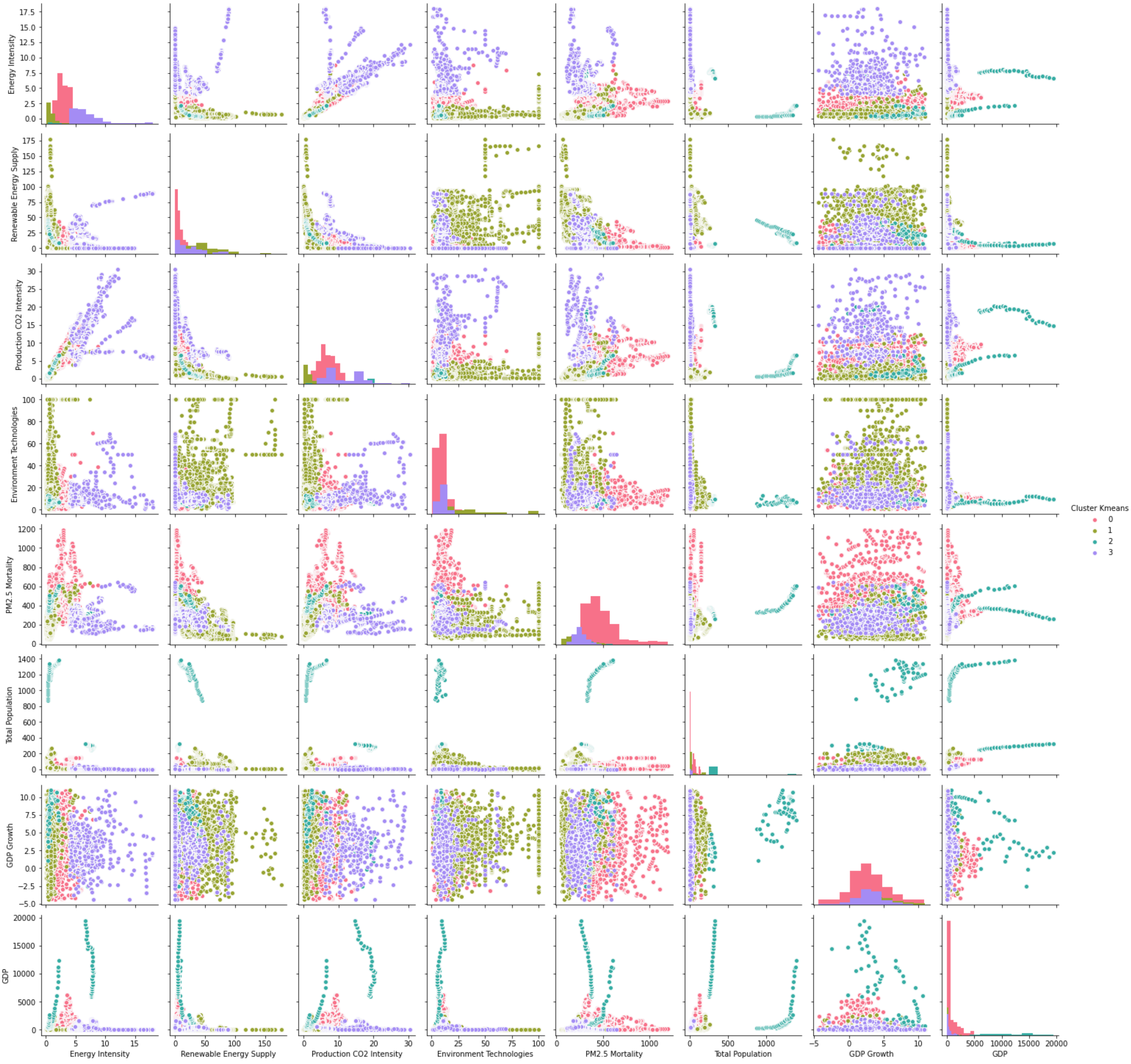
Annex 33- Country Characteristics clustering: Mean-Shift



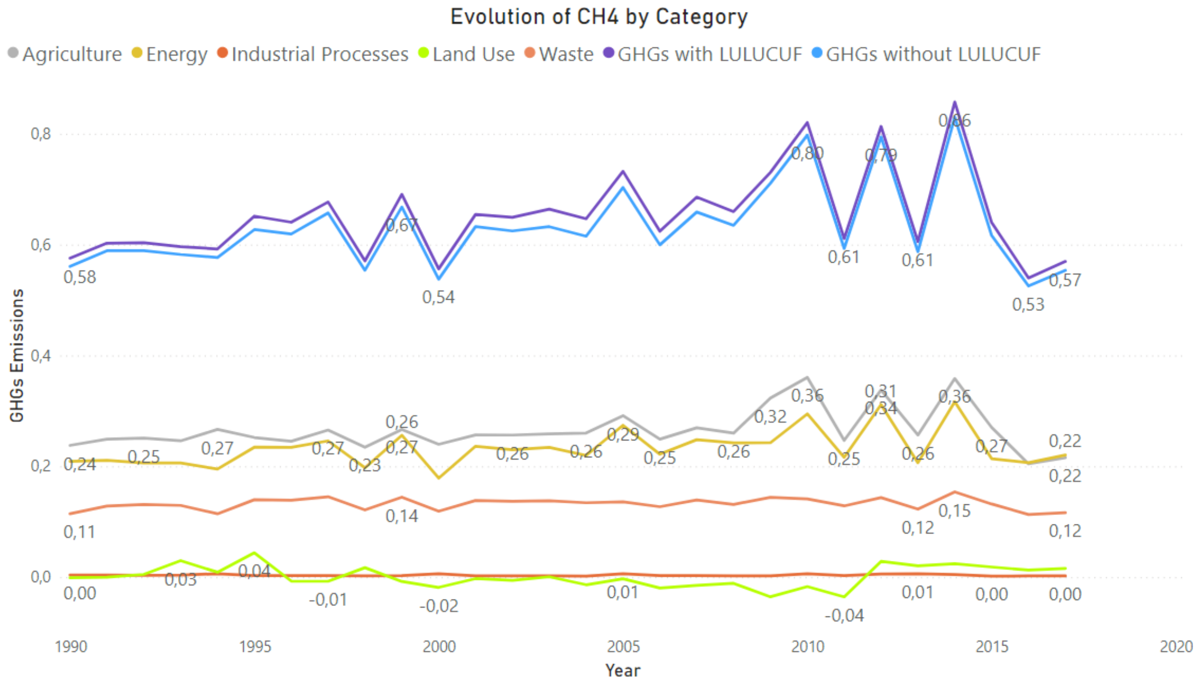
Annex 34- Country Characteristics clustering: Expectation-Maximization



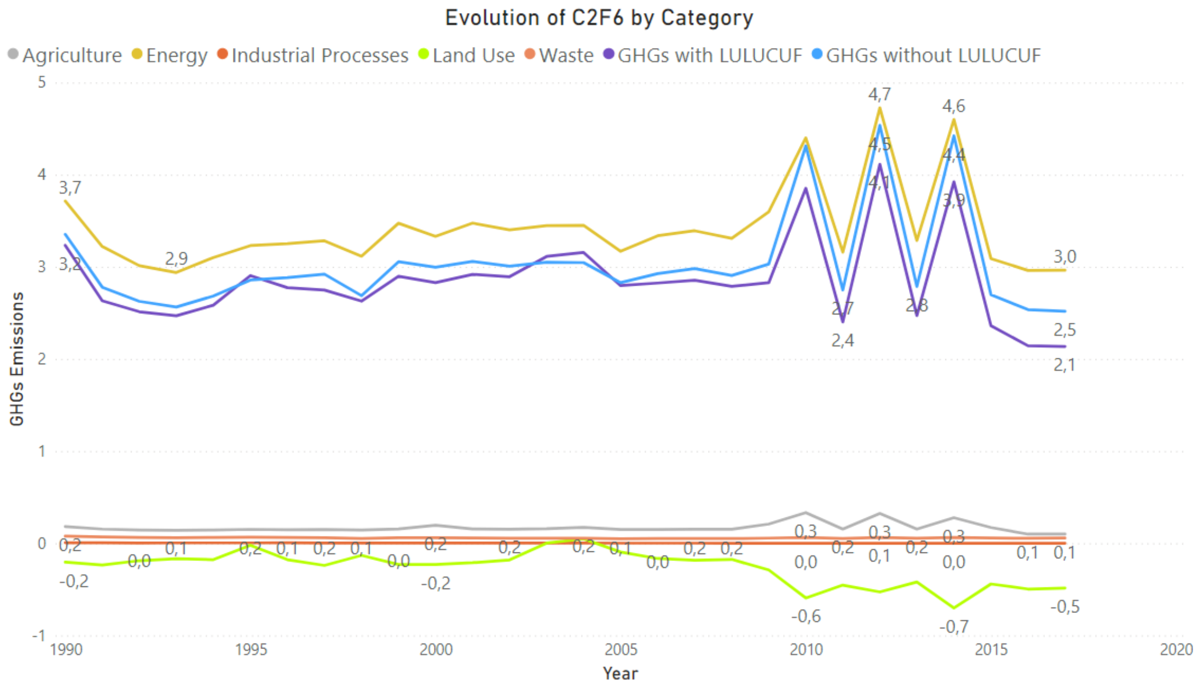
Annex 35- Country Characteristics clustering: K-Means



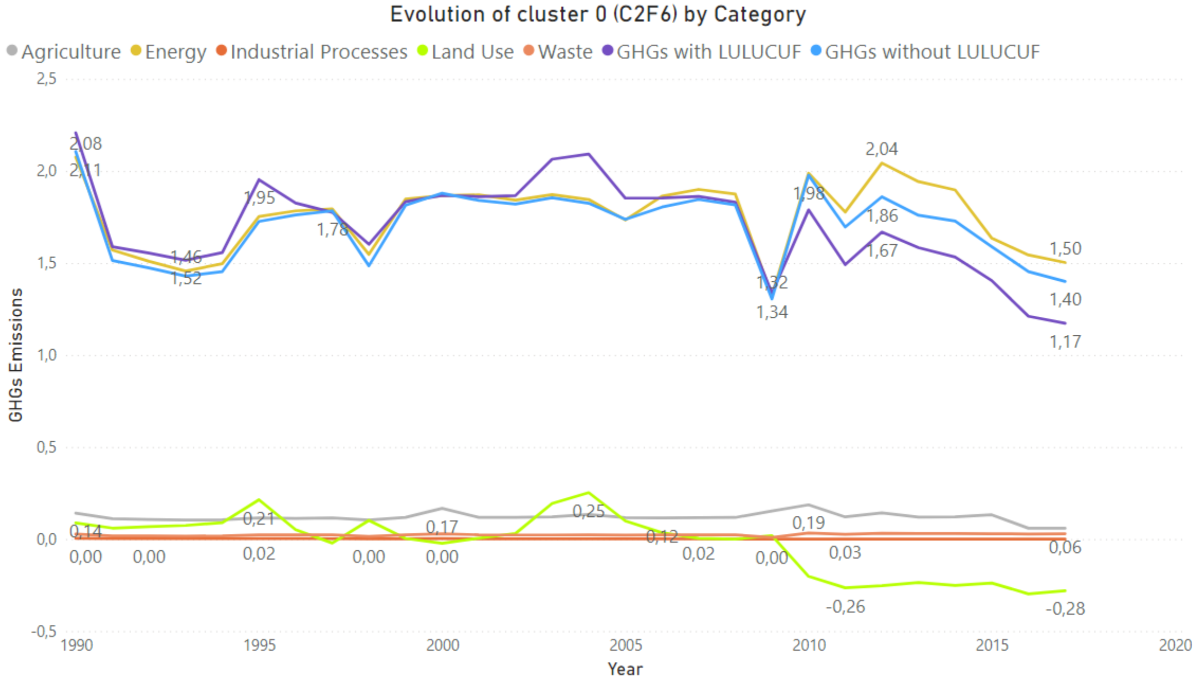
Annex 36- Evolution of CH_4 by category.



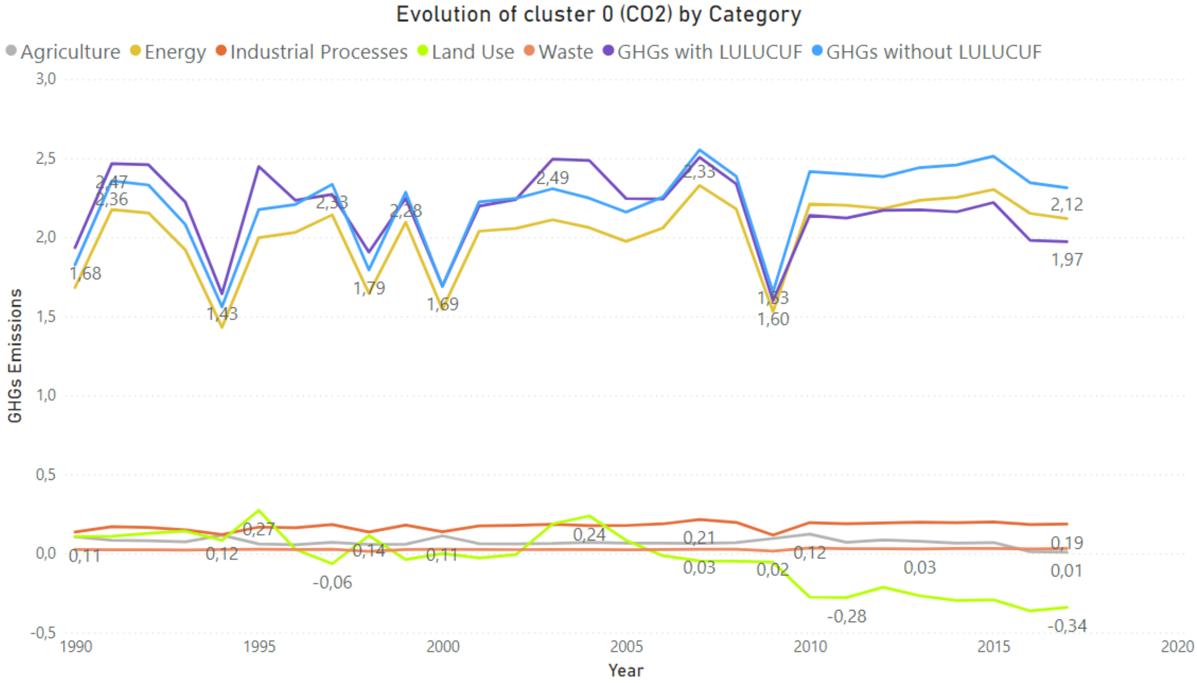
Annex 37- Evolution of C_2F_6 by category.



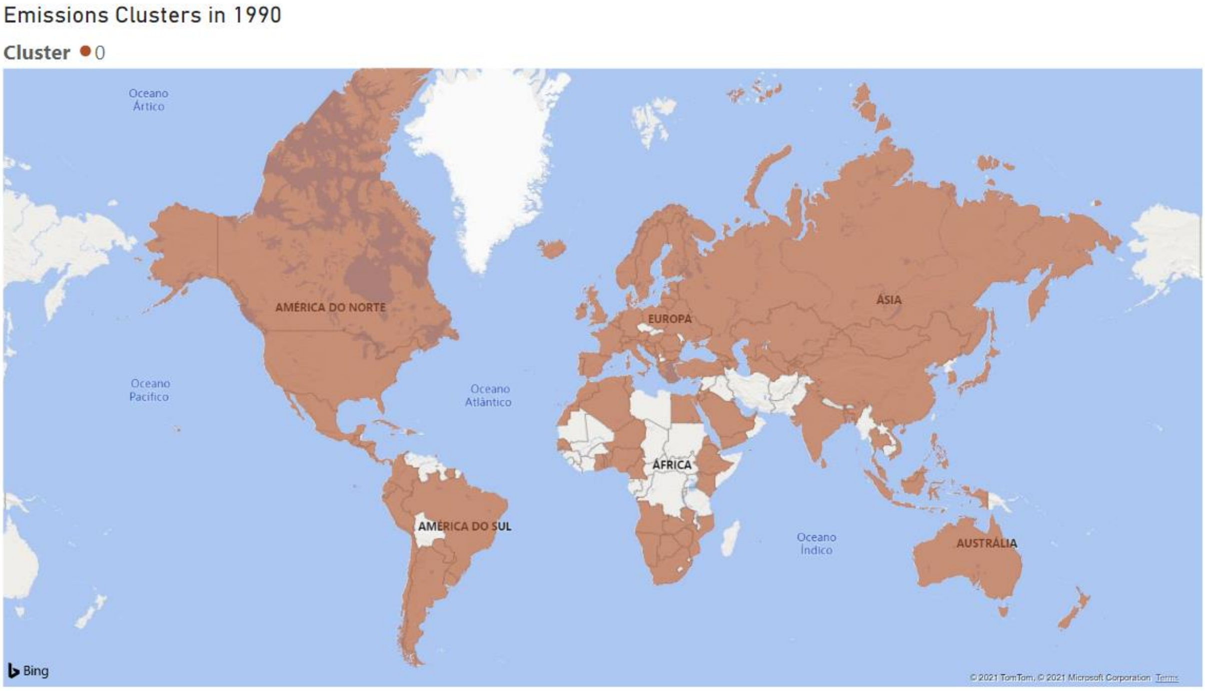
Annex 38- Evolution of C_2F_6 in cluster 0 by category.



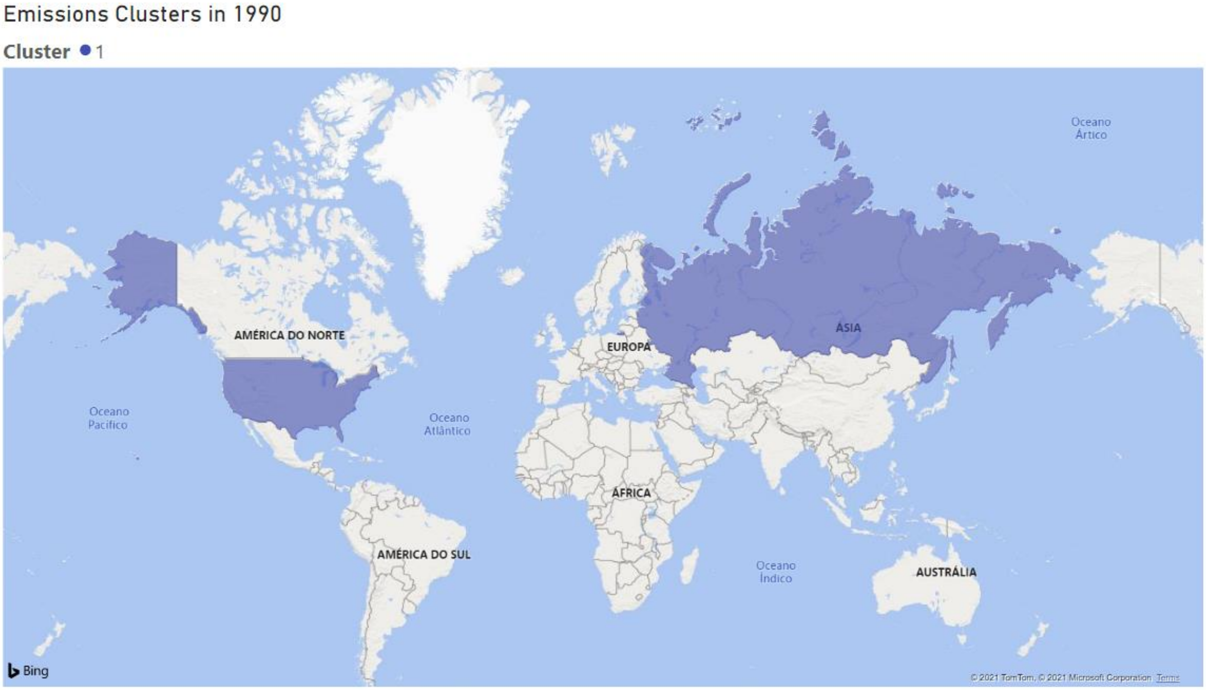
Annex 39- Evolution of CO_2 in cluster 0 by category.



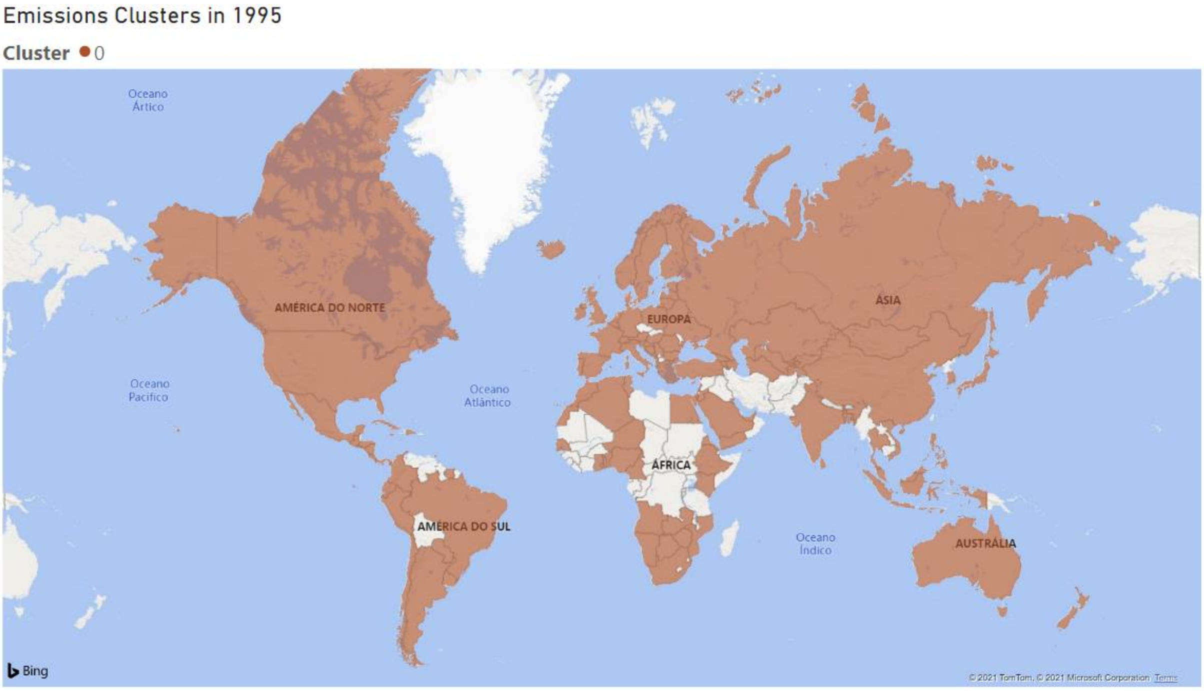
Annex 40- Countries belonging to cluster 0 in Emissions (1990)



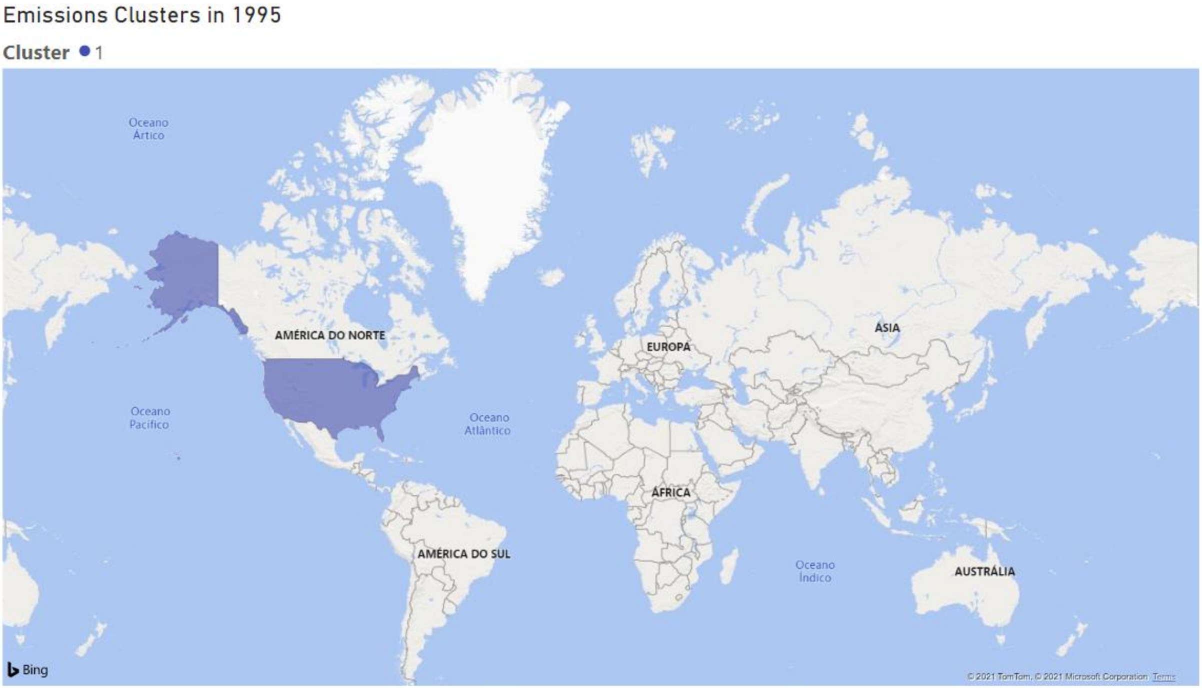
Annex 41- Countries belonging to cluster 1 in Emissions (1990)



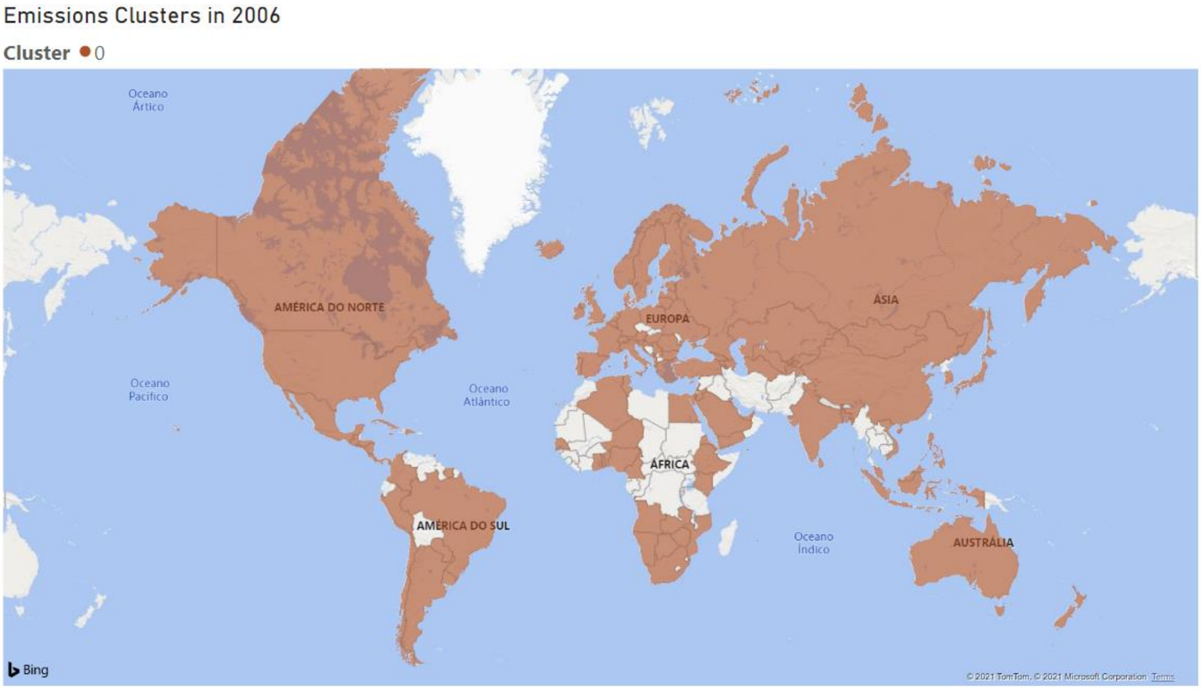
Annex 42- Countries belonging to cluster 0 in Emissions (1995)



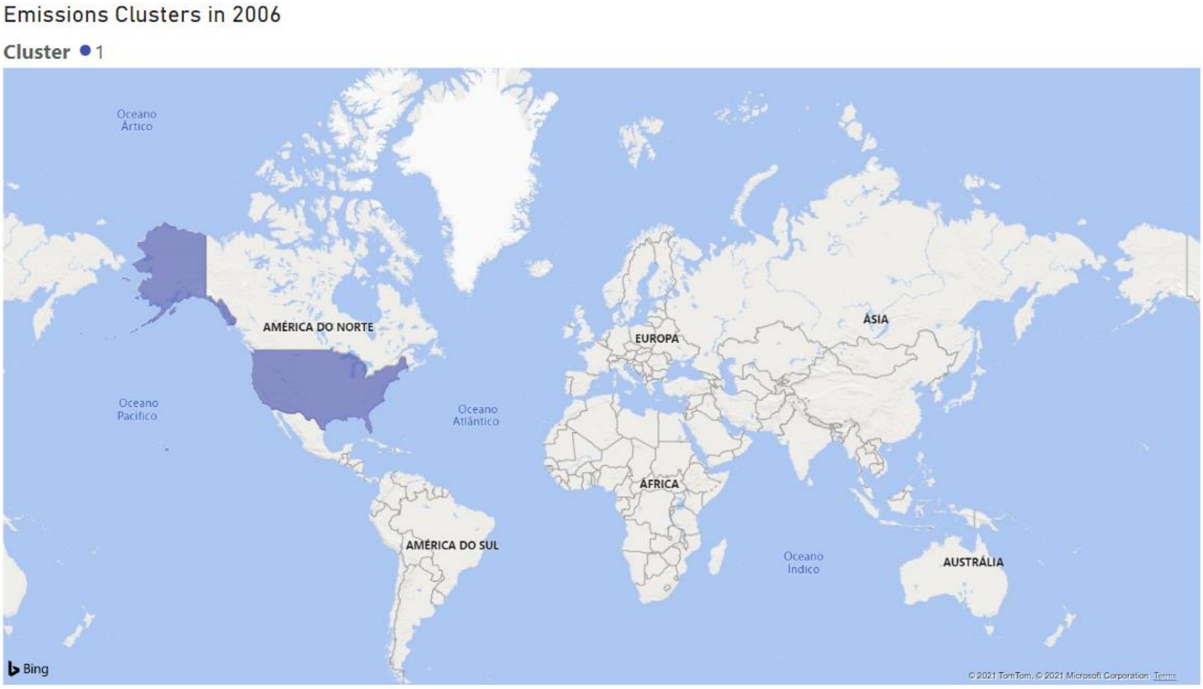
Annex 43- Countries belonging to cluster 1 in Emissions (1995)



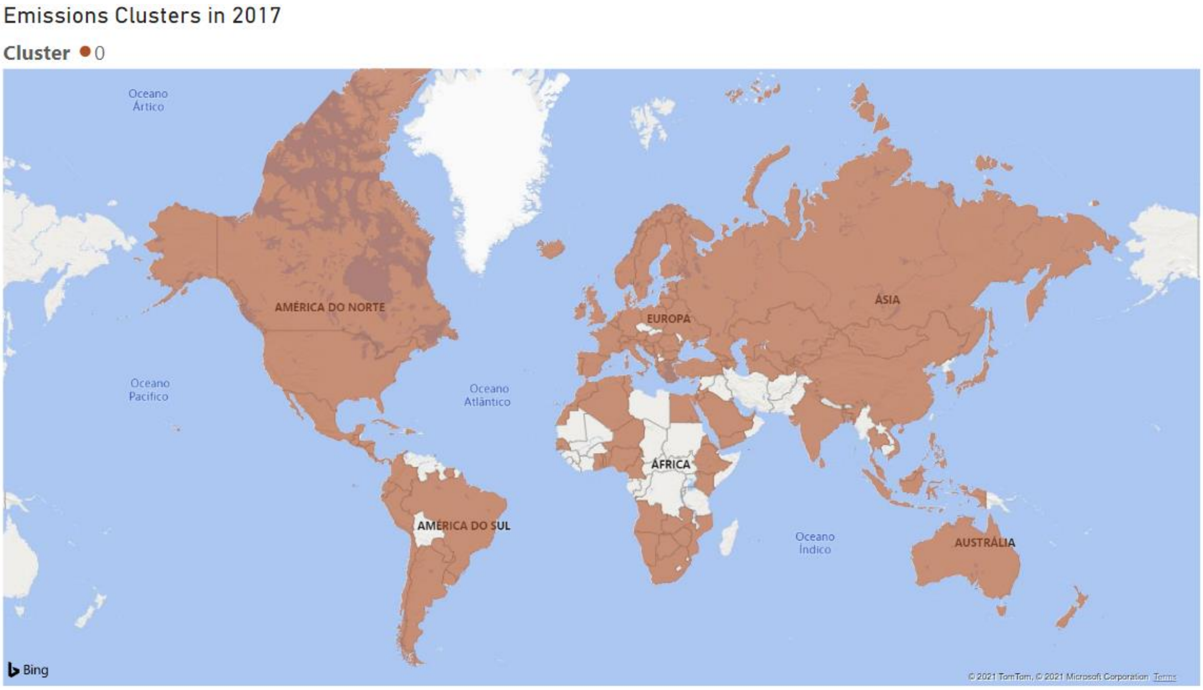
Annex 44- Countries belonging to cluster 0 in Emissions (2006)



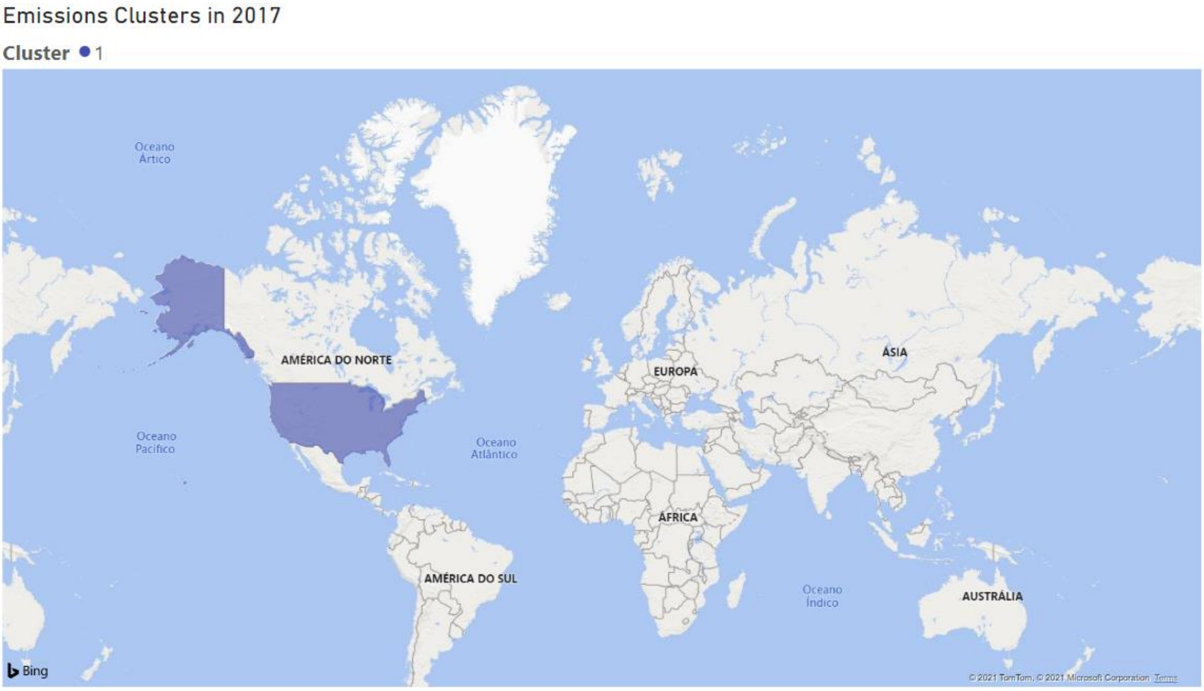
Annex 45- Countries belonging to cluster 1 in Emissions (2006)



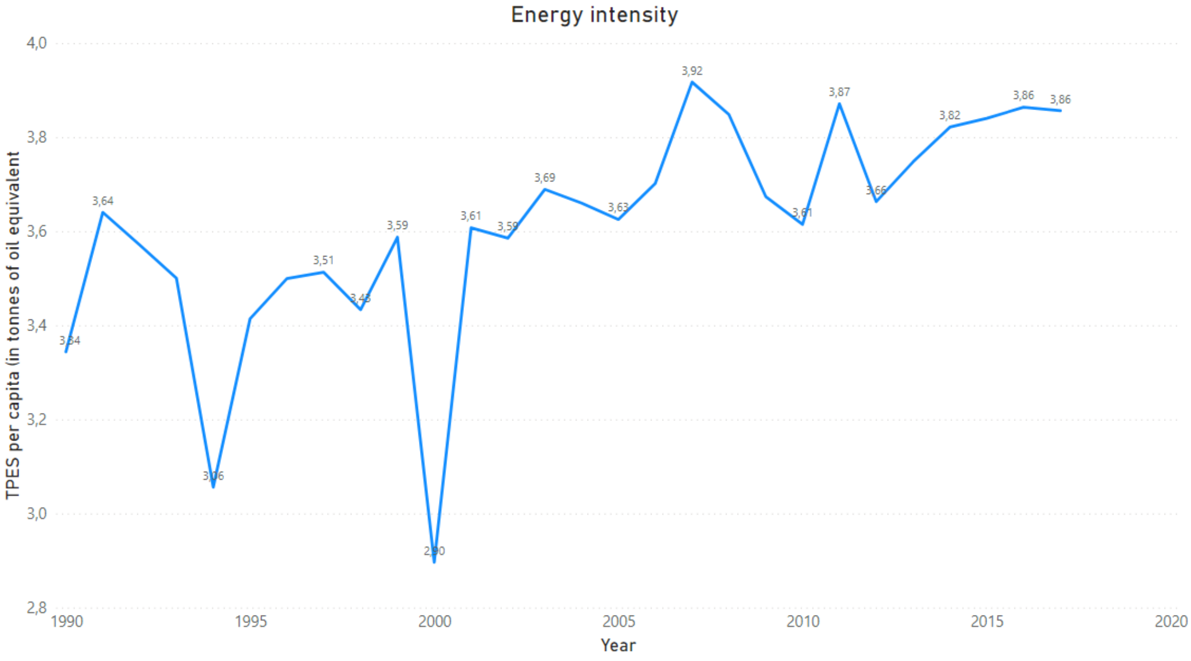
Annex 46- Countries belonging to cluster 0 in Emissions (2017)



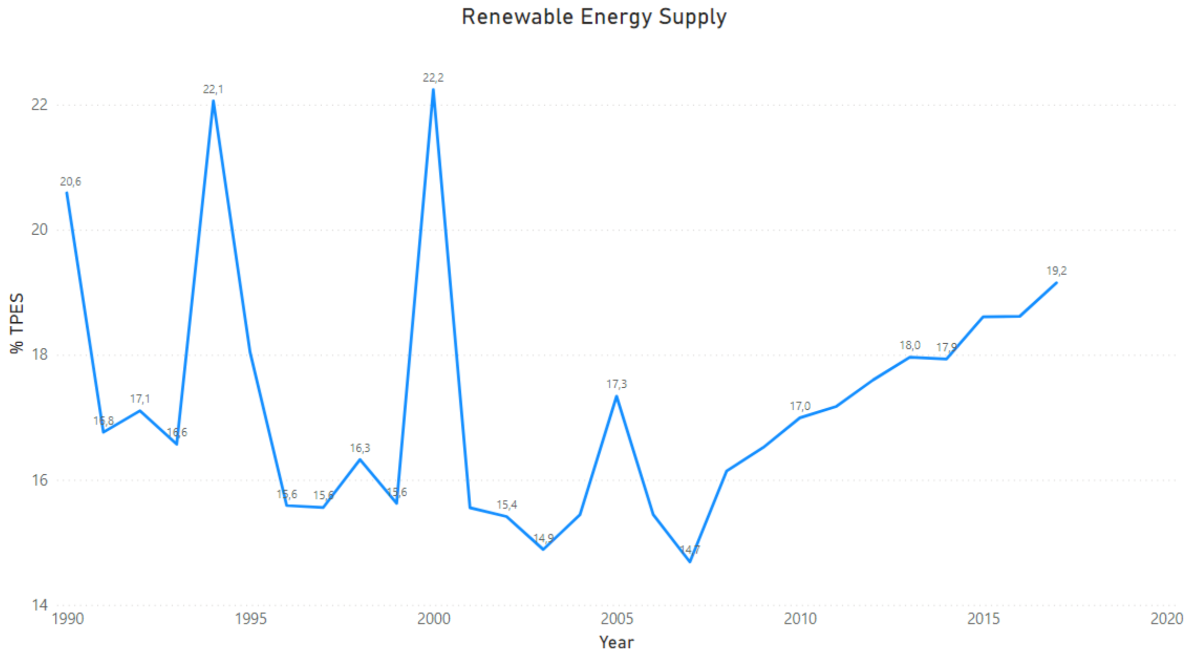
Annex 47- Countries belonging to cluster 1 in Emissions (2017)



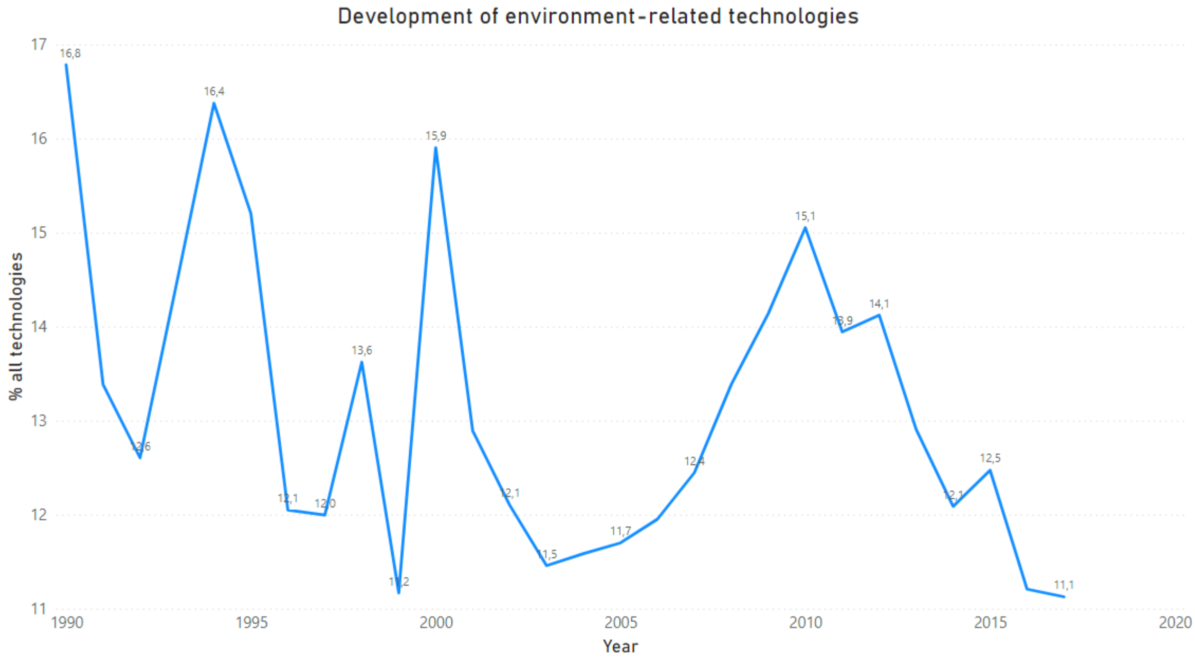
Annex 48- Evolution of Energy intensity



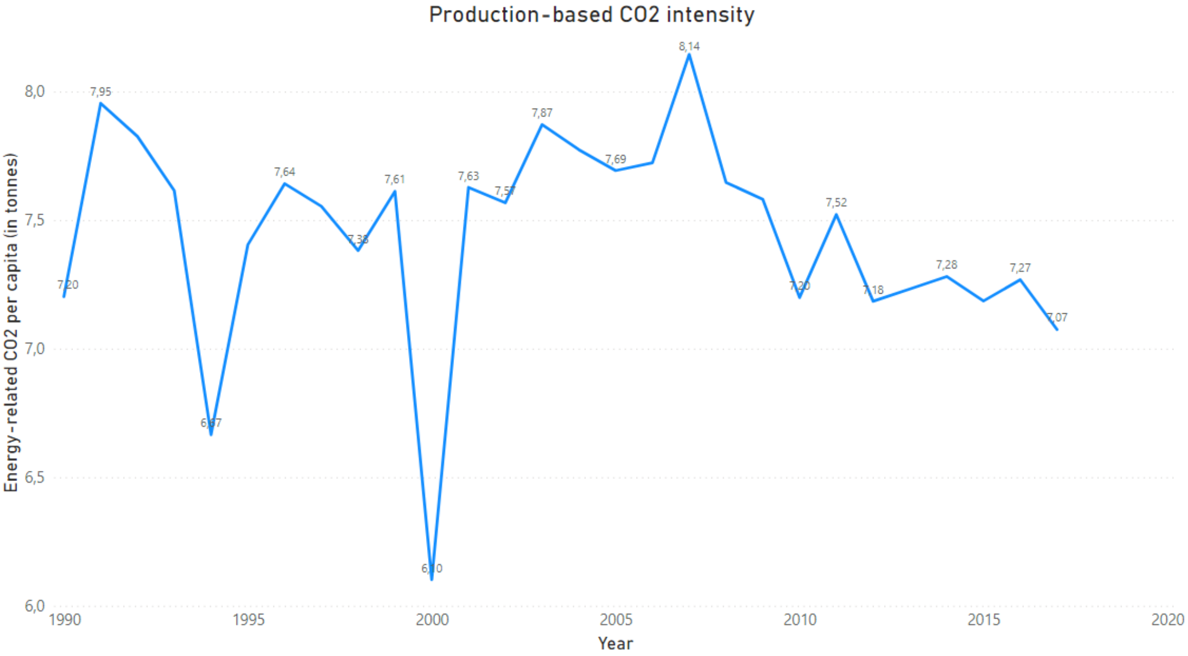
Annex 49- Evolution of Renewable Energy Supply



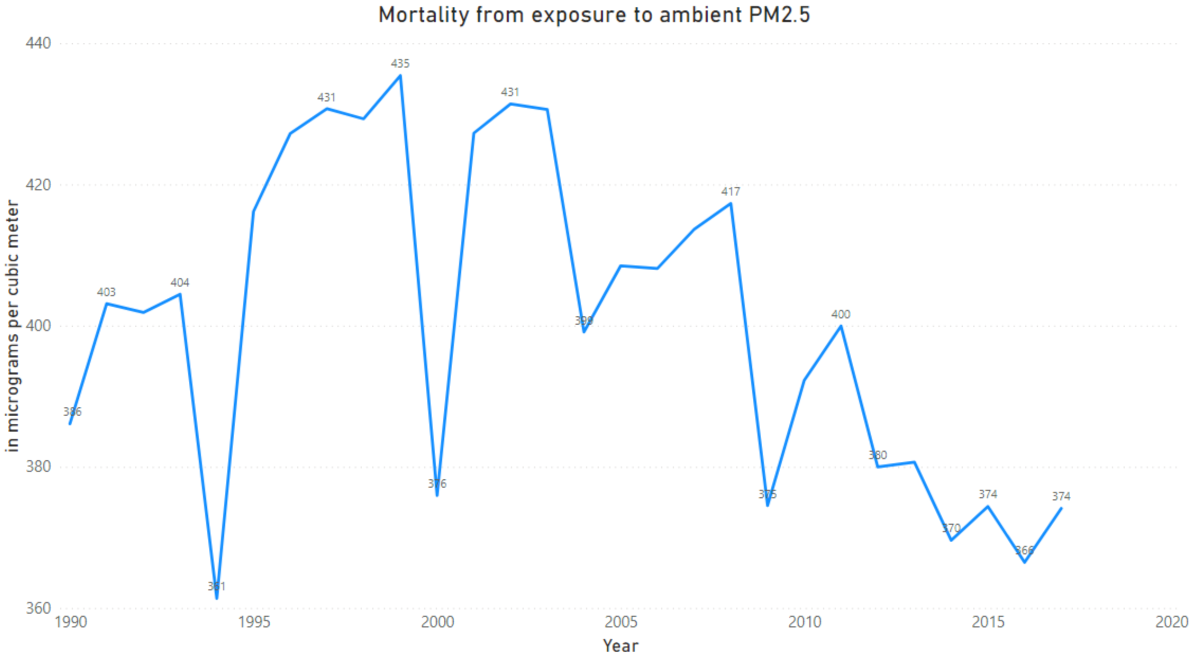
Annex 50- Evolution of development of environment-related technologies



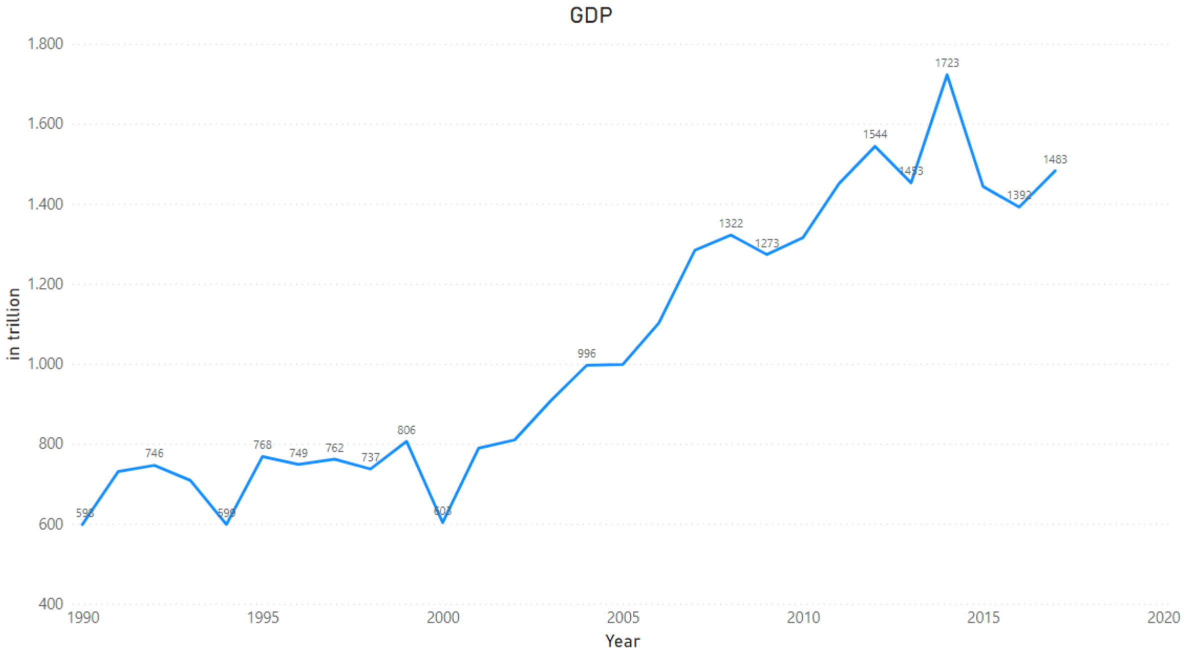
Annex 51- Evolution of Production-based CO_2 intensity



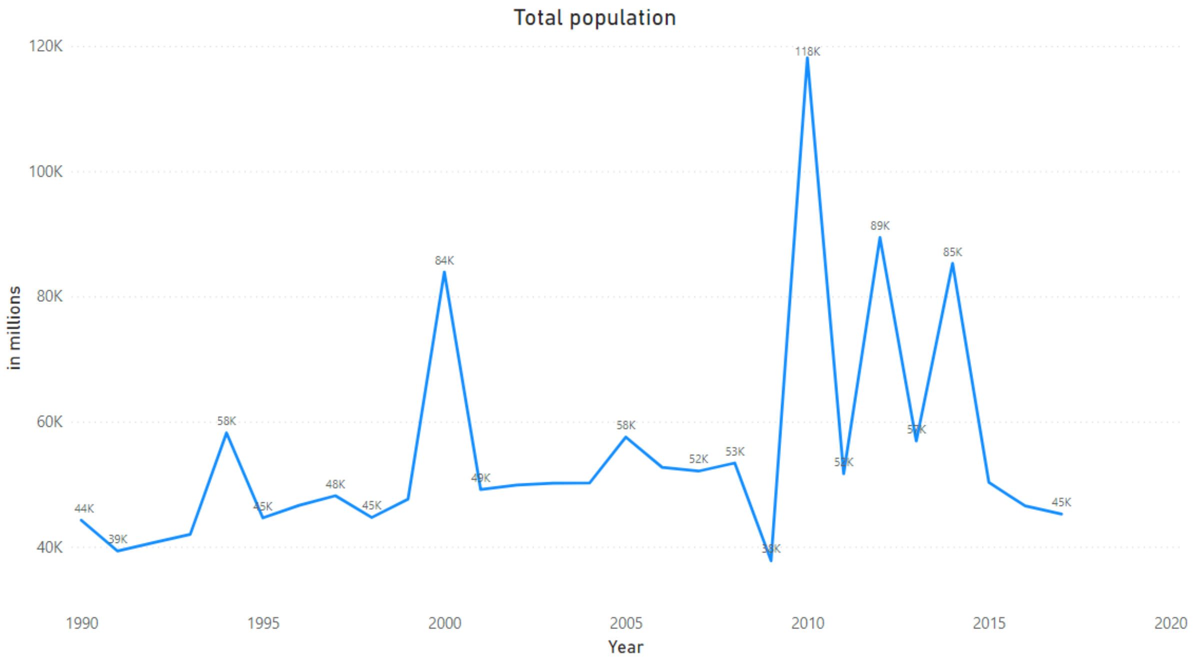
Annex 52- Evolution of Mortality from exposure tom ambient PM2.5



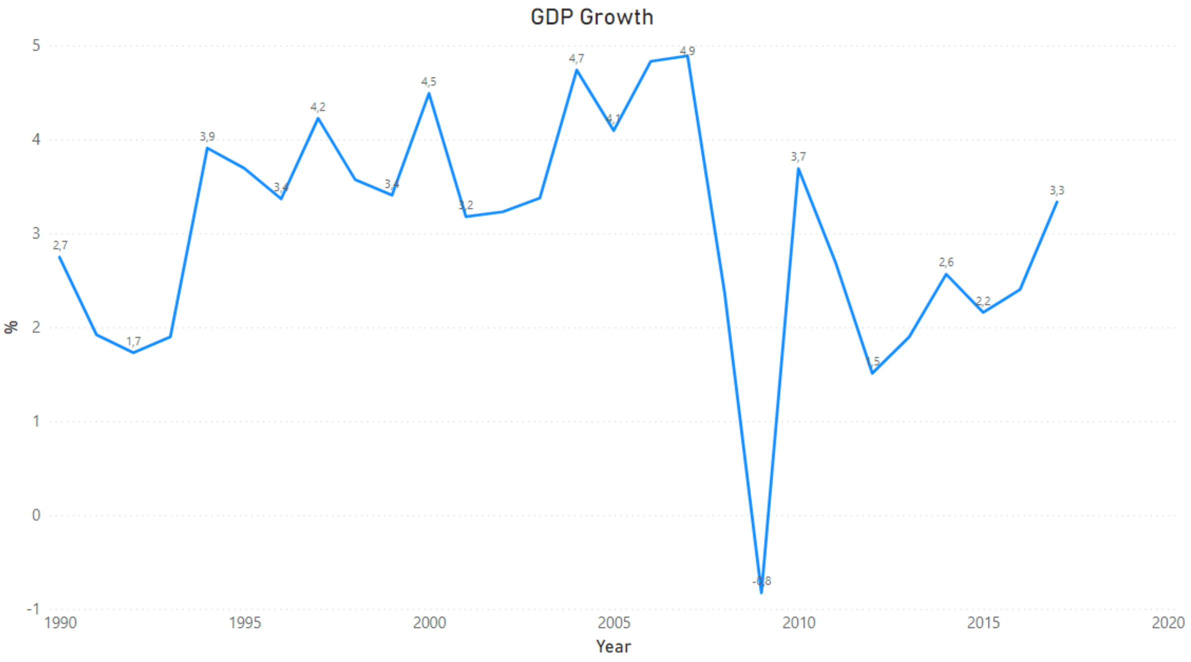
Annex 53- Evolution of GDP



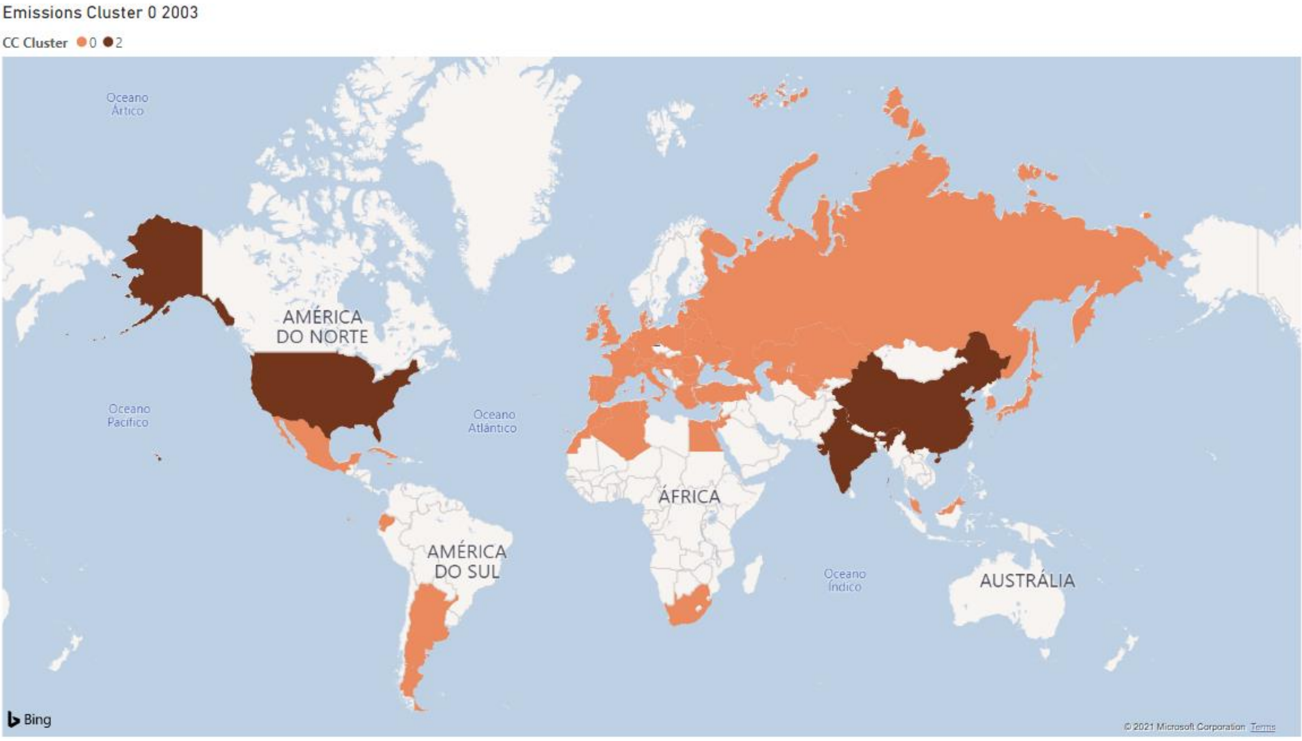
Annex 54- Evolution of Total Population



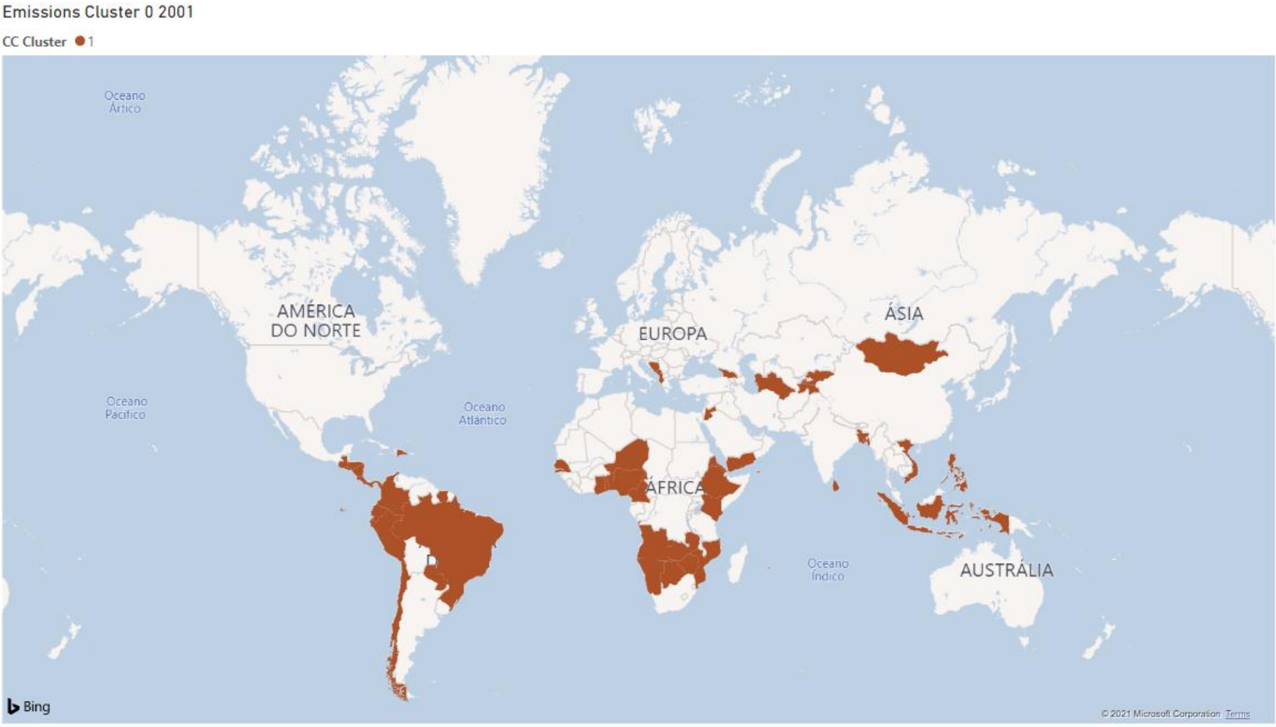
Annex 55- Evolution of GDP Growth



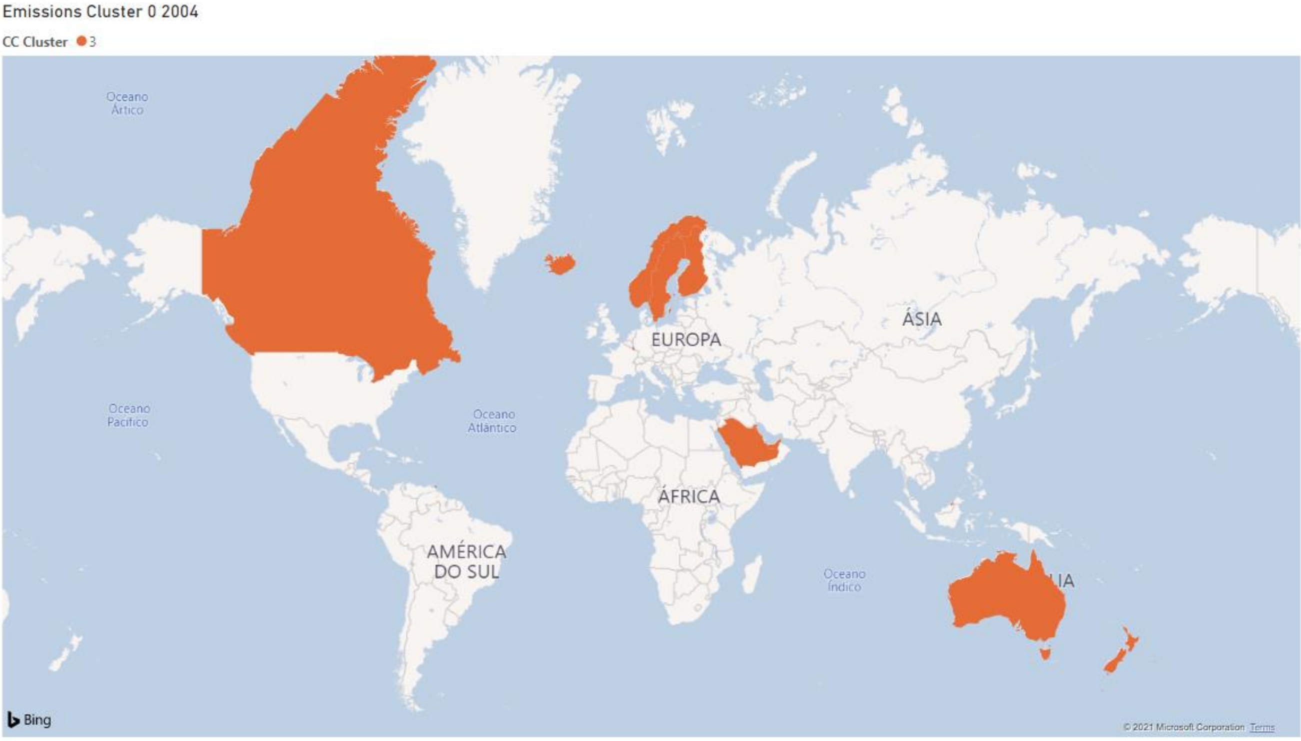
Annex 56- Contingency table clusters: Emissions C0 and Country Characteristics (CC0and CC2) in 2003



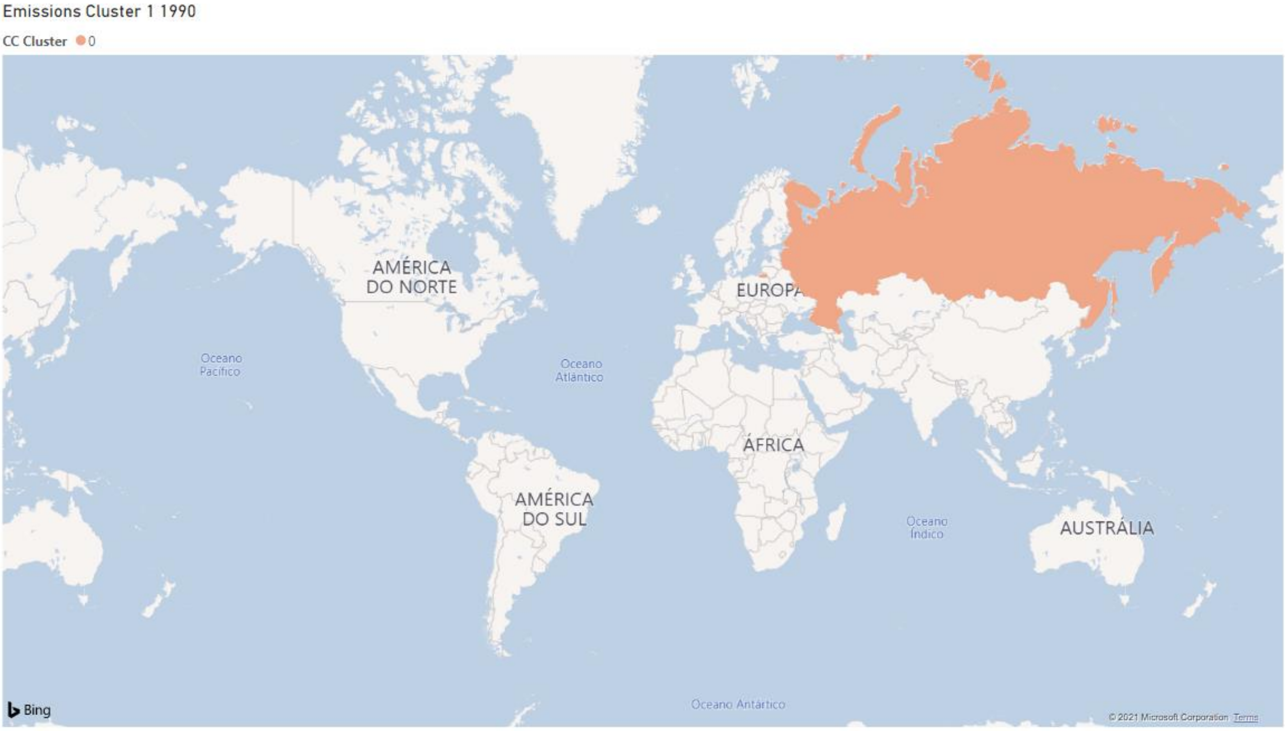
Annex 57- Contingency table clusters: Emissions C0 and Country Characteristics CC1 in 2001



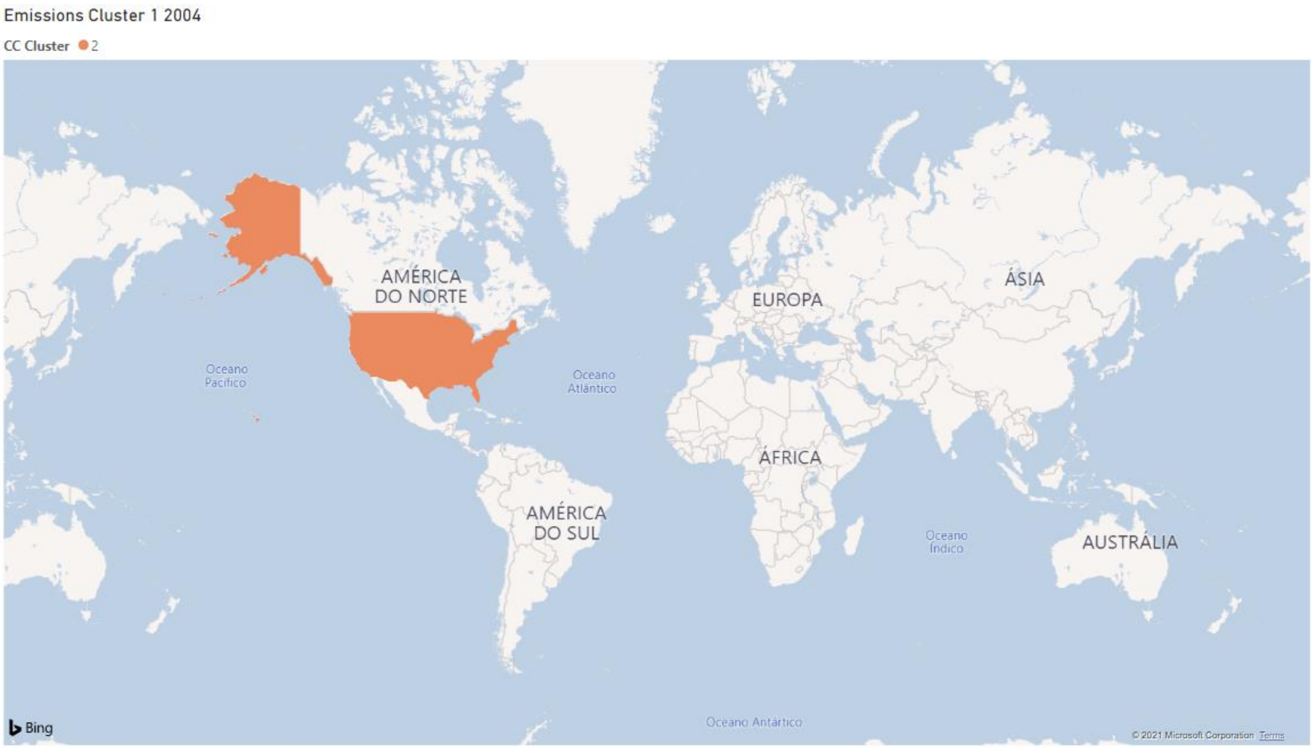
Annex 58- Contingency table clusters: Emissions C0 and Country Characteristics CC3 in 2004



Annex 59- Contingency table clusters: Emissions C1 and Country Characteristics CC0 in 1990



Annex 60- Contingency table clusters: Emissions C1 and Country Characteristics CC2 in 2004


Emissions Cluster 1 2004
CC Cluster ● 2
AMÉRICA DO NORTE
EUROPA
ÁSIA
Oceano Pacífico
Oceano Atlântico
ÁFRICA
AMÉRICA DO SUL
Oceano Índico
AUSTRÁLIA
Bing
Oceano Antártico
© 2021 Microsoft Corporation Terms