Master Thesis

Political Corruption Trajectories in 133 Countries: A Time Series Clustering Study from 2000 to 2022

Maria Fernanda Ortega Valencia
Master of Data Science for Public Policy
Hertie School
Berlin, 2024

Supervised by Prof. Asya Magazinnik

Word Count: 7600

Table of contents

Abstract 4

Introduction 5

Related work 5

Data collection and indicators selection 7

Methodology 8
- Overview . . . 8
- Clustering Techniques . . . 8
- Model Evaluation Metrics . . . 9
- Optimal Number of Clusters . . . 10
- Correlation between Sociopolitical and Economic variables and corruption clusters . . . 10

Results 11
- a) **Clustering Political Corruption** . . . 11
  - **Selection of the optimal number of clusters** . . . 11
  - **Clustering Model Selection** . . . 11
  - **Cluster Analysis and Temporal Trends** . . . 12
  - **Analysis of Sociopolitical and Economic Variables by Cluster Over Time** . . . 15
- b) **Clustering Executive, Legislative, and Judicial Corruption** . . . 17
  - **Clustering Model Selection** . . . 17
  - Executive Corruption . . . 18
  - Legislative Corruption . . . 22
  - Judicial Corruption . . . 27

Discussion of results 32
- **Clustering of Countries Based on Corruption Categories** . . . 32
- **Relationship Between Sociopolitical and economic Variables and Corruption** . . . 33

Conclusion 34

Bibliography 36

Appendix 39
- Appendix A: **Analysis of Variable Importance Scores for Political Corruption** . . . 39
  - **Table 1A: Importance Scores of Variables in Assesing Political Corruption (2000)** . . . 39
  - **Table 2A: Importance Scores of Variables in Assesing Political Corruption (2022)** . . . 40
- Appendix B: Performance Analysis of Clustering Models on Corruption Data across Executive, Legislative and Judicial branch . . . 41
  - **Table 1B: Clustering Performance Metrics for Executive Branch Corruption** . . . 41

Table 2B: Clustering Performance Metrics for Legislative Branch Corruption 41
Table 3B: Clustering Performance Metrics for Judicial Branch Corruption 42

Abstract

This study investigates the clustering of political, executive, legislative, and judicial corruption in 133 countries from 2000 to 2022, using time-series clustering techniques. The clustering was achieved through methods such as Time Series K-means with different distance metrics (Euclidean, DTW, Soft-DTW) and Kernel K-means. Model evaluations were conducted using the Silhouette Score and Davies-Bouldin Score. The research identifies distinct clusters of countries with similar corruption profiles and examines the evolution of these clusters over two decades. Our findings reveal that while some countries have recorded notable improvements in the fight against corruption, others continue to struggle with entrenched corrupt practices. In particular, Georgia and Tunisia show unique trajectories in their anti-corruption efforts. In addition, the study explores the correlation between the absence of corruption in these clusters and key socio-political and economic variables. Specifically, in single-country clusters, there are significant correlations between absence of corruption and factors such as transparency of laws, freedom of expression, access to justice and GDP per capita. The results of this research highlight the complex relationship between governance, economic variables and polical corruption.

Introduction

Over the years, the debate among national governments and international bodies on the most effective government reforms and policies to combat political corruption has intensified. This focus is driven by the recognition that corruption not only undermines democratic institutions, but also exacerbates poverty and inequality. For instance, the World Bank (2020) highlights that corruption inflates the costs and decreases the efficiency of public and private investments, which severely affects economic and social progress. In addition, the United Nations Development Program (2008) notes that corruption erodes trust in public institutions, hinders the rule of law, and can lead to political instability. Reinforcing these findings, the International Monetary Fund (2016), has stated that effective anti-corruption measures can significantly boost a country's economic growth by improving the management of public resources and the business environment. Despite these insights and global efforts to address the challenge, the complexity of corruption continues to evolve, requiring the pursuit of diverse analytical approaches to fully understand its dynamics and impact.

Building on this context, this study seeks to employ advanced statistical techniques, including time-series clustering models, to analyse data on corruption in 133 countries from 2000 to 2022. The objective is to identify how countries are clustered in terms of political corruption in general, as well as within different political branches: executive, legislative and judicial. In addition, this research examines the time trends of key socio-political and economic variables within these clusters, exploring their relationship with different branches of political corruption. The two main questions addressed are: (1) How are countries grouped according to the four categories of corruption, and which clusters show peculiar trends over time? (2) What is the relationship between key socio-political and economic variables and the four categories of corruption in the different clusters?

Identifying clusters of countries with similar corruption profiles and understanding the specific dynamics of these clusters provides detailed information that could serve as a basis for targeted anti-corruption strategies. For instance, countries within a group that exhibit high levels of judicial corruption could benefit from targeted judicial reforms and international judicial cooperation. This approach aligns with the objectives of major international agencies, such as the World Bank and the United Nations, which aim to strengthen governance and improve the effectiveness of public administration. Therefore, this study seeks to contribute to analytical research on political corruption to enable the development of informed public policies that not only respond to the state of corruption, but also seek to be predictive and preventive, with the main objective of disrupting the conditions that foster corruption.

Related work

The literature on political corruption reveals that several variables significantly influence its prevalence and occurrence. Economic development, political stability and cultural factors are repeatedly highlighted as the main determinants of corruption levels. For instance, Treisman (2000) explores the relationship between economic prosperity and corruption, arguing that wealthier nations tend to exhibit lower levels of corruption. His analysis suggests that economic development contributes to improved institutional frameworks and governance, which in turn discourages corrupt practices. With a similar argument, Mauro

(1995) relates corruption inversely to investment and economic growth, postulating that corruption not only reduces operational efficiency but also deters foreign investment, thus slowing down economic development.

On the other hand, Charron and Lapuente (2010) analyse the impact of political stability on corruption and conclude that stable political systems are associated with a reduction in corruption due to more effective governance mechanisms. Furthermore, Rothstein and Teorell (2008) highlight the crucial role of the quality of government institutions in the fight against corruption. They argue that effective, impartial and efficient institutions are essential to prevent corruption by maintaining high standards of public administration and accountability. Along those lines, Goel and Nelson (2011)elaborate on the importance of legal frameworks and their enforcement, noting that the perceived likelihood of being caught and punished is a crucial deterrent to corrupt practices, potentially more important than the severity of penalties.

With respect to cultural influences on corruption, Husted (1999) investigates how social norms and values influence perceptions and behaviours of corruption and concludes that cultural dimensions can inhibit or facilitate corrupt activities depending on the social context. Similarly, Davis and Ruhe (2003) findings suggest that higher levels of perceived corruption correlate with cultural dimensions characterised by high power distance[1], strong uncertainty avoidance and low individualism. They also argue that 'masculine societies are perceived as more corrupt than more feminine societies.' Along those lines, Uslaner (2008) emphasizes the role of societal trust and institutional fairness, arguing that higher levels of trust within a society and its institutions correlate with lower corruption levels.

In addition to studies focused on the determinants of corruption, some recent work has utilized machine and deep learning models to analyse the levels of political corruption in various countries. For instance, Ghahari et al. (2021) use artificial neural network models and time series analysis to predict corruption levels in 113 countries between 2007 and 2017. Their analysis operates both globally and within specific groups classified by development-related attributes using nonlinear autoregressive neural network models with exogenous inputs (NARX). Their findings show that NARX models can reliably predict future levels of corruption, providing practical information for policymakers and organizations to refine anti-corruption measures. In addition, Paulus and Kristoufek (2015) use clustering techniques to analyse perceptions of corruption in 134 countries based on the Corruption Perceptions Index (CPI). They identify four main clusters showing different levels of corruption and economic development. The analysis reveals a strong correlation between economic development and perceptions of corruption, with wealthier countries tending to show lower levels of perceived corruption.

[1]According to Davis and Ruhe (2023), 'power distance is the level of acceptance by a society of the unequal distribution of power in institutions.'

Data collection and indicators selection

In this research, two primary data sources have been used to explore political corruption and associated economic and governance factors: the Varieties of Democracy Institute (V-Dem) and the World Bank[2]. The following indicators of political corruption were selected from the V-Dem institute:

a. Political Corruption: This index measures the overall extent of corruption within a country's political system. It integrates perceptions and experiences of corruption across various branches of government (public sector; and executive, legislative, and judicial branches).

b. Judicial, Executive, and Legislative Corruption: These three indices measure the prevalence of corrupt activities within each governmental branch, assessing how judges, executive leaders, and legislators are influenced by bribes, personal connections, or private interests in their decision-making processes.

Regarding the political and sociopolitical and economic variables, a random forest regressor was applied to the years 2000 and 2022 to identify the most influential factors in political corruption from 20 variables,[3] considered important by the literature. Key indicators that repeatedly scored high on importance in both years included 'Freedom of Expression', 'Access to Justice for Men', 'Access to Justice for Women', and 'GDP per Capita'. Additionally, 'Health Equality' ranked second in importance for 2022, and 'Transparent Laws with Predictable Enforcement' topped the list for 2000. Therefore, these six variables were selected to be analysed in this study (Table 1A and 2A of the Appendix A show the respective scores).

Furthermore, the four corruption-related variables (political, executive, legislative and judicial) were normalized due to the use of clustering models. This allows for consistency and comparability of results across models (Han et al., 2022). Socio-political variables were also normalised, even though they were not directly involved in the clustering process. By normalising both sets of variables, a consistent framework of analysis across the different dimensions of the data is facilitated, improving the interpretability and comparability of the results. The only exception to this normalisation process was GDP per capita[4]. This variable was left unnormalised to preserve its interpretability, as its absolute values are more meaningful when not transformed. Lastly, we specifically address the inverse of the politcal corruption scores, focusing on the absence of corruption. This methodological choice improves interpretability, which is especially important when working with correlation analysis.

[2]For more information regarding the data collection and the code used in this research, see: https://github.com/MariaFernandaOrtega/Master-Thesis.git

[3]Transparent laws with predictable enforcement, Health equality, GDP per capita, Freedom of expression, Access to justice for women, Access to justice for men, Educational equality, Social class equality in respect for civil liberties, Power distributed by socioeconomic position, Freedom of discussion for women, Freedom of discussion for men, Political Violence, Power distributed by gender, Power distributed by social group, GDP growth annual percentage, civil society organizations repression, Freedom of academic and cultural expression, Government capacity to regulate online content, Political Polarization, and Government internet filtering capacity.

[4]GDP per capita adjusted for purchasing power parity (PPP) and measured in international dollars.

Methodology

Overview

This research analysis is divided into two main parts: a) the clustering of the political corruption variable for 133 countries over the period from 2000 to 2020, and b) a parallel analysis but for disaggregated data relating to legislative, executive and judicial corruption. The models used for clustering include Time Series K-means with various distance metrics (euclidean, dtw and softdtw) and Kernel K-means. Subsequently, two metrics, Silhouette Score and Davies-Bouldin Score, were used to evaluate the clustering models, and the optimal number of clusters was determined using the elbow method. Finally, correlations between the selected socio-political and economic variables and political corruption were calculated for each cluster.

Clustering Techniques

1. **Time Series K-means**: This clustering method is specifically adapted for time series data. The model partitions (n) observations into (k) clusters where each observation belongs to the cluster with the nearest mean time series pattern (Witten, Frank, & Hall, 2011). The time series K-means is executed using different distance metrics:

   - **Euclidean Distance**: Method that measures the straight-line distance between two points. For time series, it calculates the sum of squared differences between corresponding points in two sequences (Witten, Frank, & Hall, 2011).

   $$d(x, y) = \sqrt{\sum_{t=1}^{T}(x_t - y_t)^2}$$

   where, x and y are two time series of length T.

   - **Dynamic Time Warping (DTW)**: Technique that calculates an optimal match between two given sequences with certain restrictions. The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension (Bishop, 2006).

   $$DTW(x, y) = \min_{\pi} \left(\sum_{t=1}^{T}(x_{\pi(t)} - y_t)^2 \right)^{1/2}$$

   where (t) represents the warping of time indices to align the two sequences.

   - **Soft Dynamic Time Warping (Soft-DTW)**: Method that computes a variation of the traditional Dynamic Time Warping (DTW) that enhances the adaptability and applications of the original method by introducing a differentiable loss function (Cuturi & Blondel, 2017).

$$\text{SoftDTW}(x, y) = \min\left(\sum_{t=1}^{T}\sum_{s=1}^{S} D_{t,s} \cdot e^{-\gamma \cdot D_{t,s}}\right)$$

where $D_{t,s}$ represents the squared Euclidean distance between elements x_t of series x and y_s of series y. The parameter γ is a non-negative smoothing parameter that controls the softness of the minimum; larger values lead to a softer minimum.

2. **Kernel K-means**: This technique extends from the standard K-means clustering algorithm, designed to handle non-linear patterns in data through the use of kernel functions. This approach is particularly effective for complex datasets, such as time series, where the underlying patterns and relationships might not be linear or might exist in a higher-dimensional space (Soheily-Khah, Douzal-Chouakria, & Gaussier, 2016). First, a kernel function is chosen to transform the time series data. Common choices include the Gaussian kernel, polynomial kernels, sigmoid kernels and the Global Alignment Kernel (GAK) (Dhillon, Guan, & Kulis, 2004). In this study GAK is used for its demonstrated efficacy in time series classification, and its advantages over the Euclidean distance in handling temporal distortions (Cuturi, 2011). The GAK kernel is defined as:

$$K_{\text{GAK}}(X, Y) = \sum_{a,b} e^{-\frac{\|X_a - Y_b\|^2}{2\sigma^2}}$$

where: X_a and Y_b are the elements (or subsequences) of the time series X and Y, respectively. σ is a scaling parameter that adjusts the flexibility of the alignment.The exponential function ensures that closer alignments contribute more significantly to the kernel value, providing robustness against noise and small misalignments.

Furthermore, the goal of Kernel K-means is to partition the data into k clusters by minimizing the within-cluster distances in the feature space induced by the kernel. The objective function can be defined as:

$$\min \sum_{k=1}^{K} \sum_{i \in S_k} \|\phi(x_i) - \mu_k\|^2$$

Here, $\phi(x_i)$ denotes the mapping of data point x_i into the feature space, and μ_k represents the mean of cluster k in this space.

Model Evaluation Metrics

To assess the quality of the clustering models, the following metrics are employed:

1. **Silhouette Score**: Measures how similar an object is to its own cluster compared to other clusters (Rousseeuw, 1987). The silhouette score for a set of samples is given by:

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

where (a) is the mean intra-cluster distance, and (b) is the mean nearest-cluster distance for each sample. Values close to 1 indicate samples well matched to their own cluster and poorly matched to neighboring clusters.

2. **Davies-Bouldin Score**: This index signifies the average 'similarity' between clusters, where lower values mean the clusters are better separated (Davies & Bouldin, 1979):

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

σ is the average distance of all elements in cluster (i) to centroid (c_i), and d(c_i, c_j) is the distance between centroids (c_i) and (c_j).

Optimal Number of Clusters

In this research, the elbow method is used to determine the optimal number of clusters within the data. This method is especially useful for clustering algorithms, such as K-means, where the optimal number of clusters, denoted k, is not predetermined. The method is based on the principle of variance reduction: as clusters multiply, the aggregate variance within clusters-quantified by the sum of squared distances (SSD) from each point to its cluster centroid-should decrease sharply (Syakur et al., 2018).

To effectively apply the elbow method, the SSD as a function of different values of k is plotted. The SSD is mathematically expressed as:

$$SSD(k) = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

where C_i is the i-th cluster, x is a data point in cluster C_i, and μ_i is the centroid of C_i. The objective is to minimize this quantity over all clusters.

Correlation between Sociopolitical and Economic variables and corruption clusters

To explore how the absence of corruption in the political, executive, legislative and judicial spheres is related to various socio-political and economic factors, a correlational analysis is carried out between the different clusters. This analysis consists of calculating the Pearson's correlation between the absence of corruption and each of the six sociopolitical and economic variables for each group, with the objective of understanding how these variables may influence the levels of corruption within each cluster.

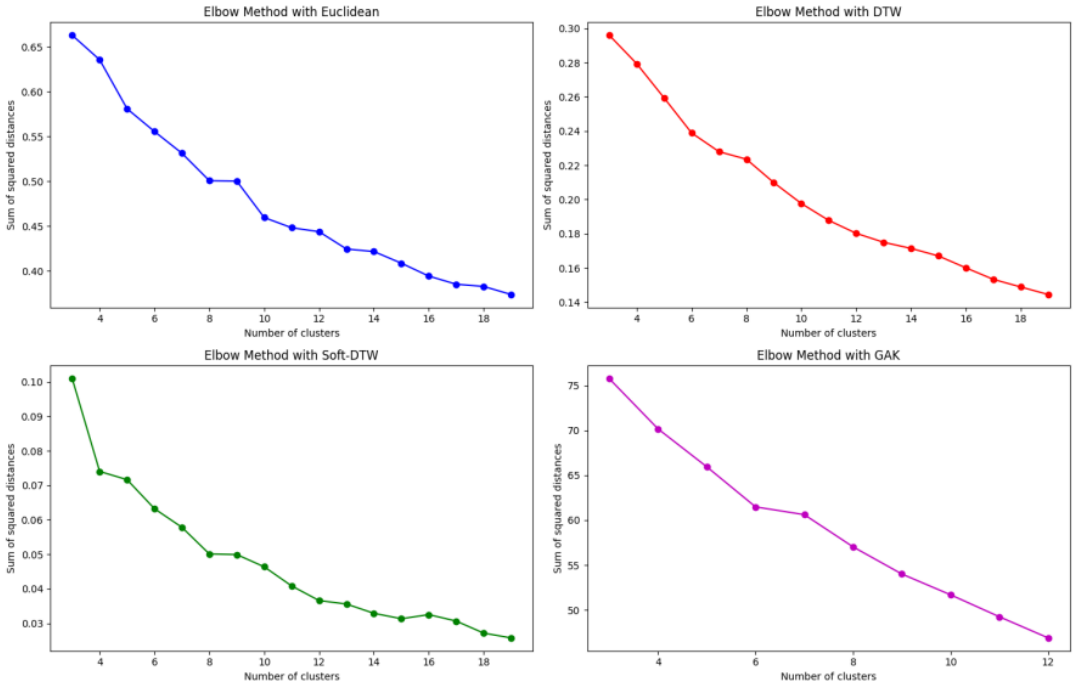
Results

a) Clustering Political Corruption

Selection of the optimal number of clusters

Figure 1 displays elbow method plots for various clustering methods, each charting the sum of squared distances against the number of clusters. The Euclidean Distance method shows the most pronounced elbow at 8-9 clusters. For Dynamic Time Warping (DTW), the elbow occurs at 6-8 clusters. Soft-DTW has a clear elbow at 8 clusters, and Kernel K-means, using the Global Alignment Kernel (GAK), presents an elbow at 6 clusters. Based on these results, the clustering interval of 6-9 was chosen to test and evaluate the models. This range reflects a pragmatic approach to capturing the intrinsic clustering of the data while maintaining simplicity and interpretability of the results.

Fig.1 Elbow Method for Various Methods



Clustering Model Selection

This section evaluates the most effective clustering algorithm for clustering the level of political corruption in the 133 countries across time. Cluster sizes, ranging from 6 to 9, were established using the elbow method. Two main metrics were used in the analysis: the Silhouette score, higher values indicate better clustering, and the Davies-Bouldin score, lower scores reflect better separation and compactness.

According to Table 1, the Kernel K-means model with 7 clusters emerges as the preferred choice for our analysis. It achieves a Davies-Bouldin index of 0.746, which is the lowest among all the tested models and configurations, indicating optimal cluster separation. Additionally, its Silhouette score of 0.426 is comparatively high, suggesting strong intra-cluster cohesion relative to the other models.

Table 1: Clustering Model Evaluaion

Model	Clusters	Silhouette	Davies-Bouldin
TimeSeriesK-means: euclidean	6	0.421	0.835
TimeSeriesK-means: dtw	6	0.448	1.150
TimeSeriesK-means: softdtw	6	0.456	0.908
Kernel K-means	6	0.441	1.417
TimeSeriesK-means: euclidean	7	0.388	0.911
TimeSeriesK-means: dtw	7	0.380	1.156
TimeSeriesK-means: softdtw	7	0.422	0.929
Kernel K-means	7	0.426	0.746
TimeSeriesK-means: euclidean	8	0.411	0.806
TimeSeriesK-means: dtw	8	0.356	1.018
TimeSeriesK-means: softdtw	8	0.382	0.998
Kernel K-means	8	0.397	1.111
TimeSeriesK-means: euclidean	9	0.417	0.821
TimeSeriesK-means: dtw	9	0.333	1.060
TimeSeriesK-means: softdtw	9	0.365	1.081
Kernel K-means	9	0.377	1.467

Cluster Analysis and Temporal Trends

The employment of the Kernel K-means clustering model in the absence of political corruption data, using 7 clusters, has resulted in different groupings of countries according to their corruption profiles between 2000 and 2022, shown in Table 2.

Table 2: Cluster of Countries by Political Corruption

Cluster	**Countries**
1	Albania, Algeria, Bolivia, Bosnia and Herzegovina, Ecuador, El Salvador, Ghana, Indonesia, Iran, Kenya, Malawi, Maldives, Mexico, Moldova, Mongolia, Morocco, North Macedonia, Philippines, Russia, Serbia, Solomon Islands, Thailand, Tunisia, Türkiye
2	Angola, Armenia, Azerbaijan, Burundi, Cambodia, Cameroon, Chad, Dominican Republic, Equatorial Guinea, Gabon, Guatemala, Honduras, Kazakhstan, Kyrgyzstan, Laos, Madagascar, Nicaragua, Nigeria, Paraguay, Republic of the Congo, Tajikistan, Togo, Uganda, Ukraine, Uzbekistan, Zimbabwe
3	Georgia
4	Botswana, Cape Verde, Costa Rica, Cyprus, Czechia, Greece, Israel, Italy, Jamaica, Latvia, Lithuania, Malta, Namibia, Oman, Seychelles, Slovenia, South Korea, Suriname, Trinidad and Tobago, United Arab Emirates
5	Belarus, Bulgaria, Colombia, Croatia, Hungary, Senegal, Slovakia, South Africa, Tanzania, Zambia
6	Argentina, Benin, Brazil, Burkina Faso, China, Eswatini, Ethiopia, Guyana, India, Kuwait, Lesotho, Malaysia, Mauritius, Montenegro, Mozambique, Panama, Peru, Romania, Rwanda, Saudi Arabia, Sri Lanka, The Gambia, Vanuatu, Vietnam
7	Australia, Austria, Barbados, Belgium, Canada, Chile, Denmark, Estonia, Finland, France, Germany, Hong Kong, Iceland, Ireland, Japan, Luxembourg, Netherlands, New Zealand, Norway, Poland, Portugal, Singapore, Spain, Sweden, Switzerland, United Kingdom, United States of America, Uruguay

The clusters, as visualized in Fig. 2 (clusters over time) and Fig.3 (the mean of the clusters over time), demonstrate varying temporal patterns of the variable studied 'absence of political corruption' across the 133 countries. Particularly noteworthy is Cluster 7, which shows the least political corruption among the countries studied, includes mainly countries from Western Europe, the Nordic region, North America, Oceania and Asia. In particular, Chile, Uruguay and Barbados represent Latin America and the Caribbean in this group. Following the same line, cluster 4 presents relatively low levels of corruption over the years, covering a wide geographic range that includes Africa, Southern Europe, the Middle East and the Caribbean, with South Korea being the only Asian representative. For both clusters their mean appears to have a fairly constant trend over time.

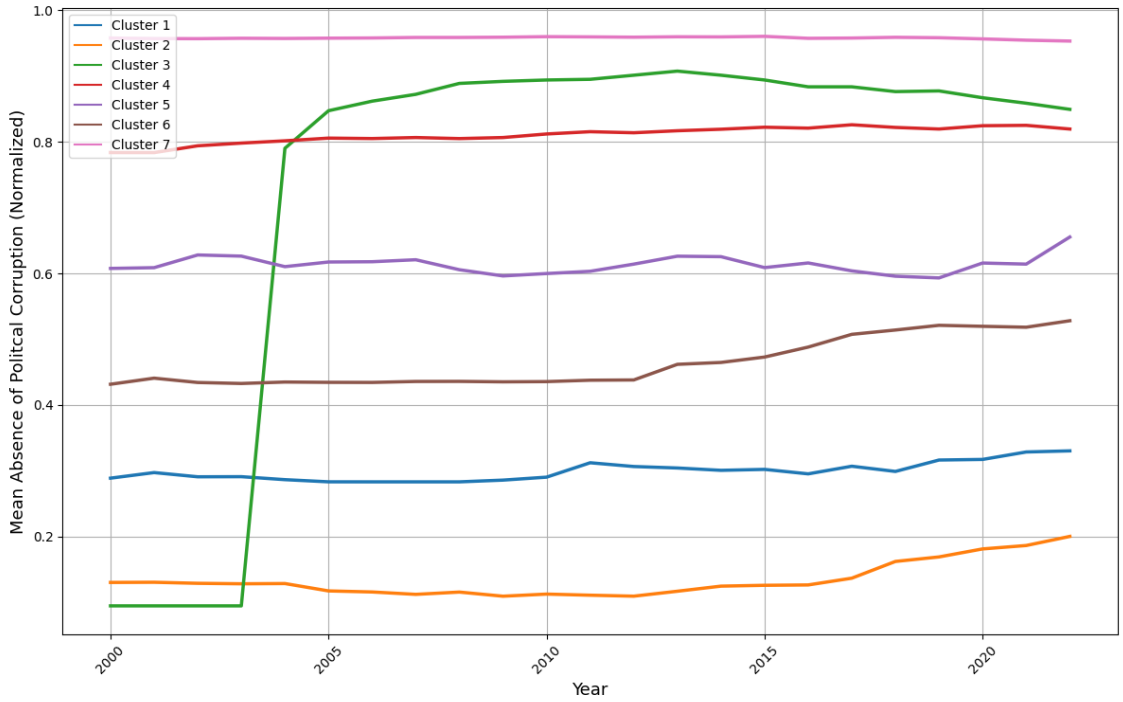
Furthermore, Clusters 5 and 6 register moderate levels of corruption over the 22-year period, compared to the other clusters. Cluster 6, in particular, has shown gradual improvements since 2012. Cluster 5 includes countries in Eastern Europe, Africa and Colombia as the only South American country. Group 6, meanwhile, includes countries in East and South Asia, Africa, South America, the Middle East and Vanuatu in Oceania.

On the other hand, cluster 2, which registers the highest levels of political corruption, has only experienced slight improvements since roughly 2015. It includes countries mainly from Africa, Latin America, Central Asia and from Eastern Europe, Ukraine. Similarly, Cluster 1 shows high levels of political corruption and a stable trend, includes a wide range of countries from the Balkans, Eastern Europe, Africa, Latin America and Asia, with Turkey as the only country from Southern Europe. Lastly, Cluster 3 stands unique

with a single country, Georgia, which exhibited a significant improvement in corruption levels around 2004.

Fig.2 Absence of Political Corruption Over Time by Cluster

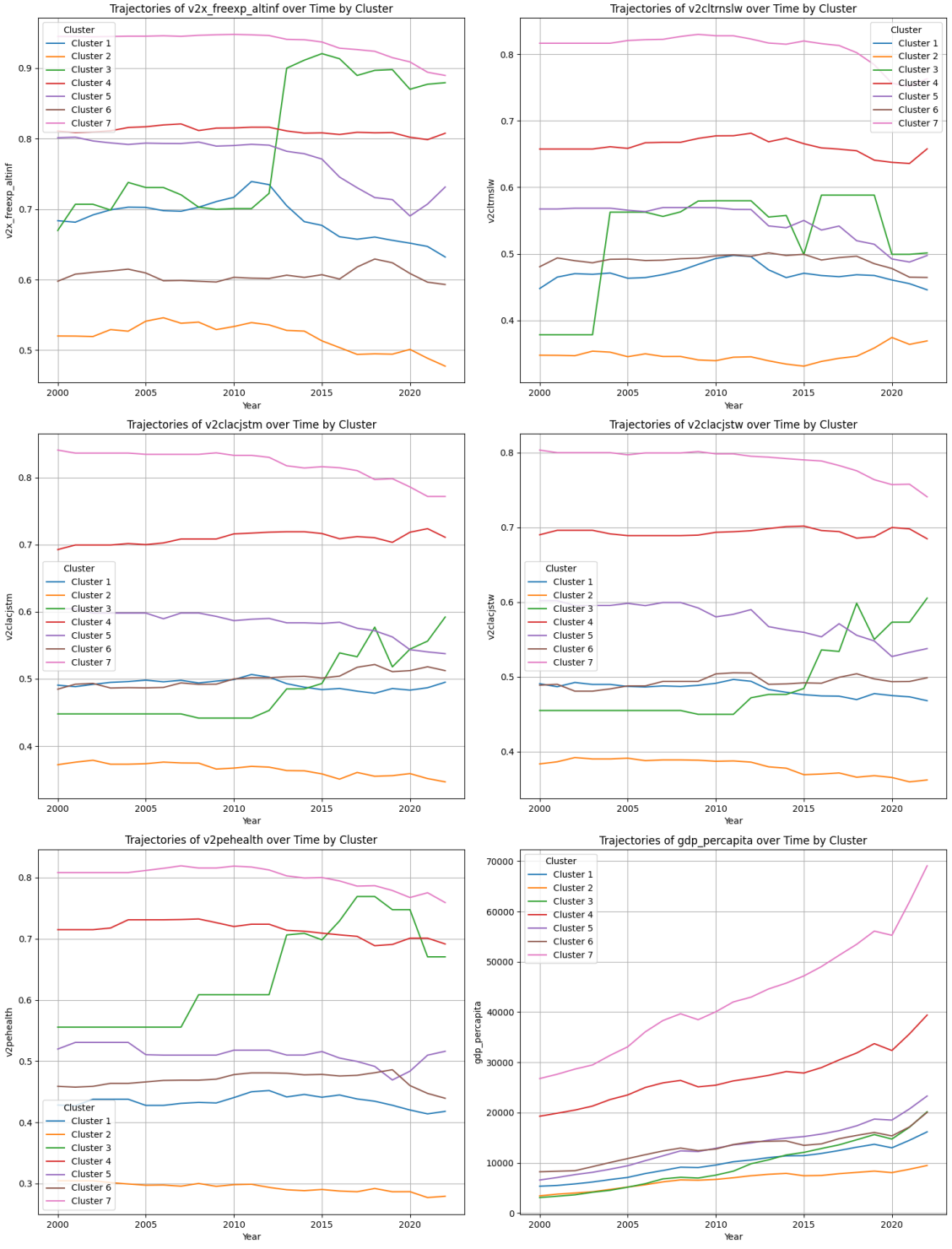
Fig.3 Mean Absence of Political Corruption by Cluster



Analysis of Sociopolitical and Economic Variables by Cluster Over Time

This section delves into the interaction between selected socio-political and economic variables and the absence of political corruption. Building on the established framework of our analysis, Figure 4 illustrates the trajectories of how variables such as freedom of expression(v2x_freexp_altinf), law transparency (v2cltrnslw), access to justice for men(v2clacjstm), access to justice for women (v2clacjstw), health equality (v2pehealth), and GDP per capita (gdp_percapita) evolve over time, showing distinct trajectories across the clusters.

Fig.4 Trajectories of Sociopolitical and Economic Variables Over Time by Cluster

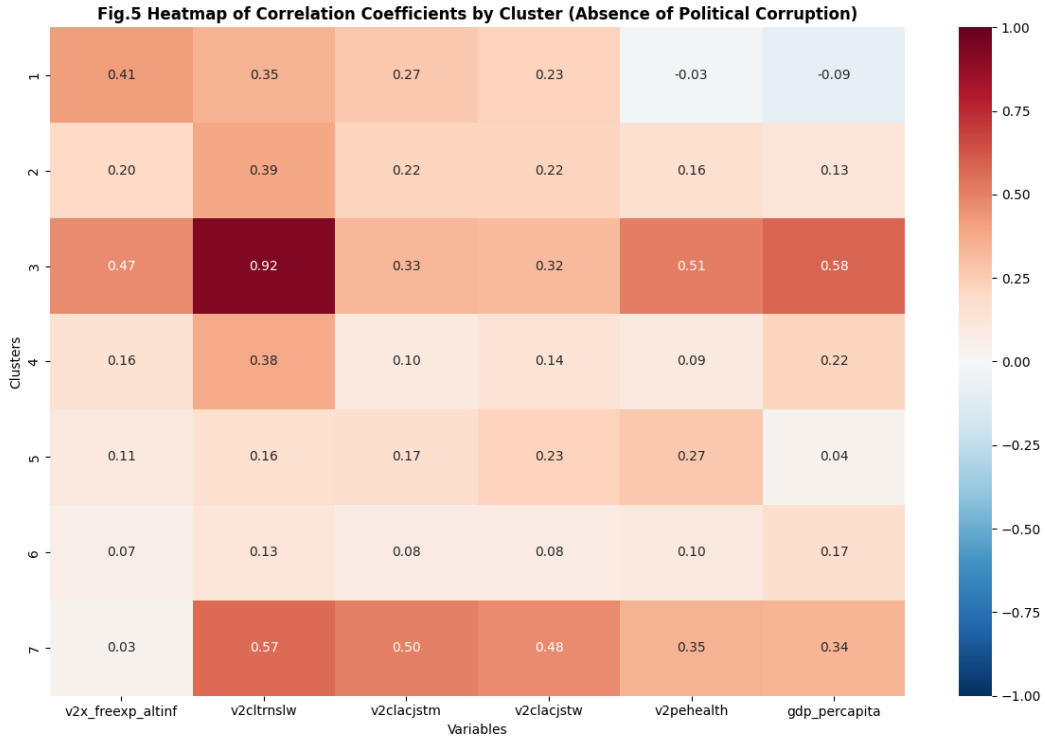


Furthermore, Figure 5 shows a heatmap with the correlation coefficients between the absence of political corruption, and each sociopolitical variable to quantify the strength and direction of relationships within each cluster:

- **Cluster 1, 2, 4, 5 and 6** show generally low correlation values across all variables, suggesting a weaker relationship between the absence of political corruption and the sociopolitical and economic variables.

- **Cluster 3**, composed by Georgia, exhibits a strong positive correlation with respect to law transparency (0.92). Health equality (0.51) and GDP per capita (0.58) also show moderately strong positive correlations.
- **Cluster 7,** consisting of countries with high levels of absence of political corruption, shows moderately strong positive correlations, especially in law transparency (0.57) and access to justice for men (0.50).

These results suggest that, among the clusters, only cluster 3 demonstrate a strong and positive relationship between political corruption and law transparency.



b) Clustering Executive, Legislative, and Judicial Corruption

Clustering Model Selection

As with political corruption, the Kernel K-means model with 7 clusters has been selected for the executive, legislative and judicial corruption data. This decision is based on the performance of the model, which has outperformed the others in at least one of the key evaluation parameters - the Silhouette and Davies-Bouldin scores - and has demonstrated competitive or superior results in the remaining metric. The tables with the evaluation results for each branch of corruption are documented in the Appendix B in Tables 1B, 2B, and 3B.

Executive Corruption

Cluster Analysis and Temporal Trends

The implementation of the Kernel K-means clustering model in the absence of corruption in the executive branch, with 7 clusters, has resulted in different groupings of countries according to their corruption profiles between 2000 and 2022, as shown in Table 3.

Table 3: Cluster of Countries by Executive Corruption

Cluster	Countries
1	Bulgaria, Colombia, Czechia, Malta, Namibia, Oman, Peru, Rwanda, Seychelles, Tanzania, Vietnam, Zambia
2	Argentina, Brazil, Burkina Faso, China, Ecuador, El Salvador, Eswatini, Ethiopia, Guyana, Hungary, India, Iran, Kuwait, Lesotho, Maldives, Mauritius, Mexico, Mongolia, Morocco, Mozambique, Panama, Serbia, South Africa, Sri Lanka, Vanuatu
3	Benin, Georgia, Romania, Tunisia
4	Albania, Algeria, Armenia, Belarus, Bolivia, Bosnia and Herzegovina, Burundi, Dominican Republic, Gabon, Ghana, Guatemala, Honduras, Indonesia, Kenya, Laos, Madagascar, Malawi, Malaysia, Moldova, Montenegro, Nicaragua, North Macedonia, Paraguay, Philippines, Russia, Saudi Arabia, Solomon Islands, Thailand, The Gambia, Togo, Türkiye, Uganda, Ukraine
5	Australia, Austria, Belgium, Canada, Chile, Denmark, Estonia, Finland, France, Germany, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Netherlands, New Zealand, Norway, Poland, Portugal, Singapore, Spain, Sweden, Switzerland, United Kingdom, United States of America, Uruguay
6	Barbados, Botswana, Cape Verde, Costa Rica, Croatia, Cyprus, Greece, Hong Kong, Israel, Jamaica, Japan, Senegal, Slovakia, Slovenia, South Korea, Suriname, Trinidad and Tobago, United Arab Emirates
7	Angola, Azerbaijan, Cambodia, Cameroon, Chad, Equatorial Guinea, Kazakhstan, Kyrgyzstan, Nigeria, Republic of the Congo, Tajikistan, Uzbekistan, Zimbabwe

Furthermore, the clusters, as visualized in Fig. 6 (clusters over time) and Fig.7 (the mean of the clusters over time), demonstrate varying temporal patterns of the variable studied 'absence of corruption in the executive branch' across the 133 counties. Firstly, cluster 1, shows slight improvement from relatively low-moderate levels of Executive corruption and encompass countries mainly from Africa and a few from Eastern Europe, Latin America, Middle East and Southeast Asia.

Cluster 2 is mainly composed by countries from Africa, Asia, Latin America, and a few from Eastern Europe (Hungary, Serbia). The corruption levels here are moderate and stable, compared to the other clusters, indicating a persistent challenge to decrease corruption in the executive branch. In contrast, Cluster 3, stands out for its significant improvement across the years. It is the smallest cluster, with only four countries, from Eastern Europe (Romania and Georgia) and Africa (Benin and Tunisia).

Furthermore, cluster 4 represents a broad mix of countries, mainly from Eastern Europe, Africa, Latin America, Southeast Asia, and Saudi Arabia. Russia, and the Solomon Islands

are also included in this cluster. In the last years analysed, it shows a slightly upward trend from low absence of corruption in the executive power to moderate levels, which suggests some progress, although challenges remain. Cluster 5 portrays a different scenario, where predominantly Western and Northern European countries, along with some countries from North America, Oceania , Latin America, Southern Europe, and Eastern Asia show an enduring absence of corruption in the executive branch.

Moreover, cluster 6, with countries from Southern Europe, the Middle East, East Asia, and Latin America and the Caribbean, exhibits relatively low levels of Executive corruption, which have remained fairly constant over time. Lastly, cluster 7, consists mainly of Central Asia and Central African countries. This is the cluster with the highest corruption levels in the Executive, however, in recent years there has been a slow trend towards improvement.

Fig.6 Absence of Corruption in the Executive Branch Over Time by Cluster

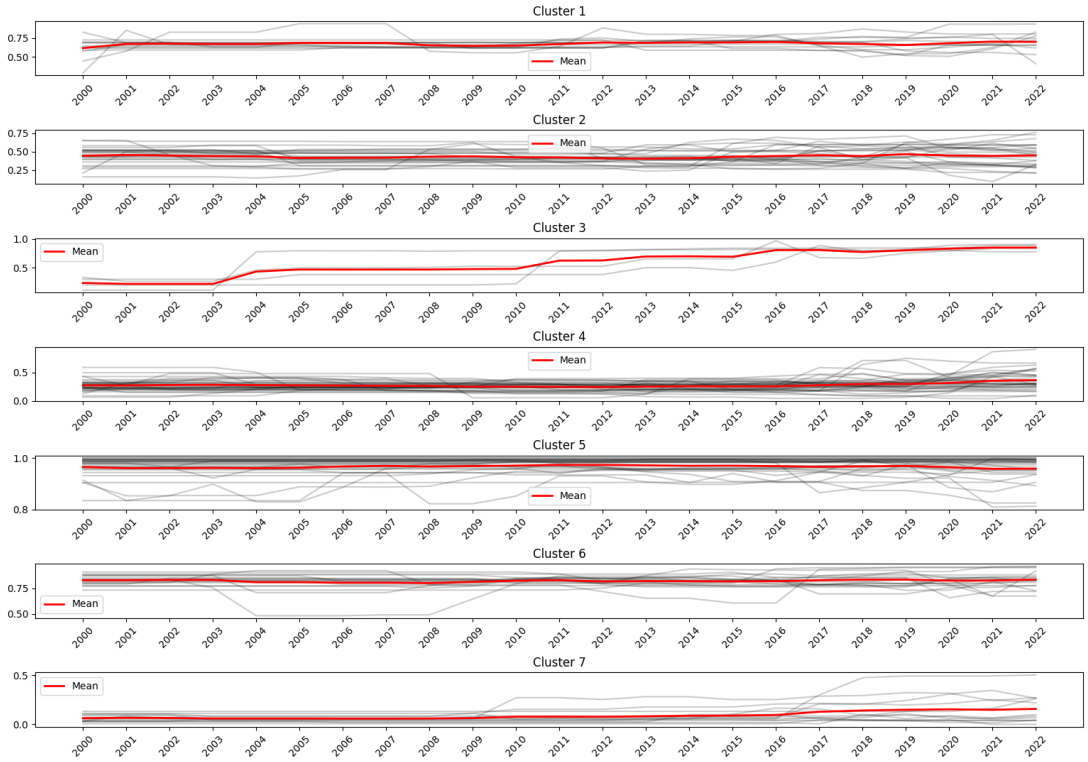
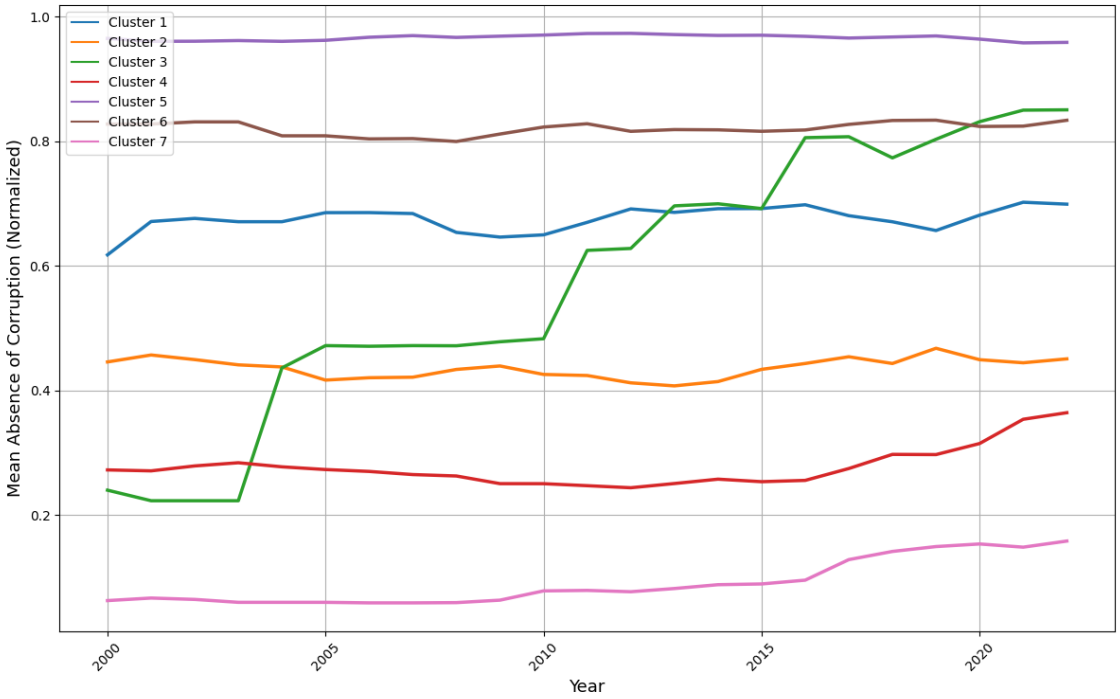


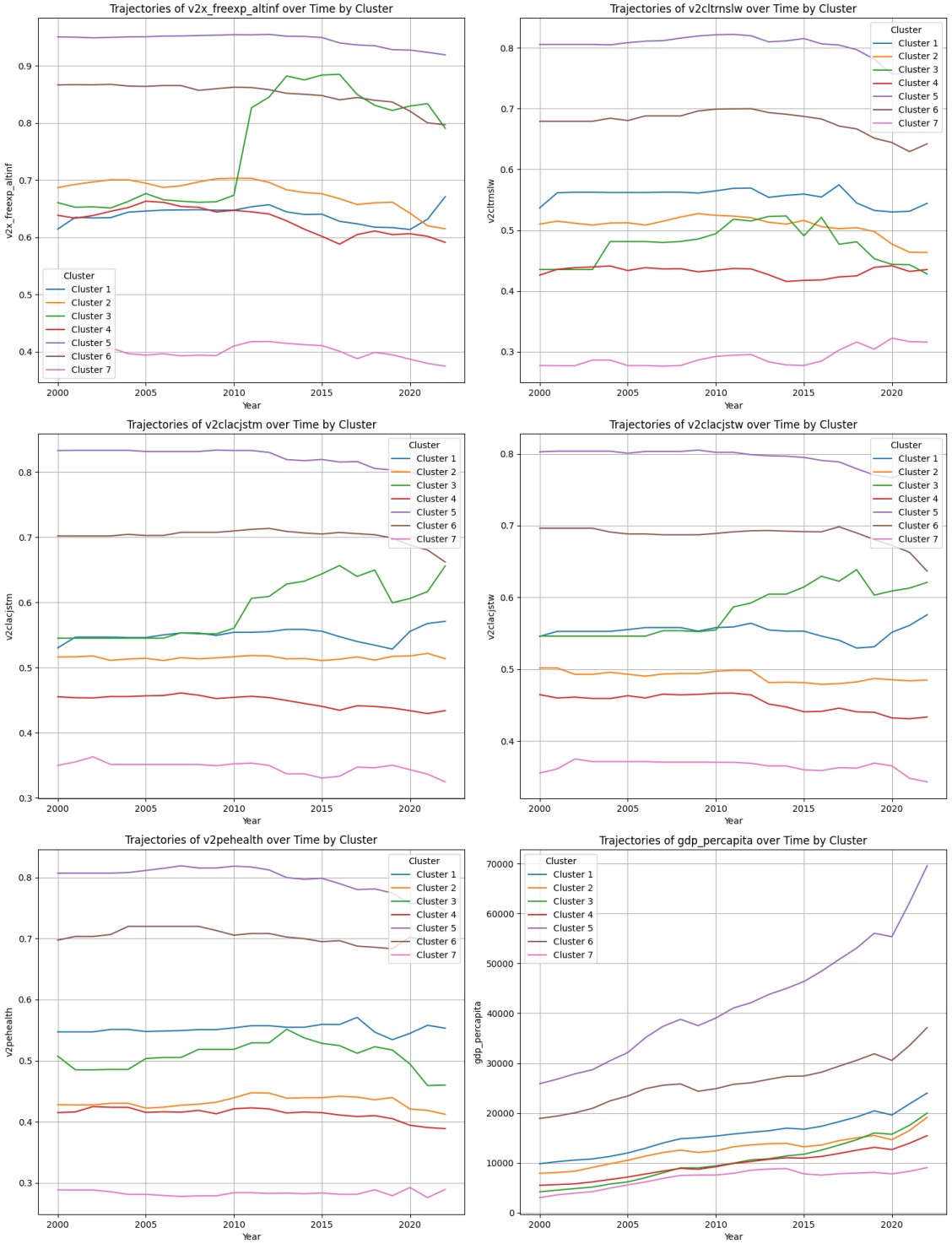
Fig.7 Mean Absence of Corruption in the Executive Branch by Cluster



Analysing Sociopolitical and economic variables across clusters

Based on the established framework of our analysis, Figure 8 illustrates the trajectories of how variables such as freedom of expression (v2x_freexp_altinf), transparency of law (v2cltrnslw), access to justice for men (v2clacjstm), access to justice for women (v2clacjstw), health equality (v2pehealth) and GDP per capita (gdp_percapita) evolve over time, showing different trends for different clusters.

Fig.8 Trajectories of Sociopolitical and Economic Variables Over Time by Cluster

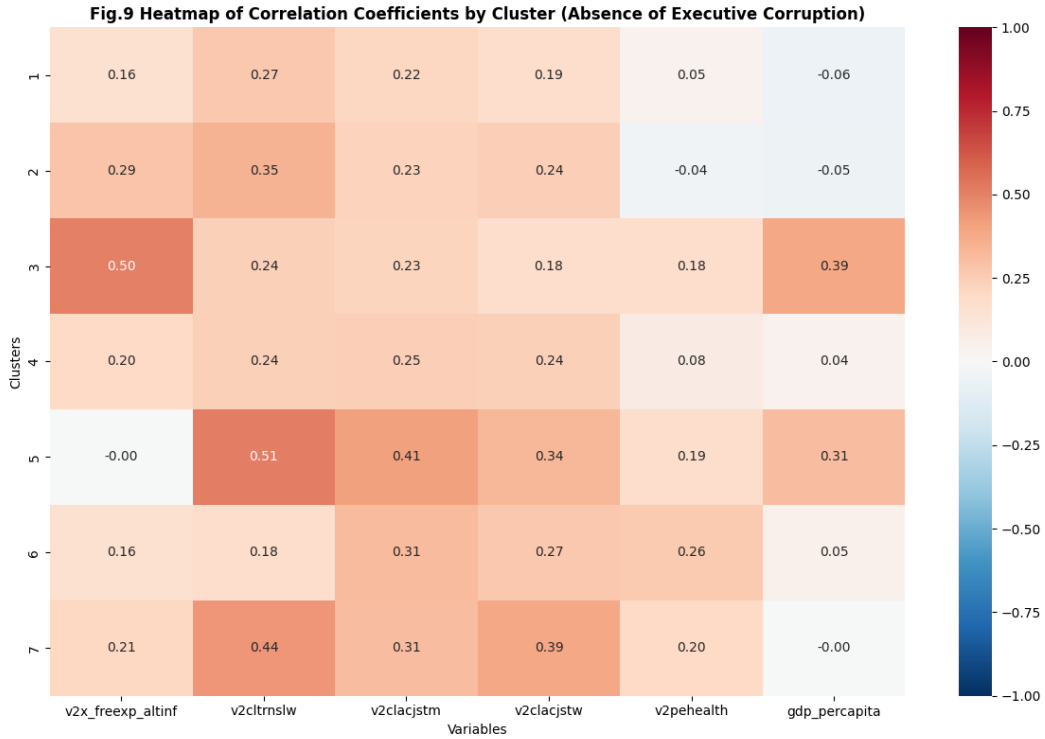


Furthermore, Figure 9 presents a heatmap of the correlation coefficients. This quantifies the strength and direction of the relationships between the absence of executive corruption and the sociopolitical variables in each cluster:

- Cluster 1, 2, 4, 6 and 7 show weaker correlation values across all variables, suggesting a weaker linear relationship between the absence of executive corruption and these sociopolitical and economic variables.

- Cluster 3, consisting only of Benin, Georgia, Romania and Tunisia, shows a moderately strong positive correlation for freedom of expression (0.50). However, it shows no strong correlations for the other variables.
- Cluster 5, composed of countries with high levels of absence of executive corruption, also shows a moderately strong positive correlation for transparency of the law (0.51) and weaker correlations for the other socio-political and economic variables.

These results illustrate that the majority of clusters demonstrate generally weaker correlations, indicating a less pronounced linear relationship between the absence of executive corruption and the sociopolitical and economic variables examined in this study.



Legislative Corruption

Cluster Analysis and Temporal Trends

Fitting the Kernel K-means clustering model to the data on the absence of corruption in the legislative branch, using 7 clusters as the selected parameter, resulted in various groupings of countries based on their corruption profiles from 2000 to 2022, as shown in Table 4.

Table 4: Cluster of Countries by Legislative Corruption

Cluster	**Countries**
1	Albania, Armenia, Brazil, Burundi, Cameroon, Ecuador, Gabon, Honduras, India, Iran, Kazakhstan, Kenya, Laos, Maldives, Mexico, Moldova, Nicaragua, Paraguay, Peru, Philippines, Republic of the Congo, Russia, Solomon Islands, Sri Lanka, Tajikistan, Thailand, Türkiye, Uganda, Ukraine, Zimbabwe
2	Tunisia
3	Australia, Barbados, Belgium, Canada, Chile, Estonia, France, Germany, Hong Kong, Iceland, Ireland, Netherlands, Slovenia, Spain, Switzerland, United Arab Emirates, United Kingdom
4	Denmark, Finland, Luxembourg, New Zealand, Norway, Singapore, Sweden, Uruguay
5	Azerbaijan, Chad, Dominican Republic, Guatemala, Indonesia, Kyrgyzstan, Madagascar, Nigeria
6	Algeria, Angola, Bolivia, Bosnia and Herzegovina, Bulgaria, Cambodia, Colombia, Croatia, El Salvador, Equatorial Guinea, Hungary, Kuwait, Malawi, Malaysia, Mauritius, Mongolia, Montenegro, Morocco, North Macedonia, Panama, Romania, Rwanda, Serbia, Slovakia, South Africa, Togo, Uzbekistan, Vanuatu
7	Argentina, Austria, Belarus, Benin, Botswana, Burkina Faso, Cape Verde, China, Costa Rica, Cyprus, Czechia, Eswatini, Ethiopia, Georgia, Ghana, Greece, Guyana, Israel, Italy, Jamaica, Japan, Latvia, Lesotho, Lithuania, Malta, Mozambique, Namibia, Oman, Poland, Portugal, Saudi Arabia, Senegal, Seychelles, South Korea, Suriname, Tanzania, The Gambia, Trinidad and Tobago, United States of America, Vietnam, Zambia

Moreover, the clusters, as visualized in Fig. 10 (clusters over time) and Fig.11 (the mean of the clusters over time), demonstrate varying temporal trends of the variable studied 'absence of corruption in the legislative branch' across the 133 counties. In the case of cluster 1, it shows a relatively low mean score, and it remains fairly constant over time. It comprises countries from a broad range of regions, including Eastern Europe, Latin America, Africa, Asia, one country of the Middle East (Iran) and one of Oceania (Solomon Islands). Conversely, cluster 2 is unique, consisting only of Tunisia from North Africa. It notably increases its absence of legislative corruption around 2010 to then have a downward trend after 2012, but maintaining moderate levels compared to the beginning of the period analysed.

On the other hand, cluster 3 is predominantly composed of Western Europe, and countries such as Canada, Australia, United Arab Emirates, Hong Kong, Chile and Barbados. It shows low-moderate levels of legislative corruption and the mean for this cluster remains constant over time. In the same line, cluster 4 shows the lowest levels of legislative corruption with respect to the other groups and remains constant over time. This cluster includes the Nordic countries and others such as New Zealand, Singapore and Uruguay.

In contrast, cluster 5 has the highest levels of legislative corruption and has worsened slightly in recent years. It includes countries mainly in Africa and Central Asia, alongside the Dominican Republic, Guatemala, and Indonesia. Furthermore, cluster 6 shows a constant mean trend over time with high-moderate leveles of legislative corruption, compared to the other clusters. It is composed mainly by countries from Africa, Asia, Eastern

Europe, and Latin America. Lastly, cluster 7 is composed of several regions, including Europe, Asia, Africa and Latin America and the Caribbean. It shows modest levels of legislative corruption compared to other clusters and shows a slight improvement over the year in terms of reducing such corruption.

Fig.10 Absence of Corruption in the Legislative Branch Over Time by Cluster

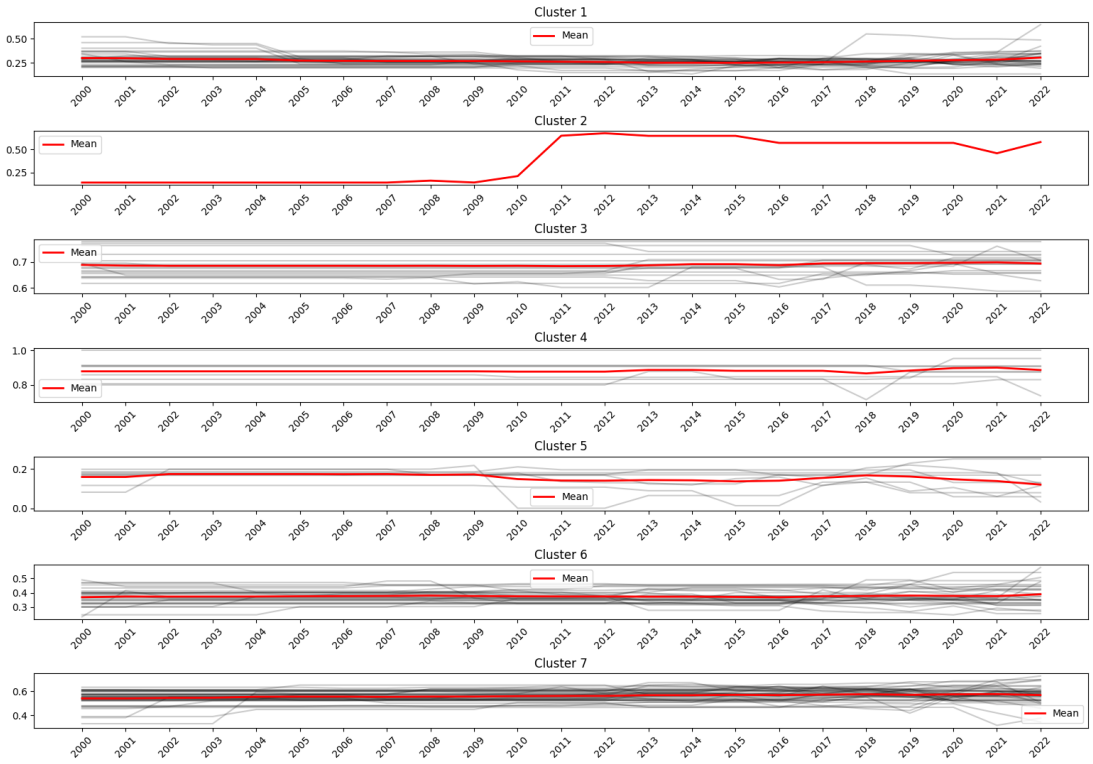
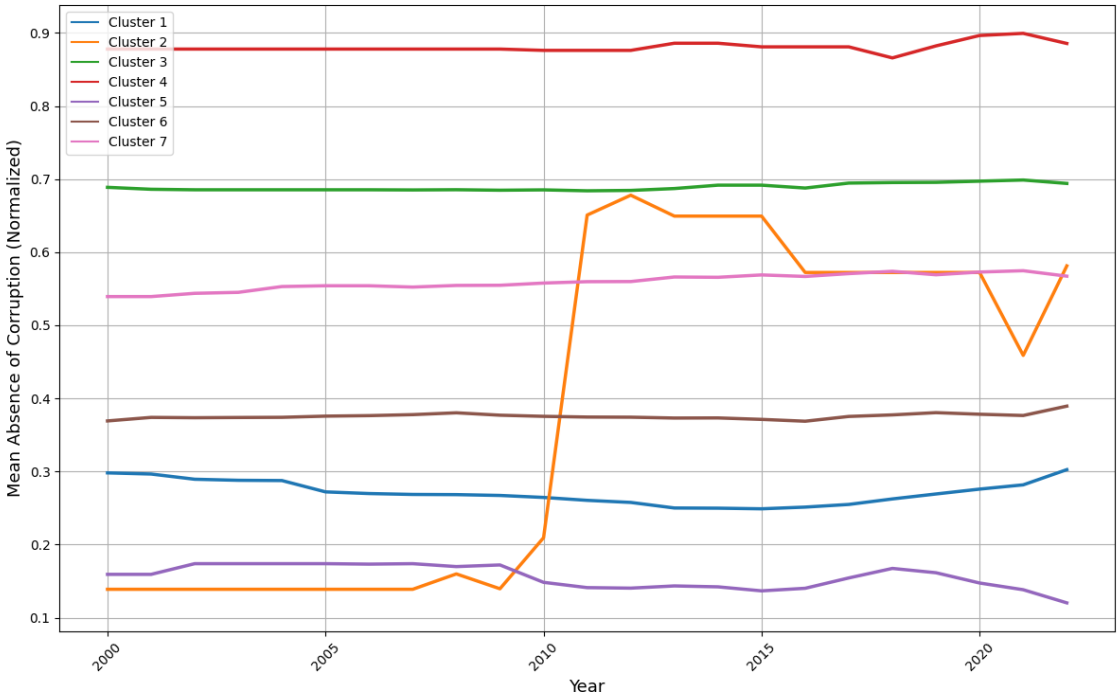


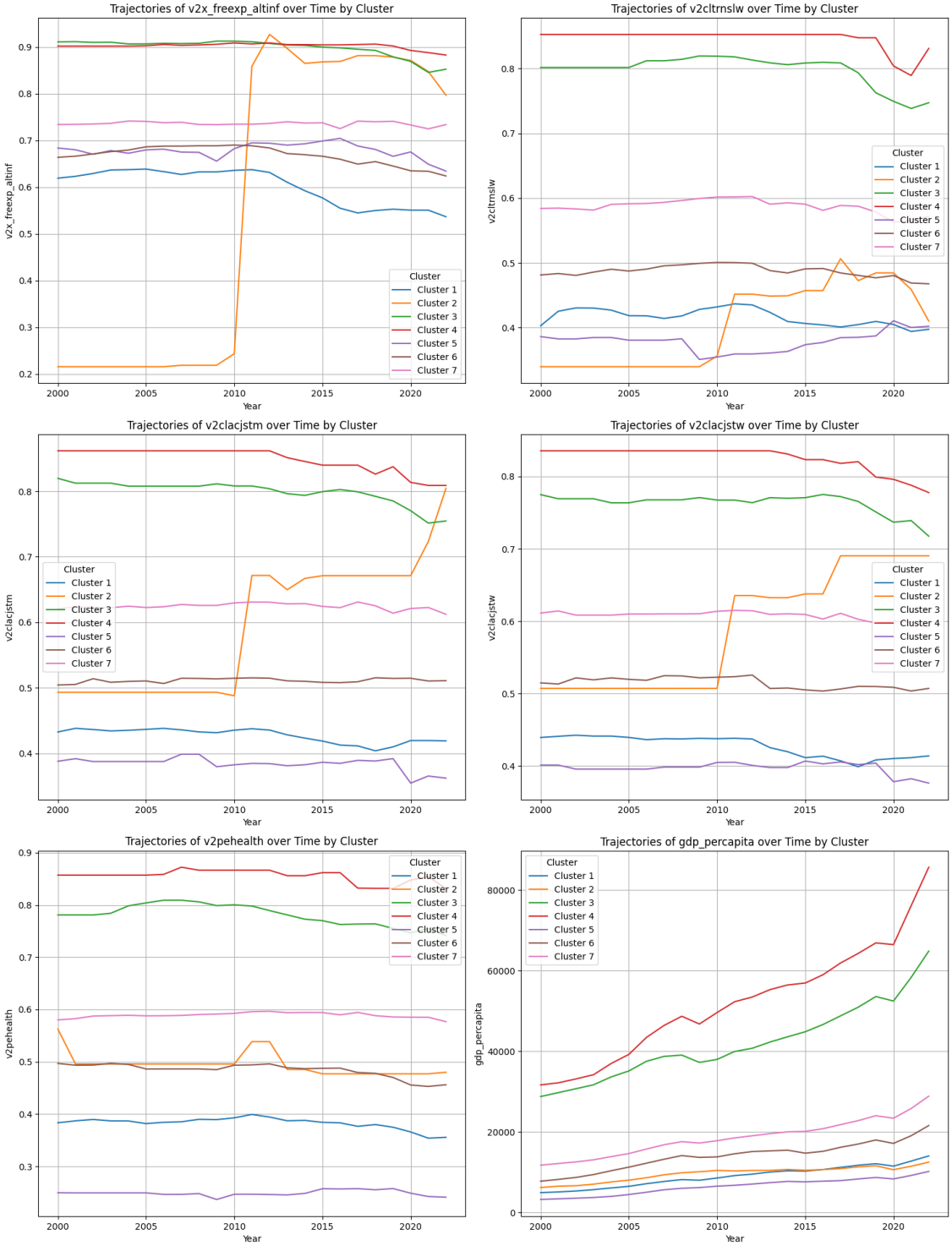
Fig.11 Mean Absence of Corruption in the Legislative Power by Cluster



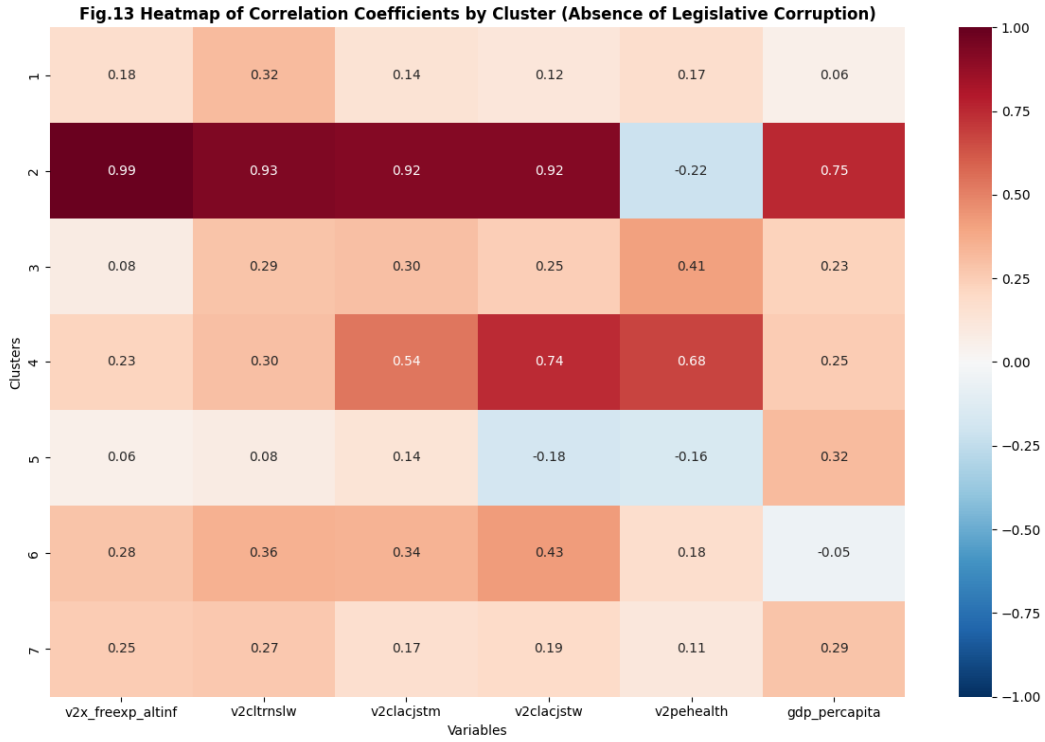
Analysing Socioeconomic variables across clusters

This section delves into the interaction between selected socio-political and economic variables and the absence of corruption in the legislative branch. Based on the established framework of our analysis, Figure 12 illustrates the trajectories of how variables such as freedom of expression (v2x_freexp_altinf), transparency of law (v2cltrnslw), access to justice for men (v2clacjstm), access to justice for women (v2clacjstw), health equality (v2pehealth) and GDP per capita (gdp_percapita) evolve over time, showing different trends for different clusters.

Fig.12 Trajectories of Sociopolitical and Economic Variables Over Time by Cluster



In addition, Figure 13 presents a heatmap of the correlation coefficients. This quantifies the strength and direction of the relationships between the absence of legislative corruption and the sociopolitical variables in each cluster:



- Clusters 1, 3, 5, 6 and 7 show weak correlations between the level of absence of corruption in the legislative brach and socio-political and economic variables.
- Cluster 2, consisting only of Tunisia, shows remarkably strong correlations for most variables, especially with respect to freedom of expression (0.99), transparency of the law (0.93), access to justice for men (0.92) and women (0.92), and GDP per capita (0.75). However, it has a weak negative correlation with health equality (-0.22).
- Cluster 4, composed of countries with high levels of absence of legislative corruption, shows strong correlations for access to justice for women (0.74) and health equality (0.68), and moderately strong with respect to access to justice for men (0.54).

These results illustrate that most clusters demonstrate generally weaker correlations, indicating a less pronounced linear relationship between the absence of legislative corruption and the sociopolitical and economic variables selected.

Judicial Corruption

Cluster Analysis and Temporal Trends

Fitting the Kernel K-means clustering model to the data on the absence of corruption in the judicial branch, using 7 clusters as the selected parameter, resulted in various groupings of countries based on their corruption profiles from 2000 to 2022, as shown in Table 5.

Table 5: Cluster of Countries by Judicial Corruption

Cluster	**Countries**
1	Barbados, Cape Verde, Costa Rica, Cyprus, Czechia, Hungary, Jamaica, Latvia, Luxembourg, Malta, Namibia, Oman, Portugal, Slovenia, South Africa, South Korea, Suriname, Trinidad and Tobago
2	Georgia
3	Albania, Algeria, Angola, Argentina, Benin, Bosnia and Herzegovina, Burkina Faso, China, Dominican Republic, Ecuador, El Salvador, Guatemala, Guyana, Indonesia, Iran, Kenya, Laos, Lesotho, Malawi, Mexico, Moldova, Mongolia, Morocco, Mozambique, North Macedonia, Panama, Peru, Romania, Russia, Rwanda, Senegal, Serbia, Tanzania, Thailand, Türkiye, Zimbabwe
4	Australia, Austria, Belgium, Botswana, Chile, Estonia, Finland, France, Germany, Ireland, Israel, Japan, Netherlands, Poland, Singapore, Spain, Switzerland, United Arab Emirates, United Kingdom, United States of America, Uruguay
5	Belarus, Brazil, Bulgaria, Colombia, Croatia, Eswatini, Greece, India, Italy, Kuwait, Lithuania, Malaysia, Mauritius, Montenegro, Saudi Arabia, Seychelles, Slovakia, Solomon Islands, Sri Lanka, The Gambia, Vanuatu, Zambia
6	Armenia, Azerbaijan, Bolivia, Burundi, Cambodia, Cameroon, Chad, Equatorial Guinea, Ethiopia, Gabon, Ghana, Honduras, Kazakhstan, Kyrgyzstan, Madagascar, Maldives, Nicaragua, Nigeria, Paraguay, Philippines, Republic of the Congo, Tajikistan, Togo, Tunisia, Uganda, Ukraine, Uzbekistan, Vietnam
7	Canada, Denmark, Hong Kong, Iceland, New Zealand, Norway, Sweden

Moreover, the clusters, as visualized in Fig. 14 (clusters over time) and Fig.15 (the mean of the clusters over time), demonstrate varying temporal trends of 'absence of corruption in the judicial branch' across the 133 counties. Firstly, cluster 1 presents an amalgamation of countries that span across the Caribbean, Sub-Saharan Africa, Southern and Eastern Europe, alongside countries like Costa Rica, Oman, Luxembourg and South Korea. These countries show a relatively moderate and stable level of judicial integrity over the two decades. On the other hand, cluster 2, which includes only Georgia, shows a significant improvement trend. This Eurasian country experienced a notable upturn in the absence of corruption in its judicial system around 2003, which then stabilized at a relatively high level.

Cluster 3 is very diverse in terms of geographic distribution. The Balkans, Eastern Europe, Sub-Saharan Africa, North Africa, East Asia, Southeast Asia and Latin America, and Iran from the Middle East are represented in this cluster. These countries have high levels of corruption in the judiciary. However, the average of this cluster shows a slight trend of improvement in recent years. In contrast, cluster 4, predominantly made up of countries from Western and Northern Europe, together with countries from North America, Oceania and some nations in Asia (Japan and Singapore), Latin America (Uruguay and Chile), the Middle East (Israel and the United Arab Emirates) and Oceania (Australia), shows a high score of absence of corruption in the judiciary, compared to the other clusters. However, it has declined slightly in recent years.

Cluster 5 incorporates countries in Southern and Eastern Europe, South and Southeast Asia, Latin America, Sub-Saharan Africa, the Middle East, and some islands in Oceania. These nations show consistently moderate levels of absence of corruption in the judicial

system, improving slightly over time. On the other hand, cluster 6 encompasses countries in Asia, Africa and Latin America. These countries show the lowest levels of absence of corruption in the judiciary and a slow trend of improvement is observed in recent years. Finally, cluster 7, consisting mainly of the Nordic countries, Canada, Hong Kong and New Zealand, exhibits very high levels of absence of corruption in their judiciary. There is a slight downward trend over the years, but these countries continue to maintain the lowest levels of corruption compared to the other countries.

Fig.14 Absence of Corruption in the Judicial Branch Over Time by Cluster

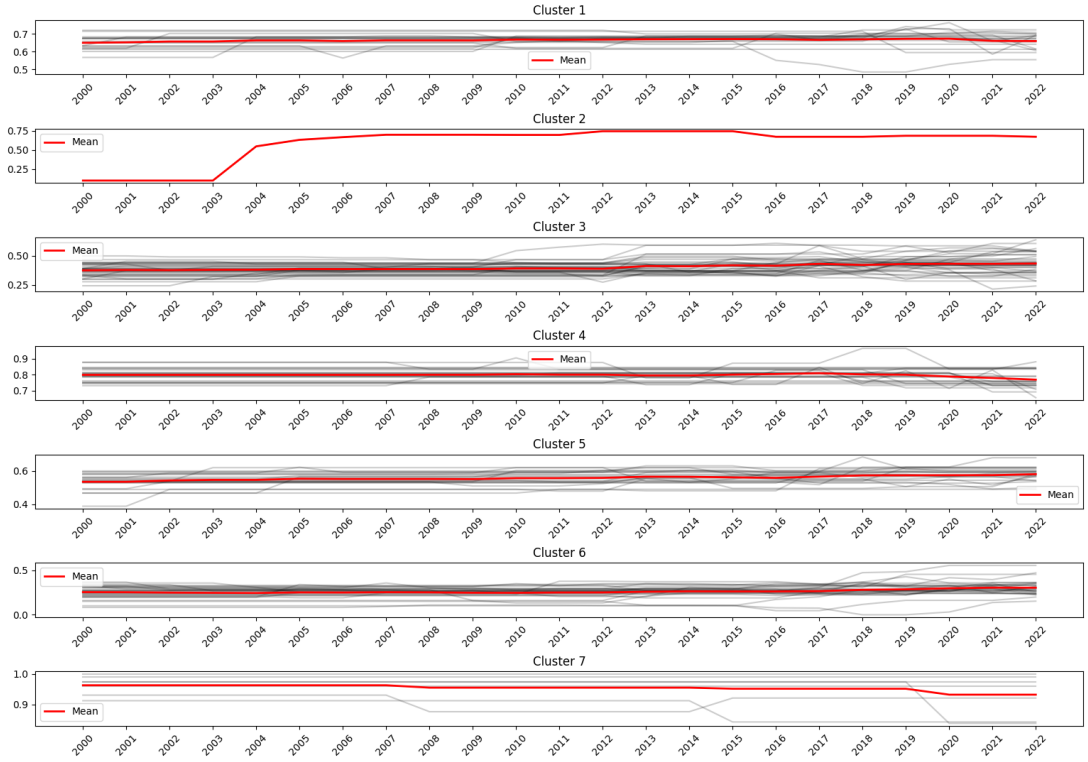
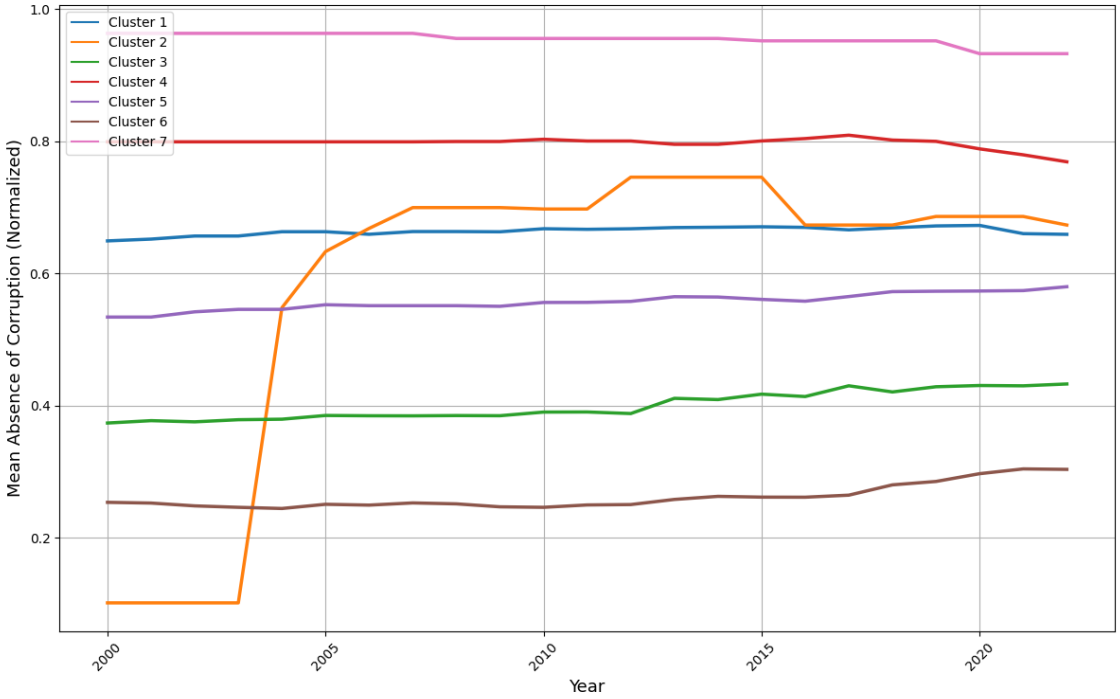


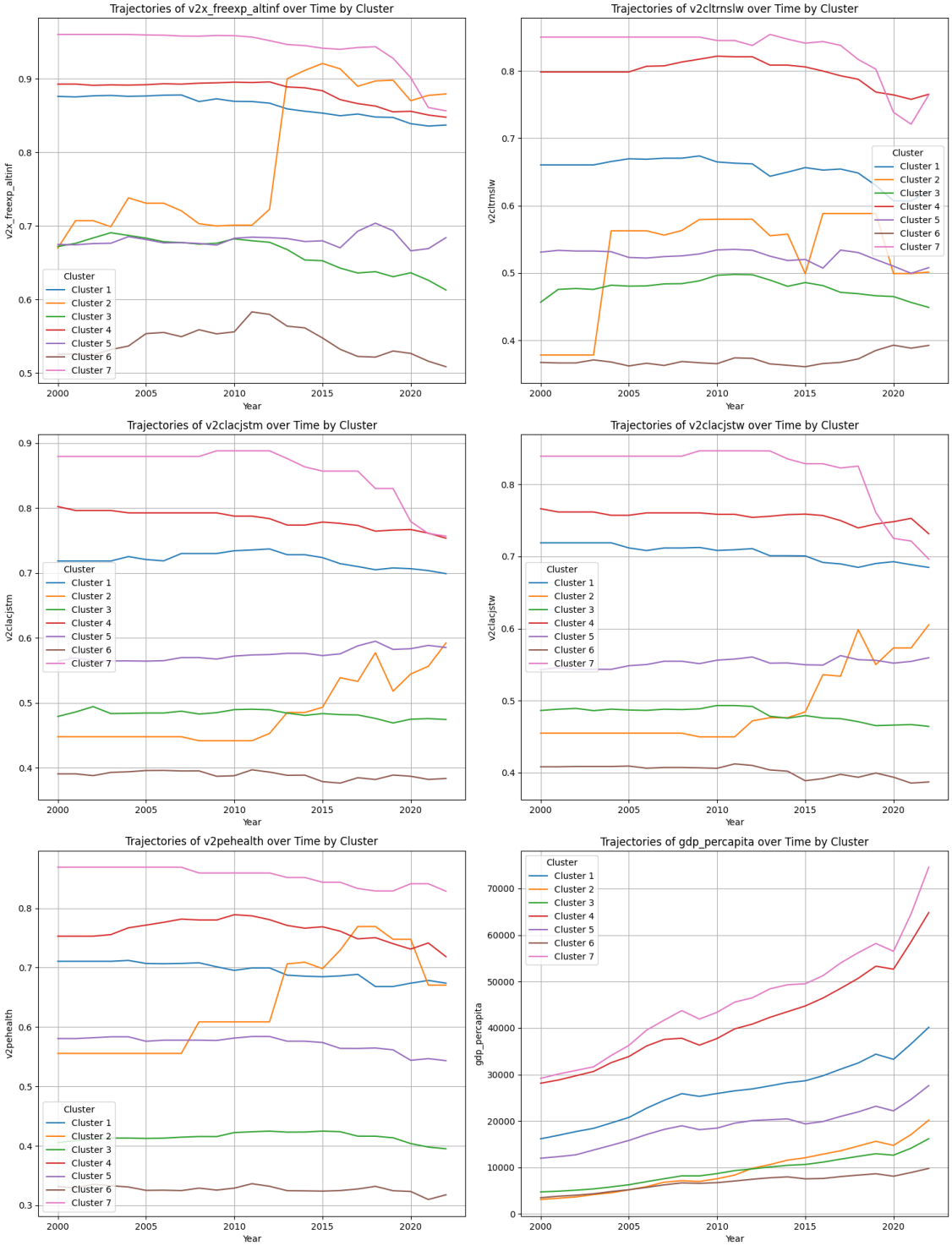
Fig.15 Mean Absence of Corruption in the Judicial Branch Over Time by Cluster



Analysing Sociopolitical and economic variables across clusters

This section delves into the interaction between selected socio-political and economic variables and the absence of corruption in the judicial branch. Based on the established framework of our analysis, Figure 16 illustrates the trajectories of how variables such as freedom of expression (v2x_freexp_altinf), transparency of law (v2cltrnslw), access to justice for men (v2clacjstm), access to justice for women (v2clacjstw), health equality (v2pehealth) and GDP per capita (gdp_percapita) evolve over time, showing different trends for different clusters.

Fig.16 Trajectories of Sociopolitical and Economic Variables Over Time by Cluster

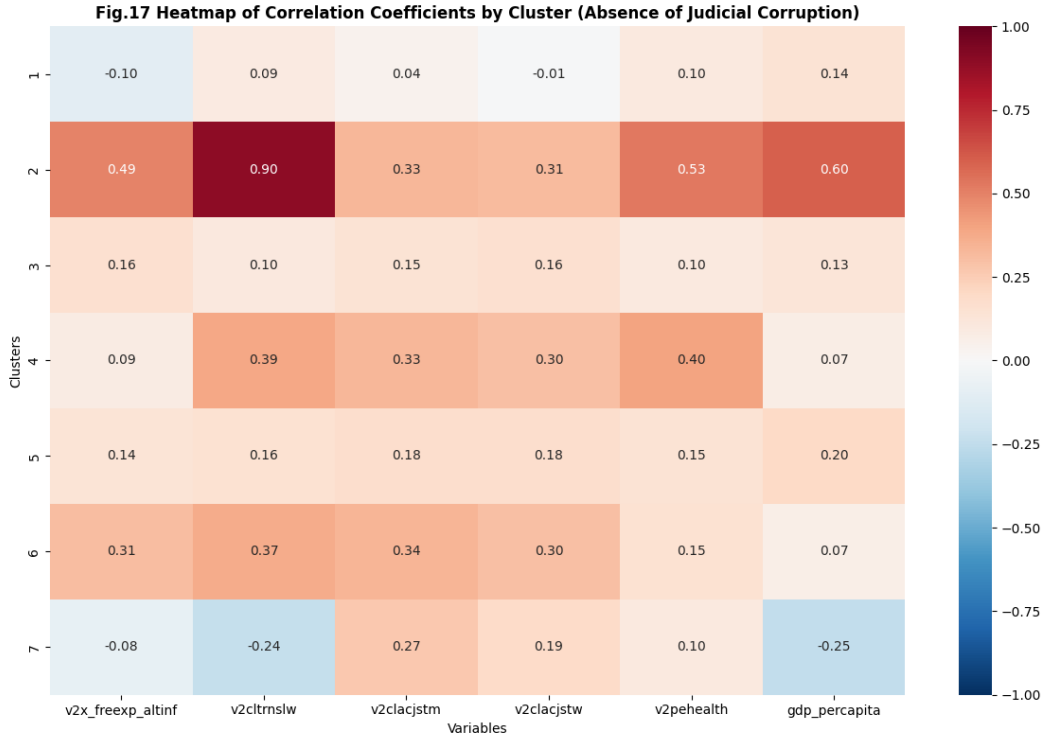


In in addition, Figure 17 presents a heatmap of the correlation coefficients. This quantifies the strength and direction of the linear relationships between absence of judicial corruption and the socio-political variables for each cluster:

- Cluster 1, 3, 4, 5, 6, 7 show weak correlations between the level of absence of corruption in the judicial branch and the selected socio-political and economic variables.

- Cluster 2 (consisting of Georgia only) stands out with a very strong positive correlation with respect to transparency of the law (0.90). In addition, health equity (0.53) and GDP per capita (0.60) show moderate positive correlations.

These results illustrate that most clusters demonstrate generally weaker correlations, indicating a less pronounced linear relationship between the absence of judicial corruption and the sociopolitical and economic variables selected.



Discussion of results

This study explores two main research questions by applying clustering models to political, executive, legislative and judicial corruption. The results provide detailed information on how countries are grouped according to corruption dynamics. In addition, the analysis sheds light on the associations between corruption and critical socio-political and economic variables for each cluster of countries.

Clustering of Countries Based on Corruption Categories

The clustering results reveal patterns in the way countries are grouped according to the four categories of corruption analysed. First, the political corruption clusters range from countries with minimal corruption, such as those in Western Europe and North America, to those with severe corruption problems, including many countries in Africa, Latin America, and parts of Asia. A notable finding is the unique position of Georgia, which stands alone in the cluster, indicating its distinctive trajectory in increasing levels of overall

political corruption absence, as it moved from having the lowest mean among all clusters during 2000 to 2003 to having the second highest mean from about 2005 onwards. This shift coincides closely with the conclusion of the Rose Revolution and the subsequent comprehensive anti-corruption reforms, which significantly reshaped Georgia's political landscape and governance practices (Engvall, 2012).

Regarding the absence of executive corruption, cluster 3, which includes countries such as Benin, Georgia, Romania, and Tunisia, is particularly interesting, showing a significant improvement over the years. From having the second lowest average among the clusters until before 2004, to having the second-best average in the last two years studied. Legislative corruption also presents an intriguing clustering, with Tunisia isolated in its own group. This is due to its improvement in increasing the absence of corruption in the legislature. Until before 2009 it had the worst average among the other 7 clusters and between 2010 and 2015 it became the third best average. It is worth mentioning that during 2020 it decreased its performance on this indicator, but it continues to maintain moderate levels compared to the other clusters. This behaviour may be due to specific legislative reforms, such as those following the Arab Spring (Muktad, 2023). Finally, the analysis of the absence of judicial corruption again highlights Georgia, grouped by itself, showing a pattern similar to that of the absence of general political corruption, as it also went from having the lowest mean of the 7 clusters on this indicator between 2000 and 2003 to having the third highest mean from around 2006 onwards.

Relationship Between Sociopolitical and economic Variables and Corruption

The relationship between key socio-political and economic variables and the absence of corruption (political, executive, legislative, and judicial) in the different clusters shows varied strengths and directions. In general, clusters composed of a single country showed the highest correlations with respect to socio-political and economic variables. This may be because clusters with several countries inherently include more variability, influenced by different national contexts, legal systems, economic conditions and political environments. A single-country cluster does not exhibit such variability, which could lead to stronger correlations between variables. It should also be noted that, as mentioned in the methodology, these 6 variables were selected with a random forest regressor model. This model is able to capture non-linear relationships and complex interactions between variables, which traditional correlation coefficients (such as Pearson's) cannot detect. Thus, if the relationship between the absence of corruption and the socio-political variables is non-linear, this correlation analysis may not fully capture the strength or nature of these relationships.

On the other hand, the clusters with the lowest levels of corruption also present positive and strong or moderately strong correlations with respect to the socio-political and economic variables. Specifically, with respect to absence of political corruption, cluster 7, composed mainly of Western European and Nordic countries, presents a moderately strong correlation for transparency in laws (0.57). Similarly, with respect to the variable absence of corruption in the legislature, cluster 4, which is made up of the Nordic countries, New Zealand, Singapore, and Uruguay, shows strong correlations for women's access to justice (0.74) and health equality (0.68) and moderately strong correlations for men's access to justice (0.54).

Lastly, with respect to the single country clusters, in Georgia, a strong positive correlation is found between transparency of the law and the absence of both political (0.92) and judicial corruption (0.90). Health equity and GDP per capita also show moderately strong positive correlations with the absence of political (0.51 for health equity, 0.58 for GDP per capita) and judicial corruption (0.53 for health equity, 0.60 for GDP per capita). These results suggest that there are associations between these factors and lower levels of corruption, although the direction or causality of these relationships cannot be determined by this analysis.

In Tunisia, freedom of expression shows an almost perfect correlation (0.99) with the variable absence of legislative corruption. Similarly, transparency of the law (0.93), access to justice for men and women (0.92 for both), and GDP per capita (0.75) also show positive and high correlations. These associations indicate that in this country when there is an increase in the absence of legislative corruption, there tends to be greater freedom of expression, greater legal transparency, and more access to justice, along with better economic performance.

Conclusion

This study analyses the clustering of political, executive, legislative and judicial corruption in 133 countries, using data from 2000 to 2022 to reveal significant patterns in corruption dynamics. In addition, correlations between key socio-political and economic variables and these corruption clusters have been calculated to better understand the relationship among them. The findings indicate a varied landscape of corruption, with some countries, like Georgia and Tunisia, showing notable improvements which not only serve as case studies for the impact of specific reforms and socio-political changes but also highlight the potential for countries to dramatically alter their corruption profiles through targeted actions.

In addition, the relationship between socio-political and economic variables and the absence of corruption for each cluster has been quantified, revealing strong correlations particularly in single-country clusters. For example, in Georgia, high correlations between transparency of the law and the absence of political and judicial corruption indicate that clear and enforceable legal frameworks are key to combating judicial corruption. Similarly, in Tunisia, the strong correlation between legislative corruption and variables such as freedom of expression, transparency of the law, access to justice and GDP per capita underlines the importance of economic development and openness of society in fostering integrity of governance.

However, some limitations must be acknowledged. First, the reliability of corruption data, which in this case is based on the perceptions of experts in each country and may be influenced by subjective biases. These biases may derive from personal experiences, cultural norms, or specific political views, which may affect the objectivity of the data. Furthermore, the study mainly identifies correlations rather than causality, which limits the ability to directly attribute changes in corruption levels to specific variables without further investigation.

Given these limitations and the findings of this study, we propose further research focusing on Georgia and Tunisia that could delve deeper into their peculiar tendency to decrease corruption. By examining these countries in isolation, insights can be gained into the

specific political, social, and economic reforms that have contributed most to their success and assess the sustainability of these changes. Furthermore, exploring how cultural factors influence the effectiveness of anti-corruption strategies in these contexts could lead to more effective and culturally tailored policies. Finally, investigating the role of technology and innovation in the fight against corruption, such as blockchain and AI-based transparency tools, given that Georgia is a pioneer in the use of blockchain for government processes, could also offer new insights into how to reduce political corruption in other countries.

Bibliography

Aghabozorgi, S., Seyed Shirkhorshidi, A., & Ying Wah, T. (2015). Time-series clustering – A decade review. *Information Systems*, *53*, 16–38. https://doi.org/10.1016/j.is.2015.04.007

Alqahtani, A., Ali, M., Xie, X., & Jones, M. W. (2021). Deep Time-Series Clustering: A Review. *Electronics*, *10*(23), Article 23. https://doi.org/10.3390/electronics10233001

Beyaert, A., García-Solanes, J., & Lopez-Gomez, L. (2022). Corruption, quality of institutions and growth. *Applied Economic Analysis*, *31*(91), 55–72. https://doi.org/10.1108/AEA-11-2021-0297

Bishop, C. M. (2006). Pattern recognition and machine learning. *Springer Google Schola*, *2*, 645–678.

Caner Türkmen, A. (2015). A review of nonnegative matrix factorization methods for clustering. *arXiv E-Prints*, arXiv-1507.

Chan, J. Y. K., Leung, A. P., & Xie, Y. (2021). Efficient High-Dimensional Kernel k-Means++ with Random Projection. *Applied Sciences*, *11*(15), Article 15. https://doi.org/10.3390/app11156963

Charron, N., & Lapuente, V. (2010). Does democracy produce quality of government? *European Journal of Political Research*, *49*(4), 443–470. https://doi.org/10.1111/j.1475-6765.2009.01906.x

Cuturi, M. (2011). Fast global alignment kernels. *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 929–936. https://openreview.net/pdf?id=r1NDZhbdWH

Cuturi, M., & Blondel, M. (2017). Soft-dtw: A differentiable loss function for time-series. *International Conference on Machine Learning*, 894–903. http://proceedings.mlr.press/v70/cuturi17a.html?source=post_page628e4799533c

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *2*, 224–227.

Deng, H., Runger, G., Tuv, E., & Vladimir, M. (2013). A time series forest for classification and feature extraction. *Information Sciences*, *239*, 142–153. https://doi.org/10.1016/j.ins.2013.02.030

Dhillon, I. S., Guan, Y., & Kulis, B. (2004). Kernel k-means: Spectral clustering and normalized cuts. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 551–556. https://doi.org/10.1145/1014052.1014118

Engvall, J. (2012). *Against the grain: How Georgia fought corruption and what it means*. Central Asia-Caucasus Institute & Silk Road Studies Program. https://www.diva-portal.org/smash/get/diva2:609285/FULLTEXT01.pdf

Ghahari, S. (2022). *Detecting and Measuring Corruption and Inefficiency in Infrastructure Projects Using Machine Learning and Data Analytics* [Thesis, Purdue University Graduate School]. https://doi.org/10.25394/PGS.15060534.v1

Ghahari, S., Queiroz, C., Labi, S., & McNeil, S. (2021). Cluster Forecasting of Corruption Using Nonlinear Autoregressive Models with Exogenous Variables (NARX)—An Artificial

Neural Network Analysis. *Sustainability*, *13*(20), Article 20. https://doi.org/10.3390/su132011366

Goel, R. K., & Nelson, M. A. (2011). Measures of corruption and determinants of US corruption. *Economics of Governance*, *12*(2), 155–176. https://doi.org/10.1007/s10101-010-0091-x

Han, J., Pei, J., & Tong, H. (2022). *Data mining: Concepts and techniques*. Morgan kaufmann. https://books.google.com/books?hl=es&lr=&id=NR1oEAAAQBAJ&oi=fnd&pg=PP1&dq=Han,+J.,+Pei,+J.,+and+Kamber,+M.+(2011).+%22Data+Mining:+Concepts+and+Techniques.%22+Morgan+Kaufmann.&ots=_N2GOFsfu_&sig=CLKMJo1TOtkZycfOuvEXL1NOJU4

Huang, X., Ye, Y., Xiong, L., Lau, R. Y. K., Jiang, N., & Wang, S. (2016). Time series k-means: A new k-means type smooth subspace clustering for time series data. *Information Sciences*, *367–368*, 1–13. https://doi.org/10.1016/j.ins.2016.05.040

Husted, B. W. (1999). Wealth, Culture, and Corruption. *Journal of International Business Studies*, *30*(2), 339–359. https://doi.org/10.1057/palgrave.jibs.8490073

International Monetary Fund (IMF). (2016). Corruption: Costs and Mitigating Strategies. *IMF*. https://www.imf.org/external/pubs/ft/sdn/2016/sdn1605.pdf

Islam, M., & Yousuf, M. (2018). *Development of a Corruption Detection Algorithm using K-means Clustering*. 1–4. https://doi.org/10.1109/ICAEEE.2018.8642985

Javed, A., Rizzo, D. M., Lee, B. S., & Gramling, R. (2023). Somtimes: Self organizing maps for time series clustering and its application to serious illness conversations. *Data Mining and Knowledge Discovery*. https://doi.org/10.1007/s10618-023-00979-9

Lakshminarayanan, S. (2020). Application of Self-Organizing Maps on Time Series Data for identifying interpretable Driving Manoeuvres. *European Transport Research Review*, *12*(1), 25. https://doi.org/10.1186/s12544-020-00421-x

Leitão, N. C. (2021). The Effects of Corruption, Renewable Energy, Trade and CO2 Emissions. *Economies*, *9*(2), Article 2. https://doi.org/10.3390/economies9020062

Maharaj, E. A., D'Urso, P., & Caiado, J. (2019). *Time Series Clustering and Classification*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429058264

Maria, N. S. B., Susilowati, I., Fathoni, S., & Mafruhah, I. (2021). The Effect of Education and Macroeconomic Variables on Corruption Index in G20 Member Countries. *Economies*, *9*(1), Article 1. https://doi.org/10.3390/economies9010023

Mauro, P. (1995). Corruption and growth. *The Quarterly Journal of Economics*, *110*(3), 681–712.

Muktad, A. (2023). Democracy, Corruption and Economic Growth Post-Arab Spring in Tunisia and Libya. *Review of Middle East Economics and Finance*, *19*(3), 153–186. https://doi.org/10.1515/rmeef-2023-0010

Ohana-Levi, N., Gao, F., Knipper, K., Kustas, W. P., Anderson, M. C., del Mar Alsina, M., Sanchez, L. A., & Karnieli, A. (2022). Time-series clustering of remote sensing retrievals for defining management zones in a vineyard. *Irrigation Science*, *40*(4), 801–815. https://doi.org/10.1007/s00271-021-00752-0

Paulus, M., & Kristoufek, L. (2015). Worldwide clustering of the corruption perception. *Physica A: Statistical Mechanics and Its Applications*, *428*, 351–358. https://doi.org/10.1016/j.physa.2015.01.065

Rothstein, B., & Teorell, J. (2008). What Is Quality of Government? A Theory of Impartial Government Institutions. *Governance*, *21*(2), 165–190. https://doi.org/10.1111/j.1468-0491.2008.00391.x

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Schölkopf, B. (2001). Statistical Learning and Kernel Methods. In G. Riccia, H.-J. Lenz, & R. Kruse (Eds.), *Data Fusion and Perception* (pp. 3–24). Springer Vienna. https://doi.org/10.1007/978-3-7091-2580-9_1

Serra, D. (2006). Empirical determinants of corruption: A sensitivity analysis. *Public Choice*, *126*(1), 225–256. https://doi.org/10.1007/s11127-006-0286-4

Soheily-Khah, S., Douzal-Chouakria, A., & Gaussier, E. (2016). Generalized k-means-based clustering for temporal data under weighted and kernel time warp. *Pattern Recognition Letters*, *75*, 63–69.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *IOP Conference Series: Materials Science and Engineering*, *336*, 012017. https://iopscience.iop.org/article/10.1088/1757-899X/336/1/012017/meta

Treisman, D. (2000). The causes of corruption: A cross-national study. *Journal of Public Economics*, *76*(3), 399–457.

UNDP (United Nations Development Programme). (2008). *Mainstreaming Anti-Corruption in Development*. https://www.undp.org/sites/g/files/zskgke326/files/publications/Mainstreaming_Anti-Corruption_in_Development.pdf

Vera, J. F., & Angulo, J. M. (2023). An MDS-based unifying approach to time series K-means clustering: Application in the dynamic time warping framework. *Stochastic Environmental Research and Risk Assessment*, *37*(12), 4555–4566. https://doi.org/10.1007/s00477-023-02470-9

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & Data, M. (2005). Practical machine learning tools and techniques. *Data Mining*, *2*(4), 403–413. https://sisis.rz.htw-berlin.de/inh2012/12401301.pdf

World Bank. (2020). *World Development Report 2020: Trading for Development in the Age of Global Value Chains* [Text/HTML]. https://www.worldbank.org/en/publication/wdr2020

Yin, R., Liu, Y., Wang, W., & Meng, D. (2022). Randomized Sketches for Clustering: Fast and Optimal Kernel k-Means. *Advances in Neural Information Processing Systems*, *35*, 6424–6436.

Appendix

Appendix A: Analysis of Variable Importance Scores for Political Corruption

Table 1A: Importance Scores of Variables in Assesing Political Corruption (2000)

Table 1A lists the importance scores assigned to various sociopolitical and economic variables[5] in the year 2000 using a random forest regression model. These scores help identify the most influential factors in political corruption as determined by the model. The variables are ranked from highest to lowest importance score.

Variable	Importance Score
v2cltrnslw	0.193269
v2clacjstm	0.136818
gdp_percapita	0.121868
v2x_freexp_altinf	0.092780
v2clacjstw	0.083582
v2clacjust	0.079436
v2csreprss	0.047846
v2pepwrses	0.041788
v2pehealth	0.037527
v2clacfree	0.030356
v2peedueq	0.028924
v2caviol_2	0.025730
v2pepwrgen	0.013215
v2cldiscw	0.012088
v2cacamps_2	0.011133
v2pepwrsoc	0.009283
v2cldiscm	0.009239
v2smgovfilcap	0.009149
v2smregcap	0.008749
gdp_growth	0.007221

[5]Transparent laws with predictable enforcement (v2cltrnslw), Health equality (v2pehealth), GDP per capita (gdp_percapita), Freedom of expression (v2x_freexp_altinf), Access to justice for women (v2clacjstw), Access to justice for men (v2clacjstm), Educational equality (v2peedueq), Social class equality in respect for civil liberties (v2clacjust), Power distributed by socioeconomic position (v2pepwrses), Freedom of discussion for women (v2cldiscw), Freedom of discussion for men (v2cldiscm), Political Violence (v2caviol_2), Power distributed by gender (v2pepwrgen), Power distributed by social group (v2pepwrsoc), GDP growth annual percentage (gdp_growth), Civil society organizations repression (v2csreprss), Freedom of academic and cultural expression (v2clacfree), Government capacity to regulate online content (v2smregcap), Political Polarization (v2cacamps_2), and Government internet filtering capacity (v2smgovfilcap).

Table 2A: Importance Scores of Variables in Assesing Political Corruption (2022)

Table 2A presents the importance scores for various sociopolitical and economic variables[6] in the year 2022, as determined by a random forest regression model. It highlights the variables that have the most substantial impact on the model's ability to assess political corruption. The indicators are ranked from highest to lowest importance score.

Variable	Importance Score
v2clacjstm	0.160451
v2pehealth	0.118136
gdp_percapita	0.112203
v2x_freexp_altinf	0.094282
v2clacjstw	0.082780
v2peedueq	0.070814
v2cltrnslw	0.069860
v2clacjust	0.045452
v2pepwrses	0.043556
v2cldiscw	0.034479
v2caviol_2	0.024388
v2pepwrgen	0.023137
v2cldiscm	0.022902
v2pepwrsoc	0.021376
gdp_growth	0.020382
v2csreprss	0.016590
v2clacfree	0.013057
v2smregcap	0.009894
v2cacamps_2	0.008615
v2smgovfilcap	0.007644

[6]Transparent laws with predictable enforcement (v2cltrnslw), Health equality (v2pehealth), GDP per capita (gdp_percapita), Freedom of expression (v2x_freexp_altinf), Access to justice for women (v2clacjstw), Access to justice for men (v2clacjstm), Educational equality (v2peedueq), Social class equality in respect for civil liberties (v2clacjust), Power distributed by socioeconomic position (v2pepwrses), Freedom of discussion for women (v2cldiscw), Freedom of discussion for men (v2cldiscm), Political Violence (v2caviol_2), Power distributed by gender (v2pepwrgen), Power distributed by social group (v2pepwrsoc), GDP growth annual percentage (gdp_growth), Civil society organizations repression (v2csreprss), Freedom of academic and cultural expression (v2clacfree), Government capacity to regulate online content (v2smregcap), Political Polarization (v2cacamps_2), and Government internet filtering capacity (v2smgovfilcap).

Appendix B: Performance Analysis of Clustering Models on Corruption Data across Executive, Legislative and Judicial branch

Table 1B: Clustering Performance Metrics for Executive Branch Corruption

Algorithm	Clusters	Silhouette	Davies-Bouldin
Kmeans euclidean	6	0.387677	1.019757
Kmeans dtw	6	0.354580	1.168293
Kmeans softdtw	6	0.395986	1.049555
KernelKmeans gak	6	0.317424	2.783482
Kmeans euclidean	7	0.379637	1.025827
Kmeans dtw	7	0.309925	1.155532
Kmeans softdtw	7	0.384886	1.042122
KernelKmeans gak	7	0.375478	0.963984
Kmeans euclidean	8	0.372650	1.046366
Kmeans dtw	8	0.310010	1.227996
Kmeans softdtw	8	0.358705	1.140538
KernelKmeans gak	8	0.368892	1.045154
Kmeans euclidean	9	0.373753	0.940649
Kmeans dtw	9	0.311077	1.394456
Kmeans softdtw	9	0.369853	1.028891
KernelKmeans gak	9	0.365957	1.064142

Table 2B: Clustering Performance Metrics for Legislative Branch Corruption

Algorithm	Clusters	Silhouette	Davies-Bouldin
Kmeans euclidean	6	0.366756	0.835434
Kmeans dtw	6	0.355962	0.861127
Kmeans softdtw	6	0.380215	0.835474
KernelKmeans gak	6	0.381049	0.834859
Kmeans euclidean	7	0.357949	0.872966
Kmeans dtw	7	0.331559	1.089425
Kmeans softdtw	7	0.333697	0.927217
KernelKmeans gak	7	0.412364	0.660586
Kmeans euclidean	8	0.348780	0.845363
Kmeans dtw	8	0.320909	0.892372
Kmeans softdtw	8	0.334301	0.925307
KernelKmeans gak	8	0.348500	0.802839
Kmeans euclidean	9	0.354484	0.757720
Kmeans dtw	9	0.300974	0.860121
Kmeans softdtw	9	0.302506	0.991433
KernelKmeans gak	9	0.275890	0.720780

Table 3B: Clustering Performance Metrics for Judicial Branch Corruption

Algorithm	Clusters	Silhouette	Davies-Bouldin
Kmeans euclidean	6	0.372260	0.831614
Kmeans dtw	6	0.340172	0.766334
Kmeans softdtw	6	0.427214	0.779549
KernelKmeans gak	6	0.413041	0.714883
Kmeans euclidean	7	0.374673	0.803799
Kmeans dtw	7	0.372457	0.724870
Kmeans softdtw	7	0.455830	0.721843
KernelKmeans gak	7	0.449791	0.598796
Kmeans euclidean	8	0.387611	0.848474
Kmeans dtw	8	0.362612	0.940231
Kmeans softdtw	8	0.370614	0.905065
KernelKmeans gak	8	0.368882	0.828662
Kmeans euclidean	9	0.364839	0.871198
Kmeans dtw	9	0.353356	0.887477
Kmeans softdtw	9	0.343769	0.957672
KernelKmeans gak	9	0.219088	0.803513