Multivariate Exploratory Analysis of data on CO2 and Greenhouse Gas Emissions

Diogo Pereira
114903
diogofop@ua.pt
Master in Data Science
Department of Mathematics
University of Aveiro

Beatriz Mesquita
115367
beatrizs.mesquita@ua.pt
Master in Data Science
Department of Mathematics
University of Aveiro

Nuno Gomes
114904
nunoteixeirag@ua.pt
Master in Data Science
Department of Mathematics
University of Aveiro

***Abstract*—This project analyzes the carbon dioxide emissions data [3] for a list of European countries using three methods: principal component analysis (PCA), factor analysis (FA), and clustering. The goal was to identify patterns or underlying factors in the data and to group the countries into clusters based on their similarity or distance from each other. PCA was used to identify the most important patterns in the data and to reduce the number of dimensions. FA was performed via PCA and Maximum Likelihood, and the resulting factor scores via PCA were used as input for the clustering analysis. A comparison was made of the data clustering (hierarchical and non-hierarchical) with two real clusters, the clusters of the factor' scores resulting from the FA and also the population cluster that was hand categorized by us.**

The results of this analysis provide insights into the structure of the carbon dioxide and greenhouse gas emissions data and can be used to better understand the relationships between the various emission variables and the countries in the dataset, which would not be possible to perceive otherwise.

***Index Terms*—carbon dioxide emissions, PCA, FA, Maximum Likelihood, clustering, K-means clustering, hierarchical clustering, factor scores**

I. Introduction

Nowadays, due to climate change, the theme of CO2 and greenhouse has emissions has been the scene of enumerations, analyses, and research around the world. [14] Carbon emission is an expression that refers to the release of greenhouse gases into the atmosphere mainly CO2(Carbon Dioxide), a gaseous chemical that causes serious imbalances in the greenhouse effect of the planet.

CO2 emissions mainly come from the burning of fossil fuels, deforestation, and land use change, among other sources. There are a range of consequences derived from CO2 emissions between melting glaciers, increased average land temperature, floods, and migration of species, among many others. [**?**]

Although these emissions are very harmful, there are some countries that find it hard to install more responsible environmental policies on the planet.

Thus, we chose to study about CO2 and greenhouse gas emissions through a dataset where the features allow the description of the problem under analysis, comparing between different European countries, "*CO2 and Greenhouse Gas Emissions*" by *Our World in Data* [3].

We begin by wanting to identify patterns or underlying factors in the carbon dioxide emissions data for the countries in the dataset using principal component analysis (PCA). So our first research question is: "**Identify patterns or underlying factors in the carbon dioxide emissions data for the variables in the dataset**". Since PCA is a dimensionality reduction technique we can use it to identify patterns and structure in a dataset by finding the directions in which the data varies the most. We then apply factor analysis via PCA and via maximum likelihood, and compare the results. We then defined the work question: "**What patterns or underlying factors can be identified in the carbon dioxide emissions data for these countries?**".

From the FA factor' scores, a non-hierarchical cluster analysis was performed. From this, we then created a "real cluster" that will allow us to make a comparison with the clustering of our original dataset and thus an evaluation. Another comparison made was with the clustering of the population column only. This was done by hand by us by sorting the population number. So our work question for this part is going to be "**What is the best comparison in order to get a good evaluation of the clustering of our data?**"

Thus, through these multivariate techniques, we are able to draw conclusions that would not otherwise be possible, which are the answers to the research questions mentioned.

II. Data Analysis and Preparation

In the preparation of this work, we use a database available on the website *Our World in Data* [3]. The dataset provided contained seventy-four features, however in the preparation of this project, we used only the most important features for our analysis. The study under analysis is dedicated to comparing the results of European countries, so we started the research with forty-four countries.

From the total features we decided to stay with:

- **cement_co2:** Annual production-based emissions CO2 from cement, measured in million tonnes;
- **CO2:** Annual total production-based emissions of carbon dioxide, measured in million tonnes;

- **coal_co2:** Annual production-based emissions of CO2 from coal, measured in million tonnes;
- **flaring_co2:** Annual production-based emissions of carbon dioxide from flaring, measured in million tonnes;
- **gas_co2:** Annual production-based emissions of CO2 from gas, measured in million tonnes;
- **land_use_change_co2:** Annual production-based emissions of carbon dioxide from land-use change, measured in million tonnes. This is based on territorial emissions;
- **methane:** Total methane emissions including land-use change and forestry, measured in million tonnes of carbon;
- **nitrous_oxide:** Total nitrous oxide emissions including land-use change and forestry, measured in million tonnes of carbon;
- **oil_co2:** Annual production-based emissions of carbon dioxide from oil, measured in million tonnes;
- **other_industry_co2:** Annual production-based emissions of CO2 from other industry sources, measured in million tonnes;
- **consuption_co2:** Annual consumption-based emissions of carbon dioxide, measured in million tonnes. Consumption-based emissions are national;
- **trade_co2:** Annual net co2 emissions embedded in trade, measured in million tonnes. Net CO2 emissions embedded in trade is the net of CO2 that is imported or exported via traded goods with an economy.

We represent each of the countries by their ISO codes, for more practical analysis when applying the visualizations. In addition, when we applied PCA we created several sub-datasets where we removed countries in order to refine/improve the performance of this technique.

For the FA application, we use the sub-dataset that we chose for the previous PCA application.

Finally, to perform cluster analysis, we created a new column that had the objective of classifying the population of each country, based on the "`population`" column we had in the raw dataset. This way, the new "`pop_cluster`" column is classified as 0 or 1. To define the classification we took the average of the population of all the thirty-three countries of the sub-dataset, if the population of a given country is greater than this average, then it is classified as 1, otherwise we classify it as 0. As said in the Introduction, this classification will be used as "real clusters" to compare with the clustering of our data.

III. Methodology

All three of these methods - principal component analysis (PCA), factor analysis (FA), and clustering - are statistical techniques that are commonly used to analyze and understand patterns and trends in multivariate data.

These methods are each based on different assumptions and use different approaches to achieve this goal. They all involve some form of data transformation, either through the use of principal components, factors, or clusters.

Finally, all of these methods are widely used in a variety of applications and are often applied to data in fields such as social science, biology, finance, and marketing, among others. They can be especially useful for analyzing large and complex datasets, where it can be challenging to understand the relationships and patterns in the data using more traditional methods of analysis.

A. Principal Component Analysis (PCA)

PCA is characterized by identifying the dimensions along which the data is more dispersed. In this way, we were able to identify the dimensions that best differentiate the dataset under analysis, that is, its principal components. These are chosen in such a way as to maximize the variance of the data while minimizing the correlations between the components.

When the patterns in the set are identified, the number of dimensions to be analyzed can be reduced without significant loss of information, as the focus is on the analysis of the main dimensions that characterize the data set.

The mathematical foundation of PCA is based on linear algebra and eigenvalue decomposition. The principal components are found by performing an eigenvalue decomposition of the covariance matrix of the data. Next we mention the steps followed to perform PCA.

1) Normalizing Data: The first step in running the PCA consists of normalizing the data, that is, standardizing a numeric dataset by centering the data at 0 and scaling it to have a standard deviation of 1. Mathematically, this is done by subtracting the mean of each column of the dataset from each value in the column and then dividing the result by the standard deviation of the column. Thus, it is given by the expression:

$$normalizeddata = \frac{\mathbf{X} - \overline{\mathbf{X}}}{\sigma(\mathbf{X})} \tag{1}$$

where $\mathbf{X}$ is a vector of values, $\sigma(X)$ is the standard deviation, and $\overline{X}$ is the mean of $\mathbf{X}$, given by:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{2}$$

where X_i are the individual values of a vector $\mathbf{X}$ and n is the total number of values per column.

Normalizing the data is often useful in statistical analysis because it allows for comparison between variables that may have different scales.

2) Covariance Matrix: Understanding the variance and covariance of the data is important in PCA, as it helps to interpret the results of the analysis and identify patterns and relationships in the data.

Variance is a measure of how spread out the data is. It is calculated as the average of the squared differences between each data point and the mean of the data. A high variance indicates that the data points are far from the mean, while a low variance indicates that the data points are close to the mean. In other words, variance indicates how dispersed the data of a dimension is in relation to the average, "ignoring"

the existence of other dimensions. It can be calculated by the expression:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{(n-1)} \tag{3}$$

The principal components are ranked by the amount of variance they explain, and the first few principal components are typically used to represent the data.

Covariance, in turn, applies to two dimensions as it is a measure of the relationship between two variables, allowing us to understand how both are related to each other. It is calculated as the average of the product of the differences between each data point and the mean of the data for each variable. It can be calculated by the expression:

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)} \tag{4}$$

This measure indicates how much the dimensions vary in relation to the average, taking into account the relationship that exists between them. When calculating the covariance between two dimensions, we can deduce how they relate to each other by performing the following analysis:

- $cov(X,Y) > 0$: If the covariance between the X and Y dimensions is positive, both dimensions are directly related to each other, that is, if the X values increase, the Y values also increase.
- $cov(X,Y) < 0$: If the covariance between the X and Y dimensions is negative, both dimensions are inversely related to each other, that is, if the X values increase, the Y values decrease.
- $cov(X,Y) = 0$: If the covariance between the X and Y dimensions is null, it is verified that the dimensions are independent of each other, as such the behavior of the X dimension does not influence the behavior of the Y dimension.

The covariance between the variables is also important, as it determines the degree to which the variables are related. If the variables are highly correlated, they will tend to vary together and will be captured by the same principal components. If the variables are not strongly correlated, they will be captured by different principal components.

3) Eigenvectors and eigenvalues: Eigenvectors and eigenvalues represent the main characteristics of a matrix.

In PCA, these are used to identify the directions in which the data varies the most. These directions are called principal components, and they are used to project the data onto a lower-dimensional space.

The eigenvectors of a matrix are defined as the non-zero vectors that, when multiplied by the matrix, resulting in a scalar multiple of the original vector. In other words, if A is a matrix and $\mathbf{x}$ is an eigenvector of A, then:

$$A\mathbf{x} = \lambda\mathbf{x} \tag{5}$$

,

where λ is a scalar value known as the eigenvalue.

The eigenvectors are characterized by the following properties:

- Only exist in square matrices (NxN);
- In an NxN matrix there are N eigenvectors;
- Are orthogonal to each other;
- They are calculated so that their length is unitary.

Eigenvalues are the scalar values that correspond to each eigenvector, and they represent the amount of variance in the data that is captured by the corresponding principal component. The eigenvectors with the highest eigenvalues are the most important ones because they capture the most variance in the data.

4) Principal Components calculation: PCA is defined as an orthogonal linear transformation that transforms the data into a new coordinate system such that the greatest variation by some scalar projection of the data comes to reside in the first principal component, the second greatest variation at the second principal component, and so on. In order to obtain the first principal component, our objective is to maximize the variance, considering a matrix X, $n \times p$ with empirical mean per column equal to zero, where each row n represents a repetition of a given test or analysis and each column p, a particular feature of the test or analysis performed. Mathematically, the transformation is done through the square of the multiplication of each existing line in the matrix **X**, by a unit vector of weights, **W**.

With this we maximize the variance through the following formula:

$$w_{(1)} = argmax\left\{\frac{w^T X^T X w}{w^T w}\right\} \tag{6}$$

5) PCA visualization: We start by creating a **cumulative scree plot**. It is a visual tool that helps to determine the number of principal components (PCs) to retain in PCA. It plots the percentage of explained variance on the y-axis against the number of PCs on the x-axis. The explained variance is the amount of variance in the data that is explained by each PC. The plot shows the cumulative explained variance as the number of PCs increases.

The point at which the plot levels off or "breaks" is known as the "elbow" of the plot. It is generally recommended to retain only the principal components up to the elbow in the plot, as the additional components do not add much additional information to the analysis.

One key difference between the classical scree plot and the cumulative scree plot is that the last shows the accumulated explained variance of all the principal components up to a certain point, while the classic scree plot shows the explained variance of each individual principal component. This means that the cumulative scree plot can give a better idea of how much of the total variance in the data is being explained by the principal components, while the classic scree plot can give a better understanding of the individual contributions of each principal component to the explained variance.

After that, we also apply a **scatterplot** and a **biplot**. These are used to visualize the results of the PCA. The scatterplot shows the points in the two-dimensional space defined by the first and second principal components, while the biplot shows both the points and the original variables in this space. The x-axis and y-axis of both plots represent the first and second principal components, respectively. The points in the scatterplot represent the observations, and the labels associated with each point are the row names of the data. In the biplot, the variables are represented by vectors, and the length of the vectors is proportional to the relative importance of the variable in explaining the variance in the data. To interpret a biplot, we can look at the position of the observations relative to each other and to the variable vectors. Observations that are close to each other are more similar to each other than observations that are far apart. We can also look at the angles between the variable vectors to understand how the variables are related to each other.

The scatterplot and biplot can be used to identify patterns and trends in the data, and to understand how the variables are related to the principal components.

To finish, we compute the correlation between the variables and the PCs. This is a measure of how much each original variable contributes to each principal component. This can help to understand which variables are driving the variance in the data and how they are related to the principal components.

6) Advantages and disadvantages to applying PCA: Advantages:

- It can reduce the complexity of a dataset by projecting the data onto a lower-dimensional space, making it easier to visualize and understand;
- It can identify patterns in the data that may not be apparent when looking at the raw data;
- It can help to identify correlated variables and remove redundancy in the data;
- It can be used to improve the accuracy of supervised learning models by reducing the number of dimensions in the data.

Disadvantages:

- It can be sensitive to the scaling of the data, so it is important to scale the data before applying PCA;
- It can be difficult to interpret the results, especially when working with high-dimensional data;
- It can potentially remove important information from the data by projecting it onto a lower-dimensional space;
- It is not suitable for data that are not continuous or that have a large number of missing values.

[7]

B. Factor Analysis (FA)

Factor analysis is based on the idea that the observed variables are correlated because they are measuring the same underlying latent variables or factors. The goal is typically to identify a smaller number of underlying factors that can explain the variance in the data, that is, to identify these latent variables and understand how they relate to the observed variables.

The theory behind factor analysis is based on the idea that the observed variables can be represented as a linear combination of the latent variables, plus some error. This can be written mathematically as:

$$Y = \Lambda F + E \tag{7}$$

where:

- **Y** is the matrix of observed variables;
- Λ is a matrix of factor loadings, which represents the weights of the latent variables on the observed variables;
- **F** is the matrix of latent variables or factors;
- **E** is the matrix of residuals or error term.

The goal of factor analysis is to estimate the matrices Λ and **F** from the observed data. This is typically done using maximum likelihood estimation, which involves finding the values of Λ and **F** that maximize the likelihood of the observed data. We can also do it via principal component analysis, as we will see.

The outcome of a factor analysis is usually a set of factors, or latent variables, that represent the common underlying structure in the data. These factors can be interpreted and used to understand the relationships between the variables in the data.

Once the matrices Λ and **F** have been estimated, the results can be interpreted to understand the relationships between the observed variables and the latent variables. The factor loadings represent the strength of the relationships between the observed variables and the latent variables, and the factors can be interpreted based on the pattern of the factor loadings. [16]

Next, we mention the steps followed to do factor analysis.

1) Compute the KMO (Kaiser-Meyer-Olkin): The KMO (Kaiser-Meyer-Olkin) measure is used to assess the suitability of data for factor analysis. It measures the proportion of variance among all the variables that can be explained by the common factors.

It is generally recommended to conduct the KMO analysis on raw data, rather than on normalized data. This is because the KMO statistic is based on the correlations between variables, and normalization can alter the relationships between those.

To calculate the KMO, we need to compute the partial correlation matrix for all pairs of variables in the dataset. Then, we calculate the KMO for each variable by taking the mean of the partial correlations for that variable. We can also calculate the global KMO by taking the mean of all the KMO values for the individual variables.

The global KMO value is a measure of the overall sampling adequacy of all the variables in the dataset. It ranges from 0 to 1, with values closer to 1 indicating a better fit for factor analysis.

Individual KMO values are calculated for each individual variable in the dataset and can range from 0 to 1 as well.

Values closer to 1 indicate that the variable has a higher degree of commonality with the other variables in the dataset and is therefore a better candidate for inclusion in the factor analysis.

Therefore, we select the variables/features that will go into the factor analysis based on the KMO results and make a comparison of this with the reference table provided by Kaiser (1970). [15]

2) Normalizing Data: Just as we did in PCA, we need to normalize our data, that is, standardize a numeric dataset by centering the data at 0 and scaling it to have a standard deviation of 1, just like we showed in equation 1.

3) Bartlett's test: In factor analysis, Bartlett's test is used to determine whether the correlations among a set of variables are significantly different from zero. This is known as the test of sphericity.

The test of sphericity is based on the assumption that the correlations among the variables are equal. If this assumption is true, then the correlations can be represented by a single factor, which is known as the common factor. If the assumption is not true, then there may be multiple factors that explain the correlations among the variables.

To perform Bartlett's test in factor analysis, we first need to calculate the sample correlations among the variables. We then use these correlations to construct a matrix of intercorrelations, which is known as the correlation matrix.

Next, we calculate the test statistic using the following formula:

$$X^2 = \frac{p(p-1)}{2} \sum_{i \neq j} \frac{r_{ij}^2}{1 - r_{ij}^2} \tag{8}$$

where p is the number of variables, r_{ij} is the sample correlation between variables i and j, and the sum is taken over all pairs of variables.

We then compare the test statistic to the chi-squared distribution with $p(p-1)/2$ degrees of freedom to determine the p-value. If the p-value is less than a predetermined significance level (e.g. 0.05), then we can reject the null hypothesis and conclude that the correlations among the variables are significantly different from zero. This suggests that there may be multiple factors influencing the variables, and further analysis is needed to identify these factors. [2]

4) Number of factors: We return information about the number of components/factors to retain with the Kaiser rule and the parallel analysis.

Different solutions are given. The classical ones are the Kaiser rule, the parallel analysis, and the usual scree test. Nongraphical solutions to the Cattell subjective scree test are also proposed: an acceleration factor (af) and the optimal coordinates index oc. The acceleration factor indicates where the elbow of the scree plot appears. It corresponds to the acceleration of the curve, i.e. the second derivative. The optimal coordinates are the extrapolated coordinates of the previous eigenvalue that allow the observed eigenvalue to go beyond this extrapolation. [6]

The results of these four strategies can be used to guide the selection of the number of factors to retain in the factor analysis. We can choose the number of factors that is consistent across these four strategies.

To determine the number of factors or components to retain based on the plot of these four solutions, we can look for the point at which the eigenvalues from the original dataset start to level off or "scree" down. This is typically interpreted as the point beyond which the eigenvalues are not significantly larger than what would be expected by chance, and therefore are likely to be due to random noise or sampling error rather than reflecting a real underlying structure in the data. The number of factors or components to retain is often taken to be the number of eigenvalues that occur before the scree. [13]

5) Different methods for factor analysis: PCA and maximum likelihood estimation can be used together in the context of factor analysis. Our goal is to identify a small number of underlying factors that explain as much of the variance in the data as possible. We have already made a brief introduction to PCA in the subsection Principal Component Analysis (PCA).

Maximum likelihood is also often used to estimate the factor loadings, which are the correlations between the observed variables and the underlying factors. It is a statistical method for estimating the parameters of a statistical model that maximizes the likelihood of the data given by the model.

Overall, PCA and maximum likelihood are two statistical techniques that can be used to identify underlying dimensions or factors in a dataset and to estimate the relationships between the variables. If the structure of our output for the two methods is consistent, then we can assume that our factor model has been applied well. However, it is important to note that in order to be able to apply factor analysis via maximum likelihood, we first have to make sure that our variables/features follow a normal distribution.

For adjusting the factor loadings to make the results easier to interpret we can use rotation. It can be performed using a variety of different methods, each with its own characteristics and assumptions:

- **No rotation:** No rotation is the default option in many statistical software packages, and it means that the results of the maximum likelihood or PCA are not rotated. This can be useful when the goal is simply to identify the underlying dimensions or factors in the data, without making any assumptions about the relationships between the variables;
- **Varimax rotation:** Varimax rotation is a method of rotating the results of maximum likelihood or PCA to maximize the sum of the variances of the squared factor loadings. It is often used when the goal is to identify a small number of underlying factors that are each associated with a distinct set of variables;
- **Quartimax rotation:** Quartimax rotation is a method of rotating the results of maximum likelihood or PCA to maximize the sum of the fourth powers of the factor loadings. It is often used when the goal is to identify a

small number of underlying factors that explain as much of the variance in the data as possible;

- **Oblimin rotation:** Oblimin rotation is a method of rotating the results of maximum likelihood or PCA to maximize the sum of the squared correlations between the factors and the variables. It is often used when the goal is to identify a small number of underlying factors that are correlated with each other, and when the assumption of orthogonality between the factors is not reasonable.

Different rotation methods may yield different results and may be more or less appropriate depending on the characteristics of the data. This way, we try multiple rotation methods and consider the interpretability of the results to determine which method is most appropriate for the data at hand.

[8]

6) Percentage of cumulative explained variance: Similar to what we did in PCA we plot the cumulative explained variance. It is the sum of the explained variance for each factor in a FA. It is a measure of the total variance in the data that is explained by the factors.

This is used to determine how many factors should be retained in the analysis. In general, it is recommended to retain only those factors that explain a significant proportion of the variance in the data.

[18]

7) Residual matrix for the correlation matrix: In this part, we analyze the values from our residues matrix. If these are small it may indicate to us that the model fits the data well. The higher the proportion of values smaller than 0.05, for example, the better.

8) Estimate the scores of each factor: To finish our factor analysis, we estimate the scores of each factor. We later use this in subsequent cluster analysis. The main idea is to extract the most information from the multivariate data.

9) Advantages and disadvantages to applying FA: Advantages:

- Factor analysis can be used to reduce the complexity of the data by identifying a smaller number of underlying factors that explain the variance in the data. This can make it easier to understand and interpret the data;
- The factors identified by factor analysis are typically more interpretable than the original variables, as they represent underlying latent constructs that are related to the variables. This can make it easier to understand the meaning of the factors and how they relate to the variables;
- Factor analysis can be used to identify and reduce multicollinearity among variables, which can improve the stability and interpretability of statistical models.

Disadvantages:

- Factor analysis relies on certain assumptions about the data, such as the existence of latent factors and the normality of the variables. If these assumptions are not met, the results of the analysis may be biased or unreliable;
- The process of reducing the data to a small number of factors may result in the loss of some of the information in the original variables. This can be a disadvantage if the factors do not adequately capture all of the important information in the data;
- Factor analysis can be a complex technique, and it may require specialized software and expertise to implement. This can be a disadvantage if we do not have access to these resources.

[4] [19]

C. Clustering

Cluster analysis is a technique of multivariate statistics that allows identifying homogeneous groups of data, based on the variables. Having this in mind, a cluster is a grouping that classifies objects based on similarity between observations, such that the points within a cluster are more similar to each other than they are to points in other clusters. [10] It is a method of unsupervised learning, which means that the clusters are formed based on the inherent structure of the data and without any prior knowledge or labels.

In this work, we use two approaches to clustering our data points: hierarchical clustering and non-hierarchical clustering (more specifically, K-means).

Starting with the K-means, it is a non-hierarchical clustering algorithm that is used to partition a set of data points into k clusters, where k is a predetermined number. The algorithm works by iteratively assigning each data point to the cluster with the nearest centroid, where the centroid is the center of the cluster. The centroids are then updated to the mean of the points in the cluster, and the process is repeated until the clusters stabilize. [5] K-means clustering is an iterative algorithm, and it is sensitive to the initial placement of the centroids. To overcome this, it is common to run the algorithm multiple times with different initializations and choose the solution that gives the lowest sum of squared distances between the points and their assigned centroids.

K-means clustering has a number of advantages, including its simplicity and speed, which make it well-suited for large datasets. However, it has some limitations, including the assumption that the clusters are spherical and that the variance within each cluster is equal. It is also sensitive to the presence of outliers, as a single outlier can greatly influence the position of the centroids. [9]

The hierarchical clustering analysis provides more than one type of data grouping, generating several possible groupings. There are two main types of hierarchical clustering:

- Agglomerative: the data points are initially treated as individual clusters, and the clusters are merged together as the analysis progresses. This is done by calculating the distance between each pair of clusters and merging the two closest clusters. The process is repeated until all of the clusters are merged into a single cluster;
- Divisive: the data is initially treated as a single cluster, and the cluster is split into smaller clusters as the analysis progresses. This is done by calculating the distance between each data point and the centroid of the cluster and assigning the data point to the cluster with the nearest

centroid. The process is repeated until each data point is in its own cluster.

In this work, we use hierarchical clustering of the divisive type.

Some advantages of using hierarchical clustering are:

- Hierarchical clustering produces a tree-like diagram called a dendrogram, which shows the relationships between the clusters. This can make it easier to interpret the results of the analysis and understand the structure of the data;
- Hierarchical clustering can handle categorical variables, which may be difficult to incorporate into other clustering algorithms;
- With hierarchical clustering, we do not need to specify the number of clusters in advance. This can be useful if we are not sure how many clusters are present in the data.

[11] Some potential disadvantages of hierarchical clustering include:

- Hierarchical clustering can be more computationally intensive than other clustering algorithms, especially for large datasets;
- Hierarchical clustering can be sensitive to noise and outliers in the data, which can affect the shape.

[1] Next, we mention the steps followed to perform K-means:

Compute the optimal number of clusters: To find the optimal number of clusters to use in our analysis, we will check for the average silhouette width.

The average silhouette width is a measure of the quality of the clusters in a clustering analysis. It is based on the silhouette of each data point, which is a measure of how similar the point is to the other points in its own cluster compared to the other clusters.

The silhouette width for a data point is calculated as follows:

$$s = \frac{b - a}{\max(a, b)} \tag{9}$$

where a is the mean distance between the data point and all of the other points in its own cluster, and b is the minimum mean distance between the data point and the points in the other clusters. The silhouette width ranges from -1 to 1, where a value of 1 indicates that the point is well separated from the other clusters, and a value of -1 indicates that the point is more similar to the other clusters than to its own cluster.

The average silhouette width is the mean of the silhouette widths for all of the data points. A high average silhouette width indicates that the clusters are well-separated and that the data points are similar to each other within the same cluster. A low average silhouette width indicates that the clusters are not well separated and that the data points are not similar to each other within the same cluster.

Therefore, by plotting the average silhouette width against the number of clusters, we choose the number of clusters that corresponds to the maximum average silhouette width.

Cluster Visualization: After choosing the best number of clusters to use, we can compute K-means and plot it on a scatterplot with the points colored according to their cluster assignments.

Now, for the hierarchical clustering we follow the steps:

Compute the optimal number of clusters: To see the best number of clusters to use, we plot the average silhouette width against the number of clusters as we did in K-means.

Cluster Visualization: After choosing the best number of clusters to use, we can compute the hierarchical clustering (divisive) and plot it on a **dendrogram**. A dendrogram is a tree-like diagram that is used to represent the hierarchical structure of a set of data. The data points are represented by leaf nodes at the bottom of the diagram. The clusters are represented by branches that connect the leaf nodes. The length of the branches is proportional to the distance between the points or clusters. This way, the height of the branches on the y-axis represents the distance between the clusters, and the length of the branches on the x-axis represents the number of observations in each cluster.

Evaluate the clustering: We begin by making the silhouette plot. A silhouette plot is a graphical representation of the silhouette of each data point in a clustering analysis. It is used to visualize the quality of the clusters and to understand the structure of the data.

In the silhouette plot, the data points are sorted by the cluster they belong to and the silhouette width. Points with high silhouette widths are plotted toward the right side of the plot, and points with low silhouette widths are plotted toward the left side. A negative value in the silhouette plot means that the value doesn't belong to that cluster.

This can be used to evaluate the quality of the clusters and to choose the optimal number of clusters for the data. It is a useful tool for understanding the structure of the data and for comparing different clustering solutions.

Another way of evaluating the clustering is by calculating the adjusted Rand index (ARI). The ARI is a measure of the similarity between two clusterings of a dataset. It is a commonly used evaluation metric for clustering algorithms and is used to compare the performance of different clustering methods.

In general, an ARI value lies between 0 and 1. A value of 1 indicates a perfect match between the two clusterings, while a value of 0 indicates that the clusterings are no better than random. An ARI value close to 1 indicates that the two clusterings are similar, while an ARI value close to 0 indicates that the clusterings are dissimilar.

[17]

IV. Results

A. PCA - Scatter plot

After we have the dataset without null values and in the desired form, we then start by computing the PCA on the normalized data. To visualize the results we start by drawing a scatter plot, as we explained in the section Principal Component Analysis (PCA). Initially, our first scatter plot with all

observations (all countries) is given in the Appendix by figure 8.

However, we can see that certain countries are quite distant and cohesive, camouflaging other patterns/ relationships that may be happening. So, we took each of these countries out (one at a time) until we could understand more information regarding the remaining data. Thus we arrive at the final scatter plot of the form:

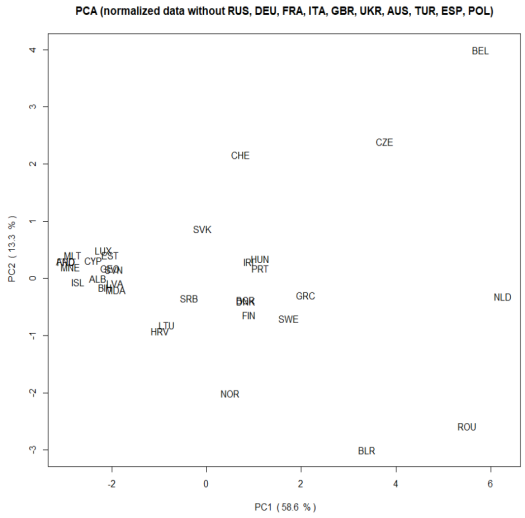

Fig. 1. PCA - Resulting scatter plot.

Thus, we reduced our observations from forty-four to thirty-three, removing, as the plot title in figure 1 says, Russia, Germany, France, Italy, Great Britain, Ukraine, Austria, Turkey, Spain, and Poland.

We can see a division in the countries, the first cluster, is closer to the axis of PC2. Beyond this, we cannot say in a clear way, but perhaps there is another central cluster with the data a bit more dispersed. However, this is just an assumption and it is not clear that this is what is happening.

B. PCA - Biplot

We now proceed to the classical biplot, according to Gabriel: [12]

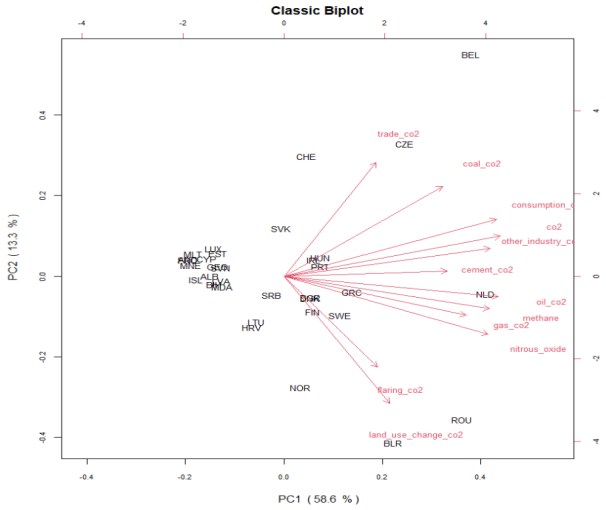

Fig. 2. PCA - Biplot.

It shows the original variables as vectors (arrows). They start at the origin [0,0] and extend to the coordinates given by the loading vector (vector of coefficients that describes the contribution of each original variable to a principal component). These vectors can be interpreted as we saw in the section Principal Component Analysis (PCA).

This way, we can see that the variables trade_co2 , flaring_co2 and land_change_co2 are the ones that contribute the least to the determination of PC1, given that their directions are closer to being parallel to the axis of PC2 than the other variables.

We can also see that the variables trade_co2, nitrous_oxide, land_change_co2 , coal_co2, and consumption are the ones with the greatest vector length, while the variables flaring_co2, cement_co2 and gas_co2 have the smallest length among the variables under study, which indicates that these are perhaps better represented in other components than PC1 and PC2.

We also saw that we can get information through the angles between the vectors of the different variables as these show their correlation in this space. In figure 2, we cannot say with certainty the existence of several clusters of variables, however, we notice the existence of a group of variables, in the middle, where the angles between them are smaller than the rest, which reveals a great correlation with each other.

C. PCA - Cumulative Explained Variance; Scree Plot

It is generally recommended to retain the number of PCs that account for at least 70-80% of the variance in the data. In this case, we get:

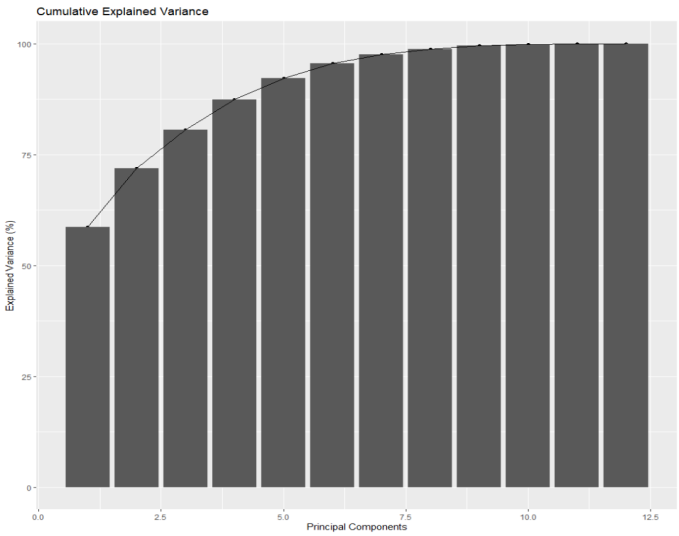

Fig. 3. PCA - Cumulative Explained Variance.

The first 2 PCs explain about 72% of the variance, and the first 3 PCs explain about 80% of the variance. It is also important to consider the interpretability of the PCs and how well they capture the underlying structure of the data. Retaining 2 PCs results in a more satisfactory level of interpretability and captures the relevant structure of the data.

D. FA - Selecting Variables

After doing the KMO, we observe that the coal_co2 variable has a value below 0.5, more precisely, a KMO equal to 0.44. The overall KMO for all variables was equal to 0.66. Removing this variable, as it does not seem appropriate to use in our analysis, we are left with an overall KMO equal to 0.78. We were thus able to see an improvement by removing this column. Furthermore, through trial and error, we were able to see that there was another variable, flaring_co2, that negatively affected our models since it was the variable with the worst results when we applied the models. Therefore, we ended up removing these two variables before performing factor analysis, in order to have the best version of the model we could get. When we normalize the data we no longer count on these variables and proceed with our analysis from there.

E. FA - Bartlett's Test

We obtain a very small p-value (2.95×10^{-84}). This output of the Bartlett test suggests that the correlation matrix of the data is not an identity matrix, since the p-value is less than the significance level of 0.05. This means that the variables in the data are correlated and factor analysis may be a suitable method for exploring these correlations.

F. FA - Selecting the number of factors

As described in the section Factor Analysis (FA), we plot the four different solutions mentioned and we get the figure 9 in Appendix.

This plot shows the eigenvalues of the observed variables on the y-axis and the number of factors on the x-axis. We can see that retaining two factors is the most consistent output between the methods. We can also confirm from the graph that we should retain two factors for our factor analysis since the intercept is still above one on the y-axis.

G. FA - Via PCA

Now, we construct several factorial models and select one (with more interpretable factors).

We explored different rotations of the results: no rotation, varimax, oblimin, and quartimax. The rotation with the best results and was the easiest to explore was quartimax. The output is presented in the Appendix by figure 10.

The output of figure 10 includes the loading vectors. The loading vectors are used to understand the patterns in the data and to identify the variables that contribute most to the variance in the data. To be able to interpret this, we should look at the values of RC1 and RC2. As we can observe the values of the variables under analysis are all present in RC1, which reveals that it is explanatory of most of the variables. However, we can see that there are two variables that have quite low values in this component, such as land_use_change_co2 and trade_co2. These present higher values in the second component, which may indicate that they are more explained by this component.

Another thing we can take from the output is the proportion of variance explained by each component, as well as the cumulative proportion of variance explained by all of the components. The proportion of variance explained can be used to evaluate the quality of the model and to determine how much of the total variance in the data is captured by the principal components.

The output also includes a test of the hypothesis that 2 components are sufficient to explain the variance in the data. The root mean square of the residuals (RMSR) is also included, which is a measure of the error in the model. A low RMSR indicates that the model is a good fit for the data. We obtained an RMSR of 0.077. Therefore, we can conclude that this model is a good fit for our data.

We also have the mean item complexity, which is a measure of the complexity of the items in the dataset. A low mean item complexity indicates that the items are relatively simple and easy to interpret. It ranges from 0 to infinity, where:

- A value of 0 indicates that all of the items in the dataset are unrelated to the factors;
- A value of 1 indicates that all of the items in the dataset are related to the factors to the same extent;
- A value greater than 1 indicates that some of the items in the dataset are related to the factors more strongly than others.

We got a mean item complexity of 1.3, which indicates that some of the items in the dataset are related to the factors more strongly than others. As we inferred from analyzing the loading vectors.

Overall, the output suggests that the principal components are able to explain a significant proportion of the variance in the data and that the model is a good fit for the data.

H. FA - Via Maximum Likelihood

As we explained in Factor Analysis (FA), to be able to perform FA via Maximum we have to make sure that our data follows a multivariate normal distribution. To do so, we can make a Q-Q plot (in an exploratory way) to check if each variable has a normal distribution). Computing the Chi-plot and the Mahalanobis distances is also a good idea. We need to check if the points are close to a straight line. If they are far apart they do not have a normal distribution.

As we can in figure 11 from the Appendix, it is suspected that the normal distribution does not fit the data, since there are points that are relatively far from the straight line.

The results of Mahalanobis distances and the Chi-square plot can also be seen in the Appendix, by figures 12 and 13 respectively.

The Chi-squared plot is similar to the Q-Q plot of the normal distribution but is now thought of as chi-squared. The points should again fit in a straight line. This is not the case. Therefore, Mahalanobis distances do not fit a Chi-squared distribution, which is what we would expect if the data had a multivariate normal distribution. This way, in conclusion, the data does not come from a multivariate normal distribution. Therefore, it is not advisable to use factor analysis via maximum likelihood.

Just to check the structure of our loading vectors, we performed the AF via maximum likelihood anyway and we got the output of figure 14 in the Appendix.

We can see that the loading vectors have the exact same structure as the one from figure 10. The output also includes the sum of the squared loadings (SS loadings) for each factor, as well as the proportion of variance explained by each factor and the cumulative proportion of variance explained by all of the factors. The proportion of variance explained can be used to evaluate the quality of the model and to determine how much of the total variance in the data is captured by the factors. In this case, the output indicates that the first factor explains 62.5% of the variance in the data, while the second factor explains an additional 11.2% of the variance. Together, the two factors explain 66.3% of the variance in the data. The remaining variance is not explained by the factors and is represented by the uniquenesses of the variables, represented by figure 15 in the Appendix.

From this, we can say that even suspecting that the multivariate normality assumption is not valid, we performed and found that the resulting factorial structure, via maximum likelihood, is equal to the one obtained when the factors were extracted using PCA. This consistency of results suggests that maximum likelihood may be robust to the normality assumption, and on the other hand, it demonstrates that the factorial structure found is strong enough since it does not change with other procedures (even if these fail in some assumptions).

I. FA - Cumulative Explained Variance; Scree Plot

The rest of our analysis will depend on the result of the factor analysis via PCA since this is the most suitable method for our data.

Just like we did in the subsection PCA - Cumulative Explained Variance; Scree Plot, we will show the plot for the cumulative explained variance, but now for the FA, to know what's the best number of factors to retain:

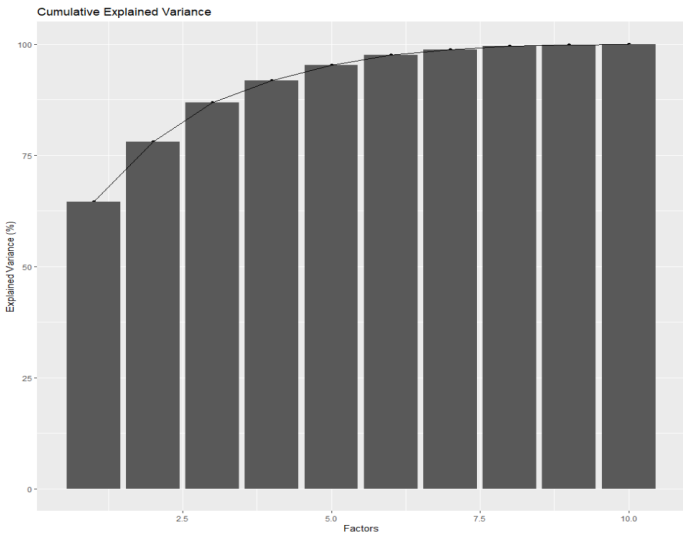

Fig. 4. FA - Cumulative Explained Variance.

The first 2 factors explain about 78% of the variance, and the first 3 factors explain about 87% of the variance. It is also important to consider the interpretability of the PCs and how well they capture the underlying structure of the data. Retaining 2 factors results in a more satisfactory level of interpretability and captures the relevant structure of the data.

J. FA - Residuals

We extracted the residuals from the results of the factor analysis. The residuals are the differences between the observed values and the fitted values of the data. We found that the proportion of absolute residuals that are less than 0.05 is about 60%. This indicates that the model is fitting the data well.

K. FA - Scores

To finish this analysis we can estimate the scores for each factor for each subject in the data. These are presented in the Appendix, in figure 16.

The matrix with the factor scores of figure 16 can be used in subsequent analyses (cluster analysis, scatterplot, etc,...). The main idea is to extract information from multivariate data.

L. PCA - Updated Biplot without the variables we dropped in FA

Let's now update the biplot that we made in the subsection PCA - Biplot, removing the variables that we removed to apply FA:

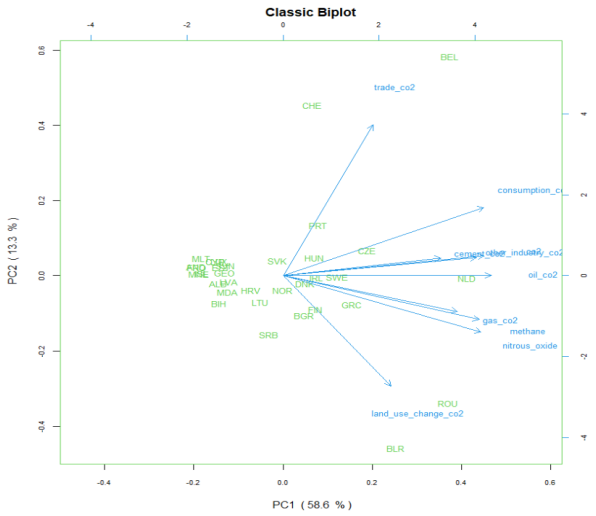

Fig. 5. FA - Updated biplot.

In this biplot we can notice some differences such as some clusters of variables closer and more cohesive than in the previous results, however the variables trade_co2 and land_use_change_co2, are still quite distant from the others, which we already expected, since they are not explained by the same factor as the others.

M. Clustering the data - Non-Hierarchical Clustering

We start by clustering the data using non-hierarchical clustering, more specifically, K-means. So, we then start by seeing what the best number of clusters is, as we explained in section Clustering, by plotting the average silhouette width in figure 17 in the Apendix. By this plot we can see that we should retain two clusters.

After this, we perform the K-means, choosing two clusters and got the resulting plot:

Fig. 6. K-means of the data.

We can see that PC1 discriminates the two clusters. When the PC1 is approximately greater than 1.25, we see the blue cluster. When it is smaller, we find the red cluster. PC2 does not discriminate between either cluster.

Plotting the silhouette in figure 18 of Appendix, we can see some negative values, which means that these values probably don't belong to this cluster. The first cluster presents better results, with an average silhouette width of 0.57. The second cluster presents an average silhouette width of 0.11.

N. Clustering the data - Hierarchical Clustering

Now we will compare the output for the K-means with the output for hierarchical clustering. We start by calculating the optimal number of clusters again, and we get that we should also use two clusters in this case, as we can see in figure 19 in the Appendix.

We then plot the dendrogram for the hierarchical clustering:

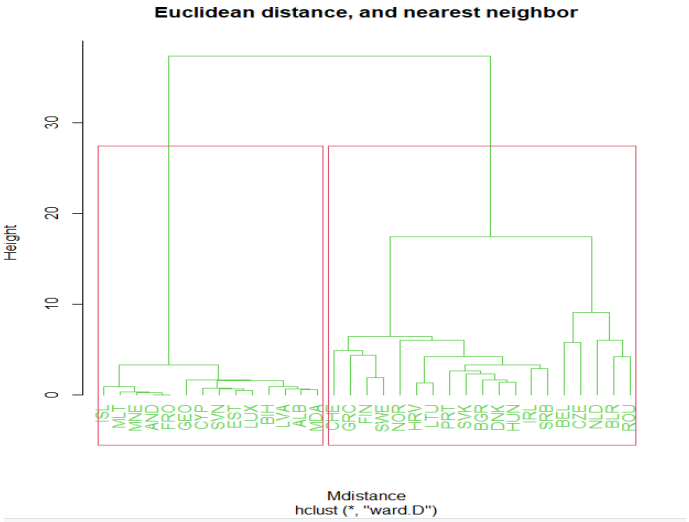

Fig. 7. Dendrogram - clustering of the data.

The first cluster shown is the least numerous and contains fourteen countries. The second cluster contains nineteen countries, being the largest cluster for the data.

As we did before, plotting the silhouette in figure 20 in Appendix, we can still see a negative value. The first cluster presents better results, with an average silhouette width of 0.57. The second cluster presents an average silhouette width of 0.14, which is relatively better, but not significant.

O. Evaluation of our clustering

As we explained in the section Clustering, to evaluate our clustering of the whole data we will perform two comparisons.

We will start by comparing with the clustering of the factor's scores we calculated in FA - Scores.

1) Comparison between the clustering of the data and the clustering of the factor's scores: To do the clustering with the factor's scores, we performed non-hierarchical clustering of these, just like we did before with the dataset.

We got that the optimal number of clusters was two, as we can see in the figure 21 in the Appendix.

This way, performing the K-means we got the result of figure 22 in Appendix.

Now, we can easily compare the results of the clustering of the data with what we got for the factor' scores.

We start by using the K-means results of the data from the section Clustering the data - Non-Hierarchical Clustering and get an ARI of 0.312 and an average silhouette width of 0.471. This indicates that the observations within the clusters are somewhat similar to each other and somewhat different from the observations in other clusters, but it is far from good.

Now we compare using the hierarchical clustering results of the section Clustering the data - Hierarchical Clustering. From this, we get an ARI of 0.185, which is much smaller than the result obtained for K-means. We got an average silhouette width of 0.505, which is slightly higher than the one obtained previously.

2) Comparison between the clustering of the data and the clustering of the population: Now let's make a comparison of the clustering of the data with a clustering of the population column handmade by us, as we explained in the section Data Analysis and Preparation.

We start by using the K-means results of the data from the section Clustering the data - Non-Hierarchical Clustering and get an ARI of 0.383, which is higher than what we got compared with the factor' scores, and an average silhouette width of 0.459, which is smaller than what we got comparing with the factor' scores.

When comparing the hierarchical clustering results of the data from the section Clustering the data - Hierarchical Clustering, we get an ARI of 0.237, which is lower than the previous one, and an average silhouette width of 0.505, which is the same we got when comparing the hierarchical clustering of the data with the results from the clustering of the factor' scores.

This comparison was made because we thought that a country with more population would have a higher emission of CO2 and greenhouse gases. However, as we can see, our results are weak, and therefore we can infer nothing about this.

V. Conclusion

In conclusion, during the elaboration of this article, we carried out three different methods with the objective of elaborating an exploratory analysis of the proposed dataset about CO2 production and its variants in European countries.

The main objective was to extract as much information as possible, allowing us to identify patterns and similarities. Throughout the study, it was necessary to make several changes, such as removing some variables and even some countries, in order to obtain better results.

From the PCA and FA results, we were able to better understand the relationships between the variables, including the variables that contribute the most to the variance of our data. We can see that in both PCA and FA, two components are sufficient to explain a large proportion of the variance in the data, which suggests that the patterns in our data are relatively simple and well-described by the models.

One minor impediment that arose in FA was that our data did not follow a multivariate normal distribution, which did not allow us to rely on the results via maximum likelihood for further analysis.

Regarding clustering, when we do its evaluation, we get a better result for ARI when compared to the population cluster defined by us. As for the average silhouette width, we obtain a better result when we do hierarchical clustering. However, these results leave much to be desired. When compared with the population clustering, we were not able to confirm that countries with larger population tend to perform more CO2 emissions, as we wished, possibly due to having eliminated the countries with the largest population. A possible improvement would be to use another type of comparisons such as GDP per capita, geographical location, or even mineral wealth.

References

[1] Advantages disadvantages of k-means and hierarchical clustering (unsupervised learning). http://santini.se/teaching/ml/2016/Lect_10/10c_UnsupervisedMethods.pdf. Accessed: 2022-12-28.

[2] cortest.bartlett: Bartlett's test that a correlation matrix is an identity matrix. https://www.rdocumentation.org/packages/psych/versions/2.2.9/topics/cortest.bartlett. Accessed: 2022-12-22.

[3] Co and greenhouse gas emissions. https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions. Accessed: 2022-12-09.

[4] Factor analysis advantages and disadvantages. https://www.ipl.org/essay/Factor-Analysis-Advantages-And-Disadvantages-F3885JNNPC486. Accessed: 2022-12-22.

[5] K-means clustering explained. https://neptune.ai/blog/K-means-clustering. Accessed: 2022-12-22.

[6] nscree: Non graphical cattel's scree test. https://www.rdocumentation.org/packages/nFactors/versions/2.4.1/topics/nScree. Accessed: 2022-12-22.

[7] O que é o pca – principal component analysis e como aplicá-lo a um conjunto de dados. https://www.bi4all.pt/noticias/blog/o-que-e-o-pca/. Accessed: 2022-12-22.

[8] principal: Principal components analysis (pca). https://www.rdocumentation.org/packages/psych/versions/2.2.9/topics/principal. Accessed: 2022-12-22.

[9] Vantagens e desvantagens de k-means. https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages. Accessed: 2022-12-22.

[10] What is clustering and different types of clustering methods. https://www.upgrad.com/blog/clustering-and-types-of-clustering-methods/. Accessed: 2022-12-22.

[11] What is hierarchical clustering? an introduction to hierarchical clustering. https://www.mygreatlearning.com/blog/hierarchical-clustering/. Accessed: 2022-12-22.

[12] K. R. Gabriel. *The Biplot Graphic Display of Matrices with Application to Principal Component Analysis*. Oxford University Press, 1971.

[13] J. C. Hayton, D. G. Allen, and V. Scarpello. Factor Retention Decisions in Exploratory Factor Analysis: a Tutorial on Parallel Analysis. *Organizational Research Methods, 7(2), 191–205*, 2004.

[14] International Energy Agency (IEA). *CO2 emissions from fuel combustion - Highlights*. 2021.

[15] Henry F. Kaiser. A second generation little jiffy. *Psychometrika volume 35, pages401–415*, 1970.

[16] Jae-On Kim, Olli Ahtola, Paul E. Spector, Joe-On Kim, and Charles W. Mueller. *Introduction to Factor Analysis: What It Is and How To Do It*. Geological Survey of New South Wales, 1978.

[17] Dao Lam and Donald C. Wunsch. *Chapter 20 - Clustering*. Elsevier, 2014.

[18] Elizabeth R. Lusczek, Nicholas E. Ingraham, Basil S. Karam, Jennifer Proper, Lianne Siegel, Erika S. Helgeson, and et al. Cumulative proportion of variance explained. 2021.

[19] B. Williams, A. Onsman, and T. Brown. Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine*, 2010.

APPENDIX

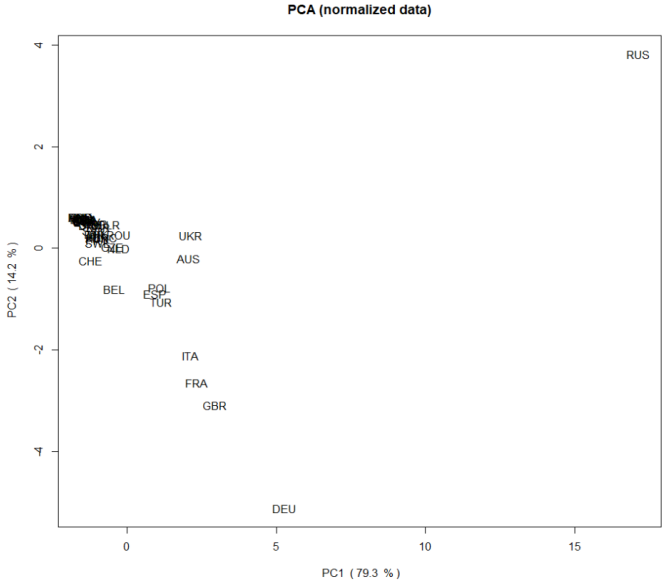

Fig. 8. PCA with all countries - Scatterplot.

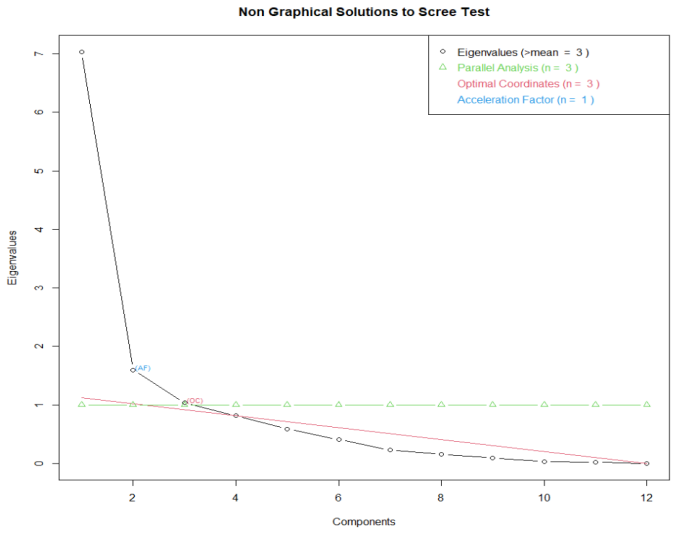

Fig. 9. FA - Scree test.

```
> print(AF,cut=0.3,digits=3)
Principal Components Analysis
Call: principal(r = rMatrix, nfactors = 2, residuals = TRUE, rotate = "quartimax",
    covar = TRUE)
Standardized loadings (pattern matrix) based upon correlation matrix
                     RC1    RC2    h2     u2  com
cement_co2         0.707        0.519 0.4809 1.07
co2                0.902        0.839 0.1609 1.06
gas_co2            0.800        0.661 0.3387 1.07
land_use_change_co2 0.522 -0.563 0.589 0.4106 1.99
methane            0.900        0.844 0.1560 1.08
nitrous_oxide      0.913        0.896 0.1045 1.15
oil_co2            0.942        0.890 0.1098 1.01
other_industry_co2 0.870        0.779 0.2209 1.06
consumption_co2    0.887  0.421 0.964 0.0363 1.43
trade_co2          0.360  0.834 0.825 0.1753 1.36

                       RC1   RC2
SS loadings          6.430 1.376
Proportion Var       0.643 0.138
Cumulative Var       0.643 0.781
Proportion Explained 0.824 0.176
Cumulative Proportion 0.824 1.000

Mean item complexity =  1.2
Test of the hypothesis that 2 components are sufficient.

The root mean square of the residuals (RMSR) is  0.077

Fit based upon off diagonal values = 0.985
```

Fig. 10. FA via PCA.

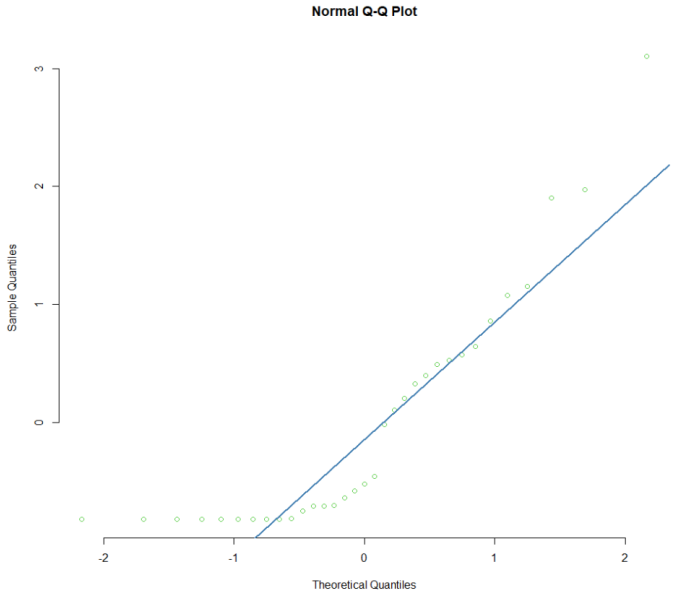

Fig. 11. QQ-plot.

```
> # Mahalanobis Distance
> mahalanobis(datax5_normalized[1:33, 1:10], colMeans(datax5_normalized[1:33, 1:10]), cov(datax5_normalized[1:33, 1:10]))
      ALB       AND       BLR       BEL       BIH       BGR       HRV       CYP       CZE       DNK       EST       FRO       FIN
 1.582370  3.537065 19.463672 20.021103  3.329898  2.340951  4.164116  3.494008 22.710398  3.927775  2.701078  2.290667  8.235537
      GEO       GRC       HUN       ISL       IRL       LVA       LTU       LUX       MLT       MDA       MNE       NLD       NOR
 5.465680 13.520737  7.073947 22.953590 20.598920  1.157755  4.888947  3.303609  2.090239  1.771321  2.036050 26.350994 23.623168
      PRT       ROU       SRB       SVK       SVN       SWE       CHE
17.604850 23.082157 11.951595  7.255384  1.139253 16.799871  9.533298
```

Fig. 12. Mahalanobis Distances.

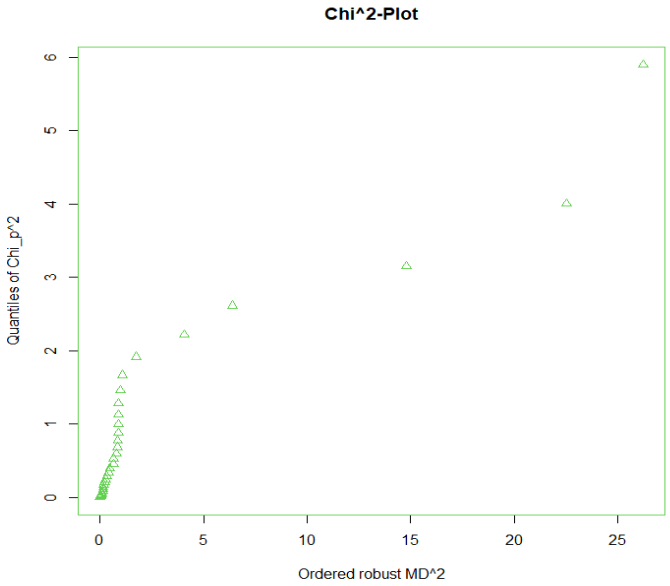

Fig. 13. Chi-plot.

```
> print(AFmv$loadings,cut=0.35,digits=3)

Loadings:
                   Factor1 Factor2
cement_co2          0.639
co2                 0.895
gas_co2             0.797
land_use_change_co2 0.445  -0.445
methane             0.901
nitrous_oxide       0.919
oil_co2             0.931
other_industry_co2  0.848
consumption_co2     0.904   0.422
trade_co2           0.386   0.701

                 Factor1 Factor2
SS loadings        6.253   1.116
Proportion Var     0.625   0.112
Cumulative Var     0.625   0.737
```

Fig. 14. FA via Maximum Likelihood.

```
> AFmv=factanal(datax5_normalized, factors=2, scores = c("Bartlett"), rotation = "quartimax"); AFmv

Call:
factanal(x = datax5_normalized, factors = 2, scores = c("Bartlett"),     rotation = "quartimax")

Uniquenesses:
         cement_co2                 co2             gas_co2 land_use_change_co2             methane       nitrous_oxide             oil_co2
              0.578               0.167               0.354               0.604               0.122               0.054               0.129
 other_industry_co2      consumption_co2          trade_co2
              0.260               0.005               0.359
```

Fig. 15. FA via Maximum Likelihood - Uniqueness.

```
> scores <- scores$scores; scores
             RC1         RC2
ALB -0.83241246 -0.169576697
AND -1.11518376  0.109206519
BLR  1.59450197 -2.695254812
BEL  1.78206587  3.387596160
BIH -0.83964490 -0.349886581
BGR  0.27467177 -0.658659706
HRV -0.44819105 -0.170872657
CYP -0.89777929  0.131211038
CZE  1.18159565  0.346856259
DNK  0.38286822 -0.149642417
EST -0.76099747  0.128405968
FRO -1.11442805  0.107596504
FIN  0.46230924  0.033650445
GEO -0.70548374 -0.231472208
GRC  0.87208551 -0.327076844
HUN  0.41179653  0.232472567
ISL -1.03640191 -0.002729511
IRL  0.63179874 -0.772784705
LVA -0.68639982 -0.046522555
LTU -0.26058778 -0.374889781
LUX -0.84146151  0.146887606
MLT -1.05183648  0.248563956
MDA -0.74777293 -0.365160308
MNE -1.07880320  0.029945314
NLD  2.65901910  0.297053957
NOR  0.02583987 -0.104933542
PRT  0.36662348  0.672669061
ROU  1.97411153 -1.734117921
SRB -0.20812302 -1.225694746
SVK -0.14506222  0.247286433
SVN -0.75094021  0.199876655
SWE  0.71009620  0.357422916
CHE  0.19212613  2.702573633
```

Fig. 16. FA - Factor's Scores.

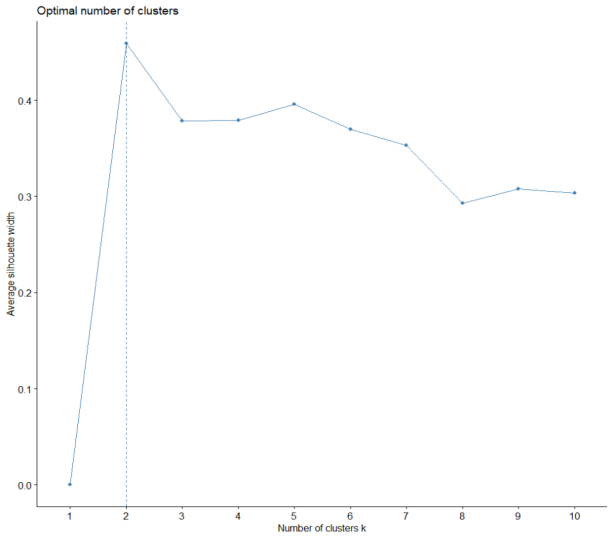

Fig. 17. Non-Hierarchical Clustering - Number of clusters.

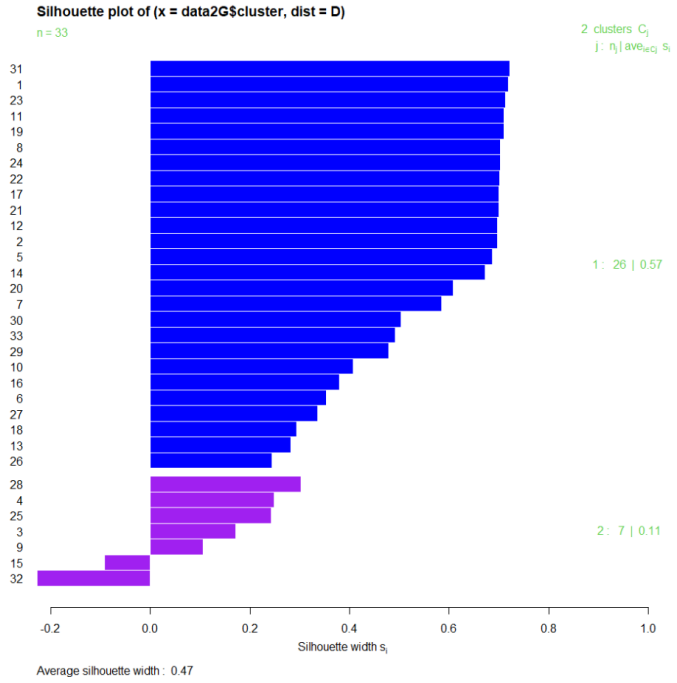

Fig. 18. Silhouette plot - Kmeans.

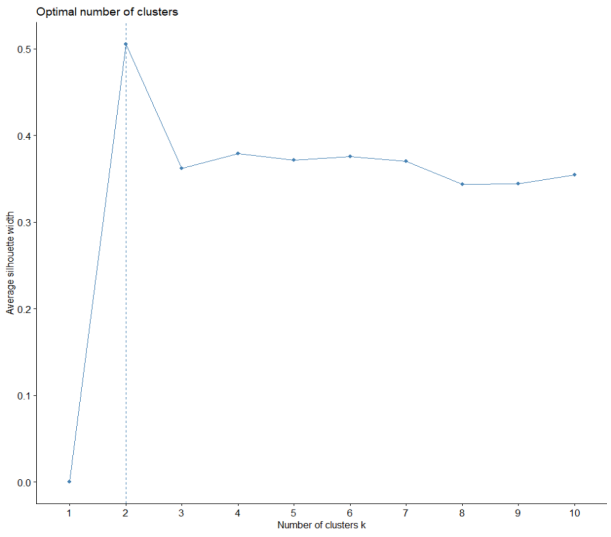

Fig. 19. Hierarchical Clustering - Number of clusters.

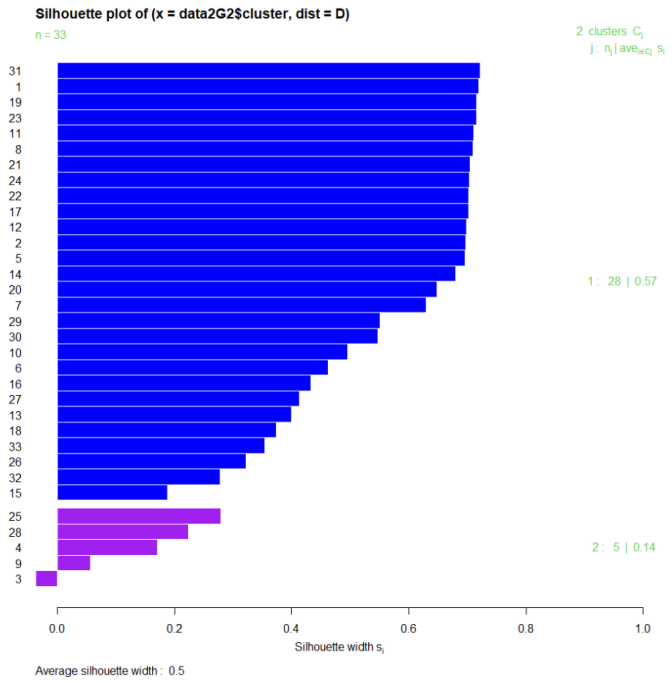

Fig. 20. Silhouette plot - Hierarchical clustering.

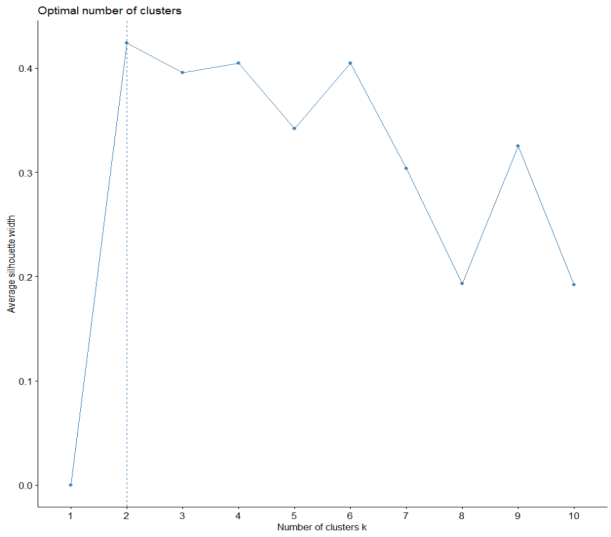

Fig. 21. K-means of factor scores - Number of clusters.

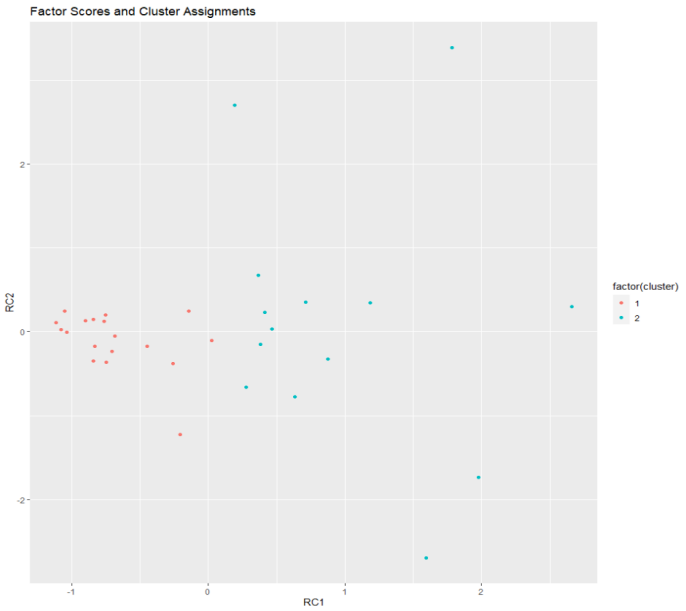

Fig. 22. K-means of factor scores.