Supervised and Unsupervised Algorithm Analysis on Economic Development

Cosimo Schiavoni - 964563
DSE Unimi - Statistical Learning

January 2025

Abstract

This project presents an in-depth examination of the economic development of 147 countries using cross-sectional data from 2021. The study is structured into two main parts.
The first part applies supervised learning algorithms such as Multiple Linear Regression, Generalized Additive Models, Random Forest, Ridge and Lasso Regression to analyze the relationship between GDP per capita, a measure of economic development, and several independent variables (including Corruption Control, Foreign Direct Investment, Political Stability, and the Democratic Index). The objective is to identify the factors that most significantly influence economic development across the dataset.
The second part of the study adopts unsupervised learning techniques, focusing on clustering. The clustering process uses the independent variables (such as Corruption Control, Political Stability, Unemployment, and Urban Population) as input to automatically generate groups of countries based on shared characteristics. The analysis seeks to determine whether these variables provide a meaningful classification of countries and whether some economic or social pattern can be described by applying those algorithms.

Keywords: Statistical Learning, Supervised Algorithms, Unsupervised Algorithms, Economic Development, Corruption Control.

Introduction

Economic development has long been a topic of interest to researchers and policy makers alike. Gross domestic product per capita (GDP PC) is often used as a measure of the economic development of nations. However, the factors that contribute to economic development are multifaceted and complex to determine. This study analyzed the relationship between GDP PC and a number of explanatory variables that include both economic and institutional factors (control of corruption, political stability, Foreign direct investment (FDI), export value index, labor force participation rate, self-employment rate, unemployment, urban population, electricity consumption, percentage of individuals using the internet, democracy index).

For this purpose, an adjusted data set of 147 countries (out of a total of 216 available countries) for the year 2021 was considered. Regarding the size of the dataset, it is important to note that the amount of datasets analyzed is adequate for the scope of the research. Indeed, a cross-sectional analysis is performed at country level and, due to the data cleaning process, almost 70% of the entire population is considered. Statistical learning, unlike machine learning, has the goal of exploring potential relationships and generating hypotheses, therefore statistical methods can be used to significant correlations or trends. For this reason, this project is not aimed at developing predictive models, the success of which usually depends on the accuracy of the prediction and not on the economic interpretation of the model. The amount of data sets could potentially affect the complexity of the model. When analysing relatively simple models (e.g. Multiple Linear Regression), fewer data points are usually sufficient to achieve good performance, so the data set obtained might be sufficient. In this case, it is important to make a careful selection of variables to avoid the curse of dimensionality (the model might become too complex for the amount of data used in the analysis). In contrast, when using complex models (e.g. random forest), larger data sets are usually required to work effectively. However, in such a case, the introduction of specific techniques (e.g. boosting, bagging or cross-validation) can help to mitigate the risk of overfitting and provide a more reliable estimate of the model's performance. In this context, it is important to keep in mind that the analysis of official data at country level should ensure the quality of observations, which is as important as the quantity of observations.

For the same reason, since the analysis is focused on macroeconomic factors, a year of data can still provide valuable insights, especially when using relatively stable variables over time. Nonetheless, any policy implications must be viewed with caution due to the temporal limitations of the dataset. Policy makers should indeed look for trends over time, while a one-year snapshot may not capture all the dynamics needed for this purpose. To gain a more comprehensive understanding of the phenomenon, it would be interesting to conduct a panel data analysis, taking into account more data and observing time effects, but this is beyond the scope of this study. By applying supervised and unsupervised statistical learning algorithms, this study has the following objectives:

Supervised Learning Analysis: to assess the impact of economic and institutional indicators on the level of economic development (GDP PC). The analysis aims to develop a comprehensive understanding of the key variables that influence economic development. By evaluating the statistical relationship between different economic and institutional variables, it is possible to point out the most significant factors that contribute to move countries out of the middle-income trap.

Unsupervised Learning Analysis: the identification of natural clusters of countries based on their economic and institutional indicators. In this sense, the aim is to uncover hidden patterns and similarities between nations based on the analysis of factors influencing economic development.

Theory Foundations

This project draws its theoretical foundations from *The Elusive Quest for Growth: Economists' Adventures and Misadventures in the Tropics* by William Easterly (2001). A former development economist at the World Bank, Easterly sets out his theory that some nations experience rapid economic growth while others lag behind. One of the main factors inhibiting economic growth is corruption, which is exacerbated by the lack of adequate economic incentives. Despite the efforts of the international community to help impoverished countries through targeted policies, investments and financial aid, most of these countries continue to live in impoverished conditions. According to Easterly, these initiatives often fail because they do not take into account key economic principles. Several key factors contribute to this phenomenon:

Unproductive Fixed Capital Investments: refers to capital allocations that do not generate significant productivity gains or promote economic development. Investments in fixed capital do not necessarily increase productivity, while investments in every-improving fixed capital usually do. If resources flow into unstrategic technologies and infrastructure, economic development can stagnate. In contrast, investing in advanced and constantly evolving technologies should increase productivity for the benefit of economic development.

Debt Forgiveness: often seen as a humanitarian necessity for highly indebted countries, could potentially lead to undesirable disadvantages. While providing immediate financial relief, it often fails to address the structural problems that led to the accumulation of debt in the first place, such as weak governance, fiscal mismanagement and an over-reliance on external borrowing. Without structural reforms, debt relief can create a cycle of dependency that encourages further borrowing in anticipation of renewed relief, thereby promoting irresponsible financial behavior (moral hazard).

Lack of Returns on Education: education is widely regarded as an important factor for economic development, but its impact is limited if there are no sound economic incentives. In many countries affected by labor market distortions, investments in education do not yield adequate returns. In fact, individuals may be discouraged from investing time and resources in their own education. Moreover, the lack of economic return on education discourages further investment in education, leading to workers migrating in search of better opportunities and perpetuating inequality.

Uneffective Birth Control Policies: efforts to control population growth in developing countries through birth control policies, are often expected to benefit long-term economic growth strategies. However, in many impoverished countries where employment opportunities are normally scarce and infant mortality rate is high, birth control policies often overlook the deeper structural causes of underdevelopment. Indeed, families in poor countries usually trade off quantity of children against investment in education, opting for high number of children with a low level of education. Moreover, birth control policies do not address the broader issues of poor governance, lack of educational infrastructure, or inadequate job creation. Therefore, economic aid that focuses on population control, without addressing structural factors, is unlikely to deliver the desired benefits.

To summarise, the failure of many international development policies is due to their inability to address fundamental economic inefficiencies. Successful development initiatives must focus on encouraging productive investment, curbing corruption, creating economic incentives, and ensuring that education and population control efforts are supported by broader institutional reforms.

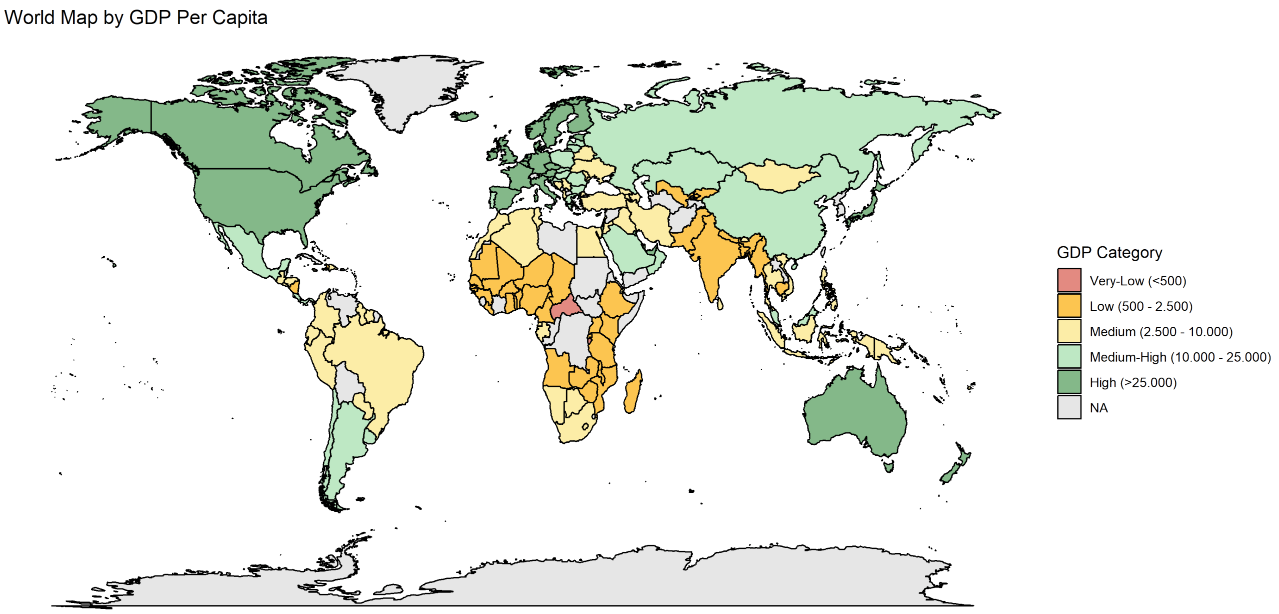


Figure 1: This figure illustrates the level of GDP per Capita for each anlyzed country.

Variables

The key variables used in the analysis include:

Gross Domestic Product per Capita (GDP PC): is the ratio between gross domestic product and population and is generally regarded as one of the most important indicators of economic development. Compared to gross domestic product (GDP), which can be used to measure a country's level of prosperity, GDP PC takes into account the size of the population, according to which prosperity should be distributed.

Corruption Control (COR): corruption is believed to be one of the main factors undermining the efficiency of state institutions, distorting the functioning of markets and having a major impact on innovation and productivity. High levels of corruption can discourage foreign investments (due to reduced effectiveness of public spending) and limit the ability to improve infrastructure and the education system.

Political Stability and Absence of Violence/Terrorism (POL): political stability is an important factor in long-term economic planning and attracting investment. Countries with an unstable political environment often suffer from unpredictable political changes and violence, which discourage both domestic and foreign investment. Political instability can be a cause of low economic growth rates.

Foreign Direct Investment (FDI), Net Inflows (% of GDP): foreign direct investment is a very important source of capital, technology and management expertise. Increased foreign direct investment can boost economic growth, create jobs and help a country develop new industries. However, reliance on FDI in low value-added sectors can lead to the middle-income trap persisting unless it encourages the development of higher value-added industries.

Export Value Index (EXP), (2015 = 100): export value is an important indicator of economic competitiveness. Countries stuck in the middle-income trap often struggle to diversify their export base and move from low value-added products to high value-added products. On the contrary, a higher export value index indicates a better performance in global trade, which can improve income by increasing foreign exchange earnings and stimulating domestic industry.

Labor Force Participation Rate (LAB), Total (% of Total Population Ages 15-64): a higher labor force participation rate usually indicates a more engaged and productive workforce. Indeed, to escape the middle-income trap, it is important to maximize the use of human capital. A low labor force participation rate, on the other hand, can hinder economic development (especially due to social inequalities or skills mismatches).

Self-Employed (SFE), Total (% of Total Employment): a high level of self-employment can often indicate a lack of employment opportunities and be a sign of an underdeveloped economy. In middle-income countries, the transition from self-employment to employment can be crucial to increasing productivity, especially if self-employment is out of necessity rather than choice.

Unemployment (UNE), Total (% of Total Labor Force): high unemployment can reduce consumption and productivity and thus impair economic development. Moreover, persistent unemployment can lead to social inequalities and affect political stability, further complicating efforts to escape the middle-income trap.

Urban Population (URB), (% of Total Population): urbanization usually correlates with higher productivity and economic development, as cities provide better access to jobs, education and services. However, if urbanization is unplanned and lacks adequate infrastructure, it can be detrimental to economic development.

Percentage of Individuals Using the Internet (INT): internet access can be considered an essential factor to be integrated into the global economy, enhancing education, innovation and improving productivity. A correct use of the internet and its related technological opportunities may stimulate economic development by creating innovative industries (e.g., IT services, e-commerce, AI) and improving productivity across traditional sectors.

Electricity Consumption (ELE), (Kilowatt-hours, Million): electricity consumption can be seen as an indicator of industrial activity and general economic development. Higher energy consumption generally indicates more intensive economic activity. In middle-income countries, inadequate or unstable energy supply can limit industrial growth and innovation, trapping the country in a low-growth scenario.

Democracy Index (DEM IND): is an index that measures the quality of democracy around the world and is based on 60 different indicators. With regard to this variable, it is not possible to declare a priori which system performs better in terms of economic development, but it is reasonable to expect different performances depending on the system of government.

Democracy Flag (DEM FL): the Democracy Flag is a categorical variable that classifies each country into the following categories based on the Democracy Index level: ”Full Democracy”, ”Weak Democracy”, ”Hybrid Regime” or ”Authoritarian Regime”.

To summarize, these variables interact in complex ways to define a country's ability to escape the middle-income trap and embrace economic development.

Data Sources

The data for this analysis were obtained from multiple offcial sources, ensuring a comprehensive and diverse set of variables to observe economic development. The main data sources include:

The Economist's Democracy Index 2021: The variables *DEM FLAG* and *DEM INDEX* were extracted from the Democracy Index report (2021), which ranks countries based on their democratic freedoms and governance structures.

The World Bank Database: Several key economic and political indicators were sourced from The World Bank, including the *Control of Corruption*, *Political Stability and Absence of Violence/Terrorism*, *Foreign Direct Investment (FDI)*, *Export Value Index (2015 = 100)*, *Labor Force Participation Rate*, *Self-Employment Rate*, *Unemployment Rate* and *Urban Population.* [1]

United Nations Statistics Division (UNSD): Data on internet usage (*Percentage of Individuals Using the Internet*) and energy consumption (*Electricity - Final Energy Consumption*) were gathered from the UNSD. [2]

Data Preprocessing

The extracted datasets were merged based on the common variable *Country Name*, ensuring each observation represented a single country. The following cleaning procedures were applied afterwards:

Handling Missing Values: missing values were handled removing the entire row.

[1] https://databank.worldbank.org/source/statistical-performance-indicators-(spi)

[2] http://data.un.org/

Yeo-Johnson Transformation: the Yeo-Johnson transformation improves linear regression model performance by stabilizing variance and making the data more normally distributed, which helps meet key assumptions like linearity, homoscedasticity and normal distribution of residuals.

Outliers: outliers detection methods like Z-score (pre-regression diagnostics) and leverage points (post-regression diagnostics) were employed. Any extreme outlier was treated by removing the entire row from the dataset.

Scaling: variables were transformed to ensure they were on comparable scales, to ensure they were more suited to run unsupervised classification models.

Polinomials Introduction: in addition to the original variables, quadratic transformations were introduced for variables that exhibited potential non-linear relationships with the dependent variable.

This cleaned and transformed dataset formed the basis for the subsequent analysis.

Supervised Learning

In this section, inference was performed based on the analysis of labeled data, meaning that the dataset includes both input data and the desired output.

Multiple Linear Regression Model

Methodology and Algorithm

Multiple linear regression model was employed to investigate the relationship between the dependent variable Y and a set of k independent variables $X_1, X_2, \ldots, X_k$. The general form of the multiple linear regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon \tag{1}$$

where:
Y is the dependent variable representing `GDP PC`.
$X_1, X_2, \ldots, X_k$ are the independent variables, such as `COR`, `POL`, `FDI`, `EXP`, `LAB`, `EMP`, `UNE`, `URB`, `INT`, `ELE` and `DEM FLAG`.
β_0 is the intercept term, representing the expected value of Y when all independent variables are equal to zero.
$\beta_1, \beta_2, \ldots, \beta_n$ are the regression coefficients, indicating the change in Y for a one-unit increase in each corresponding independent variable, holding other variables constant. Each regression coefficient represents also the slope of the line obtained by running Ordinary Least Sqares (OLS) between Y and X_i for the i-th observation.
ϵ is the error term, capturing unobserved factors that affect Y.

The estimation procedure foresees the use of Ordinary Least Squares (OLS) method to compute regression coefficients. OLS is conceived to minimize the sum of squared residuals, which is to say the differences between observed and predicted values of Y:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{2}$$

where $\hat{Y}_i$ is the predicted value of Y for the i-th observation, and n is the total number of observations.

Regarding the Statistical Significance of the model, t-test was employed. The null hypothesis is that independent variables have no effect on Y, meaning that each β_j equals zero ($H_0 : \beta_j = 0$). The t-statistic formula is the following one:

$$t_j = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)} \tag{3}$$

where $\mathrm{SE}(\hat{\beta}_j)$ is the standard error of $\hat{\beta}_j$. The null Hypothesis is rejected when the p-value is below a pre-specified threshold (e.g. 0,05). In that case it is possible to conclude that the variable is statistically significant.

Model Assumptions

Multiple linear regression can be a very powerful model due to its simplicity. However in Statistical learning it is important to consider several assumptions before running OLS. The violation of those assumptions may lead toward biased or inefficient estimates. In order to mitigate the issues, appropriate diagnostic tests must be performed. Here is reported a list of assumptions made during the development of this research:

1. Linear Relationship: It is assumed that the relationship between the dependent variable Y and the independent variables X is linear. This means that the model should be expressed as $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$, where ϵ is the error term. Otherwise, if the relationship is not linear, it is necessary to introduce non-linear techniques such as polinomial regression or trasform the data to make it linear.

2. Sample Variation in X: independent variables X_i must be variable. Formally, for some $i, j \in \{1, 2, \ldots, n\}$, it is required that $X_i \neq X_j$. Indeed, it would not be possible to estimate the relationship between X and Y without variation.

3. X and Y are Random Variables (I.I.D.): it is assumed that both the independent variable X and the dependent variable Y are independently and identically distributed (i.i.d.). This ensures that the observations are independent of each other and that each observation originates from the same underlying probability distribution.

4. X and Y are outlier free: outliers can strongly influence the OLS results by distorting the parameter estimates. It is important to recognize and handle outliers using diagnostic methods such as Z-score (pre-regression diagnostics) or leverage plots (post-regression diagnostics).

5. Exogeneity of X ($\mathbb{E}[\epsilon|X] = 0$): foresees that for the independent variables X, the error term ϵ has a conditional mean value of zero. This condition ensures that the OLS estimator $\hat{\beta}$ is unbiased. In particular, exogeneity implies that there is no correlation between X and the error term ϵ.

6. Homoskedasticity (Var$(\epsilon|X) = \sigma^2$): forsees the constant variance of the error terms. Robust standard error estimators must be applied if Heteroskedasticity is detected. To this extent, the Breusch-Pagan test is used to test for heteroscedasticity.

7. Normal Distribution of Errors: a normal distribution of the residuals ϵ was assumed. In cases where the residuals deviate from the normal distribution, transformations such as the Yeo-Johnson can be applied to stabilize the variance and achieve a distribution that is closer to normality. Yeo-Johnson transformation can handle zeros and negative values while transforming the data to approximate normality, which is the case in this research.

8. Autocorrelation of Error Terms: violates the assumption that the error terms are uncorrelated, which can affect the efficiency of the OLS estimates and lead to misleading inferences. In this respect, the Durbin-Watson test can be used to detect the presence of autocorrelation in the residuals ϵ.

9. Multicollinearity: occurs when independent variables are highly correlated with each other, inflating the standard errors of the coefficient estimates and making them unreliable. The Variance Inflation Factor (VIF) was introduced to check for multicollinearity among the independent variables. The VIF assesses how much the variance of a regression coefficient is increased due to multicollinearity. A higher degree of multicollinearity between the variables is indicated if the VIF is particularly high (a VIF value of more than 10 may indicate a significant multicollinearity).

It is important to note that proprierties from 1 to 5 guarantee that β_0 and β_1 are unbiased. In addition, proprierty 6 guardantees that that the same coefficients are Best Linear Unbiased Estimators (BLUE), according to Gauss-Markov Theorem. Moreover, when data are normally distributed or $n \to \infty$, then:

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma \sum_{i=1}^{n}(x_i - \bar{x})^2} \sim N(0, 1) \tag{4}$$

where σ^2 is estimated by the Standard Error of Regressors:

$$S^2 = \frac{\sum (e_i)^2}{n - 2} \tag{5}$$

This is an important assumption as far as it guarantees the possibility to introduce Confidence Intervals and Inference in the analysis.

Preliminary Diagnostics

Firstly, a **linearity** check was performed because the preliminary data exploration and diagnostic plots indicated a potential nonlinearity in the relationships.

In Figure 2, the plot shows a quadratic regression model fitted to the data (red curve), along with the scatter plot of the raw data points. The quadratic model captures the non-linear relationship between GDP per capita and corruption control and shows an upward trend with increasing GDP. The shaded area around the regression line represents the confidence interval, indicating that the model adequately represents the data: higher GDP per capita is associated with better corruption control, but the relationship is not strictly linear, as shown by the curvature of the regression line. Comparing the

normalized data set on the right-hand side, the graph shows a more linear trend, but even in this case a quadratic model could explain the relationship better.
The bivariate distribution of GDP per capita and corruption control shows the bivariate distribution using contour lines to represent the density of the data points. The contours indicate areas of higher density, suggesting that most data points are concentrated in certain regions. The highest density appears to correspond with higher levels of GDP per capita and control of corruption, reinforcing the trends seen in the regression plots. This graph highlights the joint distribution of the two variables and illustrates that corruption control tends to improve with increasing GDP, which is reflected in the density of points.
To test linearity, an ANOVA analysis was performed at a significance level of 0,05, comparing the linear regression model (analyzing a single variable per time) and the relative polinomial regression model (adding the quadratic term of the analyzed variables).
To account for this, the model was extended by introducing polynomial quadratic terms into the variable selection process when there was evidence of nonlinearity, which involved the following variables: `COR`,`ELE`,`INT`,`LAB`,`SFE` and `URB`. This can be used to test whether the relationship between the independent variables and the dependent variable follows a quadratic and not a purely linear form. The modified model can be expressed as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \cdots + \epsilon \tag{6}$$

The inclusion of these polynomial terms $X_1^2, X_2^2, \ldots$ helps the model to more accurately reflect the underlying relationships in the data, allowing to better capture the non-linear dynamics and improve the explanatory power of the model.

Secondly, In order to perform an automatic selection of variables, **Stepwise Forward Selection method** was employed using Akaike Information Creterion (AIC). Forward selection is a stepwise regression technique used to select a subset of variables that optimally explain the variation in the dependent variable by adding predictors one by one based on specific criteria. The objective is to build a parsimonious model by retaining only the most relevant variables. It is a useful technique for building multiple linear regression models when there are numerous predictors. However, it should be applied carefully to avoid overfitting and misinterpretation. Forward selection starts with the simplest model (the null model) and adds predictors sequentially. The process involves the following steps:

1. **Null Model:** the null model contains only the intercept:

$$Y = \beta_0 + \epsilon \tag{7}$$

2. **Best Predictor:** out of all possible predictors, select the one that, when added

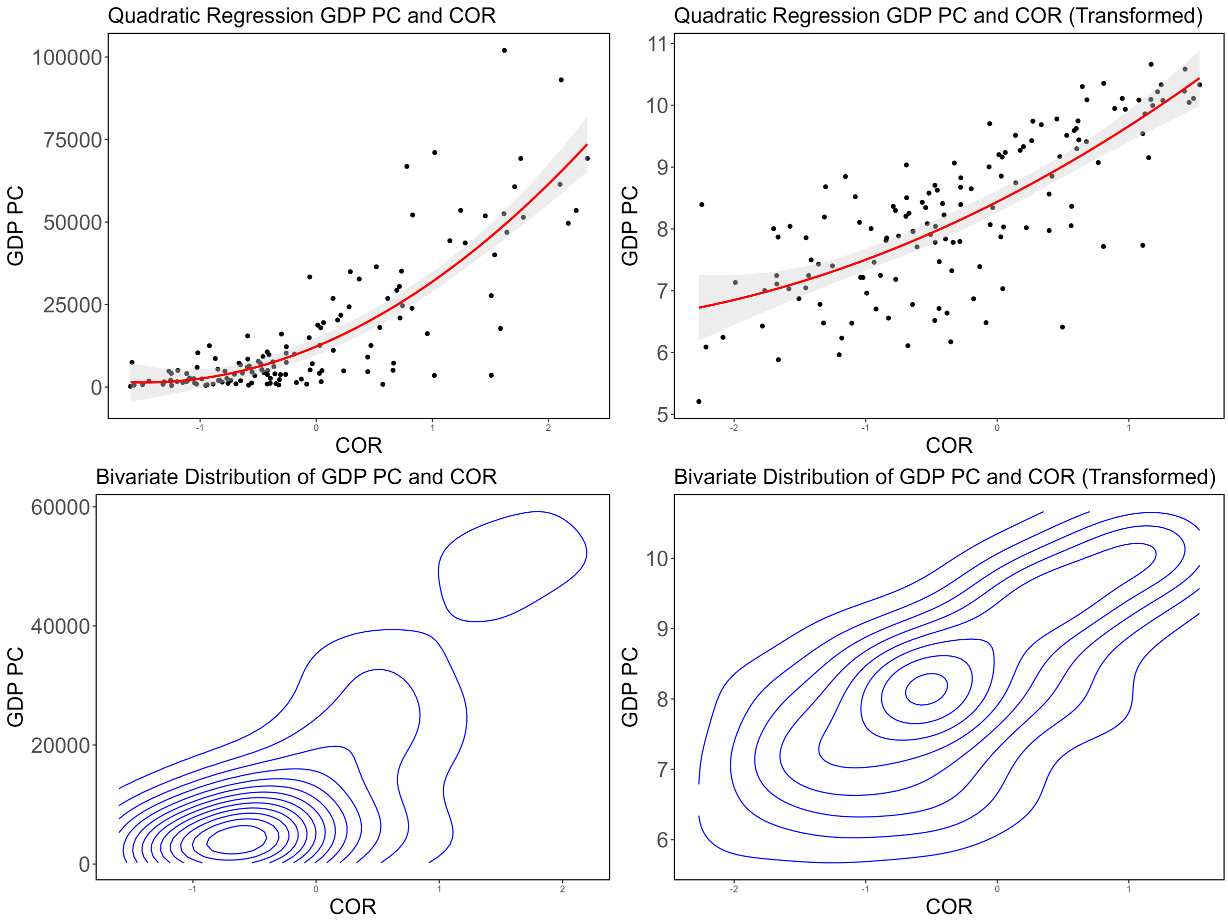


Figure 2: This figure illustrates the relationship between GDP per Capita and Corruption Control for linearity check. Original dataset (on left-side) is compared to normalized dataset (on right-side).

to the model, results in the greatest reduction in AIC: [3]

$$\mathrm{AIC} = n \log\left(\frac{\mathrm{RSS}}{n}\right) + 2p \tag{8}$$

where n is the sample size, RSS is the residual sum of squares, and p is the number of parameters (including the intercept).

For each predictor X_j, the new model is evaluated as:

$$Y = \beta_0 + \beta_j X_j + \epsilon \tag{9}$$

[3] In regression analysis, the Residual Sum of Squares (RSS) quantifies the discrepancy between observed data points and the predictions made by a model. It is calculated as the sum of the squared differences between the observed values y_i and the predicted values $\hat{y}_i$:

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where:
y_i represents the actual observed values,
$\hat{y}_i$ represents the values predicted by the regression model,
n is the total number of observations.

Minimizing the RSS is crucial in regression as it indicates the best fit, where the model's predictions are as close as possible to the actual data points. A lower RSS reflects a model that better explains the variability in the dependent variable.

The predictor with the lowest AIC is added to the model.

3. Repeat until stopping: continue adding the next best predictor until no significant improvement is experienced in model fit based on the selected criterion, then stop the process.

The output of Linearity check (with a significance level of 0,05) and Stepwise Forward Selection Method (with a significance level of 0,01), was the following model:

$$\begin{aligned} \text{GDP PC} = 24360 &+ 8477 \cdot \text{COR} + 6176 \cdot \text{COR}^2 \\ &- 721,9 \cdot \text{SFE} + 5,303 \cdot \text{SFE}^2 \\ &+ 434 \cdot \text{FDI} + \varepsilon \end{aligned} \tag{10}$$

Thirdly, a **normal distribution test of residuals** was conducted on the model obtained employing the original dataset. To this extent, the Shapiro-Wilk test was performed. The test is based on the comparison of the observed distribution of the data with a normal distribution and is particularly useful for small sample sizes. The null hypothesis of the Shapiro-Wilk test states that the data is normally distributed:

$$H_0 : \text{The data is normally distributed.}$$

The test statistic W is calculated as follows:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{11}$$

where $x_{(i)}$ are the ordered sample values, $\bar{x}$ is the sample mean, and a_i are the weights derived from the expected values of the order statistics from the standard normal distribution.

The p-value is computed based on the test statistic W. A low p-value indicates strong evidence against the null hypothesis. Typically, a significance level of $\alpha = 0,05$ is used, such tat:
If $p < \alpha$: Reject the null hypothesis, suggesting the data is not normally distributed.
If $p \geq \alpha$: Fail to reject the null hypothesis, suggesting the data may be normally distributed.

In this analysis of normality on residuals of the original dataset, the results of the Shapiro-Wilk normality test were as follows:

$$\text{W} = 0,82935, \quad \text{p-value} = 8,553 \times 10^{-12}$$

Given that W statistic (0,82935) is much smaller than 1 and that the p-value is lower than the typical threshold of 0,05, the results suggest that the null hypothesis of normality is rejected, indicating a noticeable deviation from normality.

To solve this issue, Yeo-Johnson transformation was employed:

$$y^{(\lambda)} = \begin{cases} \frac{(y+1)^\lambda - 1}{\lambda} & \text{if } y \geq 0 \text{ and } \lambda \neq 0 \\ \ln(y+1) & \text{if } y \geq 0 \text{ and } \lambda = 0 \\ \frac{(1-y)^{2-\lambda} - 1}{2-\lambda} & \text{if } y < 0 \text{ and } \lambda \neq 2 \\ -\ln(1-y) & \text{if } y < 0 \text{ and } \lambda = 2 \end{cases} \tag{12}$$

where:
y is the original value,
$y^{(\lambda)}$ is the transformed value,
λ is the transformation parameter.

A new model has then been generated employing the normalized dataset. At first, the Linearity check (with a significance level of 0,05) showed evidence of non-linearity for the following variables: `COR`, `EXP`,`INT` and `SFE`. Hence, the polinomial quadratic terms were added.
Secondly, Stepwise Forward Selection Method was performed (with a significance level of 0,01), the output of this variable selection process was a model composed by:

$$\begin{aligned} \text{GDP PC} &= 7,304 + 0,4590 \cdot \text{COR} + 0,1727 \cdot \text{COR}^2 \\ &\quad - 0,1153 \cdot \text{SFE} + 0,06674 \cdot \text{ELE} \\ &\quad + 0,003903 \cdot \text{URB} + 0,001975 \cdot \text{INT} + \varepsilon \end{aligned} \tag{13}$$

The Shapiro-Wilk test was performed again on the residuals of the new model obtained from the normalized dataset:

$$\text{W} = 0,98843, \quad \text{p-value} = 0,2627$$

The results of the test suggest that the test statistic is very close to 1, indicating that the residuals are likely close to a normal distribution, morevoer the p-value is much larger than the typical threshold of 0,05, which fails to reject the null hypothesis. This suggests that the residuals of the new model are consistent with a normal distribution.

It is therefore reasonable to conclude that the application of the Yeo-Johnson transformation allows the assumption of normality of residuals to be satisfied in this analysis. In this regard, it is crucial to note that the normalized dataset enhances the normality of variables, potentially improving the performance and interpretability of linear regression models. However, the transformation may complicate the interpretation of variables in their original scale or context.

Figure 3 shows the comparison of residuals distributions between original dataset (left-side graphs) and normalized dataset (right-side graphs), for which Yeo-Johnson transformation was employed. The evidences show how transformed dataset seems to be closer to normal distribution than original dataset (that is skewed to the left), hence being more suitable for inference purposes.
[4]

Subsequently, **outlier detection** was performed. To this extent, the Z-score technique was introduced. This technique measures how distant a particular data point is from the mean, in terms of standard deviations, detecting potential outliers in datasets. Given a dataset, the Z-score of an individual data point is calculated as:

$$Z = \frac{X - \mu}{\sigma} \tag{14}$$

where:
Z is the Z-score,
X is the value of the data point,

[4] From now on, the Mulptiple Linear Regression analysis will employ the normalized dataset.

μ is the mean of the dataset,
σ is the standard deviation of the dataset.

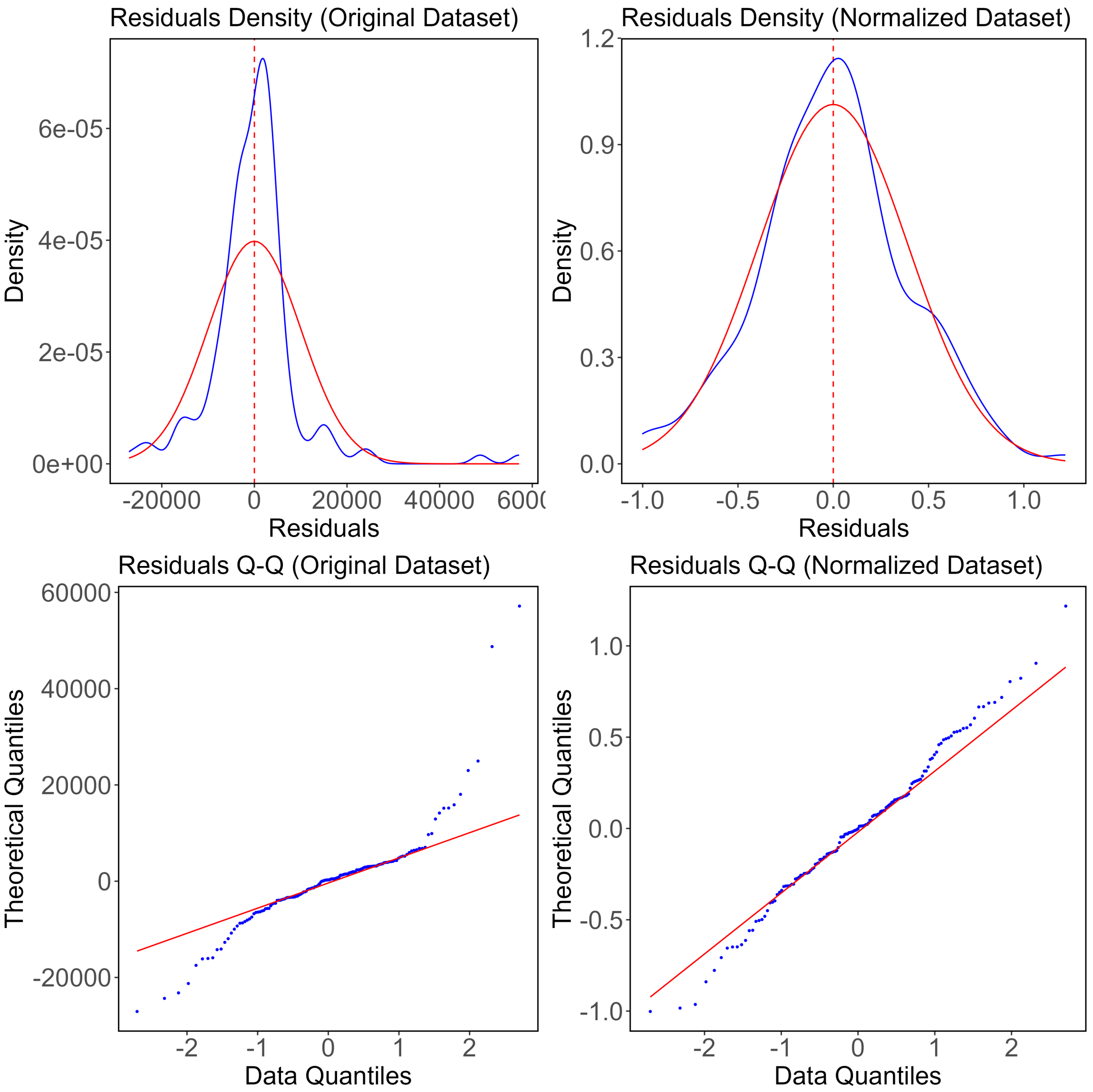


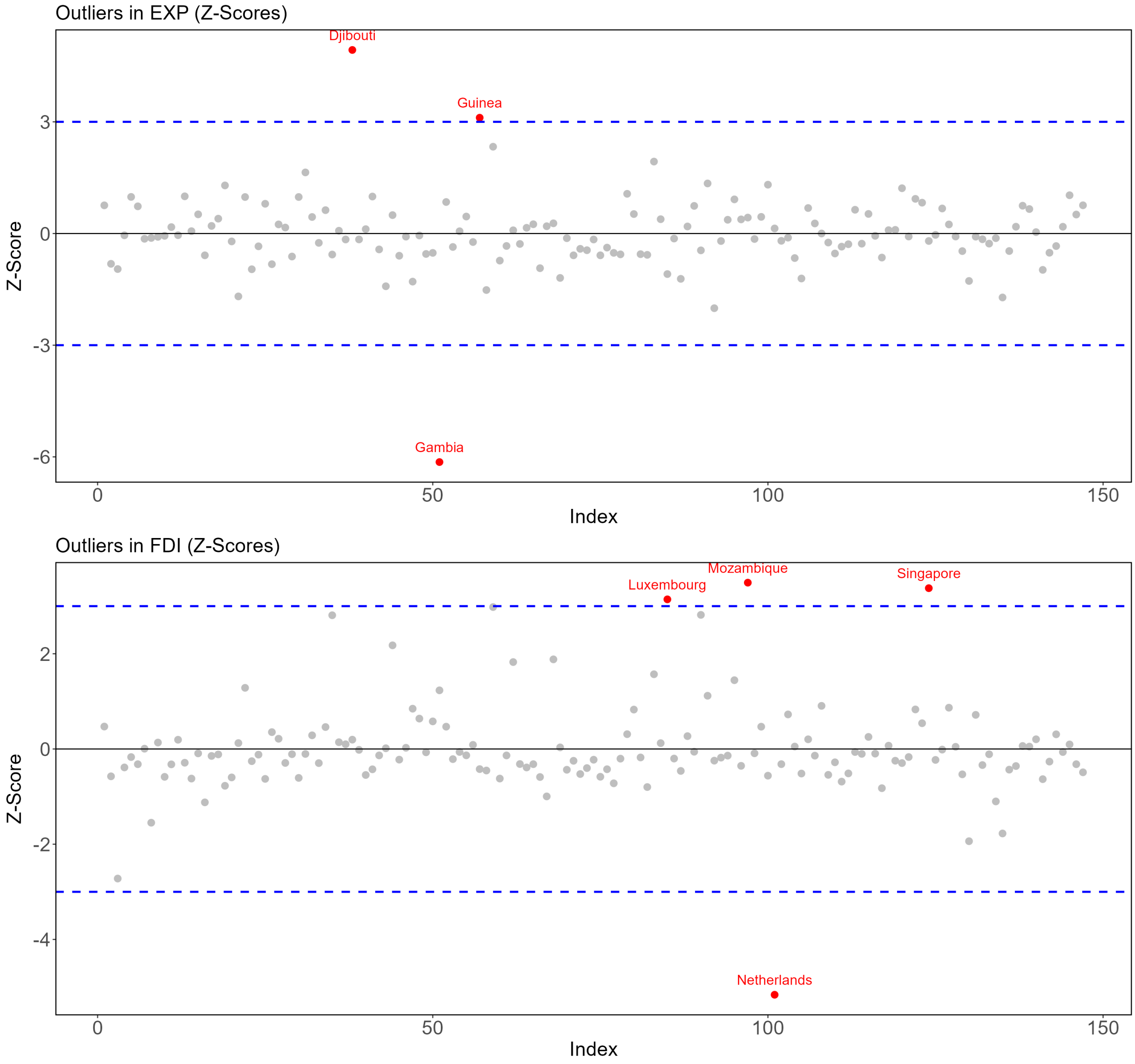
Figure 3: This figure illustrates the distribution of residuals for normalized dataset (on left-side) and normalized dataset (on right-side). The data lines in blue are compared to normal distributions in red.

In the context of outlier detection, a data point is considered an outlier if its absolute Z-score exceeds a certain threshold. A common threshold is $Z > 3$ or $Z < -3$, which means that the data point is more than 3 standard deviations away from the mean. Mathematically, outliers are defined as those points for which:

$$|Z| > 3 \tag{15}$$

Outliers can either be positive (above the mean) or negative (below the mean), and the threshold can be adjusted depending on the specific application or distribution of the data. The Z-score method assumes that the data is approximately normally distributed. When the distribution is not normal, this method may not be appropriate or may need adjustment.

The advantage of using Z-score for outlier detection is that it standardizes the data, allowing comparison across different datasets or variables, even when they have different units or scales.


Outliers in EXP (Z-Scores)

Outliers in FDI (Z-Scores)

Figure 4: This figure illustrates the variables for which outlies were identified using Z-scores. Data above or below the threshold values (dashed blue lines), which are more than 3 standard deviations away from the mean, are highlighted in red.

As shown in Figure 4, the diagnostics reported impacts for only 2 variables, the following outliers were detected: `EXP` (Djibouti, Guinea and Gambia), `FDI` (Louxemburg, Mozamique, Singapore and Netherlands). As a result, those countries were completely excluded from the analysis.

Some considerations were made regarding the assumption of **Independence and Identical Distribution (IID)** of random variables. In this context, it is important to note that data collected at the country level often do not satisfy the IID assumption. Geographic proximity and similarities in economic or social structures between countries

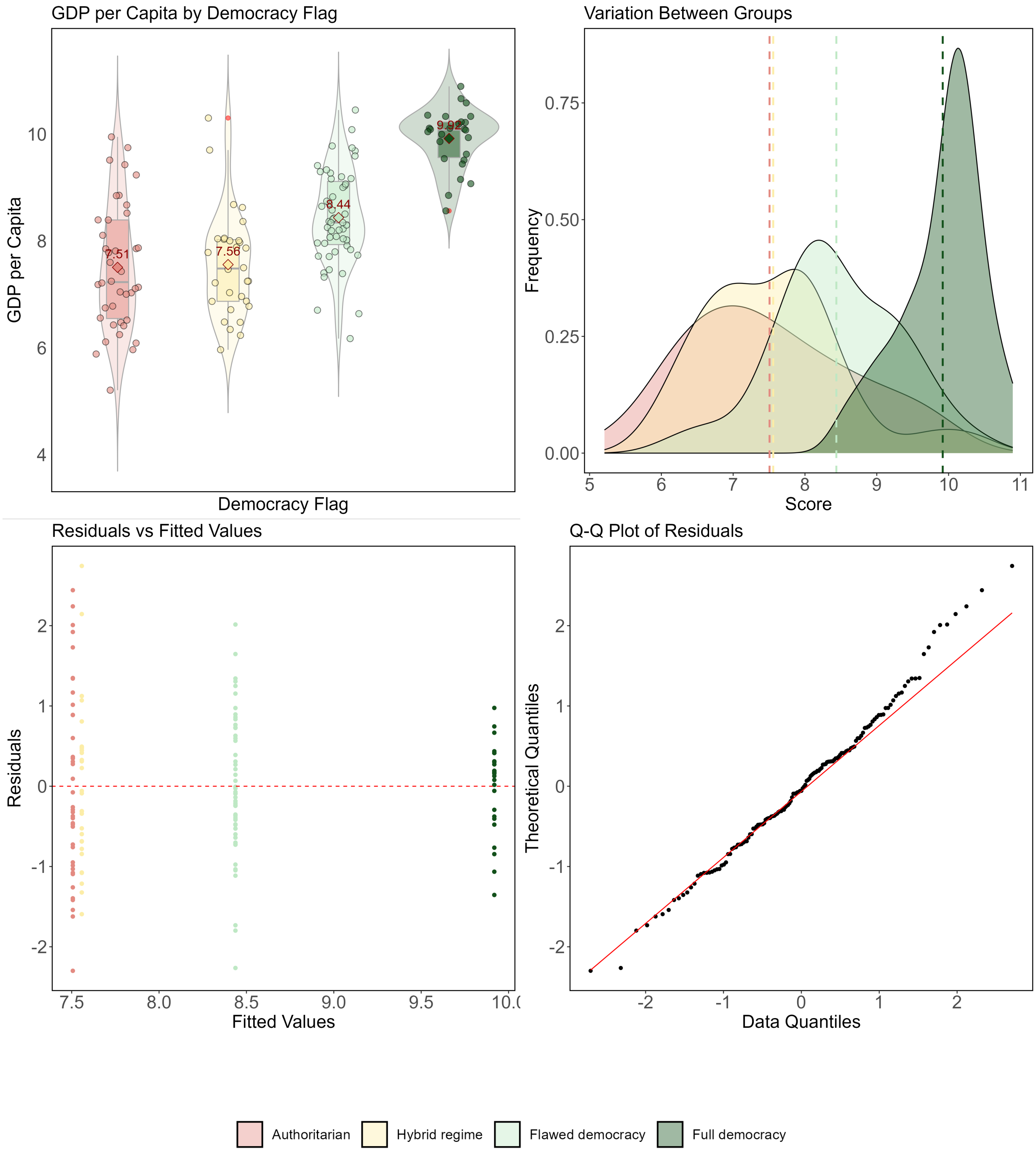


Figure 5: This figure illustrates the graphs used to support ANOVA analysis on IID Assumption check, in order to introduce Democratic Flag categorical variable into the model.

may violate the assumption of independence between observations. To address this issue, the model was enhanced with the inclusion of a categorical variable, the ”Democracy Flag.” This allows individuals within similar structural categories to be grouped, while maintaining independence between different country categories. Analysis of Variance (ANOVA) was conducted to assest the difference between groups and introduce the categorical variable into regression. Ideed, ANOVA tests the hypothesis that the means of several groups are equal. In this case, the variable `DEM FLAG` represents 4 different categories of government systems, the objective is to verify if these categories have different means in terms of the dependent variable `GDP PC`.

$$H_0 : \text{All group means are equal}(\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k)$$

$$H_1 : \text{At least one } \mu_i \text{ is different}$$

In Table 1 ANOVA results are reported, showing the variances between and within groups. The key elements are computed as follows:

Degrees of Freedom (Df): For the factor (e.g. `DEM FL`): $k-1$, where k is the number of groups. For the residuals: $N-k$, where N is the total number of observations.

Total Sum of Squares (SST): summarize the Sum of Squares Between groups (for `DEM_FL`) and the Sum of Squares Within groups (for Residuals), as follows:

$$SST = SSB + SSW \tag{16}$$

where:

$$\text{SSB} = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2 \tag{17}$$

$$\text{SSW} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 \tag{18}$$

Mean Squares (MS): is the ratio between Sum of Squares Between (for `DEM_FLAG`) and Within (for Residuals) and the realtive Degrees of Freedom, computed as follows:

$$\text{MSB} = \frac{\text{SSB}}{Df_{\text{between}}} \tag{19}$$

$$\text{MSW} = \frac{\text{SSW}}{Df_{\text{within}}} \tag{20}$$

F-value: The F-statistic is the ratio of the variance estimates:

$$F = \frac{\text{MSB}}{\text{MSW}} \tag{21}$$

p-value: The p-value indicates the probability of observing the data, given that the null hypothesis is true. A small p-value (typically $<$ 0,05) leads to the rejection of the null hypothesis.

Table 1: ANOVA Results for the Effect of `DEM_FL` on GDP PC

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
`DEM_FLAG`	3	93,33	31.111	35,63	< 2e-16 ***
`Residuals`	136	118,74	0,873	-	-

Significance codes: *** $p < 0,001$, ** $p < 0,01$, * $p < 0,05$, . $p < 0,1$, ' ' non-significant

The interpretation of the ANOVA results indicate that the `DEM FLAG` categorical variable significantly affects GDP per Capita, as evidenced by the high F-value and the extremely low p-value. More visual insights can be shown in Figure 5. The Box-Plot shows a clear trend suggesting that Full Democracies have the highest GDP per capita, while Authoritarian regimes tend to have lower economic development levels. The Violin plot reinforces the box plot findings, showing a higher density of GDP per capita in Full Democracies and a more spread-out distribution in Authoritarian regimes. Moreover, the Q-Q plot points do not significantly deviate from the line, suggesting normality.

Finally, regarding **Multicollinearity**, the correlation matrix was analyzed. to this extent, the threshold employed for identifying problematic multicollinearity is a correlation coefficient above 0,7 or below -0,7. The results shown in Figure 6 suggest 4 instances of multicollinearity, particularly involving the variables `COR`, `POL`, `SFE`, `INT` and `URB`. Regarding the correlation between `COR` and `POL`, it denotes the fact that more corrupted countires are usually expected to be more ploitically unstable. In order to address this issue, the `POL` variable was dropped off. For the rest, Variance Inflation Factor (VIF) and Generalized VIF (GVIF) values have been calculated for each independent variable in the model. The formula for VIF for a variable i is given by:

$$\text{VIF}_i = \frac{1}{1 - R_i^2} \tag{22}$$

where R_i^2 is the coefficient of determination obtained by regressing variable i against all other independent variables.
GVIF is an index adjusted for the degrees of freedom (denoted by $GVIF^{1/(2Df)}$), hence it provides a more interpretable metric for assessing multicollinearity. The followingh pattern is used for interpretation of GVIF values: VIF = 1: no multicollinearity, $1 <$ VIF < 5: moderate multicollinearity, VIF ≥ 5 or 10: high multicollinearity.

Table 2: GVIF Interpretation for each variable

Variable	GVIF	Interpretation
INT	4,40 ($GVIF^{1/(2Df)} = 2,09$)	Moderate multicollinearity; monitoring advised.
DEM FLAG	3,76 ($GVIF^{1/(2Df)} = 1,25$)	Low multicollinearity, no immediate concern.
EMP	3,69 ($GVIF^{1/(2Df)} = 1,92$)	Moderate multicollinearity, below threshold.
URB	2,48 ($GVIF^{1/(2Df)} = 1,58$)	Mild multicollinearity, not an issue.
ELE	1,55 ($GVIF^{1/(2Df)} = 1,25$)	Low multicollinearity, stable predictor.
COR	4,53 ($GVIF^{1/(2Df)} = 2,13$)	Moderate multicollinearity; monitoring advised.

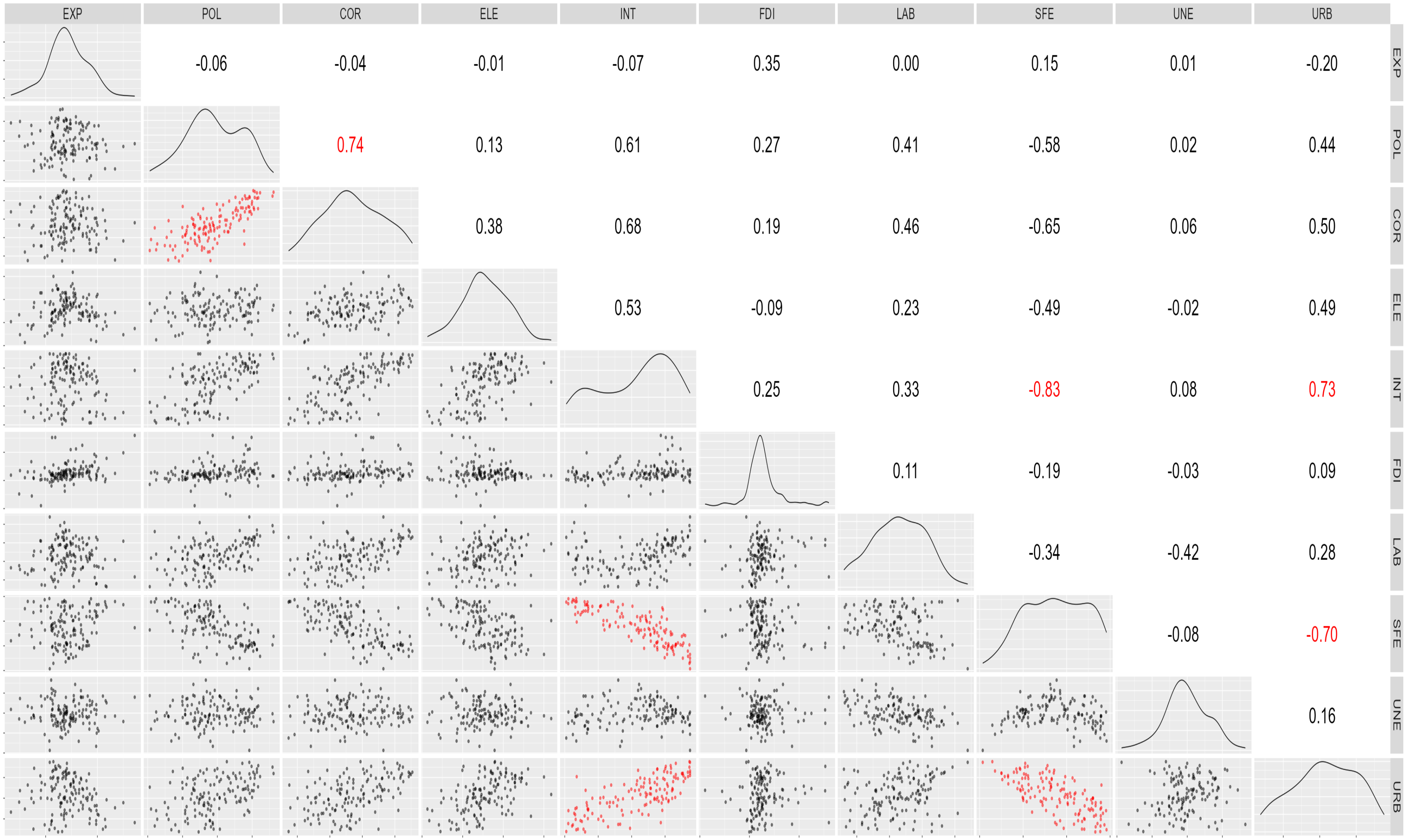


Figure 6: This Image reports the correlation index for each couple of variables in the matrix. In the lower side of the matrix are reported the scatterplots on which correlations are computed. Red values and scatterplots point out a correlation value above 0,7 or below -0,7, which can potentially detect multicollinearity. The diagonal reports the bivariate distribution for each couple of variables.

Overall, the model shows moderate multicollinearity, especially for `INT` and `COR`. However, none of the corrected VIF values exceed 5, which is often considered a threshold for concern. Therefore, while multicollinearity is present, it is not severe. Nevertheless, attention should be given to variables with higher GVIFs, as they may affect the model's interpretability and coefficient estimates.

Diagnostic on Residual Values

In Figure 7 it is possible to interpret residual diagnostic plots obtained by Model (13), which is essential for validating the assumptions of multiple linear regression model above.
1. Residuals vs. Fitted: the plot shows the residuals (the difference between observed and predicted values) against the fitted values (predicted values), with the horizontal red line indicating the zero residual level. In this analysis, residuals are randomly scattered around zero with no discernible pattern, suggesting that the variances of the error terms are equal. This implies that the model has adequately captured the relationship between the variables. If the residuals exhibited a pattern, such as a curve, it could indicate that the model was missing important predictors or that the relationship between variables was not linear. Two more outliers were detected in this plot (”Equatorial Guinea” and “Jordan”).
2. Q-Q Residuals: the Quantile-Quantile plot was used to compare the quantiles of the residuals against the quantiles of a normal distribution. Data points were mostly close to the line, especially in the center, suggesting that the residuals were approximately normally distributed.
3. Scale-Location Plot (Spread-Location): this plot was used to show the square root of the standardized residuals against the fitted values, to assess the homoscedasticity (constant variance) of the residuals. Ideally, the points should be randomly scattered without a systematic pattern. In this research, a noticeable inverted U-shape curve suggests that the variance of the residuals increases and decreases with the fitted values, indicating possible heteroscedasticity, which would violate the assumption of constant variance in linear regression. This issue will be analyzed in detail employing the Breusch-Pagan test.
4. Residuals vs. Leverage: the plot was used to compare the standardized residuals to leverage values (a measure of the distance of an observation to the center of the data). Three leverage points above the Cook's distance (dotted line) were detected. Those data points (”Equatorial Guinea”, ”Burundi” and ”Kyrgyzstan”) may influence the slope of the regression line, affecting the results of the analysis.

Normal Distribution of Residuals: the normality of residuals has already been checked in a previous step, after Outliers elimination, the Shapiro-Wilk normality test still shows evidences of normal distribution on residuals:

$$\text{W} = 0,9908, \quad \text{p-value} = 0,492$$

Since the p-value is greater than the significance level of 0,05, the null hypothesis cannot be rejected. This indicates that the residuals from the linear regression model appear to be normally distributed based on the Shapiro-Wilk test.

Exogeneity of Residuals: the following table shows the correlations between the residuals from the regression model and the independent variables.

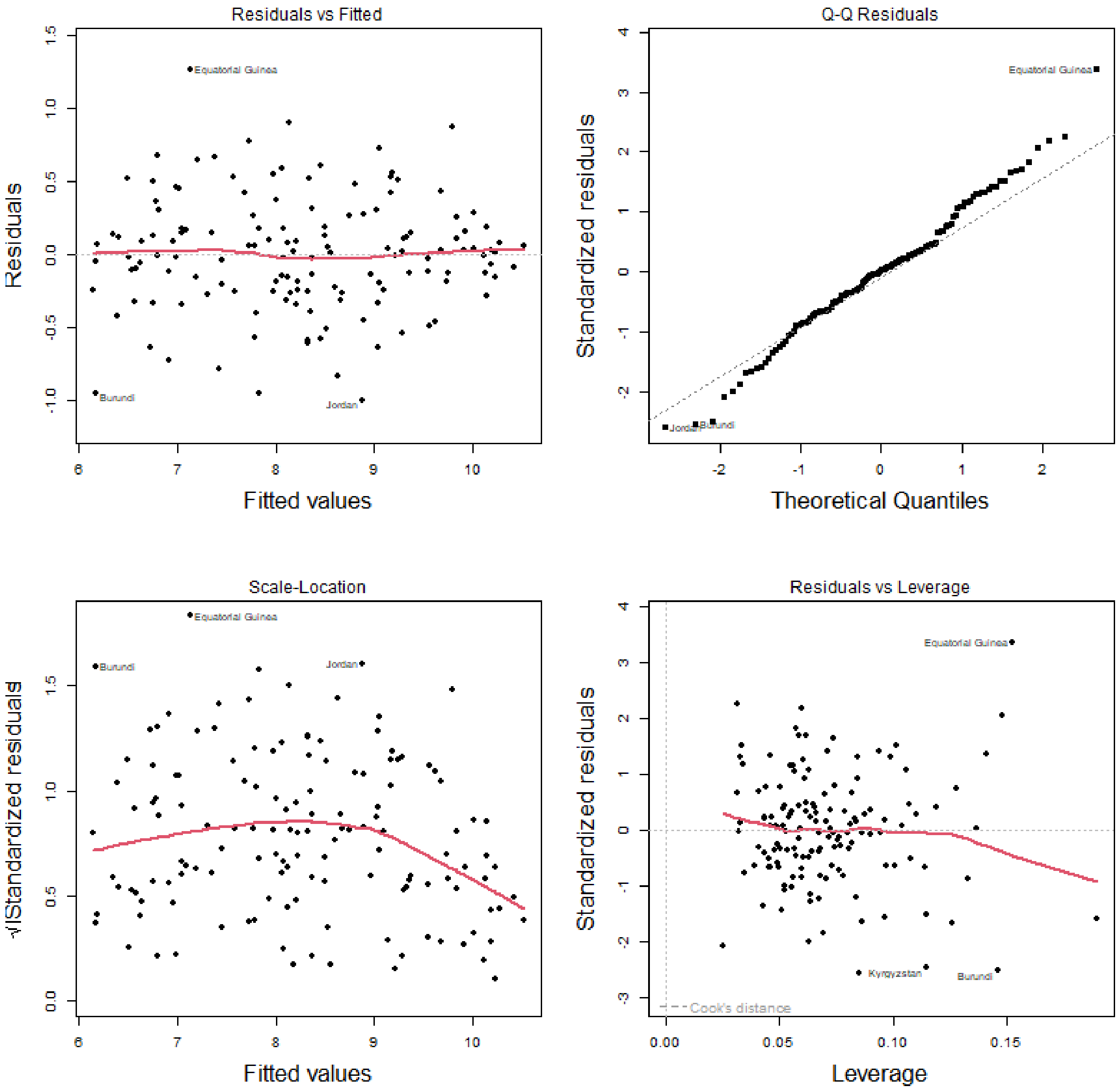


Figure 7: This figure illustrates the plots used to support residual analysis for the diagnostics post regression.

All correlations are extremely close to zero, which supports the assumption of exogeneity. Therefore, the independent variables are not correlated with the error term, and the OLS estimates are unbiased and consistent.

Autocorrelation of Residuals: the Durbin-Watson test was conducted to detect the presence of autocorrelation in the residuals from a regression analysis. Autocorrelation occurs when the residuals from one observation are correlated with the residuals from another observation. Such an evidence would violate the assumption of independence in OLS regression, leading to inefficient estimates and incorrect inferences. The Durbin-Watson statistic, denoted by d, is calculated as:

Table 3: Correlations between Independent Variables and Residuals

Variable	**Correlation with Residuals**
COR	$6,45 \times 10^{-18}$
SFE	$8,65 \times 10^{-17}$
ELE	$1,25 \times 10^{-17}$
URB	$5,85 \times 10^{-16}$
INT	$1,58 \times 10^{-17}$

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{23}$$

where:
e_t is the residual at time t,
e_{t-1} is the residual at time $t-1$,
n is the number of observations.

The Durbin-Watson statistic ranges from 0 to 4, with the following interpretation: $d = 2$ indicates no autocorrelation, $d < 2$ suggests positive autocorrelation (residuals are positively correlated) and $d > 2$ suggests negative autocorrelation (residuals are negatively correlated). The null hypothesis (H_0) of the Durbin-Watson test is that there is no autocorrelation $\rho = 0$, where ρ is the autocorrelation coefficient. The alternative hypothesis (H_1) is that autocorrelation exists ($\rho \neq 0$).

$$H_0 : \rho = 0 \quad \text{(No autocorrelation)}$$
$$H_1 : \rho \neq 0 \quad \text{(Autocorrelation exists)}$$

Lag 1 Autocor: $= 0,08322849$, Durbin-Watson St.(D-W): $= 1,82779$, p-value: $= 0,308$

The Durbin-Watson statistic is close to 2, which suggests little to no autocorrelation in the residuals. The p-value of 0,308 is higher than the common significance level (0,05), meaning that the null hypothesis cannot be rejected and invalidating any autocorrelation evindence in the residuals.

Heteroscedasticity of Residuals: the Breusch-Pagan test is a statistical test used to detect heteroscedasticity in a regression model. The null hypothesis (H_0) of the Breusch-Pagan test is that the residuals have constant variance (homoscedasticity), while the alternative hypothesis (H_1) is that the residuals have non-constant variance (heteroscedasticity).

$$H_0 : \sigma_i^2 = \sigma^2 \quad \forall i \quad \text{(Homoscedasticity)}$$
$$H_1 : \sigma_i^2 \neq \sigma^2 \quad \text{for some } i \quad \text{(Heteroscedasticity)}$$

where σ_i^2 is the variance of the residuals for the i-th observation.
The test statistic for the Breusch-Pagan test is derived from the regression of the squared residuals on the independent variables. The steps are as follows:

1. Fit the linear regression model;

2. Obtain the residuals $\hat{\varepsilon}$ from the fitted model;
3. Calculate the squared residuals $\hat{\varepsilon}^2$.
4. Regress the squared residuals on the independent variables:

$$\hat{\varepsilon}^2 = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_k X_k + u \tag{24}$$

5. The test statistic is calculated as:

$$BP = \frac{nR^2}{2} \tag{25}$$

where:
n is the number of observations
R^2 is the coefficient of determination from the regression of squared residuals.

The Breusch-Pagan test statistic follows a chi-squared distribution with degrees of freedom equal to the number of independent variables in the model k:

$$BP \sim \chi^2(k) \tag{26}$$

The results of the Breusch-Pagan test indicate the following:

$$\text{BP Statisticr:} = 17,983, \quad \text{Degrees of Freedom (df):} = 9, \quad \text{p-value:} = 0,03537$$

Since the p-value is less than 0,05, the null hypothesis is to be rejected. This suggests that there is significant evidence of heteroscedasticity in the residuals of the linear regression model, meaning that the residuals have non-constant variance. This can lead to inefficient estimates of the coefficients and can affect hypothesis tests, potentially resulting in unreliable confidence intervals and p-values. In order to address Heteroskedasticity, Outliers and Leverage Points detected in the plots of Figure 7 (”Equatorial Guinea”, ”Jordan”, ”Kyrgyzstan”, ”Burundi”) have been dropped off.

The results of the Breusch-Pagan test, after the processing of Outliers and Leverage Points detected in the previous step, indicate the following:

$$\text{BP Statisticr:} = 15,713, \quad \text{Degrees of Freedom (df):} = 11, \quad \text{p-value:} = 0,1521$$

As a consequence, the new Breusch-Pagan test p-value result (0,1521) is greater than 0,05, hence (H_0) is not to be rejected. This means that there is not enough evidence to support the presence of heteroscedasticity in the residuals of the multiple linear regression model, therefore homoscedasticity of residuals assumption is respected.

Analysis Results

In the following table, the final result of the Multiple linear regression model, run after post-regression diagnostics, using forward stepwise selection method is reported:

The linear regression model can be expressed at 0,05 significance level as:

$$\begin{aligned} \text{GDP PC} = \beta_0 &+ 0,2737 \cdot \text{COR} + 0,1145 \cdot \text{COR}^2 \\ &- 0,007559 \cdot \text{SFE}^2 + 0,06038 \cdot \text{ELE} \\ &+ 0,003226 \cdot \text{URB} + \varepsilon \end{aligned} \tag{27}$$

Table 4: Regression Results: Dependent Variable - Final Model

Variable	Estimate	Std. Error	t-value	p-value	
Intercept	7,403	3,116e-01	23,758	< 2e-16	***
I(SFE²)	-7,559e-03	1,049e-03	-7,204	4,98e-11	***
I(INT²)	1,836e-07	4,361e-07	0,421	0,674425	
DEM FLAG: Hybrid regime	5,809e-02	9,814e-02	0,592	0,555015	
DEM FLAG: Flawed democracy	2,903e-01	9,803e-02	2,961	0,003678	**
DEM FLAG: Full democracy	4,246e-01	1,582e-01	2,685	0,008256	**
ELE	6,038e-02	1,718e-02	3,516	0,000613	***
URB	3,226e-03	1,202e-03	2,683	0,008304	**
COR	2,737e-01	7,731e-02	3,540	0,000564	***
I(COR²)	1,145e-01	4,807e-02	2,383	0,018707	*
UNE	-1,083e-01	5,744e-02	-1,886	0,061652	.
INT	7,309e-04	5,061e-04	1,444	0,151190	
Residual standard error:	0,3664 on 124 degrees of freedom				
Multiple R-squared:	0,917				
Adjusted R-squared:	0,9096				
F-statistic:	124,5 on 11 and 124 DF, *p*-value: < 2,2e-16				

Significance levels: *** $p < 0,001$, ** $p < 0,01$, * $p < 0,05$, . $p < 0,1$

where the intercept β_0 is based on the category `DEM FLAG` as follows: $7,403$ for Authoritarian and Hybrid Regimes, $7,6933$ for Flawed Democracies (at a 1% significance level), and $7,8276$ for Full Democracies (at a 1% significance level). This indicates that, when independent variables are set to zero, the intercept varies depending on the governmental system, showing higher economic development levels in democratic systems. Hybrid Regime is not statistically significant, meaning there is no strong evidence that they significantly differs from Authoritarian regimes. Otherwise, Flawed Democracy and Full Democracy are significant, indicating that those countries have a different intercept with respect to Authoritarian regimes. Corruption Control and Self Employment exhibit a significant quadratic relationship with respect to GDP per capita (respectively: positive and negative relationship), while Electricity Consumption and Urban Population show a positive linear relationships. Among these, Corruption Control appears to be the most influential factor on Economic Development in this analysis.
The model fit statistics indicate that approximately 91,17% of the variance in the dependent variable is explained by the independent variables (Multiple R^2), with strong significance (p-value $< 2,2 \times 10^{-16}$). The Adjusted R-squared 90,96% accounts for the number of predictors in the model and remains high, confirming that the model is not overfitting due to the number of variables. This is a very high proportion, suggesting that the model has a strong fit to the data.The F-statistic (124,5) with a highly significant p-value (< 2,2e-16) indicates that the overall model is statistically significant.

Ridge and Lasso Regression

The Stepwise Forward selection method used in multiple linear regression, can be very useful to improve prediction accuracy in certain cases, such as when only a few covariates have a strong relationship with the outcome, however it may increase prediction error in other cases. The Ridge and Lasso regression methods are both regularization techniques used to address problems such as multicollinearity or curse of dimensionality in linear regression models and to enhance prediction accuracy. These methods add a penalty term to the OLS regression objective function, which helps prevent overfitting and improves model interpretability. Lasso forces the sum of the absolute value of the regression coefficients to be less than a fixed value, reducing certain coefficients to zero and performing variable selection. Ridge regression, which also shrinks the size of the coefficients performing regularizzation, does not reduce coefficients to zero (it does not perform variable selection). While both Ridge and Lasso regression methods aim to reduce overfitting by applying penalties, they serve different purposes.

Ridge regression: is particularly useful when there are many small/medium-sized effects, as it retains all predictors but shrinks their coefficients. It modifies the ordinary least squares objective function by adding a L_2 penalty. The objective function for Ridge regression is defined as:

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{28}$$

where y_i is the dependent variable, X_i is the vector of independent variables, β is the vector of coefficients and $\lambda \geq 0$ is the regularization parameter that controls the strength of the penalty. As λ increases, the coefficients β are shrunk towards zero, which reduces model complexity and multicollinearity, leading to more stable estimates.

Lasso regression: is beneficial when there are many predictors, some of which may not be relevant, as it tends to exclude them from the model entirely. It uses a L_1 penalty, which promotes sparsity in the coefficient estimates. The objective function for Lasso regression is defined as:

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{29}$$

In this equation, the terms are defined similarly to those in Ridge regression. The main difference lies in the penalty term, where the L_1 norm forces some coefficients to be exactly zero, effectively performing variable selection.

The Ridge and Lasso regression methods were applied to the normalized dataset, employing cross-validation to point out the best λ parameter.

Figure 8 shows the cross-validated Mean Squared Error (MSE) as a function of the regularization parameter λ, represented in logarithmic scale $\log(\lambda)$. The red dots indicate the MSE at different values of λ, while the dotted vertical lines mark the optimal values of λ based on the model's performance. [5] In Table 5 are reported the key results from the two models. The coefficients reflect the magnitude and direction of each predictor's effect on the dependent variable, `GDP PC`.

[5] the first vertical line corresponds to the λ value that minimizes the MSE, representing the best-performing model. The second vertical line indicates the largest λ within one standard error of the minimum MSE, which provides a simpler model with comparable performance.

Table 5: Comparison of Ridge and Lasso Regression Coefficients

Variable	**Ridge Coefficient**	**Lasso Coefficient**
Intercept	8,8831	7,283
DEM_INDEX	0,0709	0,0741
POL	0,1814	0,1554
COR	0,1273	0,1066
ELE	0,0853	0,0713
INT	0,0008	0,0009
LAB	$4,4512 \times 10^{-6}$	$2,1606 \times 10^{-6}$
SFE	-0,1029	-0,1191
URB	0,0036	0,0030
EXP	-0,8698	0
FDI	0,0114	0
UNE	-0,0103	0
Lasso Log(λ): -3,4693		
MSE:	0,1280	
Residual Standard Error:	0,3754 on 124 degrees of freedom	
R-Squared:	0,9128	
Adjusted R-Squared:	0,9058	
F-Statistic:	118,0423 on 124 DF, p-value $< 3,6274 \times 10^{-60}$	
Ridge Log(λ): -1,8644		
MSE:	0,1274	
Residual Standard Error:	0,3745 on 124 degrees of freedom	
R-Squared:	0,9137	
Adjusted R-Squared:	0,9058	
F-Statistic:	119,4974 on 124 DF, p-value $< 1,8212 \times 10^{-60}$	

The Ridge and Lasso regression models showed comparable predictive performances, as reflected in their metrics. Ridge achieves a marginally lower Mean Squared Error (MSE) of 0,1274 compared to 0,1280 for Lasso, indicating a slight edge in accuracy. Similarly, the R-squared values are very close, with Ridge explaining 91,37% of the variance and Lasso explaining 91,28%. Both models reported similar residual standard errors, with 0,3745 for Ridge and 0,3754 for Lasso. Regarding the intecepts, the models showed different intercept values, with the Ridge model showing a higher estimate (8,8831), while the Lasso model's intercept (7,283) is the lowest.

As expected, Ridge Regression included all predictors, even those with minimal coefficients, as the model does not shrink coefficients to zero. This resulted in smaller but retained coefficients for variables such as `LAB`, `FDI`, and `EXP`. On the contrary, Lasso Regression selectively removed variables by setting coefficients to zero for `EXP`, `FDI`, and `UNE`, indicating that these variables are not useful in this model. The Ridge Regression output suggests that variables related to governance quality (`COR` and `POL`) and infrastructure (`ELE`) have a positive relationship with the dependent variable, whereas factors like unemployment and self-employment show a negative association.

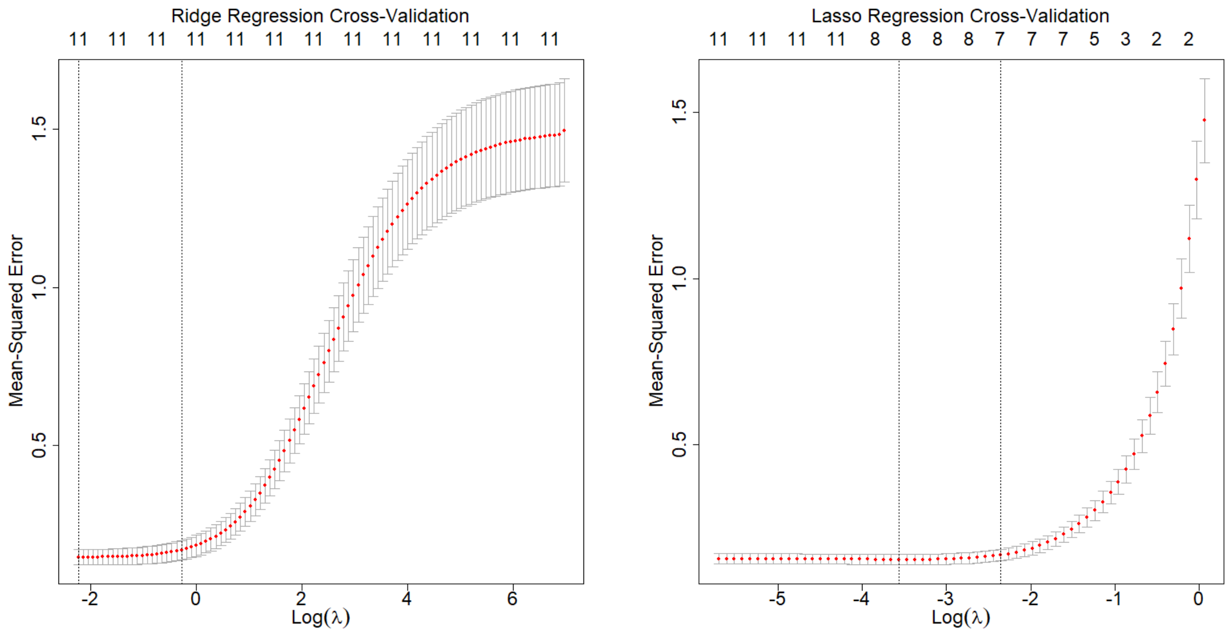


Figure 8: This figure illustrates the result of cross validation analysis between MSE and $\log(\lambda)$ for both Ridge and Lasso Regressions.

Generalized Additive Models (GAM)

Generalized Additive Models (GAM) provide a general framework extending a standard linear model by allowing non-linear functions of each of the variables. As linear regression models, they allow for both countinuous and categorical variables.

$$g(E[Y]) = \beta_0 + \sum_{j=1}^{p} f_j(X_j) \tag{30}$$

where g is a link function and f_j are smooth functions estimated using splines. Splines are piecewise polynomial functions used to create smooth curves that can fit complex relationships in data. They allow flexibility in modeling while maintaining control over the smoothness of the function.

Basis Splines (B-splines): are defined by a set of control points and a degree d. The B-spline basis functions $N_{i,d}(t)$ are constructed such that:

$$S(t) = \sum_{i=0}^{n} c_i N_{i,d}(t) \tag{31}$$

where $S(t)$ is the spline function, c_i are the coefficients and n is the number of basis functions. B-splines maintain local control, meaning adjustments to a control point affect only a limited range of the spline.

Natural Splines: natural splines are a type of cubic spline that is constrained to be linear beyond the boundary knots. This constraint ensures that the spline does not oscillate wildly outside the range of the data. The natural spline is defined by specifying the knots and the polynomial degree:

$$S(t) = \sum_{i=0}^{k} c_i B_i(t) \tag{32}$$

where $B_i(t)$ are the B-spline basis functions, and the additional constraints ensure linearity at the boundaries.

Smooth Splines: use a penalty on the roughness of the fitted curve. The objective function for a smooth spline is given by:

$$\hat{S} = \arg\min_{S} \left\{ \sum_{i=1}^{n} (y_i - S(t_i))^2 + \lambda \int S''(t)^2 dt \right\} \tag{33}$$

where λ controls the trade-off between the fidelity to the data and the smoothness of the spline, where $S''(t)$ is the second derivative of the spline.

In this research, GAM were employed using the normalized dataset to model the relationship between predictors and the response variable `GDP PC` while capturing potential non-linear dependencies. GAM extends the concept of splines by enabling the response variable to depend on smooth functions of the predictors, offering flexibility without requiring a fully parametric specification. An optimization algorithm was used to identify the best configuration for the GAM. For each predictor, two forms were considered: its linear representation and its smoothed representation (s(x)). At each iteration, the algorithm evaluated models that incorporate each predictor in both forms and calculates their AIC. The predictor form (linear or smoothed) minimizing the AIC was retained, ensuring that only the most optimized version of each variable was included in the final model. This iterative process provided a robust framework for building an optimized GAM and assessing the statistical significance of each predictor's relationship with `GDP PC` . The resulting plots in figure 9 illustrate the effect of each predictor on the response variable, showcasing both the linear and non-linear contributions to the model. In Table 6 are reported the reults of the GAM.

Table 6: GAM Results

Model Statistics	
MSE	0,0922
Residual Standard Error	0,3168
R-Squared	0,9373
Adjusted R-Squared	0,9323
F-Statistic	187,1471 on 10 and 125 DF, p-value <3,5573e-70

The GAM achieves a high R-squared value, indicating that it explains approximately 93,73% of the variance in GDP per capita. The Adjusted R-squared of 93,23% further confirms that the model is robust, even after accounting for the number of predictors. The Mean Squared Error (MSE) of 0,0922 and the Residual Standard Error (RSE) of 0,3168 indicate a low level of error, reflecting high predictive accuracy. The F-statistic of 187,15 with a p-value < 3,56e-70 shows that the model as a whole is highly significant, strongly rejecting the null hypothesis that the predictors have no effect on GDP per capita.

Table 7: ANOVA Results for Parametric Effects

Predictor	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
s(SFE)	1	152,099	152,099	1262,6192	< 2,2e-16	***
s(INT)	1	14,474	14,474	120,1517	< 2,2e-16	***
s(COR)	1	6,004	6,004	49,8377	1,583e-10	***
s(ELE)	1	3,542	3,542	29,4058	3,515e-07	***
URB	1	1,153	1,153	9,5709	0,0025065	**
POL	1	1,797	1,797	14,9174	0,0001901	***
s(FDI)	1	0,185	0,185	1,5325	0,2183755	
DEM_FL	3	1,062	0,354	2,9398	0,0363707	*
Residuals	110	13,251	0,120	-	-	

Significance levels: *** $p < 0,001$, ** $p < 0,01$, * $p < 0,05$, . $p < 0,1$

Table 8: ANOVA Results for Nonparametric Effects

Predictor	Npar	Df	Npar F	Pr(F)	
(Intercept)	-	-	-	-	
s(SFE)	3	-	0,9696	0,40987	
s(INT)	3	-	1,1265	0,34161	
s(COR)	3	-	2,3622	0,07520	.
s(ELE)	3	-	1,8257	0,14668	
URB	-	-	-	-	
POL	-	-	-	-	
s(FDI)	3	-	3,9077	0,01074	*
DEM_FL	-	-	-	-	

Significance levels: *** $p < 0,001$, ** $p < 0,01$, * $p < 0,05$, . $p < 0,1$

The results showed in Table 7 assess whether adding the smoothing term (as a whole) significantly improves the model fit compared to a model without that term. A significant parameter in the parametric ANOVA indicates that the predictor, modeled through a smooth function, is important in explaining variance. In contrast, the results showed in Table 8 examine the adequacy of the chosen smoothing function itself. Indeed, a non-significant result in the nonparametric ANOVA means that the specific form of smooth used might not improve model fit significantly or that it could be adequately modeled with fewer degrees of freedom.
It is important to state that in GAM, the interpretation of estimated coefficients differs from traditional linear regression. For the smooth terms in a GAM, there are generally no straightforward standard errors, t-values, or p-values associated with each estimated "coefficient" in the usual sense. in fact, coefficients for smooth terms don't have a direct interpretation due to the non-parametric nature of the estimation. Instead, the significance of a smooth term is assessed as a whole using tests like an ANOVA for the smooth term's contribution to the model. [6] Among the variables, smoothed Self-Employment

[6] GAM softwares typically provide approximate p-values for the overall smooth term, but these reflect

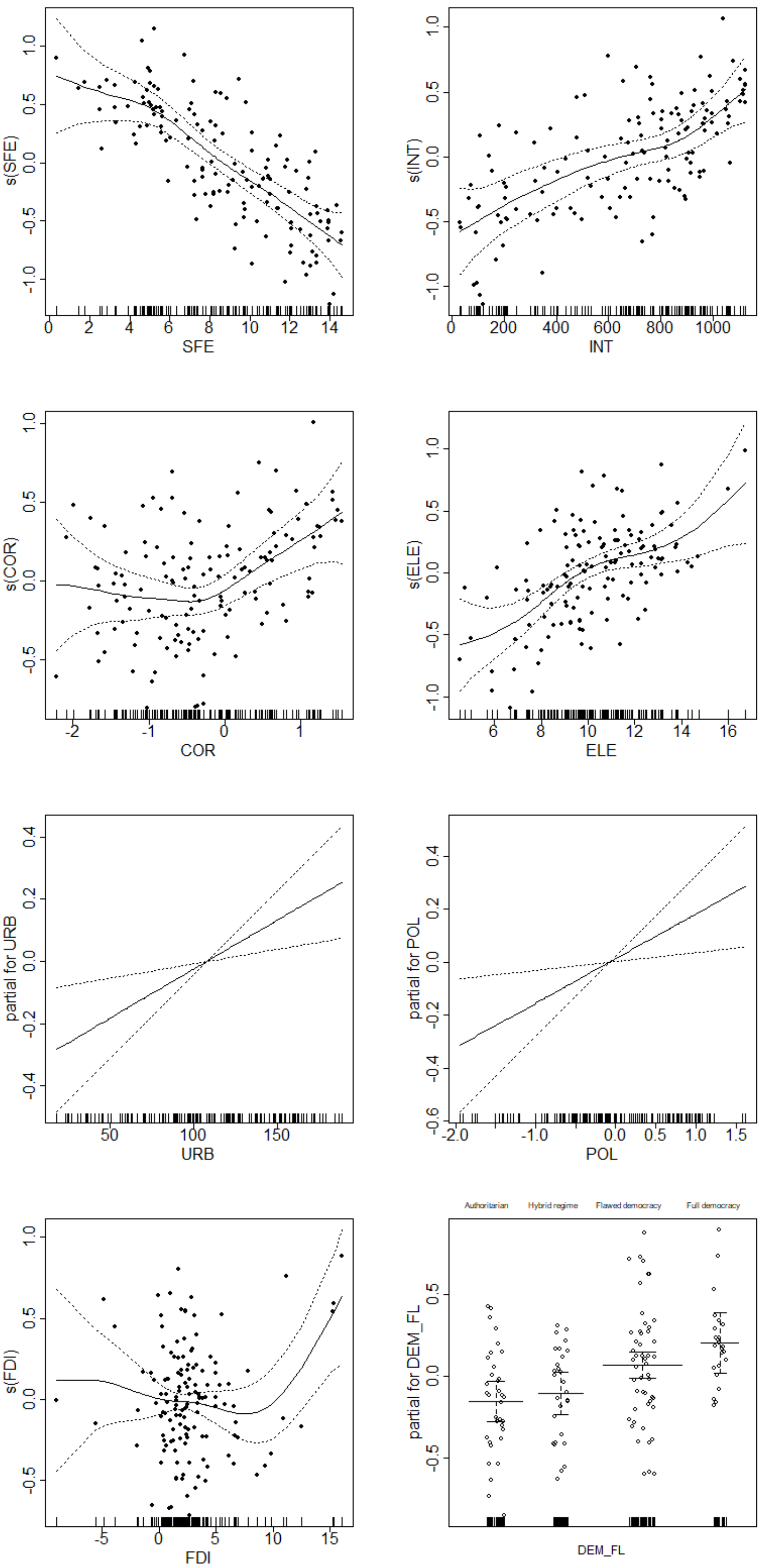


Figure 9: This immage illustrates the partial effects of predictors on the t-test results for GDP per capita. Each plot points out the result of GAM analysis for each variable.

the smooth term's contribution rather than individual coefficients.

(s(SFE)) and Internet Usage (s(INT)) were the most important, followed by smoothed Corruption Control(s(COR)), which showed lower importance in this model. The categorical variable DEM_FL reported the effect of different political regimes. Countries with full democracies show the highest t-statistics, followed by flawed democracies, indicating that these regimes are more significantly associated with GDP PC. The larger error bars for hybrid and authoritarian regimes suggested greater variability or uncertainty in their statistical significance with respect to the dependent variable.

Random Forest

Random Forest is an ensemble method that originates from **Regression Trees**, a non-parametric approach to regression that splits the data into subsets based on feature values, allowing for different predictions in different regions of the input space. The decision tree is formed by recursively partitioning the dataset into subsets based on a feature that minimizes the residual sum of squares (RSS). The key concept in building a regression tree is to find the best split at each node. The objective function can be represented as:

$$RSS = \sum_{i=1}^{N} (y_i - \hat{y})^2 \tag{34}$$

where y_i are the actual values, $\hat{y}$ is the predicted value for the subset, and N is the number of observations in that subset. The splitting criterion can be optimized using methods like least squares, where the algorithm looks for splits that minimize the RSS in the resulting child nodes.

Decision trees tend to overfit the training data, creating complex models that may capture noise. To address this issue, *pruning* can be used to reduce the size of the tree by removing splits that do not significantly improve predictive performance. This helps to balance the trade-off between model complexity (variance) and generalization ability (bias).

The process of pruning is often guided by *cross-validation*, where the dataset is divided into training and validation sets. For each potential tree size k, the cross-validated error is calculated on the validation set, and the optimal size is chosen by minimizing the error.

For classification trees, the *misclassification error* is the proportion of observations incorrectly classified by the model. It is calculated as follows:

$$\text{Misclassification Error} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(y_i \neq \hat{y}_i) \tag{35}$$

where N is the total number of observations, y_i is the true class label for observation i, and $\hat{y}_i$ is the predicted class. The indicator function $\mathbb{I}(y_i \neq \hat{y}_i)$ equals 1 if the prediction is incorrect and 0 otherwise.

Firstly, a regression tree algorithm was run to assess the base model. The optimal size was found at a depth level of 3, indicating that further growth did not significantly improve the model's performance. Limiting tree depth can be useful for balancing complexity and interpretability, with sufficient terminal node sample sizes reducing the risk of overfitting. In Figure 10 it is possible to see the tree generated by the algorithm.

The dataset was split into training set and the test set, let N represent the total number of observations, and let n_{train} and n_{test} represent the number of observations in

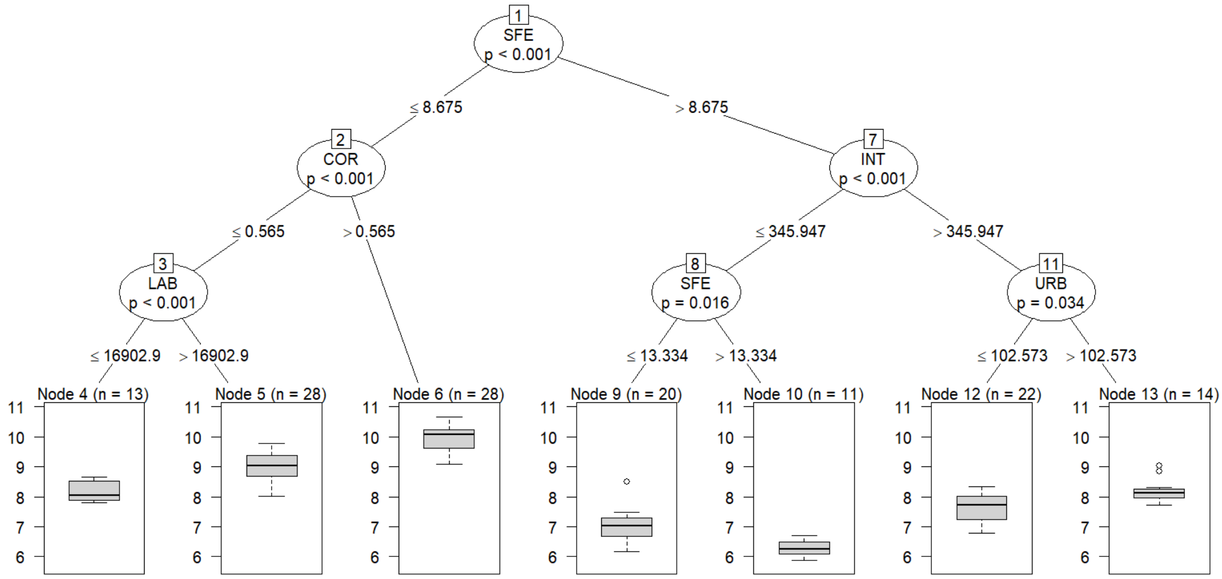


Figure 10: This plot illustrates the result of regression tree analysis for GDP per Capita.

the training and test sets, respectively:

$$n_{\text{train}} + n_{\text{test}} = N \tag{36}$$

In this analysis, $n_{\text{train}} = 30$ and $n_{\text{test}} = 106 \quad (N - 30)$.

The regression tree was fitted to predict `GDP PC` using the training set through the following formula:

$$\hat{Y}_i = f(X_i) = \sum_{m=1}^{M} c_m \mathbb{I}(X_i \in R_m) \tag{37}$$

where $\hat{Y}_i$ is the predicted value of `GDP PC`, c_m is the mean of the dependent variable in region R_m, and $\mathbb{I}(X_i \in R_m)$ is an indicator function that assigns observation i to region R_m. In Figure 11 it is possible to see the model complexity selection method and the prediction fit of the trained algorithm.

The tree splits the data to minimize the sum of squared residuals within each region:

$$\sum_{i \in R_m} (Y_i - \hat{Y}_R)^2 \tag{38}$$

The predicted values on the test set, $\hat{Y}_{test}$ were generated using the trained model:

$$\hat{Y}_{test} = f(X_{test}) \tag{39}$$

To assess the model's performance, the Mean Squared Error (MSE) metric can be used:

$$\text{MSE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (Y_i^{\text{test}} - \hat{Y}_i^{\text{test}})^2 \tag{40}$$

where Y_i^{test} are the actual values of `GDP PC` in the test set, and $\hat{Y}_i^{\text{test}}$ are the predicted values.

The MSE of 0,3805704 showed a relatively low average squared error, suggesting the tree performs well in capturing relationships between predictors and the target variable. Significant splitting variables `SFE`, `COR`, `INT`) at higher levels indicate strong influence on the response variable.

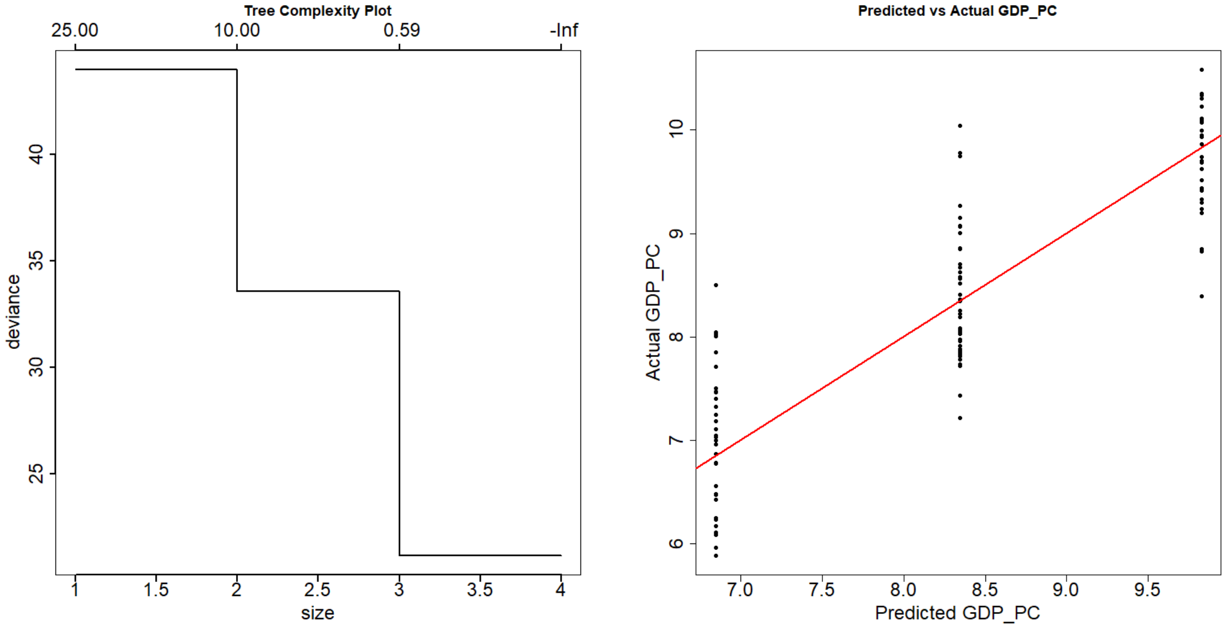


Figure 11: This figure illustrates the complexity analysis (left-side plot) for the regression tree, and the comparison of predicted against actual values (right-side plot) to assess the performance of regression tree algorithm.

Secondly, **Bagging (Bootstrap Aggregating)**, an ensemble method that enhances the stability and accuracy of algorithms was applied. It works generating multiple bootstrap samples D_b from the training dataset D and training a separate regression tree model M_b on each sample. For regression, the final prediction $\hat{Y}$ for an observation X is obtained by averaging the predictions from all B trees:

$$\hat{Y} = \frac{1}{B} \sum_{b=1}^{B} \hat{Y}^{(b)}(X) \tag{41}$$

where $\hat{Y}^{(b)}(X)$ is the prediction from the b-th tree.

This averaging helps reduce the variance of the predictions and often leads to better performance compared to a single model.

The performance of the bagging model is evaluated using the mean of squared residuals and explained variance. The mean of squared residuals (MSR) is computed as:

$$\text{MSR} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{42}$$

where Y_i is the true value of the target variable and $\hat{Y}_i$ is the predicted value. In this case, the MSR is given as:

$$\text{MSR} = 0,2872498$$

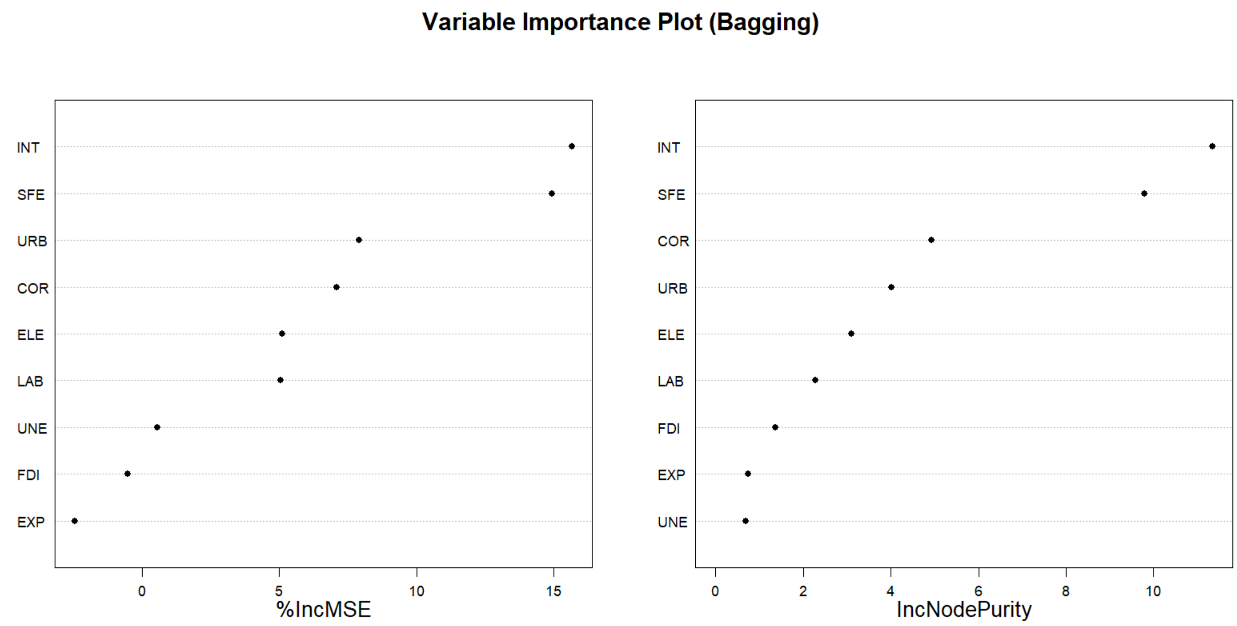


Figure 12: The immage illustrates the list of predictors, ordered by importance for Bagging algorithm.

This represents the average squared difference between the predicted and actual values of `GDP PC`, and a lower value indicates better model performance.

The percentage of variance explained by the model is another important metric and is given by:

$$\text{Var\_explained} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{43}$$

where $\bar{Y}$ is the mean of the true values. The percentage of variance explained indicates how much of the variability in the target variable is captured by the model. In this example:

$$\text{Var\_explained} = 78,87\%$$

This means that the bagging model explained in this analysis approximately 78,87% of the variance in `GDP PC`, based on the predictors used.

The bagging model trained on the normalized dataset uses 500 trees, with all 9 variables considered at each split (`EXP`, `COR`, `ELE`, `INT`, `FDI`, `LAB`, `SFE`, `UNE`, `URB`). The low mean of squared residuals (0,2872498) suggests that the model fits the data well, and the 78,87% of variance explained indicates that the model is able to capture a significant portion of the variation in `GDP PC`. However, there remains 21,13% unexplained variance, which could be due to factors not included in the model or noise in the data.

Random Forest is an advanced ensemble method based on bagging, which constructs a multitude of decision trees during training and outputs the mean prediction (for regression) from individual trees. It generates multiple bootstrap samples D_b from the training dataset D and trains a separate regression tree model M_b by selecting a random subset of m features at each node, where $m < p$ (the total number of features). Hence, with respect to Bagging, it introduces an extra layer of randomness by selecting a subset of features at each split, which decorrelates the trees, reduces variance, minimizes overfitting, and improves prediction accuracy through model aggregation.

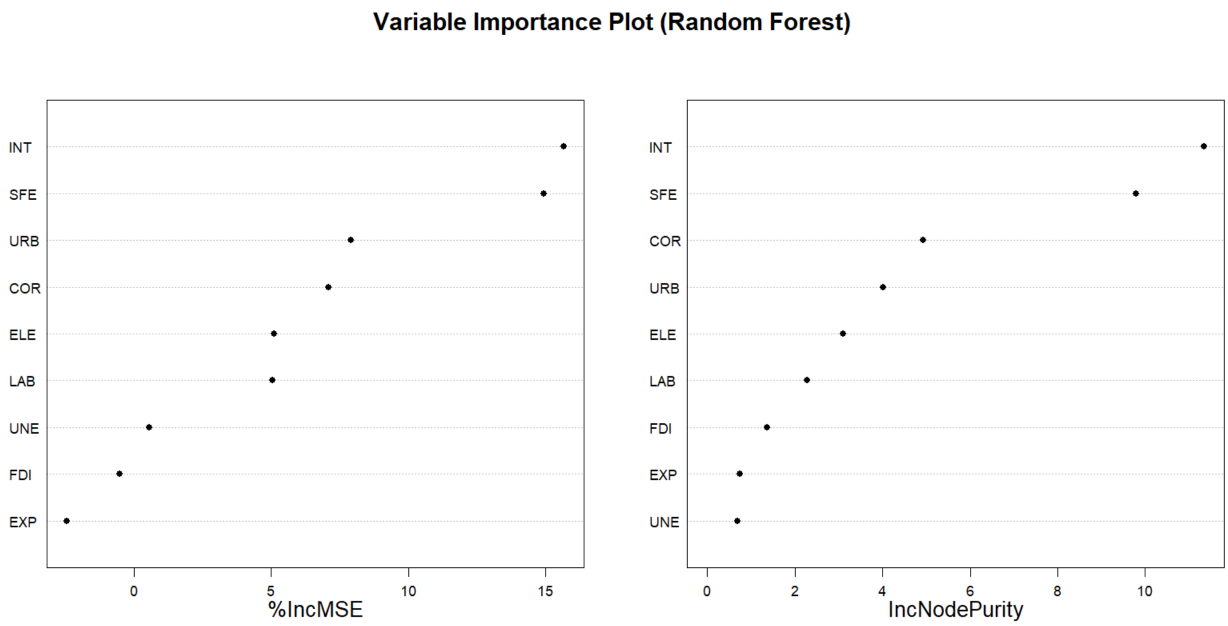


Figure 13: The immage illustrates the list of predictors, ordered by importance for Random Forest algorithm.

The random forest model applied in this analysis was a regression model predicting `GDP_PC` based on a set of 9 explanatory variables. The model employed 500 regression trees in the forest, using only a set of 3 predictors randomly selected at each split (this is different with respect to Bagging, where the complete set of 9 predictors were used at each split).

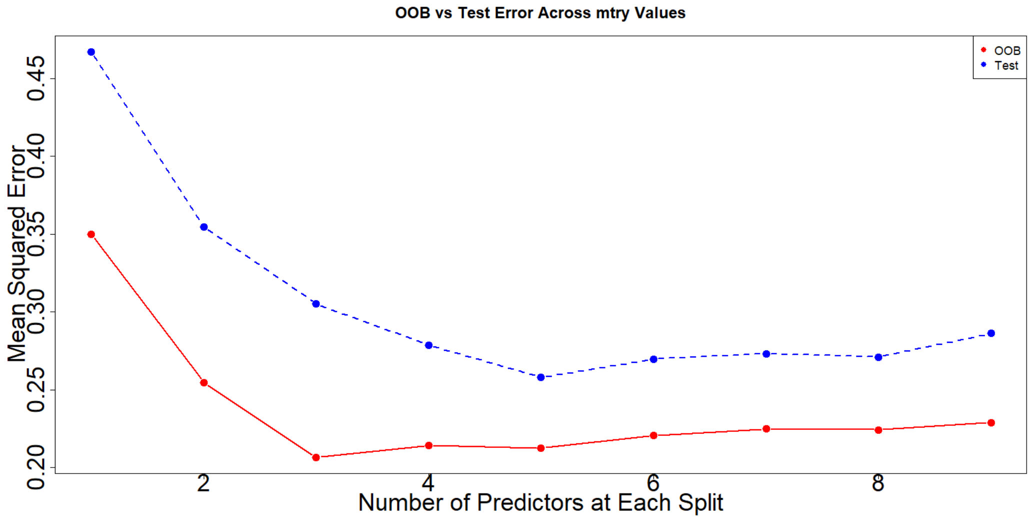


Figure 14: This immage illustrates the comparison between Out-Of-Bag and Test error from Random Forest analysis.

The MSR for this model was:

$$\text{MSR} = 0.3006377$$

The percentage of variance explained by this model was:

$$\text{Var\_explained} = 77.89\%$$

It is possible to noitce in Figure 14, that Out-Of-Bag (OOB) error was higher than test error at lower set of random Variables used in Random Forset, reflecting its bias-corrected nature. The gap narrows as the parameter of random predictors increased, with a marginal effect after 3 or 4 random parameters, indicating effective learning and good generalization to test data.

Figure 15 shows the predictions of trained Bagging and Random Forest models. In this research, Bagging slightly outperformed Random Forest in terms of MSE (0.2872 vs. 0.3006) and % variance explained (78.87% vs. 77.89%), with a negligible difference. This was probably due to the fact that Bagging uses all variables at each split, leading to potentially more optimal splits but reduced tree diversity. Random Forest, on the other hand, limits the number of variables, enhancing robustness and reducing overfitting, especially with correlated features. While Bagging may achieve slightly better accuracy, Random Forest offers better generalizability and computational efficiency, making it preferable when interpretability and overfitting are concerns. Bagging may be better suited for maximizing predictive accuracy when overfitting is not an issue.

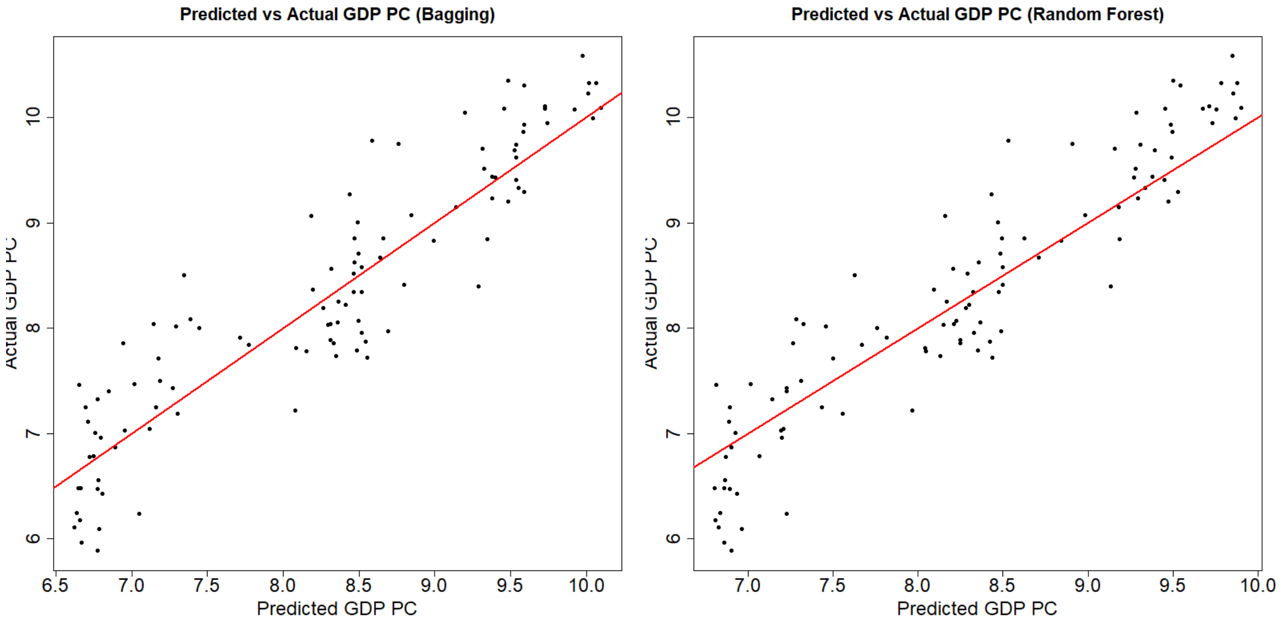


Figure 15: This figure illustrates the comparison of predicted against actual values for Bagging (left-side plot) and for Random Forest (right-side plot) to assess the performance of both the algorithms.

In Figure 12 and Figure 13 it is possible to see the importance rank of variables in both Bagging and random Forest Models.

Conclusion

When comparing the Supervised Statistical Learning models applied to the normalized dataset of 147 countries, it is possible to summarize 3 findings:

1. Model Accuracy: The GAM stands out as the most accurate model. With an MSE of 0,0922, a Residual Standard Error (RSE) of 0,3168, and an Adjusted R-squared of 0,9323, it outperformed all other methods in terms of minimizing error and explaining the variance in GDP per capita. The Multiple Linear Regression (MLR) model also performed well, achieving an Adjusted R-squared of 0,9096, however its higher RSE of 0,3664 makes it slightly less precise compared to GAM. Ridge and Lasso regression models achieve similar performance to MLR, with slightly lower Adjusted R-squared values of 0,9058 for both and marginally higher MSE and RSE values (around 0,1274-0,1280 and 0,3745–0,3754, respectively). While reducing the potential for overfitting, the performance of these regularized models did not surpass that of GAM. On the other hand, Bagging and Random Forest methods explained a much lower proportion of variance, with Bagging achieving a slightly better MSE (0,2872) and variance explained (78,87%) compared to Random Forest (MSE of 0,3006 and 77,89% variance explained). Despite their ability to handle complex interactions, these ensemble methods fall short in terms of raw accuracy compared to GAM and MLR.

2. Most Important Predictors: Across the different statistical learning models used to analyze GDP per capita, the importance of variables varied slightly depending on the methodology, yet some common patterns emerged. For Ridge regression, the top three variables are `EXP`, `POL`, and `COR`. This suggests that Ridge, which retains all variables and reduces overfitting through regularization, highlights the broader importance of trade (`EXP`) and institutional stability (`POL`) alongside corruption control (`COR`). In contrast, Lasso regression, which penalizes and shrinks less important coefficients to zero, identifies `SFE`, `POL`, and `COR` as the most important predictors, emphasizing Self Employement's nonlinear contribution while confirming the relevance of political and institutional variables. For Multiple Linear Regression (MLR), the top three variables are `SFE`, `ELE`, and `COR`. Random Forest, prioritized `INT`, `SFE`, and `COR`, showcasing the importance of both nonlinear predictors (`INT` and `SFE`) and corruption. Similarly, the GAM, which incorporates nonlinear smoothers, identified `SFE`, `INT`, and `COR` as the top contributors, reinforcing the importance of capturing nonlinear relationships for explaining economic development.

3. Theory foundations Empirical Confirmation: Overall, `SFE` and `COR` emerge consistently across almost all models as critical variables, highlighting their pervasive influence on economic development. Based on the observed behavior of Self-Employment and Corruption Control in this analysis, the findings lend greater credibility to the theoretical framework proposed by W. Easterly in The Elusive Quest for Growth (2001), specifically, the results align with arguments regarding the critical role of institutional quality and structural economic factors in shaping a country's developing trajectory.

Unsupervised Learning

In this section, inference was performed based on the analysis of unlabeled data, meaning that the dataset includes only input data, excluding desired output.

K-Means Clustering

K-Means Clustering is one of the main unsupervised learning algorithms. In this research the algorithm of Hartigan and Wong (1979) was employed. The algorithm uses an iterative method for partitioning n observations into k clusters, minimizing the within-cluster sum of squares $W(C_k)$:

$$\min_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} W(C_k) \right\} \tag{44}$$

$$W(C_k) \ = \sum_{j=1}^{k} \sum_{i \in C_j} \|x_i - \mu_j\|^2, \tag{45}$$

where: C_j is the set of points in cluster j, $\mu_j = \frac{1}{|C_j|} \sum_{i \in C_j} x_i$ is the mean of cluster j, $\|x_i - \mu_j\|^2$ represents the squared Euclidean distance between a point x_i and the cluster centroid μ_j.

Combinigng equation 44 and 45, it is possible to synthesize the minimization problem in one sigle formula:

$$\min_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} \sum_{j=1}^{k} \sum_{i \in C_j} \|x_i - \mu_j\|^2 \right\} \tag{46}$$

The Hartigan and Wong algorithm iteratively adjusts the cluster assignments and cluster centroids based on the following steps:

Initialization: choose k initial centroids $\{\mu_1, \mu_2, \ldots, \mu_k\}$, typically by random selection or some heuristic.

Assignment Step: assign each point x_i to the cluster C_j with the nearest centroid:

$$C_j = \{x_i : \|x_i - \mu_j\|^2 \leq \|x_i - \mu_l\|^2 \, \forall l = 1, \ldots, k\}. \tag{47}$$

Update Step: for each cluster C_j, update the centroid to the mean of the points assigned to that cluster:

$$\mu_j = \frac{1}{|C_j|} \sum_{i \in C_j} x_i. \tag{48}$$

Optimization: iteratively reassign points to clusters and update centroids. For each point x_i, compute the potential change in $W(C_k)$ if x_i is moved to another cluster C_l. If moving x_i reduces the $W(C_k)$, reassign it to C_l. Formally, the change in $W(C_k)$, $\Delta W(C_k)$, is:

$$\Delta W(C_k) = \|x_i - \mu_j\|^2 - \|x_i - \mu_l\|^2, \tag{49}$$

where j is the current cluster and l is the candidate cluster.

Convergence: repeat the assignment and update steps until convergence, which occurs when no points change clusters anymore, or the change in $W(C_k)$ is below a threshold. The algorithm ensures convergence to a local minimum of the $W(C_k)$. However, as the cost function is non-convex, it is important to bear in mind that the solution depends on the initial choice of centroids.

The process begins by applying the K-Means clustering algorithm to a multivariate, scaled dataset. The original dataset was standardized in order to fix differences in scales between variables. Numeric columns of the dataset were standardized by centering and scaling each variable to have a mean of 0 and a standard deviation of 1. This process ensures that all variables are on the same scale, which is particularly useful for methods sensitive to variable magnitude (e.g. PCA clustering).

Given a dataset, the standardization for each element x_{ij} in the dataset is computed as:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{50}$$

where:
x_{ij} is the value of the i-th observation in the j-th column.
$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ is the mean of the j-th column.
$s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$is the standard deviation of the j-th column.

After applying standardization, each column of the dataset satisfies the following properties:

$$\text{Mean} = 0, \quad \text{Standard Deviation} = 1$$

This transformation is performed independently for each column in the dataset.

Since the dataset contains multiple dimensions, direct graphical representation is infeasible. To overcome this, PCA is employed to reduce the dimensions of the clusterized data while preserving the maximum variance possible [7].

in Figure 16, it is shown the result of the K-means Cluster with 3 Clusters in a 2D Visualization, using the following independent variables: `DEM IND`, `EXP`, `POL`, `COR`, `IND`, `UNE` and `URB`. The initial scatter plot illustrates the presence of three clusters, with data points colored based on their assigned clusters. Ellipses are added to depict the approximate boundaries of the clusters. While the plot reveals the clustering structure, it is evident that the ellipses overlap slightly, likely due to the complexity and dimensionality of the data set.

To address the visualisation limitations of the 2D representation, a 3D scatter plot Figure 17 is provided. By incorporating a third principal component, the visualization unveils a clearer separation between clusters. This additional dimension reduces the overlap observed in 2D, making the distinct boundaries between clusters more apparent, enhancing interpretability and supporting the validity of the K-Means algorithm's results.

[7] To this extent, it is important to notice that PCA does not affect the partitioning in clusters, since clustering is performed before applying PCA to the data. PCA is only used to enhance data visualization

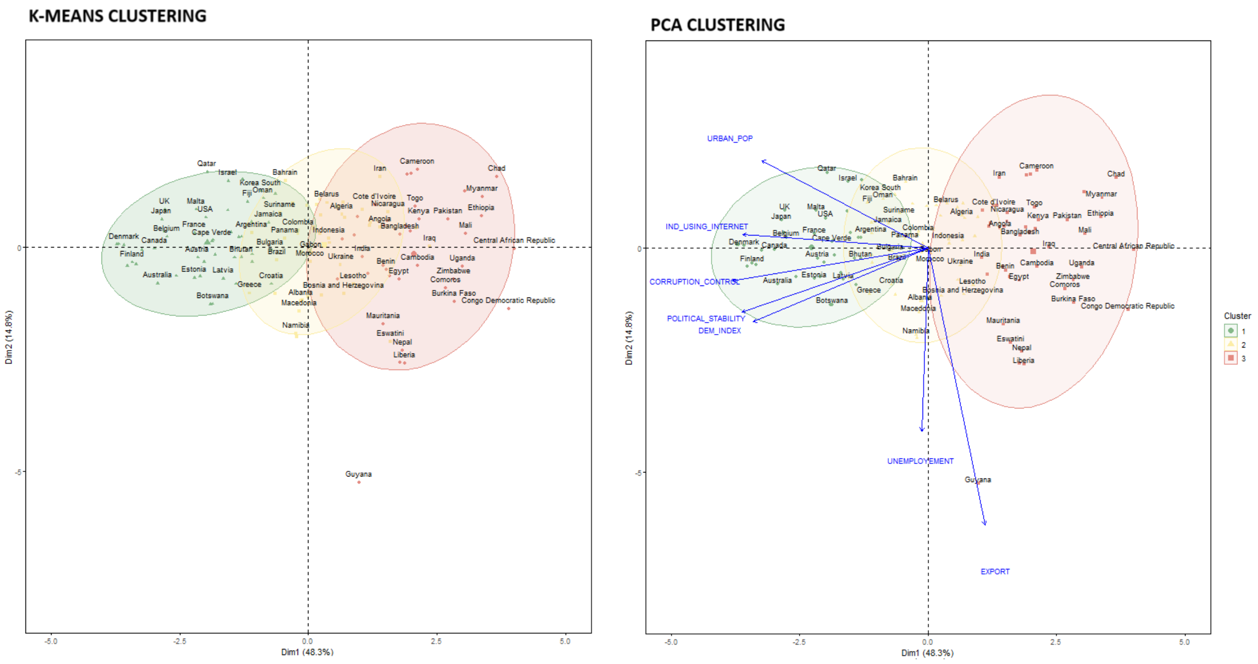


Figure 16: This figure illustrates the biplot for K-Means clustering (left-side plot) and PCA clustering (right-side plot)

The 3D plot provides a more intuitive understanding of the data structure, demonstrating that while 2D visualizations can serve as a preliminary tool, higher-dimensional plots are crucial for interpreting complex datasets. These findings reinforce the importance of combining clustering algorithms with dimensionality reduction techniques for effective data exploration and communication.

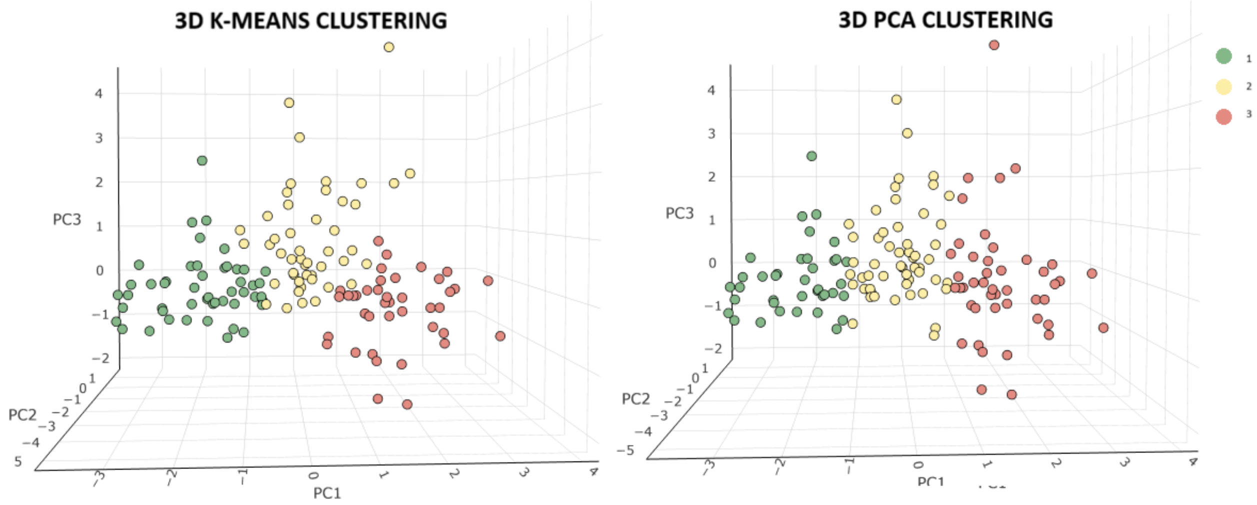


Figure 17: This figure illustrates the 3D Plot for K-Means clustering (left-side plot) and PCA clustering (right-side plot).

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction method employed to preserve the most variance as possible. PCA transforms the original variables into a new set of uncorrelated variables called principal components, which are linear combinations of the original variables.

The PCA process involves the following steps:

Standardization: center the data by subtracting the mean and scaling to unit variance. This step is critical as far as not scaled data would result in unbalanced effects of principal components. Observing the variance of the variables, it can be noticed that some variables have higher values (`EXP` = 147,18, `IND` = 68,81 and `URB` = 61,27 are much greater than `DEM IND` = 5,54, `POL` = -0,15, `COR` = -0,06 and `UNE` = 7,45), hence most of the principal components that are observe would be driven by them.

$$Z_i = \frac{X_i - \mu}{\sigma} \tag{51}$$

Covariance Matrix Computation: Calculate the covariance matrix C of the standardized data:

$$C = \frac{1}{n-1} Z^T Z \tag{52}$$

Eigenvalue and Eigenvector Calculation: solve the eigenvalues λ and eigenvectors v of the covariance matrix:

$$Cv = \lambda v \tag{53}$$

Principal Component Selection: Select the top k eigenvectors corresponding to the largest eigenvalues to form the projection matrix W.

Transformation: Project the original data onto the principal component space:

$$Y = ZW \tag{54}$$

In this research, Principal Component Analysis (PCA) was performed to better understand the underlying structure of the data, trying to figure out which clustering configuration could be retrieved by reducing the dimensionality of the dataset. The results of the PCA are summarized below.

Table 9: PCA Component Summary

Component	Standard Deviation	Proportion of Variance	Cumulative Proportion
PC1	1,8387	0,48296	0,48296
PC2	1,0189	0,14832	0,63128
PC3	1,0056	0,14445	0,77573
PC4	0,8714	0,10848	0,88420
PC5	0,5911	0,04991	0,93412
PC6	0,5008	0,03583	0,96995
PC7	0,4587	0,03005	1,00000

PC1 showed the highest standard deviation of 1,8387 and captured around 48,29% of the total variance in the dataset, indicating to represent the most significant underlying pattern. PC2 explained an additional 14,83% of the variance, bringing the cumulative explained variance to 63,13% with the first two components. PC3 added 14,4%, meaning that the first three components together can explain 77,57% of the total variance. PC4 explanied almost 88,42% of the total variance, capturing most of the meaningful structure in the data. PC5, PC6 and PC7 explained smaller portions of the variance, each contributing less than 5%, indicating diminishing returns from adding more components beyond the first four.

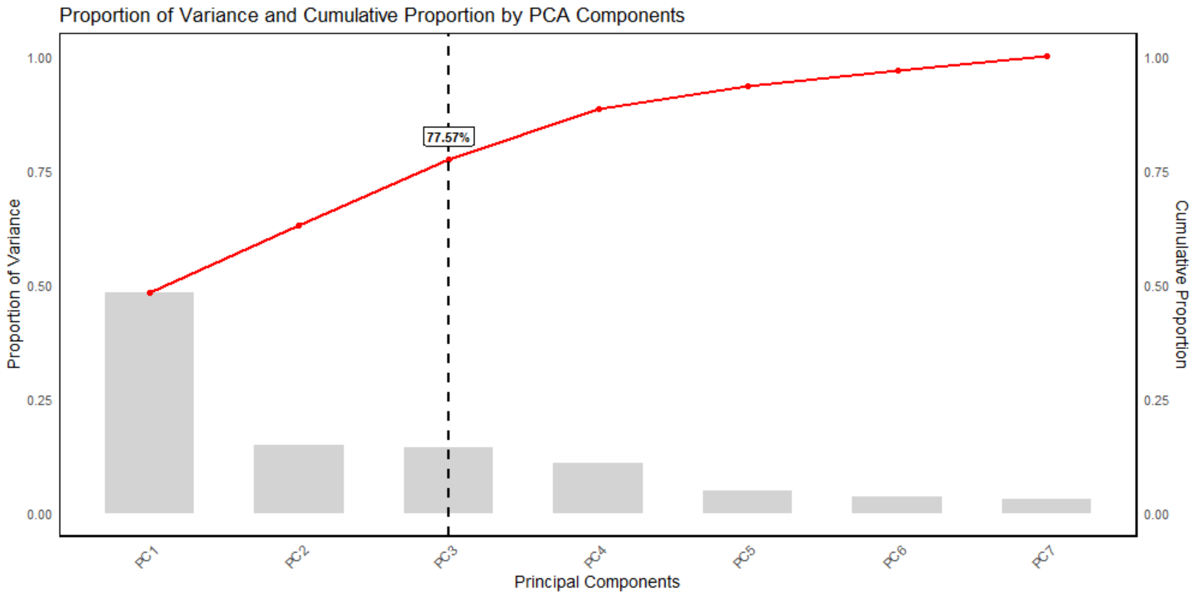


Figure 18: This plot illustrates the cumulative explained variance (red line), which is the cumulative sum of gray bars, which express the explained variance for each single Principal Component.

Given that the first 3 components explained about 77,57% of the variance, it is reasonable to retain them for further analysis or dimensionality reduction, allowing for a 3D representation of clusters. Including components beyond PC3 or PC4 may introduce unnecessary noise or complexity without adding much explanatory power to the model.

At first, as shown in Figure 16, a biplot was created by plotting the first two principal components, allowing for visualization of data clusters in a 2D scatterplot. Clusters were generated by applying K-Means clustering technique on the PCA-reduced data [8].

The biplot shows the relationships between countries and key variables based on PCA: PC1 is the horizontal axis and it explains the largest proportion of the variance in the dataset. Countries further to the right on PC1 tend to be more influenced by the variables with large loadings on this axis, such as high Corruption Control, Political Stability, and other positive economic indicators. PC2 is the vertical axis and explains the second-largest proportion of variance. Countries further up on PC2 tend to be influenced more by variables like Export and Unemployment. The blue arrows represent the loadings of the original variables, longer arrows normally indicate stronger contributions of those variables to the components. Export seems to be positively associated with PC2, meaning countries that score higher on this component have stronger export-oriented economies. Unemployment has a strong influence on PC2 as well, but it followed a different direction

[8] To this extent, it is important to notice that PCA affects the partitioning in clusters, since clustering is performed after applying PCA to the data. In this case, differently to the previous method, PCA elaborate the dataset and clustering is used to show clusters it in a partitioned plot

compared to export. Corruption Control, Political Stability, and Democracy Index were all closely aligned, indicating that they contributed similarly to both components, likely reflecting institutional strength and governance. Industrial Use of the Internet and Urban Population also played a significant role in driving PC1, aligning with countries that are more developed or urbanized.

The plot effectively divided countries based on their economic and institutional characteristics. Cluster 1 (Developed countries) like USA, UK, Germany, Australia, Canada, Japan, France, and Norway were clustered to the left, reflecting positive associations with Corruption Control, Political Stability, and Urban Population. Cluster 2 (Developing countries) like Brazil, Argentina, India, China, and Russia were more centrally located, showing moderate values for these variables. Cluster 3 (Underdeveloped or developing countries), such as Bangladesh, Ethiopia, Chad, Madagascar, Central African Republic, and Niger were positioned on the right and bottom, suggesting lower scores on factors such as Corruption Control and Political Stability, as well as higher Unemployment rates. In Figure 17 is provided a 3D representations, in which also PC3 can be observed, better highlighting how countries are grouped based on their similarities across socio-economic dimensions.

Agglomerative Hierarchical Clustering

Hierarchical clustering is a clustering technique that does not require pre-specifying the number of clusters, which is a potential advantage with respect to K-Means. As opposite to K-Means, it produces an attractive tree-based representation of the observations grouped into clusters, called a dendrogram. In this reasearch, the bottom-up or agglomerative clustering approach was employed. This basically foresees that the dendogram is built starting from the leaves and combining clusters up to the trunk. The algorithm starts by treating each observation as a single cluster, then it iteratively merges the two most similar clusters based on a dissimilarity measure (In this reasearch Euclidian distance was employed as a dissimilarity measure). This process continues until all observations merge into a single cluster and the dendogram is completed. At this point, it is possible to cut the dendogram at a specific level on vertical axis, in oreder to obtain a reasonable group of clusters. It is important to bear in mind that it not possible to draw conclusions about the similarity of two observations based on their proximity along the horizontal axis. Indeed, there are 2^{n-1} possible reorderings of the dendrogram, where n is the number of leaves. This is because at each of the $n-1$ points where merging occurs, the position of the two merged branches could be swapped without affecting the meaning of the dendrogram. One of the most important element in hierarchical clustering is about defining the dissimilarity between groups of observations, which is normally achieved through linkage methods. The choice of linkage and dissimilarity measure significantly influences the dendrogram's structure and interpretability. The main steps performed by the algorithm are:

Initializzation: Start with each data point as its own cluster.

Distance Calculation: Compute the distance between every pair of clusters using a chosen linkage criterion.

Merging Clusters: Merge the two closest clusters. The choice of distance can be defined by several linkage criteria:

Complete Linkage: Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the largest of these dissimilarities.

$$D(A, B) = \max\{d(a, b) : a \in A, b \in B\} \tag{55}$$

Single Linkage: Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the smallest of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.

$$D(A, B) = \min\{d(a, b) : a \in A, b \in B\} \tag{56}$$

Average Linkage: Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster X and the observations in cluster Y, and record the average of these dissimilarities

$$D(A, B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b) \tag{57}$$

Centroid linkage: Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable inversions.

$$D(A, B) = d(\overline{a}, \overline{b}) \tag{58}$$

Ward's Method: This method does not directly define a measure of distance between two points or clusters. It is an ANOVA-based approach. One-way univariate ANOVAs are done for each variable with groups defined by the clusters at that stage of the process. At each stage, two clusters merge that provide the smallest increase in the combined error sum of squares.

$$D(A, B) = \frac{1}{2} \left(\|A \cup B\|^2 - \|A\|^2 - \|B\|^2 \right) \tag{59}$$

This formula calculates the change in variance (or inertia) caused by merging two clusters: $\|A \cup B\|^2$: The total variance of the new merged cluster. $\|A\|^2 + \|B\|^2$: The combined variance of the two clusters before merging. Subtracting the latter from the former measures the "additional variance" introduced by combining A and B. The formula's division by 2 normalizes this change, reflecting the contribution from both clusters.

Repeat: Continue the merging process until only one cluster remains or until a stopping criterion is reached.

In this reasearch, all the afore mentioned methods have been tested. As a result, only Ward's method provided an acceptable clustering result, which can be interpreted for categorizing countries into developed, developing, and underdeveloped. In Figure 19, the results are good even if not always perfect. Indeed, it is possible to find some wrong classification cases, such as Bhutan and Malaysia classified as developed countries, or Haiti

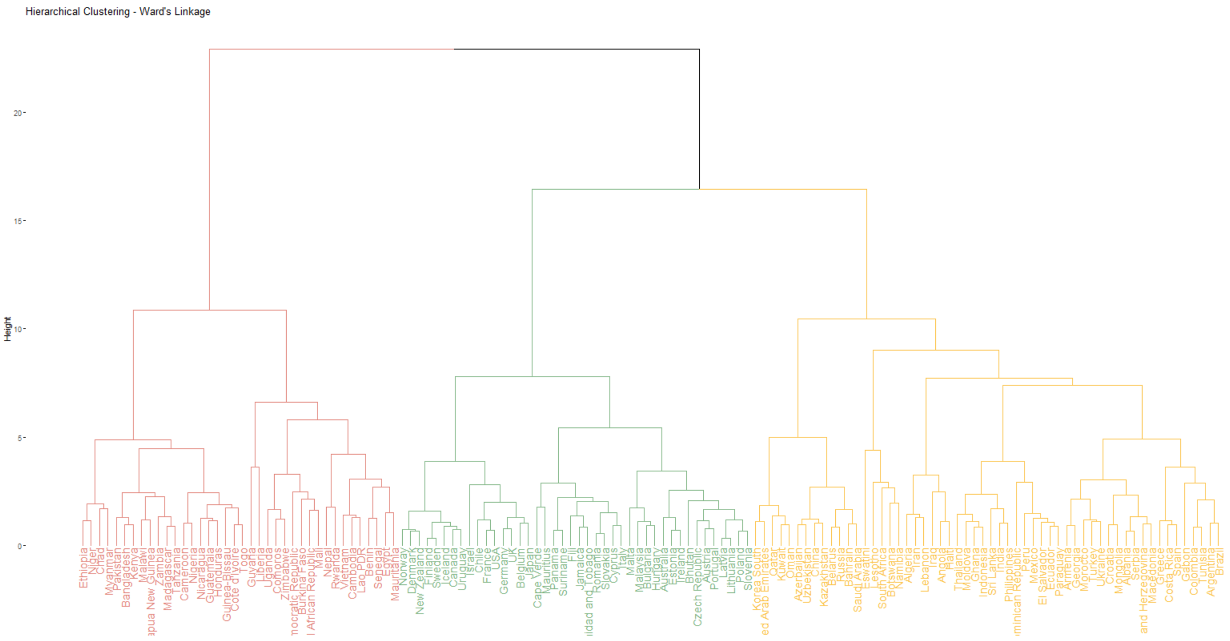


Figure 19: This figure illustrates the dendogram obtained through Ward's Method for Agglomerative Hierarchical Clustering Algorithm.

and Sri Lanka calssidied as developing economies. On the contrary, the other methods (Complete Linkage, Single Linkage and Average Linkage) do not yield meaningful clusters. The fact that Ward's method produces better clustering results suggests that the dataset likely exhibit strong within-cluster similarity (in terms of variance) and distinct cluster separation. In this regard, socioeconomic indicators (such as GDP, unemployment, political stability, etc.) likely exhibit a structure where countries that are economically similar are close in terms of variance, making Ward's variance-minimizing approach a good fit.

Conclusion

When comparing the Unsupervised Statistical Learning models applied to the scaled dataset of 147 countries, it is possible to summarize 3 findings:

1. Model Performance: the combination of K-Means clustering and PCA proves highly effective in identifying and interpreting groupings of countries based on their economic and governance indicators. K-Means excels in handling high-dimensional data and providing stable classifications, while PCA enhances interpretability by reducing dimensionality for easier visualization. Despite minor discrepancies in PCA projections, the hybrid approach ensures robustness and improves the clarity of cluster separability. Regarding Agglomerative Hierarchical Clustering, only Wasrd's method provided a good result for the scope of this project, which suggests that the algorithm in general is not suited for this kind of analysis.

2. Most Important Predictors: the analysis highlighted governance quality (e.g. corruption control and political stability), infrastructure (e.g. electricity consumption and urbanization), and economic indicators (e.g. unemployment) as the most significant variables in distinguishing between clusters. These predictors align with theoretical expectations, effectively differentiating developed, middle-income, and low-income countries, as observed in the consistent composition of clusters across methods.

3. Theory Foundation Empirical Confirmation: the findings confirmed established economic theories linking governance quality, infrastructure and economic development. For instance, the alignment between K-Means and PCA clusters validates the stability of classifications based on these predictors, especially for developed nations (e.g. Australia and Germany) and low-income economies (e.g. Ethiopia and Chad). This theoretical consistency underscores the importance of investing in governance and infrastructure as key drivers of economic development.

References

G. James, D. Witten, T. Hastie, R. Tibshirani (2023) - ”An Introduction to Statistical Learning with Applications in R” Second Edition;

https://online.stat.psu.edu/stat505/lesson/14/14.4

https://www.stat.cmu.edu/ cshalizi/350/2008/lectures/08/lecture-08.pdf

https://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/Collinearity#: :text=More