

**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: 7500

## Table of contents

<b>Abstract</b>	<b>3</b>
<b>Introduction</b>	<b>4</b>
<b>Related work</b>	<b>4</b>
<b>Data collection and variables selection</b>	<b>5</b>
<b>Methodology</b>	<b>6</b>
Overview . . . . .	6
Clustering Techniques . . . . .	7
Model Evaluation Metrics . . . . .	8
Optimal Number of Clusters . . . . .	9
Correlation between Sociopolitical and Economic variables and corruption clusters . . . . .	9
<b>Results</b>	<b>9</b>
a) <b>Clustering Political Corruption</b> . . . . .	9
b) <b>Clustering Executive, Legislative, and Judicial Corruption</b> . . . . .	15
<b>Clustering Model Selection</b> . . . . .	15
Executive Corruption . . . . .	16
Legislative Corruption . . . . .	20
Judicial Corruption . . . . .	25
<b>Discussion of results</b>	<b>30</b>
<b>Clustering of Countries Based on Corruption Categories</b> . . . . .	30
<b>Relationship Between Sociopolitical and economic Variables and Cor-         ruption</b> . . . . .	31
<b>Conclusion</b>	<b>32</b>
Acknowledgments . . . . .	33
References . . . . .	33
Signed statement of authorship . . . . .	33
Signed library submission form . . . . .	33

## 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 shown marked improvements in curbing corruption, others continue to struggle with entrenched corrupt practices. Particularly, Georgia and Tunisia display unique trajectories in their anti-corruption efforts. Furthermore, the study explores the correlation between the absence of corruption in these clusters and key socio-political and economic variables, selected through a random forest regressor model. Particularly in single-country clusters, there are significant correlations between corruption levels and factors such as transparency of laws, freedom of expression, access to justice, and economic performance. These insights highlight the potential of targeted anti-corruption strategies tailored to specific national contexts and underscore the complex relationship between governance, economic variables, and corruption.

## Introduction

In recent years, the global community has increasingly recognized corruption as a major obstacle to socioeconomic development, affecting various spheres of governance in all countries. This pervasive problem not only undermines democratic institutions, but also exacerbates poverty and inequality. The World Bank (2020) highlights that corruption inflates the costs and decreases the efficiency of public and private investments, which severely undermines 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 employs 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 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 corruption. The two main questions addressed are: (1) How are countries grouped according to the four categories of corruption analysed, 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?

By identifying clusters of countries with similar corruption profiles and understanding the specific dynamics of these clusters, the study provides detailed insights that can serve as a basis for targeted anti-corruption strategies. For example, countries within a group that exhibit high levels of judicial corruption could benefit from targeted judicial reforms, increased legal training 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, the findings of this study could be instrumental in developing policies that are not only responsive to the state of corruption, but also predictive and preventive, with the aim of disrupting the conditions that foster corruption.

## Related work

The literature on 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, asserting 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, strong uncertainty avoidance and low individualism. In addition, they argue that masculine societies are perceived as more corrupt than more feminine societies. Moreover, 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 focusing on corruption determinants, some recent works have explored the clustering of countries based on corruption levels. 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 non-linear autoregressive neural network models with exogenous inputs. Their findings show that NARX models can reliably predict future levels of corruption, providing practical information for policy makers and organisations to refine anti-corruption measures.

Furthermore, (Paulus & Kristoufek, 2015) use average linkage clustering 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.

## **Data collection and variables 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. The following indicators of political corruption were selected from the V-Dem institute:

a. Political Corruption Index: This is a composite index that 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 significant influencers of political corruption from 20 variables<sup>1</sup> deemed important by the literature. Key variables 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 for further analysis (Table 1 of the Appendix).

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. The reason for this was to allow comparison of the scores of these variables with those derived from the corruption clusters. 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<sup>2</sup>. This variable was left unnormalized to preserve its interpretability, as its absolute values are often directly indicative of the economic situation and are more meaningful when not transformed. Lastly, we specifically address the inverse of the corruption scores, focusing on the absence of corruption. This methodological choice improves interpretability, which is especially important when working with correlation analysis.

## 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

---

<sup>1</sup>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.

<sup>2</sup>GDP per capita adjusted for purchasing power parity (PPP) and measured in international dollars.

optimal number of clusters was determined using the elbow method. Finally, correlations between the selected socio-political and economic variables and 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. 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.

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

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.

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

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.

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

where  $D_{t,s}$  represents the squared Euclidean distance between elements  $x_t$  of series  $x$  and  $y_s$  of series  $y$ . The parameter  $\gamma$  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.

First, a kernel function is chosen to transform the time series data. Common choices include the Gaussian kernel (also known as the RBF kernel), polynomial kernels, and sigmoid kernels. In this study the Gaussian kernel is used for its ability to handle varied shapes of data and is defined as:

$$K(x, y) = \exp \left( -\frac{\|x - y\|^2}{2\sigma^2} \right)$$

where,  $x$  and  $y$  are two time series data points,  $\|x - y\|^2$  represents the squared Euclidean distance between them, and  $\sigma$  is a parameter that controls the width of the kernel. Each element  $K_{ij}$  of the matrix represents the similarity between the  $i$ -th and  $j$ -th time series in the dataset.

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  $\mu_k$  represents the mean of cluster  $k$  in this space. The mean is computed as:

$$\mu_k = \frac{1}{|S_k|} \sum_{j \in S_k} \phi(x_j)$$

The similarity  $s_{ij}$  between any two points  $x_i$  and  $x_j$  in the feature space is computed using the kernel function  $K$ , as:

$$s_{ij} = K(x_i, x_j)$$

During the clustering process, each data point  $x_i$  is assigned to the cluster  $k$  that minimizes the squared distance to the cluster mean  $\mu_k$ , which is computed as:

$$\begin{aligned} &\text{Assign } x_i \text{ to cluster } k \text{ that minimizes } \|\phi(x_i) - \mu_k\|^2 = \\ &\sum_{j \in S_k} (K(x_i, x_i) - 2K(x_i, x_j) + K(x_j, x_j)) \end{aligned}$$

## 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. 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:

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

$\sigma$  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.

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  $\mu_i$  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 groups. This analysis consists of calculating the Pearson correlation between the absence of corruption and 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.

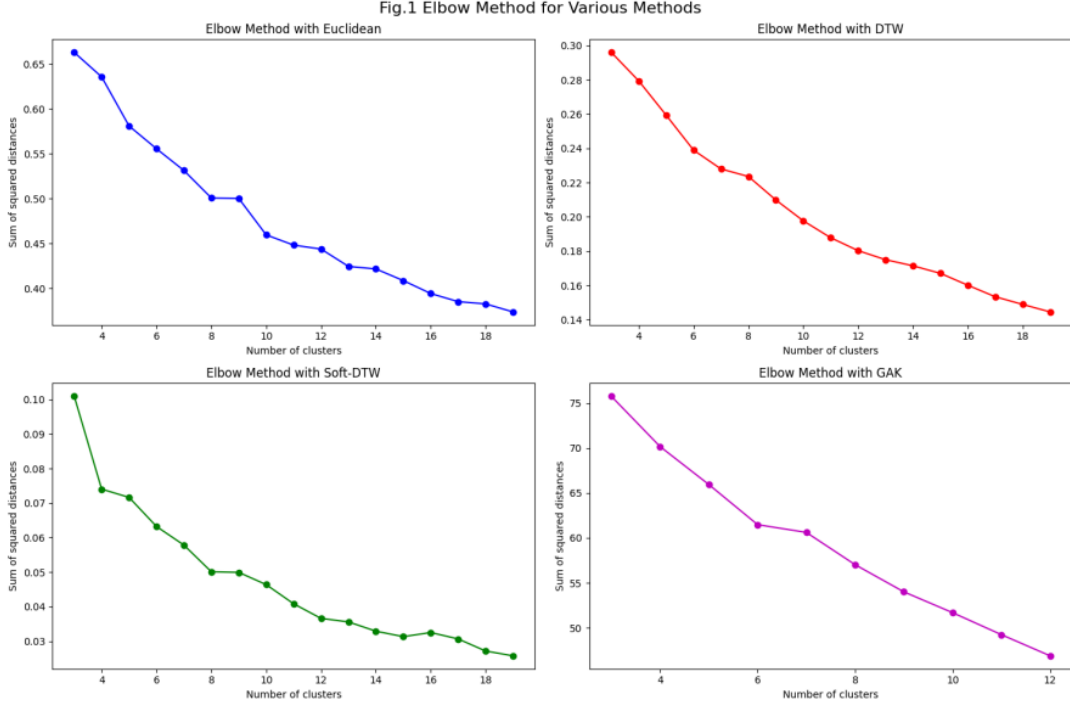
## Results

### a) Clustering Political Corruption

#### Selection of the optimal number of clusters

Figure 1 displays elbow plots for various clustering methods, each charting the sum of squared distances against the number of clusters using different metrics. The Euclidean Distance method shows the most pronounced elbow at 8-9 clusters. For Dynamic Time Warping (DTW), the elbow occurs at 7-8 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.



## Clustering Model Selection

This section evaluates the most effective clustering algorithm for grouping the level of political corruption in the 133 countries. 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.

## 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 1: Clustering Model Evaluation

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

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

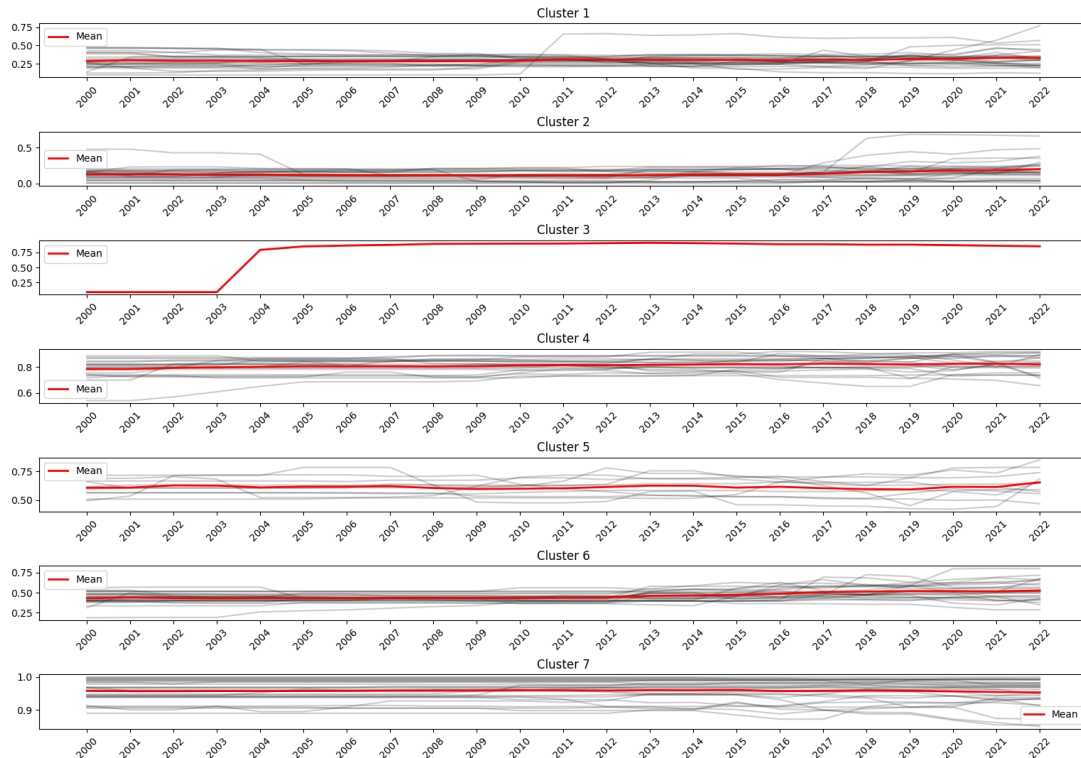
The clusters, as visualized in Fig. 2 (clusters over time) and Fig.3 (the mean of the clusters over time), demonstrate varying temporal patterns of the variable studied ‘absence of

political corruption’ across the 133 countries analysed. 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.

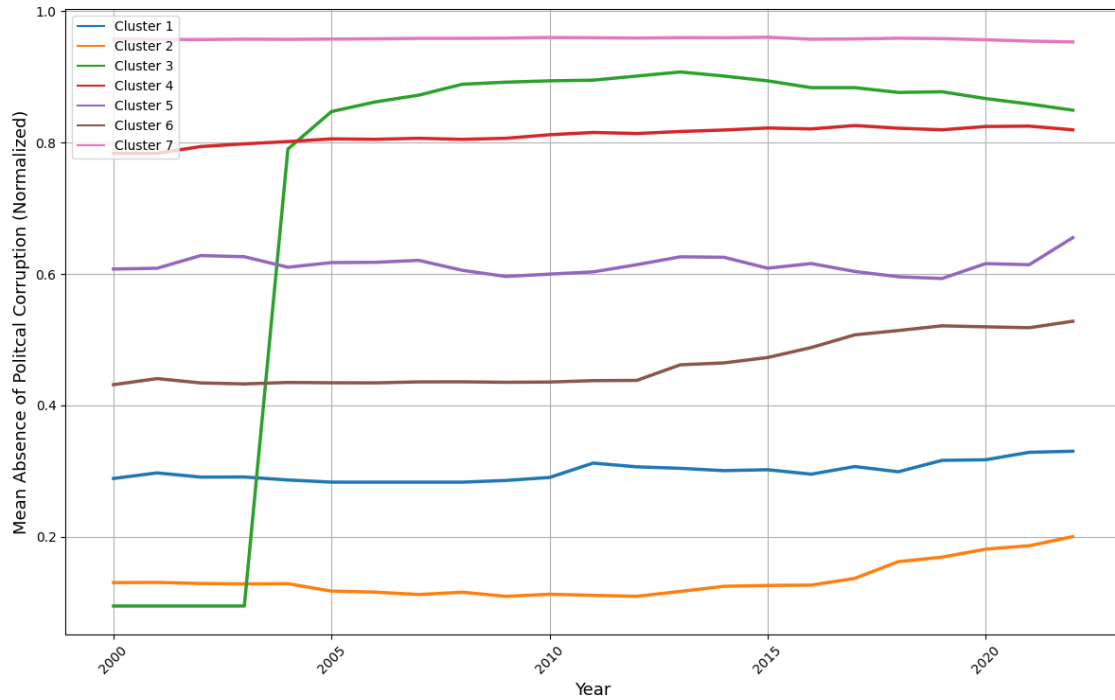
Furthermore, Clusters 5 and 6 register moderate levels of corruption over the 22-year period, compared to the other clusters. Cluster 6, in particular, has shown gradual improvements since 2012. Cluster 5 includes countries in Eastern Europe, Africa and Colombia as the only South American country. Group 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 European Ukraine. Similarly, Cluster 1 shows high levels of 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**



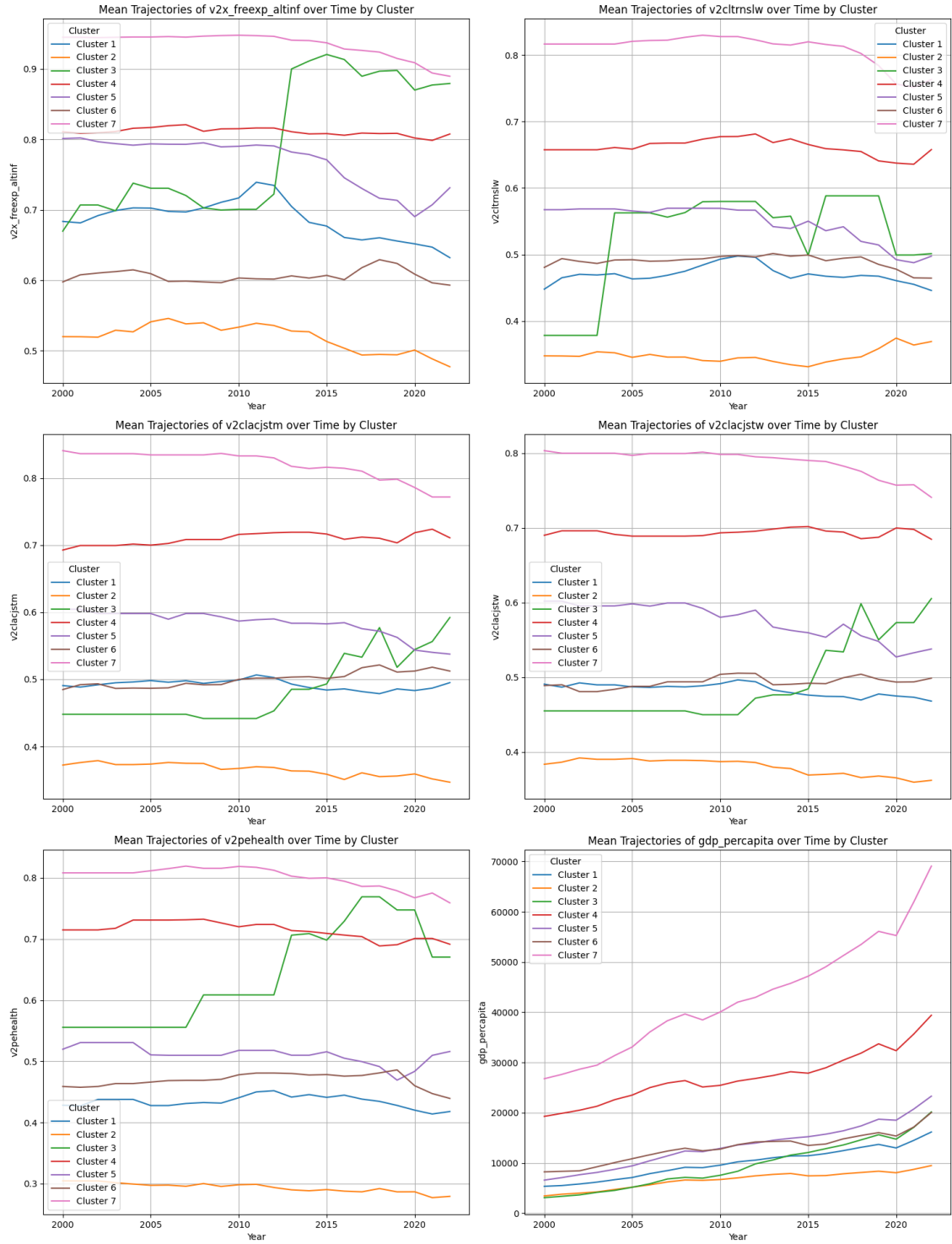
**Fig.3 Mean Absence of Political Corruption by Cluster**



### 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**

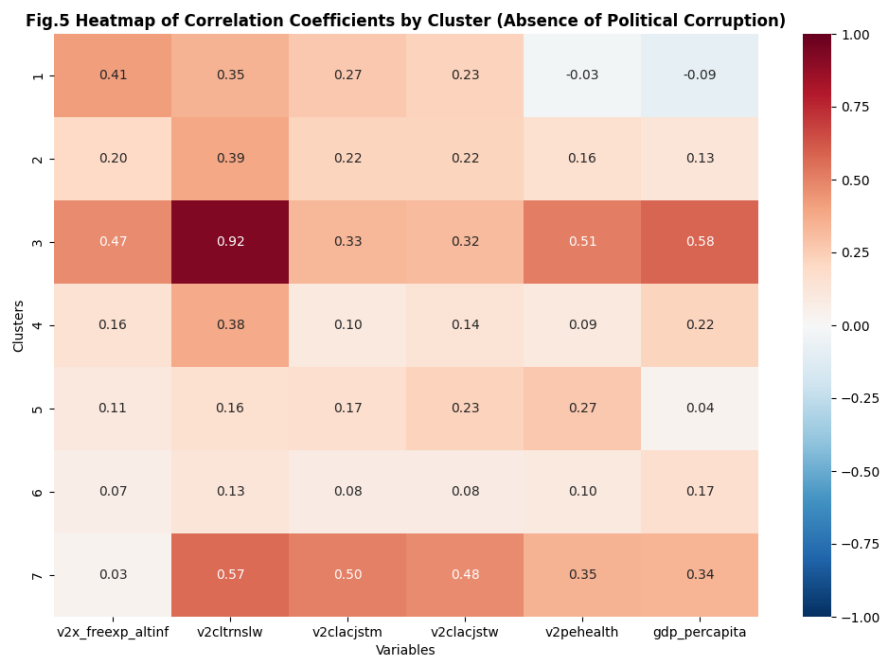


Furthermore, Figure 5 shows a heatmap with the correlation coefficients between the absence of political corruption, and each sociopolitical variable to quantify the strength and direction of relationships within each cluster:

- **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 correlations, especially in law transparency (0.57) and access to justice for men (0.50)

These results collectively suggest that 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 executive, legislative and judicial corruption. This decision is based on the performance of the model, which has outperformed the others in at least one of the key performance parameters - the Silhouette and Davies-Bouldin scores - and has demonstrated competitive or superior results in the remaining metric. The detailed results and comparative analysis of the models for the evaluation scores for each branch of corruption are documented in the Appendix in Tables 5, 6, and 7.

## 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 analysed. Firstly, **Cluster 1**, shows slight improvement from relatively moderate levels to low 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).

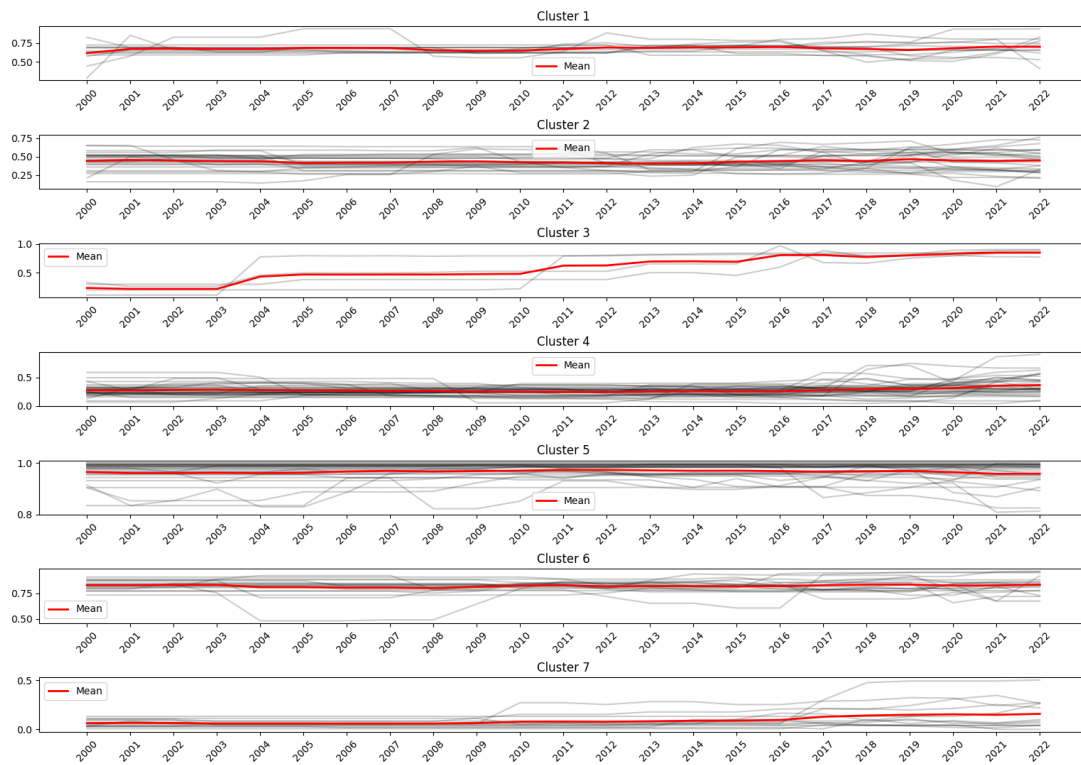
Furthermore, **Cluster 4** represents a broad mix of countries, mainly from Eastern Europe, Africa, Latin America, Southeast Asia, and Saudi Arabia, Russia, and Solomon Islands.



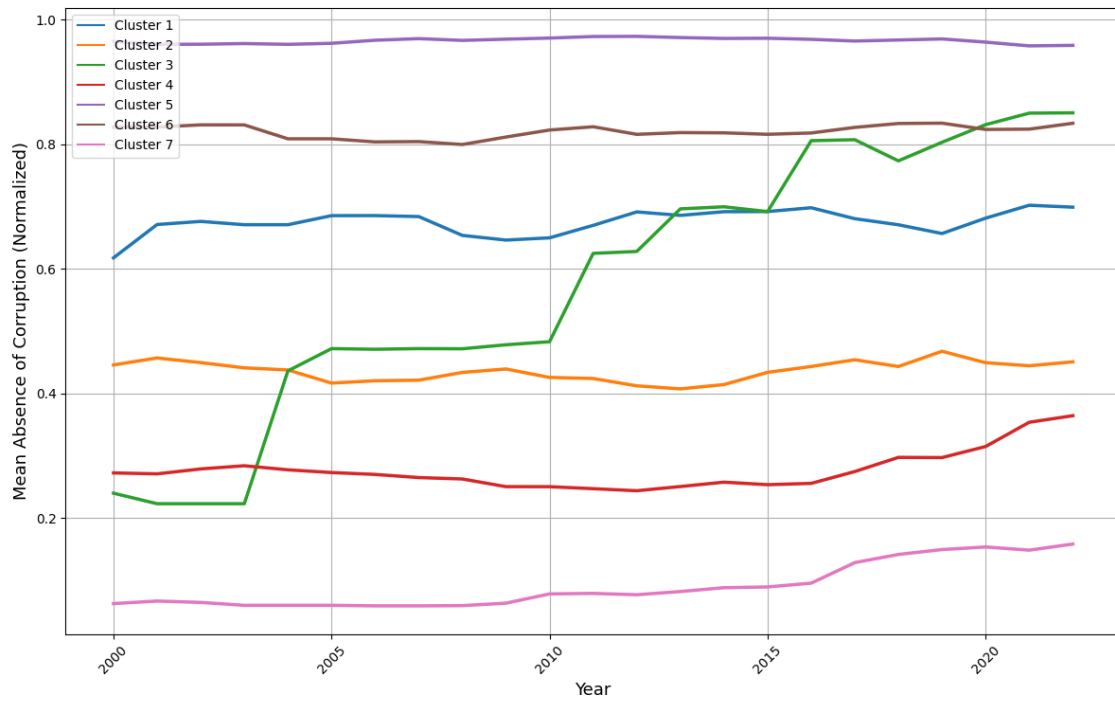
In the last years analysed, shows a slightly upward trend from low absence of corruption in the executive power to moderate levels, which suggests some progress, although challenges remain. **Cluster 5** portrays a different scenario, where predominantly Western and Northern European countries, along 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, and it's improving slightly over time. Lastly, **Cluster 7**, consisting 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**



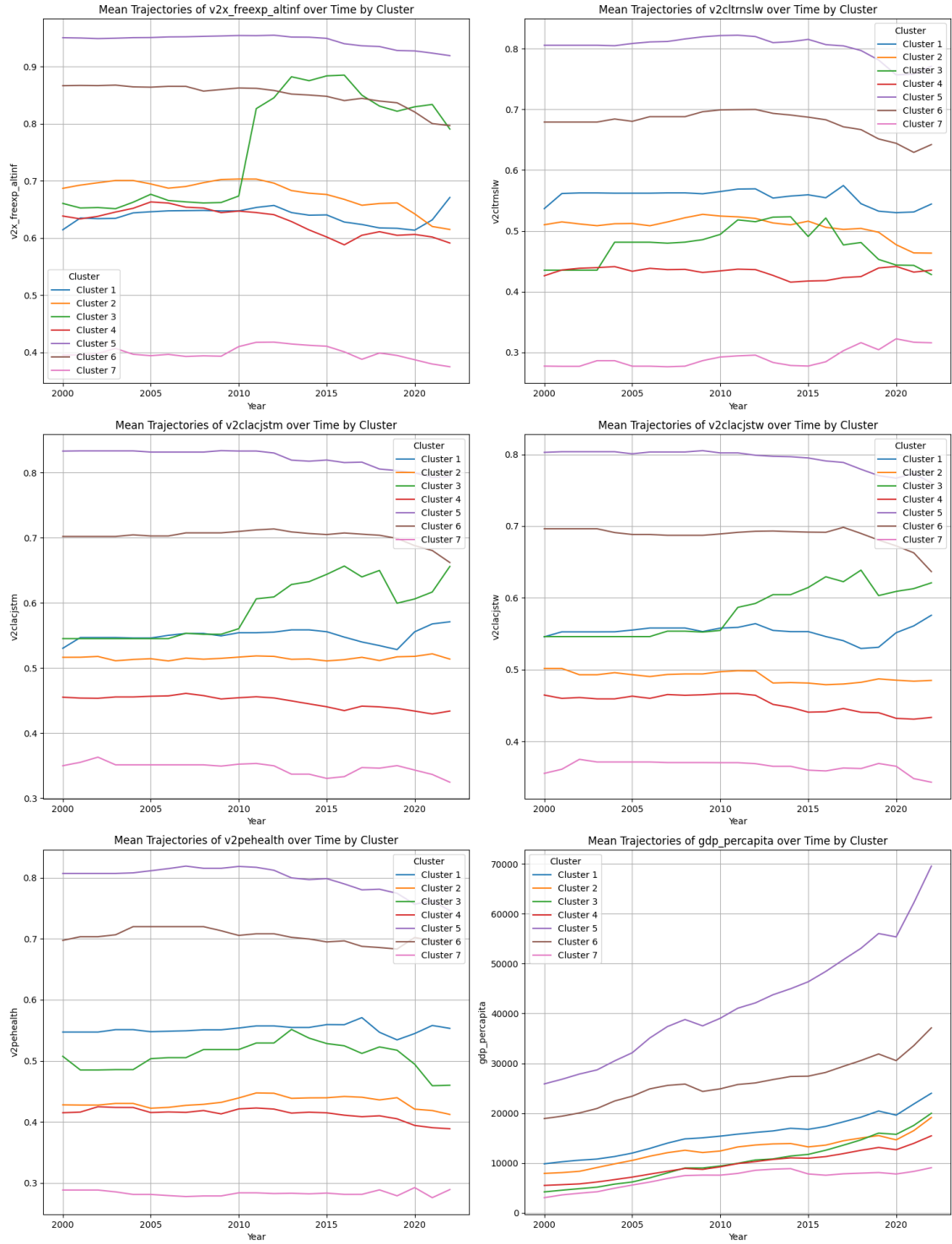
**Fig.7 Mean Absence of Corruption in the Executive Branch by Cluster**



### **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**

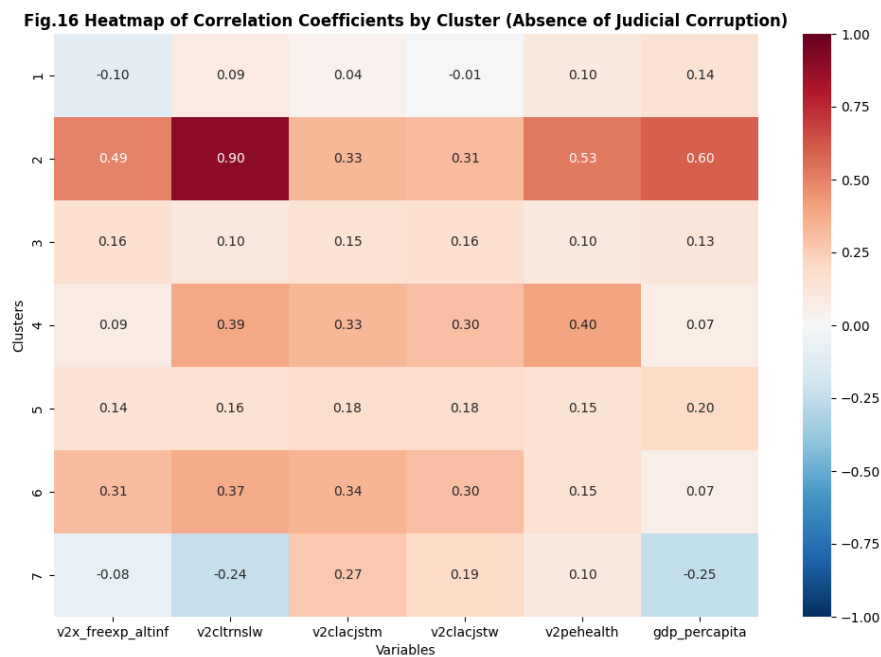


Furthermore, Figure 9 presents a heat map of the correlation coefficients. This quantifies the strength and direction of the relationships between the absence of executive corruption and the sociopolitical variables in each cluster:

- 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.

- Group 3, consisting only of Benin, Georgia, Romania and Tunisia, shows a moderately strong positive correlation for freedom of expression (0.50). However, it shows virtually no strong correlations for the other variables.
- Group 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

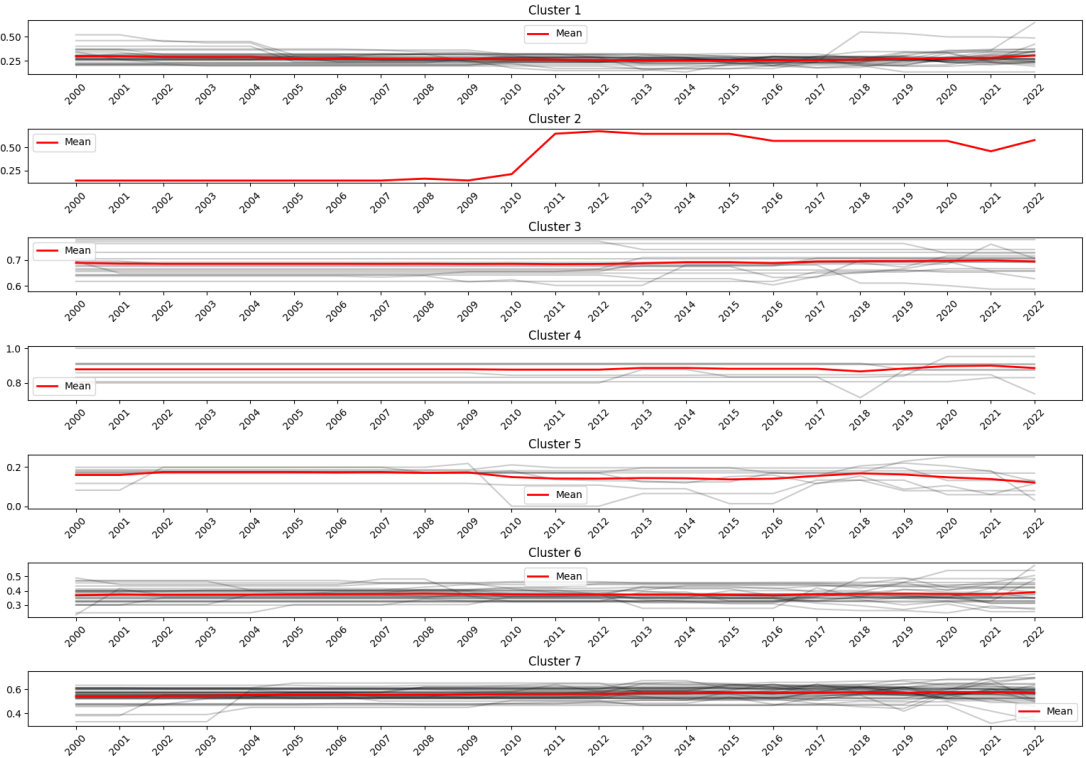
Moreover, the clusters, as visualized in Fig. 9 (clusters over time) and Fig.10 (the mean of the clusters over time), demonstrate varying temporal trends of the variable studied 'absence of corruption in the legislative branch' across the 133 counties. In the case of **Cluster 1**, it shows a relatively 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 to 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.

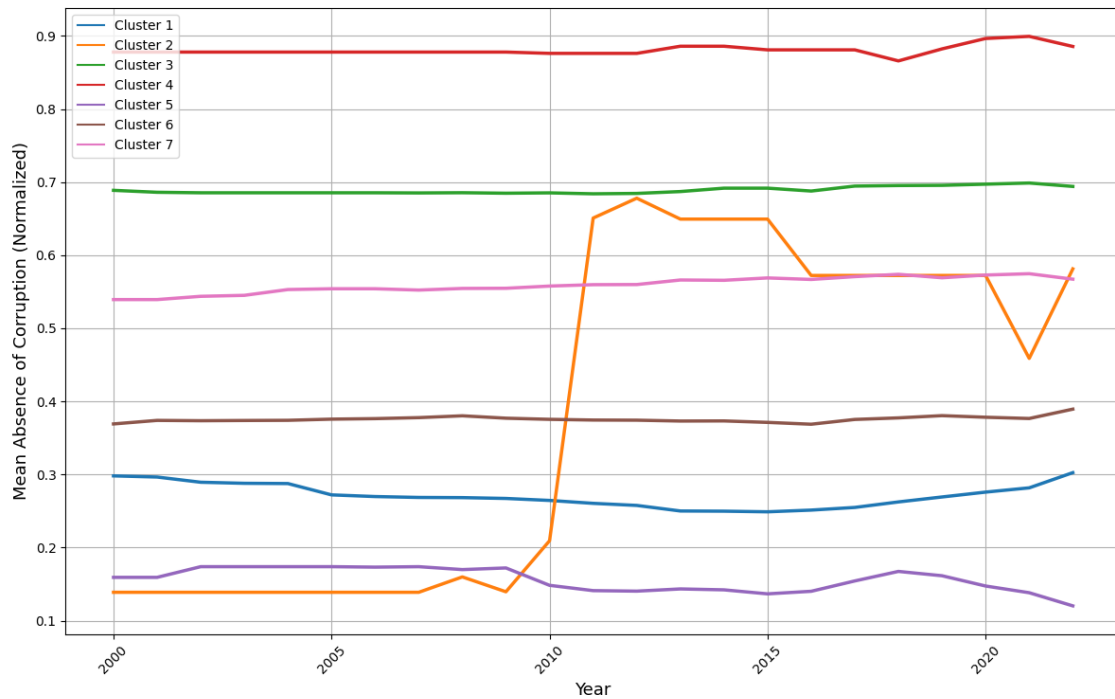
In contrast, **Cluster 5** has the highest levels of legislative corruption and has worsened slightly in recent years. It includes countries mainly in Africa and Central Asia, alongside the Dominican Republic, Guatemala, and Indonesia. Furthermore, **Cluster 6** shows a constant mean trend over time with a moderate to high level 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.9 Absence of Corruption in the Legislative Branch Over Time by Cluster**



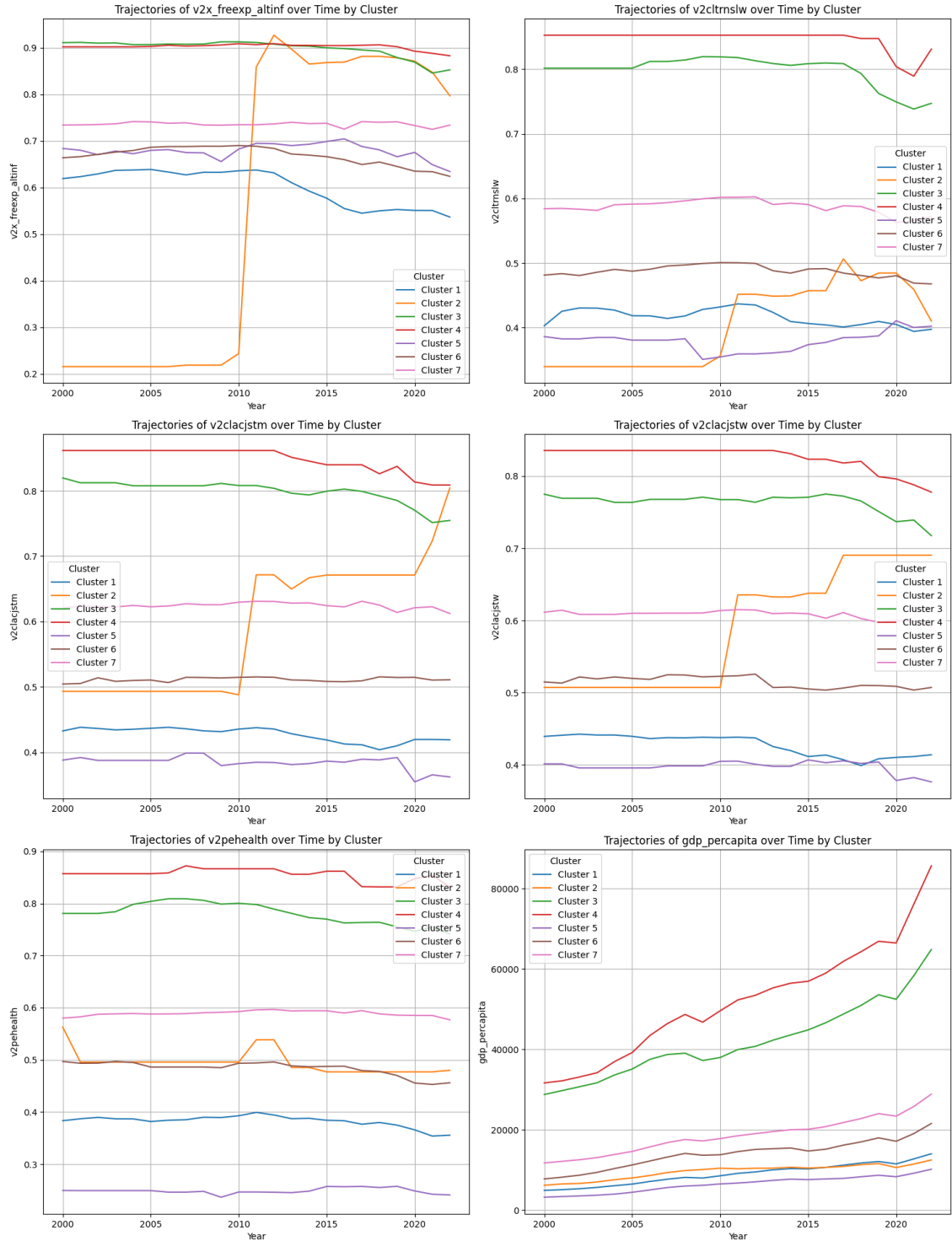
**Fig.10 Mean Absence of Corruption in the Legislative Power by Cluster**



### 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 11 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.11 Trajectories of Sociopolitical and Economic Variables Over Time by Cluster**



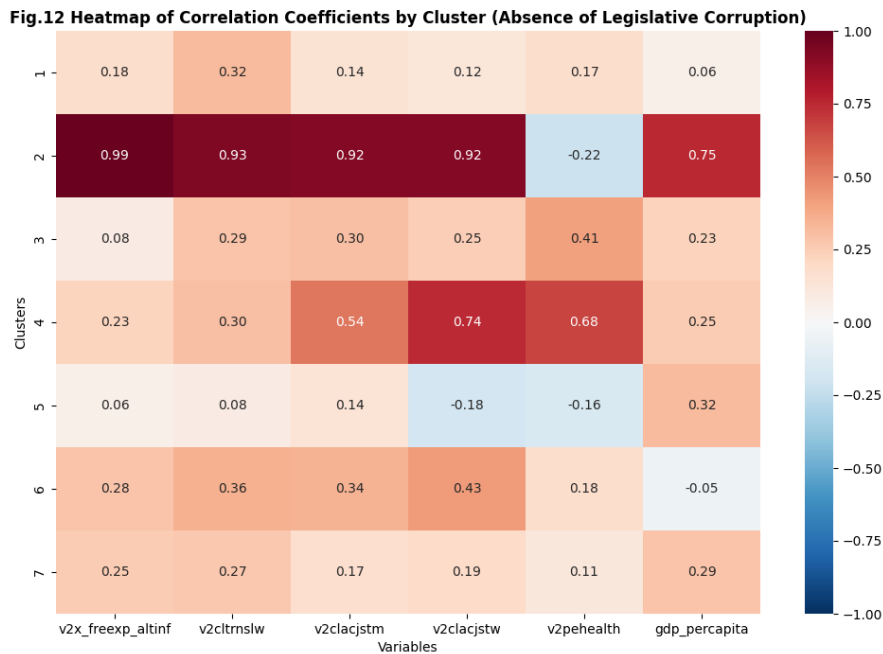
In addition, Figure 12 presents a heat map of the correlation coefficients. This quantifies the strength and direction of the relationships between the absence of legislative corruption and the sociopolitical variables in each cluster:

- Clusters 1, 3, 5, 6 and 7 show weak correlations between the level of absence of corruption in the legislature 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

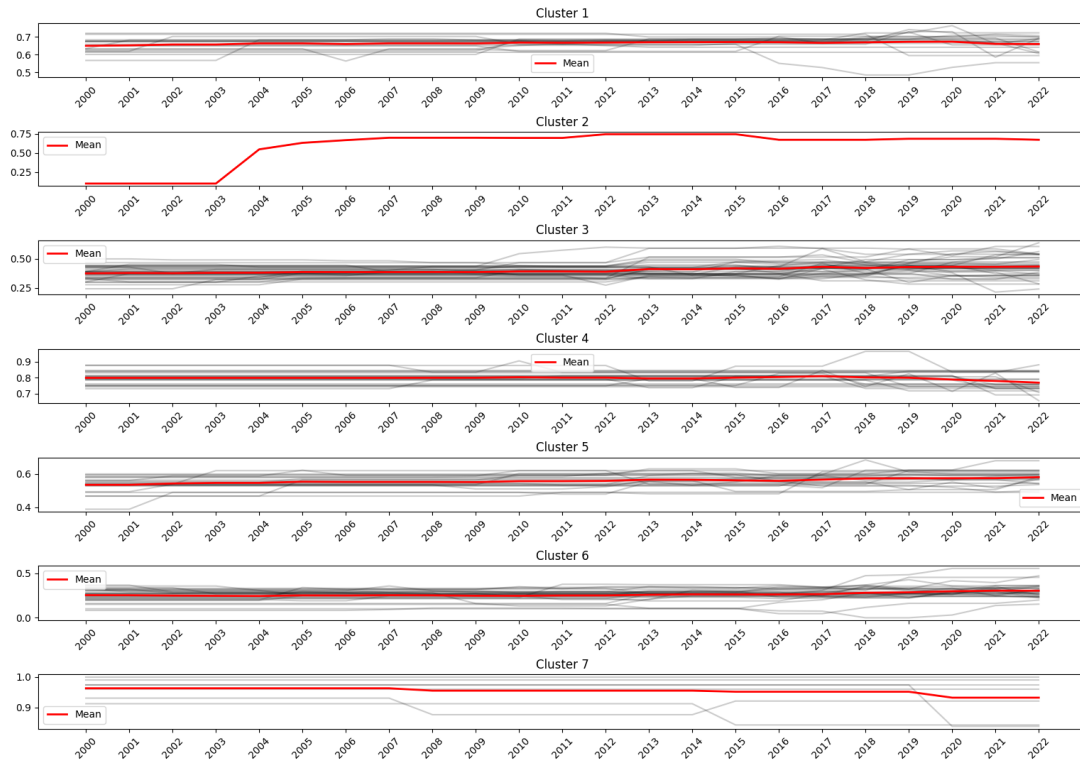
Moreover, the clusters, as visualized in Fig. 13 (clusters over time) and Fig.14 (the mean of the clusters over time), demonstrate varying temporal trends of 'absence of corruption in the judicial branch' across the 133 countries. Firstly, **Cluster 1** presents an amalgamation of countries that span 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** is 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). This cluster shows a high score of absence of corruption in the judiciary, compared to the other clusters. However, it has declined slightly in recent years.

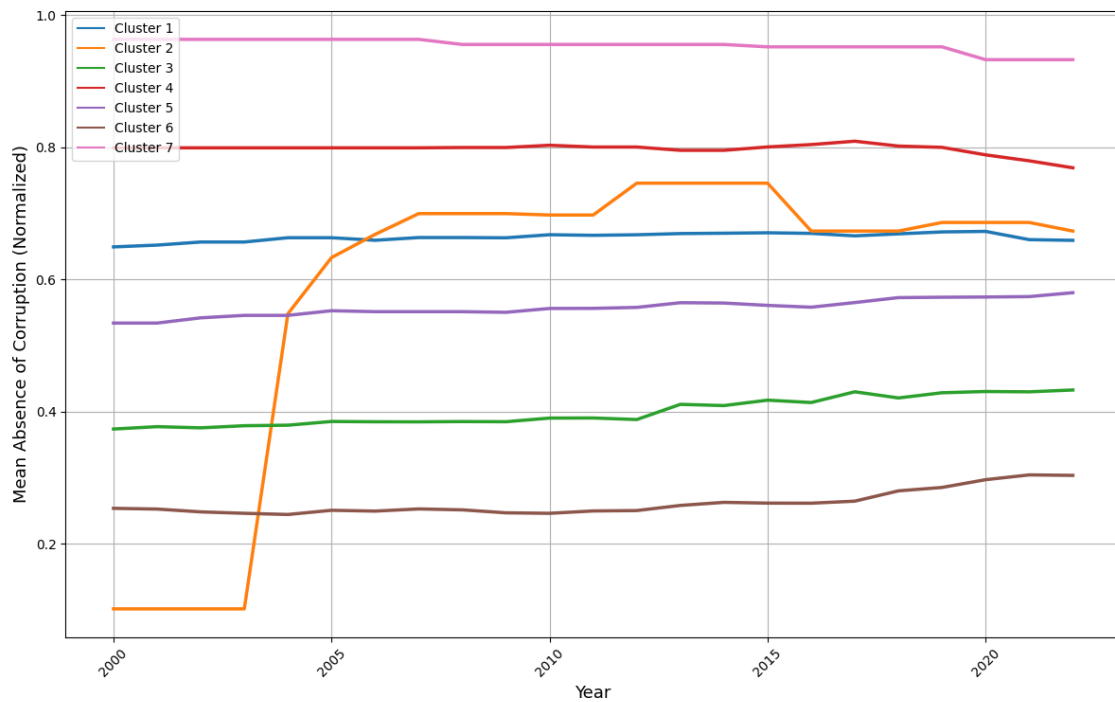
Cluster 5 incorporates countries in Southern and Eastern Europe, South and Southeast Asia, Latin America, Sub-Saharan Africa, the Middle East, and some islands in Oceania.

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.13 Absence of Corruption in the Judicial Branch Over Time by Cluster**



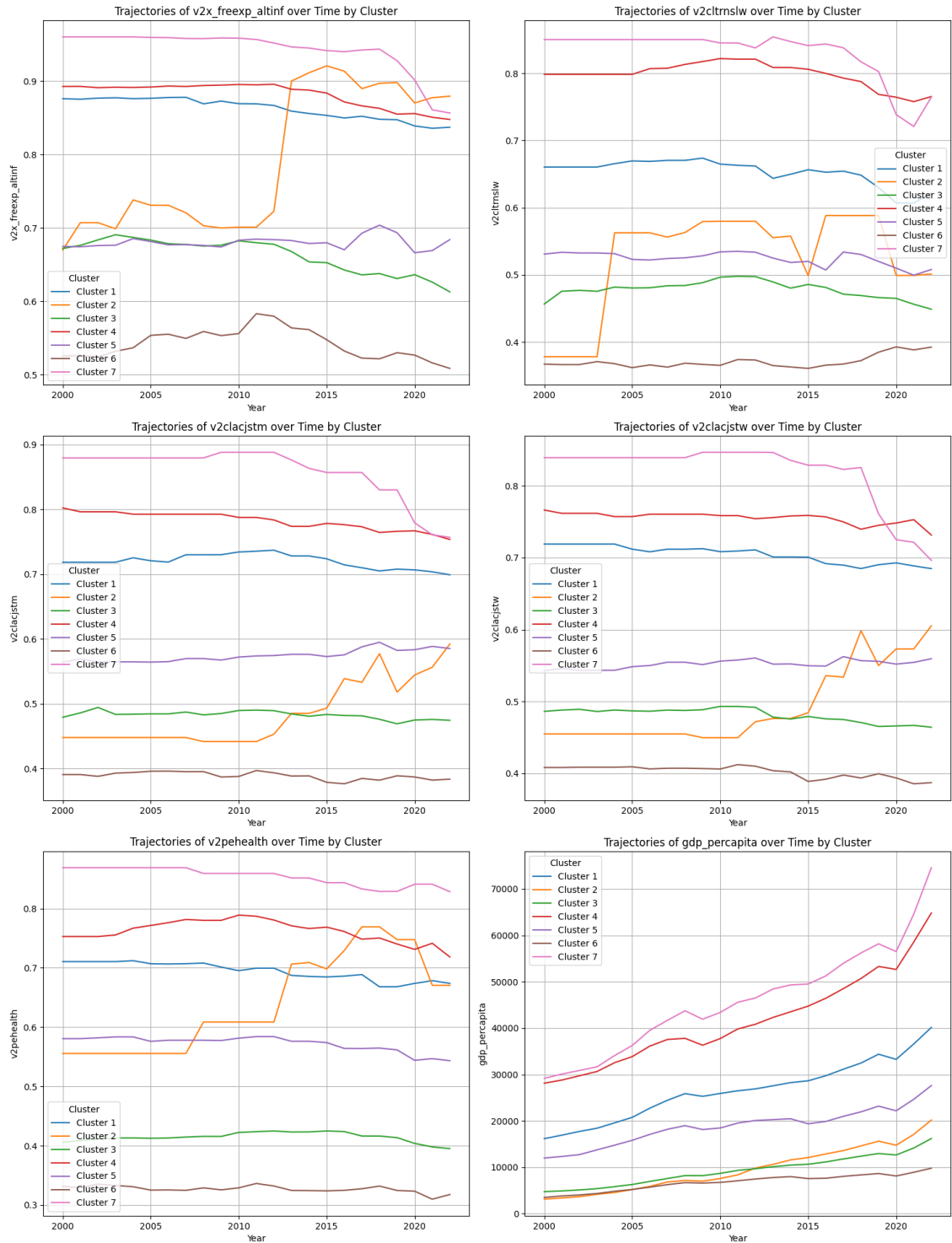
**Fig.14 Mean Absence of Corruption in the Judicial Branch Over Time by Cluster**

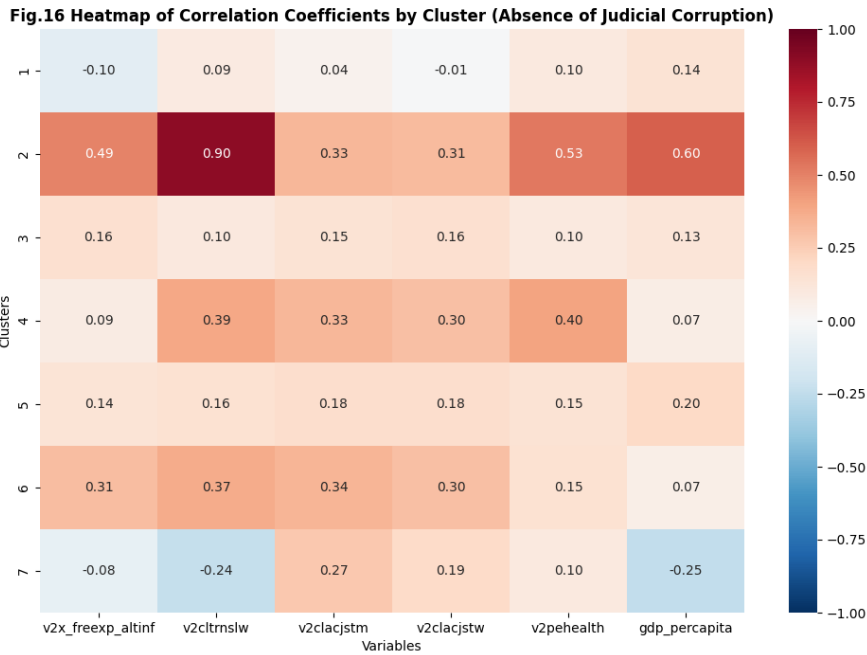


### **Analysing Sociopolitical and economic variables across clusters**

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 15 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.15 Trajectories of Sociopolitical and Economic Variables Over Time by Cluster**





## 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.

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. 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 type of 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 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 socio-political and economic variables. Specifically, with respect to the absence of political corruption, cluster 7, composed mainly of Western Europe and the 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 a decrease in legislative corruption, there tends to be greater freedom of expression, greater legal transparency, and more equal 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, an analysis of the relationships between key socio-political and economic variables and these corruption clusters has been provided in order to better understand the relationship between 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 within clusters 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, it would be worth exploring the issue further. In particular, research focusing on Georgia and Tunisia could delve deeper into their unique trends in reducing 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 ways to reduce corruption.



**Aknowledgments**

**References**

**Signed statement of authorship**

**Signed library submission form**