Treball de fi de màster

Climate Change and Environmental Data Science Study

Cognoms: Czerwińska

Nom: Adrianna

Titulació: Màster en Ciència i Tecnologia de la Sostenibilitat

Director/a: Miquel Sànchez-Marrè

Data de lectura: Octubre 2021

Abstract

The environmental and climatic conditions of the Earth have been gradually changing during the last few decades due to various factors such as the use of chemical fertilizers in agriculture, urbanization and deforestation. The increase in the use of vehicles and the number of factories are also responsible for the increasing global warming, because they produce the biggest amount of CO_2.

This master's thesis seeks to prove that these changes are real by using machine learning techniques to analyze the data. It makes use of preprocessing and data analysis as well. For the purpose of this dissertation, 50 random countries were chosen with various climatic and environmental variables over a time span of 26 years (1990-2015).

Exploratory data analysis helped to visualize and better understand the changes in specific topics and the overall trend for each country and for the world. These topics are air pollution, greenhouse gases, climate change and land cover. Clustering of the countries using the K-Means method showed the dependence between the variables and helped group countries into 3 classes, each representing a specific level of sustainability.

Later, by using the association rules, specifically the Apriori Algorithm, some hidden associations were discovered among variables such as the CO_2 emissions per capita and Human Development Index, the country's sustainability level and Human Development Index or the land dominance (human-made or natural) and the CO_2 emissions per capita. The forecasting technique Multiple Linear Regression was used to calculate the average CO_2 emissions in the world in 2050, based on some factors like Human Development Index, Air Quality Index and the percentage of human-made land dominance.

The results showed how the climate and environmental changes were evolving over the years. The data helped understand how big the problem is. Overall results were better than expected, it seems that some countries have put a good amount of effort into improving their environmental policies. However, there are still some countries (like China and the USA) that need to work on these policies. Averaging countries' results can be misleading because some countries are doing well while others are not.

Keywords: climate change, data analysis, data science, environment, machine learning, sustainability.

Resumen

Las condiciones ambientales y climáticas de la Tierra han ido cambiando gradualmente durante las últimas décadas debido a diversos factores como el uso de fertilizantes químicos en la agricultura, la urbanización y la deforestación. El aumento en el uso de vehículos y el número de fábricas también son responsables del aumento del calentamiento global, porque producen la mayor cantidad de CO_2.

Esta tesis pretende demostrar que estos cambios son reales mediante el uso de técnicas de aprendizaje automático para analizar los datos. También hace uso de preprocesamiento y análisis de datos. Para el propósito de esta tesis, se eligieron 50 países al azar con diversas variables climáticas y ambientales durante un período de tiempo de 26 años (1990-2015).

El análisis de datos exploratorios ayudó a visualizar y comprender mejor los cambios en temas específicos y la tendencia general de cada país y del mundo. Estos temas son la contaminación del aire, los gases de efecto invernadero, el cambio climático y la cobertura del suelo. La agrupación de países usando el método K-Means mostró la dependencia entre las variables y ayudó a agrupar a los países en 3 clases, cada una de las cuales representa un nivel específico de sostenibilidad.

Posteriormente, mediante el uso de las reglas de asociación, específicamente el Algoritmo Apriori, se descubrieron algunas asociaciones ocultas entre variables como las emisiones de CO_2 per cápita y el Índice de Desarrollo Humano, el nivel de sostenibilidad del país y el Índice de Desarrollo Humano o el dominio de la tierra (humana o natural) y las emisiones de CO_2 per cápita. Se utilizó la técnica de pronóstico Regresión Lineal Múltiple para calcular el promedio de las emisiones de CO_2 en el mundo en 2050, con base en algunos factores como el Índice de Desarrollo Humano, el Índice de Calidad del Aire y el porcentaje de dominio de la tierra creado por el hombre.

Los resultados han demostrado que los cambios climáticos y ambientales han evolucionado a lo largo de los años. Los datos ayudaron a entender cómo de grande es el problema. Los resultados generales han sido mejor de lo esperado, parece que algunos países han hecho un gran esfuerzo para mejorar sus políticas medioambientales. Sin embargo, todavía hay algunos países (como la China y los EEUU) que necesitan trabajar en estas políticas. Hacer un promedio de los resultados de los países puede ser engañoso, ya que algunos países van bien, mientras que otros no.

Palabras clave: cambio climático, análisis de datos, ciencia de datos, medio ambiente, aprendizaje automático, sostenibilidad.

Resum

Les condicions ambientals i climàtiques de la Terra han anat canviant gradualment durant les darreres dècades a causa de diversos factors com l'ús de fertilitzants químics a l'agricultura, la urbanització i la desforestació. L'augment de l'ús de vehicles i el nombre de fàbriques també són responsables de l'augment de l'escalfament global, ja que produeixen la major quantitat de CO_2.

Aquesta tesi pretén demostrar que aquests canvis són reals mitjançant l'ús de tècniques d'aprenentatge automàtic per analitzar les dades. També fa ús del preprocessament i de l'anàlisi de dades. Als efectes d'aquesta dissertació, es van triar 50 països aleatoris amb diverses variables climàtiques i ambientals en un període de 26 anys (1990-2015).

L'anàlisi de dades exploratòries va ajudar a visualitzar i comprendre millor els canvis en temes específics i la tendència general de cada país i del món. Aquests temes són la contaminació atmosfèrica, els gasos d'efecte hivernacle, el canvi climàtic i la cobertura del sòl. El clústering dels països usant el metode K-Means va mostrar la dependència entre les variables i va ajudar a agrupar els països en 3 classes, cadascuna representant un nivell específic de sostenibilitat.

Posteriorment, mitjançant l'ús de les regles d'associació, específicament l'Algoritme Apriori, es van descobrir algunes associacions ocultes entre variables com les emissions de CO_2 per càpita i l'Índex de Desenvolupament Humà, el nivell de sostenibilitat del país i l'Índex de Desenvolupament Humà o la dominància de la terra (de creació humana o natural) i les emissions de CO_2 per càpita. La tècnica de predicció de Regressió Lineal Múltiple es va utilitzar per calcular la mitjana d'emissions de CO_2 al món el 2050, basant-se en alguns factors com l'Índex de Desenvolupament Humà, l'índex de Qualitat de l'Aire i el percentatge de dominació de la terra feta per humans.

Els resultats han demostrat que els canvis climàtics i ambientals han evolucionat al llarg dels anys. Les dades van ajudar a entendre com de gran és el problema. Els resultats generals han estat millor del que s'esperava, sembla que alguns països han fet un gran esforç per millorar les seves polítiques mediambientals. Tot i així, encara hi ha alguns països (com la Xina i els EUA) que necessiten treballar en aquestes polítiques. Fer una mitjana dels resultats dels països pot ser enganyós, ja que alguns països van bé, mentre que d'altres no.

Paraules clau: canvi climàtic, anàlisi de dades, ciència de dades, medi ambient, aprenentatge automàtic, sostenibilitat.

Acknowledgment

I would like to thank my advisor Miquel Sànchez-Marrè for letting me work on this topic and dedicating his time to explain complex things. I am very grateful for the inspirations, reflections and ideas that have guided me. I deeply appreciate all the help and support I received while developing my thesis. None of this would be possible without that.

During my master's studies on Sustainable Science and Technology I have also met so many amazing people. Special expressions of gratitude to Martí Rosàs for helping me at the beginning of my journey of my master, for the incredible knowledge given and for being not only a great professor but also a great person. To Karina Gibert i Oliveras and Miquel Sànchez-Marrè for instilling in me a passion for programming. Without the classes of data science I would never have discovered my fondness to analyze and model the data. I am very grateful for expanding my horizons. For all the professors separately, thank you so much for teaching me new things, opening my eyes for very important topics and all the patience you showed during those 2 years.

Last but not least, I want to thank all my master's friends for making this studies an amazing journey I will never forget.

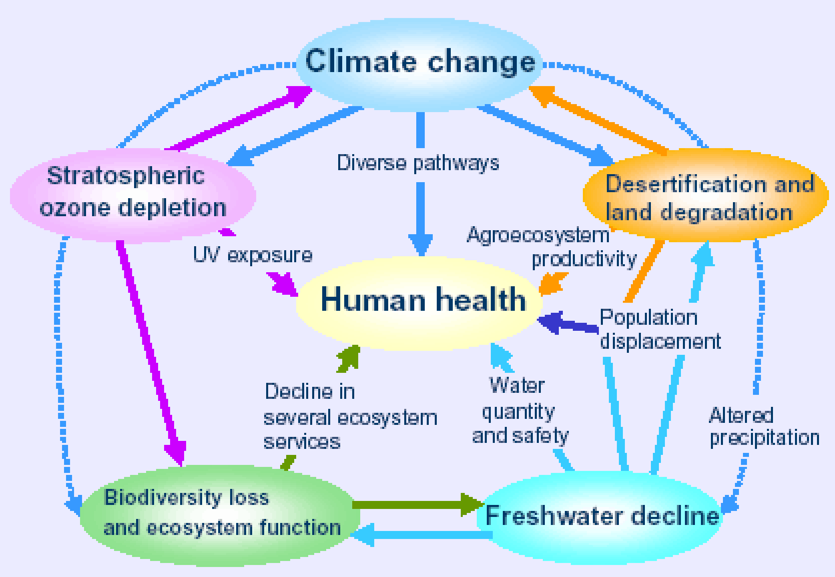
Contents

List of figures

List of tables

1. Introduction

Environmental change is a change or interference of the environment and is mostly caused by human influences and natural ecological processes (natural disasters or animal interactions). These changes include climate change, ozone emissions, biodiversity changes, hydrological and freshwater changes, deforestation or urbanization.



Climate change
Stratospheric ozone depletion
Diverse pathways
Desertification and land degradation
Agroecosystem productivity
UV exposure
Human health
Population displacement
Decline in several ecosystem services
Water quantity and safety
Altered precipitation
Biodiversity loss and ecosystem function
Freshwater decline

Figure 1. Environmental changes that affect human health. (WHO | Global environmental change, 2021)

Moreover, as shown on figure 1, this phenomenon not only makes our environment poorer but is also very dangerous for human health (WHO | Global environmental change, 2021). Sadly, environmental change can be observed with the naked eye and all the factors that contribute to it are heavily dependent on one another. As evidence, let's get climate change and other effects on the record.

Climate is a weather condition (regional or global) that occurs in a given region and its pattern is observed over the years or even decades. That includes temperature, rainfall, snow, wind, floods etc.

With that being said, climate change describes a long-term change in average weather patterns in a given region that lasts for an extended period. These changes can be also observed globally (Shaftel, 2021). Climate change is caused by factors such as biotic processes, variations in solar radiation received by Earth, plate tectonics, and volcanic eruptions. Some uncertainties in climate change are given additional attention because they may have considerable impact on human activities. Moreover, natural variability of global weather patterns is frequently identified as important, as is human-induced global warming. It is sure to say that climate change is a fact and there is plenty of evidence confirming this phenomenon (Climate Change: Vital Signs of the Planet, 2021), such as:

- warmer temperatures,
- faster ice melting,
- rising sea levels,

- change in precipitation or snowfall patterns,
- shrinking mountain glaciers,
- occurrence of extreme events,
- ocean acidification.

Climate change is caused by emissions of air pollutants and greenhouse gases. According to NASA Climate Center[1], the most dangerous substances are:

- water vapor,
- carbon dioxide (CO_2),
- methane,
- nitrous oxide,
- chlorofluorocarbons (CFCs).

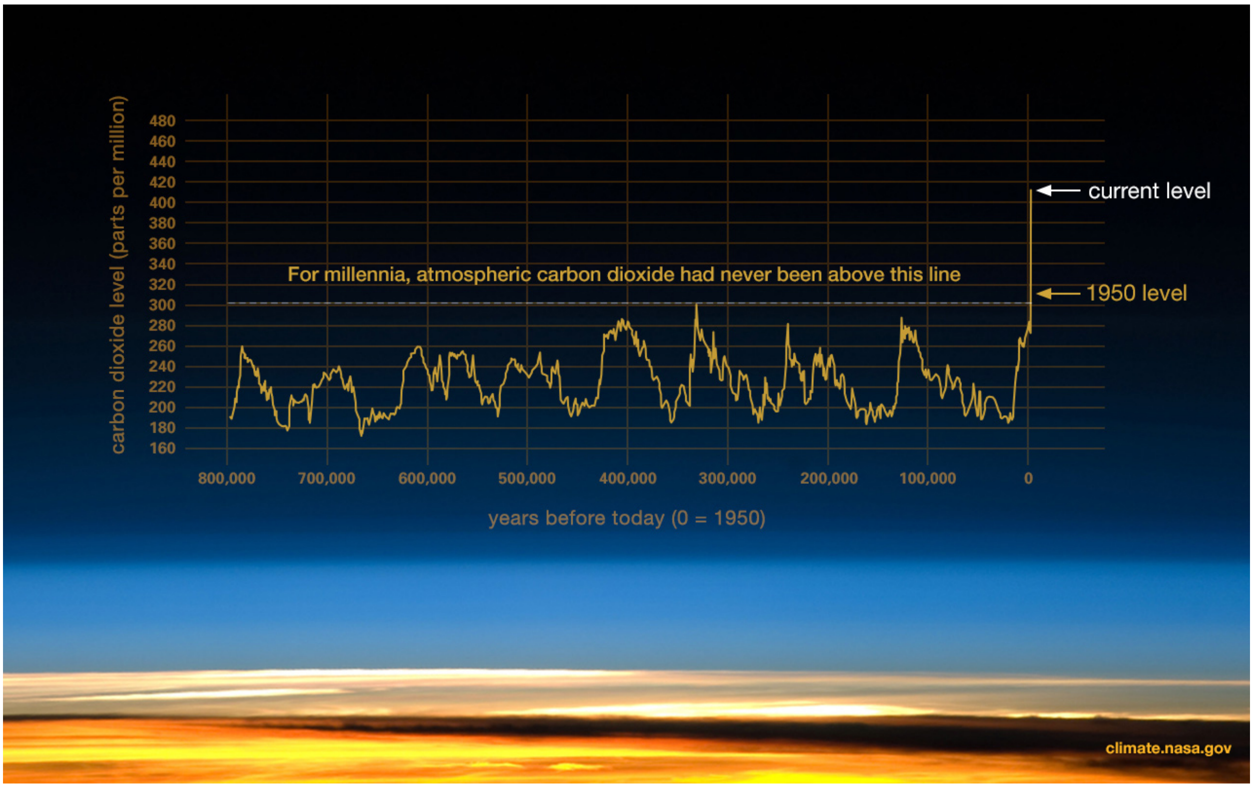


Figure 2. Global CO_2 emissions. (Climate Change: Vital Signs of the Planet, 2021)

Carbon dioxide (CO_2) is the pioneer in contribution to climate change. Figure 2 shows how the emissions were changing throughout the millennia. It is clearly visible that the current level of CO_2 in the atmosphere is twice higher than the maximum level in the past. Currently, the concentration of CO_2 is about 0.04% (412 ppm) by volume. It has increased by 47% since the Industrial Revolution began (280 ppm) (Climate Change: Vital Signs of the Planet, 2021). CO_2 emissions are caused both by nature and human activity. From the natural side, it can be caused by volcano eruptions or decomposition processes.

[1] https://climate.nasa.gov/

Humans contribute by deforestation, burning fossil fuels (coal and oil) or changing the use of the land. Also by industrial activities and transportation.

Water vapor, on the other hand, is the most important GHG in the atmosphere. It is the answer for all the climatic changes happening there. When climate changes, the temperature is getting warmer, clouds and precipitation are changing their patterns - it can all be observed due to the presence of water vapor (Water Vapor, 2021).

The climate change results are now becoming more stark than ever. Increasingly, the global warming news is just as likely to be accompanied by devastating floods or historic ice melt as it currently is by heat waves. A lot of global changes will be observed also in the future. Most importantly, climate change will continue throughout the next decades. In consequence, temperatures will keep rising and due to this there will be more droughts and bigger heat waves. Growing and frost-free season will become longer, precipitation patterns will change and hurricanes will appear more often and will be stronger. Moreover, sea levels will keep rising and the ice cap will melt at a dizzying pace (Climate Change: Vital Signs of the Planet, 2021).

1.1. Goal

As described above, climate change policies and related environmental aspects are outstanding sustainability problems in the world. In this report, the main goal is to provide a detailed, intelligent data analysis evaluation of how these topics are being managed in different countries in the world, at different years, to check the evolution of these topics' management.

This document describes the analysis of different datasets from different countries, with the relevant information to extract same profiling characteristics, with associative data mining models which outline some interesting relationships among several indicators, and with some predictive/discriminant models which provide the estimation of some relevant indicators.

2. Background

The choice of countries and relevant indicators was the most crucial part of the thesis. Without them, the whole project would not happen. With that being said, this chapter will focus on this issue. Furthermore, to proceed with data science processes, some of the machine learning algorithms that will be used in the thesis will be explained.

2.1. Countries and relevant indicators selection

For the purpose of this project, the data was collected for 50 countries, initially between the years 1970 and 2015. These countries are listed in table 1. In case of some indicators, some of the data (e.g. forest area or air pollution indicators) or years (Estonia, Russia, Uzbekistan) are missing. Furthermore, in the case of Montenegro, the data collected is only from 2006 onwards. It is because Montenegro has been an independent country only since 2006.

Afghanistan	Costa Rica	India	Morocco	Spain
Angola	Cuba	Indonesia	Netherlands	Sri Lanka
Argentina	Egypt	Israel	New Zealand	Switzerland
Australia	El Salvador	Italy	Nigeria	Tanzania
Austria	Estonia	Japan	Norway	Thailand
Brazil	Fiji	Kenya	Philippines	Turkey
Canada	France	Kuwait	Poland	United Arab Emirates
Chad	Germany	Madagascar	Romania	United Kingdom
China	Ghana	Mexico	Russia	United States of America
Colombia	Guatemala	Montenegro	South Africa	Uzbekistan

Table 1. Chosen countries.

There was not any particular reason why the mentioned countries were chosen. The only rule that was being followed was to have at least two countries from each continent to get a general picture of the situation on the Earth. To make the work clearer, these countries will be classified based on the Human Development Index (HDI).

Human Development Index is an index created by the United Nations. It ranks countries around the globe by their level of human development, gauged by the average achievements in a country in three fundamental aspects of human development (longevity, knowledge and a decent standard of living), regardless of the countries' income levels. The HDI makes equalizing differences easily visible by providing a single number that can be used to rank countries against each other according to their level of human development (Human Development Index - Wikipedia, 2021).

As per 2015, the classification was as follows (Nations Online, 2021):

- **very high (0.800–1.000):**

Argentina, Australia, Austria, Canada, Estonia, France, Germany, Israel, Italy, Japan, Kuwait, Montenegro, Netherlands, New Zealand, Norway, Poland, Romania, Russia, Spain, Switzerland, United Arab Emirates, United Kingdom, United States of America;

- **high (0.700–0.799):**

Brazil, China, Colombia, Costa Rica, Cuba, Fiji, Mexico, Sri Lanka, Thailand, Turkey, Uzbekistan;

- **medium (0.550–0.699):**

Egypt, El Salvador, Ghana, Guatemala, India, Indonesia, Kenya, Morocco, Philippines, South Africa;

- **low (0.350–0.549):**

Afghanistan, Angola, Chad, Madagascar, Nigeria, Tanzania.

It is clearly seen that most of the countries have very high or high HDI. It includes all of the European countries and all big economies or developing countries. Only 16 countries have medium or low HDI and they are mostly African countries (10 out of 16) or countries with low development levels. Based on the classification above, the analysis will be made to see if the development of the countries plays any particular role in environmental and climate change changes.

Nextly, chosen indicators were divided into six groups that correspond to different issues. All of the indicators affect the environment and/or climate change to some extent. Moreover, they influence human health which directly connects with sustainability. The following groups are distinguished:

1. **Air pollution**

Outdoor air quality affects both public health and the environment, both indirectly and directly. For health, the most common symptoms of polluted air are problems with lungs or eyes (direct) or contamination of food products (indirect). When it comes to the environment, air pollution contributes to acid rains, one of the most destructive phenomena in the world. Also, corrosive air pollutants (sulfur dioxide, carbon dioxide) can damage vehicles or buildings, which indirectly leads to pollution of the environment (Topics, 2021). The most common air pollutants are (Air pollution - Wikipedia, 2021):

- carbon monoxide (CO) - result of burning fuels (natural gas, coal or wood), the highest percentage of its emissions to the atmosphere comes from vehicles which results in creating smog;
- ammonia (NH_3) - product of agricultural waste with characteristic odor, the hazardous part of ammonia is that it reacts with nitrogen oxide and sulfur oxide to create other, secondary particles;
- nitrogen oxide (NO_x) - effect of high temperature combustion and the product of thunderstorms with a specific sharp and biting odor, it makes the environment acidic and corrosive;

- sulfur dioxide (SO_2) - product of volcano eruptions and industrial processes, the biggest contributor to acid rains when oxidized with NO_2;
- ozone (O_3) - the most important factor in the ozone layer, when its levels are too high, it contributes to smog and global warming.

Besides chemical compounds, there is also another, very dangerous group of air pollutants: particulate matter. Particulate matter is a mixture of solid particles and liquid droplets found in the air. It includes dust, soot, smoke or dirt, and sometimes they are big enough to be seen normally. Others, on the other hand, are so small that they can only be seen through a microscope. “Most particles form in the atmosphere as a result of complex reactions of chemicals such as sulfur dioxide and nitrogen oxides, which are pollutants emitted from power plants, industries and automobiles.” The smaller the diameter of the particle, the more serious health risks it can cause (Particulate Matter (PM) Basics | US EPA, 2021). There are two main particulate matters:

- PM2.5 - fine inhalable particles, with diameters that are generally 2.5 micrometers and smaller,
- PM10 - inhalable particles, with diameters that are generally 10 micrometers and smaller.

For air quality to be sustainable, the air pollutants must not show any significant direct or indirect health threats to people and the environment.

2. Climate change

Climate change refers to a significant increase of surface temperatures all around the world, and can be directly connected to high concentrations of greenhouse gases. Climate change has been occurring on Earth for billions of years, but recently has begun occurring at an alarming rate because of human activities. The most important factors that can mean climate change are:

- temperature - an important scientific measurement that expresses how hot or cold something is, it can be measured with a thermometer in three different scales: Celsius scale (°C), Fahrenheit scale (°F) and Kelvin scale (K) (Temperature - Wikipedia, 2021). Measurement that helps to visualize climate changes by showing that overall, the worldwide temperature is rising.
- Precipitation - “any product of the condensation of atmospheric water vapor that falls under gravitational pull from clouds”. It is not only rain, it can also include drizzling, sleet, snow, ice pellets, graupel and hail. This process happens when a part of the atmosphere becomes saturated with water vapor. In consequence, the water condenses and falls (Precipitation - Wikipedia, 2021). Changes in its pattern reflect climate change.

3. Greenhouse gases

Greenhouse gases (GHG) are the gases in the atmosphere that trap heat near Earth's surface, contributing to global warming. Although some gases occur naturally, produced by plants and animals as part of their metabolic processes, human activity, such as burning fossil fuels, industrial processes and agriculture, has led to an increase in the concentration of many of these gases, notably carbon dioxide, methane and nitrous oxide (Overview of Greenhouse Gases | US EPA, 2021).

- Carbon dioxide (CO_2) - product of deforestation and burning fossil fuels (coal, oil, natural gas or wood) for energy, but also it is a result of respiration and volcanic eruptions, its high levels in the atmosphere cause the global temperature rise, it is removed from the atmosphere during photosynthesis processes (Air pollution - Wikipedia, 2021);
- methane (CH_4) - emitted by animals in the process of digestion and also a component of natural gas that overall brings a lot of good to the environment. However if it is released to the atmosphere before being burned, it traps the heat in the atmosphere and contributes to climate change (Methane and the Environment | SoCalGas, 2021);
- nitrous oxide (N_2O) - result of agricultural processes like animal waste or fertilization of the soil, it depletes ozone in the stratosphere, contributing to the ozone hole (Thompson et al., 2019).

4. Land cover

Land cover is the natural or human-built environment at ground level such as vegetation, land use or impervious surface. Land cover types are used for distinguishing between types of geographical area covered with trees, urban areas, agricultural areas, bodies of water etc. Some of them are:

- country land - total size of the country containing land area and inland waters (although it does not fall into the category of land cover, it is needed for the purpose of this project);
- land area - the land of the country that does not include inland waters;
- inland waters - the amount of permanent water bodies (like rivers, lakes or reservoirs) that are inland in the country (Glossary: Inland waters, 2021);
- forest land - “ecosystems that have a tree crown density of 10% or more and are stocked with trees capable of producing timber or other wood products.” (Forest Land, 2021);
- agricultural land - the land base where agriculture is practiced;
- urbanized land - built-up area, both urban and rural, where there is density of human structures such as houses, commercial buildings, roads, bridges, and railways (National Geographic Society, 2021).

2.2. Metadata gathering

Data collected for this project had multiple different sources and is fully numerical. Moreover, in order to create final datasets, few sources’ datasets had to be manually downloaded and combined into one. Datasets are coded in Microsoft Excel (Microsoft Corporation, 2021). The premise of this thesis was to use the annual, historical data between the years 1970 and 2015. However, as mentioned in section 2.1., in some cases, data from earlier years was not available and there is a lot of missing data between 1970 and 1990. To obtain more clear analysis, it was decided to proceed with the data from 1990 to 2015.

Each dataset also contains variables like country and year, which are not included in the descriptions.

Firstly, a general dataset was created. This dataset contains information about a country's HDI between the years 1990-2015. HDI data was collected from the United Nations Development Programme website (Human Development Data Center | Human Development Reports, 2021).

Variable	Description
hdi	Human Development Index value

Table 2. Structure of hdi dataset.

1. Air pollution

Data for air pollution was obtained from following websites: EDGAR (EDGAR - The Emissions Database for Global Atmospheric Research, 2021) for all emissions in tons, OECD (Air and climate - Air pollution exposure - OECD Data, 2021) for PM2.5 exposure in micrograms per cubic meter, Data Bank (WDI Database Archives (beta) | DataBank, 2021) for PM10 exposure in micrograms per cubic meter and State of Global Air (Explore the Data | State of Global Air, 2021) for Ozone in parts per billion. Each variable with emissions in tons contained emissions from different sectors. To obtain the total number of emissions, these variables were summarized using Excel. In the case of PM2.5 ($\mu g/m^3$), PM10 ($\mu g/m^3$) and O_3 (ppb), no additional calculations were needed. The final dataset contains information of annual emissions of each air pollutant in tons between 1990 and 2015.

The biggest challenge was to obtain Air Quality Index indicators in suitable units, that is why there is only data for PM2.5, PM10 and O_3. Even so, this data has a lot of missing values.

Variable	Description
co.em	Carbon monoxide emissions in tons
nh3.em	Ammonia emissions in tons
nox.em	Nitrogen oxide emissions in tons
so2.em	Sulfur dioxide emissions in tons
pm2.5.em	Particulate matter 2.5 emissions in tons
pm10.em	Particulate matter 10 emissions in tons
pm2.5.exp	Particulate matter 2.5 exposure in micrograms per cubic meter
pm10.exp	Particulate matter 10 exposure in micrograms per cubic meter
o3.exp	Ozone exposure in parts per billion

Table 3. Structure of air pollution dataset.

2. Climate change

Climatic data were obtained from the CEDA archive (CEDA Archive, 2021), more specifically from Crudata website (Crudata, 2021). This data describes annual temperature and precipitation in the timespan of 117 years (1901-2018). However, as mentioned before, data used in this project falls in the years 1990-2015. Not necessary years were deleted. What is more, data was presented for each month, not year. To obtain annual data for temperature and precipitation, the average calculations for each year had to be done using the following Excel formula:

=AVERAGE(number1; [number2]; ...)

Variable	Description
min.temp	Minimum temperature in Celsius degrees
avg.min.temp	Average minimum temperature in Celsius degrees
avg.temp	Average temperature in Celsius degrees
avg.max.temp	Average maximum temperature in Celsius degrees
max.temp	Maximum temperature in Celsius degrees
min.prec	Minimum precipitation in depth in millimeters
avg.prec	Average precipitation in depth in millimeters
max.prec	Maximum precipitation in depth in millimeters
total.prec	Total precipitation in depth in millimeters

Table 4. Structure of climate change dataset.

3. Greenhouse gases

Data collected for the greenhouse gases dataset also comes from the EDGAR website (EDGAR - The Emissions Database for Global Atmospheric Research, 2021). Each variable contained emissions from different sectors. For the purpose of this project, all emissions were summarized using Excel. Thus, the final dataset contains information of annual emissions of each greenhouse gas in tons and emissions per capita in tons between the years 1990-2015, with some exceptions.

Variable	Description
co2.em	Carbon dioxide emissions in tons
ch4.em	Methane emissions in tons of CO_2 equivalent
n2o.em	Nitrous oxide emissions in tons of CO_2 equivalent
co2.cap	Carbon dioxide emissions in tons per capita
ch4.cap	Methane emissions in tons of CO_2 equivalent per capita
n2o.cap	Nitrous oxide emissions in tons of CO_2 equivalent per capita

Table 5. Structure of greenhouse gases dataset.

4. Land cover

Data for all variables, except urbanized land, was downloaded from the FAO website (FAOSTAT, 2021). Urbanized land data comes from GHSL website (Global Human Settlement - Degree of urbanisation - European Commission, 2021). This dataset contains the most missing values. For example, urbanized land occurs only for years 1990, 2000 and 2015. Forest land starts only from 1990 or even later. Nevertheless, collected data shows perfectly the environmental changes.

Variable	Description
country.land	Country land in square kilometers
land.area	Land area in square kilometers
inland.waters	Inland waters in square kilometers
forest.land	Forest land in square kilometers
agricult.land	Agricultural land in square kilometers
urban.land	Urbanized land in square kilometers

Table 6. Structure of land cover dataset.

2.3. Techniques used

2.3.1. Preprocessing

2.3.1.1. Data cleaning

Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a dataset. It refers to identifying incomplete (missing data), incorrect, irregular (outliers), inaccurate or irrelevant (duplicates) parts of the data and then replacing, modifying, or deleting the ‘dirty’ data.

2.3.1.2. Interpolation

Interpolation is an estimation method of finding new data points based on the range of a set of known data points. It has different kinds like ‘linear’, ‘nearest’, ‘zero’, ‘slinear’, ‘quadratic’, ‘cubic’) (scipy.interpolate.interp1d — SciPy v0.14.0 Reference Guide, 2021). This method is commonly used to deal with missing data in the datasets.

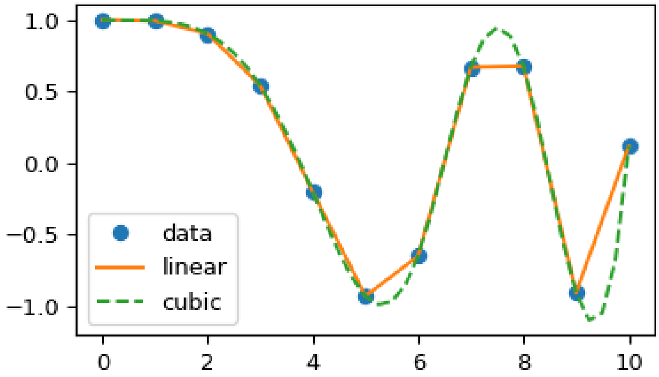


Figure 3. Different interpolation methods (linear and cubic). (Interpolation (scipy.interpolate) — SciPy v1.7.1 Manual, 2021)

2.3.1.3. New variables creation

Creation of new variables helps to obtain a bigger picture of the dataset or to have an insight on some information that is not directly in the dataset. For example, having the population and CO_2 emissions, one wants to know the CO_2 emissions per capita. In this case the creation of a new variable is needed where CO_2 emissions per capita will equal CO_2 emissions divided by population.

2.3.1.4. Correlation

Correlation is a measure that indicates the dependence between two or more variables. The most common methods are the Pearson, Spearman and Kendall indexes. However, in this thesis only Pearson correlation will be used.

Pearson correlation coefficient measures the strength of linear correlation (and its direction - whether positive, negative or neutral) between two numeric variables. It is represented by r and it only assumes values between -1 and 1. It can be calculated using the following formula:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

Interpreting the value of r (Correlation: straight to the point, 2021):

- 0.9-1 - very strong correlation;
- 0.7-0.9 - strong correlation;
- 0.5-0.7 - moderate correlation;
- 0.3-0.5 - weak correlation;
- 0-0.3 - negligible correlation.

It was introduced by Francis Galton in the 1880s but developed by Karl Pearson. Later its mathematical formula was derived and published by Auguste Bravais in 1844 (Pearson correlation coefficient - Wikipedia, 2021).

2.3.1.5. Dimensionality reduction

Dimensionality reduction uses an algorithm to reduce the size of an input data to a smaller size while keeping the integration of data. It is used when there are too many dimensions in the dataset.

2.3.2. Machine learning

Machine learning algorithms are broadly used for any type of classification or prediction models. There are two main types of these algorithms: supervised and unsupervised learning. The main focus will be on supervised and unsupervised learning algorithms as shown on figure 3 (A Brief Introduction to Unsupervised Learning, 2021), (Supervised vs. Unsupervised Learning: What's the Difference?, 2021).

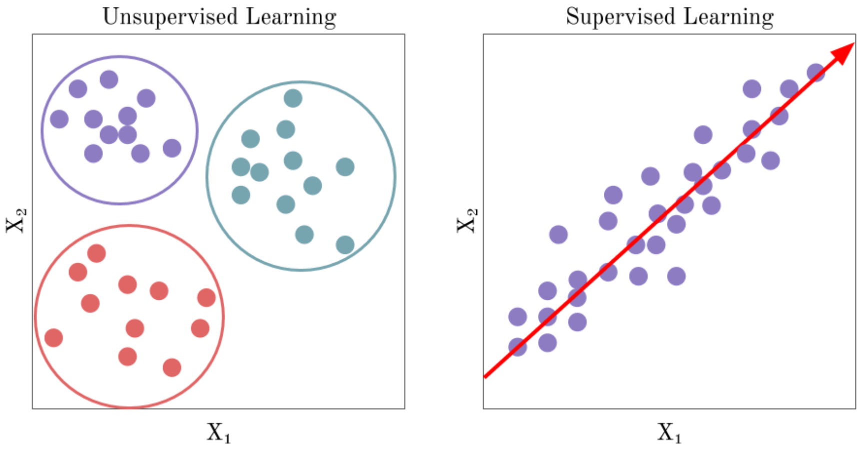


Figure 4. Unsupervised and supervised learning algorithms. (A Brief Introduction to Unsupervised Learning, 2021)

1. **Supervised learning** - a machine learning approach that uses target variables (categorical or numeric) to create a model that measures its accuracy and learns over time. The datasets are trained to classify the data or to predict the outcomes and desired levels of accuracy under supervision. Examples are: Regression, Decision Tree, Random Forest, KNN, Logistic Regression, etc.

The main tasks of supervised learning are:

- classification - the use of an algorithm to classify the test data correctly, for example separation between spam and normal email;
- regression or prediction - the use of an algorithm to understand the relationship between numeric dependent and numeric independent variables.

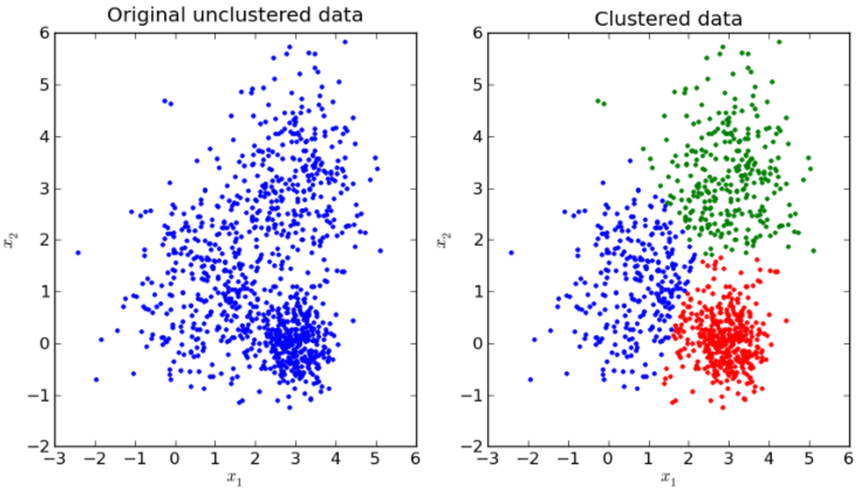
2. **Unsupervised learning** - machine learning approach that uses unlabeled variables to analyze possible relationships among variables, such as association rules, or among observations, such as clustering techniques. These algorithms discover hidden patterns without direct human supervision. There is no outcome to predict or estimate. Examples are: Apriori algorithm, K-Means clustering, Hierarchical Clustering.

The main tasks of unsupervised learning are:

- clustering - the use of an algorithm to group unlabeled data based on their similarities or differences;
- association - the use of an algorithm to find relationships between variables in a dataset by using different rules.

2.3.2.1. K-Means Clustering

Clustering is the process of grouping objects of a dataset in a way that objects from the same group (cluster) present more similarities between each other than with objects from another group. (Understanding the concept of Hierarchical clustering Technique, 2021)



Original unclustered data

Clustered data

Figure 5. Clusters. (Understanding the concept of Hierarchical clustering Technique, 2021)

K-Means Clustering is an unsupervised machine learning algorithm that is used to identify pre-defined clusters of data objects in a dataset. It segregates the data based on having similar features or common patterns. The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid (K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks, 2021). It works as follows:

1. Specify the number of clusters K.
2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids (assignment of data points to clusters is not changing).

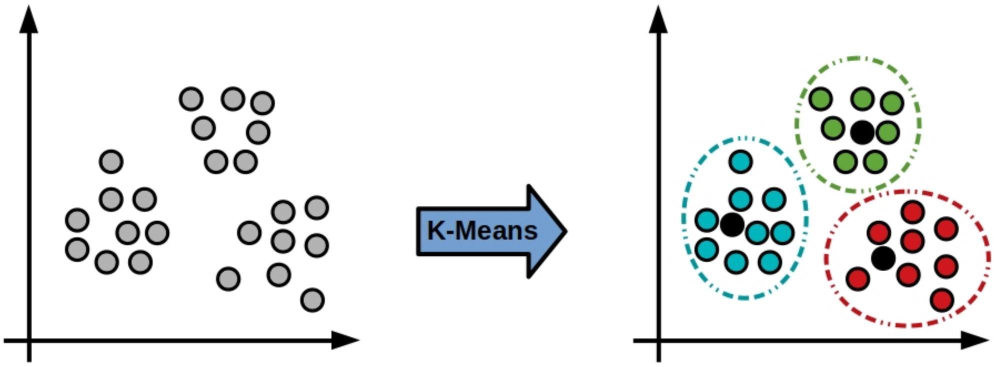


Figure 6. K-Means Clustering. (K-Means: One Of The Simplest Clustering Algorithms – Fly Spaceships With Your Mind, 2021)

K-Means as a term was first used by James MacQueen in 1967, however it was introduced by Hugo Steinhaus in 1956. The K-Means algorithm was proposed by Stuart Lloyd of Bell Labs in 1957 but not published anywhere. Only in 1965, Edward W. Forgy published basically the same method (k-means clustering - Wikipedia, 2021).

2.3.2.2. Association rules analysis

Association rules analysis is a rule-based machine learning method for discovering interesting relations between variables in large datasets. It is intended to identify strong rules discovered in datasets using some measures of interest. The goal is to obtain a set of association rules which express the correlation among attributes, from a dataset of item transactions. The dataset should have enough number of transactions in order that the correlation appears a sufficient number of times. The original data matrix (unsupervised, but could be supervised) is used as an input and the output is a set of association rules satisfying a minimum support and a minimum confidence.

They are evaluated with 3 main measures (Ng, 2021):

- Support - an indication of how frequently the itemset I appears in the dataset. It is measured by the proportion of dataset in which an itemset appears, itemsets can have more than one item.

$$support(I) = \frac{number\ of\ appearances\ containing\ I}{total\ number\ of\ appearances}$$

- Confidence - an indication of how often the rule has been found to be true, in other words how likely item Y is appearing when item X is appearing. It is measured by the proportion of the dataset with item X, in which item Y also appears. It has one drawback: it tends to misrepresent the importance of an association.

$$confidence(X \rightarrow Y) = \frac{number\ of\ appearances\ containing\ X\ and\ Y}{number\ of\ appearances\ containing\ X}$$

- Lift - computes the ratio between the rule’s confidence and the support of the itemset in the rule consequent. Ix says how likely item Y is appearing when item X is appearing, while controlling for how popular item Y is.

$$lift(X \rightarrow Y) = \frac{number\ of\ appearances\ containing\ X\ and\ Y}{number\ of\ appearances\ containing\ X\ *\ number\ of\ appearances\ containing\ XY}$$

Association rules analysis was introduced by Rakesh Agrawal, Tomasz Imieliński and Arun Swami in order to discover the regularities between the products in a large-scale transaction data that was collected by the systems in supermarkets. That is why it is mainly used for market basket analysis (Association rule learning - Wikipedia, 2021).

Apriori Algorithm

Apriori Algorithm is an algorithm for association rule mining. Its goal is to identify the frequent individual items in the dataset and extend them into larger item sets as long as these items appear often in the database. It uses a “bottom up” approach, where subsets are extended one item at a time and groups are tested against the data. If there is no more option of extension, the algorithm stops.

Data scientists Agrawal and Srikant introduced the Apriori Algorithm in 1994 (Apriori algorithm - Wikipedia, 2021).

2.3.2.3. Linear Regression

Linear Regression is an algorithm based on supervised learning. It is used to model relationships between a dependent variable (y) and one or more independent variables (x). It can be used to predict or forecast variables. The function of this model is as follows:

$$\mathbf{y = \theta_1 + \theta_2 \cdot x}$$

where:

x - input training data;

y - data labels;

$\boldsymbol{\theta_1}$ - intercept;

$\boldsymbol{\theta_2}$ - coefficient of x.

The main idea of this model is to find a line of the best fits representing two or more variables as shown on figure 7. The most desirable result is when the values (x and y) are as close as possible to the line. Once these variables are found, we get the best fit line for our model.

Linear Regression divides into two main types:

- Simple Linear Regression - characterized by one independent variable;
- Multiple Linear Regression - characterized by multiple independent variables.

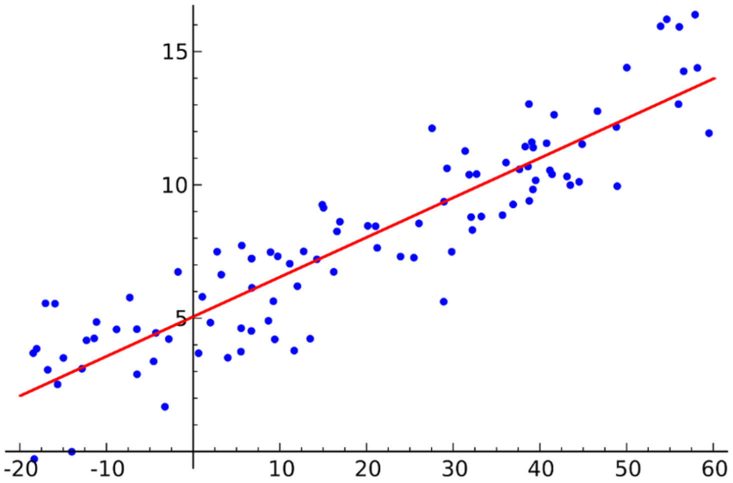


Figure 7. Linear regression. (Introduction to Machine Learning Algorithms: Linear Regression, 2021)

Linear Regression was introduced by Legendre (1805) and Gauss (1809) to find a fit line for a prediction of planetary movement (Linear regression - Wikipedia, 2021).

3. Objective

The main objective of this thesis is to prove that climatic and environmental change is real by providing the analysis of countries' profiles. The analysis will be performed during the time span of 26 years (1990-2015) for 50 chosen countries. The whole analysis will be based on air pollution, climate change, land and greenhouse gases related indicators on an annual basis. In addition, each country will be classified based on how sustainable it was each year. To do so, special classification will be designed. Depending on the country, this analysis will reveal that this phenomenon is more or less visible, yet it is happening no matter how small the change is.

To obtain satisfactory results for the analysis, machine learning models, among others, will be used. Firstly, the *data cleaning* process will be performed during the initial *pre-processing stage*. Then, Interpolation will be used to deal with missing data. *Interpolation* will be done using the 'cubic' kind because it has the best fit line. 'Cubic' refers to a spline interpolation of first, second or third order (scipy.interpolate.interp1d — SciPy v0.14.0 Reference Guide, 2021). Before grouping countries by sustainability level using conditional statements in Python, there is a need to *create new variables* based on which the classification will be performed. This will be done using basic calculation methods in Python. As a result of all the processes above, the sustainability level classification will be performed based on the conditions described in section 4.3.

Before doing the analysis of the results, there is a need to calculate correlation coefficient amongst all variables. It is needed to see which variables are highly correlated, and thus can be removed. However, this move will not affect the next part - exploratory data analysis. In this part the overall trends and changes will be described. Next, K-Means Clustering will be performed to find some classes and profiles amongst variables. Further, the other part of this thesis will be to find if there are any association patterns between different variables using the Apriori Algorithm over the years. For the final step of the thesis, Multiple Linear Regression will be used to forecast CO_2 emissions in 2050.

Annex includes the link to a Github repository where all the codes and datasets can be found. There are also detailed results for each country. Initially, it was planned to include all of the countries in this dissertation but finally it was decided to proceed with the general analysis instead, as there is too much data, graphs and information to include.

4. Methodology

4.1. Data selection

To start with, data collected covered the years from 1970 to 2015. After data collection and a preliminary analysis, it was seen that there was a lot of missing data between the years 1970-1990. Hence, in order to avoid the huge missing values gap, it was decided to proceed only with the years 1990-2015 for the main analysis. Variables chosen for the analysis were selected for four specific topics: air pollution, climate change, greenhouse gases and land cover. In addition, there is data containing info about HDI.

4.2. Pre-processing

The initial preprocessing was done manually during the data collection. Although it was not hard to do, it was the most time-consuming and crucial part of the pre-processing. More complicated things were done using Python programming language (Python.org, 2021).

To implement the Python analysis, the following libraries will be used in this thesis:

Python library	Function
Pandas	Data manipulation
Numpy	Data manipulation and computation with arrays
Matplotlib, Seaborn, Plotly	Data visualization
Mlxtend, Sklearn	Machine learning
Scipy	Data computation and algebra

Table 7. Python libraries used in the master thesis.

4.2.1. Data cleaning

As mentioned above, initial data cleaning was done during data collection. It focused mostly on cleaning unnecessary years with a lot of missing values. Also, some very basic calculations were performed in the meantime using Excel. The goal was to obtain the best, cleanest datasets possible.

4.2.2. Missing values treatment

Since there is a considerable number of missing values, it will be handled by using a cubic spline interpolation method to fill them as mentioned in section 3. It is done to the most important variables to obtain more clear analysis.

4.2.3. New variables creation

For each dataset there will be new variables created. The reason for creating new variables is to use them on the country classification part and for some separate analysis as well.

4.2.3.1. Country classification variable

Country classification variables aim to show their sustainability level based on assumed conditions. Countries will be classified as sustainable, quite sustainable and not sustainable.

4.2.4. Data merge

The final data frame will be created by merging together four datasets created before. What is more, the final country classification variable will be based on an already existing country classification variable for each dataset.

4.2.5. Correlation and dimensionality reduction

Correlation will help to determine if two variables in a dataset are related in any way. Strongly correlated variables will be removed from the dataset to prevent future duplication of the information. It will provide a dimensionality reduction that will help to obtain more accurate results.

4.3. Exploratory data analysis

Exploratory data analysis (EDA) is an approach used to analyze and investigate datasets. Its goal is to summarize main characteristics of datasets by using statistical graphics and other data visualization methods. Also, it aims to formulate possible hypotheses related to the data that can lead to new analyses (Exploratory data analysis - Wikipedia, 2021). The main use of EDA is to spot relationships between the variables that cannot be seen by doing formal modelling or hypothesis testing. Moreover, EDA helps to develop a deeper understanding of the data. Questions asked during the whole process navigate as tools that are guiding the investigation.

This analysis method was developed by John Turkey in the 1970s. He was an American mathematician and statistician that encouraged other statisticians to explore the data more. His best known developments were the Fast Fourier Transform (FFT) algorithm and box plot (John Tukey - Wikipedia, 2021).

Since the dataset contains 50 different countries, the EDA will be done for overall trends over the years and characteristics of the indicators in starting and ending years, not for the countries individually. However, in the Annex, there will be results for each country. The characteristics analyzed will be:

- air pollutants emissions;
- air quality index;
- maximum and maximum temperature;
- total precipitation;
- climate change patterns;
- greenhouse gases emissions
- co2 emissions per capita;
- urbanized and forest area;
- land dominance;
- hdi.

4.4. Data mining

4.4.1. Profiling of the different countries according to the indicators

K-Means Clustering will find clusters of variables. It will aim to show the similarities of the objects. It will be done for variables after dimensionality reduction. Moreover, each variable will be compared against each other. Two chosen variables will be represented on a scatter plot to make a profiling of the clusters.

It will be done for 2 cases: one, for each country each year and second, for each country in general (with average values of each variable).

The step by step procedure for this part of the thesis will be as follows:

1. Normalization of the data to bring all the variables to the same scale.
2. Plotting the intra-cluster similarity to decide the number of clusters using the elbow method. The Elbow method is a method used to determine the number of clusters in a dataset (Elbow method (clustering) - Wikipedia, 2021).

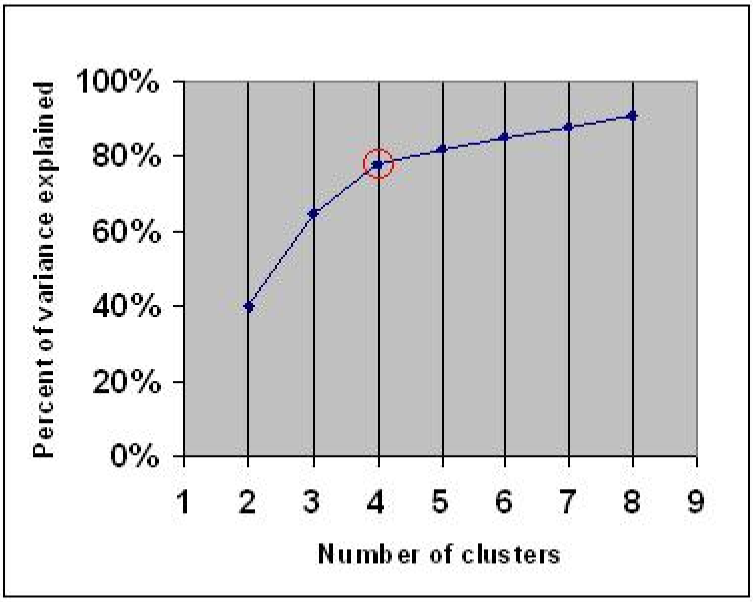


Figure 8. Elbow method. (Elbow method (clustering) - Wikipedia, 2021)

3. Calculating centroids for each variable in the cluster.
4. Obtaining the number of instances in each cluster.
5. Application of K-Mean Clustering and visualization.
6. Profiling of the clusters.
7. Implementation of K-Means Clustering using:
   - number of clusters - 3;
   - seed - 15;
   - minit - points.

4.4.2. Associative models study

The Apriori Algorithm will be used to identify how often the variables appear with each other. It will be done for some chosen variables to see the associations amongst them but also, most importantly, for the country's sustainability level and HDI level to see if there is any significant association.

The variables that will be compared are:

- co2 per capita (sust.level.ghg) and hdi (hdi.level);
- aqi (aqi.bucket) and hdi (hdi.level);
- land dominance (land.dom) and co2 per capita (sust.level.ghg);
- country's sustainability level (sust.level) and hdi (hdi.level).

The step by step procedure for this part of the thesis will be as follows:

1. Data pre-processing by doing One-Hot-Encoding to the data. It is needed to have a 0/1 or True/False dataset.
2. Application of the Apriori Algorithm in which the minimum support required for the itemset to be selected is 0.05.
3. Validation of Association Rules using:
   - the metric of interest as 'confidence' and the threshold at 65% and 80%;
   - the metric of interest as 'lift' and the threshold of 1 and 1,3.

   It is to compare which metrics and thresholds give better results.

4.4.3. Predictive models study

Predictive model study was selected to forecast changes in CO_2 emissions in 2050. This method uses the numerical variables to predict the outcome. Specifically, the Multiple Linear Regression was used because the aim was to predict the CO_2 emissions based on few, not one, variables.

4.3.3.1 Multiple Linear Regression model

The Multiple Linear Regression model will predict CO_2 emissions in 2050 based on the year, Human Development Index, Air Quality Index and percentage of human-made land. The goal will be to predict the average CO_2 emissions in the world in 2050, but also for each country separately.

The variables that will be used are:

- Year (year) - 2050;
- Air Quality Index (aqi) - average value of 123;
- Human Development Index (hdi) - average value of 0.673;
- Percentage of human-made land dominance (urban.agricult.perc) - average value of 44.

The step by step procedure for this part of the thesis will be as follows:

1. Calculating the average values of selected variables.
2. Selecting dependent (co2.em) and independent variables (year, aqi, hdi, urban.agricult.perc).
3. Fitting the model.
4. Predicting the output.

5. Results

5.1. Data selection

As mentioned in section 4., the final dataset contains values selected between the year 1990 and 2019, for 50 different countries. There are 31 variables selected in total. The topics covered by the variables are: HDI, air pollution, climate change, greenhouse gases and land cover.

5.2. Data preprocessing

5.2.1. Missing values treatment

- **HDI dataset**

Even if there is some missing data for HDI, it was decided not to interpolate the missing values. Instead, the missing values were filled with 0 which will be treated as no data.

- **Air pollution dataset**

Since the data obtained for O_3, PM2.5 and PM10 was only for some of the years between the years 1990-2015, the missing data were calculated using the interpolation with the cubic method.

What is more, in order to avoid negative values of these variables, the following equation was used:

*PM10 exposure = 0 - (PM10 exposure * 0.5)*

Each following value after the year 2011, if it was lower than 0, was 0.5 times higher than the previous one.

- **Climate change dataset**

There is no missing data in this dataset.

- **Greenhouse gases dataset**

There is no missing data in this dataset.

- **Land cover dataset**

In general, there was no missing data in this data set. The only variable with missing values was urbanized land (uran.land). Data collected for urbanized land was only for the years 1990, 2000 and 2015. To achieve more compact data, missing years were filled using the interpolation with the cubic method.

5.2.2. New variables creation

- **HDI dataset**

HDI level (hdi.level)

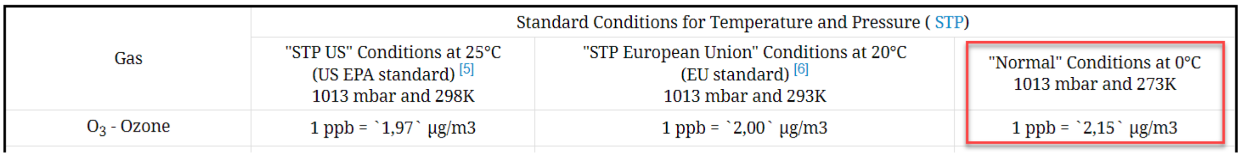
Human Development Index levels were created based on the following conditions:

1) if hdi was above 0.8 - very high hdi;
2) if hdi was between 0.8 and 0.7 - high hdi;
3) if hdi between 0.7 and 0.55 - medium hdi;
4) if hdi was below 0.55 - low hdi;
5) if hdi was 0 - no data.

- **Air pollution dataset**

O_3 exposure (o3.exp)

The unit of measurement of ozone exposure was in parts per billion (ppb). For the purpose of further calculations, it had to be converted into micrograms per cubic meter (µg/m³).



Gas	Standard Conditions for Temperature and Pressure (STP)		
	"STP US" Conditions at 25°C (US EPA standard) [5] 1013 mbar and 298K	"STP European Union" Conditions at 20°C (EU standard) [6] 1013 mbar and 293K	"Normal" Conditions at 0°C 1013 mbar and 273K
O_3 - Ozone	1 ppb = `1,97` µg/m3	1 ppb = `2,00` µg/m3	1 ppb = `2,15` µg/m3

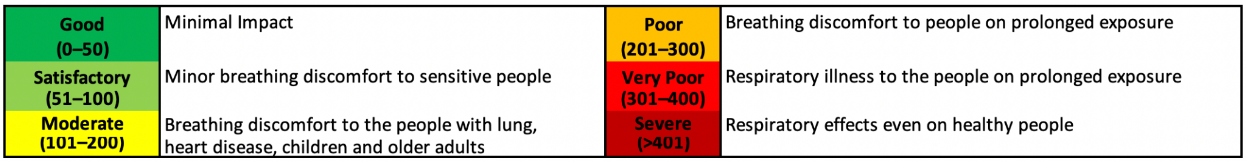
Figure 9. Ozone conversion method. (World Air Quality Index Project, 2021)

Figure 9 shows how the calculation was done. The use of "Normal" conditions is determined by the fact that the counties are spreaded worldwide (World Air Quality Index Project, 2021).

Air quality index (aqi)

To calculate the air quality index, the following conditions had to be compiled: both PM2.5 and PM10 values and at least one of SO_2, NO_x, NH_3, CO or O_3 values had to exist. In case of the gathered data, the values used to calculate AQI were PM2.5, PM10 and O_3. The final AQI is the maximum sub-index among used indices.

Air quality buckets (aqi.bucket)



Good (0–50)	Minimal Impact	Poor (201–300)	Breathing discomfort to people on prolonged exposure
Satisfactory (51–100)	Minor breathing discomfort to sensitive people	Very Poor (301–400)	Respiratory illness to the people on prolonged exposure
Moderate (101–200)	Breathing discomfort to the people with lung, heart disease, children and older adults	Severe (>401)	Respiratory effects even on healthy people

Figure 10. Air pollution levels. (Calculating AQI (Air Quality Index) Tutorial, 2021)

In order to calculate the AQI levels (buckets) the maximum values of PM2.5, PM10 or O_3 were taken into account. Next, the AQI level was dependent on conditions that can be seen on figure 10.

- **Climate change dataset**

Temperature pattern (temp.pat) and precipitation pattern (prec.pat)

To receive the trendline patterns, the table had to be created manually with the variables based on the trendline direction.

The assumption was that if the trendline of the temperature and precipitation patterns wass clearly going up, it had an upward trend. If the difference between start and end of the trendline was almost not visible, it was considered a stable trend. If the trendline was clearly going down, the trend was decreasing.

- **Land cover dataset**

Nature percentage (forest.inland.perc)

It was calculated as a sum of forest area and inland water and later as a percentage of country area.

Human percentage (urban.agri.perc)

It was calculated as a sum of urban and agricultural land and later as a percentage of country area.

Land dominance (land.dom)

Land dominance is a variable that shows which influence is stronger in the country - the one of nature or of the human. This variable has three conditions: human, nature and balance.

- human - the percentage of the urban and agricultural land (urban.agri.perc) is higher than the percentage of forest area and inland waters (forest.inland.perc);
- nature - the percentage of the forest area and inland waters (forest.inland.perc) is higher than the percentage of urban and agricultural land (urban.agri.perc);
- balance - the percentage of both is between 45 and 55.

5.2.2.1. Country classification

- **Air pollution dataset**

The most important indicator in the air pollution dataset was the air quality index. The better the air quality, the more sustainable the country is.

AQI bucket (aqi.bucket)	**Classification (sust.level.aqi)**
Good	Sustainable
Satisfactory	Quite sustainable
Moderate	Quite sustainable
Poor	Not sustainable
Very poor	Not sustainable
Severe	Not sustainable

Table 8. AQI bucket classification.

- **Climate change dataset**

To recognize the sustainability level when it comes to climate change, the pattern change of temperature was taken into consideration. It is crucial to highlight that this classification is for a 1990-2015 change.

In order to have a consistent dataset without NaN values, it will be assumed that the final sustainability level for the country in 2015 will be the same in all previous years. If the country in 2015 was sustainable, it means it has carried out good practices on climate change since 1990 and vice versa.

Minimum and maximum temperature patterns depended on the trend lines. If the trendline was going up, it meant that the temperatures are rising which is not a good sign. That is why this pattern is not sustainable. On the other hand, if temperatures are dropping, it can be seen as a positive sign and a sustainable pattern. If the trend is more or less steady - it is considered quite sustainable.

Minimum temperature pattern (min.temp.pat)	**Classification (sust.level.min)**
Down	Sustainable
Balance	Quite sustainable
Up	Not sustainable

Table 9. Minimum temperature pattern classification.

Maximum temperature pattern (max.temp.pat)	**Classification (sust.level.max)**
Down	Sustainable
Balance	Quite sustainable
Up	Not sustainable

Table 10. Maximum temperature pattern classification.

The final temperature pattern and the classification were based on the following conditions:

Temperature pattern (temp.pat)		**Classification (sust.level.temp)**
Classification min.temp.pat	**Classification max.temp.pat**	
Sustainable	Sustainable	Sustainable
Sustainable	Quite sustainable	Quite sustainable
Sustainable	Not sustainable	Quite sustainable
Quite sustainable	Sustainable	Quite sustainable
Quite sustainable	Quite sustainable	Quite sustainable
Quite sustainable	Not sustainable	Not sustainable
Not sustainable	Sustainable	Quite sustainable
Not sustainable	Quite sustainable	Not sustainable
Not sustainable	Not sustainable	Not sustainable

Table 11. Temperature pattern classification.

- **Greenhouse gases dataset**

CO_2 is the most powerful greenhouse gas. Its emissions say a lot about the country's profile. Based on 2016 data[2], the average global CO_2 emissions per capita were 4,8 tons.

CO_2 emissions per capita (co2.cons.cap)	**Classification (sust.level.ghg)**
≤ 4,8 tons	Sustainable
4,9 - 8,2 tons	Quite sustainable
≥ 8,3 tons	Not sustainable

Table 12. CO_2 emissions per capita classification.

- **Land cover dataset**

To classify countries depending on land area, percentage land domination of forest land or urbanized land and agricultural land will be taken into consideration. The more natural domination, the more sustainable the country.

[2] https://www.worldometers.info/co2-emissions/co2-emissions-per-capita/

Land dominance (land.dom)	Classification (sust.level.land)
Nature	Sustainable
Balance	Quite sustainable
Human	Not sustainable

Table 13. Land dominance classification.

5.2.3. Data merge

Final classification of country's sustainability level was created taking into account only the following classifications: sust.lev.land (land cover dataset), sust.lev.ghg (greenhouse gases dataset) and sust.lev.aqi (air pollution dataset). It is because, as mentioned above, the classification of climate change dataset is only for the year 2015.

Conditions	Classification (sust.level)
3x sustainable 2x sustainable & 1x quite sustainable	Sustainable
2x sustainable & 1x not sustainable 2x quite sustainable & 1x not sustainable 3x quite sustainable 2x quite sustainable & 1x sustainable 1x sustainable, 1x quite sustainable & 1x not sustainable	Quite sustainable
3x not sustainable 2x not sustainable & 1x quite sustainable 2x not sustainable & 1x sustainable	Not sustainable

Table 14. Final sustainability level classification.

After merging all above dataset with hdi and climate patterns dataset together, there are many columns that were just a support for creating other columns. In order to have a more consistent dataset, these unnecessary columns will be deleted.

forest.perc	inland.perc	urban.perc	agricult.perc	o3.aqi
pm2.5.aqi	sust.level.min	sust.level.max	checks	pm10.aqi

Table 15. Columns deleted from the final dataset.

After deleting the unnecessary columns and creating the final classification level for each country each year, the final dataset contains *1300 rows* and *45 columns*, where the index is a country name and a year.

5.2.4. Correlation and dimensionality reduction

Pearson correlation coefficient was calculated for the final dataset. At the beginning, the correlation matrix contained 44 variables. In order to reduce the dimension of the dataset and prevent the information duplication, variables with high correlation (more than r=0.7) were removed.

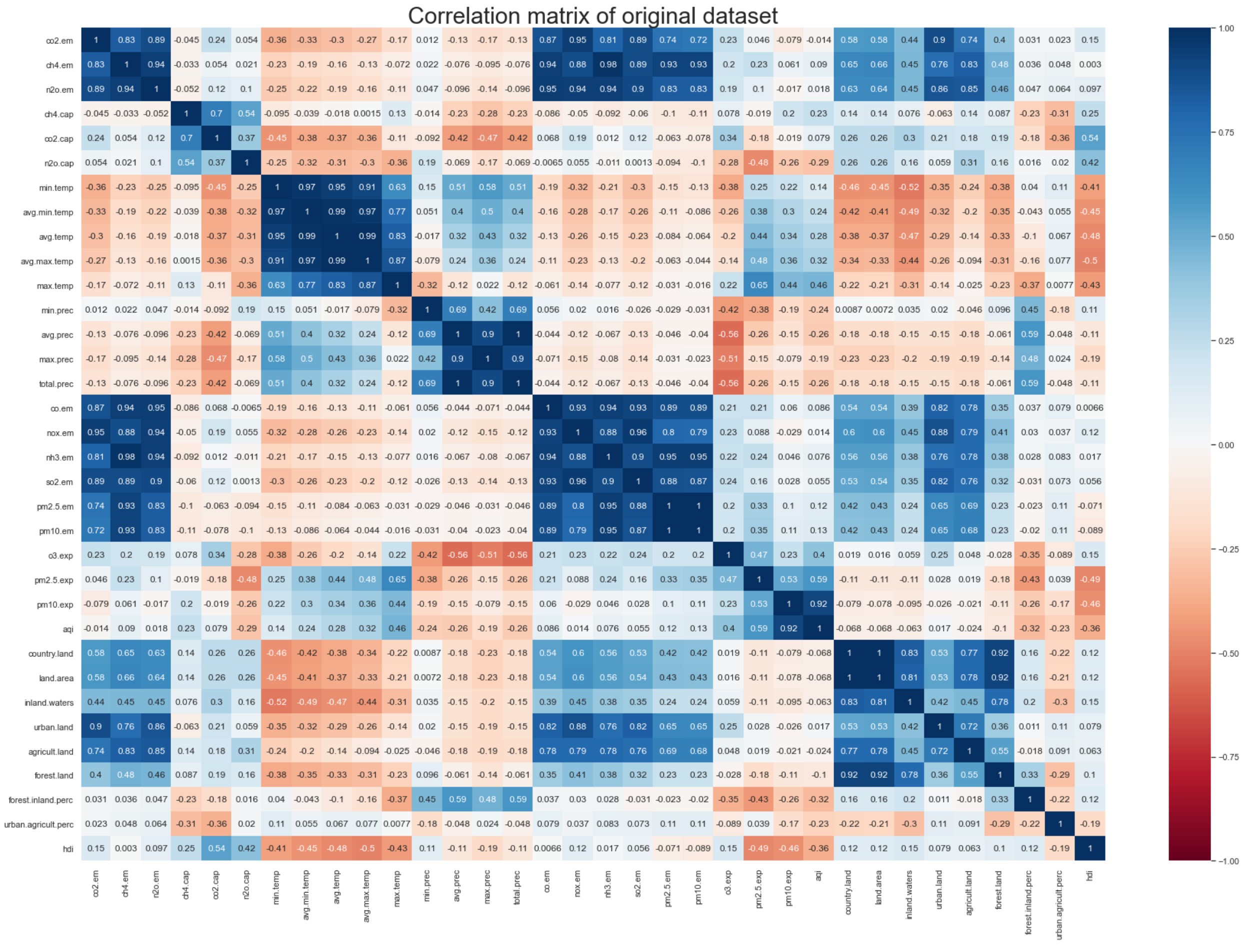


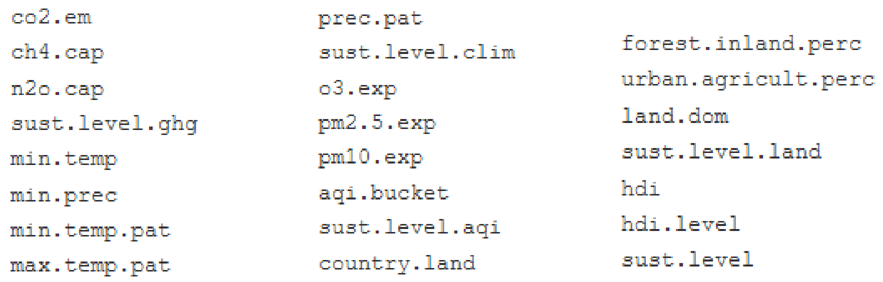
Figure 11. Correlation coefficient matrix of original dataset.

As expected, the highest correlation is shown by values from the same datasets or very similar topics. Few observations can be made here:

1. Greenhouse gases emissions (co2.em, n2o.em, ch4.em) are highly correlated with each other but also with air pollution emitants (pm2.5.em, pm10.em, co.em, nox.em, etc.), with r>0.72;
2. Emission values show moderate to high correlation with land data variables (urban.land, land.area, etc.), with correlation from 0.4 to 0.9;
3. Greenhouse gases emissions are not correlated at all with greenhouse gases emissions per capita;
4. Maximum temperature is relatively close to having a strong correlation with PM2.5 exposure (r=0.65);
5. HDI has moderate correlation with CO_2 emissions per capita and almost no correlation with other variables (except temperature variables with low correlation between -0.41 and -0.5 and particulate matter with also low correlation with r=-0.46 and r=-0.49);

As expected, values from the same datasets present the highest correlations. But it is also interesting to see that HDI has no high correlation with other values. Before the analysis, it was expected to be the opposite.

As mentioned before, highly correlated columns were removed. The final dataset now contains 23 variables. The columns are as follows:



```
co2.em
ch4.cap
n2o.cap
sust.level.ghg
min.temp
min.prec
min.temp.pat
max.temp.pat
prec.pat
sust.level.clim
o3.exp
pm2.5.exp
pm10.exp
aqi.bucket
sust.level.aqi
country.land
forest.inland.perc
urban.agricult.perc
land.dom
sust.level.land
hdi
hdi.level
sust.level
```

Figure 12. Columns used for modeling.

After the dimension reduction, correlation was done for the reduced dataset. It can be observed that in almost 99% cases, these variables show weak to almost no correlation. Some of them still show moderate correlation, like ch4.cap with n2o.cap (r=0.54), country.land with co2.em (r=0.58) or pm2.5.em with pm10.em (r=0.53).

This step was necessary to further simplify the modeling process.

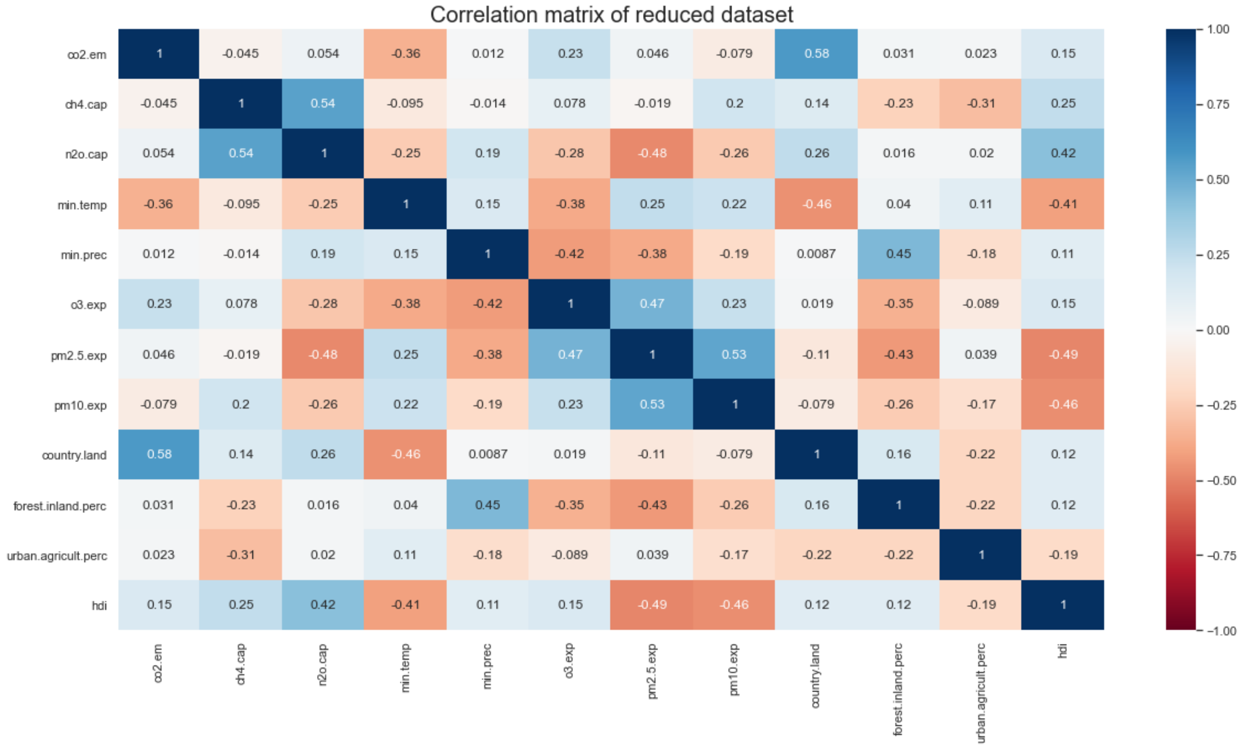


Figure 13. Correlation coefficient matrix of reduced dataset.

5.2. Exploratory data analysis

This analysis of the results aimed to prove the hypothesis that the climate and environmental changes are real. They could be positive or negative, depending on the variables.

- **HDI dataset**

Human Development Index describes how developed a society is in a given country. Obviously, the higher the number, the better. One would think that a more developed society will have better sustainability knowledge and the whole country in general will be more sustainable.

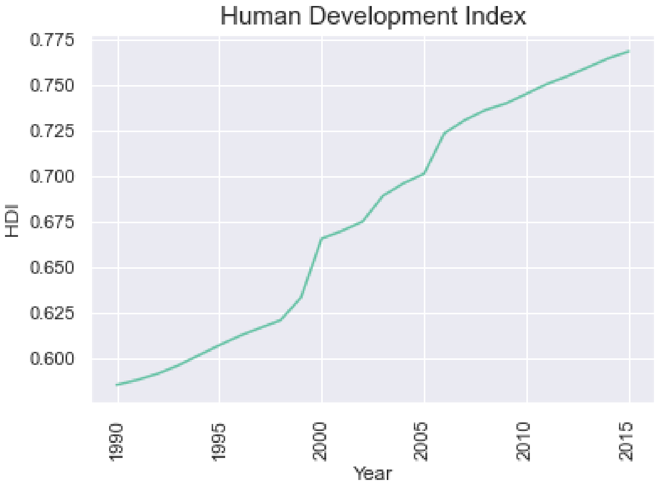


Figure 14. Human Development Index over the years.

As the figure above shows, over the years, the average HDI was steadily growing. It can mean that countries carried out a lot of practices to make their society more developed. Having all 50 countries grouped together, it is safe to say that average HDI for them is high, at the level of approximately 0.77. Overall, the average HDI was 0.678.

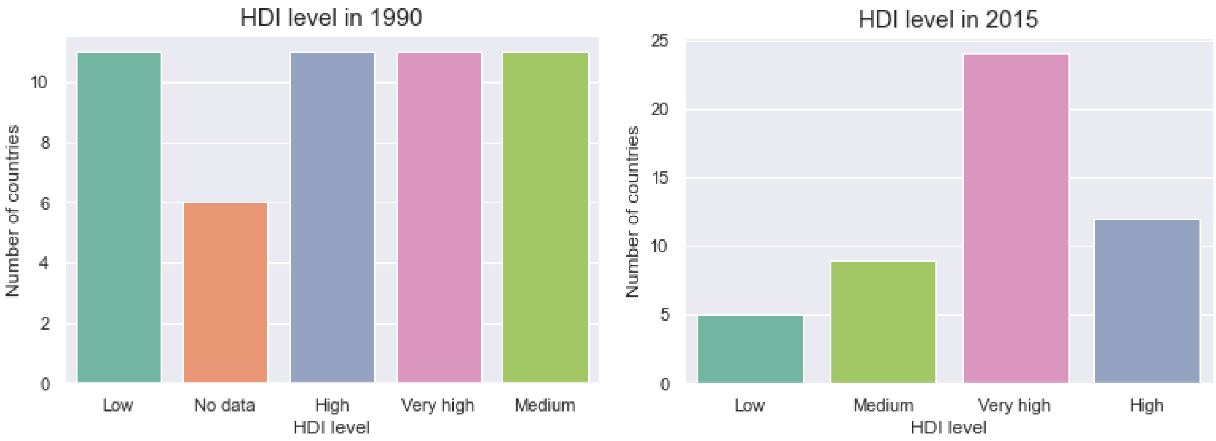


Figure 15. HDI level comparison of 1990 and 2015.

The comparison of countries' HDI levels in 1990 and 2015 looks promising. In 1990, there were the same numbers of low, medium, high and very high HDI levels, with the number of 11 each (44 in total). Also, there were 6 countries with no data available about HDI level. In 2015, low HDI dropped from 11 to 5, medium dropped from 11 to 9, high increased from 11 to 12 and very high increased from 11 to 24. What can it mean for sustainability? According to Neumayer, countries with high or very high HDI tend to be strongly unsustainable, mostly because of their high, not sustainable, CO_2 emissions. On the other hand, countries with the lowest HDI appear to be sustainable or quite sustainable (Neumayer, 2010).

- **Air pollution dataset**

The main pollutants that were taken into account when calculating air quality index were Ozone, PM2.5 and PM10.

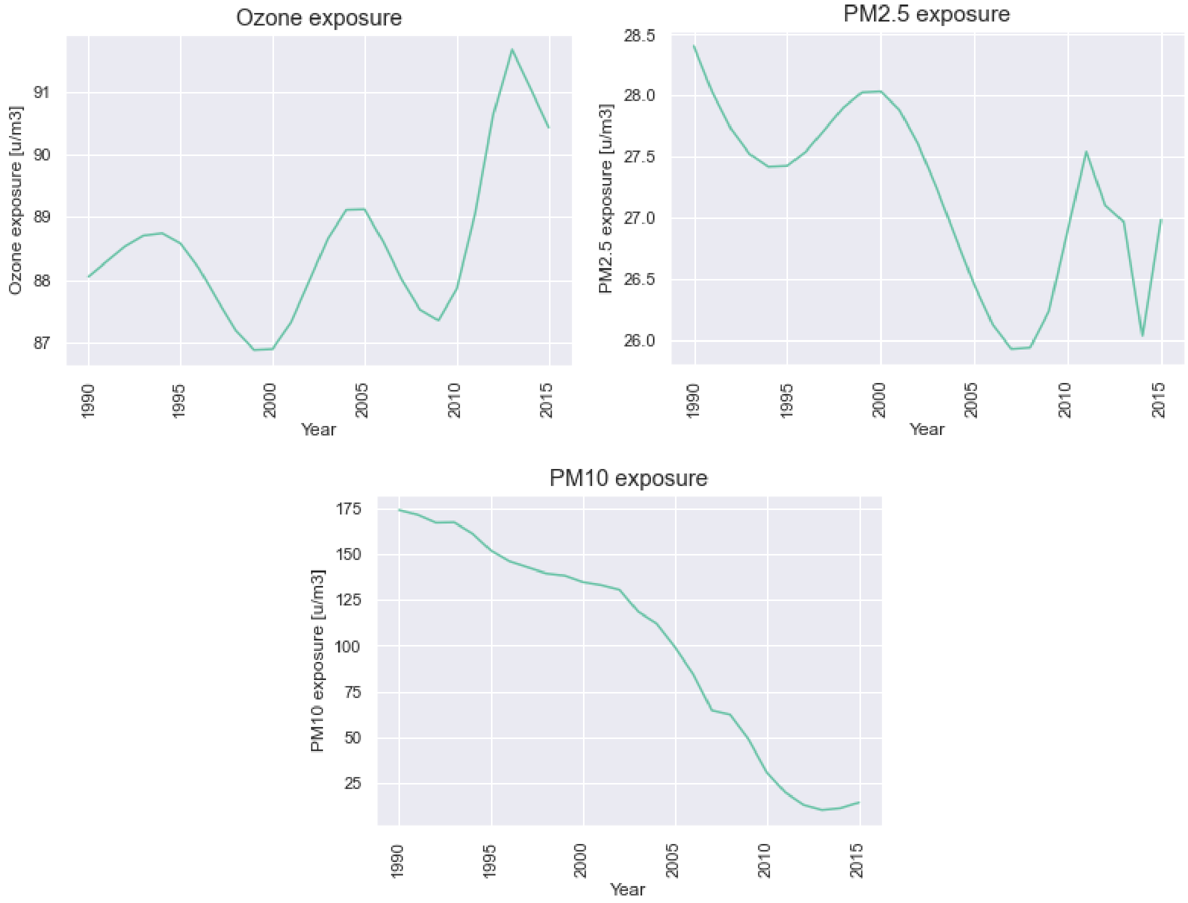


Figure 16. Ozone, PM2.5 and PM10 exposure over the years.

As shown on figure 16, over the years, the values of Ozone and PM2.5 exposure were fluctuating. In general, Ozone exposure was higher in 2015 than in 1990 and PM2.5 exposure was the opposite. PM10 exposure was steadily decreasing just to increase a tiny bit in 2015.

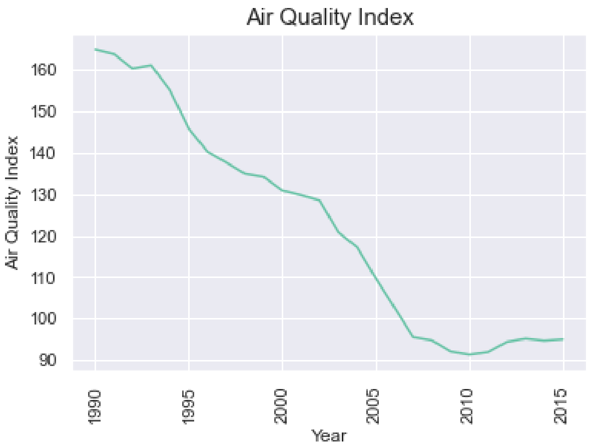


Figure 17. Air quality index over the years.

After all, the air quality index was decreasing over the years from the level of around 165 just to stabilize a bit in 2017 around the value of 95.

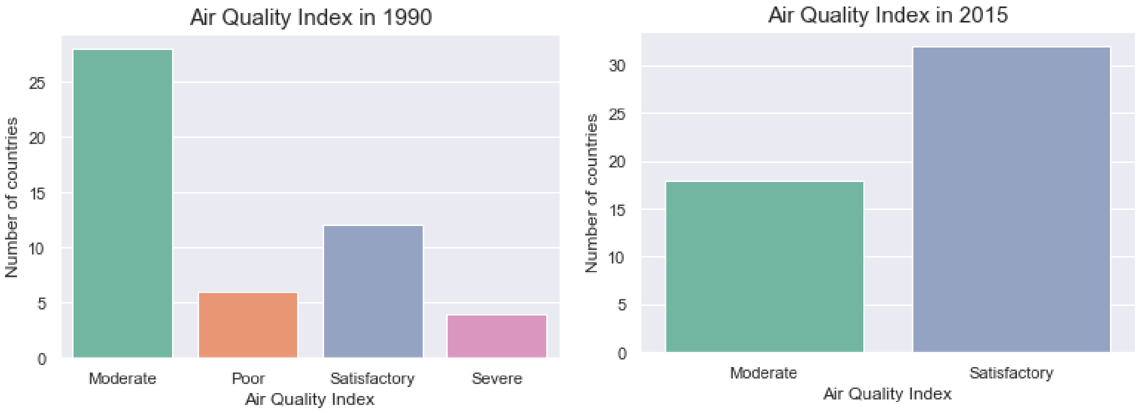


Figure 18. Air quality index comparison of 1990 and 2015.

In general, it is visible that the countries took a step forward to improve the air quality. Back in 1990, there were 10 countries with questionable air quality indexes (poor and severe). 28 countries had moderate AQI and only 12 - satisfactory. In 2015, 17 countries had moderate AQI and 33 - satisfactory. There is no country that has a good air quality index. Nevertheless, the change is huge and it can be a result of a series of policies established by countries' governments when it comes to the emission of air pollutants. Moreover, it can mean that people became more aware of the environment and, for example, stopped using cars all the time.

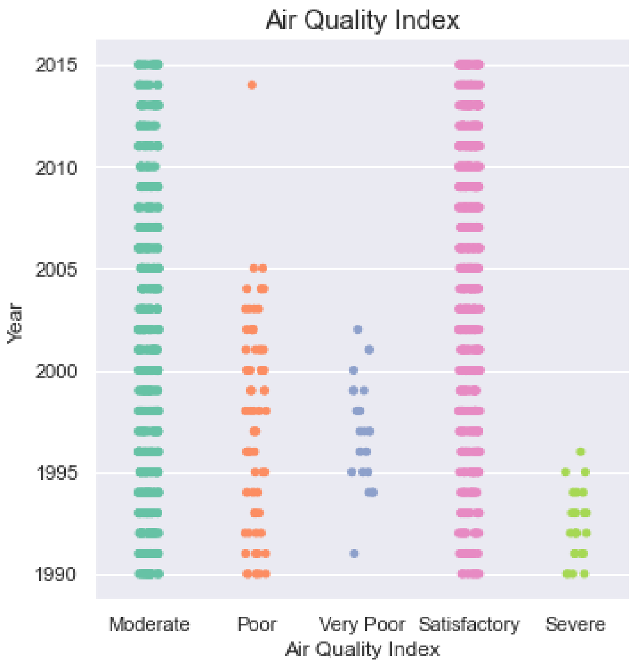


Figure 19. Air quality index distribution over the years.

It is clearly visible that from severe, satisfactory, poor and very poor air quality levels, countries started to have only moderate or satisfactory levels which shows significant improvement.

- **Climate change dataset**

Rising temperatures or higher precipitation are the signs of climate changes all over the world. Even though in some countries, the temperature or precipitation patterns were going down or were steady at the same level, overall pattern change shows an upward movement.

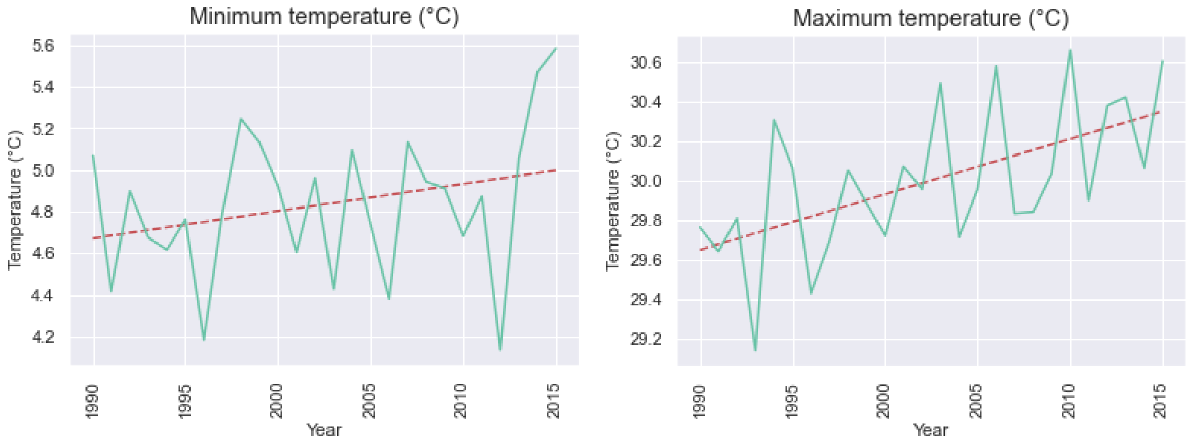


Figure 20. Minimum and maximum temperature pattern over the years.

Minimum temperature as well as maximum temperature increased over the 26 years. The average minimum temperature rose for about 1.3ºC (from around 4.7ºC to 5ºC), whereas average maximum temperature rose for about 0.8ºC (from around 29.6ºC to almost 30.4ºC)

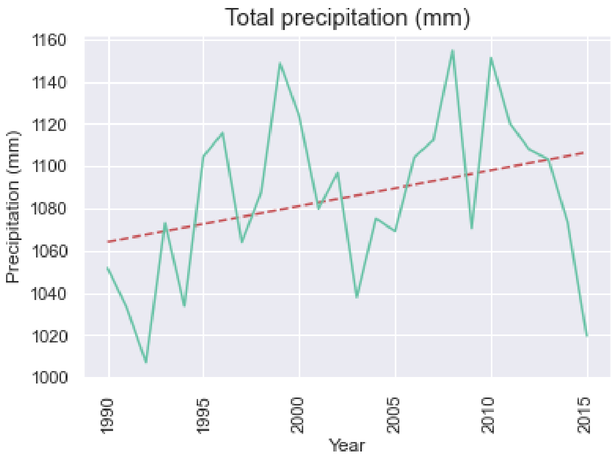


Figure 21. Total precipitation pattern over the years.

When it comes to total precipitation, the change is also visible. Even if the precipitation value in 2015 is lower than in 1990, over the years, total average precipitation increased from 1065mm to 1105mm in depth.

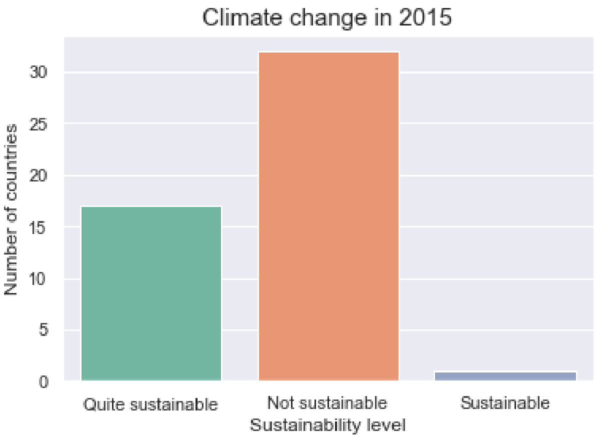


Figure 22. Climate changes in 2015.

From all 50 countries, only one of them had sustainable climate change which means that the temperature pattern was going down. It was Thailand. A stable temperature change could be seen in 17 countries and in 32 of them, this change was not sustainable (upward pattern). However, the classification adopted in section 5.1.3. does not reflect the reality 100%. The downtrend in the case of Thailand might as well be treated as a not sustainable one. Nevertheless, on account of this dissertation, this classification is considered as the reference.

- **Greenhouse gases dataset**

Greenhouse gases, like CO_2, CH_4 and N,O, are the most common pollutants. Overall trends show the emissions in tons whereas the sustainability level refers to emissions of CO_2 per capita.

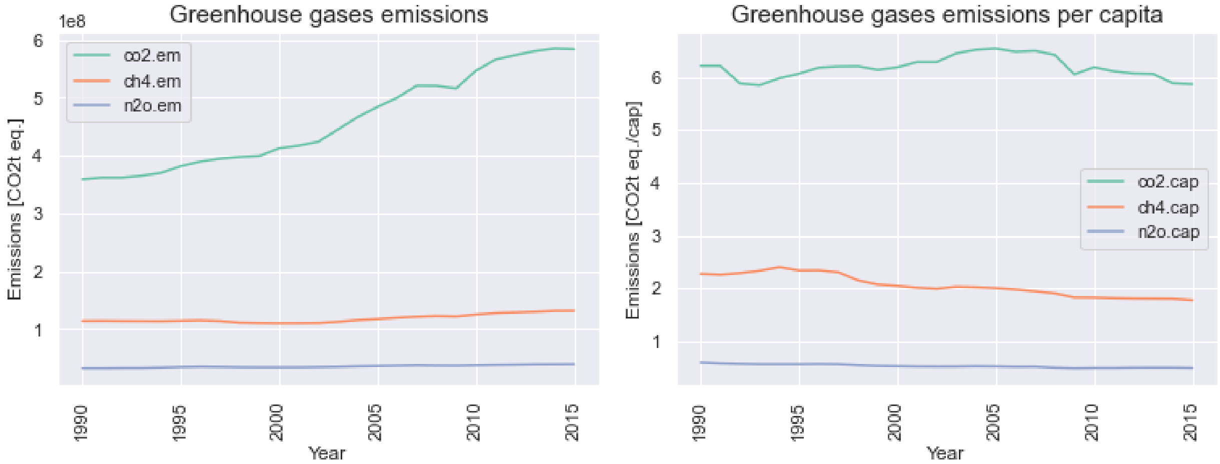


Figure 23. Greenhouse gases emissions over the years.

Looking at figure 23 one can say that the greenhouse gases emissions per capita are stable over the years, especially CO_2 per capita emissions with the value of 6.6t approximately. However, when taking a closer look at the countries separately (code in Annex), it is clearly visible that this general trend is wrong. There are a lot of countries with very high values and a lot of them with very low ones. Thus, the average value is not reliable at all. It can be treated only as a general reference. On the other hand, this trend can show that even if the value is not too reliable, it stays the same over the years.

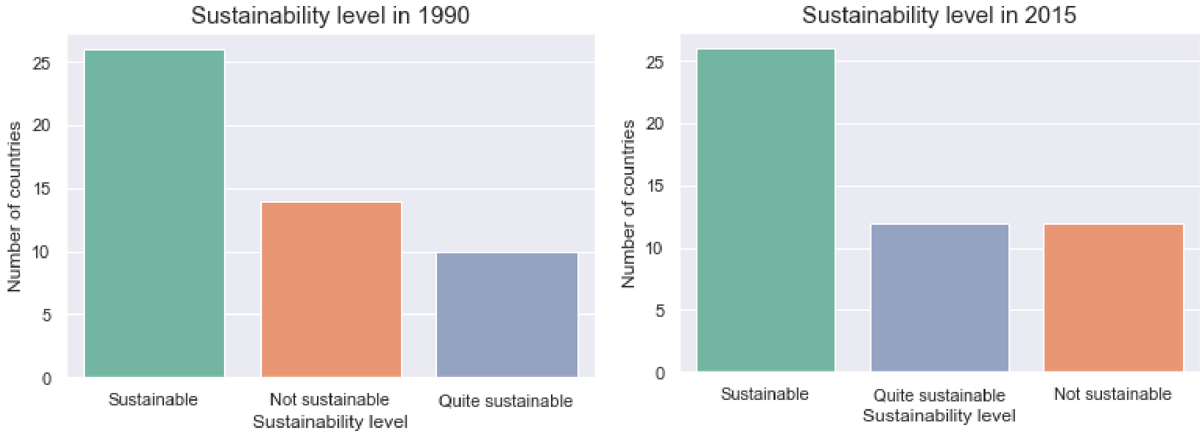


Figure 24. Sustainability level of CO2 emissions per capita comparison of 1990 and 2015.

As can be seen in the figure above, 26 countries in 1990 and 2015 had sustainable levels of CO_2 emissions per capita. Over the years, the number of countries with not sustainable emissions reduced from 14 to 12 and quite sustainable ones rose from 10 to 12. This information can mean that countries, as in cases of air pollutants, made steps forward and started implementing the policies and spread the awareness amongst the citizens.

- **Land cover dataset**

The main objective of analysis of the land cover was to observe how human factors (like urbanized land) affect nature.

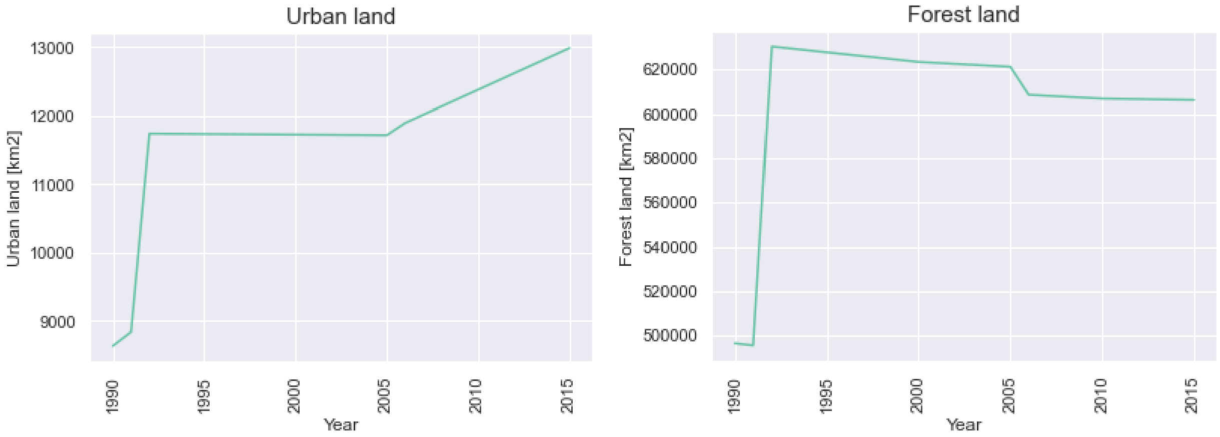


Figure 25. Urbanized land and forest area over the years.

Figure above shows that when the size of urbanized land was increasing, the size of forest area was decreasing. It is commonly known that human impact is almost always negative when it comes to nature. However, one has to take into account that there are some countries whose data was available only from 1992 or 2006.

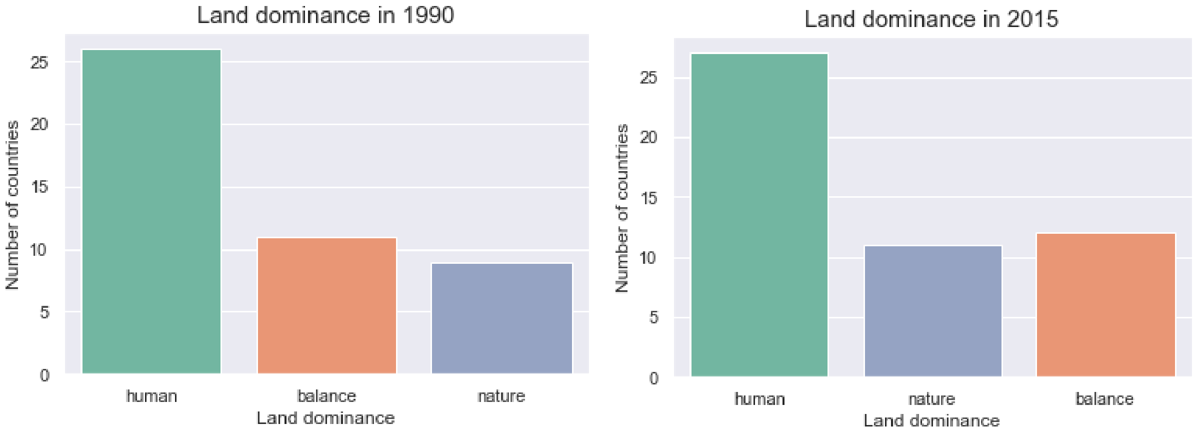


Figure 26. Sustainability level of land dominance comparison of 1990 and 2015.

Unfortunately, over the years, human dominance rose in 1 country. From 26 countries in 1990 to 27 countries in 2015. Moreover, nature dominance rose from 9 countries to 11 and balanced dominance decreased from 11 to 12. The countries are developing and spreading their urbanized areas. Even if the nature dominance increased slightly, the overall trend points out that the practices connected to land cover are not sustainable, as more than 50% is dominated by humans and around 25% is in balance with nature and humans.

5.3. Data mining

5.3.1. Profiling of the different countries according to the indicators

K-Means Clustering was performed for each country each year separately and also for each country in general. In the first case, each country was evaluated based on the year. It aimed to check how countries have improved (or not) over the years. In the second case, the average value of each variable over the years for each country was taken into account to check the general distribution of countries.

In order to determine the optimal number of clusters for both cases, the elbow method was used.

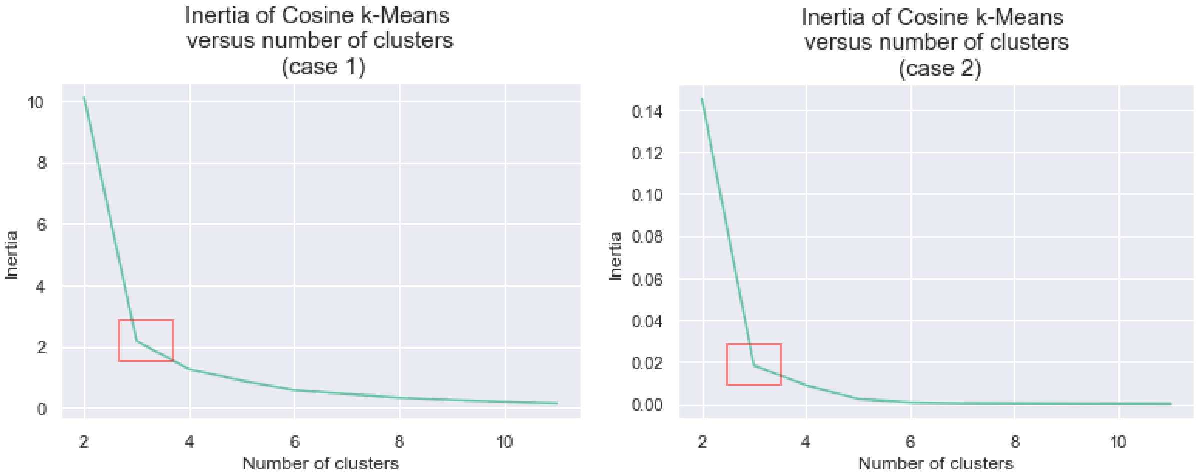


Figure 27. Elbow method for choosing an optimal number of clusters.

Figure 27 indicates that in both cases the optimal number of clusters would be 3. It is also a good sign, as the principal assumption of the thesis was to classify countries into three groups: sustainable, quite sustainable and not sustainable ones.

- **Country-year (case 1)**

The cluster distribution is as follows:

- cluster 1 - 40 countries-years;
- cluster 2 - 97 countries-years,
- cluster 3 - 1141 countries-years.

After receiving the number of countries in each cluster, the next step was to calculate the centroids for each variable of each cluster.

Variable	Cluster 1 (40 countries-years)	Cluster 2 (97 countries-years)	Cluster 3 (1141 countries-years)
co2.em	6426815000	1589117000	156991400
ch4.cap	1,76	1,26	2,13
n2o.cap	0,77	0,34	0,53
min.temp	-9,96	-8,15	6,71
min.prec	32,04	26,42	26,23
o3.exp	107,24	100,65	86,66
pm2.5.exp	26,08	31,20	28,89
pm10.exp	66,83	98,70	104,25
country.land	9659981	6222195	1131476
forest.inland.perc	31,88	45,38	31,04
urban.agricult.perc	48,41	35,22	45,06
hdi	0,81	0,73	0,66

Table 16. Description of centroids for each variable in each cluster.

All variables in table 16 are important factors in climate and environmental changes. Analyzing the table above, countries-years from cluster 1 are the biggest ones, from cluster 2 are medium-sized and from cluster 3 - the smallest ones. Observations in cluster 1 have the highest values of CO_2 emissions that are around 4,2 times higher than in cluster 2 and almost 41 times higher than in cluster 3. Those are also countries-years with the highest percentage of human-made land (48%) and 32% of natural land, whereas in cluster 2 there is the highest percentage of natural land (45%) and the lowest percentage of human-made land (35%). In cluster 3, there is 31% of natural land and 45% of human-made one. Ozone exposure is the highest in cluster 1, the lowest in cluster 3. Emissions of other greenhouse gases (CH_4 and N_2O) are the highest in cluster 1 and 3, and the lowest in cluster 2. In cluster 1, the exposure to PM10 and PM2.5 is the lowest, whereas in cluster 2 and 3 is very high. It may seem surprising, as countries-years in cluster 1 seemed to be the worst ones. When it comes to the minimum temperature and minimum precipitation, in cluster one there are the highest minimum precipitation and the lowest minimum temperature. In cluster 2, both of these values are in the middle, and in cluster 3, minimum temperature is the highest and minimum precipitation - the lowest. What is surprising, that countries with the highest, negative impact, have the highest HDI so it could be assumed these are the most developed countries. Cluster 2 has medium values of HDI, so there should be medium developed countries. Again, cluster 3 has the lowest HDI, which can mean there are least developed countries.

Further, 2 variables (CO_2 emissions and country area) for each year were taken and compared against each other to see how the clusters are distributed on the scatterplot.

The main problem faced here was the labeling. If performed in Python, the whole figure would be illegible. It was decided to make labels by hand, for the most important countries-years/changes.

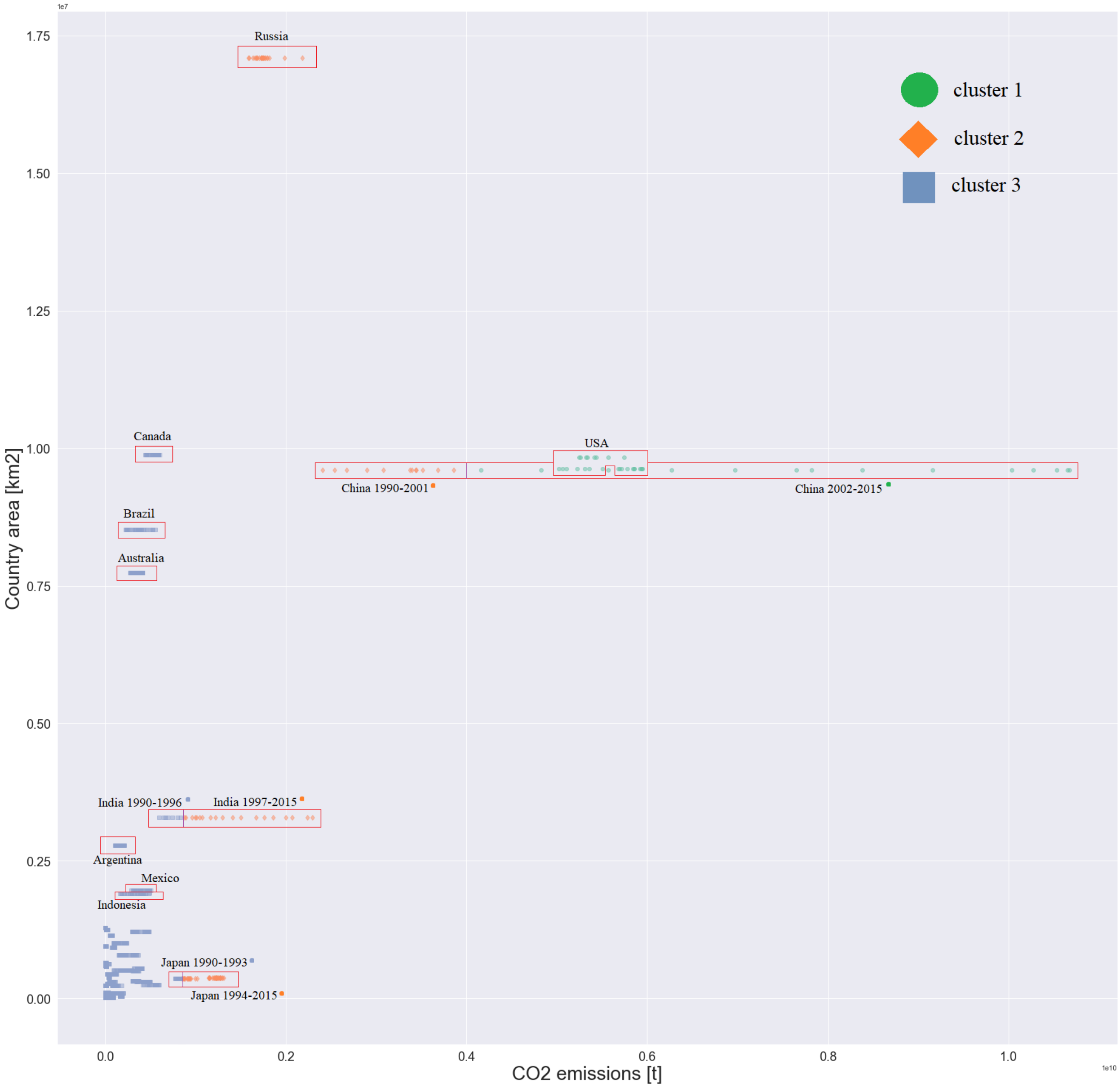


Figure 28. Clustering of each country each year depending on CO_2 emissions and country areas.

Most countries-years observations (1141) are in cluster 3 (blue). It seems that it is the cluster with overall the lowest CO_2 emissions and the lowest country land area. Some exceptions are the countries with large land areas but still low emissions (Canada, Brazil, Australia). Countries, that did not change the cluster over the years are: Canada, United Kingdom, Italy, Mexico, South Africa, France, Brazil, Australia, Indonesia, Poland, Spain, Turkey, Thailand, Netherlands, Argentina, Egypt, United Arab Emirates, Uzbekistan, Romania, Nigeria, Philippines, Austria, Colombia, Kuwait, Montenegro, Israel, Switzerland, Norway, Morocco, New Zealand, Cuba, Estonia, Angola, Sri Lanka, Guatemala, Kenya, Ghana, El Salvador, Costa Rica, Tanzania, Afghanistan, Madagascar, Fiji, Chad. This cluster also contains Japan in 1990-1992 and India in 1990-1996. Both of these countries, over the later years, were emitting more CO_2, thus they were later moved to cluster 2. There is no specific classification as to why these countries are there. They vary from developed, developing to least developed countries.

Cluster 1 and cluster 2 are definitely the small ones, with 44 and 97 observations, respectively.

Countries from cluster 2 (orange) are: Russia, India in 1997-2015, Japan in 1993-2015, Germany and China in 1990-2001. This cluster contains both developing (Russia, India, China) and developed countries (Japan, Germany). There is a relationship between small, developed countries producing more or less the same amount of CO_2 emissions as big, developing countries. As mentioned in case of cluster 1, Japan and India had higher emissions in later years, that is why they were assigned to this cluster. Interesting is the fact that this cluster also contains China in 1990-2001.

In cluster 1 (green) there are China in 2002-2015 and the United States of America (44 observations). While China quadrupled its CO_2 emissions between 1990 and 2015 (from around 2.5 to 10.6 billion tons), the USA had more or less steady emissions over the years, around 5.5 billion tons. That is why China was moved from cluster 1 to cluster 2. These are two of the biggest and most powerful economies in the world so it makes sense that they were put together as the biggest pollutants. It is also not surprising that for all the years they were the most unsustainable countries.

When describing the cluster based on the centroids, some assumptions were made about the level of countries' development for each cluster. However, further analysis of the scatterplot shown, that these assumptions were wrong. Cluster 1 does not necessarily contain the most developed countries, there is one developed and one developing country. In cluster 2 is the same situation. Cluster 3 is the most diverse one, as it contains countries from all development levels.

At the end, from all the information gathered together, the final interpretation can be made. The profiling of the countries is as follows:

Cluster	Number of instances	Cluster size	Cluster label
Cluster 1	40	Small	Not sustainable
Cluster 2	97	Small	Quite sustainable
Cluster 3	1141	Large	Sustainable

Table 17. Profiling of the clusters.

- **Country (case 2)**

The cluster distribution is as follows:

- cluster 1 - 4 countries;
- cluster 2 - 2 countries,
- cluster 3 - 44 countries.

After receiving the number of countries in each cluster, the next step was to calculate the centroids for each variable of each cluster.

Variable	**Cluster 1** **(2 countries)**	**Cluster 2** **(4 countries)**	**Cluster 3** **(44 countries)**
co2.em	5680212000	1272965000	146238700
ch4.cap	1,57	1,27	2,14
n2o.cap	0,65	0,35	0,53
min.temp	-10,63	-6,27	6,57
min.prec	26,98	26,33 ⬇	26,43 ⬆
o3.exp	106,64	99,61	86,66
pm2.5.exp	31,57 ⬆	29,64 ⬆	26,71 ⬇
pm10.exp	109,31 ⬆	85,38 ⬇	102,56 ⬆
country.land	9646139	5280117	1106326
forest.inland.perc	28,99 ⬇	46,36	45,06 ⬆
urban.agricult.perc	50,00	35,61	45,41
hdi	0,76	0,75	0,66

Table 18. Description of centroids for each variable in each cluster.

In comparison to countries-years evolution, countries alone have slightly different results. All the variables are distributed very similarly, however, there are some changes. As expected, countries in cluster 1 started to have the highest PM10 and PM2.5 exposure, when before they were the lowest ones. Now, cluster 2 and 3 countries have more or less the same values as before. Minimum precipitation shows almost no change with the values being basically the same. Last change is that natural land decreased in countries in cluster 1 and increased in countries in cluster 3.

Again, the same as before, 2 variables (CO_2 emissions and country area) were taken and compared against each other to see how the clusters are distributed on the scatterplot.

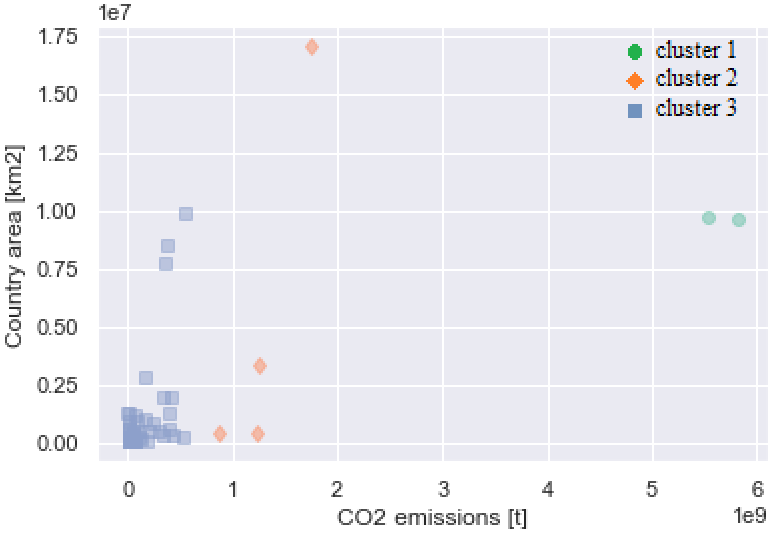


Figure 29. Clustering of each country depending on CO_2 emissions and country areas.

Most countries are in *cluster 3*. These are the following countries: Canada, United Kingdom, Italy, Mexico, South Africa, France, Brazil, Australia, Indonesia, Poland, Spain, Turkey, Thailand, Netherlands, Argentina, Egypt, United Arab Emirates, Uzbekistan, Romania, Nigeria, Philippines, Austria, Colombia, Kuwait, Montenegro, Israel, Switzerland, Norway, Morocco, New Zealand, Cuba, Estonia, Angola, Sri Lanka, Guatemala, Kenya, Ghana, El Salvador, Costa Rica, Tanzania, Afghanistan, Madagascar, Fiji, Chad. All the countries that were in this cluster over all the years, stayed here.

Countries from *cluster 1* are Russia, India, Japan and Germany. Since India and Japan were in both cluster 3 and cluster 1 during the evolution, the final classification assigned them to cluster 1.

In *cluster 2* there are China and the United States of America. As mentioned in case 1, these countries are not a surprise. Because of growing CO_2 emissions in China over the years, it was assigned to cluster 2 for overall comparison.

Based on the centroids table interpretation (table 19) and cluster distribution on figure 28, the final interpretation can be made. The profiling of the countries is as follows:

Cluster	Number of instances	Cluster size	Cluster label
Cluster 1	2	Small	Not sustainable
Cluster 2	4	Small	Quite sustainable
Cluster 3	44	Large	Sustainable

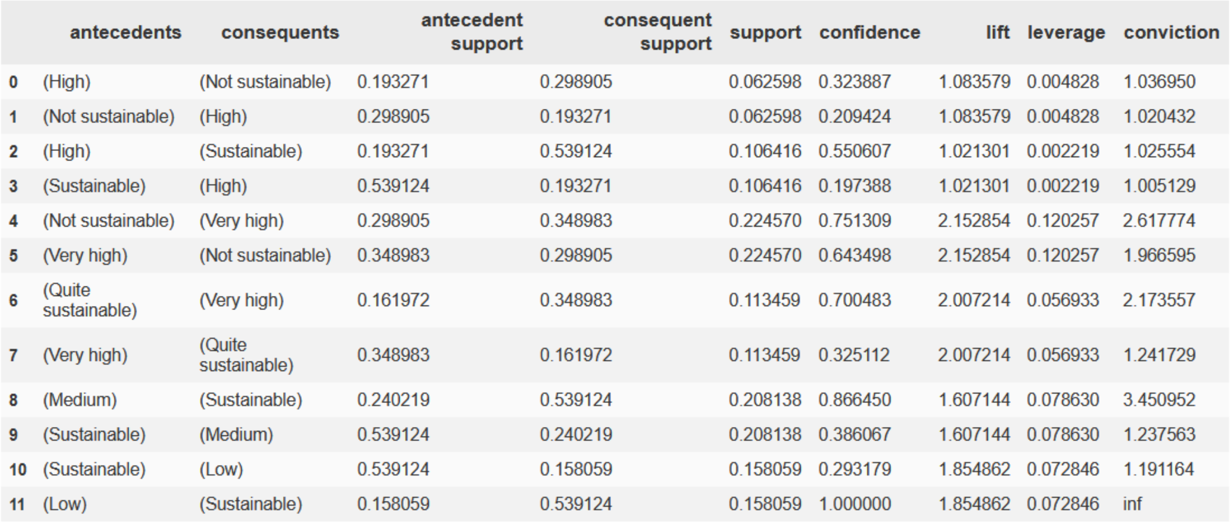
Table 19. Profiling of the clusters.

To sum up everything, both cases showed that countries in cluster 1 are not sustainable, in cluster 2 are quite sustainable and in cluster 3 - sustainable. In general, there are only 3 countries that changed the clusters over the years. These are India and Japan, which moved from cluster 3 to cluster 2 and China, which moved from cluster 2 to cluster 1. That is why, in the general classification (case 2), those countries were assigned to clusters with worse sustainability levels.

5.3.2. Associative models study

As mentioned in section 4., this part was done for chosen pairs of variables. The most important factors to evaluate the association rules obtained are support, confidence and lift.

- **co2 per capita (sust.level.ghg) and hdi (hdi.level)**

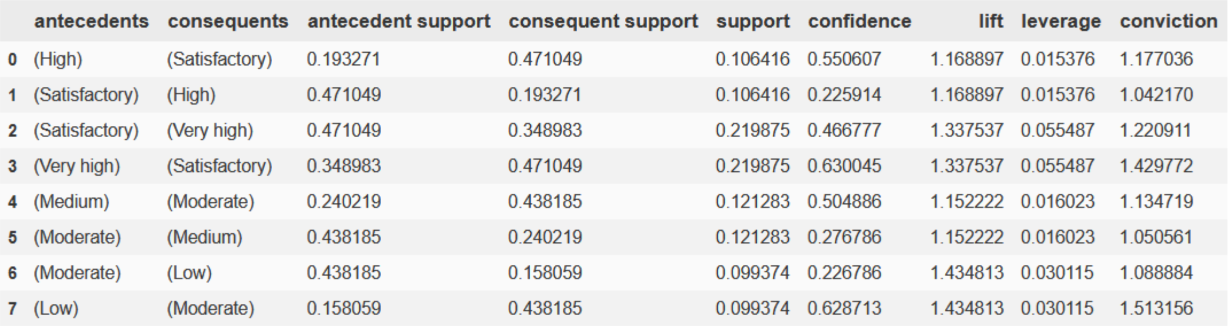


	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(High)	(Not sustainable)	0.193271	0.298905	0.062598	0.323887	1.083579	0.004828	1.036950
1	(Not sustainable)	(High)	0.298905	0.193271	0.062598	0.209424	1.083579	0.004828	1.020432
2	(High)	(Sustainable)	0.193271	0.539124	0.106416	0.550607	1.021301	0.002219	1.025554
3	(Sustainable)	(High)	0.539124	0.193271	0.106416	0.197388	1.021301	0.002219	1.005129
4	(Not sustainable)	(Very high)	0.298905	0.348983	0.224570	0.751309	2.152854	0.120257	2.617774
5	(Very high)	(Not sustainable)	0.348983	0.298905	0.224570	0.643498	2.152854	0.120257	1.966595
6	(Quite sustainable)	(Very high)	0.161972	0.348983	0.113459	0.700483	2.007214	0.056933	2.173557
7	(Very high)	(Quite sustainable)	0.348983	0.161972	0.113459	0.325112	2.007214	0.056933	1.241729
8	(Medium)	(Sustainable)	0.240219	0.539124	0.208138	0.866450	1.607144	0.078630	3.450952
9	(Sustainable)	(Medium)	0.539124	0.240219	0.208138	0.386067	1.607144	0.078630	1.237563
10	(Sustainable)	(Low)	0.539124	0.158059	0.158059	0.293179	1.854862	0.072846	1.191164
11	(Low)	(Sustainable)	0.158059	0.539124	0.158059	1.000000	1.854862	0.072846	inf

Figure 30. Association rules for country's CO_2 emissions per capita and HDI.

What can be observed is that if a country has low HDI it has a 85% increase in expectation that it will have a sustainable level of CO_2 emissions per capita. This is true in about 16% of the cases (as support indicates). Interesting fact is that there is 100% confidence that if a country has low HDI, its emissions are sustainable (but not the opposite!). Moreover, if it has medium HDI, it is 61% more likely to be sustainable as well. It happens in about 21% of the cases. Countries with high HDI have only around 2% of the expectation to have sustainable emissions. It is surprising, because one can say that more developed countries will have lower emissions (thus be sustainable) based on their policies, citizen awareness et cetera but it is only in 10% of the cases. It is actually the opposite. Very high HDI shows almost a 100% likelihood that these countries will have higher emissions per capita and it is supported by 64-75% of confidence. Also the support indicates that it happens in 22% of the cases. Similarly, high HDI is 84% more likely to have high emissios, with 20-32% confidence. However, it happens only in 6% of the cases.

- **aqi (aqi.bucket) and hdi (hdi.level)**

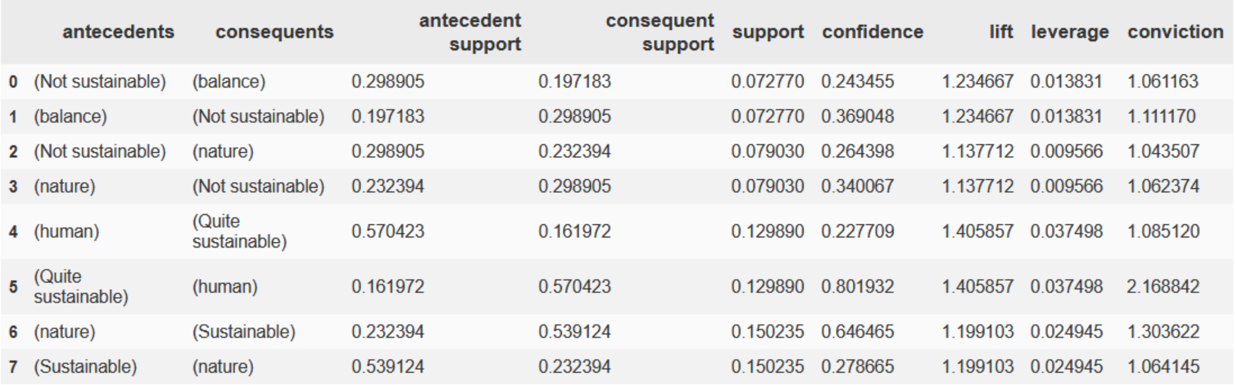


	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(High)	(Satisfactory)	0.193271	0.471049	0.106416	0.550607	1.168897	0.015376	1.177036
1	(Satisfactory)	(High)	0.471049	0.193271	0.106416	0.225914	1.168897	0.015376	1.042170
2	(Satisfactory)	(Very high)	0.471049	0.348983	0.219875	0.466777	1.337537	0.055487	1.220911
3	(Very high)	(Satisfactory)	0.348983	0.471049	0.219875	0.630045	1.337537	0.055487	1.429772
4	(Medium)	(Moderate)	0.240219	0.438185	0.121283	0.504886	1.152222	0.016023	1.134719
5	(Moderate)	(Medium)	0.438185	0.240219	0.121283	0.276786	1.152222	0.016023	1.050561
6	(Moderate)	(Low)	0.438185	0.158059	0.099374	0.226786	1.434813	0.030115	1.088884
7	(Low)	(Moderate)	0.158059	0.438185	0.099374	0.628713	1.434813	0.030115	1.513156

Figure 31. Association rules for a country's AQI buckets and HDI.

Countries with moderate AQI and low HDI happen in only 9% of the cases. However, around 43% of likelihood indicates that if a country has moderate AQI, it also has low HDI. There is a pretty high confidence of 63% indicating that the country's medium HDI will have moderate AQI. Also, moderate AQI happens in 12% of the cases while it has only 15% chances to happen. Countries with high and very high HDI have satisfactory AQI. Very high HDI and satisfactory AQI have 33% of expectations. It happens in 22% of the cases too. Only 10% of the cases show countries with high HDI and satisfactory AQI. They have 16% of chances to happen. In general, higher confidence is when HDI is antecedent and AQI - consequent.

- **land dominance (land.dom) and co2 per capita (sust.level.ghg)**

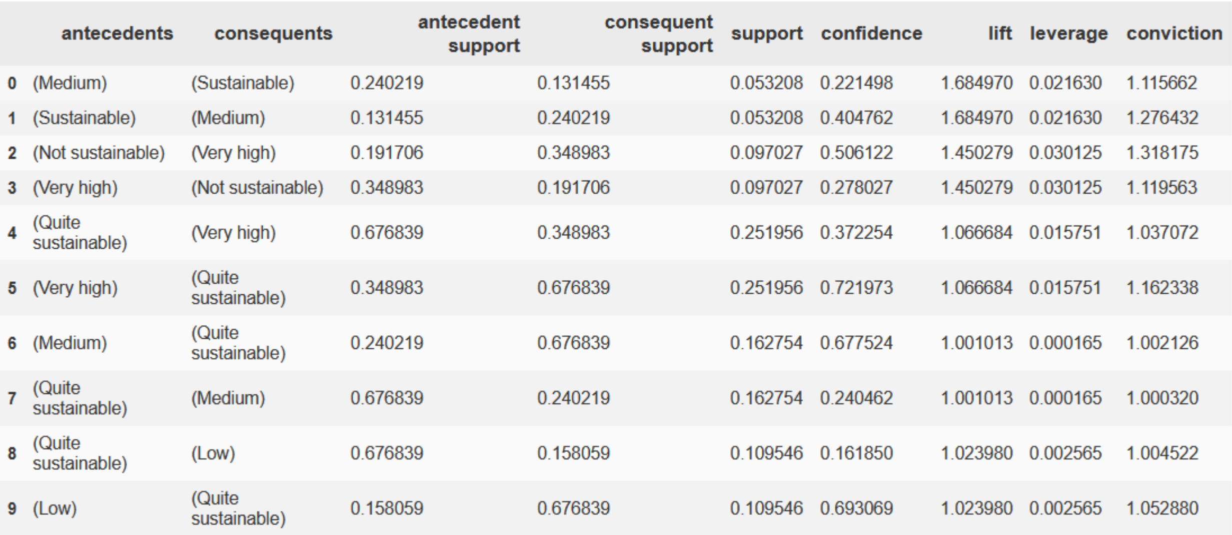


	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Not sustainable)	(balance)	0.298905	0.197183	0.072770	0.243455	1.234667	0.013831	1.061163
1	(balance)	(Not sustainable)	0.197183	0.298905	0.072770	0.369048	1.234667	0.013831	1.111170
2	(Not sustainable)	(nature)	0.298905	0.232394	0.079030	0.264398	1.137712	0.009566	1.043507
3	(nature)	(Not sustainable)	0.232394	0.298905	0.079030	0.340067	1.137712	0.009566	1.062374
4	(human)	(Quite sustainable)	0.570423	0.161972	0.129890	0.227709	1.405857	0.037498	1.085120
5	(Quite sustainable)	(human)	0.161972	0.570423	0.129890	0.801932	1.405857	0.037498	2.168842
6	(nature)	(Sustainable)	0.232394	0.539124	0.150235	0.646465	1.199103	0.024945	1.303622
7	(Sustainable)	(nature)	0.539124	0.232394	0.150235	0.278665	1.199103	0.024945	1.064145

Figure 32. Association rules for a country's CO_2 emissions per capita and land dominance.

Figure above indicates that a country with natural land dominance has 64% confidence to have sustainable CO_2 emissions (but not the opposite), however it is expected to be in 19%. Pair like this happens only in about 15% of the cases. What is surprising, is that there is a 34% likelihood that countries with natural land dominance will also have very high emissions. It would seem the other way around. It occurs in 8% of the cases and has a 14% chance of happening. There is 80% confidence that countries with high emissions are dominated by human-made land. It is expected to happen in 41%. Support says that 13% of the countries that have high CO_2 emissions also have land dominated by human-made. Balance between nature and human dominance occurs in 7% of the cases, whilst the likelihood that it happens is 37%. This itemset has a 23% increase in expectation.

- **country’s sustainability level (sust.level) and hdi (hdi.level)**



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Medium)	(Sustainable)	0.240219	0.131455	0.053208	0.221498	1.684970	0.021630	1.115662
1	(Sustainable)	(Medium)	0.131455	0.240219	0.053208	0.404762	1.684970	0.021630	1.276432
2	(Not sustainable)	(Very high)	0.191706	0.348983	0.097027	0.506122	1.450279	0.030125	1.318175
3	(Very high)	(Not sustainable)	0.348983	0.191706	0.097027	0.278027	1.450279	0.030125	1.119563
4	(Quite sustainable)	(Very high)	0.676839	0.348983	0.251956	0.372254	1.066684	0.015751	1.037072
5	(Very high)	(Quite sustainable)	0.348983	0.676839	0.251956	0.721973	1.066684	0.015751	1.162338
6	(Medium)	(Quite sustainable)	0.240219	0.676839	0.162754	0.677524	1.001013	0.000165	1.002126
7	(Quite sustainable)	(Medium)	0.676839	0.240219	0.162754	0.240462	1.001013	0.000165	1.000320
8	(Quite sustainable)	(Low)	0.676839	0.158059	0.109546	0.161850	1.023980	0.002565	1.004522
9	(Low)	(Quite sustainable)	0.158059	0.676839	0.109546	0.693069	1.023980	0.002565	1.052880

Figure 33. Association rules for a country's sustainability level and HDI.

It can be observed that if a country has low HDI it has almost no existing expectations that it will be quite sustainable. However, it has a 69% likelihood to happen and it happens in 11% of the cases. What is also interesting, that medium and very high HDI have also high confidence - 68% and 72% respectively. Both of them have almost no expectations that they will be quite sustainable. Nevertheless, they occur in 16% and 25% of the cases, respectively. Quite surprising is the fact that countries with very high HDI are not sustainable, as it is expected to happen in 45%. The confidence that it will happen is 27% or 50%. The highest probability, of 68%, has the occurrence of sustainable countries with medium HDI. Countries that are sustainable are 40% confident to have medium HDI. This happens in 5% of the cases.

5.3.3. Predictive models study

It was decided that the variable CO_2 would be a good choice for predicting its value, as it is a variable creating a high low pollution scenario.

In addition, it was decided to try a linear model, because of the rather linear relationship between CO_2 emissions and year, percentage of human-made land, HDI and AQI.

5.3.3.1. Multiple Linear Regression model

CO_2 prediction for 2050 was based on 4 variables: year, Human Development Index, Air Quality Index and the percentage of human-made land dominance. After calculating the average value of each variable over the years, the following results were obtained:

- Human Development Index - 0.673;
- Air Quality Index - 123;
- percentage of human-made land dominance - 44.

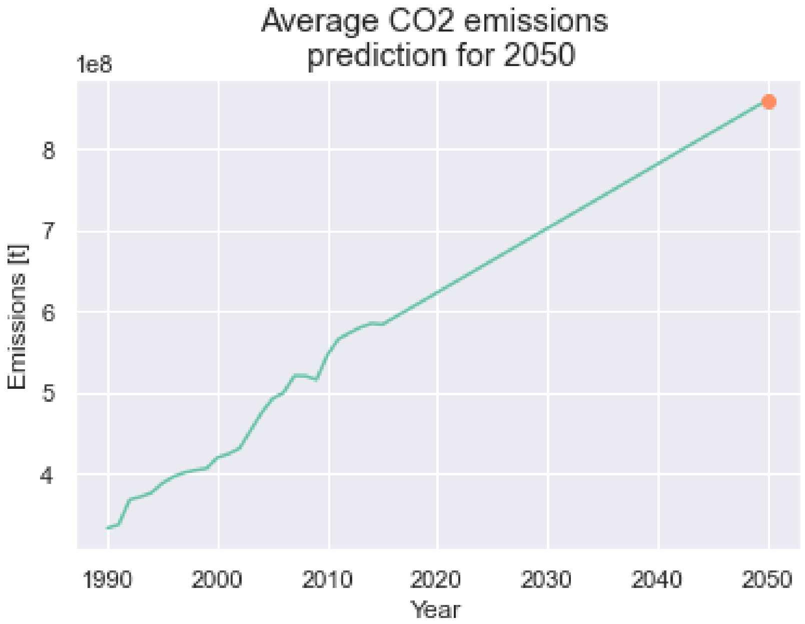


Figure 34. Prediction of average CO_2 emissions in 2050..

It can be observed that in 2050, the average CO_2 emissions in the world will be 1,5 times higher than in 2015. In 2015, the average CO_2 emissions were 5,8 billion tons. Based on the calculations, this value in 2050 will achieve almost 9 billion tons of CO_2 emissions.

This prediction is a great example of climate and environmental changes over the world. Emissions of CO_2 are the most important factor when it comes to evaluating these changes. As mentioned in the introduction, CO_2 emissions have the highest levels ever than at any point in the past 2000 or even more years. Unfortunately, it does not seem that it is going to improve. On the contrary, the emissions will be higher and higher each year.

6. Analysis of results

6.1. Discussion

The analysis of the different countries and variables helped to have an insight on the climate and environmental changes that are happening in the world and it has generated various curious results. Results obtained had proven that these changes are very visible and progressive. The changes in overall trends are not so drastic, but they are there. Although when looking at trends of each country separately, one can be shocked how big the actual changes are.

Performance of exploratory data analysis helped to uncover interesting trends in overall analysis. Average Human Development Index increased over the years as well as the number of countries with high and very high HDIs. Air pollution, specifically the Air Quality Index, improved significantly. From countries that had 5 out of 6 levels of AQI in 1990 (bad and good ones), the change showed only 2 levels (satisfactory and moderate, good ones) in 2015. Unfortunately, none of the countries at any point of the 26 years had good AQI. When it comes to temperature change, it is clearly visible that minimum and maximum temperatures have risen over the years. It is a bad sign as it indicates that global warming is more and more evident. Also, the total precipitation rose. Climate change patterns took into account only temperatures and unfortunately, in 2015 only one country had lower temperatures than before. It was Thailand. More than 30 countries showed a rise in the temperatures and the rest of them were quite steady (temperatures did not change much or at all). General CO_2 emissions have risen but CO_2 emissions per capita slightly decreased. Anyway, it cannot be the reason to be satisfied, because later forecasting showed that the emissions will be way higher in 2050. Other greenhouse gases emissions also have risen but emissions per capita slightly decreased. It may be due to increasing human-made land percentage and decreasing the natural one.

Clustering helped to identify which countries in which years belonged to which cluster. The vast majority belonged to sustainable cluster. These countries varied from developed, to developing and least developed ones. Quite sustainable and not sustainable clusters were definitely the small ones, with only 6 countries (4 and 2, respectively). Those contained developed and developing countries. It is not surprising that countries from not sustainable clusters are two of the biggest economies in the world - the United States of America and China. It can be suspected that they are responsible for the majority of negative climate and environmental changes. What is surprising, that only 3 countries, over the time span of 26 years, changed clusters. However, this change was not positive. India and Japan moved from being sustainable ones to being quite sustainable, while China from quite sustainable, became not sustainable.

Association rules uncovered some interesting relationships between variables. For example, countries with very high HDI are most probably the ones to be quite or not sustainable. Also, medium HDI showed high confidence for the countries to be quite sustainable or sustainable. The most surprising was the fact that countries with low HDI are the ones that are quite sustainable. In all the cases it was thought to be the opposite. The higher the country's HDI, the higher the supposed awareness and ability to implement new policies, thus the higher sustainability level. However, this analysis refuted this.

As mentioned, Multiple Linear Regression was used to perform forecasting of average CO_2 emissions in 2050. Disturbing is the fact that in 2050 these emissions will be around 1.5 times higher than they were in 2015.

6.2. Limitations

During the process of the research, there were many limitations faced. First and most important, was the availability of the data, especially for earlier years than only 1990. A lot of variables had data available from 1970 onwards but also a lot of them only from 1990.

What is more, even if the final data contained information from the years 1990-2015, some of the years were still missing. To avoid missing values, interpolation was performed. However, this could not be treated as a true reflection of the overall trends.

Another critical limitation was that there were some assumptions that had to be made on the way. For example, when it comes to urbanized land, the data was available only for a few years. Performed interpolation could only assume how these values were changing every year, but it is not the real picture. Another important assumption made was that developed countries should be the most sustainable ones et cetera, but later in the thesis it was refuted. Therefore, the final results have to be treated with a distance as they do not represent 100% reality.

Lack of expert knowledge in the field, a lot of assumptions and not enough data available were considerable limitations that resulted in not 100% real results.

7. Conclusions

The dissertation was carried out by collecting and analyzing the information about the climate and environmental changes in 50 different countries in the time span of 26 years. This study used actual data collected from various websites and covered topics of air pollution, climate change, greenhouse gases and land cover. Analysis performed on this data aimed to confirm the main hypothesis that climate and environmental changes are real. Based on the results received, it can be concluded that these changes are very visible and disturbingly progressive.

During the project, some thoughts were made about expected outcomes. It was believed from the beginning that this thesis will prove the established hypothesis. However, some of the results were surprising. The fact that most unsustainable countries are the ones most developed and the ones least developed are in fact sustainable was not the answer that was expected. What is more, no such drastic changes were expected when analyzing each country separately.

This climate and environmental change analysis should help to better understand the impact that climate change is going to have (or already having), and what actions countries may need to take to adapt to a changing environment. It also should open the eyes of everyone who thinks that climate and environmental changes are not real. Even though the analysis was performed for 50 randomly selected countries, it is sure that any country that would be selected would show bigger or smaller changes. Of course, there are countries whose practises are incredibly effective but to change something, more countries have to be like this.

As a final thought, it is worth mentioning that even if these results do not reflect the reality in 100%, they give a serious insight on how accelerating the changes are. There are a lot of things to think about and there is still a lot of room for improvement from both countries as well as individuals' sides.

References

1. Al-Obaidi, A. and NguyenHuynh, T., 2018. Renewable vs. conventional energy: which wins the race to sustainable development?. IOP Conference Series: Materials Science and Engineering, 434, p.012310.
2. Analytics Vidhya. 2021. KNN Algorithm | What is KNN Algorithm | How does KNN Function. [online] Available at: <https://www.analyticsvidhya.com/blog/2021/04/simple-understanding-and-implementation-of-knn-algorithm/> [Accessed 1 September 2021].
3. bp global. 2021. Statistical Review of World Energy | Energy economics | Home. [online] Available at: <https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html> [Accessed 16 May 2021].
4. Climate Change: Vital Signs of the Planet. 2021. Climate Change: Vital Signs of the Planet. [online] Available at: <https://climate.nasa.gov/> [Accessed 16 May 2021].
5. Crudata.uea.ac.uk. 2021. High-resolution gridded datasets. [online] Available at: <https://crudata.uea.ac.uk/cru/data/hrg/> [Accessed 16 May 2021].
6. Data.ceda.ac.uk. 2021. [online] Available at: <https://data.ceda.ac.uk/badc/cru/data/cru_cy/cru_cy_4.03/data> [Accessed 16 May 2021].
7. Data.worldbank.org. 2021. Population, total | Data. [online] Available at: <https://data.worldbank.org/indicator/SP.POP.TOTL> [Accessed 26 May 2021].
8. Databank.worldbank.org. 2021. WDI Database Archives (beta) | DataBank. [online] Available at: <https://databank.worldbank.org/reports.aspx?source=1277&series=EN.ATM.PM10.MC.M3> [Accessed 16 May 2021].
9. Docs.scipy.org. 2021. Interpolation (scipy.interpolate) — SciPy v1.7.1 Manual. [online] Available at: <https://docs.scipy.org/doc/scipy/reference/tutorial/interpolate.html?highlight=scipy%20interpolate%20bisplrep> [Accessed 31 August 2021].
10. Earthobservatory.nasa.gov. 2021. Water Vapor. [online] Available at: <https://earthobservatory.nasa.gov/global-maps/MYDAL2_M_SKY_WV> [Accessed 20 May 2021].
11. Edgar.jrc.ec.europa.eu. 2021. EDGAR - The Emissions Database for Global Atmospheric Research. [online] Available at: <https://edgar.jrc.ec.europa.eu/country_profile> [Accessed 16 May 2021].
12. Eia.gov. 2021. International - U.S. Energy Information Administration (EIA). [online] Available at: <https://www.eia.gov/international/data/world> [Accessed 19 May 2021].

13. En.wikipedia.org. 2021. Air pollution - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/Air_pollution> [Accessed 16 May 2021].
14. En.wikipedia.org. 2021. Apriori algorithm - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/Apriori_algorithm> [Accessed 25 August 2021]
15. En.wikipedia.org. 2021. *Elbow method (clustering) - Wikipedia*. [online] Available at: <https://en.wikipedia.org/wiki/Elbow_method_(clustering)> [Accessed 15 September 2021].
16. En.wikipedia.org. 2021. Exploratory data analysis - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/Exploratory_data_analysis> [Accessed 31 August 2021].
17. En.wikipedia.org. 2021. Human Development Index - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/Human_Development_Index> [Accessed 16 May 2021].
18. En.wikipedia.org. 2021. John Tukey - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/John_Tukey> [Accessed 31 August 2021].
19. En.wikipedia.org. 2021. *k-means clustering - Wikipedia*. [online] Available at: <https://en.wikipedia.org/wiki/K-means_clustering> [Accessed 10 September 2021].
20. En.wikipedia.org. 2021. k-nearest neighbors algorithm - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm> [Accessed 1 September 2021].
21. En.wikipedia.org. 2021. Linear regression - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/Linear_regression#History> [Accessed 31 August 2021].
22. En.wikipedia.org. 2021. Pearson correlation coefficient - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/Pearson_correlation_coefficient> [Accessed 4 September 2021].
23. En.wikipedia.org. 2021. Precipitation - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/Precipitation> [Accessed 16 May 2021].
24. En.wikipedia.org. 2021. Random forest - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/Random_forest> [Accessed 31 August 2021].
25. En.wikipedia.org. 2021. Temperature - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/Temperature> [Accessed 16 May 2021].
26. Encyclopedia Britannica. 2021. natural gas | Definition, Discovery, Reserves, & Facts. [online] Available at: <https://www.britannica.com/science/natural-gas> [Accessed 17 May 2021].
27. Engineering Education (EngEd) Program | Section. 2021. Introduction to Random Forest in Machine Learning. [online] Available at: <https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/> [Accessed 31 August 2021].
28. Fao.org. 2021. FAOSTAT. [online] Available at: <http://www.fao.org/faostat/en/#data/RL/visualize> [Accessed 16 May 2021].
29. Fly spaceships with your mind. 2021. *K-Means: One Of The Simplest Clustering Algorithms – Fly Spaceships With Your Mind*. [online] Available at: <https://starship-knowledge.com/k-means> [Accessed 10 September 2021].

30. Ghsl.jrc.ec.europa.eu. 2021. Global Human Settlement - Degree of urbanisation - European Commission. [online] Available at: <https://ghsl.jrc.ec.europa.eu/CFS.php> [Accessed 16 May 2021].
31. Greenfacts.org. 2021. Glossary: Inland waters. [online] Available at: <https://www.greenfacts.org/glossary/ghi/inland-waters.htm> [Accessed 16 May 2021].
32. Hq.nasa.gov. 2021. Forest Land. [online] Available at: <https://www.hq.nasa.gov/iwgsdi/Forest_Land.html> [Accessed 16 May 2021].
33. Ibm.com. 2021. Supervised vs. Unsupervised Learning: What's the Difference?. [online] Available at: <https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning> [Accessed 26 July 2021].
34. Igual, L. and Seguí, S., 2017. Introduction to data science. A Python Approach to Concepts, Techniques and Applications.
35. Kaggle.com. 2021. Calculating AQI (Air Quality Index) Tutorial. [online] Available at: <https://www.kaggle.com/rohanrao/calculating-aqi-air-quality-index-tutorial> [Accessed 19 August 2021].
36. KNN Classification using Scikit-learn, Data Camp, 2021. [online] Available at: <https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn> [Accessed 1 September 2021].
37. Medium. 2021. A Brief Introduction to Unsupervised Learning. [online] Available at: <https://towardsdatascience.com/a-brief-introduction-to-unsupervised-learning-20db46445283> [Accessed 26 July 2021].
38. Medium. 2021. Correlation: straight to the point. [online] Available at: <https://medium.com/brdata/correlation-straight-to-the-point-e692ab601f4c> [Accessed 4 September 2021].
39. Medium. 2021. Introduction to Machine Learning Algorithms: Linear Regression. [online] Available at: <https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a> [Accessed 28 July 2021].
40. Medium. 2021. Understanding Random Forest. [online] Available at: <https://towardsdatascience.com/understanding-random-forest-58381e0602d2> [Accessed 31 August 2021].
41. Medium. 2021. Understanding the concept of Hierarchical clustering Technique. [online] Available at: <https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec> [Accessed 18 August 2021]
42. Microsoft Corporation. (2021). Microsoft Excel (No. 2019). [online] Available at: <https://office.microsoft.com/excel> [Accessed 16 May 2021].

43. Nations Online, 2021. [online] Available at: <https://www.nationsonline.org/oneworld/human_development.htm> [Accessed 26 August 2021].
44. Neumayer, E., 2010. Human Development and Sustainability. Human Development Research Paper 2010/05.
45. Ng, A., 2021. Association Rules and the Apriori Algorithm: A Tutorial - KDnuggets. [online] KDnuggets. Available at: <https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html> [Accessed 25 August 2021]
46. Python.org. 2021. Welcome to Python.org. [online] Available at: <https://www.python.org/> [Accessed 19 May 2021].
47. Shaftel, H., 2021. Overview: Weather, Global Warming and Climate Change. [online] Climate Change: Vital Signs of the Planet. Available at: <https://climate.nasa.gov/resources/global-warming-vs-climate-change/> [Accessed 16 May 2021].
48. Socalgas.com. 2021. Methane and the Environment | SoCalGas. [online] Available at: <https://www.socalgas.com/stay-safe/methane-emissions/methane-and-the-environment> [Accessed 16 May 2021].
49. Society, N., 2021. non-renewable energy. [online] National Geographic Society. Available at: <https://www.nationalgeographic.org/encyclopedia/non-renewable-energy/> [Accessed 17 May 2021].
50. Society, N., 2021. urban area. [online] National Geographic Society. Available at: <https://www.nationalgeographic.org/encyclopedia/urban-area/> [Accessed 16 May 2021].
51. Stateofglobalair.org. 2021. Explore the Data | State of Global Air. [online] Available at: <https://www.stateofglobalair.org/data/#/air/tabl> [Accessed 16 May 2021].
52. Statistics Globe. 2021. Regression Imputation (Stochastic vs. Deterministic & R Example). [online] Available at: <https://statisticsglobe.com/regression-imputation-stochastic-vs-deterministic/> [Accessed 28 July 2021].
53. theOECD. 2021. Air and climate - Air pollution exposure - OECD Data. [online] Available at: <https://data.oecd.org/air/air-pollution-exposure.htm> [Accessed 16 May 2021].
54. Thompson, R., Lassaletta, L., Patra, P., Wilson, C., Wells, K., Gressent, A., Koffi, E., Chipperfield, M., Winiwarter, W., Davidson, E., Tian, H. and Canadell, J., 2019. Acceleration of global N2O emissions seen from two decades of atmospheric inversion. Nature Climate Change, 9(12), pp.993-998.
55. Topics, H., 2021. Air Pollution: MedlinePlus. [online] Medlineplus.gov. Available at: <https://medlineplus.gov/airpollution.html> [Accessed 16 May 2021].

56. Twi-global.com. 2021. What is Green Energy? (Definition, Types and Examples). [online] Available at: <https://www.twi-global.com/technical-knowledge/faqs/what-is-green-energy> [Accessed 16 May 2021].
57. US EPA. 2021. Overview of Greenhouse Gases | US EPA. [online] Available at: <https://www.epa.gov/ghgemissions/overview-greenhouse-gases> [Accessed 16 May 2021].
58. US EPA. 2021. Particulate Matter (PM) Basics | US EPA. [online] Available at: <https://www.epa.gov/pm-pollution/particulate-matter-pm-basics> [Accessed 16 May 2021].
59. Who.int. 2021. WHO | Global environmental change. [online] Available at: <https://www.who.int/globalchange/environment/en/> [Accessed 16 May 2021].
60. World Air Quality Index Project, 2021. Ozone AQI: Using concentrations in milligrams or ppb?. [online] aqicn.org. Available at: <https://aqicn.org/faq/2015-09-06/ozone-aqi-using-concentrations-in-milligrams-or-ppb/> [Accessed 16 May 2021].

Annex

Whole code and all datasets can be found in the github repository below. Please read README.md first in order to know which file is about what.

https://github.com/adaczerwinska/TFM-Adrianna-Czerwinska