Thesis sections

Research Question

"How do the temporal dynamics of corruption vary across different clusters of countries, and do countries with similar temporal patterns of socioeconomic factors exhibit similar corruption behaviors over time? Additionally, what insights can be gained from identifying outliers within these temporal clusters?"

Data Sources and Scope

The dataset utilized in this study comprises a comprehensive collection of variables sourced primarily from reputable institutions such as the Varieties of Democracy (V-Dem), The World Bank, and Transparency International. Covering a period spanning from 2000 to 2022.

Methodology

1. Data Preparation:

Initially, the dataset is refined by choosing only those variables that possess available data for the targeted years and countries. Moreover, to enable meaningful comparisons, the data is normalized to ensure that all variables operate on a consistent scale.

2. Variable Selection using Random Forest

With corruption levels designated as the target variable and all other variables as features, a Random Forest model is trained on the dataset. Subsequently, the feature importance scores generated by the model are analyzed to identify the most influential variables contributing to predicting corruption levels. Based on these scores, the top-ranking variables are selected as inputs for subsequent clustering analysis, thus focusing the analysis on the most salient factors related to corruption dynamics.

3. Clustering Algorithm

Following the identification of the most significant variables, the dataset undergoes clustering based on countries' corruption dynamics and social, political and economic attributes. To accomplish this, the Dynamic Time Warping (DTW) clustering algorithm is employed. DTW is particularly well-suited for analyzing time series data, as it accounts for temporal shifts and distortions in the data, thereby providing a more accurate measure of similarity between time series trajectories. By clustering countries based on DTW distances, the methodology aims to discern distinct groups with similar corruption behaviors over time, facilitating a nuanced understanding of corruption dynamics.

Two possible approaches:

- Multivariate Time Series Clustering: performing clustering using a multivariate time series that includes the Political Corruption Index and the key socio-economic variables. This method will cluster countries based on the combined dynamics of corruption and socio-economic factors over time.
- Cluster Profiling: once clusters are determined based on corruption, profiling each cluster by averaging the socio-economic variables within it. This profiling can help identify what socio-economic conditions are common in clusters with improving or worsening corruption trends.

5. Outlier Detection

In addition to cluster analysis, efforts are made to identify outliers within clusters – countries exhibiting anomalous corruption behaviors relative to their cluster peers. By analyzing these outliers, the methodology seeks to gain insights into the factors driving their divergent corruption trajectories. Furthermore, investigate any outliers within clusters where the socio-economic variables deviate significantly from the rest of the cluster.

6. Validation and Sensitivity Analysis

To validate the reliability of the findings, rigorous validation and sensitivity analyses are conducted. Cross-validation techniques, such as k-fold validation are employed to assess the stability of Random Forest model. Furthermore, validating the TimeSeriesKMeans model in your corruption dynamics study involves several key strategies to ensure the reliability and applicability of your clustering results. Using the silhouette score provides insight into the cohesion and separation of clusters, indicating how well the time series are matched to their clusters and differentiated from others.

Preliminary results

Random Forest model

In the preliminary phase of our analysis, we applied a Random Forest regressor model to data from the years 2000 and 2022 to identify the variables most predictive of corruption levels. The initial findings have shed light on several key socio-economic factors and their importance scores, indicating their predictive power regarding corruption. For the year 2022, the variables with the highest importance scores were related to health equality (v2pehealth), access to justice for women (v2clacjstw) and for men (v2clacjstm), transparent laws with predictable enforcement (v2cltrnslw), and educational equality (v2peedueq). These results underscore the significant influence of societal well-being and institutional integrity on corruption levels.

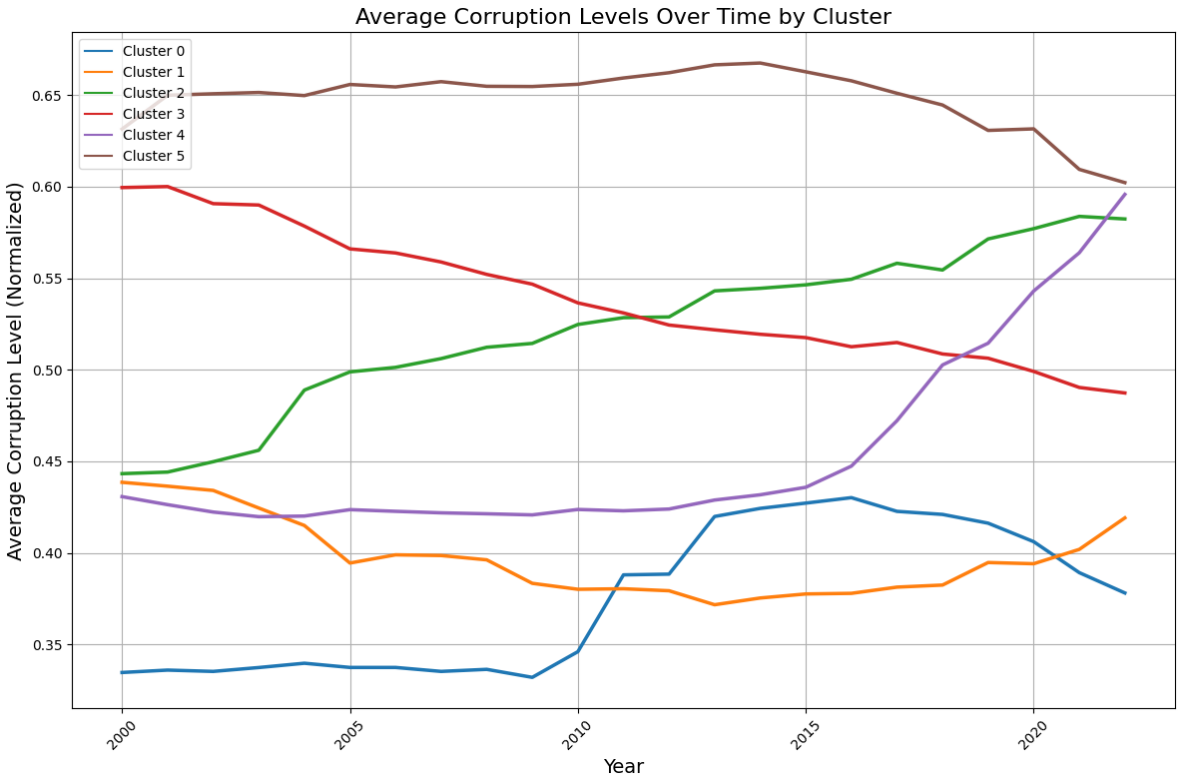
	importance
v2pehealth	0.153056
v2clacjstw	0.142310
v2clacjstm	0.119850
v2cltrnslw	0.108975
v2peedueq	0.093329
v2x_freexp_altinf	0.080084
v2clacjust	0.051069
v2pepwrses	0.048926
v2cldiscw	0.046032
v2pepwrgen	0.041331
v2cldiscm	0.032760
v2pepwrsoc	0.028329
v2clacfree	0.015821
gdp_growth	0.015032
v2smregcap	0.012411
v2smgovfilcap	0.010685

For the year 2000, the model identified "Transparent laws with predictable enforcement" (v2cltrnslw) as the most significant predictor of corruption, with an importance score of 0.202370, indicating a critical emphasis on the clarity and enforceability of laws at that time. Access to justice for men (v2clacjstm) and women (v2clacjstw) followed closely, with scores of 0.177814 and 0.144972 respectively. "Freedom of expression" (v2x_freexp_altinf) also had a high score of 0.097561 and "Health equality" (v2pehealth) with 0.069139.

	importance
v2cltrnslw	0.202370
v2clacjstm	0.177814
v2clacjstw	0.144972
v2x_freexp_altinf	0.097561
v2pehealth	0.069139
v2clacjust	0.055560
v2pepwrses	0.037870
v2cldiscw	0.037238
v2cldiscm	0.033212
v2clacfree	0.032396
v2peedueq	0.031704
v2pepwrgen	0.023601
v2smregcap	0.016226
v2smgovfilcap	0.015914
v2pepwrsoc	0.012913
gdp_growth	0.011510

Other variables such as freedom of discussion for men and women, freedom of academic and cultural expressio, socioeconomic power disparities, gender power disparities and GDP growth also contributed to the model but with lesser importance scores. More variables will be tested in the model in the next stages of the research.

Time series clustering



The time series clustering for the variable Political Corruption Index, where a higher score indicates lower corruption, reveals distinct trends across six identified clusters from 2000 to 2022.

Cluster 0 demonstrates a unique pattern, with a significant increase in integrity scores around 2015 but declining sharply afterward. This could reflect a crisis of corruption that was effectively addressed but later regressed, pointing to the complexity of sustaining anti-corruption progress.

Cluster 1 starts and remains in the lower middle range throughout, suggesting a persistent struggle with corruption and potentially ineffective or inconsistent anti-corruption efforts.

Cluster 2 starts with lower scores, indicative of higher corruption, but shows remarkable improvement across the years, likely reflecting successful anti-corruption measures or societal reforms. This cluster's upward trajectory indicates a positive shift in political transparency.

Cluster 3 shows a significant deterioration over time, starting with moderate scores but experiencing a steady decline, suggesting a worsening corruption scenario, possibly due to weakening institutions or political upheavals.

Cluster 4 stays stable over time until 2015 when it starts to increase, presenting an improve in the political corruption index.

Cluster 5 presents a consistently high level of political integrity, maintaining a leading position with the least corruption throughout the observed period. This cluster demonstrates the potential effects of stable and effective anti-corruption frameworks.

Countries by clustering:

Cluster 0: Burkina Faso, Burma/Myanmar, Cambodia, Canada, Chad, Cuba, Egypt, El Salvador, Gabon, Greece, Indonesia, Iraq, Israel, Ivory Coast, Jamaica, Mali, Mauritius, Niger, Republic of the Congo, Senegal, Tunisia, Uzbekistan, Zimbabwe

Cluster 1: Argentina, Bulgaria, Burundi, Cameroon, Central African Republic, Chile, Croatia, Democratic Republic of the Congo, Ethiopia, Haiti, Madagascar, Malta, Mauritania, Mexico, Nigeria, Oman, South Africa, Spain, Sri Lanka, Suriname, Thailand, Turkmenistan

Cluster 2: Bolivia, China, Colombia, Ecuador, Estonia, Finland, France, Georgia, Guatemala, Guinea-Bissau, Ireland, Italy, Kenya, Kyrgyzstan, Latvia, Lesotho, Libya, Lithuania, Luxembourg, Maldives, Montenegro, Morocco, Nepal, Paraguay, Romania, Russia, Rwanda, Seychelles, South Korea, Togo, United Arab Emirates

Cluster 3: Bangladesh, Costa Rica, Equatorial Guinea, Germany, Hong Kong, Hungary, India, Iran, Jordan, Mongolia, Nicaragua, North Macedonia, Norway, Pakistan, Papua New Guinea, Poland, Portugal, Solomon Islands, Sweden, Türkiye, United Kingdom

Cluster 4: Albania, Angola, Armenia, Australia, Azerbaijan, Barbados, Benin, Czechia, Dominican Republic, Guinea, Guyana, Honduras, Japan, Kazakhstan, Kuwait, Laos, Malaysia, Moldova, New Zealand, Panama, Saudi Arabia, Sierra Leone, Singapore, Slovakia, Sudan, Switzerland, Tajikistan, Tanzania, The Gambia, Trinidad and Tobago, Uganda, Ukraine, Vanuatu, Vietnam

Cluster 5: Algeria, Austria, Belarus, Belgium, Bosnia and Herzegovina, Botswana, Brazil, Cape Verde, Comoros, Cyprus, Denmark, Eswatini, Fiji, Ghana, Iceland, Malawi, Mozambique, Namibia, Netherlands, Palestine/West Bank, Peru, Philippines, Serbia, Slovenia, United States of America, Uruguay, Zambia