**Data code for: “Evaluating the Economic Impacts of the G20 Compact Initiative: Evidence from Causal Inference Using Advanced Machine Learning Techniques”**

Tosin Kolajo Gbadegesin and Nadege Desiree Yameogo

**Data and Code Availability Statement**

The paper uses public and non-confidential data on control of corruption, political stability and absence of violence or terrorism, regulatory quality, rule of law, voice and accountability, and government effectiveness published by the World Bank Worldwide Governance Indicator (WGI). <https://www.worldbank.org/en/publication/worldwide-governance-indicators>.

The paper uses public and non-confidential data on trade per cent of GDP), exports (current USD), imports (current USD), net trade, population (gross), GDP growth rate, net FDI inflows (BoP, current US$), FDI (% of GDP), Consumer Prices, exchange rate, access to electricity, net official development assistance and their relation to GDP (NetODA and NetODAGDP), lending rate, domestic credit to the private sector, government expenditure (percentage of GDP), GDP deflator, GDP per capita (constant US dollars), gross domestic savings (percentage of GDP), gross capital formation (percentage of GDP), natural resource rent, and GDP per capita (constant US dollars) published by World Bank Development indicator. <https://databank.worldbank.org/source/world-development-indicators>

The paper uses public and non-confidential data on inflation (consumer prices per cent), urban population (percentage of the total population), and secondary school enrolment (per cent of the total population) published by the African Development Bank’s open data for Africa. Employment-to-population ratio data is obtained from ILO while that of FDI Inflows (current prices US$) is obtained from the United Nations Conference on Trade and Development. <https://dataportal.opendataforafrica.org/data/?source=AfDB> / <https://ilostat.ilo.org/data/>

**Computational Requirements**

**Software and Hardware Requirements**

* **Software**:
  + **Python**: We used Jupyter Notebook from Anaconda Distribution 2023.03. This version uses the base environment Python 3.10 and offers Python 3.8 and Python 3.9 support. It also includes conda 23.1.01 and Anaconda Navigator 2.4.01.
  + **Packages**: Numerous packages are required, and they are all specified at the beginning of each code script.
* **OS**:
  + **Windows 11 Pro**: This was our primary operating system. Other versions of Windows, as well as Mac and Linux, should be compatible.
* **Hardware**:
  + **Model:** Dell Latitude 7420
  + **CPU**: 11th Gen Intel(R) Core(TM) i5-1145G7 @ 2.60GHz 1.50 GHz
  + 16.0 GB / 64-bit operating system, x64-based processor

## Downloading and opening the replication files

Link to the data and replicate code is available at: https://github.com/TosinSDGs/CwA-Impact-Study. CwA Analysis.ipynb contains the code for reproducing the tables and figures, CwA Analysis Robustness.ipynb contain the code for performing the robustness analysis and CwA Final Data.xlsx is the data used.

## Data preparation and reformatting

For the full details regarding the construction of all the specific variables used in the analysis, please refer to the paper.

After assembling the data, the initial steps focused on preparing the dataset by creating necessary dummy variables. A new column, CwACountries, was added to indicate whether a country was part of the Compact with Africa (CwA) group, and this was converted into a 0 or 1 dummy variable. Additionally, a CovidDummy column was created to represent the years 2021 and 2022, capturing the impact of the COVID-19 pandemic. For missing data, zeros in specified columns were replaced with NaN, and mean imputation was applied within each country group to ensure completeness. Descriptive statistics were then generated to provide an overview of key economic indicators and assess data completeness, including the analysis of missing values.

**Control Group Selection**

The process of selecting control groups for each key outcome variable involved several steps to ensure robust and comparable analysis. First, logistic regression was used to calculate propensity scores based on pre-treatment data, estimating the likelihood of CwA participation. These scores were integrated into the dataset, and the Nearest Neighbors algorithm identified control countries with the closest propensity scores to the treated countries. This procedure was repeated for all outcome variables to maintain consistency.

After matching, subset DataFrames were created for each treated and control country pair, allowing for focused analysis of treatment effects on specific variables. These subsets were then combined to increase the number of observations, enhancing the statistical power of the analysis. Following this, the data for each key outcome variable was standardized, excluding certain identifiers and categorical variables, to ensure all continuous variables were on a comparable scale. The standardized data was combined with the original categorical data to create a comprehensive dataset for further analysis.

Next, the combined DataFrames were used to calculate the Standardized Mean Difference (SMD) for propensity scores, a key metric for assessing the balance between treated and control groups. The SMD results were compiled into a comprehensive data frame and visualized with a bar plot to represent the balance achieved across outcome variables. This entire process was systematically repeated for all key outcome variables to ensure thoroughness and consistency in the analysis.

**Figure 1**

To replicate Figure 1, create customized visualizations to compare key outcome variables between CwA and non-CwA countries over time. Map each variable to its appropriate title and unit, and plot them in a grid of subplots for side-by-side comparison. Display data for CwA countries with solid blue lines and non-CwA countries with dashed orange lines, highlighting the CwA initiative period (2017-2022) in red. Add legends and labels for clarity, and save the final visualizations as a PNG file named comparison\_metrics1.png.

**Figure 2**

After completing the standardization process, calculate the Standardized Mean Difference (SMD) for the propensity scores across all key outcome variables to assess the balance between treated and control groups. Use a function to compute the SMD for each outcome variable, comparing the mean differences between treated and control groups, adjusted by their pooled standard deviation. Compile these results into a data frame for clear visualization, then generate a bar plot to display the SMD for each outcome variable, with a threshold line indicating acceptable levels of imbalance. Save the plot as a PNG file for documentation.

**Tables 1 and 2**

To replicate Tables 1 and 2, use Targeted Maximum Likelihood Estimation (TMLE) to estimate the treatment effects of the CwA initiative on key outcome variables, employing machine learning models such as Gradient Boosting, Random Forest, and XGBoost. Begin by selecting these models to account for data complexity and potential confounding factors. Apply the TMLE method for each model by fitting both outcome and treatment models, updating predicted outcomes using the clever covariate approach, and calculating the treatment effect. To ensure robustness, conduct a bootstrap procedure with 1,000 iterations, resampling the data and recalculating the treatment effect for each sample. Assess model performance using metrics like R-squared, RMSE, and MAE. Compile the bootstrap analysis results, including mean treatment effects, standard errors, confidence intervals, and other metrics, into a summary table. Repeat this process systematically for all key outcome variables to ensure consistency and accuracy in the findings presented in Tables 1 and 2.

**Figure 3**

To replicate Figure 3, generate a series of subplots to visually represent the estimated treatment effects of the CwA initiative across key outcome variables: Net FDI Inflow, FDI Inflow, Gross Capital Formation, Employment, GDP per Capita, and Export. Assign each outcome variable to a specific subplot within a 2x3 grid, and plot the treatment effects alongside 95% confidence intervals to show the precision of the estimates. Include a horizontal line at zero to indicate the null effect. Organize the final layout for clarity and save it as a PNG file titled combined\_treatment\_effects\_plot.png.

**Figure 4**

To replicate Figure 4, generate residual plots to evaluate the performance of the models used to estimate treatment effects for key outcome variables: Net FDI Inflow, FDI Inflow, Gross Capital Formation, Employment, GDP per Capita, and Export. For each outcome variable, plot the residuals (the difference between observed and predicted values) against the predicted values for different models. Arrange these plots in a 3x2 grid, with each subplot representing one of the outcome variables. Include a horizontal line at zero to indicate where residuals equal zero, which helps in assessing model accuracy and bias. Organize the final layout for clarity and save it as a PNG file titled residual\_plots\_combined.png.