Data Science Replication Study

Team A

Data Science for Business

Team A:

Mohamad Abdulla

Anvith Amin

Manjunath Mallikarjun Kendhuli

Papers Reviewed

- Paper 1: Predicting Employee Attrition (IBM)
- Paper 2: Data Analytics for Optimizing and Predicting Employee Performance
- Paper 3: Migration and Innovation: Learning from Patent and Inventor Data
- Challenges faced during the project

Selected Paper

"The Political Economy of Green Industrial Policy"

Juhász et al., 2022

- Used Global Trade Alert (GTA) database
- Three key figures showing green policy trends in G20 countries

---

Problems Faced

- Unclear objectives at the beginning
- Extremely large and complex datasets
- GitHub deployment issues

Replication of Figure 1

- **Title:** Green Industrial Policy Activity in G20 Countries (2010–2022)
- **What it shows:**
  - Annual green policy activity for Middle-income vs. High-income countries
  - Indexed to 2010–2012 average = 100
  - High-income line is scaled (divided by 5) for visual comparison
- **Axes:**
  - Left Y-axis: Middle-income index
  - Right Y-axis: High-income index (scaled)

Replication Figure

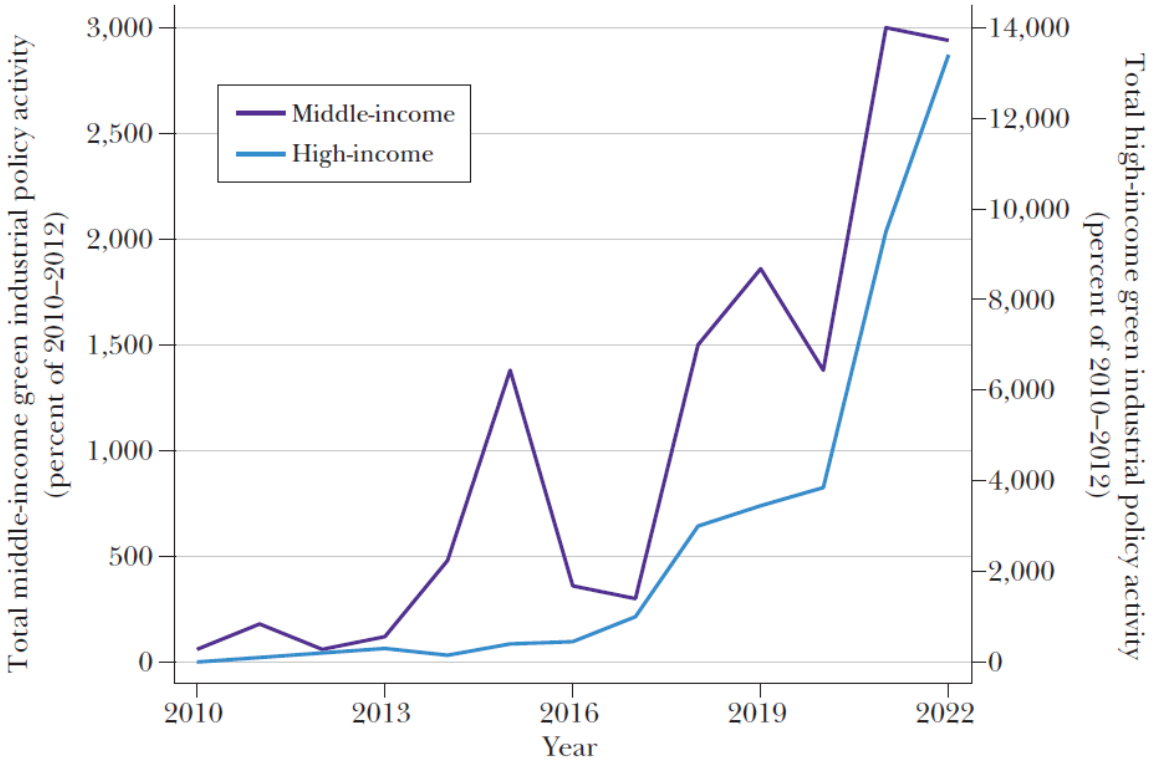

Figure 1: Fig 1: Green Industrial Policy Activity in G20 Countries, 2010–2022

Code Logic Summary

- **Step 1: Load and clean raw data**
  - Import original `IP_G20.dta` file
  - Filter valid rows and deduplicate by MeasureID–Year–Country
- **Step 2: Identify green policies**
  - Use keywords like *climate, emission, renewable* to flag green measures
- **Step 3: Add income group classification**

  - Load World Bank Excel data
  - Reshape to long format and convert fiscal to calendar years
  - Merge with green policy data by country and year

- **Step 4: Standardize income group labels**
  - Map `H` to "High-income", `LM/UM` to "Middle-income"
  - Remove unmatched or missing classifications
- **Step 5: Count policies per year**
  - Group by year and income group
  - Count number of green policies announced
- **Step 6: Compute 2010–2012 baseline**
  - Calculate average policy count in 2010–2012 for each group
- **Step 7: Index calculation**
  - Create index: `(policy_count / baseline_avg) * 100`
  - Expresses annual activity relative to baseline (baseline = 100)

- **Step 8: Visualization**
  - Plot both income groups on one chart
  - Scale high-income index by `/5` on secondary Y-axis for comparison

---

R Code for Replication

Output of Figure 1 Replication

Replication of Figure 2

- **Title:** Top Five Green Industrial Policy Instruments across G20 Economies by Income Group (2010–2022)
- **What it shows:**
  - Distribution of green industrial policies by instrument type (e.g. financial grant, state loan) -Comparison between High-income and Middle-income G20 countries -Focuses only on the top five most frequent instruments within each group -Measures are shown as shares of total green policy activity, normalized within each group
- **Axes:**
  - Left Y-axis: Income group (*High-income / Middle-income*)
  - Right Y-axis: Share of green policies by instrument type

Replication Figure

Figure 2
Top Five Green Industrial Policy Instruments across G20 Economies by Income Group, 2010–2022

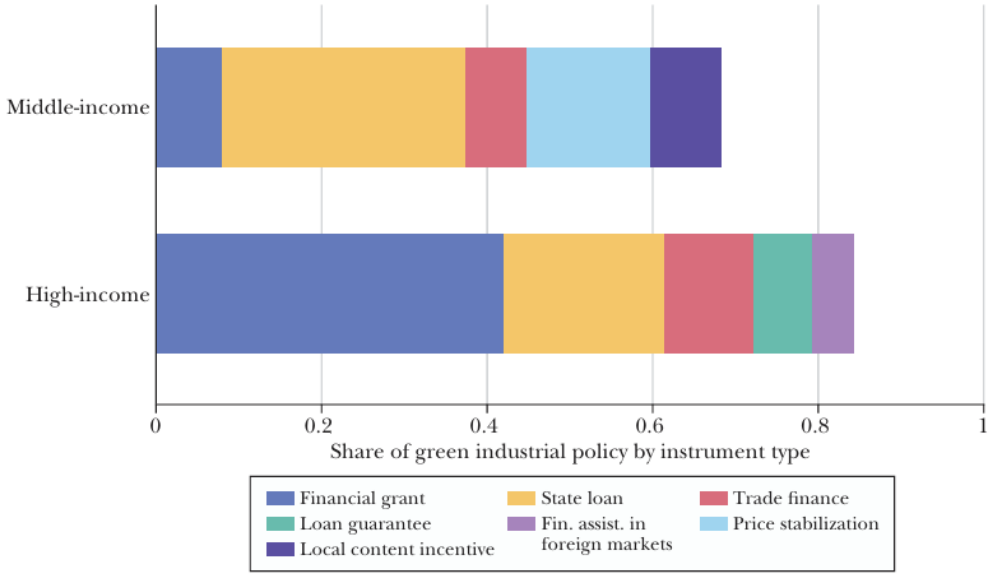

Figure 2: Figure 2: Top five industrial policy intruments

Code Logic Summary – Figure 2

- **Step 1: Load and prepare data**
  - Import `IP_G20.dta` (policy dataset) and `wb.xlsx` (income classification)
  - Standardize column names using `clean_names()`
- **Step 2: Filter relevant policies**
  - Keep policies from 2010–2022 with non-missing descriptions
  - These represent green or environmentally relevant measures

---

- **Step 3: Assign income group**
  - Use a fixed list to classify countries as *High-income* or *Middle-income*
  - Add this classification to each policy record
- **Step 4: Identify top 5 policy instruments**
  - Count frequency of each policy tool (`measure_type`)
  - Select the top 5 most common types separately for each income group
- **Step 5: Compute usage shares**
  - Within each group, calculate how much each of the top 5 instruments was used
  - Expressed as a share of total green policies in that group (0.0 to 1.0)

- **Step 6: Visualize with stacked bar chart**
  - Plot horizontal bars showing instrument composition by income group
  - Use `coord_flip()` to flip axes and `number_format()` to show decimals
- **Step 7: Display output**
  - Render the plot with minimal styling and a grouped color legend

---

R Code for Replication

Output of Figure 2 Replication

Challenges faces

- Uploaded local data files to GitHub and linked using raw URLs so others could run the code.
- Replaced `percent_format()` with `number_format()` to show axis labels as decimals.
- Renamed output to `index.html` so GitHub Pages would display the updated version.

Future Work with this Replication

- **Clean and verify all country names**
  - Match them correctly with World Bank data to fix missing values
- **Improve keyword filtering**
  - Refine how we identify green policies using better or more complete keywords
- **Match Stata version exactly**
  - Compare our R code outputs with the original Stata graphs for full accuracy

---

- **Check missing data**
  - Investigate why some years or countries have fewer policies than expected
- **Automate income group assignment**
  - Instead of manual grouping, use official classification files from the World Bank