Data Science Replication Study

Team A

Table of contents

1 Data Science for Business

1.1 Team A:

1.1.1 Mohamad Abdulla

1.1.2 Anvith Amin

1.1.3 Manjunath Mallikarjun Kendhuli

---

2 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

---

3 Selected Paper

3.1 "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

---

4 Problems Faced

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

---

5 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)

---

6 Replication Figure

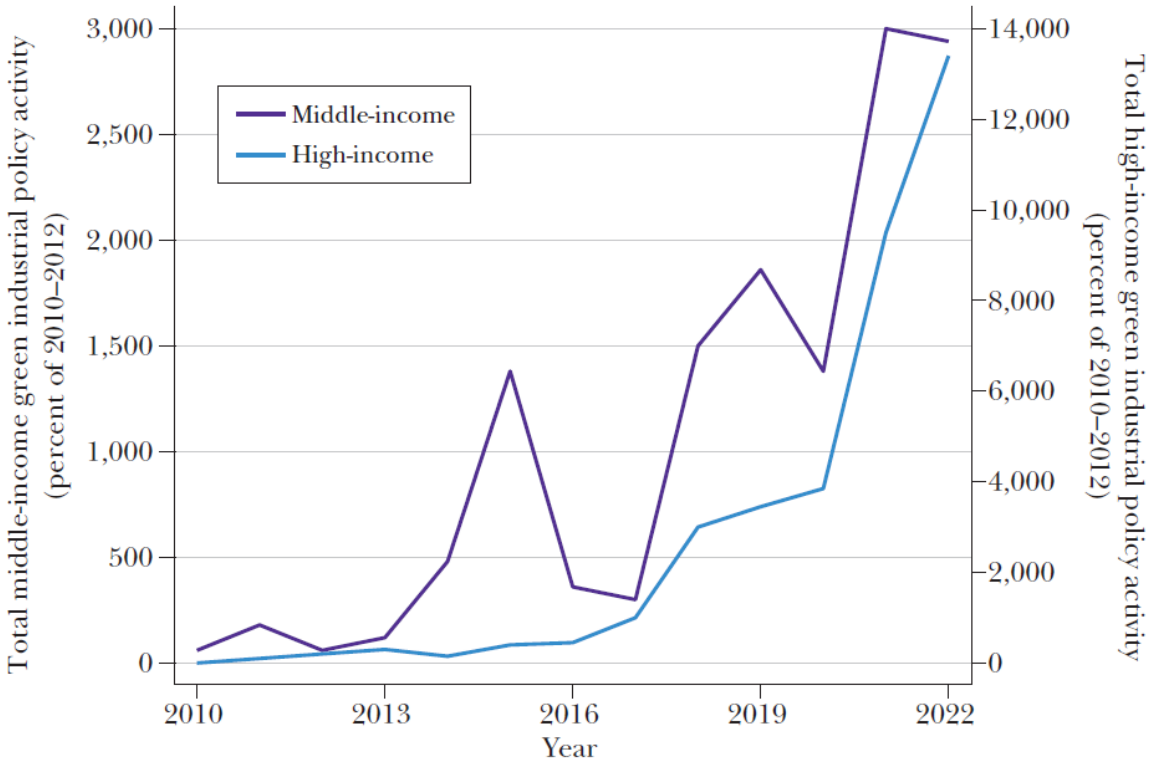

Figure 1: Fig 1: Green Industrial Policy Activity in G20 Countries, 2010–2022

---

7 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

---

8 R Code for Replication