Scientific Research Review

RESEARCH Open Access

Uncovering Global Emission Regimes Using Unsupervised Learning: A Structural and Dynamic Analysis

Rabiah Eladwiah Abdul Rahim[1,*]

** Correspondence: College of Business, Universiti Tenaga National (UniTen), Jalan Ikram, Kuala Lumpur, Malaysia; Rabiah@uniten.edu.my*

1 College of Business, Universiti Tenaga National (UniTen), Jalan Ikram, Kuala Lumpur, Malaysia; Rabiah@uniten.edu.my

Abstract: Understanding the structural composition of greenhouse gas (GHG) emissions is critical for effective climate mitigation, yet existing research largely focuses on aggregate levels rather than the multivariate structure of sectoral emissions. This study develops a data-driven framework to identify emission regimes using a global panel dataset for 172 countries over 1990–2023. Sectoral emissions are transformed into compositional vectors to capture relative contributions, followed by Principal Component Analysis (PCA) and K-means clustering to uncover latent structural regimes. The first two principal components explain 37.9% of total variance and reveal clear gradients distinguishing agriculture-dominated vs. industrial-energy systems and fossil extraction vs. urban-consumption emissions. Five distinct regimes are identified based on silhouette optimization ($\bar{s}$ = 0.38). A first-order Markov transition analysis shows strong persistence, with diagonal probabilities ranging from 0.876 to 0.973, indicating high path dependence and limited short-term mobility across regimes. However, gradual transitions toward mixed emission systems suggest slow structural diversification. Robustness checks confirm the stability of the identified regimes, with repeated K-means runs producing identical cluster assignments (ARI = 1.00), substantial agreement with hierarchical clustering (ARI = 0.57), a Calinski–Harabasz Index of 2161.84, and a Davies–Bouldin Index of 0.88. Overall, the findings show that global emission systems aren't all the same—they're built differently across regions, tend to stick to their patterns, and only change slowly over time. That really matters because it means we can't just look at the big-picture trends. Instead, we need to pay attention to these distinct regimes if we want to truly understand how emissions evolve in the long run.

Keywords: Extreme rainfall prediction, Soil moisture, NASA POWER data, Bangladesh climate modeling

Check for updates

Citation: Rahim, R. E. A. (2026). Uncovering Global Emission Regimes Using Unsupervised Learning: A Structural and Dynamic Analysis. *Scientific Research Review*, 1(1), 32-46. https://doi.org/10.66348/srr.26.a75

Received: 2026-05-07
Revised: 2026-08-12
Accepted: 2026-08-30
Online Available: 2026-09-30
Published: 2026-09-30

Copyright: © 2026 by the authors. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Global greenhouse gas (GHG) emissions continue to be a leading cause of anthropogenic climate change, creating substantial challenges for the implementation of international mitigation measures such as those stipulated in the Paris Agreement. There is an increasing body of literature that has explored the dynamics and sectoral structure of emissions across countries and regions. For instance, Lamb et al. (2021) provide a comprehensive synthesis of global emission trends, highlighting persistent growth across energy, industry, transport, and agricultural sectors, with only limited evidence of structural decarbonization. Earlier work by du Can and Price (2008) similarly documents long-term sectoral trends in global energy use and emissions, emphasizing the role of economic activity and sector-specific demand in shaping global emission pathways. Sector-specific studies also indicate significant heterogeneity in emission drivers. Blok et al. (2020) estimate substantial emission reduction potential across sectors, particularly in energy production and agriculture, while Salat (2009) and Haberl et al. (2023) highlight the critical role of built environments and infrastructure in shaping energy demand and emissions. At the macroeconomic level, the relationship between economic growth and emissions has also been widely explored. For example, Martín-Ortega and González-Sánchez (2025) analyze the Environmental Kuznets

Curve (EKC) hypothesis using sectoral GDP data, demonstrating heterogeneous long-run relationships between economic activity and emissions across countries.

A large body of literature is devoted to determining the causes of emission change through decomposition. Methods such as Structural Decomposition Analysis (SDA) and Index Decomposition Analysis (IDA) have been widely applied to disentangle the contributions of technological change, energy intensity, and structural shifts (Su & Ang, 2012; Qu, 2020). Empirical applications, including Hastuti et al. (2021), show that factors such as scale effects, energy efficiency, and structural demand are critical factors affecting emission trajectories. Although these methods can offer valuable insights into the drivers of emission change, they are mainly designed to explain changes in emission levels and may not fully capture the multidimensional structure of sectoral emission systems. This distinction is important because countries with similar aggregate emission levels may have very different underlying sectoral emission structures, and therefore may face different mitigation priorities. Studying the composition of emissions can consequently provide a more informative basis for understanding emission heterogeneity and designing targeted climate policies.

Over the last few years, machine learning and unsupervised clustering algorithms have become increasingly useful for analyzing complex environmental data. Studies such as Jiménez-Preciado et al. (2024) and Paramarta et al. (2025) apply Principal Component Analysis (PCA) and clustering algorithms to identify emission patterns across countries, demonstrating the effectiveness of dimensionality reduction in handling high-dimensional emission data. Similarly, Ganzenmüller et al. (2019) employ neural network-based approaches to classify emission pathways and identify structural shifts over time. Clustering approaches have also been applied in sector-specific contexts, such as the analysis of China's power industry emissions (Wen & Hao, 2024) and regional emission characteristics using fuzzy clustering techniques (Yu et al., 2012). These studies highlight the potential of unsupervised learning methods for uncovering latent structures within emission data.

In addition to the classification of emission patterns, the dynamics of emission systems have become the subject of increasing interest. Markov chain models provide transition-based approaches for examining how emission patterns change over time. For example, Chen et al. (2025) examine the evolution of emission regimes across Chinese provinces, demonstrating the importance of economic growth and structural factors in driving transitions. Broader perspectives on emission dynamics and policy implications are provided by studies such as Afuye and Tolios (2026) and Moghtader et al. (2026), which emphasize the persistence of emission-growth linkages and the challenges of achieving climate targets. Furthermore, complexity-based approaches (Mercure et al., 2016) highlight the importance of heterogeneity and path dependence in understanding sustainability transitions.

Even with these advances, there are still some critical gaps. First, the existing literature concentrates either on decomposition of emission drivers or on clustering based on aggregate indicators, in either case without fully leveraging the compositional nature of sectoral emissions. Second, although emission patterns have been classified using clustering approaches, these studies often overlook the temporal dynamics of the identified patterns. Third, dynamic analyses are generally confined to particular regions or sectors rather than being offered in a global, multi-sectoral context. More importantly, previous applications of PCA and clustering have primarily focused on identifying similarities or differences in emission patterns, while less attention has been given to treating the relative sectoral composition itself as an emission fingerprint and then tracking the persistence and transition of the resulting regimes over time.

To fill these gaps, this paper suggests an emission fingerprinting framework in which each country-year observation is modeled as a compositional vector of sectoral emissions. This framework is used to identify and describe specific emission regimes and examine how these regimes evolve over time by incorporating dimensionality reduction using PCA, unsupervised clustering methods, and dynamic transition analysis. Unlike approaches that focus primarily on aggregate emission magnitudes or static classifications, the proposed framework combines compositional representation, PCA-based dimensionality reduction,

K-means clustering, and Markov transition analysis within a single global multi-sectoral framework. This allows the analysis to move from identifying structural differences between countries to examining how these structures persist, diversify, and transition over time. In particular, the study seeks to (i) categorize countries in terms of the structural composition of emissions, (ii) identify dominant emission regimes, and (iii) examine the nature of persistence and transition of these regimes over time.

The results add to the literature in three fundamental aspects. First, they offer an innovative view of global emission heterogeneity by emphasizing structural composition instead of aggregate levels. Second, they add a dynamic aspect to emission analysis by quantifying changes in emission regimes over time. Third, they provide information on the stability and transformation of emission systems, which have significant implications for designing targeted and effective climate policies. Together, these contributions provide a framework for understanding not only how much countries emit, but also how their emissions are structurally organized and how these structures change over time.

2. Data and Methodology

2.1. Data Structure and Notation

Let $i = 1,\ldots\ldots N$ index countries and $t = 1,\ldots..T$ index time (years). The dataset consists of annual observations of sectoral greenhouse gas (GHG) emissions for each country-year pair over the period 1990--2023, obtained from the Our World in Data (OWID) database, based on Climate Watch (2026).

For each country i at time t emissions are observed across $K = 10$ sectors. The sectoral emission vector is defined as:

$$x_{it} = \left(x_{it}^{(1)}, x_{it}^{(2)}, \ldots, x_{it}^{(K)}\right), \quad (1)$$

where $x_{it}^{(k)}$ denotes emissions (in tonnes of CO_2) from sector k. The sectors include agriculture, waste, buildings, industry, manufacturing and construction, transport, electricity and heat, fugitive emissions, other fuel combustion, and aviation and shipping.

Stacking across all observations yields a high-dimensional panel dataset:

$$X \in R^{\{(N \times T) \times K\}}, \quad (2)$$

which forms the basis for subsequent multivariate analysis. This representation allows each country-year observation to be treated as a point in a K dimensional emission space.

2.2. Sample Construction and Preprocessing

To ensure consistency and comparability, several preprocessing steps were applied to construct the analytical sample. First, aggregate entities such as global regions (e.g., World, Asia, Europe) and income groups were filtered out, leaving only sovereign country-level observations.

Second, observations with missing values in any of the sectoral variables were removed to maintain a complete multivariate structure. This ensures that all observations lie in the same K-dimensional space, which is necessary for dimensionality reduction and clustering algorithms.

Third, the land-use change and forestry (LULCF) sector was excluded from the analysis because it exhibited substantial volatility and contained negative emission values. In the original dataset, LULCF includes both positive and negative values, reflecting net sources and removals of greenhouse gases. Because the present analysis represents emissions as compositional shares, each component must be non-negative and contribute to a

common total. Negative values would therefore make the construction of valid compositional emission fingerprints problematic and could distort the interpretation of relative sectoral contributions. The exclusion of LULCF means that the identified emission regimes should be interpreted as reflecting the structural composition of the ten included emission sectors rather than the complete national greenhouse gas balance including land-use-related removals.

Following these procedures, the final dataset contained 5,522 country-year observations from 172 countries over the 1990–2023 period, providing a consistent analytical panel for the subsequent multivariate analysis.

2.3. Construction of Emission Fingerprints

To characterize the structural composition of emissions, each country-year observation is represented as a high-dimensional emission vector. For country i at time t, the sectoral emission vector is defined as:

$$x_{it} = \left(x_{it}^{(1)}, x_{it}^{(2)}, \dots, x_{it}^{(K)}\right), \tag{3}$$

where $x_{it}^{(k)}$ denotes emissions (in tonnes of CO_2) from sector k. Since the primary interest lies in the relative contribution of each sector rather than absolute emission levels, the emission vectors are transformed into compositional shares. Specifically, the emission share of sector k for country i at time t is defined as:

$$s_{it}^{k} = \frac{x_{it}^{k}}{\Sigma_{k=1}^{K} x_{it}^{k}}, \; for\; k = 1, \dots, K \tag{4}$$

such that:

$$\Sigma_{k=1}^{K} s_{it}^{k} = 1. \tag{5}$$

The resulting vector

$$s_{it} = (s_{it}^{1}, s_{it}^{2}, \dots, s_{it}^{K}), \tag{6}$$

represents the emission fingerprint of a country-year observation, capturing the proportional structure of emissions across sectors. This representation allows country-year observations with different overall emission magnitudes to be compared according to the relative composition of their sectoral emissions. The resulting data lie in the simplex space, consistent with the statistical framework for compositional data (Aitchison, 1982).

2.4. Dimensionality Reduction via Principal Component Analysis

Given the high dimensionality of the emission fingerprints, Principal Component Analysis (PCA) is employed to reduce dimensionality while preserving the dominant variance structure in the data. Because the sectoral shares are measured on comparable proportional scales but may exhibit different variances, the emission-share matrix is standardized before PCA.

Let $S \in \mathbb{R}^{M \times K}$ denote the matrix of emission shares, where $M = N \times T$ is the total number of observations. After standardization, PCA seeks a linear transformation:

$$Z = SW, \tag{7}$$

where $W \in \mathbb{R}^{K \times K}$ is a matrix of eigenvectors derived from the covariance matrix of S, and Z represents the transformed data in the principal component space.

The principal components are arranged in descending order of explained variance, such that the initial few components account for the largest portion of the overall variance. The most significant components, which explain a substantial proportion of the variance,

are retained for further analysis, as is standard practice in multivariate statistics (Jolliffe, 2002).

In the present analysis, the first four principal components were retained for clustering, while the first two components were used for visualization and interpretation of the dominant structural gradients.

2.5. Clustering Methodology

To identify distinct emission regimes, K-means clustering is applied to the reduced-dimensional representation of the data. The objective of K-means is to partition the observations into K_c clusters by minimizing the within-cluster sum of squares (Hartigan & Wong, 1979):

$$\min_{C_1,\ldots,C_{K_c}} \Sigma_{k=1}^{K_c} \Sigma_{i \in C_k} ||z_i - \mu_k||^2, \quad (8)$$

Where C_k denotes the set of observations assigned to cluster *k*, z_i is the vector of principal component scores for observation *i*, and μ_k is the centroid of cluster *k.*

The combination of silhouette analysis and the elbow method was used to determine the optimal number of clusters *Kc*. The elbow method examines the marginal decrease in within-cluster variance as the number of clusters is increased, whereas the silhouette score measures the amount of separation between clusters (Rousseeuw, 1987). These methods offer complementary measures for determining a well-separated and interpretable clustering framework.

2.6. Cluster Interpretation and Emission Regimes

Following the application of K-means clustering, each cluster is characterized by its centroid, representing the average emission fingerprint within that cluster. Formally, the centroid of cluster k is defined as:

$$\mu_k = \left(\frac{1}{|C_k|}\right) \Sigma_{i \in C_k} s_i \quad (9)$$

Where C_k denotes the set of observations assigned to cluster *k* and s_i is the compositional emission vector for observation *i*.

Each centroid $\mu_k = (\mu_k^1, \mu_k^2, \ldots, \mu_k^K)$ represents the average share of emissions across sectors, thereby providing a structural signature of the corresponding emission regime. Based on the dominant sectoral contributions, clusters are interpreted as distinct emission regimes, including:

- Agriculture-dominated systems, characterized by a high proportion of emissions from the agricultural sector;
- Energy-dominated systems, where emissions are primarily driven by electricity, heat, and industrial activities;
- Mixed or diversified systems, where emissions are more evenly distributed across multiple sectors.

Based on the dominant sectoral contributions, the clusters are interpreted as distinct emission regimes. The interpretation is based on the sectoral shares represented by the cluster centroids rather than on aggregate emission levels. This approach allows the statistical clustering results to be linked to meaningful structural characteristics of national emission systems.

2.7. Dynamic Transition Analysis

To examine the temporal evolution of emission regimes, a first-order Markov transition framework is employed. Let $C_{it} \in \{1, \dots, K_c\}$ denote the cluster membership of country i at time t. The transition probability from cluster r to cluster s is defined as:

$$P_{rs} = \Pr(C_{t+1} = s \mid C_t = r), \tag{10}$$

where P_{rs} represents the probability that an observation transitions from regime r at time t to regime s at time $t + 1$.

The transition probabilities are estimated empirically using observed transitions across consecutive time periods:

$$\hat{P}_{rs} = \frac{n_{rs}}{\sum_{s=1}^{K_c} n_{rs}}, \tag{11}$$

where n_{rs} denotes the number of observed transitions from cluster r to cluster s. The resulting transition matrix P summarizes the persistence and mobility of emission regimes over time.

This framework allows for the assessment of regime stability (diagonal dominance of P) and structural transformation (off-diagonal transitions), providing insights into the dynamic behavior of global emission systems. The use of Markov processes for modeling state transitions follows the classical framework of Anderson and Goodman (1957)

2.8. Robustness and Validation

To assess the reliability and robustness of the clustering results, several complementary validation procedures were implemented. First, cluster quality was evaluated using the silhouette coefficient. Second, clustering stability was assessed by comparing cluster assignments across repeated K-means runs with different random initializations using the Adjusted Rand Index (ARI). Third, robustness across clustering algorithms was evaluated by comparing the K-means solution with a Ward hierarchical clustering solution using the ARI. Finally, internal cluster validity was assessed using the Calinski–Harabasz and Davies–Bouldin indices, which provide complementary measures of between-cluster separation and within-cluster compactness.

First, cluster quality was evaluated using the silhouette coefficient. For each observation i, the silhouette value is defined as:

$$s(i) = \frac{(b(i) - a(i))}{\max\{a(i), b(i)\}}, \tag{12}$$

where $a(i)$ denotes the average distance between observation i and all other points within the same cluster, and $b(i)$ denotes the minimum average distance between observation i and points in the nearest neighboring cluster. The silhouette score lies in the interval [−1, 1], with higher values indicating better cluster separation (Rousseeuw, 1987).

Second, clustering stability was assessed by comparing cluster assignments across multiple runs of the K-means algorithm with different random initializations. Let C(1) and C(2) denote two clusterings obtained from different initializations. Their agreement was evaluated using the Adjusted Rand Index (ARI), which measures the similarity between two partitions while correcting for agreement that may occur by chance. Higher ARI values indicate greater consistency between clustering solutions, with an ARI of 1 indicating identical partitions.

Finally, an alternative clustering method based on hierarchical clustering with Ward's linkage was applied (Ward, 1963). The consistency between K-means and hierarchical clustering solutions was evaluated through cross-tabulation of cluster assignments. A high level of agreement across methods indicates that the identified emission regimes are not driven by a specific algorithmic choice, but reflect underlying structural patterns in the data (Hennig et al., 2015).

3. Results and Discussion

3.1. Descriptive Overview

The cleaned data have 5,522 observations in 172 countries in the 1990-2023 period. Each of the observations is a compositional vector of greenhouse gas emissions in ten sectors, and is normalized in the sense that share in one sector plus share in another sector, etc., sums to unity. Such a structure of composition would guarantee comparability between countries with very different levels of emissions.

First-time investigation shows that there is a great level of heterogeneity in the sectoral contributions. Although there are countries with a dominance of agriculture in the emission profile, some countries have a strong share of energy-related sectors in the emission profile: electricity generation, transport, and manufacturing. This heterogeneity motivates the use of dimensionality reduction and clustering techniques to uncover latent emission regimes.

3.2. Principal Component Analysis

Principal Component Analysis (PCA) was applied to the standardized sectoral emission shares to identify the major dimensions of variation in the global emission fingerprints. The first principal component (PC1) explained 23.91% of the total variance, while PC2 explained 13.97%, giving a cumulative contribution of 37.88% for the first two components. PC3 and PC4 explained a further 11.74% and 11.13%, respectively. Together, the first four principal components accounted for 60.75% of the total variance and were therefore retained as the reduced-dimensional representation for subsequent clustering.

Table 1. Explained variance of the principal components

Principal component	Standard deviation	Proportion of variance (%)	Cumulative variance (%)
PC1	1.546	23.91	23.91
PC2	1.182	13.97	37.88
PC3	1.083	11.74	49.62
PC4	1.055	11.13	60.75
PC5	1.013	10.27	71.01

The first two components provide the most readily interpretable dimensions of the emission structure. PC1 is strongly associated with a contrast between agriculture and industrial-energy-related sectors. Agriculture has a large negative loading on PC1 (−0.582), whereas electricity and heat (0.488), manufacturing and construction (0.393), transport (0.338), and industry (0.293) have positive loadings. PC2 is dominated by fugitive emissions, which has a loading of 0.693, while transport (−0.375), industry (−0.342), and waste (−0.329) contribute negatively. These patterns suggest that the main structural variation in the data reflects differences between agrarian and industrial-energy-oriented systems, as well as differences associated with fossil extraction and other emission structures.

Figure 1 presents the scree plot and the cumulative variance explained by the principal components. The relatively larger contribution of the first four components supports their use as a reduced-dimensional representation for clustering, while the first two components provide a useful two-dimensional visualization of the major emission gradients.

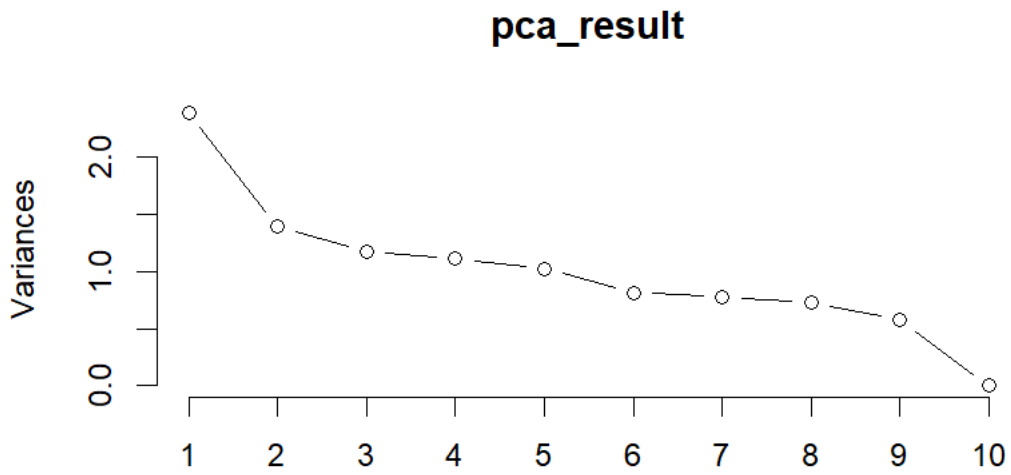
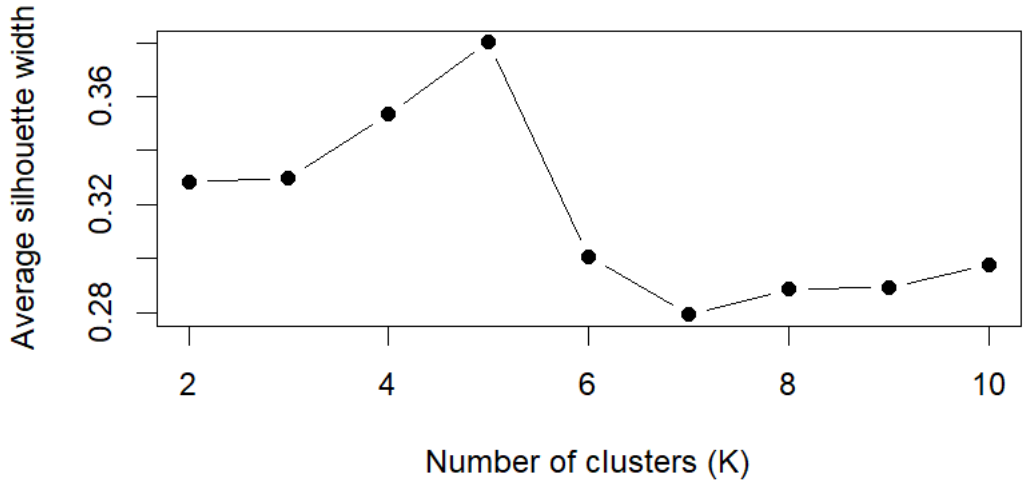


Figure 1: Scree plot showing the variance explained by each principal component

3.2. Determination of the Number of Clusters

The optimal number of clusters was assessed using both the elbow method and the average silhouette coefficient. The elbow plot showed a clear reduction in within-cluster sum of squares as the number of clusters increased, while the silhouette analysis indicated meaningful separation among the candidate cluster solutions. Based on the combined evidence from these diagnostics and the interpretability of the resulting emission structures, a five-cluster solution was selected. The average silhouette width for the five-cluster solution was 0.380, indicating moderate but meaningful separation among the identified regimes.



Figuer 2: Average silhouette width across different cluster sizes

The average silhouette score of 0.38 indicates moderate but meaningful cluster separation.

The resulting clusters exhibit the following sizes:

• Cluster 1: 611 observations

- Cluster 2: 1,668 observations
- Cluster 3: 258 observations
- Cluster 4: 174 observations
- Cluster 5: 2,811 observations

3.3. Characterization of Emission Regimes

The K-means analysis identified five distinct emission regimes based on the first four principal components. The distribution of observations across the clusters was uneven. Cluster 1 contained 611 observations (11.1%), Cluster 2 contained 1,668 observations (30.2%), Cluster 3 contained 258 observations (4.7%), Cluster 4 contained 174 observations (3.2%), and Cluster 5 contained 2,811 observations (50.9%). The sectoral composition of the cluster centroids provides the basis for interpreting these regimes.

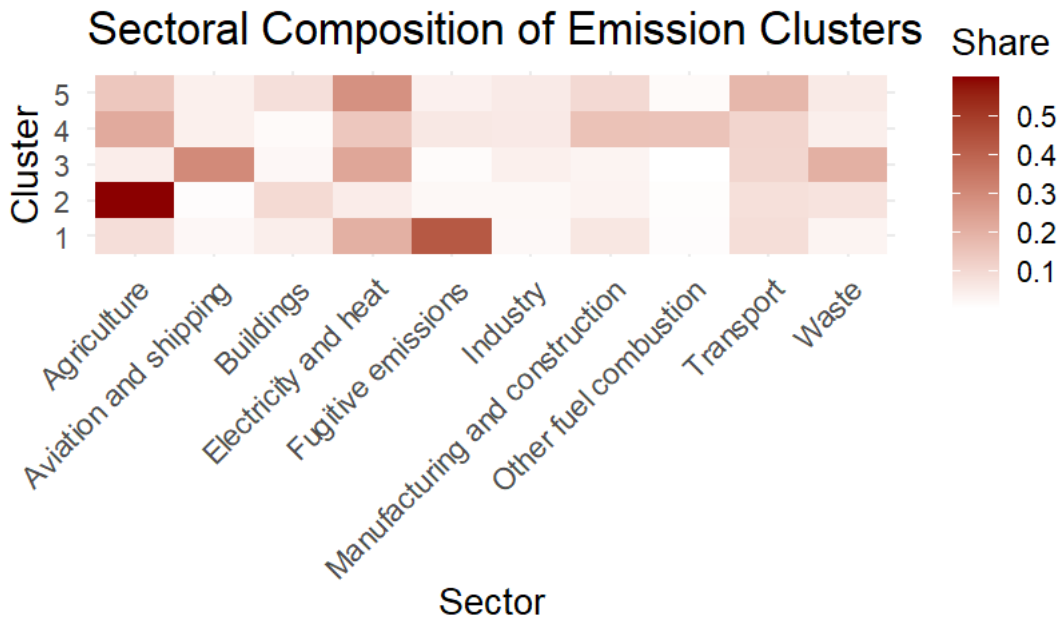


Figure 3: Sectoral composition of emission regimes based on cluster centroids

The five regimes show clear differences in their sectoral emission structures. Cluster 1 is characterized by a dominant contribution from fugitive emissions (42.6%), followed by electricity and heat (20.2%). Cluster 2 is strongly agriculture-dominated, with agriculture accounting for 59.7% of its emission composition. Cluster 3 has relatively high contributions from electricity and heat (22.6%) and waste (20.2%). Cluster 4, the smallest regime, shows relatively high contributions from agriculture (21.6%), manufacturing and construction (15.7%), and electricity and heat (14.3%). Cluster 5, which accounts for more than half of all observations (50.9%), is characterized by a relatively diversified structure, with electricity and heat (28.0%) and transport (18.4%) making the largest contributions.

Overall, the cluster profiles indicate that countries differ substantially in the structural composition of their emissions. While some regimes are strongly dominated by a particular sector, such as agriculture or fugitive emissions, others show a more diversified combination of energy, transport, industrial, and agricultural sources.

Cluster	Representative countries from the current analysis
1	Venezuela, Syria, Yemen, Algeria, Gabon
2	Nicaragua, Bangladesh, Albania, Cambodia, Zambia
3	Mauritius, Panama, Seychelles, Malta, Cape Verde

4	Kyrgyzstan, Moldova, Bosnia and Herzegovina, Ukraine, Iceland
5	Chile, Egypt, Dominican Republic, Mexico, Tunisia

To provide a more tangible interpretation of the identified regimes, representative country-year observations were identified based on their proximity to the corresponding cluster centroids. These examples provide a country-level illustration of the structural characteristics represented by each regime.

3.4. Temporal Dynamics of Emission Regimes

The temporal dynamics of the identified emission regimes were examined using a first-order Markov transition framework. The resulting transition matrix shows a strong degree of persistence within all five regimes. The probability of remaining in the same regime ranges from 0.876 to 0.973. Cluster 5 shows the highest persistence, with a probability of 0.973 of remaining in the same regime from one year to the next, followed by Cluster 2 (0.967), Cluster 3 (0.952), and Cluster 1 (0.938). Cluster 4 has the lowest persistence probability at 0.876, indicating comparatively greater mobility.

Table 2. First-order transition probabilities between emission regimes

Current regime	Regime 1	Regime 2	Regime 3	Regime 4	Regime 5
Regime 1	0.938	0.008	0.000	0.002	0.052
Regime 2	0.001	0.967	0.002	0.000	0.030
Regime 3	0.000	0.012	0.952	0.000	0.036
Regime 4	0.018	0.012	0.000	0.876	0.094
Regime 5	0.009	0.009	0.003	0.006	0.973

Although the diagonal probabilities indicate substantial persistence, the off-diagonal probabilities provide evidence of gradual structural change. The largest transition probability outside the diagonal occurs from Cluster 4 to Cluster 5 (0.094), followed by transitions from Cluster 1 to Cluster 5 (0.052), Cluster 3 to Cluster 5 (0.036), and Cluster 2 to Cluster 5 (0.030). Overall, the transition structure suggests that countries generally remain within their existing emission regimes over short periods, while movement toward more diversified or energy-mixed structures occurs gradually rather than through abrupt changes.

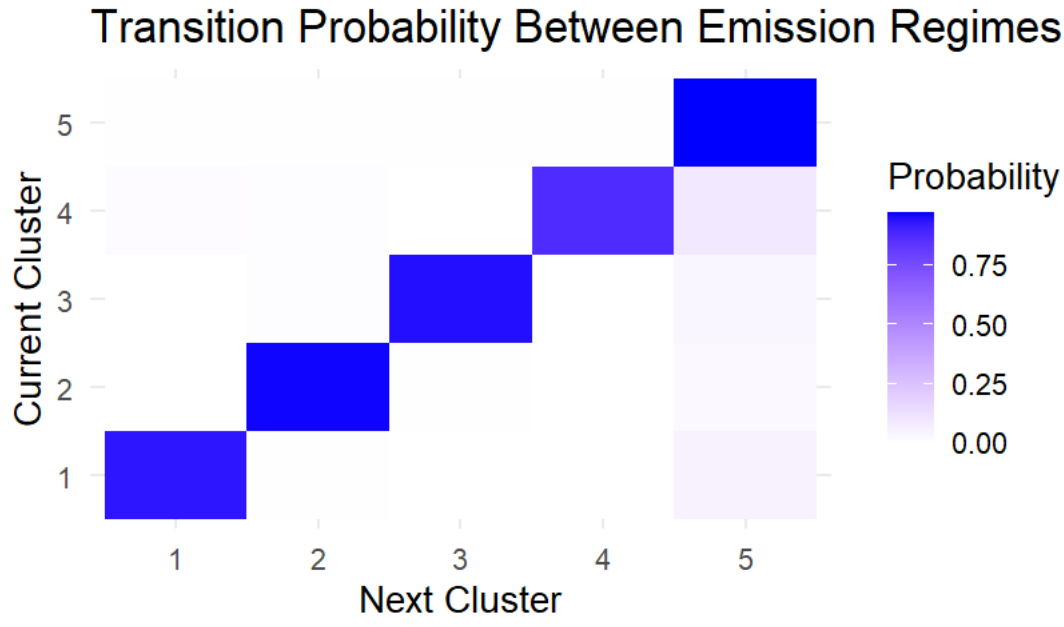


Figure 4: Transition probability matrix between emission regimes.

3.5. Robustness and Validation

Several complementary diagnostics were used to assess the robustness and quality of the five-cluster solution. The average silhouette width was 0.380, indicating moderate but meaningful separation among the identified regimes. Repeated K-means runs with different random initializations produced identical cluster assignments, yielding an Adjusted Rand Index (ARI) of 1.000. This indicates perfect stability of the clustering solution with respect to random initialization.

The K-means solution was also compared with a Ward hierarchical clustering solution. The resulting ARI was 0.571, indicating moderate to substantial agreement between the two clustering approaches. Although the two algorithms do not produce identical partitions, the degree of agreement suggests that the broad structural patterns identified by K-means are not entirely dependent on a single clustering algorithm.

Table 3. Robustness and internal validation of the five-cluster solution

Validation measure	Result	Interpretation
Average silhouette width	0.380	Moderate cluster separation
ARI: repeated K-means	1.000	Perfect stability
ARI: K-means vs. Ward	0.571	Moderate to substantial agreement
Calinski–Harabasz Index	2161.84	Strong between-cluster separation relative to within-cluster dispersion
Davies–Bouldin Index	0.877	Good cluster compactness and separation

The internal validation indices provide further support for the quality of the clustering solution. The Calinski–Harabasz Index was 2161.84, indicating substantial between-cluster dispersion relative to within-cluster dispersion. The Davies–Bouldin Index was 0.877, with the relatively low value indicating good separation and compactness of the identified clusters. Taken together, the silhouette coefficient, repeated-run ARI, cross-method ARI, Calinski–Harabasz Index, and Davies–Bouldin Index provide consistent evidence that the five-regime solution is reasonably stable and structurally meaningful.

4. Discussion

4.1. Synthesis of Key Findings

This study provides a novel empirical characterization of global greenhouse gas emission structures by integrating compositional analysis, dimensionality reduction, clustering, and dynamic transition modeling. The findings highlight three core points. First, the agrarian-industrial transition axis can be well summarized in the first principal component of the emission structures across countries. Second, five different emission regimes are identified based on heterogeneous sectoral composition among countries. Third, the dynamic analysis shows that these regimes are highly persistent over time, implying limited mobility across emission structures. The five regimes also show strong clustering stability, with repeated K-means runs producing identical assignments (ARI = 1.00).

The overall implications of these findings are that there are stable and structurally differentiated emission pathways at the global level.

4.2. Structural Interpretation of Emission Regimes

The clusters identified are associated with different economic and structural forms of national economies. The agriculture-dominated regime reflects economies where land-based activities are the biggest source of emissions. By contrast, the energy-mixed and industrial regimes are associated with emission structures that are more concentrated in electricity production, transport, and manufacturing.

This pattern is consistent with the broader literature on structural transformation. Lamb et al. (2021) and du Can and Price (2008) emphasize the importance of sectoral transformation, economic activity, and changes in energy use in shaping emission pathways**.** The results further reinforce the idea that emission structures are closely tied to stages of economic development and sectoral composition.

4.3. Persistence and Path Dependency in Emission Structures

One of the important contributions of the research is the dynamic approach to emission regimes. The estimated transition probabilities suggest that countries are highly persistent within their respective clusters, with diagonal probabilities ranging from 0.876 to 0.973.

This observation implies that emission structures are highly path-dependent, reflecting long-term investments in infrastructure, energy systems, and production technologies. Such persistence aligns with the concept of technological and institutional lock-in, which Mercure et al. (2016) discuss in relation to path dependence and sustainability transitions. Once established, emission patterns tend to reinforce themselves, making rapid structural transitions difficult.

The limited transitions observed, primarily toward energy-mixed systems, indicate gradual rather than abrupt changes in emission composition. This finding is also consistent with the transition-based evidence reported by Chen et al. (2025), who highlight the role of structural and economic factors in shaping the evolution of emission regimes.

4.4. Relation to Existing Empirical Literature

The findings contribute to a growing body of research employing unsupervised learning techniques to identify emission patterns across countries. Jiménez-Preciado et al. (2024) and Paramarta et al. (2025) demonstrate the usefulness of PCA and clustering approaches for identifying latent emission structures across countries**.** However, most existing work focuses on static classifications.

By incorporating a dynamic Markov framework, this study extends the literature by capturing the temporal evolution of emission regimes. This approach complements traditional decomposition methods such as Structural Decomposition Analysis (SDA) and Index Decomposition Analysis (IDA), which focus on drivers of emission changes but do not explicitly model regime transitions. Su and Ang (2012) and Qu (2020) show how decomposition methods can identify the contributions of technological, structural, and energy-related factors to emission changes, while the present framework focuses on the structural regimes and their transitions over time. The additional robustness analysis also supports the reliability of the identified regimes, with an ARI of 0.571 between K-means and Ward clustering, a Calinski–Harabasz Index of 2161.84, and a Davies–Bouldin Index of 0.877.

4.5. Policy Implications

The discovery of different emission regimes has significant climate policy implications. The findings indicate that a single, one-size-fits-all policy scheme is unlikely to work, considering the fact that the emission structure of different countries is heterogeneous.

In economies that are largely based on agriculture, mitigation efforts should focus on sustainable land use, methane reduction, and agricultural efficiency. Conversely, industrial and energy-intensive regimes need policies that focus on energy transition, electrification, and technological innovation.

The differences across regimes suggest that mitigation priorities should be aligned with the dominant sources within each emission structure rather than based only on aggregate emission levels. This high persistence in emission regimes further suggests that policy interventions should be long term and structural in nature. Without more fundamental change in economic and energy systems, short-term actions are not likely to cause significant changes in the composition of emissions.

4.6. Methodological Contribution

This study contributes methodologically by integrating compositional data analysis with unsupervised machine learning and stochastic process modeling. By transforming

emissions into compositional shares, the analysis ensures scale invariance and cross-country comparability, consistent with the statistical framework proposed by Aitchison (1982).

The combined use of PCA and K-means clustering allows for effective identification of latent emission structures, while the Markov transition framework provides insights into their temporal stability. The robustness results, including perfect agreement across repeated K-means runs and complementary validation using alternative clustering and internal validity measures, further support the reliability of this integrated framework. This approach offers a flexible and scalable framework for analyzing complex, high-dimensional environmental data.

4.7. Limitations and Future Research

The study has a number of limitations despite its contributions. First, the use of compositional shares abstracts from absolute emission levels, which can obscure changes in total emission magnitude. Second, the clustering methodology is descriptive and does not provide causal relationships between economic variables and emission structures. Third, the aggregation of sectors may mask within-sector heterogeneity.

The framework could be expanded in future studies by adding explanatory variables such as GDP, energy intensity, and policy indicators to model the drivers of regime transitions. Integrating socioeconomic and policy variables could also help explain why countries move between emission regimes and what factors support or hinder structural transitions. Probabilistic clustering techniques or dynamically evolving clustering techniques may further offer insights into overlapping or changing emission patterns.

4.8. Conclusion

This analysis provides a complete empirical framework of the structural composition and the structural evolution of global greenhouse gas emissions. The analysis has identified five different emission regimes, which provide evidence of the heterogeneity of sectoral emission structures across countries. The robustness analysis further supports the stability of these regimes, with repeated K-means runs producing identical cluster assignments (ARI = 1.00) and moderate agreement between K-means and Ward clustering (ARI = 0.57).

The results show that the global emission patterns are not haphazard but are structured into stable and interpretable regimes, ranging from systems dominated by agriculture to systems dominated by energy-intensive and diversified sources. The dynamic analysis also shows high persistence in these regimes, with the probability of remaining within the same regime ranging from 0.876 to 0.973, and this is an indication that the emission structures are highly path-dependent and resilient to short-run change.

Notably, the findings demonstrate a gradual movement towards mixed systems of emissions, which are indicative of larger trends of industrialization and structural change. Nevertheless, the small mobility between regimes highlights the difficulties of speedy decarbonization without profound structural and technological transformations. The relatively strong cluster validation results, including a Calinski–Harabasz Index of 2161.84 and a Davies–Bouldin Index of 0.877, further support the meaningful separation and stability of the identified regimes.

Policy-wise, the evidence indicates that successful climate policy requires adaptation to the underlying emission structure of each country. Prescriptions of uniform policies can hardly work in conditions of heterogeneity like this one. Rather, it needs targeted, sector-specific, and long-term interventions to enable meaningful transitions towards low-carbon pathways. The persistence of the regimes also suggests that such interventions need to address the underlying infrastructure, energy systems, and production structures rather than relying only on short-term measures.

On the whole, this research offers both methodological and substantive contributions to the study of climate policy design by providing a framework that scales to identify and

track emission regimes, thus providing a basis to develop more subtle and evidence-based climate policy design.

Declarations

Ethics and Guidelines: Not Applicable.

Consent to participate: Not Applicable.

Consent to publish: The authors have provided consent to publish.

Competing interests: The authors declare no competing interests.

Data availability statement: Data will be made available on reasonable request from the corresponding author.

Funding: This research received no external funding.

Clinical Trial Number: Not Applicable.

Declaration of using generative AI: During the preparation of this work the author(s) used ChatGPT in order to correct the grammatical errors. After using this tool/service, the author(s) reviewed and edited the content as needed and take full responsibility for the content.

Author Contributions: R.E.A.R.: Conceptualization, methodology, writing—original draft preparation, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

References

Afuye, G. A., & Tolios, G. (2026). Quantitative impacts and patterns of greenhouse gas emission trends: Implications for global change and policy. *Carbon Management*, *17*(1), Article 2622340. https://doi.org/10.1080/17583004.2026.2622340

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *44*(2), 139–160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x

Anderson, T. W., & Goodman, L. A. (1957). Statistical inference about Markov chains. *The Annals of Mathematical Statistics*, *28*(1), 89–110. https://doi.org/10.1214/aoms/1177707039

Blok, K., Afanador, A., Van Der Hoorn, I., Berg, T., Edelenbosch, O. Y., & Van Vuuren, D. P. (2020). Assessment of sectoral greenhouse gas emission reduction potentials for 2030. *Energies*, *13*(4), Article 943. https://doi.org/10.3390/en13040943

Chen, S., Kang, X., & Park, S. Y. (2025). Transitions of carbon dioxide emissions in China: K-means clustering and discrete endogenous Markov chain approach. *Climate*, *13*(8), Article 165. https://doi.org/10.3390/cli13080165

du Can, S. d. I. R., & Price, L. (2008). Sectoral trends in global energy use and greenhouse gas emissions. *Energy Policy*, *36*(4), 1386–1403. https://doi.org/10.1016/j.enpol.2007.12.017

Ganzenmüller, R., Pradhan, P., & Kropp, J. P. (2019). Sectoral performance analysis of national greenhouse gas emission inventories by means of neural networks. *Science of the Total Environment*, *656*, 80–89. https://doi.org/10.1016/j.scitotenv.2018.11.311

Haberl, H., Löw, M., Perez-Laborda, A., Matej, S., Plank, B., Wiedenhofer, D., Creutzig, F., Erb, K.-H., & Duro, J. A. (2023). Built structures influence patterns of energy demand and CO_2 emissions across countries. *Nature Communications*, *14*(1), Article 3898. https://doi.org/10.1038/s41467-023-39728-3

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108. https://doi.org/10.2307/2346830

Hastuti, S. H., Hartono, D., Putranti, T. M., & Imansyah, M. H. (2021). The drivers of energy-related CO_2 emission changes in Indonesia: Structural decomposition analysis. *Environmental Science and Pollution Research*, *28*(8), 9965–9978. https://doi.org/10.1007/s11356-020-11414-7

Jiménez-Preciado, A. L., Cruz-Aké, S., & Venegas-Martínez, F. (2024). Identification of patterns in CO_2 emissions among 208 countries: K-means clustering combined with PCA and non-linear t-SNE visualization. *Mathematics*, *12*(16), Article 2591. https://doi.org/10.3390/math12162591

Jolliffe, I. T. (2002). Graphical representation of data using principal components. In *Principal component analysis* (2nd ed., pp. 64–91). Springer. https://doi.org/10.1007/b98835

Lamb, W. F., Wiedmann, T., Pongratz, J., Andrew, R., Crippa, M., Olivier, J. G., Wiedenhofer, D., Mattioli, G., Khourdajie, A. A., House, J., et al. (2021). A review of trends and drivers of greenhouse gas emissions by sector from 1990 to 2018. *Environmental Research Letters*, *16*(7), Article 073005. https://doi.org/10.1088/1748-9326/abee4e

Martín-Ortega, J. L., & González-Sánchez, M. (2025). Sectoral composition of GDP and greenhouse gas emissions: An empirical analysis in EU27. *Environment, Development and Sustainability*, *27*(2), 3899–3926. https://doi.org/10.1007/s10668-023-04048-5

Mercure, J.-F., Pollitt, H., Bassi, A. M., Viñuales, J. E., & Edwards, N. R. (2016). Modelling complex systems of heterogeneous agents to better design sustainability transitions policy. *Global Environmental Change*, *37*, 102–115. https://doi.org/10.1016/j.gloenvcha.2016.02.003

Moghtader, B., Ulengin, F., Yıldız, E. H., Aktas, E., & Topcu, Y. I. (2026). Trade flows and carbon emissions: How far are we from the climate targets? *International Journal of Sustainable Transportation*, 1–14. https://doi.org/10.1080/15568318.2026.2649318

Paramarta, V., Rahman, A., Priska, L., Roken Gurning, R., & Ayu Purwati, W. (2025). Comparative Analysis of K-Means and Gaussian Mixture Model in Clustering Global CO2 Emissions. *Bit-Tech*, *8*(1), 998–1008. https://doi.org/10.32877/bt.v8i1.2805

Qu, S. (2020). The decomposition analysis of carbon emissions: Theoretical basis, methods and their evaluations. *Chinese Journal of Urban and Environmental Studies*, *8*(4), Article 2050020. https://doi.org/10.1142/S2345748120500207

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Salat, S. (2009). Energy loads, CO_2 emissions and building stocks: Morphologies, typologies, energy systems and behaviour. *Building Research & Information*, *37*(5–6), 598–609. https://doi.org/10.1080/09613210903162126

Su, B., & Ang, B. W. (2012). Structural decomposition analysis applied to energy and emissions: Aggregation issues. *Economic Systems Research*, *24*(3), 299–317. https://doi.org/10.1080/09535314.2012.677997

Wen, L., & Hao, Y. (2024). Factor decomposition and clustering analysis of CO_2 emissions from China's power industry based on Shapley value. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, *46*(1), 9362–9378. https://doi.org/10.1080/15567036.2020.1776795

Yu, S., Wei, Y.-M., Fan, J., Zhang, X., & Wang, K. (2012). Exploring the regional characteristics of inter-provincial CO_2 emissions in China: An improved fuzzy clustering analysis based on particle swarm optimization. *Applied Energy*, *92*, 552–562. https://doi.org/10.1016/j.apenergy.2011.11.068

Disclaimer: All statements, viewpoints, and data featured in the publications are exclusively those of the individual author(s) and contributor(s), not of SRR and/or its editor(s). SRR and/or the editor(s) absolve themselves of any liability for harm to individuals or property that might arise from any concepts, methods, instructions, or products mentioned in the content.