README:
"Global inequality of opportunity" replication

Introduction

The replication consists of the following files:

Code files (4):

- "1_pure_replication.R"
- "2_cross-continent_analysis.R"
- "3_income_importance_for_brazilian_migration.R"
- "Plots.R"

Data files (13):

- "final08_1.RData"
- "BRA_expats.xlsx"
- "country_mapping.RData"
- "languages.csv"
- "forms.csv"
- "parameters.csv"
- "Brazil_Immigration_1884-1953.csv"
- "DEMIG_VISA_Database_version_1.4.xlsx"
- "IMPICDatasetV2_1980-2018.dta"
- "undesa_pd_2024_ims_stock_by_sex_destination_and_origin.xlsx"
- "Coup_data_2.2.0.csv"
- "WYD_reg.xlsx"

- “replication_results_part1.RData”

All code and data files are used across four stages, which can be run separately, independently, and can be found at the following github repository: Milanovic-2015 replication

The original data set, as well as Stata codes, are available at dataverse.harvard.edu.

1 Pure Replication

1.1 Data

The data file is called “final08_1.RData”. It is a R file which contains all the variables needed to run the regressions and get the results reported in the paper.

1.1.1 Variables:

- **contcod** = country three-letter World Bank acronym.
- **group** = income percentile (running from 1=poorest percentile to 100=richest percentile). A percentile contains 1% of a country's population.
- **maxgroup** = total number of groups for a country. All but one country have 100 groups, that is 100 percentiles.
- **inc** = real annual per capita income of a given country/percentile in the year 2008. The values are in 2005 international (ppp) dollars. Calculated from countries' household surveys conducted in the year given by the variable survey_year. If survey year is not 2008, the values are brought into 2008 values by the use of country's Consumer Price Index.
- **survey_year** = year in which household survey was conducted. (About 2/3 of surveys are conducted in the benchmark year 2008; almost 90% of surveys are conducted in the window 2007-09.)
- **year** = year into which all the variables are “benchmarked” (2008).
- **lninc** = ln(inc).
- **gdpppp** = GDP per capita for the year 2008; expressed in 2005 international (ppp) dollars. The data are from World Bank World Development Indicators.
- **lngdppp** = ln(gdpppp).
- **pop** = population size (in million) of each country/percentile in the year 2008. The data are from World Bank World Development Indicators.
- **gini** = country's Gini coefficient.
- **ayos** = average number of years of education if age>15, by country; from 2012 World Bank World Development Indicators.

All the variables are original to the paper (WYD).

All the other variables used in the paper are created from these variables.

1.2 Code

The code used is provided in “1_pure_replication.R”.

All the R packages needed to run the code are available on CRAN and should be automatically downloaded and loaded by simply running the code.

Finally, the results obtained by running the code.

1.3 Final Note

In order to run the replication, the user needs to have downloaded the comprehensive dataset used (Branko’s WYD), available under “Datasets/final08_1.RData” as well as the code file, available under “R_scripts/1_pure_replication.R”.

2 Cross-Continent Analysis

2.1 Data

The data file is named "final08_1.RData". This section contains no regressions or statistical models; its sole purpose is to analyze the data grouped by continent.

2.2 Code

The code used is provided in "2_cross-continent_analysis.R".

All required R packages are available on CRAN and will be installed and loaded automatically when the code is run.

Finally, the results obtained by running the code (output files include "WYD_reg.xlsx" and "WYD_cont.xlsx").

2.2.1 Variables

All variables are the same as those considered in Part 1.

2.3 Final Note

The user needs to have only two files in order to replicate all the results from this part. The data file ("Datasets/final08_1.RData") and the code file ("R_scripts/2_cross-continent_analysis.R")

3 Income Importance for Brazilian Migration

3.1 Data

This section consists of organizing the final dataset used to feed the gravity model. Because the data were gathered from multiple sources, several files were used. The data files for this part include:

- "BRA_expats.xlsx": number of Brazilian expats living in each country (Brazilian Ministry of Foreing Affairs);
- "country_mapping.RData": correction of country names between WYD and the remaining data sets utilized;
- "languages.csv", "forms.csv" & "parameters.csv": ASJP is a large cross-linguistic wordlist resource. The latest release is distributed via Zenodo, and it provides downloadable wordlists suitable for distance/proximity measures (Zenodo);
- "Brazil_Immigration_1884-1953.csv": number of migrants, per nationality and year, that arrived in Brazil between 1884 and 1953 (Brazilian Institute of Geography and Statistics);
- "DEMIG_VISA_Database_version_1.4.xlsx": visa and exit permit requirements of 214 countries for travellers of 237 countries over four decades (International Migration Institute);
- "IMPICDatasetV2_1980-2018.dta": quantitative indices to measure immigration policies in most OECD countries and for the time period 1980-2018 (Immigration Policies in Comparison);
- "undesa_pd_2024_ims_stock_by_sex_destination_and_origin.xlsx": estimates of the total number of international migrants by sex, as well as their places of origin and destination, for 233 countries and areas(UN International Migrant Stock);
- "Coup_data_2.2.0.csv": outcomes of coup events (i.e., realized, unrealized, or conspiracy), the type of actor(s) who initiated the coup (i.e., military, rebels, etc.), as well as the fate of the deposed leader (Cline Center Coup d'État Project Dataset);
- "WYD_reg.xlsx": original Branko's WYD data set with the addition of the variable "reg", indicating the region/continent of each country listed;
- "gravity_df.RData": used in the final gravity PPML model (combination of all other data sets gathered for such part).

3.1.1 Variables:

The final dataset used for the regression comprehends the following variables:

- **contcod** = country three-letter World Bank acronym.

- **cont** = country's official name.
- **reg** = country's region/continent.
- **expats** = number of Brazilian expats legally living in the country.
- **language_dist** = linguistic distance index from Brazilian-Portuguese (PT-BR).
- **cult_dist** = cultural proximity index based on foreign migration waves to Brazil (between 1884 and 1953).
- **geo_dist** = geographical-geodesic distance from Brazil (by distance to closest border/shoreline). The data are from the R package *rnaturalearth*.
- **visa** = baseline mobility friction (based on entry visa requirement).
- **mig_policy** = general policy restrictiveness at destination/country with respect to migration.
- **ln_diaspora** = Brazil-born stock in that country prior to 2008 (network effects → the idea is that a bigger Brazilian diaspora in a given country lowers costs such as general information, housing, job search, etc).
- **n_coups** = number of realized and attempted coups/conspiracies in the country (proxy for political instability).
- **ln_inc** = ln(inc).
- **gini** = country's Gini coefficient.
- **pop** = population size (in million) of each country in the year 2008.
- **ln_dist** = ln(geo_dist).

3.2 Code

The code used is provided in "3_income_importance_for_brazilian_migration.R"

All the R packages needed to run the code are available on CRAN and should be automatically downloaded and loaded by simply running the code.

Finally, the results obtained by running the code.

3.3 Final Note

The third part is the longest to complete, as it documents the entire data-collection process.

A more direct approach is also possible: the dataset used in the final model is available at "Datasets/gravity_df.RData". Thus, loading that file together with the script at "R_scripts/3_income_importance_for_brazilian_migration.RData" is sufficient to replicate the outcomes of Part 3.

4 Plots

4.1 Data

The data file is named “replication_results_part1.RData”. This section contains no regressions or statistical models; its sole purpose is to replicate the plots utilized throughout the project.

4.2 Code

The code used is provided in “Plots.R”

All the R packages needed to run the code are available on CRAN and should be automatically downloaded and loaded by simply running the code.

Finally, the results obtained by running the code are presented. All plots are automatically saved using the *ggsave* function after each chunk is executed.

4.3 Final Note

The user needs to have only two files in order to replicate all the results from this part. The data file (“Datasets/replication_results_part1.RData”) and the code file (“R_scripts/Plots.R”).