


Decoding CO₂: A Statistical Dive into Global Emissions



AMS 597 - Statistical Computing
Prof. Silvia Sharna



Devaansh Kataria
Garima Prachi
Laurence Caradonna
Prerna Sachdeva
Rishabh Gosain

INTRODUCTION

This project explores the relationship between **CO₂ emissions**, **economic growth**, and **demographic indicators** using a comprehensive dataset spanning multiple countries and decades. With rising environmental concerns, understanding how these factors interact is crucial for shaping sustainable development policies.


Our analysis is guided by three key research questions:

1. **How do environmental and demographic factors influence economic performance**, as measured by GDP, when analyzed in relative (percentage) terms?
2. **Do CO₂ and economic patterns cluster more clearly by income level or by geographic region?**
3. **How have the relationships between CO₂ emissions and economic indicators changed over time since 1982**, and are there identifiable breakpoints that correspond to major global economic or policy events?

To address these questions, we first apply **linear regression** on log-transformed GDP to evaluate how **CO₂ emissions**, **energy use**, and **population** impact economic output in percentage terms. We then use **K-Means clustering** to identify whether countries group more clearly by **income level** or **region** based on environmental and economic indicators. Lastly, we conduct **time series analysis with structural break testing** to trace how the CO₂–GDP relationship has shifted since 1982, revealing key breakpoints linked to global events. All modeling is done in **R**, focusing on reproducibility and interpretability.



DATA DESCRIPTION

- **Source:**
 - The dataset was obtained from Our World in Data (OWID), available publicly on [GitHub](#). It compiles data from trusted institutions including the Global Carbon Project, Energy Institute, and U.S. Energy Information Administration (EIA).
 - **Why This Dataset Was Chosen:**
 - It offers global coverage from 1750 to present, includes key variables like CO₂ emissions, GDP, population, and energy use, and allows for robust analyses across time and geography.
 - **Why This Dataset Was Chosen:**
 - The dataset supports the project's goals of studying emission trends, economic performance, and structural changes through regression, clustering, and time series analysis.
- 



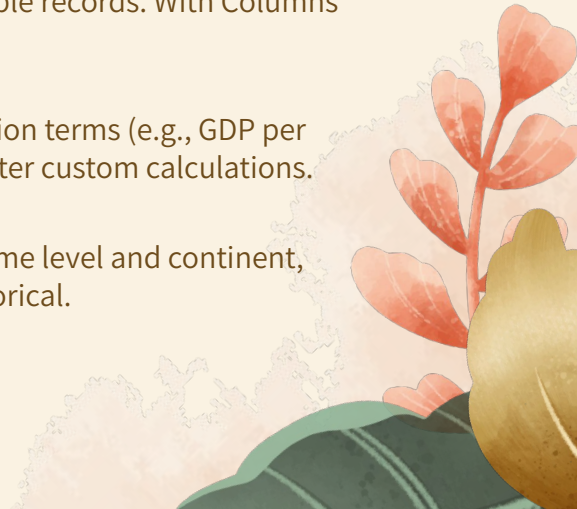
DATA PREPROCESSING

Timeframe & Filtering: The dataset originally had 50,590 rows spanning from the 1800s, but the analysis was restricted to data post-1981 to avoid recession-era volatility and focus on modern CO₂ trends.

Handling Missing Data: Rows with missing GDP or CO₂ values (about 40% of the dataset) were removed due to the volume and prevalence across both small and large countries, leaving 6,000+ usable records. With Columns missing less than 5%, data was imputed using linear regression

Data Simplification: Composite (e.g., continental) entries and precalculated interaction terms (e.g., GDP per capita, cumulative emissions) were removed to maintain consistency and allow for later custom calculations.

Final Dataset: After cleaning and enrichment with categorical variables for 1982 income level and continent, the final dataset contained 5,859 rows with 18 variables — 16 continuous and 2 categorical.



METHODOLOGY

After cleaning the data, we are addressing the following research question:

How do environmental and demographic factors influence economic performance, as measured by GDP, when expressed in relative (percentage) terms?

We employ linear regression on log-transformed GDP to precisely quantify the percentage impact of key predictors, such as CO₂ emissions, energy consumption, and population characteristics, on economic output.

- **Model Specification**

We applied a linear regression model on the log-transformed GDP. This transformation converts absolute changes into percentage changes, allowing the regression coefficients to be directly interpreted as the approximate impact on economic performance. The model incorporates several predictors—including environmental factors (e.g., CO₂ emissions per unit energy) and demographic variables (e.g., population, income level, and continent).

- **Data Partitioning & Performance Metrics**

The dataset was divided into a training set (70%) for model development and a testing set (30%) for evaluation. Model performance was measured using key metrics such as RMSE, MAE, and R-squared, which quantify the accuracy of the log(GDP) predictions.

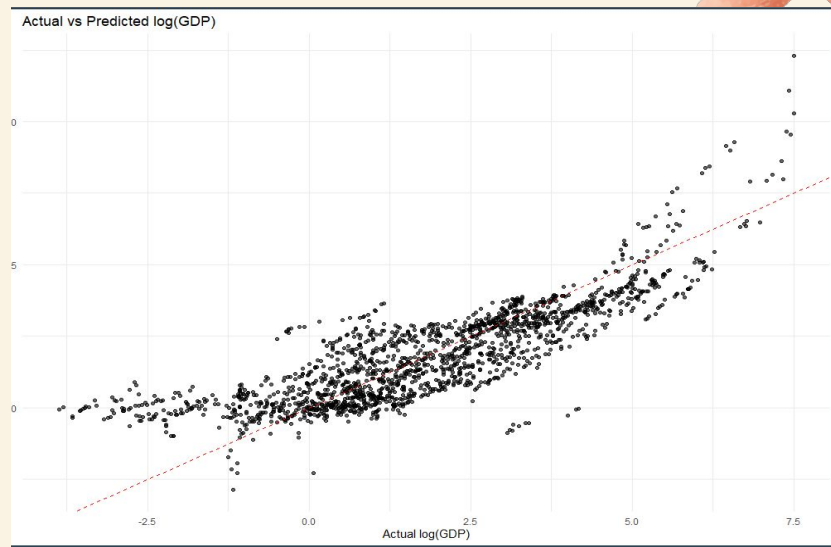
INTERPRETATION

- **Understanding Coefficient Estimates**

Due to the log transformation of GDP, the model's coefficients represent the approximate percentage change in GDP for a one-unit change in each predictor. For example, a coefficient of 0.03 indicates an approximate 3% change in GDP for a one-unit increase in the corresponding predictor. This formulation directly aligns with policy-relevant interpretations since it presents economic impacts in relative (percentage) terms rather than absolute figures.

- **Visual Validation & Predictor Significance**

The actual versus predicted log(GDP) scatter plot shows points clustering around the 45° reference line, confirming that the model predictions closely follow the observed data. Additionally, the statistical summary indicates that many predictors have significant p-values (e.g., income levels and continent indicators are highly significant), underscoring their reliable influence on economic performance. This further clarifies which environmental and demographic factors are key drivers of GDP variation.



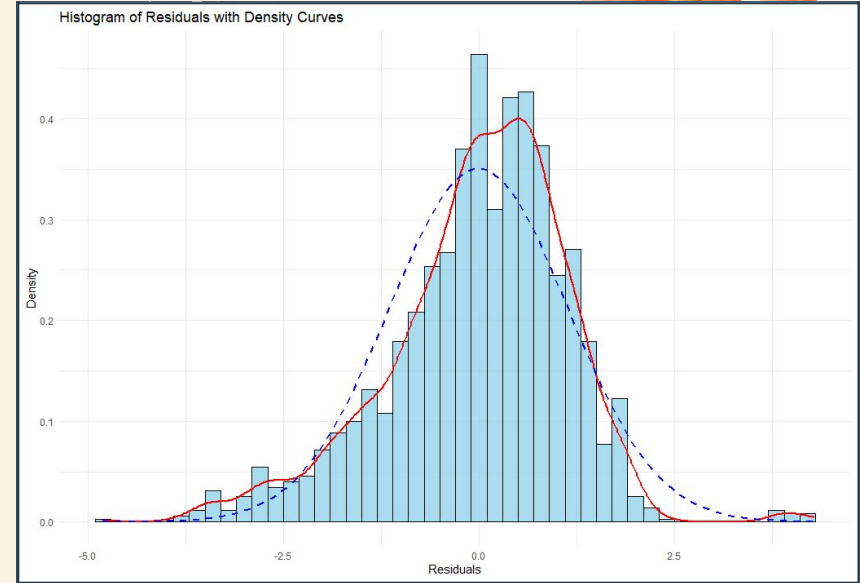
ANALYSIS

- **Residual Diagnostics & Density Curves**

We evaluated the model's assumptions by generating a histogram of the residuals, overlaid with both a kernel density estimate (to represent the empirical distribution) and a theoretical normal curve calculated from the residuals' mean and standard deviation. The close alignment between the empirical and theoretical curves indicates that the residuals are approximately normally distributed—an important assumption for the validity of linear regression.

- **Assessment of Model Fit**

The model achieved an R-squared value of approximately 0.692, suggesting that around 69% of the variability in log-transformed GDP is explained by the predictors included. Moreover, the RMSE and MAE values provide a measure of the average prediction error on the log scale, reinforcing the overall model reliability.



METHODOLOGY

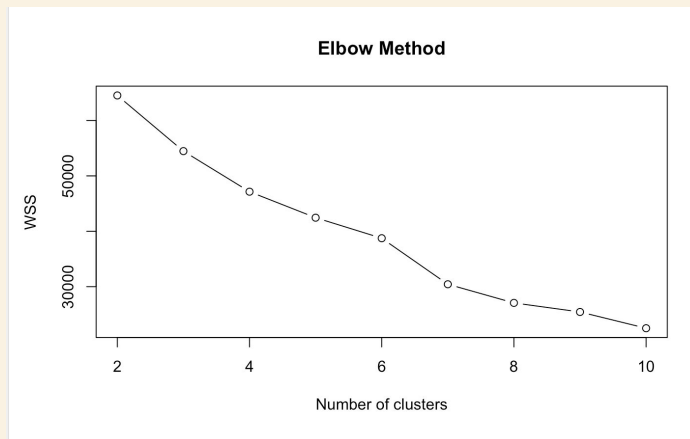
Research Question:

Do CO₂ and economic patterns cluster more clearly by income level or by continent?

Our K-means clustering analysis aimed to uncover natural groupings of countries based on their CO₂ emissions and economic indicators—without using predefined labels like income level or continent. We applied principal component analysis (PCA) to reduce high-dimensional data for visualization and evaluated the clustering performance against both **income level** and **continent**.

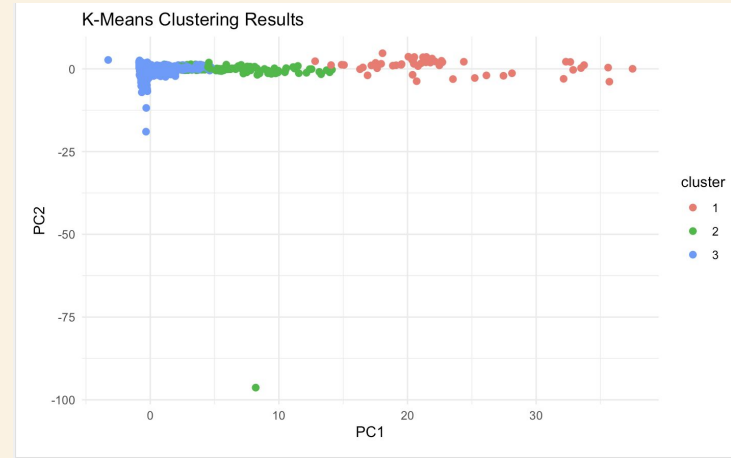
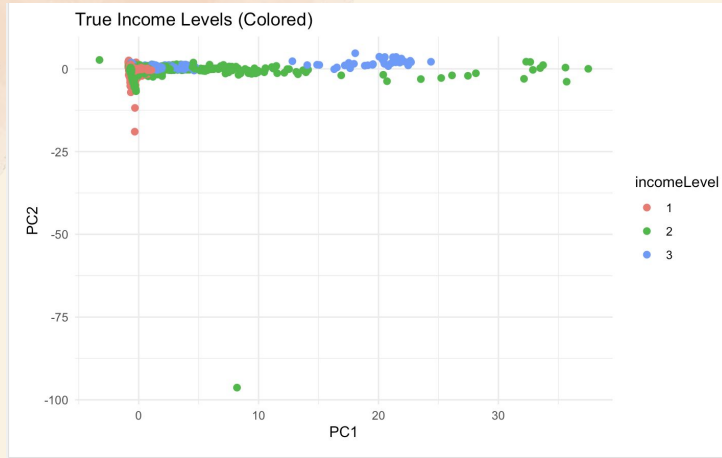
- **Optimal Clusters—Elbow Method**
Used the Elbow Method to determine the ideal number of clusters. The “elbow point” suggested **k = 3**, aligning with income level categories.
- **K-means Clustering**
Applied K-means to group countries by economic and environmental profiles. Each country was assigned to a cluster based on feature similarity.
- **PCA for Visualization**
Used PCA to reduce dimensions and plotted the first two components to visualize clusters and compare them with true income levels.
- **Performance Evaluation**
Mapped clusters to actual income groups using majority vote. Evaluated performance using a confusion matrix, with an accuracy of ~54%.

INTERPRETATION



The K-means clustering revealed that countries tend to group based on shared environmental and economic characteristics such as CO₂ emissions, energy usage, GDP, and population. While the clusters often matched known income levels, there were notable overlaps—particularly among middle-income nations. This indicates that income level alone doesn't fully reflect a country's sustainability or development profile. The clustering provides deeper insights into global patterns, uncovering hidden similarities and differences that could guide more targeted policy-making or sustainability efforts.

ANALYSIS

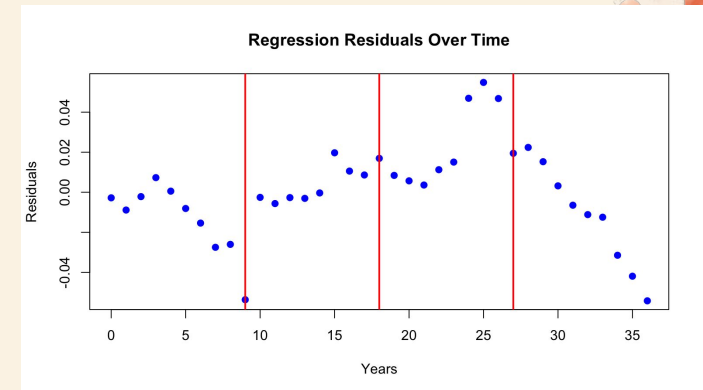
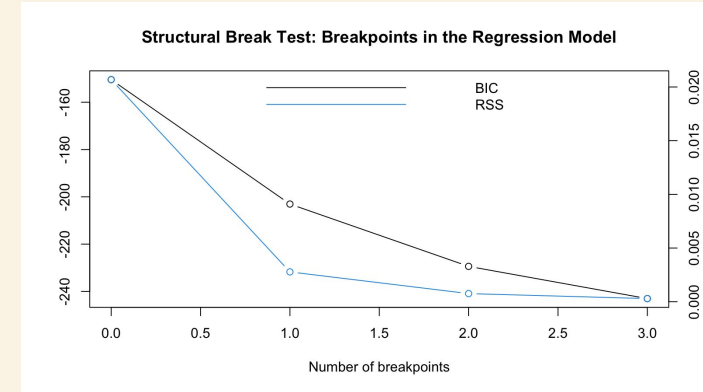


The K-means model ($k = 3$) grouped countries based on environmental and economic features. Clusters mostly aligned with income levels, especially for high-income nations. However, overlaps were seen in middle- and low-income groups. PCA plots showed moderate separation, and the model achieved ~54% accuracy. The analysis reveals that CO_2 and energy patterns offer deeper insights than income labels alone.

METHODOLOGY

Research Question:

- How have the relationships between CO₂ emissions and economic indicators changed over time since 1982, and are there identifiable breakpoints that correspond to notable global economic or policy events?
- **Aggregated Regression & Structural Break Testing:**
 - **Baseline Model:** Fit a linear regression using log(GDP) as the dependent variable with key predictors.
 - **Breakpoint Analysis:** Use the strucchange package to detect significant shifts in relationships over time.
- **Time Series Modeling (ARIMAX):**
 - **ARIMA Errors:** Capture temporal autocorrelation within the data.
 - **External Regressors:** Incorporate exogenous variables (e.g., CO₂ metrics, population) to improve forecast accuracy.
- **Segmented Analysis:**
 - **Data Segmentation:** Divide the series into regimes based on detected breakpoints.



INTERPRETATION

Series: train
Regression with ARIMA(1,0,0) errors

Coefficients:

	ar1	intercept	co2	cement_co2	land_use_change_co2	population
	0.9656	1.6996	0.0016	0.0546	0e+00	0.0361
s.e.	0.0411	0.1807	0.0005	0.0123	4e-04	0.0061

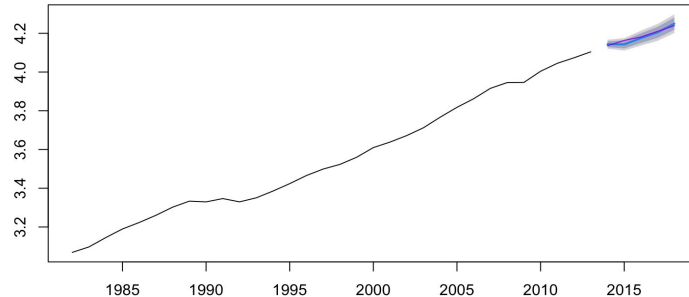
$\sigma^2 = 0.0001379$: log likelihood = 98.8

AIC=-183.59 AICc=-178.92 BIC=-173.33

Training set error measures:

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	0.00238415	0.01058431	0.008111678	0.073785	0.2274011	0.2334902	0.121173

ARIMAX Forecast (Training Data: 1982-2013; Test Data: 2014-?)



- **Accurate Forecasting:** The ARIMAX model closely tracks the observed upward trend in log(GDP) with narrow confidence intervals.
- **Temporal Dynamics:** A high autoregressive coefficient confirms strong temporal dependency, enhancing forecast reliability.
- **Regime Shifts:** Differences in the segmented forecasts illustrate distinct regimes, reflecting shifts in economic dynamics due to global events.
- **Practical Implications:** These insights support potential policy analysis and further economic investigations.

ANALYSIS

Potential Global Event Alignment

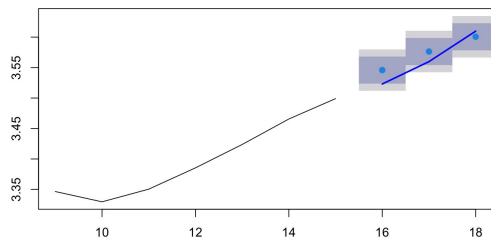
- **1991:** Fall of the Soviet Union – economic reorientation and shifts in energy policies.
- **2000:** Dot-com bubble – rapid tech growth altering global investment and emissions patterns.
- **2009:** Global financial crisis – economic recession impacting industrial output and environmental regulation.

These events likely contributed to changes in the relationship between CO₂ emissions and GDP.

Policy & Economic Implications

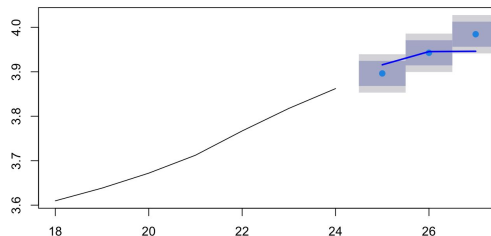
- Identified regimes can inform targeted policy decisions. For example, more stringent environmental regulations may be warranted during periods of high emission elasticity.
- The evolving dynamics suggest a need for adaptive strategies in economic planning.
- Further country-specific analysis is essential for localized policy insights.

ARIMAX Forecast for Segment 2 (9 to 18)



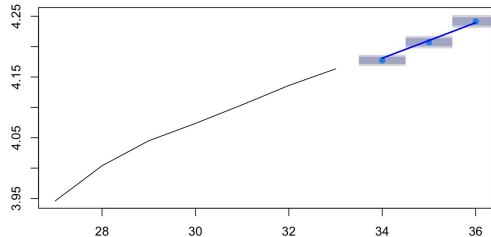
Segment 2: Transitional Regime, Moderate Accuracy

ARIMAX Forecast for Segment 3 (18 to 27)



Segment 3: "Simpler Model, High Volatility"

ARIMAX Forecast for Segment 4 (27 to 36)



Segment 4: "Stable Growth, Low Error"

CONCLUSION AND LIMITATIONS

Conclusion

Linear Regression Findings:

CO₂ emissions, energy use, and population are significant predictors of GDP, with the model providing strong interpretability through percentage-based effects.

Clustering Insights:

K-Means clustering revealed that countries do not always align with traditional income groups, especially in the middle-income range, suggesting more nuanced development patterns.

Time Series Observations:

Structural break analysis identified key shifts in the CO₂-GDP relationship over time, with later periods showing stronger and more stable associations.

Limitations

Modeling

Linear regression assumes linearity, homoscedasticity, and normality of residuals, which may not hold consistently across all countries or variables.

Assumptions:

Clustering

K-Means assumes spherical clusters and is sensitive to initialization and scaling, which may limit the accuracy of cluster interpretation.

Constraints:

Aggregation

&

Interpretability:

Time series analysis on aggregated global data may mask country-level differences, and while breakpoints show **when** changes happen, they don't explain **why** without further context.

REFERENCES

- The World Bank annual report 1982 (English). Washington, D.C. : World Bank Group.
<http://documents.worldbank.org/curated/en/458551468765615887>
- G. C. Chow, "Tests of equality between sets of coefficients in two linear regressions," *Econometrica*, vol. 28, no. 3, pp. 591–605, Jul. 1960.
- A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, Jun. 2010.
- M. H. Pesaran, Y. Shin, and R. J. Smith, "Bounds testing approaches to the analysis of level relationships," *Journal of Applied Econometrics*, vol. 16, no. 3, pp. 289–326, May–Jun. 2001
- A. Banerjee, R. L. Lumsdaine, and J. H. Stock, "Recursive and sequential tests of the unit-root and trend-break hypotheses: Theory and international evidence," *Journal of Business and Economic Statistics*, vol. 10, no. 3, pp. 271–287, Jul. 1992.
- R. J. Hyndman and E. Wang, "Characteristic-based clustering for time series data," *Data Mining and Knowledge Discovery*, vol. 13, no. 3, pp. 335–364, Nov. 2006.
- H. Kim and Y. Choi, "Linear regression analysis of energy consumption in wireless sensor networks," *IEEE Communications Letters*, vol. 12, no. 4, pp. 234–236, Apr. 2008.