Regional Inequalities in Europe

A Clustering Approach to the East–West Divide

Nicole Gemelli, 880002
Federica Romano, 931429
Davide Tortorella, 926136
Francesca Verna, 880290

Academic Year 2024–2025

Contents

1 Introduction

Understanding socio-economic and political disparities across European regions has long been a focus of both academic research and policy-making. Despite decades of European integration, persistent differences remain between Eastern and Western Europe in terms of income, employment, health outcomes, educational attainment, and political freedoms. These differences are important because they influence economic development, social cohesion, and the capacity of regions to adapt to global challenges.

The aim of this project is to quantify and analyze the East-West divide in Europe through a multi-dimensional clustering approach, combining socio-economic indicators with measures of political freedom. Specifically, this study seeks to answer the following research question:

To what extent does there exist a gap between Eastern and Western Europe in terms of socio-economic conditions and political freedoms? An analysis based on clustering using income, employment, health, education, and political indicators.

To address this question, we combined data from the OECD Regional Well-Being dataset with Freedom House political freedom scores, providing a comprehensive view of both material well-being and civil liberties. Socio-economic indicators include household disposable income per capita, employment rates, life expectancy, educational attainment, access to broadband, and measures of social support, among others. Political freedom is measured via overall Freedom House scores and its subcomponents: political rights and civil liberties.

1.1 Background and Motivation

The economic and political landscape of Europe has been shaped by historical, social, and institutional factors that differ markedly between Eastern and Western regions. While Western European countries generally benefitted from early industrialization, stronger institutions, and higher integration into global markets, many Eastern European countries experienced delayed economic development due to historical transitions, including centrally planned economies and post-communist restructuring.

These historical divergences manifest in contemporary disparities across multiple dimensions:

- Income and employment: Western regions typically show higher household disposable income and lower unemployment rates.
- Health and well-being: Life expectancy, access to healthcare, and subjective well-being measures are often more favorable in the West.
- Education and Technology: Educational attainment, internet infrastructure, and digital access exhibit persistent East-West gaps.
- Political freedoms: Civil liberties and political rights, measured by Freedom House indices, also differ, reflecting varying levels of democratic consolidation.

Understanding these disparities is critical for several reasons. First, it informs regional policy and resource allocation within the European Union. Second, it provides insights into the social and economic mechanisms that sustain or reduce inequality. Finally, quantifying the relationship between political freedoms and socio-economic outcomes can help assess the potential benefits of institutional reforms or targeted interventions.

1.2 Methods Overview

Our analysis followed a multi-step approach designed to uncover patterns at both the regional and country level. First, we harmonized and preprocessed the data to ensure consistency across countries and regions, addressing missing values and standardizing measures. We then reduced the complexity of the dataset using Principal Component Analysis (PCA), thus capturing the main dimensions of variation in socio-economic characteristics. This facilitated clustering, allowing us to group regions with similar profiles and to examine how Eastern and Western regions distribute across these clusters.

Beyond clustering, we assessed internal cohesion within countries, measuring the degree of variation among regions and testing whether patterns differ according to historical EU membership or political freedom. Finally, we explored scenario-based counterfactuals using predictive modeling, simulating how changes in selected institutional and infrastructural indicators in Eastern regions could influence predicted economic outcomes. These analyses provided descriptive insights into potential mechanisms of inequality, without claiming causal inference.

Through this combination of descriptive statistics, clustering, and predictive modeling, the project offers a structured, multi-faced view of regional inequality in Europe. It highlights not only the current socio-economic gaps but also the institutional and structural factors that may shape future regional development.

2 Dataset and Exploratory Data Analysis

2.1 First Dataset

The first dataset used in this project comes from the OECD Regional Well-Being database, which provides indicators of quality of life across regions in OECD (Organization for Economic Co-operation and Development) countries. Data refer to 2021 or the last available year.

The dataset was retrieved directly from the OECD data portal. Data are available, through the button 'Download the data', at the following link:
https://www.oecd.org/en/data/tools/oecd-regional-well-being.html
The data used in this study come from the section *'See the value of the indicators for all OECD regions and for the most recent year and for the beginning year'*

The dataset consists of 225 rows and 17 columns. Each row represents a region within one of the European OECD member countries included in the study. The 26 countries covered are:
Austria, Belgium, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Netherlands, Norway, Poland, Portugal, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, United Kingdom.

The following variables are included:

- country: the country to which the region belongs.
- region: the specific region within the country.
- code: the official code identifying the region.
- disp_income_pc: average annual disposable income per person.

- employment_rate: percentage of the working-age population currently employed.
- life_expectancy: average number of years a person is expected to live at birth.
- secondary_edu_pct: percentage of the adult population having completed at least secondary education.
- homicide_rate: number of intentional homicides per 100,000 inhabitants.
- mortality_rate: mortality rate per 1,000 people.
- air_pollution: average annual exposure to particulate matter (PM 2.5), measured in $\mu g/m^3$.
- voter_turnout: percentage of registered voters participating in the most recent national elections.
- broadband_access: percentage of households with broadband internet access.
- internet_speed: deviation from OECD average, in percentage, in Mbps.
- number_rooms: average number of rooms available per person within households.
- social_support: percentage of people reporting they have someone to rely on in times of need.
- life_satisfaction: average self-reported life satisfaction on a scale from 0 to 10.

Only minimal preprocessing was performed. Specifically, all variables except for country, region, and code were converted into numeric format to ensure consistency across regions and allow further statistical analysis.

Lastly, to facilitate our analysis, we introduced a new categorical variable, 'macro_area', classifying each country as either "East" or "West" Europe.
This division is not arbitrary, but rather reflects the historical and political distinction inherited from the Cold War, with the Eastern bloc countries (Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Slovak Republic, Slovenia) grouped under "East", and the remaining OECD European countries under "West" (Austria, Belgium, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom).

2.2 Second Dataset

The second dataset comes from Freedom House, an international non-governmental organization based in Washington, D.C. that publishes an annual report that evaluates the state of civil liberties and political rights across countries worldwide.

The full dataset is publicly available at the following link:
https://freedomhouse.org/report/freedom-world
under the voice 'All Data, FIW 2013-2024 (Excel Download)'. For the purposes of this work, we restricted the dataset to the year 2021 to ensure temporal consistency with the OECD Well-Being data.

Although the original Freedom House dataset provides a wide range of indicators, we retained only four key variables:

- country: the country name, slightly modified to match the OECD dataset (e.g., "Czechia" replaced with "Czech Republic"). This variable was essential for merging the two datasets.

- freedom_score: an overall score ranging from 0 to 100, where higher values indicate greater levels of freedom.
- political_rights: originally measured on a scale from 1 (most free) to 7 (least free), this variable evaluates electoral processes, political pluralism, and participation. For consistency with other indicators and to preserve an increasing order, the scale has been inverted: in our analysis, 1 corresponds to the least free and 7 to the most free.
- civil_liberties: originally measured on a scale from 1 (most free) to 7 (least free), this indicator assesses freedom of expression and belief, associational rights, rule of law, and personal autonomy. Here too, the scale has been inverted.

After merging the Freedom House data with the OECD Regional Well-Being dataset, the final integrated dataset consists of 225 rows and 20 columns. Each row represents a European region, enriched with both well-being indicators and measures of political and civil freedom.

2.3 Handling Missing Values

Overall, the dataset is well-structured and contains very few missing values. Missing observations are limited to only seven variables, with relatively low percentages across the entire dataset:

- life_satisfaction: 5 missing (2.22%)
- social_support: 5 missing (2.22%)
- number_rooms: 3 missing (1.33%)
- secondary_edu_pct: 3 missing (1.33%)
- broadband_access: 2 missing (0.89%)
- life_expectancy: 2 missing (0.89%)
- employment_rate: 1 missing (0.44%)

Given the limited extent of missing data, we opted for an imputation strategy based on the country median. Specifically, for each numeric variable, missing values were replaced with the median of the corresponding country.

Being aware of the fact that within-country regional disparities may be smoothed out using a country-level value, we still decided to opt for this imputation to ensure internal consistency within each country.

In summary, the choice of country median imputation represents a balanced compromise between robustness, interpretability, and methodological simplicity, given the very low percentage of missing values.

2.4 West and East: Definition

The aim of this section is to investigate whether systematic differences exist between Eastern and Western Europe with respect to socio-economic well-being, democratic participation, and political freedom.

To this end, an overview of the distributions of all the variables under consideration is here presented:

Variable	West Min	West Q1	West Median	West Q3	West Max
air_pollution	4.10	8.30	9.60	11.20	19.70
broadband_access	75.00	87.00	91.00	96.00	100.00
civil_liberties	6.00	7.00	7.00	7.00	7.00
disp_income_pc	7452.00	20031.00	21810.00	24015.00	66131.00
employment_rate	32.20	63.90	70.40	75.70	84.30
freedom_score	87.00	90.00	93.00	96.00	100.00
homicide_rate	0.00	0.50	0.70	1.00	9.60
internet_speed	-87.15	-30.68	-11.00	18.04	68.90
life_expectancy	73.30	81.15	82.10	82.60	84.00
life_satisfaction	5.20	6.50	7.00	7.40	7.90
mortality_rate	5.33	6.75	7.29	7.93	13.80
number_rooms	0.98	1.60	1.89	2.00	2.46
political_rights	7.00	7.00	7.00	7.00	7.00
secondary_edu_pct	39.80	69.35	79.80	84.98	94.30
social_support	67.40	88.57	91.95	93.82	100.00
voter_turnout	34.35	64.22	71.96	78.78	92.16

Table 1: Summary statistics (Min, Q1, Median, Q3, Max) of variables for Western Europe.

Variable	East Min	East Q1	East Median	East Q3	East Max
air_pollution	5.20	10.00	13.50	16.20	24.80
broadband_access	77.60	87.00	89.00	92.00	96.10
civil_liberties	5.00	6.00	6.00	7.00	7.00
disp_income_pc	9336.00	14451.75	16176.50	17240.50	26002.00
employment_rate	61.40	68.10	70.80	73.35	78.90
freedom_score	69.00	82.00	89.00	90.25	95.00
homicide_rate	0.00	0.98	1.30	1.82	6.10
internet_speed	-60.03	-41.48	-23.32	-0.41	43.48
life_expectancy	72.60	75.38	76.20	77.60	81.80
life_satisfaction	5.60	6.07	6.30	6.43	7.30
mortality_rate	7.57	10.82	11.98	12.87	16.16
number_rooms	1.10	1.20	1.42	1.60	1.95
political_rights	5.00	6.00	7.00	7.00	7.00
secondary_edu_pct	78.20	89.20	92.00	94.62	98.00
social_support	5.80	88.00	90.60	92.93	96.20
voter_turnout	42.63	53.38	60.88	64.05	75.42

Table 2: Summary statistics (Min, Q1, Median, Q3, Max) of variables for Eastern Europe.

Overall, the comparison between Western and Eastern Europe highlights clear differences in socio-economic and well-being indicators. Western countries tend to display higher disposable income, life expectancy, life satisfaction, and freedom-related measures, together with lower mortality and homicide rates. Conversely, Eastern countries show higher air pollution, mortality, and homicide rates. Some indicators, such as broadband access and secondary education, appear relatively closer between the two regions, although Western Europe generally maintains a slight advantage.

Next, focusing more specifically, we compared the distribution of four key variables across the two macro-areas: disposable income, life expectancy, voter turnout, and freedom score.

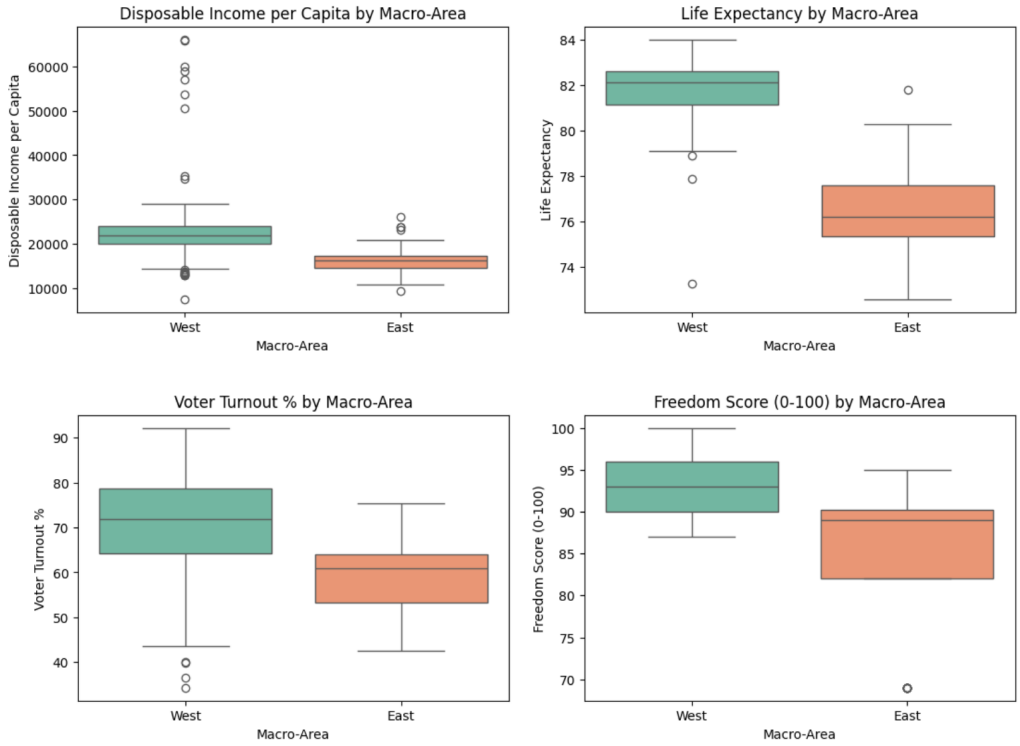

Figure 1: Distribution of four key variables across the two macro-areas.

Disposable Income per Capita: Western Europe exhibits substantially higher income levels, with not only the median but even the third quartile in the West lying well above the first quartile in the East. Moreover, income variability is greater in the West, reflecting stronger economic heterogeneity among regions. Eastern Europe, instead, shows more compact but consistently lower values.

Life Expectancy: Life expectancy is consistently higher in Western Europe compared to Eastern Europe. The median value in the West is above 82 years, while in the East it is around 76 years. Moreover, the interquartile range of the West lies entirely above that of the East, indicating a systematic gap rather than just differences in extremes.

Voter Turnout: Western Europe shows higher participation in national elections, with a median voter turnout around 71% significantly above that of Eastern Europe. Additionally, turnout is more dispersed in the West, meaning that some regions achieve very high levels of participation while others fall closer to Eastern averages.

Freedom Score: Western Europe consistently achieves higher scores. Eastern Europe, while generally close, shows a lower median and a few clear outliers at the bottom, signaling that democratic and civil liberties are not as uniformly guaranteed as in Western regions.

2.5 West VS East: T-test and Mann-Whitney test

In order to statistically assess whether the observed differences between Eastern and Western Europe are significant across the considered indicators, we applied two complementary hypothesis tests: the t-test and the Mann–Whitney U test. Both methods allow us to evaluate whether the distributions of the two groups differ systematically, but they rely on different assumptions and provide distinct perspectives on the data.

The t-test (with Welch's correction) assumes that data are approximately normally distributed within each group and that variances may differ across groups. It compares the null hypothesis, under which, the two populations share the same mean, against the alternative: the two populations don't share the same mean.

The Mann–Whitney U test, by contrast, is a non-parametric alternative that does not rely on the assumption of normality. Instead of focusing on means, it assesses whether two independent samples originate from the same underlying distribution. Formally, the null hypothesis states that the probability distribution of the two groups is identical, while the alternative hypothesis implies a systematic shift in values between them. It evaluates whether observations in one group tend to be larger or smaller than those in the other.

Applying both tests allows us to cross-check the robustness of our findings: consistent significance across the two methods would provide strong evidence of systematic East–West disparities.
The results of these tests are reported in Table 2. For each variable, we provide group means (East vs. West), together with p-values from both the t-test and the Mann–Whitney test.

Variable	Mean East	Mean West	t-test p-value	Mann-Whitney p-value
life_expectancy	76.49	81.84	2.05e-36	1.50e-28
mortality_rate	11.78	7.41	2.45e-33	1.66e-28
secondary_edu_pct	91.24	76.76	7.48e-32	2.97e-22
number_rooms	1.40	1.81	4.33e-19	2.06e-15
disp_income_pc	16205.28	23057.48	4.59e-16	5.30e-16
life_satisfaction	6.29	6.85	3.08e-15	4.63e-11
voter_turnout	58.88	70.23	1.42e-12	2.10e-11
freedom_score	85.47	93.32	1.36e-10	8.68e-13
civil_liberties	6.18	6.81	7.92e-10	2.31e-13
air_pollution	13.26	9.72	8.51e-08	9.51e-09
political_rights	6.45	7.00	1.96e-07	1.96e-18
homicide_rate	1.63	0.92	8.54e-05	1.15e-09
broadband_access	88.92	90.94	4.51e-03	3.87e-03
internet_speed	-20.24	-8.78	1.17e-02	1.09e-02
employment_rate	70.70	68.67	2.96e-02	7.31e-01
social_support	87.17	91.02	6.70e-02	1.66e-02

Table 3: Comparison of East and West Europe: group means and p-values from t-test and Mann–Whitney test.

The results of both the t-test and the Mann–Whitney test confirm the existence of systematic and statistically significant differences between Eastern and Western Europe across most of the socio-economic and political indicators considered.

The strongest disparities emerge for life expectancy, mortality rate, and secondary education attainment, where p-values are the lowest in both tests. Western regions exhibit substantially higher life expectancy and, being strongly associated, lower mortality rates, while Eastern regions report consistently higher levels of secondary education completion, which represents the only indicator in which the East clearly outperforms the West. Similarly, variables related to material well-being, such as disposable income per capita and number of rooms per person, show large and highly significant gaps, with Western Europe enjoying a clear advantage.
Political dimensions follow the same pattern: freedom scores, civil liberties, and political rights all display significantly better values in the West.
Two variables, however, stand out as less consistent. Employment rate shows a weaker significant difference in the t-test, while the Mann–Whitney test does not reject the null hypothesis (considering $\alpha = 0.05$), suggesting broadly comparable employment levels between East and West. Social support also displays mixed evidence, with a significant difference in the Mann–Whitney test but not in the t-test, pointing to potential distributional differences rather than a clear gap in averages.

Overall, the convergence of results across both the parametric and non-parametric approaches strength-

ens the robustness of these findings.

To complement the hypothesis tests, Figure 2 presents a ranked visualization of the indicators based on the magnitude of the differences between the two macro-areas. After computing group means for Eastern and Western Europe, each variable was rescaled to a common [0,1] range using min–max normalization. On this normalized scale, we then calculated the difference between East and West, which allows for a direct comparison of disparities across heterogeneous indicators.

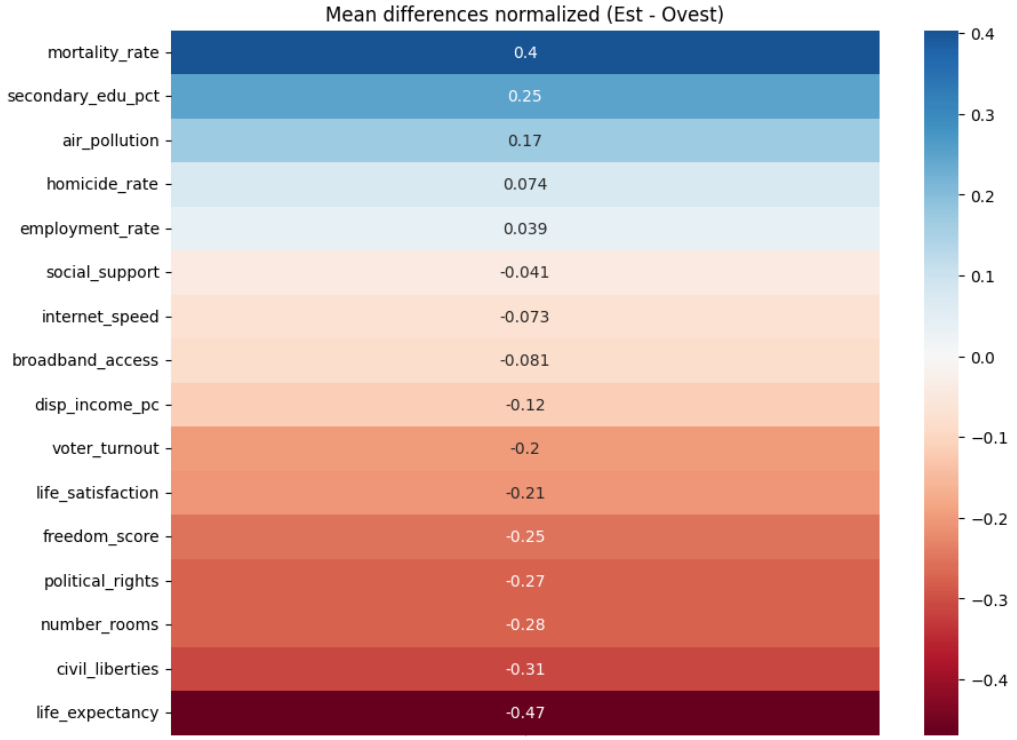

Figure 2: Normalized East–West Differences Across Indicators

2.6 Internal Cohesion

The purpose of this section is to move from the regional to the national level, by assessing the degree of internal cohesion within each country. Specifically, we construct a cohesion index that quantifies how much variability exists among the regions belonging to the same country. A low value of this index indicates that regions are relatively homogeneous in their socio-economic and political characteristics (high cohesion), while a high value reflects stronger heterogeneity across regions (low cohesion).

As a first step, we reduced the dimensionality of the dataset through Principal Component Analysis (PCA). All variables were standardized prior to the analysis, in order to ensure comparability across indicators with different scales. We imposed the requirement that the selected components explain at least 85% of the total variance. This condition led to the retention of 7 principal components, which summarize the main directions of variability in the data while preserving most of the original information.

The cohesion index was constructed as the average of two complementary measures of within-country variability:

1. **Dispersion in PCA space**: for each country, we computed the centroid of its regions in the PCA space and measured the average Euclidean distance of the regions from this centroid.

2. **Dispersion in standardized original variables**: in parallel, we calculated the standard deviation of each standardized indicator within every country, and then averaged across indicators. This provides a measure of the average within-country variability across all dimensions of the dataset.

The overall **Cohesion Index** is then defined as the arithmetic mean of these two measures.
By construction, lower values of the Cohesion Index indicate higher internal cohesion (regions are similar), while higher values correspond to stronger internal heterogeneity.

Figure 3 presents the Cohesion Index for all countries in the sample, sorted in ascending order. Countries on the left of the chart, such as Luxembourg, Denmark, and Iceland, display highly homogeneous regional profiles, indicating a strong degree of internal cohesion. In contrast, countries on the right, most notably Estonia, France, and Italy, are characterized by markedly higher internal dispersion, suggesting pronounced regional divides.

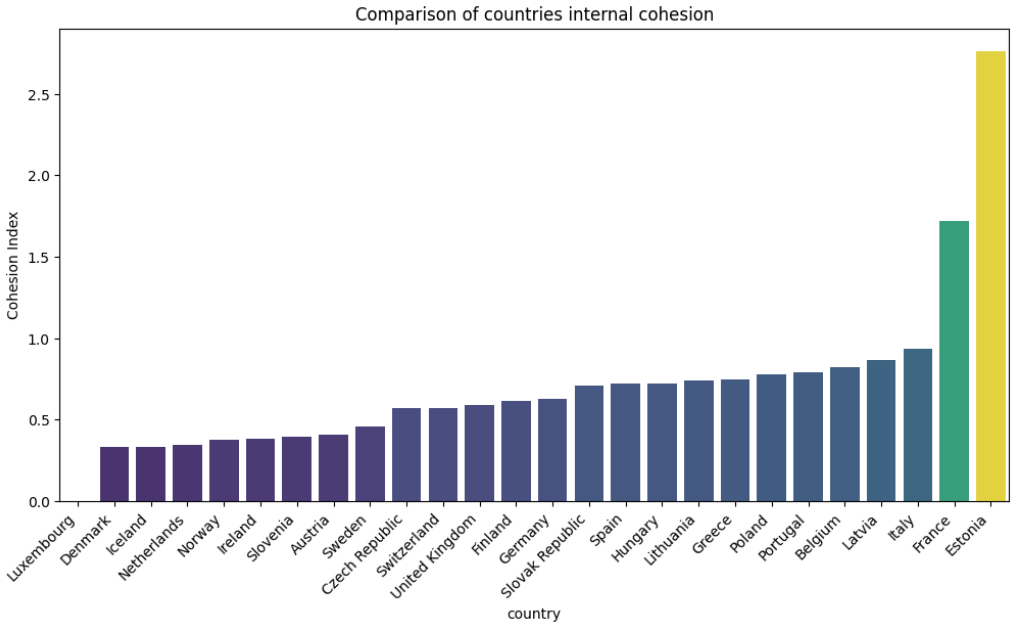

Figure 3: Countries Internal Cohesion

2.6.1 Internal Cohesion by EU group and by Freedom category

To further explore the determinants of internal cohesion, we classified countries according to two complementary dimensions: European Union membership status and political freedom category.

EU group classification. Countries were divided into three groups.
The first group includes the founding members and early enlargements of the European Union, commonly referred to as *EU15*: *Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden, United Kingdom.* The second group comprises the Central and Eastern European countries that joined in 2004, labeled as *EU post-2004*: *Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Slovak Republic, Slovenia.* Finally, all remaining OECD European countries not belonging to the EU are grouped under *Non-EU*: *Iceland, Norway, Switzerland.*

Freedom category classification. We also classified countries into categories of political freedom based on the *Freedom House* score. Countries with a score greater or equal to 70 are labeled as *Free*, those between 35 and 69 as *Partly Free*, and those below 35 as *Not Free*. In our sample, no OECD European country fell into the *Not Free* category. This classification broadly corresponds to international standards used by Freedom House, although it is worth noting that all OECD members considered are already consolidated democracies. Thus, the distinction between *Free* and *Partly Free* should be interpreted with caution, as it might reflect more subtle institutional differences rather than sharp democratic divides.

Results. Table 4 reports the median cohesion index for the three EU groups, while Table 5 summarizes results by freedom category. The results indicate that *Non-EU* countries exhibit the lowest internal dispersion, with a cohesion index of 0.38. This suggests that countries like Iceland, Norway, and Switzerland are relatively homogeneous internally. Among EU members, the *EU15* countries appear more cohesive (0.61) than the *EU post-2004* group (0.73), which shows stronger internal divides across regions.

When turning to political freedom, the distinction is more nuanced: countries classified as *Free* show a lower cohesion index (0.61), while those labeled *Partly Free* present higher internal heterogeneity (0.72).

EU Group	Median Cohesion Index
EU post-2004	0.73
EU15	0.61
Non-EU	0.38

Table 4: Median cohesion index by EU group.

Freedom Category	Median Cohesion Index
Free	0.61
Partly Free	0.72

Table 5: Median cohesion index by Freedom House category.

Overall, both classifications point in the same direction: countries from the Eastern enlargement of 2004 and those with lower levels of political freedom tend to display higher internal dispersion among regions. Conversely, long-standing EU members and non-EU OECD countries appear more cohesive, suggesting that historical integration and stronger institutional frameworks may contribute to reducing within-country disparities.

3 Experiments and Results

3.1 K-Means

To identify potential patterns and groupings within the data, we employed the K-Means clustering algorithm. We chose this algorithm due to its simplicity and interpretability, trying to provide a first approximation of structural differences between European regions. The goal here is to assess whether countries naturally form clusters that align with, or diverge from, the East–West divide highlighted in previous analyses.

3.1.1 Finding Best K

As discussed in the previous section, a Principal Component Analysis (PCA) had already been performed on the standardized variables. We retained the number of components necessary to explain at least 85% of the total variance, which resulted in 7 principal components summarizing the main directions of variability in the data while preserving most of the original information.

To determine the most appropriate value of K, we iterated over values from 2 to 10. For each configuration, we computed two diagnostic metrics:

- *Inertia (Elbow method)*: it measures the total within-cluster variance. Lower values indicate tighter clusters, but inertia always decreases as K increases.
- *Silhouette score*: evaluates the cohesion and separation of clusters, ranging from -1 (poor clustering) to +1 (well-separated clusters).

The Elbow method did not provide a clear cutoff in our data, reflecting one of the limitations of K-Means: it tends to produce clusters even when the data structure is not well-suited for partitioning, and the inertia curve may not exhibit a distinctive elbow. The Silhouette method, however, offered a more robust criterion, identifying the optimal number of clusters as K=2.

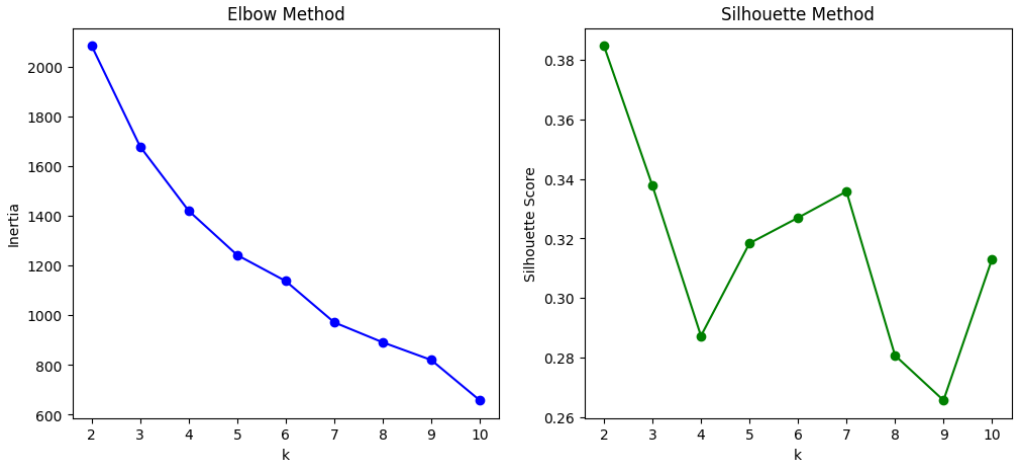

Figure 4: Elbow and Shilouette curves

Based on this result, we ran the final K-Means model with two clusters. Figure 4 shows the Silhouette and Elbow analysis, while Figure 5 provides a two-dimensional visualization of the clustering solution, plotting the first two principal components (PC1 and PC2). Although only a projection of the full 7-dimensional space, this scatterplot offers an intuitive representation of how the regions are grouped under the chosen clustering solution.

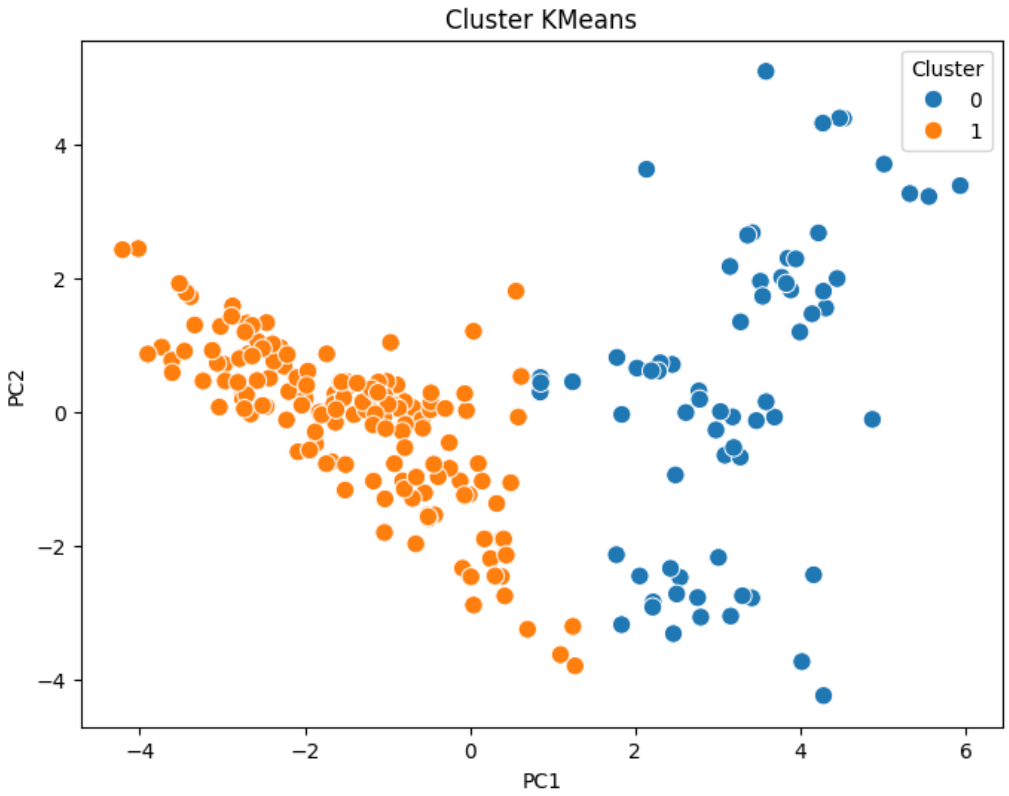

Figure 5: K-Means Clustering of European Regions on the First Two Principal Components

3.1.2 Cluster Analysis Results

The application of K-Means with K=2 revealed two distinct groups of European regions. While the partition is necessarily a simplification of the multidimensional structure of the data, the analysis allows us to highlight systematic contrasts between the clusters in terms of the indicators. In what follows, we present a comparative overview of the two groups through both a graphical and a tabular summary, focusing on the variables that most strongly differentiate them.

Figure 6 provides a radar chart comparing the normalized means of the two clusters across a set of 15 key variables. The radar plot highlights a clear asymmetry between the two clusters. One cluster achieves globally better results across most indicators, with the exceptions of *air pollution*, *homicide rate*, and *secondary education percentage*, where the other cluster performs relatively better (though not always in a desirable direction, as in the case of higher homicide rates and air pollution). The visualization underscores how clusters capture broad patterns rather than uniform profiles: each group displays strengths and weaknesses across different dimensions. Notably, one of the areas where the difference between clusters remains consistently high is that of political dimensions: *political rights*, *civil liberties*, and *freedom score*, which strongly characterize the overall separation.

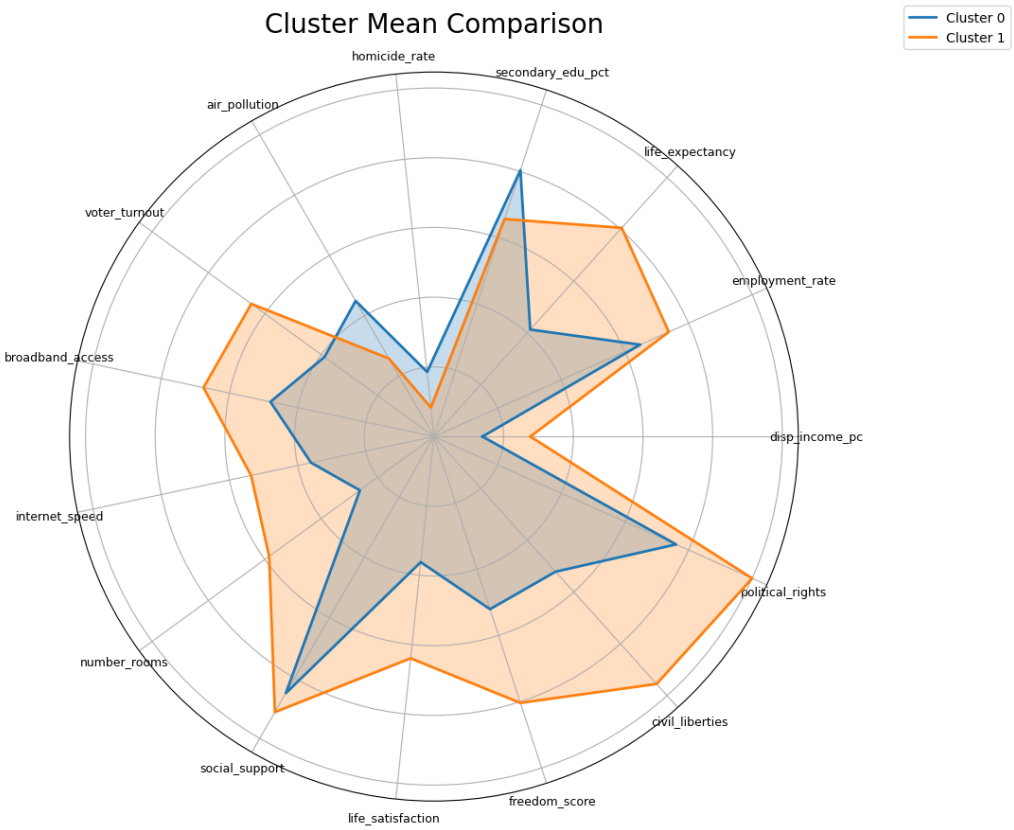

Figure 6: Radar plot of normalized cluster means across selected variables

To systematically identify which variables drive the separation between clusters, we computed the normalized difference in mean values across clusters. This measure indicates how much the clusters diverge on a standardized scale [0,1]. Interestingly, political and well-being indicators such as *civil liberties*, *freedom score*, and *life satisfaction* dominate the ranking, suggesting that institutional quality and perceived freedoms play a crucial role in shaping the East–West divide. In addition, demographic factors like *life expectancy* (and, conversely, *mortality rate*) also emerge as highly relevant, further underlining the multidimensional nature of the separation.

3.1.3 Supervised Approach

After applying the unsupervised K-Means clustering, we now compare the obtained clusters with the East-West labels that were assigned at the beginning of the analysis. This supervised validation allows us to assess to what extent the clustering solution aligns with the actual macro-regional division.

Figure 7 presents a scatterplot of the first two principal components, where colors represent the cluster assignment and markers indicate the true East/West labels.

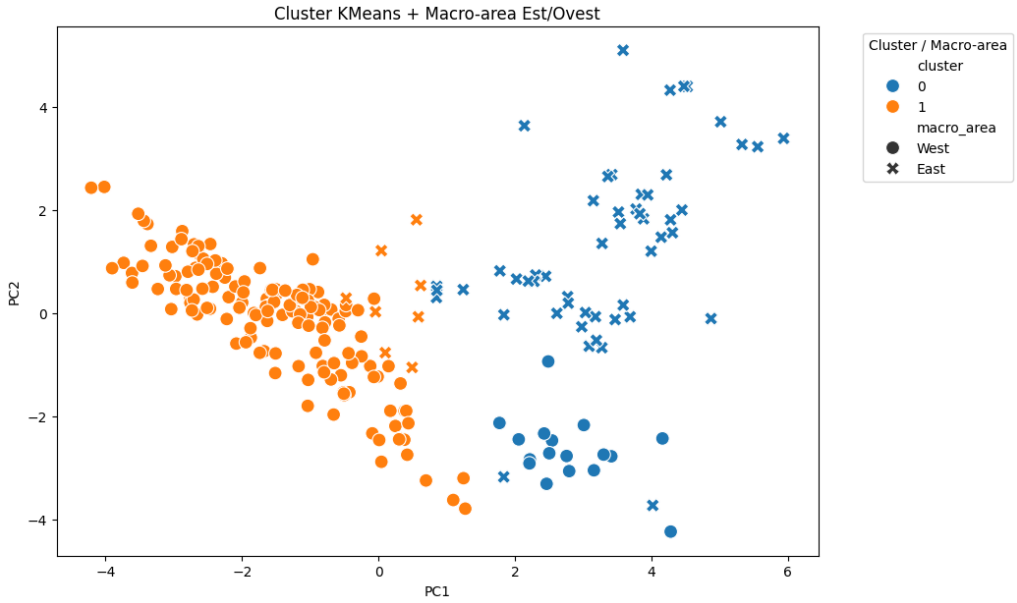

Figure 7: Comparison of K-Means clustering with East/West macro-area labels on the first two principal components.

To quantify the relationship between clusters and macro-areas, we computed the distribution of East and West regions within each cluster. Table 6 reports the absolute frequencies, while Table 7 shows the corresponding percentages.

	East	West
Cluster 0	52	17
Cluster 1	8	148

Table 6: Counts per cluster / macro-area.

	East	West
Cluster 0	75.36%	24.64%
Cluster 1	5.13%	94.87%

Table 7: Percentages per cluster / macro-area.

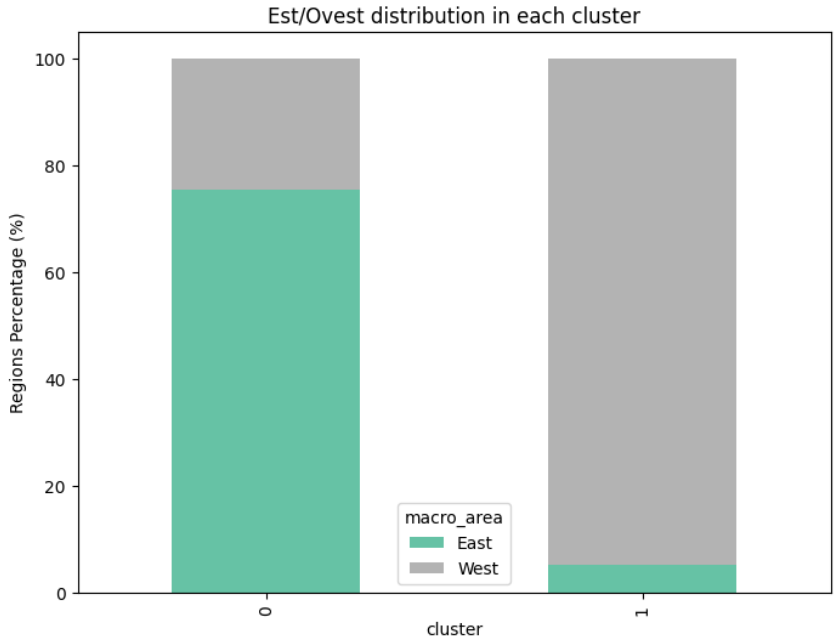

Figure 8: Distribution of East/West regions within each cluster (percentages).

The results highlight a correspondence between the clustering solution and the East/West macro-division. Cluster 0 is predominantly composed of Eastern regions, while Cluster 1 is overwhelmingly Western. Nevertheless, there are some misclassifications: a non-negligible share of Western regions falls into Cluster 0, and a small number of Eastern regions are included in Cluster 1. This confirms that the East/West divide is the main line of separation, but some exceptions reveal additional internal diversity.

To summarize the degree of alignment between clusters and true labels, we computed the *clustering purity*.

Purity is defined as the proportion of observations that are correctly assigned to the dominant class within each cluster, providing a measure of how well the clustering matches the original labels.
In our case, the purity reaches a value of **0.89**, meaning that 89% of the regions are correctly classified with respect to their East/West macro-area. This confirms that the unsupervised K-Means solution, despite not using the labels, is highly consistent with the actual geopolitical division.

3.1.4 Supervised Approach: Bridge Regions

Although the clustering solution aligns very closely with the East/West division, the misclassified regions are particularly interesting from an analytical perspective. These cases often correspond to "bridge" areas whose socioeconomic and political profiles combine features of both groups. These misclassifications provide evidence on the internal diversity of Europe. They highlight that the East/West division, while dominant, is not absolute.

In total, there are 25 such bridge regions. Table 8 reports the country, region, macro-area, assigned cluster, and the dominant macro-area of that cluster.

Country	Region	Macro-area	Cluster	Cluster Dominant Area
France	French Guiana	West	0	East
France	Guadeloupe	West	0	East
France	Martinique	West	0	East
France	Mayotte	West	0	East
Greece	Attica	West	0	East
Greece	Central Greece	West	0	East
Greece	Central Macedonia	West	0	East
Greece	Crete	West	0	East
Greece	Eastern Macedonia, Thrace	West	0	East
Greece	Epirus	West	0	East
Greece	Ionian Islands	West	0	East
Greece	North Aegean	West	0	East
Greece	Peloponnese	West	0	East
Greece	South Aegean	West	0	East
Greece	Thessaly	West	0	East
Greece	Western Greece	West	0	East
Greece	Western Macedonia	West	0	East
Czech Republic	Prague	East	1	West
Czech Republic	Southwest	East	1	West
Estonia	North Estonia	East	1	West
Estonia	Southern Estonia	East	1	West
Estonia	West Estonia	East	1	West
Slovak Republic	Bratislava	East	1	West
Slovenia	Eastern Slovenia	East	1	West
Slovenia	Western Slovenia	East	1	West

Table 8: List of the 25 bridge regions, i.e., regions assigned to a cluster whose dominant macro-area differs from their own.

A closer look at the table shows that the French bridge regions correspond to overseas territories (French Guiana, Guadeloupe, Martinique, Mayotte), which are geographically distant from continental Europe. Their assignment to the Eastern cluster may reflect certain structural disadvantages, such as lower economic development and limited infrastructure, compared to continental regions. Conversely, some Eastern regions are classified in the Western cluster due to their higher development and urbanization, as exemplified by the capitals Prague and Bratislava. This highlights how socioeconomic and political characteristics can create "bridge" regions that combine features of both clusters.

Building on the analysis of the 25 bridge regions, we computed the distance of each region from the centroid of the cluster to which it was assigned. By measuring this distance in the PCA-reduced space, we can assess how centrally located a misclassified region is within its (incorrect) cluster.
Regions that are both misclassified and close to the centroid of the wrong cluster may represent extreme examples of structural similarity to that cluster. In other words, these regions could be seen as particularly "virtuous" or "outlier" cases: despite belonging to one macro-area, their characteristics make them strongly aligned with the opposite cluster.

Table 9 reports the ten bridge regions with the smallest distance to their assigned cluster centroid.

Country	Region	Macro-area	Cluster	Distance to Centroid
Estonia	North Estonia	East	1	2.12
Slovenia	Western Slovenia	East	1	2.13
Greece	Attica	West	0	2.39
Estonia	Southern Estonia	East	1	2.80
Czech Republic	Prague	East	1	2.80
Greece	Central Macedonia	West	0	3.00
Slovenia	Eastern Slovenia	East	1	3.04
Estonia	West Estonia	East	1	3.09
Czech Republic	Southwest	East	1	3.14
Greece	Peloponnese	West	0	3.16

Table 9: Bridge regions closest to the centroid of their assigned cluster.

Examining these central bridge regions provides interesting insights. For instance, the Estonian regions and the Czech capitals Prague are misclassified into the Western cluster, yet they lie close to its centroid, suggesting that their economic and political profiles are highly aligned with Western characteristics. Similarly, Greek regions such as Attica and Peloponnese are assigned to the Eastern cluster but are central within it, highlighting structural similarities with Eastern regions despite their Western macro-area label.

3.1.5 K-Means without Political Freedom Indicators

In the previous section, the clustering analysis was performed using the full set of socio-economic and political indicators. While this approach provides a comprehensive overview of regional disparities, it may raise the concern that political freedom variables: *freedom score*, *political rights*, and *civil liberties*, could dominate the separation, artificially reinforcing the East–West divide.
To address this issue, we re-estimated the K-Means clustering after removing the three Freedom House indicators.

The procedure mirrored the previous pipeline: after scaling the variables and applying PCA to retain at least 85% of the total variance, we ran the K-Means algorithm with the same optimal number of clusters ($K = 2$). The resulting cluster assignments were then compared with those obtained in the full model.

To quantify the similarity between the two solutions, we computed the *Adjusted Rand Index* (ARI), which measures the agreement between two clustering partitions, corrected for random chance. The ARI ranges from 0 (random agreement) to 1 (perfect consistency). In our case, the ARI reached a value of **0.82**, indicating a strong alignment between the two clustering solutions. This suggests that, even in the absence of political freedom indicators, the East–West divide remains evident.

Table 10 reports the contingency table between clusters obtained with and without political indicators. While the majority of regions maintain the same assignment, a non-negligible set of regions switch clusters. These results demonstrate that the observed East–West separation is not solely driven by political freedom variables. Although they contribute significantly to the clustering, socio-economic indicators alone are sufficient to reproduce a similar partition of European regions.

	without_0	without_1
with_0	68	1
with_1	17	139

Table 10: Confusion matrix between cluster assignments with and without political freedom indicators.

A detailed list of the 18 regions that changed their cluster assignment is provided in Table 11.

Country	Region	Macro Area	k_{with}	k_{wo}
Czech Republic	Southwest	East	1	0
Estonia	North Estonia	East	1	0
Estonia	West Estonia	East	1	0
Estonia	Southern Estonia	East	1	0
Germany	Bremen	West	1	0
Hungary	Budapest	East	0	1
Italy	Campania	West	1	0
Italy	Apulia	West	1	0
Italy	Basilicata	West	1	0
Italy	Calabria	West	1	0
Italy	Sicily	West	1	0
Portugal	Algarve	West	1	0
Portugal	Central Portugal	West	1	0
Portugal	Alentejo	West	1	0
Portugal	Azores (autonomous region)	West	1	0
Portugal	Madeira (autonomous region)	West	1	0
Slovak Republic	Bratislava	East	1	0
Slovenia	Eastern Slovenia	East	1	0

Table 11: Regions switching cluster assignment between the full model and the specification without political freedom indicators.

Interestingly, some of the switching regions belong to Southern Italy and Portugal. In particular, the Italian "Mezzogiorno" (Campania, Apulia, Basilicata, Calabria, Sicily) is systematically reassigned from the “Western” cluster to the “Eastern” one, when political freedom indicators are excluded. This suggests that, while Italy as a whole is firmly anchored in the West when democratic performance is taken into account, its Southern regions appear structurally closer to Eastern European standards in purely socio-economic terms. A similar dynamic emerges for Portugal, where peripheral regions, including the Algarve and the autonomous islands, are reclassified towards the Eastern cluster.

3.2 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that identifies dense groups of observations, while labeling as noise those points that lie isolated in sparse regions. The algorithm depends on two parameters: ε (eps), the radius used to search for neighbouring points, and *min_ samples*, the minimum number of points required to form a dense region. DBSCAN has two main advantages for our application: (i) it does not require specifying the number of clusters

in advance, and (ii) it explicitly identifies outliers (noise), which can be important when working with geographically and socio-economically heterogeneous regions.

3.2.1 Implementation details

We ran DBSCAN on the PCA-reduced representation. To choose a sensible ε we computed the 5-nearest-neighbour distances for each point, plotted the sorted distances (the K-distance plot) and inspected the "knee" of the curve. Based on that plot we selected

$$\varepsilon = 2.25, \qquad \text{min\_samples} = 5.$$

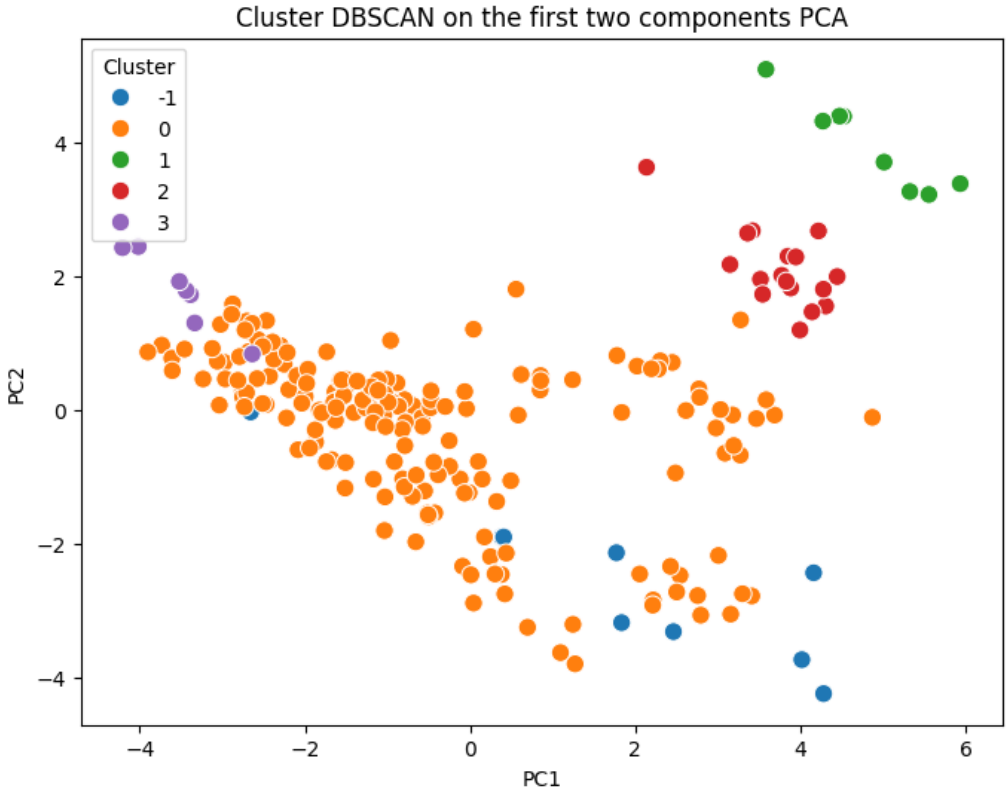

Figure 9: DBSCAN cluster allocation on the first two components of PCA.

3.2.2 Interpreting cluster archetypes

DBSCAN produced five distinct labels in our run: $-1, 0, 1, 2, 3$.

Cluster	East	West	Total
-1	2	7	9
0	33	151	184
1	8	0	8
2	17	0	17
3	0	7	7

Table 12: DBSCAN cluster counts by macro-area (East / West).

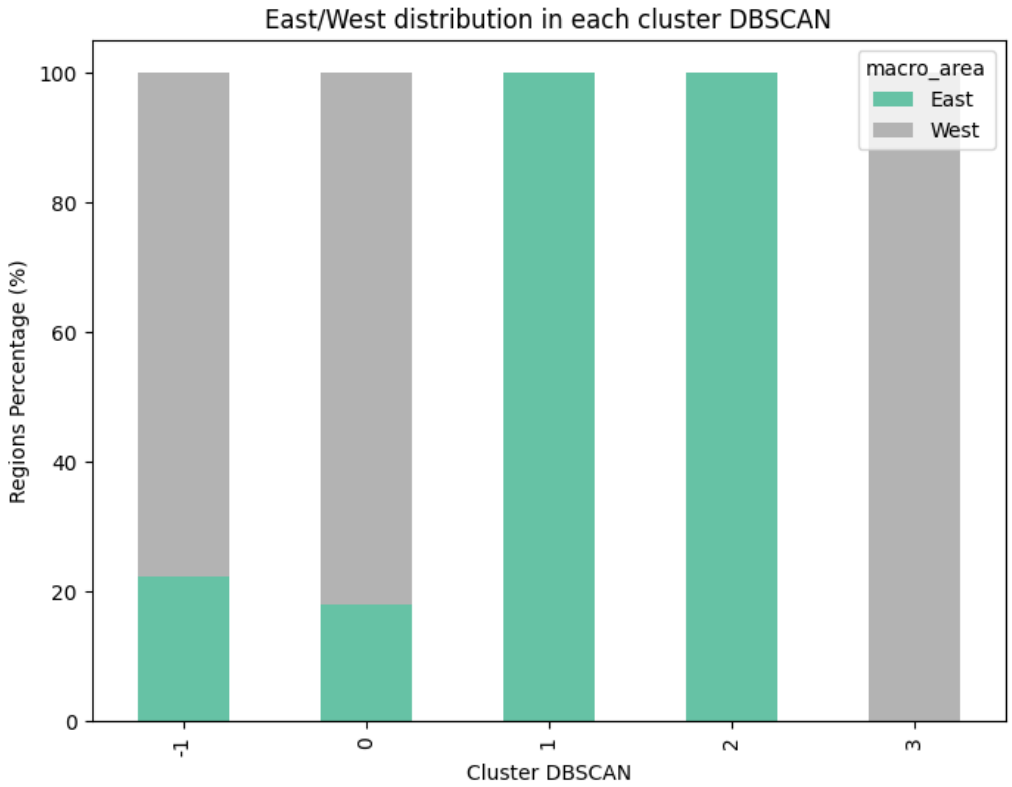

Figure 10: Distribution of DBSCAN clusters across macro-areas (East/West).

The overall clustering purity (matching the dominant macro-area per cluster) is ≈ 0.84, indicating a strong alignment with the East/West division while still revealing internal heterogeneity and noise.

To characterize each cluster, we inspected mean values of the original socio-economic and political indicators. The main patterns are:

- **Cluster 3 (very high-income Western group)** Mean disposable income per capita is exceptionally high, together with very high broadband access and above-average internet speed. This cluster likely contains very wealthy regions (e.g. Switzerland / Luxembourg in our dataset) and is clearly distinct from other Western regions.
- **Cluster 0 (typical Western)** This is the large Western cluster with intermediate-to-high levels of income, life expectancy and freedom indicators. It captures the broad majority of continental Western regions.
- **Cluster 1 (lower-income Eastern group)** Cluster 1 has relatively low disposable income and lower institutional scores. This group corresponds to Eastern regions with weaker institutional and material outcomes.
- **Cluster 2 (Eastern group with high educational attainment)** Cluster 2 is Eastern as well, but shows very high secondary schooling level, moderate incomes and somewhat different health indicators.
- **Cluster -1 (Noise / outliers)** The nine outliers include several French overseas regions and some small/island territories (e.g. Guadeloupe, Martinique, French Guiana, La Réunion, Mayotte, Corsica, Åland) as well as a couple of Estonian subregions. These areas are either geographically

remote or atypical in their socio-economic profile, so DBSCAN reasonably treats them as low-density points rather than members of a large homogeneous cluster.

3.2.3 How DBSCAN complements the K-Means findings

Compared to K-Means (which returned an unrefined two-cluster solution aligned with East vs West), DBSCAN provides three useful refinements:

1. it explicitly *flags outliers* that K-Means must nevertheless assign to some cluster;
2. it *splits the Eastern bloc* into at least two distinct density-based groups, revealing internal heterogeneity: one characterized by lower institutional quality and income (cluster 1) and another with high formal education but still lagging on some welfare measures (cluster 2);
3. it isolates a *small very-wealthy Western cluster* (cluster 3) that K-Means' two-cluster partition tended to merge into the broader Western group.

3.2.4 Caveats and robustness checks

DBSCAN is sensitive to the choice of ε and *min_samples*. Our eps was selected by visual inspection of the 5-NN distance plot; however, the knee is not always unambiguous. We therefore recommend reporting sensitivity checks (varying ε in a plausible range and testing *min_samples* = 4–8) and, if useful, repeating the analysis on the original standardized variables (rather than PCA scores) to verify stability.

3.2.5 Summary

In sum, DBSCAN confirms the prominent East–West structuring of European regions while adding two types of insight: (i) it isolates geographically or structurally atypical regions as noise, and (ii) it uncovers intra-Eastern heterogeneity that K-Means' coarse partition missed. These features make DBSCAN a useful complement to centroid-based algorithms when the aim is both to recover major divisions and to highlight exceptions and dense subgroups deserving targeted policy attention.

3.3 Counterfactual Analysis

Counterfactual analysis is a simulation approach used to assess how outcomes would change if certain explanatory variables were altered, while holding all others constant. In our case, we first estimated a baseline prediction of regional *disposable income per capita* using a Random Forest model. We chose this variable as the target because it provides a comprehensive proxy for the overall economic well-being of households, and is directly comparable across countries and regions (Figure 11).

Model training was carried out on the full sample of European regions, using all available socio-economic and institutional features. Predictive performance was assessed through 5-fold cross-validation, yielding an average R^2 of 0.77, indicating that the model captures a substantial share of income variation. To quantify feature relevance, we further computed permutation importance scores, which confirmed that both structural and institutional variables play a central role.

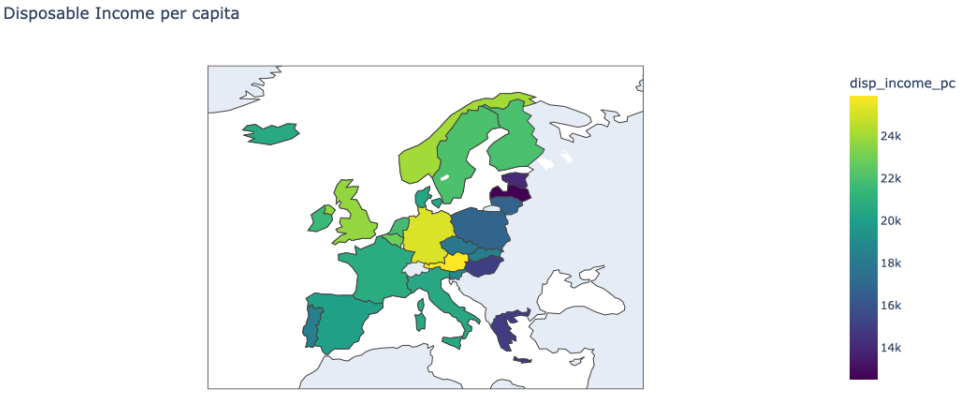

Figure 11: Baseline disposable income per capita across European regions.

Note: Switzerland and Luxembourg were removed from the map to better visualize differences among the other countries, as their exceptionally high values would otherwise compress the color scale.

3.3.1 Defined scenarios

Building on this baseline model, counterfactual scenarios were generated by replacing selected Eastern-region variables with the corresponding Western-region averages, while keeping all other covariates unchanged. For each scenario, predictions were re-estimated and the uplift in disposable income computed. To ensure robust inference, we repeated this procedure under bootstrap resampling ($n = 200$), reporting median effects and 95% confidence intervals.

For each scenario, we specify the variables subject to imputation.

- **Scenario 1 – Civic & Digital:** voter turnout, broadband access, Internet speed.
- **Scenario 2 – Welfare & Health:** mortality rate, life expectancy, life satisfaction, social support.
- **Scenario 3 – Human Capital & Employment:** employment rate, attainment of secondary education.
- **Scenario 4 – Institutional Strength:** freedom score, civil liberties, political rights.

The results are reported in terms of the predicted increase in disposable income per capita for the eastern regions, both in absolute and relative terms compared to the observed baseline.

3.3.2 General Results

Bootstrap estimates show that the largest potential uplift emerges in Scenario 2, with a median increase of €4,407 (95% CI: €3,462 – €5,376), corresponding to roughly +27.2% over the baseline. More moderate gains are associated with Scenario 1 and Scenario 4, each yielding an uplift of around €1,200–1,300 (+7.8% to +8.0%). By contrast, Scenario 3 displays only marginal effects, with the median estimate close to zero (95% CI: –€401 to +€712), corresponding to +0.7% and not statistically significant.

Scenario	2.5%	Median	97.5%	Median % uplift
Scenario 1 – Civic & Digital	€621	€1,291	€2,217	+7.97%
Scenario 2 – Welfare & Health	€3,462	€4,408	€5,376	+27.20%
Scenario 3 – Human Capital & Employment	-€401	€110	€712	+0.68%
Scenario 4 – Institutional Strength	€646	€1,268	€2,265	+7.83%

Table 13: Counterfactual uplift in Eastern regions (absolute and relative to baseline). Values are bootstrap medians with 95% confidence intervals.

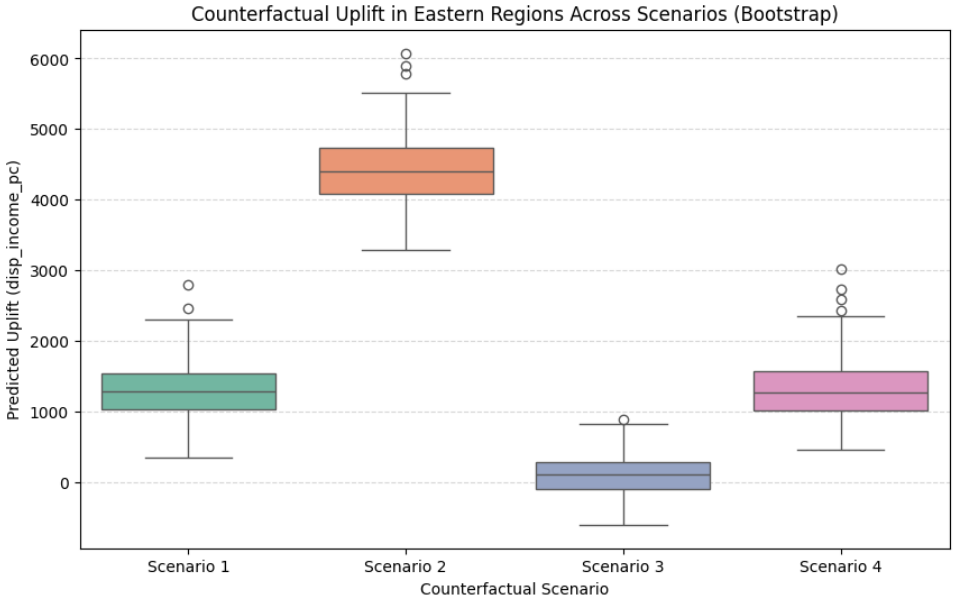

Figure 12: Counterfactual uplift in Eastern regions across scenarios (bootstrap).

3.3.3 Country-Level and Regional Heterogeneity

The counterfactual results reveal marked heterogeneity both across countries and within individual regions of Eastern Europe.

At the **country level**, Figure 13 shows that the size and even the sign of the predicted uplift vary substantially depending on the scenario.

Under Scenario 1 – Civic & Digital, the strongest positive effects are observed in Estonia and the Czech Republic, while Poland and Hungary experience only minor gains. In Scenario 2 – Welfare & Health, improvements are widespread across all Eastern countries, with Latvia and Estonia showing the largest relative increases in disposable income. By contrast, Scenario 3 – Human Capital & Employment produces negligible or even slightly negative changes in several countries (Slovak Republic, Poland, Lithuania), suggesting that structural constraints in labor markets may limit the returns from human capital improvements alone. Finally, in Scenario 4 – Institutional Strength, Hungary, Latvia, and Lithuania display the highest potential uplifts, whereas Estonia and Slovenia show almost no change, reflecting their already higher levels of institutional quality.

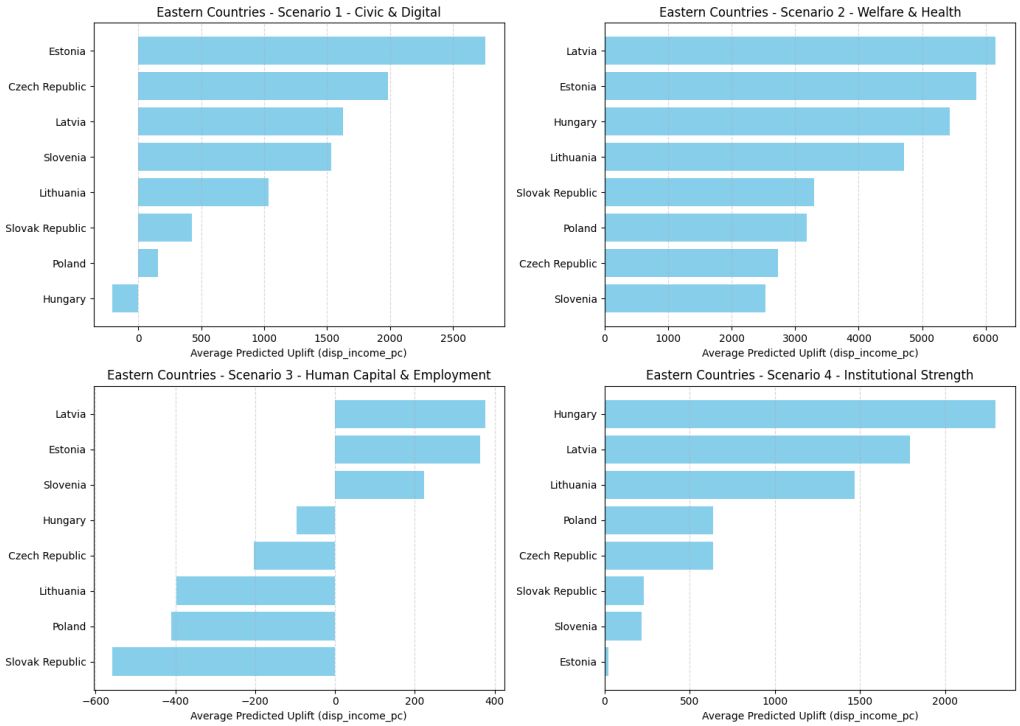

Figure 13: Average predicted uplift by Eastern country across scenarios

At the **regional level**, the distribution of benefits is even more uneven. Figure 14 reports the ten Eastern regions with the highest predicted income gains in each scenario.

Consistent with the country-level findings, Estonian regions dominate in Scenario 1, where digital and civic engagement variables are key drivers. In Scenario 2, several Latvian regions — most notably Latgale — rank at the top, together with disadvantaged areas of Hungary and Lithuania. In Scenario 3, gains are again concentrated in Latvia (Latgale, Kurzeme, Vidzeme), highlighting that the benefits from improvements in education and employment are highly localized rather than widespread. Finally, Scenario 4 emphasizes institutional reforms, with the top positions occupied by regions in Latvia, Lithuania, and Hungary, underlining the uneven distribution of institutional weaknesses within Eastern Europe.

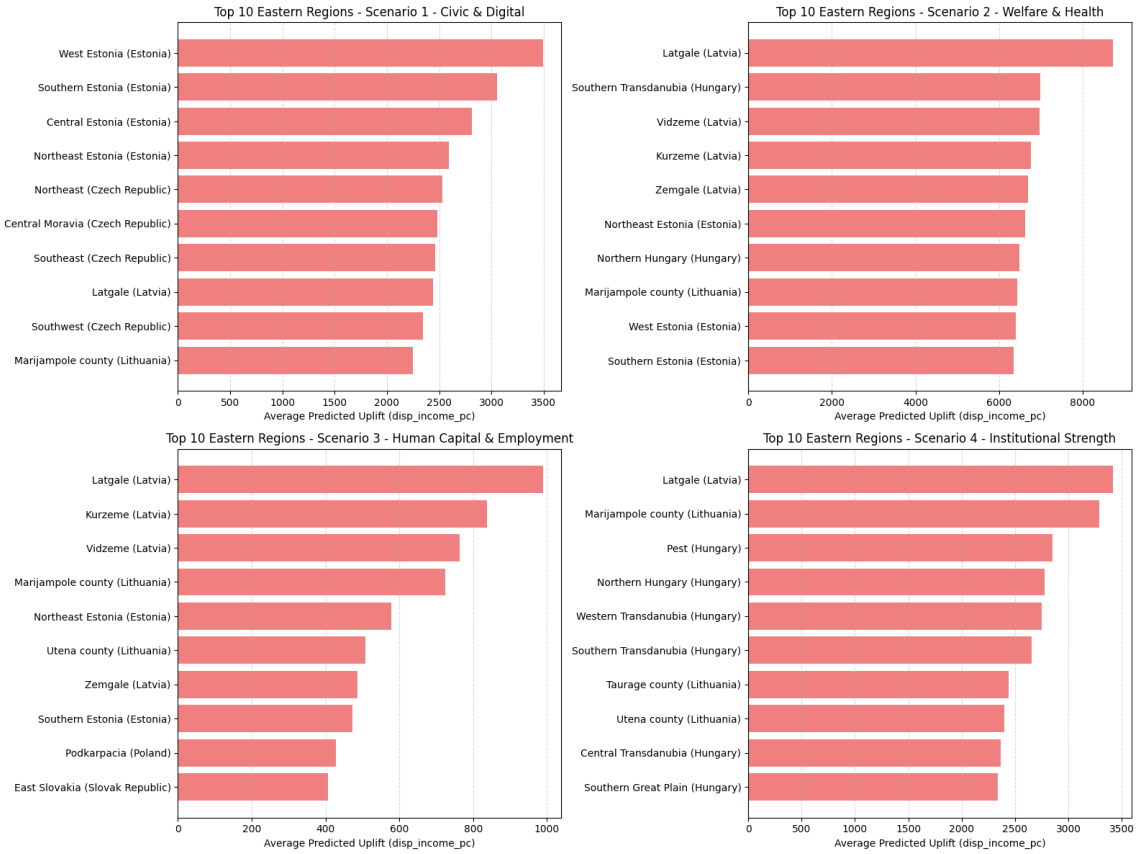

Figure 14: Top 10 Eastern regions by predicted uplift across scenarios.

3.3.4 Interpretation

The counterfactual analysis shows that welfare and health-related improvements (Scenario 2) deliver the broadest and most consistent gains across Eastern Europe, while other interventions display a much more selective geography. Digital access and civic participation benefit primarily the Baltic states, institutional reforms appear most relevant for parts of Hungary and Latvia, and improvements in education and employment yield only limited advantages concentrated in a few regions.

Overall, the drivers of income convergence in Eastern Europe emerge as diverse and context-dependent: rather than pointing to a single dominant pathway, the results highlight a mosaic of policy levers whose effectiveness varies across space. This underscores that future cohesion efforts should not only prioritize broad welfare improvements but also remain flexible enough to integrate region-specific strategies, ensuring that growth potential is unlocked in both leading and lagging areas.

4 Conclusion

Our analysis aimed to quantify the extent of the East–West divide in Europe by jointly considering socio-economic conditions and political freedoms.

The clustering results provide strong evidence of such a divide: K-Means with K=2 reached a clustering purity of 0.89, assigning 75% of Eastern regions to the “Eastern” cluster and almost 95% of Western regions to the “Western” one. In addition, DBSCAN, with an overall purity of around 0.84, not only confirmed this macro-division but also revealed relevant heterogeneity especially within Eastern Europe, distinguishing low-income areas from better-educated yet still disadvantaged regions. These findings

show that the gap is systematic and multidimensional.

The counterfactual analysis highlights which policy levers hold the greatest potential to reduce this divide. Welfare and health improvements stand out as the most effective intervention, with an estimated median uplift of €4,407 (+27.2%) in disposable income per capita for Eastern regions. On the other hand, digital participation and institutional reforms also deliver meaningful but smaller gains (+7/8%), while education and employment policies alone show negligible effects.
Taken together, these results suggest that investments in welfare and health systems should be prioritized promote alignment, with digital and institutional reforms serving as complementary tools for supporting cohesion across Europe.