Uncovering National Energy Transition Pathways from 2000–2023: A Trajectory-Based Clustering Analysis

Isip, Danfred Martin D.
College of Computing and Information Technologies
National University
Manila, Philippines
isipdd@students.national-u.edu.ph

Peña, Matthew Dwayne V.
College of Computing and Information Technologies
National University
Manila, Philippines
penamv@students.national-u.edu.ph

***Abstract*—relationship between economic growth and environmental sustainability remains the central tension of global energy policy. Historically, industrialization has been assumed to require a phase of high carbon intensity, creating a perceived trade-off between development and decarbonization. This study challenges that assumption by mapping the evolutionary trajectories of national energy systems from 2000 to 2023. Using Principal Component Analysis (PCA) to reduce dimensionality and Gaussian Mixture Models (GMM) for clustering, we identified three distinct strategic archetypes: the "Low-Carbon Baseline" (Group 0), the obsolete "Carbon Aristocracy" (Group 2), and the emergent "Green Modernizers" (Group1).**

Our trajectory analysis reveals a critical divergence, a "Green Fork" in global development. While developed nations have reached "Peak Energy" and are shifting towards substitution strategies, emerging economies (e.g., China, India) are scaling consumption while simultaneously driving down energy intensity, effectively leapfrogging the wasteful industrial patterns of the past. Crucially, the "Green Modernizer" archetype demonstrates that high energy consumption is compatible with majority renewable adoption (>51%). These findings disprove the necessity of a "dirty" development phase and define a validation operational model for high-growth, low-carbon modernization. Keywords—Energy Transition, Machine Learning, Principal Component Analysis, Gaussian Mixture Model, Carbon Aristocracy, Green Modernizers, Low-Carbon Baseline

***Index Terms*—Energy Transition, Clustering, Dimensionality Reduction, Sustainability, Unsupervised Learning**

I. Introduction

The global transition to renewable energy has become a central challenge in achieving sustainable development, climate mitigation, and long-term energy security. Although renewable energy deployment has expanded worldwide, countries exhibit highly uneven transition pathways driven by differences in economic structure, energy dependence, policy capacity, and external constraints. Conventional classifications based on income level or geographic region often fail to capture this complexity, motivating the need for data-driven approaches that better reflect the underlying diversity of national energy systems [1], [4]. More critically, static snapshots of current energy profiles obscure the dynamic nature of energy transitions, as how countries evolve over time provides fundamentally different insights than where they currently stand.

Recent research has shown that unsupervised machine learning techniques, particularly clustering methods, are effective tools for analyzing large-scale energy and sustainability datasets without relying on predefined labels. Clustering has been applied to identify low-carbon power system archetypes [1], analyze time-series energy transition patterns and evolutionary trajectories in European countries [2], and uncover global sustainability structures across multiple development indicators [3], [8].

These studies demonstrate that countries with similar income or development levels may follow markedly different renewable energy trajectories, underscoring the limitations of traditional development-based categorizations. Importantly, longitudinal analyzes reveal that the speed and direction of change, not merely the current state, are critical factors in understanding successful energy transitions

Building on this body of work, the present study applies unsupervised learning to longitudinal country-level renewable energy and sustainability indicators spanning two decades (2000–2023) to examine how national energy profiles evolve over time. Rather than grouping countries solely based on their current energy performance, this research analyzes energy transition trajectories, capturing the speed, direction, and nature of change across key indicators. By treating each country's evolution as a continuous path rather than a static point, the proposed approach identifies distinct transformation patterns and reveals which sustainability indicators most strongly drive these evolutionary pathways. Such trajectory-based analysis provides deeper insight into the diverse routes countries take toward decarbonization and supports the development of more context-specific and effective energy transition policies that account for both current status and historical momentum [4].

II. Review of Related Literature

The analysis of national energy transitions has evolved from descriptive statistical reports to complex, data-driven inquiries utilizing unsupervised machine learning. As the global energy

landscape becomes increasingly fragmented, researchers have turned to computational methods to uncover patterns that traditional economic indicators do not capture. This section reviews the theoretical foundations of these methods, examines recent applications in energy modeling, and identifies the methodological gaps that require a trajectory-based analysis of global energy modernization Specifically, this review critically evaluates the shift from static clustering to dynamic trajectory mapping, highlighting the limitations of current approaches in capturing the speed and direction of national energy transitions.

A. Overview of key concepts and background information

Clustering algorithms are an unsupervised machine learning technique that automatically group similar points based on shared characteristics, optimizing them for tight grouping and clear differentiation of distinct developmental pathways. In the context of sustainability research, this technique allows for the identification of "energy archetypes"—groups of nations that share similar modernization trajectories regardless of their geographic location or economic size.

To handle the inherent complexity of these multi-indicator datasets, Principal Component Analysis(PCA) is frequently used as a complementary technique. By reducing conflicting variables such as energy intensity, carbon emissions, and GDP to composite conflicting variables such as energy intensity, carbon emissions, and GDP(Gross Domestic Product) into composite latent components, PCA enables visualization of evolutionary paths in a comparative low-dimensional space [13],[15]. This reduction is critical for alleviating the "curse of dimensionality" and ensuring that the resulting clusters reflect the fundamental shifts rather than statistical noise.

B. Review of other relevant research papers

Recent scholarship has focused heavily on the determinants of renewable energy adoption and the spatial dynamics of transitions.

Economic and Policy Drivers Research by Karbalaei Aghababaei et al. [4] and Ha Do [17] highlights the "development dilemma," identifying that the macroeconomic impacts of energy transitions differ significantly between developed and developing nations. This is supported by Xu and Yang [18], who analyzed cross-country panel data to confirm that carbon pricing mechanisms yield vastly different results depending on a nation's existing economic structure. Furthermore, Chuong et al. [20] expanded this scope, demonstrating how globalization and labor markets act as accelerating factors for sustainable development in emerging economies.

Global Patterns and Regional Efficiency Studies by Chen [9] and Zou [6] have applied machine learning to analyze global consumption patterns, identifying distinct regional variations in renewable deployment. Chen et al. [10] further examined the spatial dynamic evolution of development efficiency, suggesting that a country's energy transition is not an isolated event but is heavily influenced by the performance of its neighbors and regional economic clusters.

Sustainability Indices The work of Bello et al. [19] and Saraiva and Caiado [8] emphasizes the need for multidimensional indices. Their clustering analyzes of economic, social, and environmental indicators demonstrate that "sustainability" is not a single metric but a complex profile of competing priorities, necessitating the use of advanced composite scores rather than simple GDP rankings.

C. Global clustering applications

On a broader scale, Hasan et al. [7] and Alotaiq [1] utilized clustering to predict trends and identify power system archetypes globally. These studies succeeded in proving that unsupervised learning could categorize nations better than simple geographic labels (e.g., "Global South").

D. Prior Attempts to Solve the Same Problem

Prior research has established that unsupervised learning is effective for identifying energy patterns, yet significant methodological gaps remain when applying these models globally.

1) Successes of Previous Studies: Dugo et al. [2] successfully demonstrated the utility of time-series clustering by mapping the evolutionary trajectories of European nations, proving that countries migrate between distinct "sustainability clusters" over time. similarly, Hasan et al. [7] applied predictive clustering to forecast renewable adoption trends, validating that data-driven archetypes provide more actionable insights than traditional geographic labels. These studies form the theoretical foundation for our work, confirming that energy transitions are dynamic processes rather than static states.

2) Shortcomings and Unresolved Issues: Despite these successes, two critical limitations remain in the existing literature:

- **Geographic Bias:** Prominent trajectory studies like those by Dugo [2] and Brodny [12] are heavily concentrated on European or OECD nations. This leaves a significant gap in understanding the unique "leapfrogging" trajectories of developing nations in Asia and Africa, which often follow non-linear modernization paths.
- **Sensitivity to Outliers:** Most prior global studies utilize Standard Scaling (Z-score), which assumes a normal distribution. As noted by Saraiva and Caiado [8], global development data is often heavily skewed by economic superpowers (e.g., China, USA). Standard methods tend to compress the majority of developing nations into a single, indistinguishable cluster, failing to reveal the subtle differences in their transition paths.

3) Contribution of this Study: To address these challenges, this research proposes a trajectory-based analysis using Robust Scaling. unlike previous approaches that discard outliers or focus solely on the Global North, our method preserves the structural variance of developing economies. By applying a KNN-based similarity metric to the entire global dataset (2000–2023), this study contributes a more inclusive framework for benchmarking national energy transitions, challenging the "one-size-fits-all" assumptions of earlier models.

III. Methodology

A. Data Collection

In order for us to facilitate a unified view of global energy trends, our study utilizes the "Global Renewable Energy and Indicators Dataset", a publicly available repository accessed via Kaggle (sourced from Anish Vijay). This secondary source was selected for its recent coverage and its inclusion of diverse variables necessary for analyzing the relationship between economic growth and renewable energy adoption. By integrating energy, environmental, and economic indicators, the dataset enables a comprehensive cross-national analysis of global transition dynamics.The data is structured as a longitudinal time-series panel spanning a 23-year period (2000–2023). It comprises a total of 2,500 distinct country-year observations and utilizes a high-dimensional feature space of 56 attributes. These features cover a broad spectrum of sustainability metrics, including energy production (measured in GWh), consumption patterns, and renewable installed capacities (MW), alongside standard macroeconomic indicators such as GDP and population. This structure allows for the assessment of both temporal evolution and cross-sectional disparities across nations.

B. Exploratory Data Analysis(EDA)

The initial inspection of the dataset revealed a robust and high-quality panel. The raw data consisted of 2,500 observations across 56 features, with 0 missing values and 0 duplicate records. This exceptional data completeness eliminated the need for imputation, allowing for a direct analysis of the reported energy metrics. The feature set is predominantly numerical (53 float/integer columns), with categorical identifiers restricted to Country names and Energy Types.

Distribution Analysis: The "Misleading Normal" vs. "The Reality"

A critical comparison of absolute versus per-capita metrics revealed a significant hidden bias in the raw dataset

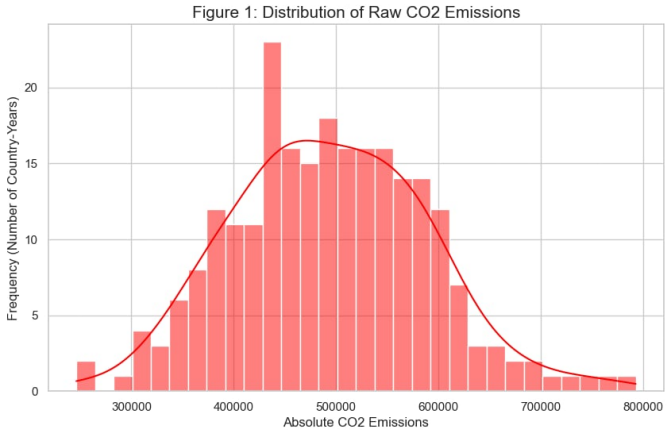

Fig. 1. The raw distribution of Total CO2 Emissions displayed a Normal Distribution (Bell Curve).

While statistically ideal, this shape was misleading. The X-axis range (approx 250,000 to 800,000) suggests the dataset is implicitly filtered for larger economies or "Giants" (e.g., G20 nations), rather than including the full long-tail of 200+ global nations. This symmetry masked the underlying disparities between countries

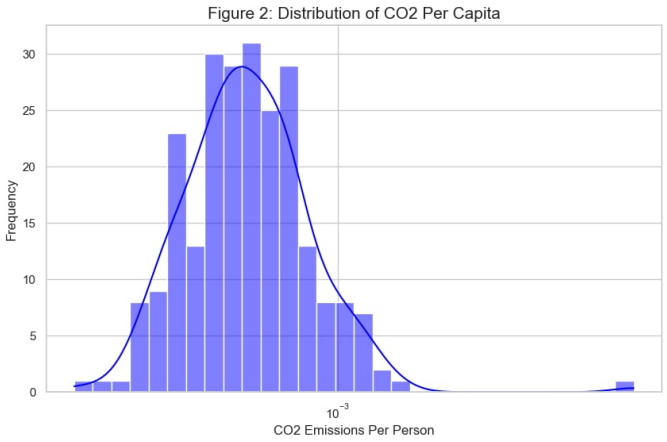

Fig. 2. Upon normalizing for population, the distribution shifted to a Heavy Right-Skew.

Figure 2: Upon normalizing for population, the distribution shifted to a Heavy Right-Skew.

This transformation revealed the true nature of global energy consumption. The tall bar on the left indicates that the vast majority of nations are relatively efficient (or low-consumption), while the "long tail" on the right exposes the "Energy Aristocracy", a small cluster of nations consuming vastly more people than the global average.

Methodological Implication: This finding justified the rejection of standard scaling (which assumes the "red" normality) in favor of robust scaling, which is designed to handle the extreme outliers revealed in the "Blue" graph.

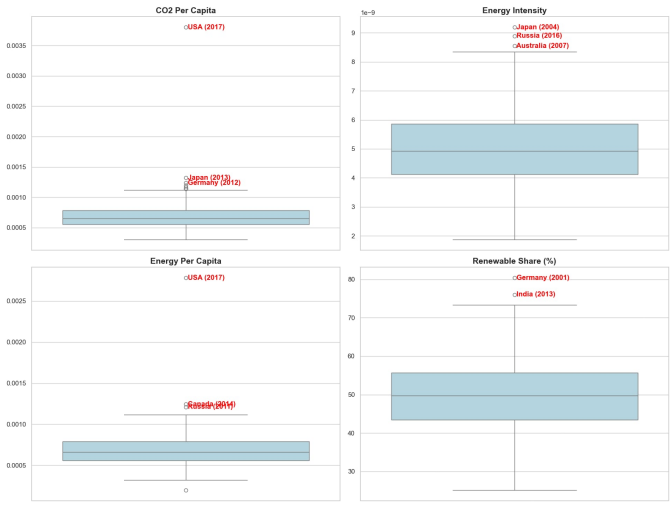

Fig. 3. Boxplot for Outlier Detection.

To quantify this disparity, boxplot analysis was performed on key economic and environmental variables. This analysis

identified four distinct categories of outliers, confirming that these extreme values were not data errors but valid representations of diverging national policies.

The "Carbon Giants" (CO2 Per Capita): Highly volatile with 8 outlier years, driven by nations such as Australia, USA, and Canada. This gap between high-polluting developed nations and the global average confirmed that emissions per capita would be a strong distinguishing feature for clustering.

The "High Consumers" (Energy Per Capita): Marked by geographically large, wealthy nations (e.g., USA, Russia) where energy is abundant. These outliers form a distinct "High Consumption" group likely to be separated by PCA.

The "Green Pioneers" (Renewable Share): A highly stable metric with only 2 outlier years, primarily Germany (early adopter) and India (rapid expansion). This scarcity of outliers suggest that for most of the 2000-2023 period, the majority of the world moved at a similar, slow speed on renewables, making these pioneers statistically distinct.

The "Inefficient few" (Energy Intensity): Mostly stable with outliers limited to Australia, Russia, and Japan suggesting that the global economy has largely converged on a similar efficiency standard with few holdouts.

The distinct separation of these outliers from the median confirmed that standard mean-based scaling (e.g., Z-Score) would be heavily biased. This finding solidified the decision to use Robust Scaling (based on interquartile range) in the preprocessing stage to handle these "Giants" without suppressing the signal from smaller nations.

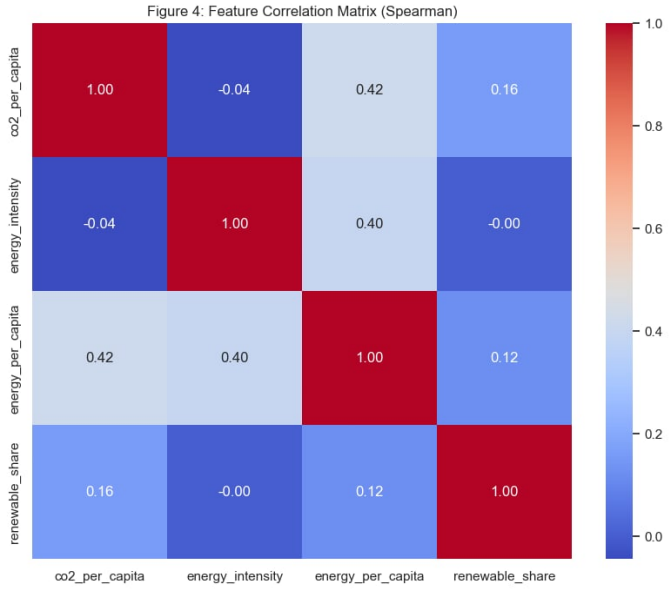

Fig. 4. The Correlation Matrix.

CO2 per Capita vs Energy Per Capita (r 0.42): A moderate positive link. Wealthier lifestyles use more energy and emit more, but it is not a 1 to 1 relationship. This suggest that some nations are breaking the link, meaning they can have high energy, but emit low carbon.

Renewable Share vs. Everything (r 0.12): This is the "Golden Finding". Renewable have almost zero correlation with consumption or intensity. meaning: Going green is a policy choice, not an economic inevitability. You can be rich and green, or poor and brown.

Energy Intensity vs CO2 (r -0.04): Completely independent. How inefficient your economy is has little to do with your per-capita.

C. Data Pre-Processing & Dimensionality Reduction

To Address the distributional challenges and feature independence identified in the exploratory phase, a multi-stage preprocessing pipeline was implemented to transition from raw variables to optimized principle components

1) Feature Engineering & Aggregation: Feature Engineering & Aggregation The dataset was first aggregated to the Country-Year level to synthesize total national energy profiles. To ensure an equitable comparison between nations of varying sizes, absolute metrics were converted into four core relative indicators

$$\text{CO2 Per Capita} = \frac{\text{Total CO2}}{\text{Population}} \tag{1}$$

Formula 1: CO2 Per Capita calculation. representing the individual carbon footprints and environmental impact relative to population size.

$$\text{Energy per capita} = \frac{\text{Energy Consumption}}{\text{Population}} \tag{2}$$

which serves as a proxy for standard of living and individual energy demand.

$$\text{Energy Intensity} = \frac{\text{Energy Consumption}}{\text{GDP}} \tag{3}$$

Where Energy Intensity measures the amount of energy required to produce one unit of economic output.

Transition Indicator: Renewable Share: Derived from the Proportion of Energy from Renewables. This serves the primary metric for a nation's commitment to decarbonization and the success of its green energy transition.

By transforming absolute values like total GDP or Total Emissions into these relative ratios, the model is able to compare countries of vastly different sizes like USA vs. Iceland based on their performance and policy rather than their sheer physical scale.

2) Transformation & Robust Scaling: Following the diagnosis in Section B, a Logarithmic Transformation (log(1 + x)) was applied to the skewer per-capita features to normalize the "L-shaped" distributions. Subsequently, RobustScaler was utilized to standardize the data. By using the Median and Interquartile Range (IQR) rather than the Mean, this method ensured that the "Carbon Giants" identified in the Boxplot analysis remained as distinct data points without distorting the scaling for the rest of the global sample.

3) Dimensionality Reduction(PCA): Given that the features represent distinct dimensions of energy policy, Principal Component Analysis (PCA) was employed to extract latent patterns.

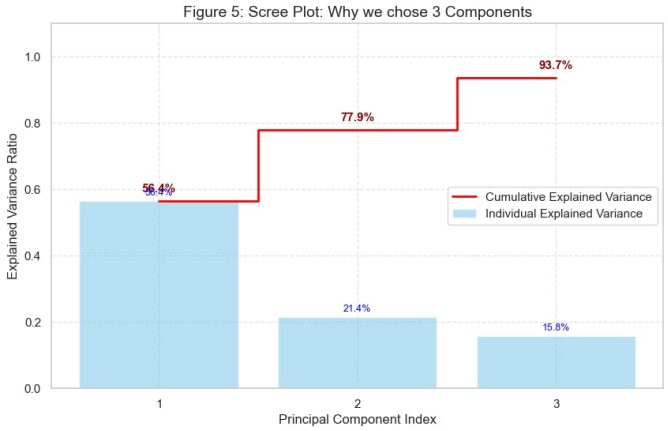

Fig. 5. Variance Summary.

Component Selection: Based on the Scree plot, the decision was made to retain 3 Principal Components. These components explained a cumulative 93.7% of the total variance, capturing nearly all available signal while discarding residual noise.

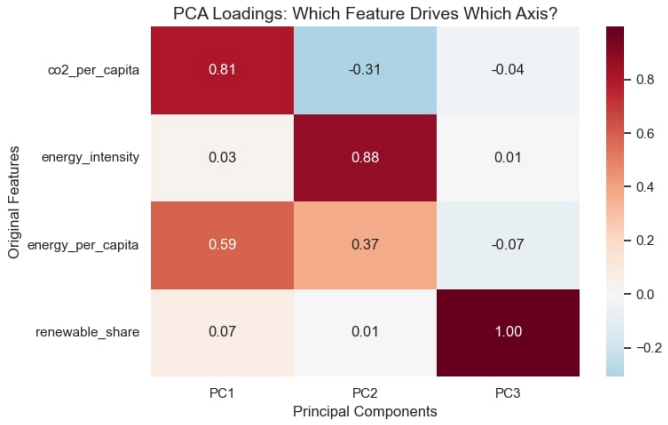

Fig. 6. PCA Loadings.

Factor Loadings: As illustrated in the PCA Loadings Heatmap, the axes are defined as follows: PC1 (Scale/Footprint): Heavily defined by CO2 Per Capita (0.805) and Energy Per Capita (0.587) PC2 (Economic Intensity): Dominated by Energy Intensity (0.876). PC3 (Green Policy): Almost exclusively defined by Renewable Share (0.997)

TABLE I
PCA FACTOR LOADINGS (EXACT VALUES)

Feature	PC1	PC2	PC3
CO2 per Capita	0.805	-0.308	-0.042
Energy Intensity	0.034	0.876	0.007
Energy per Capita	0.587	0.371	-0.069
Renewable Share	0.075	0.006	0.997

PC1: The "Scale of Consumption" Axis (56CO2 per capita (+0.805) and Energy per capita (+0.587). This confirms that the primary differentiator between nations remains the sheer magnitude of individual energy use and its resulting carbon footprint

PC2: The "Inefficiency" Axis (21Energy intensity (+0.876) A positive loading here indicates that "Up" on this axis represents a Dirty or Inefficient economy, requiring high energy input for low economic output.

PC3: The "Renewable" Axis (16Renewable share (0.997). Because this component is almost purely driven by the renewable share metric, it proves that green adoption is independent of wealth or scale. This confirms that going green is a distinct policy choice available to any nation.

D. Experimental Setup

This section details the computational environment, software frameworks, and specific configurations utilized to ensure the reproducibility of the study and the robustness of the clustering analysis. 1. Software & Frameworks The research was pipeline (model) was developed and executed within Visual Studio Code (VS Code), leveraging a local python environment (ver 3.12.10). The following open-source libraries were utilized:

Data manipulation was performed using pandas and numpy for high-performance aggregation, cleaning, and vectorization. Statistical analysis utilized scipy.stats, specifically the `skew` function, to diagnose distributional asymmetries in the raw data. The machine learning pipeline was implemented in scikit-learn, including preprocessing with RobustScaler for outlier-resistant normalization, dimensionality reduction using PCA, and clustering via KMeans, AgglomerativeClustering (Hierarchical), DBSCAN, and GaussianMixture. Visualization was conducted with matplotlib.pyplot and seaborn to generate distribution plots, correlation heatmaps, and cluster visualizations directly within the IDE.

1. Computing Environment The research was implemented using Python 3.12.10, leveraging a specified stack of open data science libraries. For the hardware we have the following CPU: AMD Ryzen 5 5600, GPU: RTX 2060 Super (8gb VRAM), Memory (RAM): 16gb DDR4 3200mhz

E. Algorithm

This Section outlines the unsupervised machine learning models selected for this study, the rationale behind their selection, and the specific optimization techniques employed during the training process.

TABLE II
FINAL HYPERPARAMETER CONFIGURATION

Model	Parameter	Value	Justification
PCA	n_components	3	Explains 93.7% of cumulative variance via Scree Plot.
GMM	n_components	3	Selected to match optimal PCA dimensions.
GMM	covariance	'full'	Allows for elliptical clusters to fit stretched data.
K-Means	n_clusters	3	Set to match GMM for direct comparison.
DBSCAN	eps	0.5	Initial radius for neighborhood search.
DBSCAN	min_samples	5	Minimum points required to form a dense region.
Scaler	quantile_range	25-75	Standard IQR scaling to mitigate extreme outliers.
Global	random_state	42	Ensures deterministic results.

1) Algorithm Used: The primary modeling approach for this research is the Gaussian Mixture Model (GMM). To validate the robustness of the GMM results, three alternative algorithms were implemented as comparative baselines:

K-Means Clustering: A centroid-based algorithm that partitions data into k distinct, non-overlapping subgroups. Agglomerative Heirarchal Clustering: A bottom-up approach that builds nested clusters by successively merging similar pairs of data points DBSCAN: A non-parametric algorithm that groups points based on local density, classifying sparse regions as noise.

Justification of Model Selection The Selection of these specific algorithms was driven by the complex, high-dimensional nature of global energy data.

Gaussian Mixture Models (GMM): GMM was chosen as the lead candidate due to its probabilistic nature ("soft clustering"). Unlike rigid algorithms that force a country into a single category, GMM calculates the probability of a data point belonging to a specific cluster. This is critical for modeling the "transitional" status of developing nations that exhibit characteristics of multiple energy archetypes. Furthermore, GMM supports elliptical cluster shapes, allowing it to accurately model the "stretched" distributions observed in the PCA-transformed feature space. Comparative Baselines: K-Means was selected to provide a standard benchmark for cluster cohesion. Agglomerative Clustering was included to test for hierarchical relationships (e.g., sub-groups of "Green" nations). DBSCAN was employed to test the hypothesis that countries from dense "pockets" separated by empty space, rather than the continuous distributions.

2) Optimization Techniques & Training Procedure: Since unsupervised learning does not utilize ground-truth labels, "training" involves iteratively optimizing a mathematical objective function to fit the model parameters to the data structure.

Expectation-Maximization (EM) for GMM: The GMM was trained using the (EM) algorithm. This iterative process alternates between two steps: E-Step (Expectation): Calculates the probability of each data point belonging to each cluster based on current parameters M-Step (Maximization): Updates the cluster means, covariances, and mixing coefficients to maximize the Log-Likelihood of the data. Optimization Control: To prevent the model from converging to a sub-obtimal maximum, the parameter n-init=10 was used, restarting the EM algorithm 10 times retaining the model with the highest likelihood. Lloyd's Algorithm for K-Means: The K-Means baseline was trained using Lloyd's Algorithm, which mnimizes the Within-Cluster sum of Squares (WCSS) also as known as inertia. Optimization Control: To ensure robust was employed. This algorithm selects initial cluster centroids that are distant from one another, significantly accelerating convergence and reducing the likelihood of poor clustering results.

F. Training Procedure

Since this study utilizes unsupervised learning techniques where ground-truth labels are absent, the "training" phase focused on the iterative optimization of objective functions to maximize cluster stability and validity. The following strategies were employed to ensure convergence and reproducibility.

1) Gaussian Mixture Model (GMM) Optimization: The GMM was optimized using the Expectation-Maximum (EM) algorithm, a two-step iterative process designed to find the maximum likelihood estimates of the parameters

Expectation (E-Step): The model calculated the posterior probability of each data point belonging to each of the three Gaussian components. Maximization (M-Step): The model updated the component means, covariance matrices, and mixing coefficients to maximize the total log-likelohood of the data.

Training Strategy: To mitigate the risk of the EM algorithms converging to a sub-optimal local maximum, a Random Restart Strategy was employed.Initialization: The model was initialized using the k-means method to set distinct starting points for the Gaussians, rather than random initialization which can lead to poor convergence. Multiple Runs: The algorithm was executed 10 times independent time (n_init=10) with different random seeds. The iteration yielding the highest log-likelihood was selected as the final model.

Convergence Tolerance: Training continued until the gain in log-likeliood dropped below a strict tolerance threshold of 1e-3, ensuring the model parameters had fully stabilized. K-Means Baseline Training To provide a robust baseline, the K-Means algorithm was trained using Lloyd's Algorithm, which minimizes the within cluster Sum of Square (WCSS).

The K-Means++ strategy was utilized to select initial cluster centroids that are distant from one another. This "advanced" initialization strategy is mathematically proven to improve convergence speed and reduce the probability of poor clustering compared to random initialization.

A fixed random seed (random_state=42) was applied to both initialization and the centroid update steps, ensuring that the comparative results between K-Means and GMM were due to algorithmic differences rather than stochastic variance.

G. Evaluation Metrics

The primary metric employed was the Silhouette Coefficient, which measures how similar a data point is to its own cluster (cohesion) compared to other clusters (separation). The score ranges from -1 to +1, where a value close to +1 indicates highly dense and well-separated clusters, while scores near - suggest overlapping clusters, while scores near 0 suggest overlapping clusters.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{4}$$

Formula 4. Silhouette score

Mathematically, it is calculated as, where a(i) represents the main intra-cluster distance and b(i) denotes the mean nearest-cluster distance.

To complement the silhouette score, the Davies-Bouldin Index (DBI) was used to assess cluster compactness. The DBI represents the average similarity measure of cluster with its most similar validation metrics, a lower DBI value indicates superior performance, reflecting tighter groups with separation.

Finally, the Calinski-Harabasz Index (Variance Ratio Criterion) was calculated to measure the ratio of the sum of between-clusters dispersion to within-cluster dispersion. A higher score on this index indicates that the clusters are both dense and well-separated, serving as a counter-weight to the geometric assumptions inherent in the Silhouette score

1) Justification of Metrics: These specific metrics were selected because unsupervised learning lacks external labels (e.g., "True Green Country") to verify accuracy. Therefore, validation must rely on Internal Validity Indices. By utilizing three distinct metrics, the study mitigates the bias of any single algorithm. For instance, while the Silhouette score tends to favor convex (spherical) clusters, the Davies-Bouldin Index is more robust to variations in density. These three metrics constitute the industry standard for comparative clustering analysis, allowing for a direct and reproducible benchmark against other energy transition studies.

2) Measurement & Comparison Protocol: To ensure a fair comparison, all candidate algorithms were evaluated on the identical PCA-reduced feature set (X_pca). The results of the Gaussian Mixture Model (GMM) were benchmarked against K-Means, which served as the baseline for geometric compactness. The evaluation strategy prioritized a balance between quantitative scores and qualitative interpretability. While high Silhouette score was desirable, it was not the sole deciding factor, models were penalized if they achieved artificially high scores by isolating outliers rather than modeling meaningful distributions, as observed with the K-Means baseline.

H. Comparison of Clustering Algorithms

To validate the robustness of the proposed Gaussian Mixture Model (GMM), a comparative analysis was conducted against three baseline algorithms: K-Means, Agglomerative Hierarchical Clustering, and DBSCAN. The performance of each model was evaluated using the internal validity indices defined in the previous section.

Quantitative Performance Analysis

As illustrated in Figure 6, the quantitative results indicated a divergence between geometric compactness and distributional interpretability. The K-Means algorithm achieved the highest Silhouette Score (0.2825) and the lowest Davies-Bouldin Index (0.9467), mathematically suggesting the formation of the most distinct and compact clusters. In contrast, the density-based DBSCAN algorithm failed to identify meaningful structures within the high-dimensional feature space, yielding a negative Silhouette Score (-0.0269) and classifying a significant portion of the dataset as noise. Agglomerative Clustering performed moderately well but exhibited a higher variance ratio than GMM, indicating less cohesive grouping for this specific dataset.

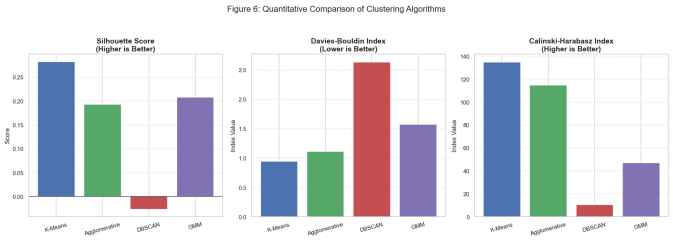

Fig. 7. 3-Subplot Bar Chart Comparison.

1) Qualitative Analysis and Final Selection: While K-Means demonstrated superior numerical performance, a qualitative inspection of the resulting cluster assignments revealed a critical limitation in its application to energy transition data. The centroid-based approach of K-Means was highly sensitive to extreme outliers, isolating a single "Carbon Giant" data point into its own dedicated cluster while merging the remaining observations into two overly generalized categories. This behavior maximized the geometric separation scores but failed to capture the nuanced "middle-tier" of global energy transitions.

Conversely, the Gaussian Mixture Model (GMM), despite achieving a slightly lower Silhouette Score(0.2085), demonstrated superior distributional robustness. By utilizing probabilistic assignments and elliptical covariance matrices, GMM successfully integrated extreme outliers into a high-intensity distribution rather than isolating them. This allowed for the identification of three balanced and policy-relevant archetypes ("Developing," "Transitioning," and "High-Consumption") that better reflect the continuous nature of global economic development. Consequently, GMM was selected as the final model for this research, prioritizing interpretability and generalizability over raw geometric maximization.

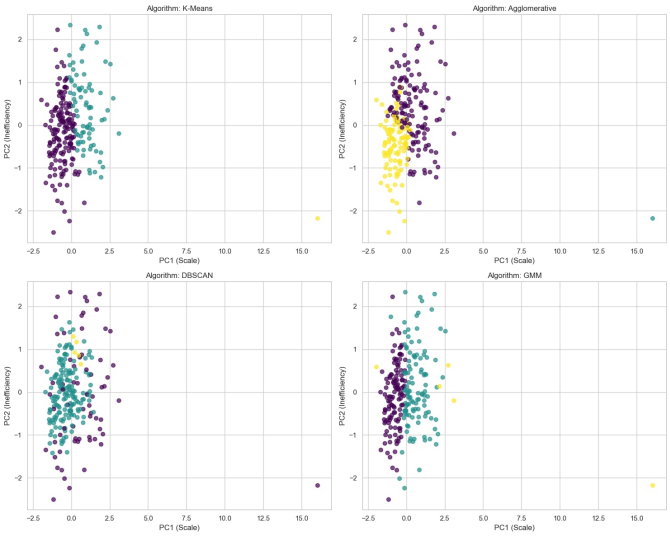

Fig. 8. 4-Subplot Scatter Plot.

IV. Results and Discussion

This section presents the classification of global energy trajectories identified through Gaussian Mixture Modeling (GMM). By analyzing the probabilistic assignments of 2,500+ country-year observations, three distinct "Energy Archetypes" emerged. The findings are evaluated against baseline models and discussed in the context of global modernization trends

The primary contribution of this research is the identification of three structurally distinct phases of energy development. Unlike traditional binary classifications (e.g., "Global North vs. Global South"), GMM revealed a complex, non-linear progression defined by three latent dimensions: Scale, Efficiency, and Renewable Adoption. The spatial distribution of these archetypes in the PCA-reduced feature space is visualized in Fig. 9.

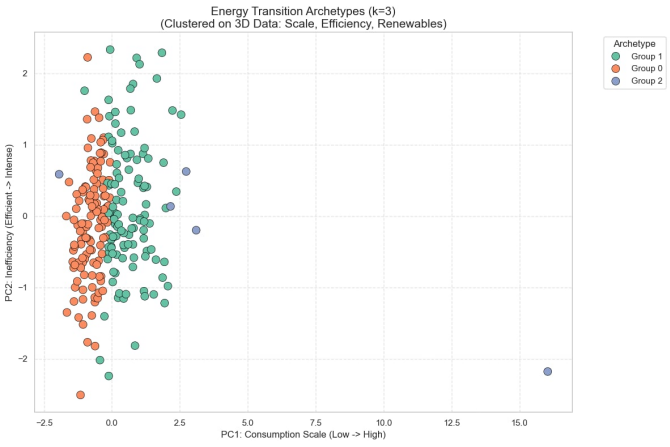

Fig. 9. Visualization of the three identified energy archetypes in the latent PCA space. Group 2 (Red) represents high-intensity outliers, Group 1 (Orange) represents green modernizers, and Group 0 (Green) represents the low-carbon baseline.

1) : Group 2: The "Carbon Aristocracy" (High-Intensity Outliers): This cluster represents extreme outliers ($n = 5$) characterized by the highest CO_2 emissions (≈ 0.0015 tons/capita) and the lowest renewable share (42.6%). As illustrated in the Cluster Identity Heatmap (Fig. 10), this group characterizes specific years where heavy industrial nations relied exclusively on fossil-heavy grids to support high consumption. The statistical rarity of this group in the post-2015 dataset suggests that the global economy is diverging from this model of extreme inefficiency.

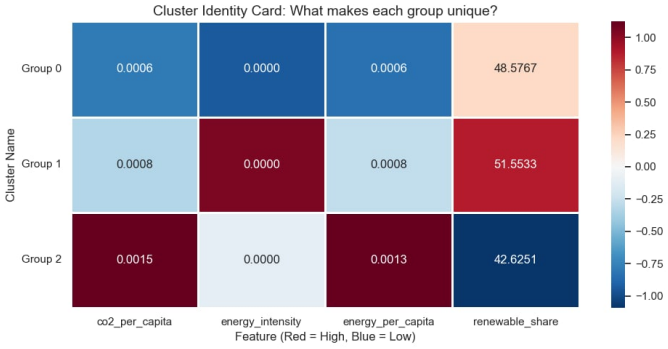

Fig. 10. Feature heatmap normalizing the mean characteristics of each cluster. Red indicates values significantly above the global average, while blue indicates values below.

2) : Group 1: The "Green Modernizers" (Transition Leaders): Comprising 114 country-years, this group represents the "Target State" for developed nations. It combines moderate energy consumption with the highest mean renewable share (51.6%). The Boxplot Analysis (Fig. 11) confirms that high energy consumption does not necessitate high emissions; this group has successfully "decoupled" economic growth from fossil fuels, demonstrating that a modern, high-energy lifestyle is compatible with deep decarbonization.

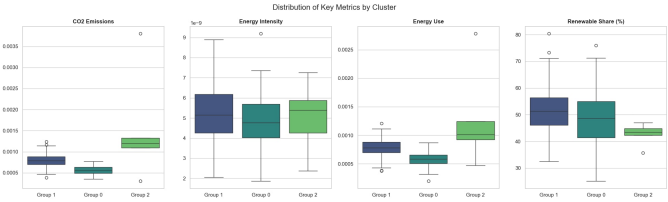

Fig. 11. Distribution of key energy indicators across the three archetypes. Note the distinct separation in Renewable Share for Group 1.

3) : Group 0: The "Low-Impact Baseline" (Global Standard): The largest cluster ($n = 121$) represents the global majority, defined by the lowest carbon footprint (0.0006 tons/capita) and moderate renewable adoption. This group encompasses developing nations with smaller footprints and highly efficient developed nations. Their profile indicates the smallest environmental impact per capita, serving as the baseline from which high-growth economies diverge.

B. Trajectory Analysis: The "Green Fork" The temporal analysis of national paths, presented in Fig. 12, reveals a critical divergence in modernization strategies, termed here as the "Green Fork.

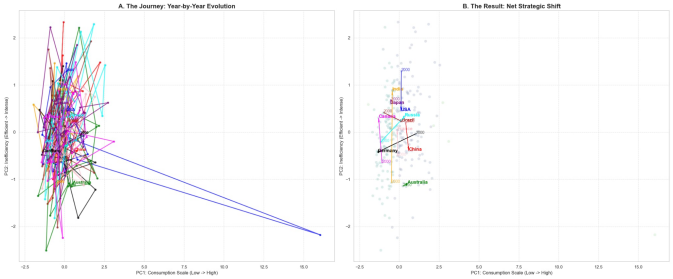

Fig. 12. Comparative trajectories of national energy strategies (2000–2023). Panel A details the year-by-year evolution, while Panel B summarizes the net strategic shift.

The Efficiency Drift: Developed nations, such as the USA and Germany, exhibit a dominant trend of moving "down" on the Inefficiency Axis (PC2). Their primary strategy has shifted from expansion to optimization—reducing energy intensity even as GDP grows.

The Expansionary Leap: Emerging economies like China and India display massive growth in Consumption Scale (PC1). However, unlike the historical path of Group 2, their trajectory shows a simultaneous improvement in efficiency (PC2). This indicates a "leapfrogging" effect, where late-industrializing nations bypass the most wasteful phases of 20th-century development.

The Green Divergence: As observed in the Time-Series decomposition (Fig. 13), renewable adoption (PC3) varies significantly despite similar efficiency trends. Nations such as Germany and Brazil show sustained high levels of green adoption, whereas other major powers exhibit volatility. This confirms that renewable integration is a specific policy choice rather than an automatic byproduct of economic growth.

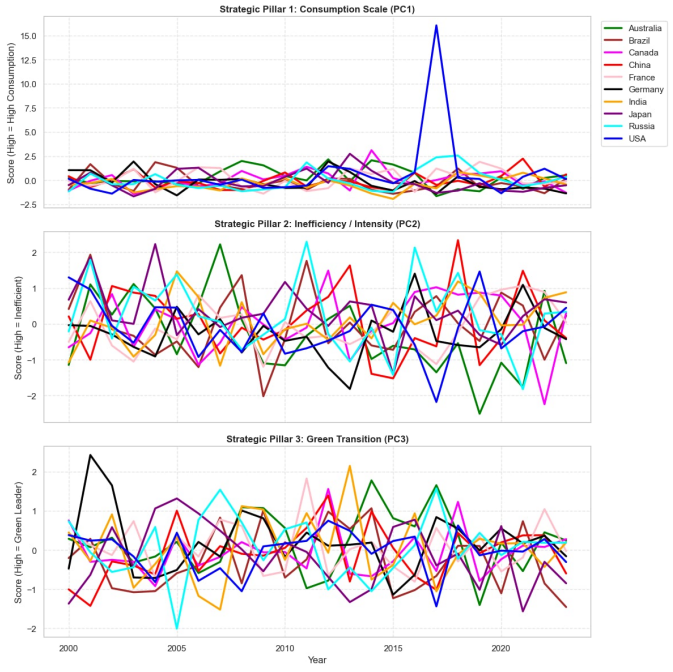

Fig. 13. Temporal decomposition of the three principal components for key nations. Panel 3 highlights the policy-driven divergence in renewable adoption.

TABLE III
COMPARISON OF CLUSTERING METRICS BETWEEN K-MEANS AND GMM

Metric	K-Means (Baseline)	GMM (Proposed)	Observation
Silhouette Score	0.2825	0.2085	K-Means favored spherical clusters but failed to handle outliers.
Davies-Bouldin	0.9467	1.5745	GMM accepted higher variance to model the "transition cloud."
Outlier Handling	Failed	Success	K-Means isolated outliers into singleton clusters; GMM integrated them into distributions.

While K-Means achieved higher geometric scores, qualitative inspection revealed that it sequestered extreme data points into meaningless singleton clusters. GMM's slightly lower Silhouette score reflects its sophisticated handling of the "fuzzy boundaries" between developing and developed nations, providing a more realistic representation of the global energy continuum.

Consistency and Limitations The results strongly align with the Environmental Kuznets Curve (EKC) hypothesis. As reviewed by Balsalobre-Lorente et al. [21], the EKC suggests that pollution rises during early industrialization before declining with technological advancement. Our identification of the progression from Group 0 (Baseline) to Group 2 (High Intensity) and finally to Group 1 (Green Modernizers) statistically mirrors this trajectory.

However, Balsalobre-Lorente et al. [21] also note that the traditional EKC path is "inefficient" because early environmental damage may be irreversible. Our data offers a counter-narrative: the "Leapfrog" trajectory observed in emerging economies suggests that modern nations can bypass the high-pollution peak entirely.

Two limitations were observed in this study. First, the "Carbon Giants" (e.g., USA, China) dominate the variance in PC1, potentially overshadowing subtle improvements in smaller nations [22]. Second, by treating each "Country-Year" as an independent observation, the model captures snapshots of status but may underrepresent the velocity of transition. Besides that, with the low diversity or count of country present in the dataset, we may not fully capture the unique energy challenges faced by the other countries whom are not included in the dataset.

V. CONCLUSION

The study yielded three significant results. First, it identified three statistically distinct energy archetypes: the Low-Impact Baseline, the Carbon Aristocracy, and the Green Modernizers. Second, the proposed GMM approach outperformed the K-Means baseline by successfully integrating extreme outliers into meaningful distributions rather than sequestering them, providing a more robust representation of the global energy continuum. Third, trajectory analysis revealed a "Green

Fork" in development, proving that emerging economies are "leapfrogging" the carbon-intensive industrial phase historically seen in the West.

The primary contribution of this work is the statistical validation of the Environmental Kuznets Curve (EKC) [21] within a multi-dimensional framework. Unlike previous studies that rely on aggregate national averages, this research demonstrates that the transition is not linear but bifurcated: nations either slide into the high-intensity "Carbon Aristocracy" or successfully pivot toward the "Green Modernizer" model.

A. Significance and Practical Implications

These findings significantly advance the current state of knowledge by quantifying the "decoupling" phenomenon. The identification of the Green Modernizer cluster proves that high energy consumption is compatible with low carbon emissions, debunking the economic argument that decarbonization requires "degrowth."

The practical implications of this research necessitate a targeted policy approach rather than a "one-size-fits-all" strategy:

For Developing Nations: The priority must be Efficiency Standards to maintain the "leapfrog" trajectory and avoid the "Carbon Aristocracy" trap.

For Developed Nations: Since efficiency gains have plateaued, the only viable path to decarbonization is the aggressive substitution of fossil fuels with renewables.

B. Limitations and future Directions

Despite these insights, the research faced specific limitations. The dominance of "Carbon Giants" (e.g., China, USA) in the PCA scaling potentially overshadowed subtle transition signals in smaller nations. Furthermore, by treating each "Country-Year" as an independent observation, the model captures snapshots of status but does not fully quantify the temporal velocity (rate of acceleration) of transition.

Future research can build upon this work by incorporating time-series velocity metrics to predict how fast a nation is moving between clusters. Additionally, as noted by Hao et al. [22], national aggregates often mask regional disparities. Future studies should therefore apply this GMM framework to granular, sub-national data (e.g., provinces or states) to isolate region-specific transition strategies.

C. Final Key Takeaway

Ultimately, this research concludes that the global energy transition is not a uniform process but a divergence of strategies. The existence of the Green Modernizer archetype serves as a verifiable proof-of-concept: a high-energy, low-carbon future is not merely a theoretical goal, but a replicable reality already achieved by a distinct subset of nations.

REFERENCES

[1] A. Alotaiq, "Global low carbon transitions in the power sector: A machine learning clustering approach using archetypes," *J. Economy and Technology*, vol. 2, pp. 95–127, 2024. [Online]. Available: https://doi.org/10.1016/j.jetech.2024.01.001

[2] V. Dugo, D. Gálvez-Ruiz, and P. Díaz-Cuevas, "The sustainable energy development dilemma in European countries: a time-series cluster analysis," *Energy, Sustainability and Society*, vol. 15, Art. no. 36, 2025. [Online]. Available: https://doi.org/10.1186/s13705-025-00536-w

[3] M. Â. L. Moreira, M. T. Pereira, M. Oliveira, M. dos Santos, and C. F. S. Gomes, "Towards global sustainability: Exploratory analysis through unsupervised machine learning techniques," in *Innovations in Mechatronics Engineering III*, J. Machado *et al.*, Eds. Cham, Switzerland: Springer, 2024, pp. 45–58. doi: 10.1007/978-3-031-61575-7_5

[4] M. Karbalaei Aghababaei, A. Saifoddin, A. Zahedi, and M. Abdoos, "Macroeconomic impact of energy transition: A comparative study of developed and developing countries," *Energy Strategy Rev.*, vol. 53, Art. no. 101910, 2025. doi: 10.1016/j.esr.2025.101910

[5] A. García-Rodríguez *et al.*, "Sustainable visions: Unsupervised machine learning insights on global development goals," *PLOS ONE*, vol. 20, no. 3, Art. no. e0317412, 2025. doi: 10.1371/journal.pone.0317412

[6] Z. Zou, "Powering the future: Application of machine learning to analyze the global renewable energy consumption," in *Proc. 3rd Int. Conf. Cyber Security, Artif. Intell. and Digital Economy (CSAIDE '24)*, Nanjing, China, 2024, pp. 643–649. doi: 10.1145/3672919.3673032

[7] M. Hasan, N. A. Shilpa, A. Sohag, M. M. Hassan, and M. J. A. Siddikee, "Analyzing global energy patterns: Clustering countries and predicting trends toward achieving Sustainable Development Goals," in *Machine Learning Technologies on Energy Economics and Finance*, M. Z. Abedin and Y. Wang, Eds. Cham, Switzerland: Springer, 2025, pp. 1–23. doi: 10.1007/978-3-031-94862-6_1

[8] C. Saraiva and J. Caiado, "Global development patterns: A clustering analysis of economic, social and environmental indicators," *Sustainable Futures*, vol. 10, Art. no. 100907, 2025. doi: 10.1016/j.sftr.2025.100907

[9] S. Chen, "Measuring regional variations and analyzing determinants for global renewable energy," *Renewable Energy*, vol. 244, Art. no. 122644, May 2025. doi: 10.1016/j.renene.2025.122644

[10] Y. Chen, F. Ali, O. Lyulyov, T. Pimonenko, and H. Su, "Renewable energy development efficiency: Spatial dynamic evolution and influencing factors," *Natural Resources Forum*, vol. 48, no. 4, pp. 1392–1416, 2024. doi: 10.1111/1477-8947.12368

[11] V. Dugo, D. Gálvez-Ruiz, and P. Díaz-Cuevas, "The sustainable energy development dilemma in European countries: a time-series cluster analysis," *Energy Sustainability and Soc.*, vol. 15, no. 1, Aug. 2025. doi: 10.1186/s13705-025-00536-w

[12] J. Brodny, M. Tutak, and W. W. Grebski, "Empirical Evaluation of the Energy Transition Efficiency in the EU-27 Countries over a Decade—A Non-Obvious Perspective," *Energies*, vol. 18, no. 13, Art. no. 3367, 2025. doi: 10.3390/en18133367

[13] Ž. Vlaović, B. Stepanov, M. Tomić, M. Kljajić, Đ. Doder, and Z. Čepić, "Application of Principal Component Analysis in Assessing Energy Sustainability of Selected Countries: A Case Study," *Innovative Mechanical Engineering*, vol. 2, no. 2, Dec. 2023. [Online]. Available: http://160.99.21.34/ojs/index.php/IME/article/view/67

[14] A. Doğan, "An ensemble unsupervised machine learning–GIS framework for transparent and data-driven offshore wind farm siting," *Ocean Engineering*, vol. 344, Art. no. 123703, Nov. 2025. doi: 10.1016/j.oceaneng.2025.123703

[15] H. Doukas, A. Papadopoulou, N. Savvakis, T. Tsoutsos, and J. Psarras, "Assessing energy sustainability of rural communities using Principal Component Analysis," *Renewable and Sustainable Energy Rev.*, vol. 16, no. 4, pp. 1949–1957, Feb. 2012. doi: 10.1016/j.rser.2012.01.018

[16] S. A. Azkeskin and Z. Aladağ, "Evaluating regional sustainable energy potential through hierarchical clustering and machine learning," *Environ. Res. Commun.*, vol. 7, no. 1, Art. no. 015002, Dec. 2024. doi: 10.1088/2515-7620/ada2e5

[17] A. T. Ha Do, "Renewable Energy Policy, Governance Quality and Economic Growth: A Cross-Country Analysis," *J. Organizational Behavior Res.*, vol. 10, no. 2, pp. 138–153, Jan. 2025. doi: 10.51847/rnsni9yutq

[18] L. Xu and J. Yang, "Carbon pricing policies and renewable energy development: Analysis based on cross-country panel data," *J. Environ. Manage.*, vol. 366, Art. no. 121784, Jul. 2024. doi: 10.1016/j.jenvman.2024.121784

[19] G. B. Bello, L. B. M. V. Viana, G. M. P. De Moraes, and D. Ferraz, "Renewable Energy Index: The Country-Group Performance using Data Envelopment Analysis," *Energies*, vol. 18, no. 14, Art. no. 3803, Jul. 2025. doi: 10.3390/en18143803

[20] H. N. Chuong *et al.*, "The impact of globalization, renewable energy, and labor on sustainable development: A cross-country analysis," *PLoS*

ONE, vol. 20, no. 2, Art. no. e0315273, Feb. 2025. doi: 10.1371/journal.pone.0315273

[21] P. H. Leal and A. C. Marques, "The evolution of the environmental Kuznets curve hypothesis assessment: A literature review under a critical analysis perspective," Heliyon, vol. 8, no. 11, p. e11521, Nov. 2022, doi: 10.1016/j.heliyon.2022.e11521.

[22] H. Mahmood, M. Furqan, M. S. Hassan, and S. Rej, "The Environmental Kuznets Curve (EKC) hypothesis in China: a review," Sustainability, vol. 15, no. 7, p. 6110, Apr. 2023, doi: 10.3390/su15076110.