Hierarchical clustering in World Energy Consumption data

Bc. Daniela Pillárová, Bc. Nicol Fedurcová

November 30, 2025

Abstract

This study explores the application of hierarchical clustering techniques to global energy consumption data, aiming to uncover patterns in electricity usage relative to socioeconomic indicators such as GDP and population size. Using data from 28 countries, we analyzed energy consumption across various renewable and non-renewable sources. Through a series of clustering experiments, we identified the weighted-cityblock configuration as the most effective clustering approach for our chosen data representatives, based on cophenetic coefficient values. Hierarchical Clustering achieved a marginally higher Silhouette Score (0.44) compared to K-means (0.41), highlighting its effectiveness in capturing nuanced patterns within the dataset. Key parameters, including oil consumption, wind consumption, and GDP, were instrumental in revealing distinct clusters, underscoring the relationship between economic development and energy consumption trends.

1 Aims

The primary aim of this study is to apply hierarchical clustering methods to analyze global energy consumption data, particularly focusing on electricity usage. We aim to uncover whether hierarchical clustering can effectively reveal groupings of countries based on their electricity and broader energy consumption profiles, providing insights into how factors such as economic development and population size influence energy consumption patterns globally.

2 Hypothesis

Our hypothesis is that countries with larger populations will exhibit higher electricity consumption levels compared to smaller ones. Our dataset includes energy consumption data across various renewable and non-renewable sources. We expect that hierarchical clustering will reveal distinct clusters in which countries with higher GDP will demonstrate greater consumption of renewable energy sources (such as solar, wind, and hydro energy) than those with lower GDP. Additionally, we anticipate that countries with both higher GDP and larger populations will generate more electricity overall.

3 Dataset description

The available dataset we used, World Electricity Consumption, is sourced from Kaggle and contains comprehensive information about global energy consumption, focusing on both renewable and non-renewable energy sources.

For our analysis, we specifically selected data for 28 countries, which include both developed and developing nations, such as Slovakia, France, Czechia, Spain, Germany, Croatia, Greece, Norway, Venezuela, Colombia, Vietnam, Thailand, Japan, Denmark, Hungary, Estonia, Poland, South Korea, Mexico, Egypt, Morocco, Turkey, Italy, Portugal, Peru, Lithuania, Latvia, Slovenia, Saudi Arabia. The dataset includes various energy consumption metrics from different kinds of sources like solar, wind, gas, hydro, and oil. In addition, the dataset provides key socioeconomic indicators like population and GDP, offering a broader context for analyzing energy consumption patterns. The data is from the year 2015, allowing us to examine how countries with varying levels of development and population sizes consumed and generated electricity (see table 1).

Column	Description	Unit
country	Geographic location.	String
year	Year of observation (2015).	Integer
population	Population by country, based on data and estimates from different sources.	Integer
gdp	Gross domestic product (GDP) - This data is adjusted for inflation and differences in the cost of living between countries.	Decimal
electricity_demand	Measured in terawatt-hours.	Decimal
electricity_generation	Total electricity generation - Measured in terawatt-hours.	Decimal
gas_consumption	Primary energy consumption from gas - Measured in terawatt-hours.	Decimal
hydro_consumption	Primary energy consumption from hydropower - Measured in terawatt-hours, using the substitution method.	Decimal
oil_consumption	Primary energy consumption from oil - Measured in terawatt-hours.	Decimal
solar_consumption	Primary energy consumption from solar power - Measured in terawatt-hours, using the substitution method.	Decimal
wind_consumption	Primary energy consumption from wind power - Measured in terawatt-hours, using the substitution method.	Decimal

Table 1: Columns description.

4 Methods and algorithms

4.1 Dataset Preparation

We began by preparing the dataset through several key steps. First, we filtered the dataset to include only the countries and parameters relevant to our analysis. Next, we standardized the dataset to ensure comparability across variables with different units and scales, which allowed each variable to have a uniform influence on the clustering process. To address missing values, we filled any null entries with zeros, maintaining consistency in the dataset and simplifying processing without introducing bias.

4.2 Initial Data Exploration

In the initial phase of data exploration, we generated a preliminary dendrogram using the ward linkage method and the default euclidean metric, allowing us to observe basic clustering patterns within the data (see figure 1).

Additionally, we created a cluster heatmap based on this initial dendrogram, which provided a visual overview of the clustering structures in the dataset. This heatmap helped us to familiarize ourselves with both the clustering methods and the data distribution (see figure 2).

4.3 Hierarchical Clustering Analysis

For the hierarchical clustering analysis, we experimented with multiple linkage methods and distance metrics to explore different clustering structures. To demonstrate our experimentations we created a combined dendrogram representation that shows all combinations of linkage methods and distance metrics that we tried (see figure 8). The same way we created a cummulated view of clustermaps (see figure 9).

To visually investigate relationships within the data, we created pair plots for each combination of columns in the dataset, enabling us to assess patterns and distinctions in the initial clusters (see figure 10).

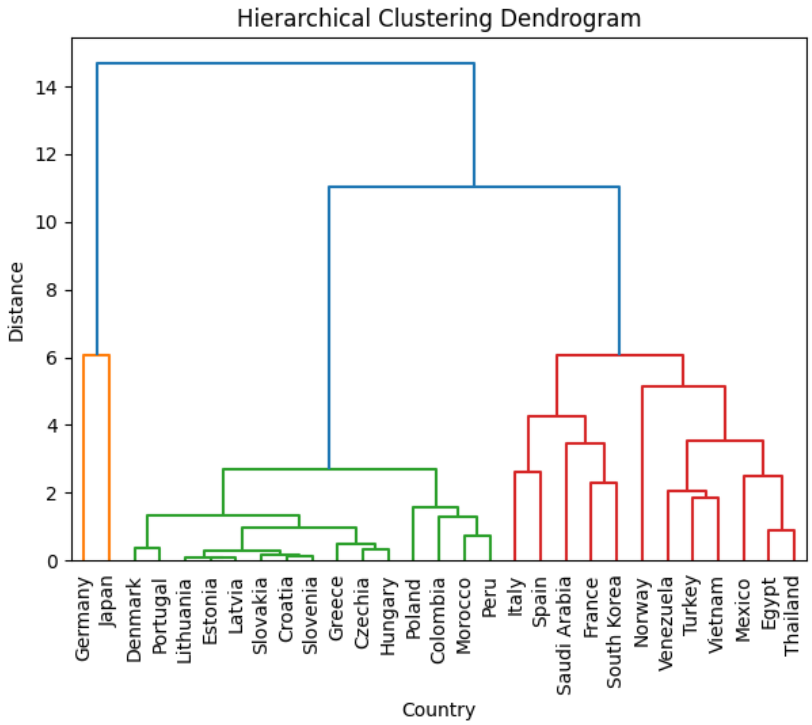

Figure 1: Initial dedrogram.

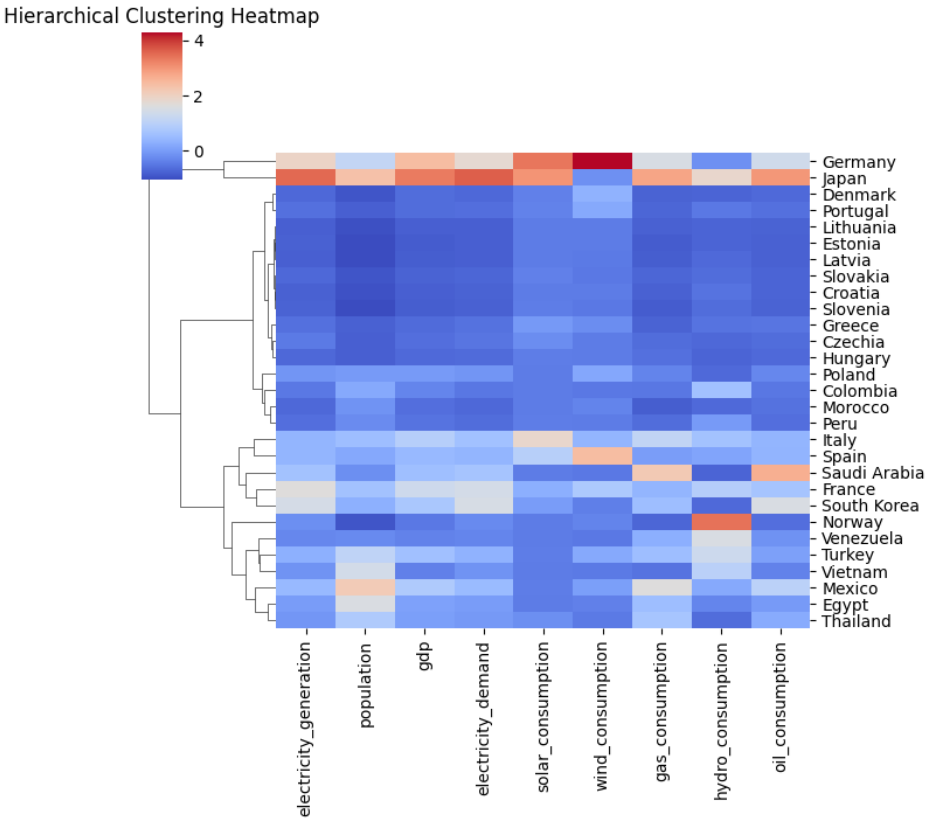

Figure 2: Initial hierarchical clustering heatmap.

Afterwards we have chosen tree combinations of linkage methods and distance metrics, that will allow us to experiment with different tresholds and criterias. To be able to compare this, we have chosen 3 columns as representatives of our data: gdp, wind consumption, oil consumption. The methods included single, complete, ward, average, centroid, and weighted linkage, while the distance metrics tested were euclidean, cityblock, and cosine. Specifically, we generated dendrograms for selected method-metric pairs, including ward-euclidean, single-cosine, and complete-cityblock, to capture variations in clustering behavior. Based on these combinations, we applied different thresholds to refine clustering: a threshold of 3 for ward-euclidean, 0.2 for single-cosine, and 8 for complete-cityblock. To further customize the clustering process, we used criteria such as maxclust and distance, aiming to optimize cluster formation across different configurations.

4.4 K-means Clustering

To be able to compare the clusters created by Hierarchical clustering, we tried the K-means method on the three chosen column representatnts of our data: gdp, wind consumption, oil consumption.

4.5 Evaluation Metrics

To evaluate the effectiveness and consistency of our clustering approach, we calculated two key coefficients for all method-metric combinations. First, we computed cophenetic coefficients to determine how well the hierarchical clustering preserved the pairwise distances in the data, providing a measure of clustering fidelity. Additionally, we calculated inconsistency coefficients to assess the stability and variability within each hierarchical clustering result, offering insights into the robustness of the clusters formed across different method-metric pairings.

As for the comparison of the Hierarchical Clustering algorithm and the K-means Clustering algorithm we decided to use Silhouette score. It has allowed us to measure cluster quality.

5 Chosen parameters and assumptions

For this analysis, we selected key parameters, such as oil consumption, wind consumption, and GDP. We attempted to identify and highlight distinct clusters of countries that exhibit similar energy consumption behaviors and economic profiles. Our analysis is based on data from the year 2015, which we assume to be relatively representative of the trends in energy consumption at the time. We further assume that the data is accurate and consistent across countries, despite potential variations in reporting methodologies.

6 Results

The clustering analysis revealed distinct grouping patterns across three different parameter combinations, resulting in three unique cluster distributions: ward-euclidian with treshold 3 and criterion maxclust (see figure 3), single-cosine with treshold 0.3 and criterion distance (see figure 4), weighted-cityblock with treshold 10 and criterion distance (see figure 5). Among these, the clustering configuration using the weighted linkage method and cityblock distance metric emerged as the most effective, based on cophenetic coefficient value 0.91 (see figure 6), indicating the highest accuracy of the three chosen pairs.

Kmeans clustering resulted on the same data with different clusters (see figure 7). When evaluating clustering quality using the Silhouette Score, we observed that Hierarchical Clustering achieved a score of 0.44, slightly outperforming K-means, which scored 0.41.

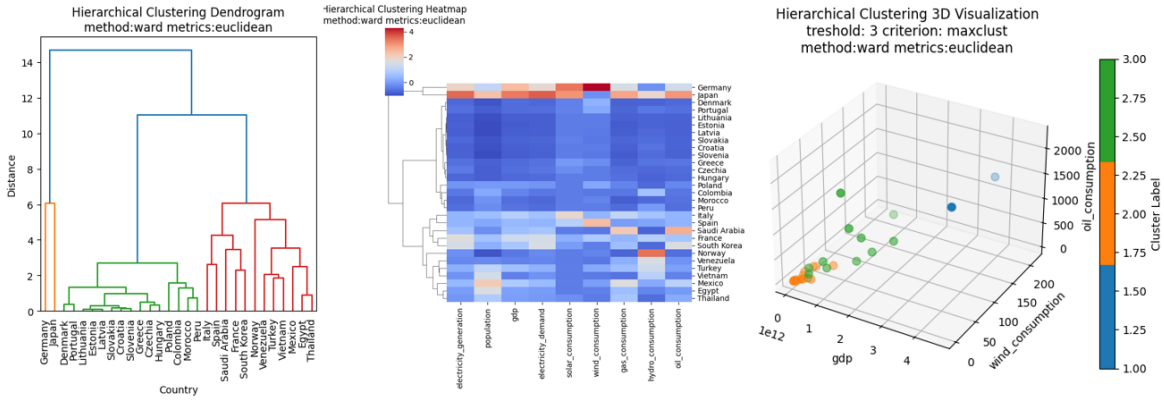

Figure 3: Results of ward-euclidian with treshold 3 and criterion maxclust.

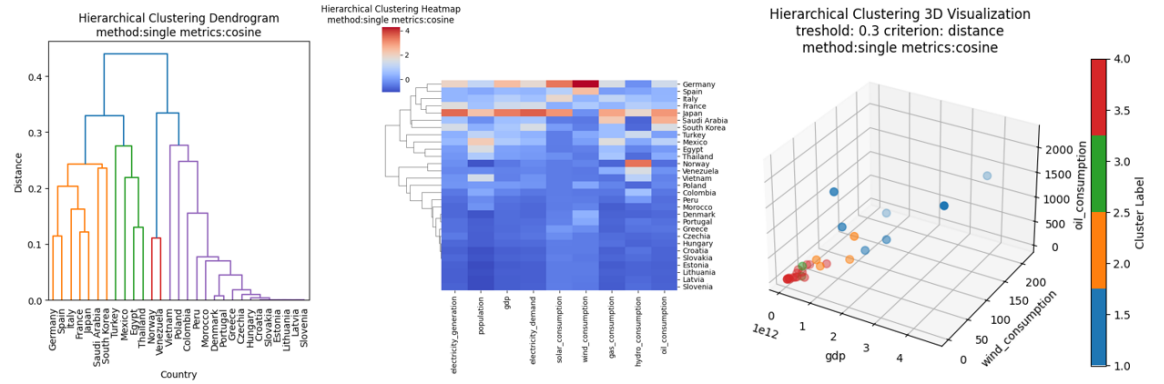

Figure 4: Results of single-cosine with treshold 0.3 and criterion distance.

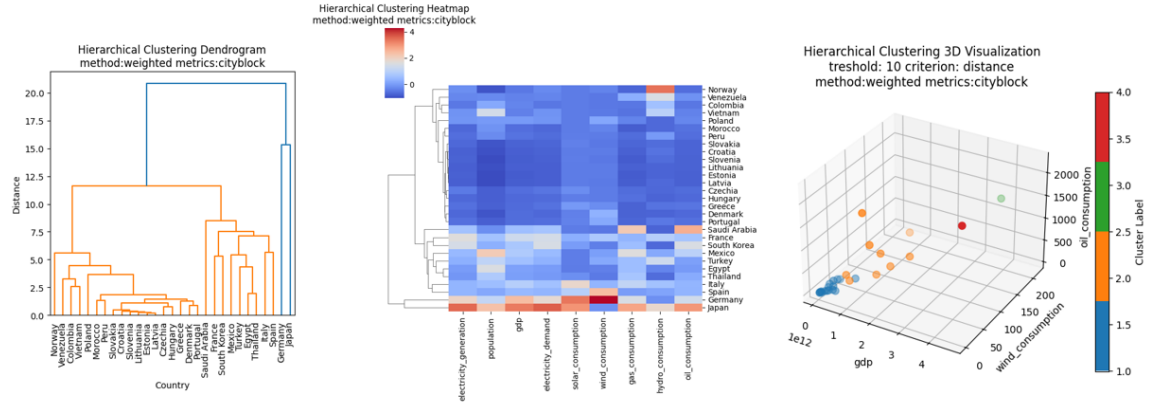

Figure 5: Results of weighted-cityblock with treshold 10 and criterion distance.

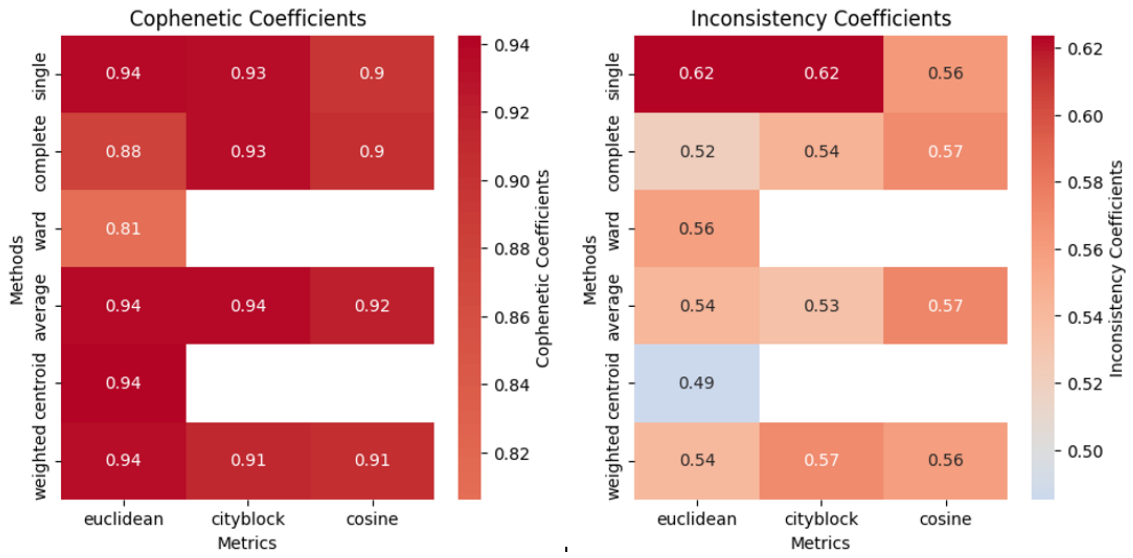

Figure 6: Results of Cophenetic Coefficients and Inconsistency Coefficients.

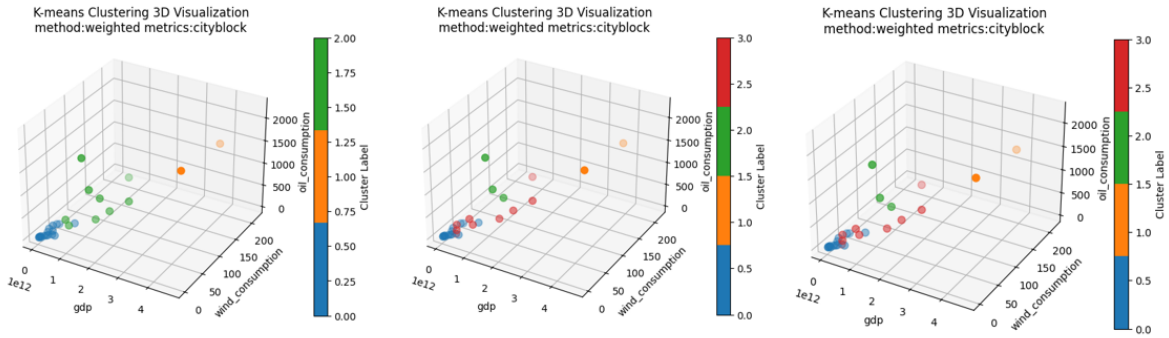

Figure 7: Results of clusters K-means.

7 Discussion

Our clustering analysis provides valuable insights into the data's natural groupings, as seen across three distinct clustering parameter combinations. Each configuration — ward-euclidean with a threshold of 3 (criterion: maxclust), single-cosine with a threshold of 0.3 (criterion: distance), and weighted-cityblock with a threshold of 10 (criterion: distance) — produced unique cluster distributions, highlighting how variations in linkage methods and distance metrics impact cluster formation. Among these, the weighted-cityblock configuration demonstrated the highest accuracy to the data structure, as reflected by its superior cophenetic correlation coefficient, suggesting its highest accuracy of the three chosen pairs.

When it comes to the comparison of Hierarchical Clustering and K-means method, the Silhouette Scores suggest that while both clustering methods provide reasonable separation, Hierarchical Clustering offers a marginally better-defined cluster structure for this dataset. This comparison underscores the effectiveness of hierarchical approaches in capturing nuanced patterns within the World Energy Consumption data.

8 Appendix

The code is provided as an attachment.

9 Attachments

Here are provided the large figures discussed in the study.

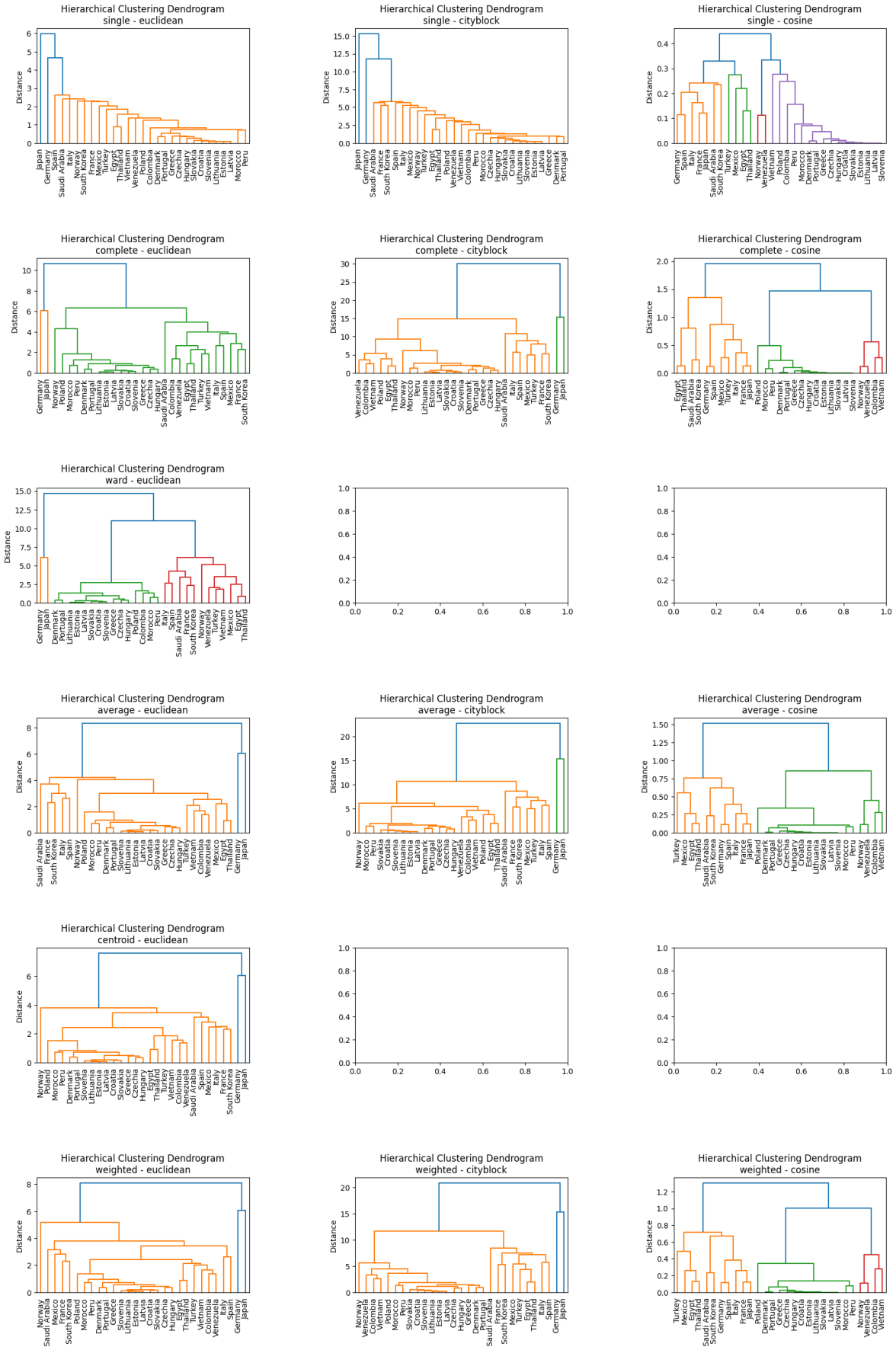

Figure 8: All combinations of dendrograms using different metrics and methods.

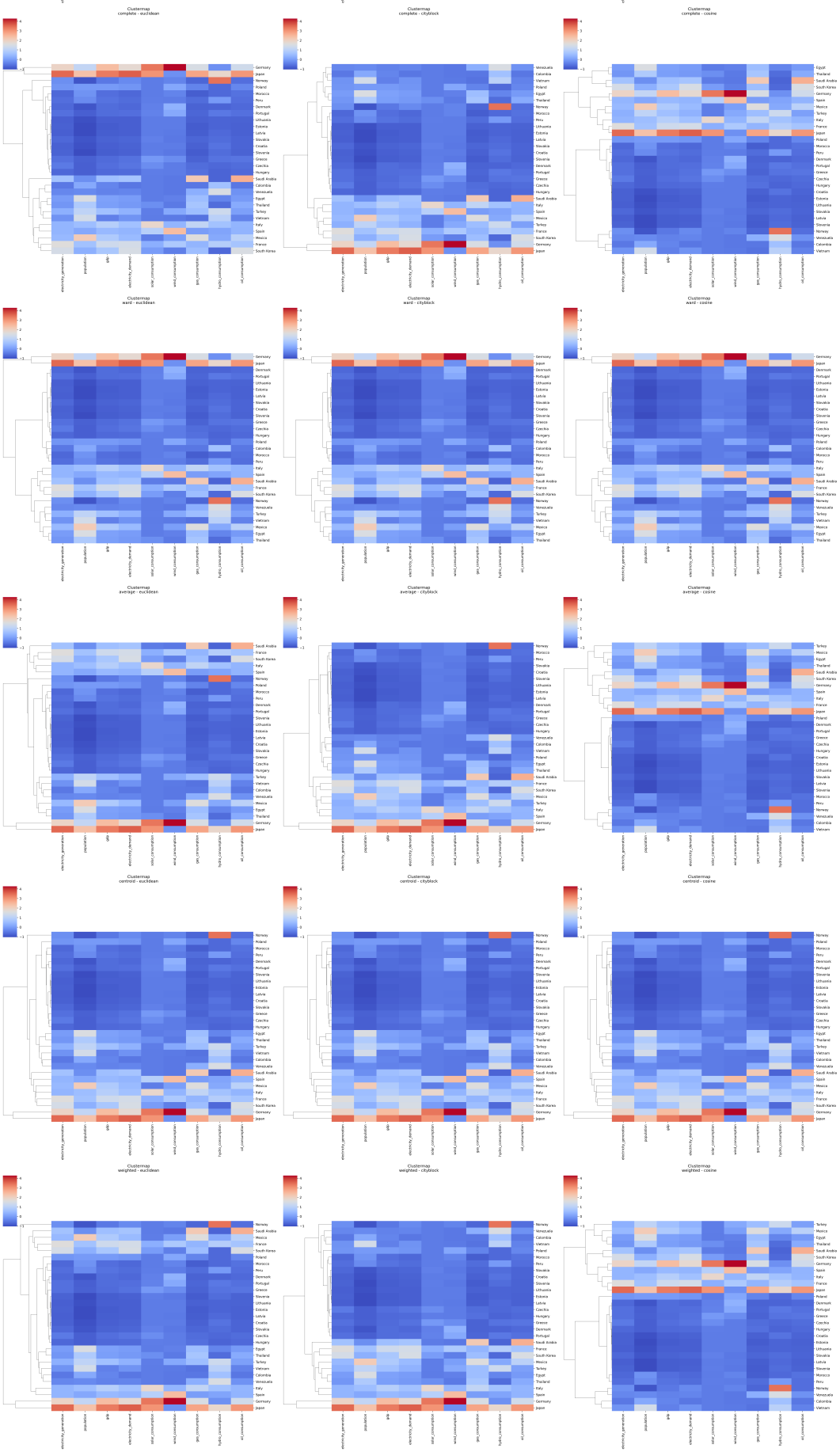

Figure 9: All combinations of cluster heat maps using different metrics and methods.

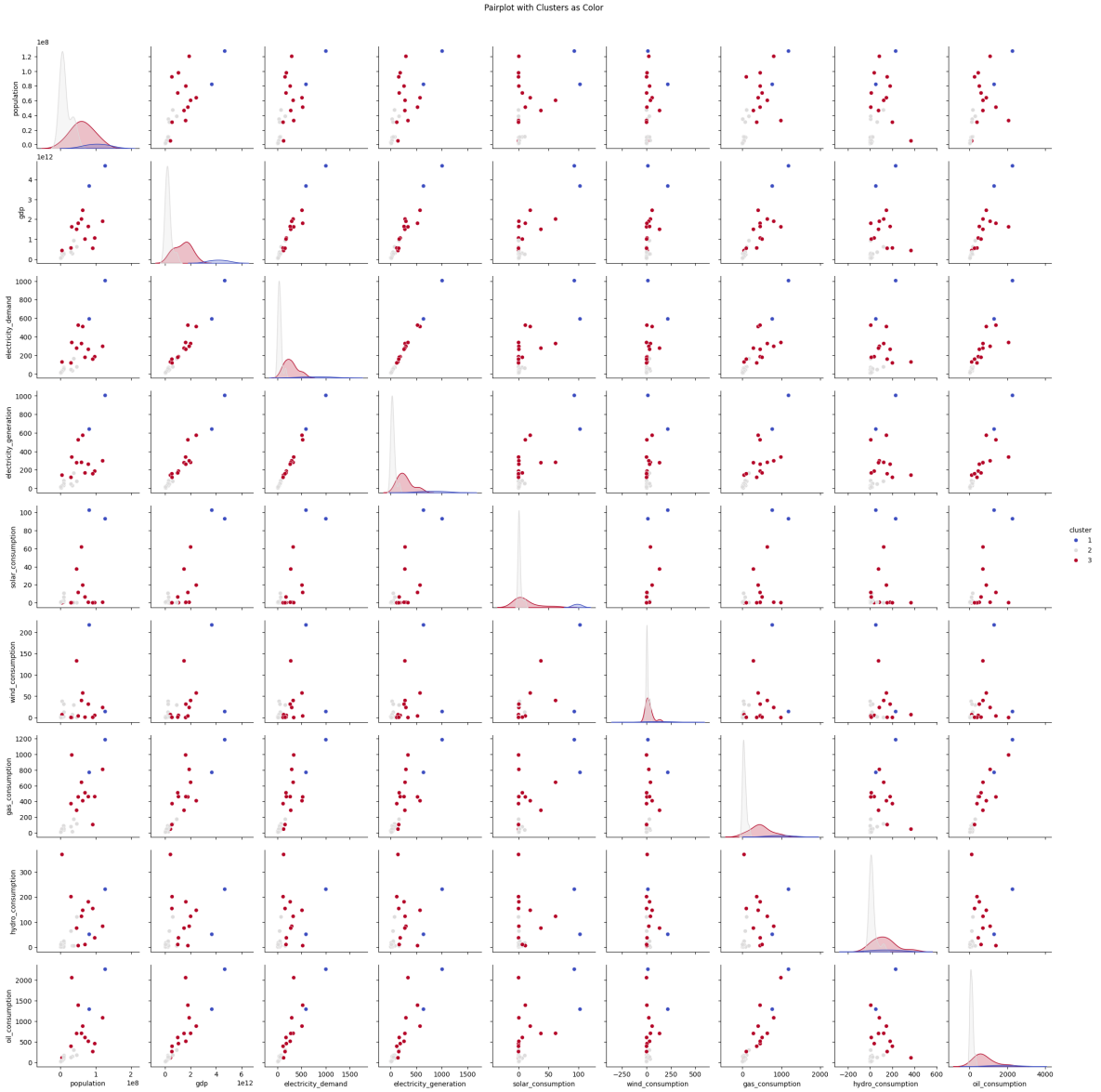

Figure 10: Pair plots for each combination of columns.