Step-by-step guide to the Findex code

Overview

This repository includes instructions and code to replicate figures in the 2025 Global Findex using R. All data in the report can be found using the country-level data. We use proprietary data from Gallup.

Data Availability

We will publish our data in the microdata library and also publish the country/region/demographic-level averages after the report launch.

Raw data from Gallup CANNOT be made publicly available.

Data Sources

1. Filename: World_Bank_032125.csv (currently compressed)
    a. Source: Global Findex: https://www.worldbank.org/en/publication/globalfindex
    b. NOT publicly available
2. Filename: Findex Module Mapping 2024.xlsx
    a. Source: Global Findex: https://www.worldbank.org/en/publication/globalfindex
    b. Not publicly available
3. Filename: strataharmonize_11202012.csv
    a. Source: Global Findex: https://www.worldbank.org/en/publication/globalfindex
    b. Not publicly available
4. Filename: Final WDI.xlsx
    a. World development indicators database accessed via API last on February 17, 2025
    b. Indicators accessed:
        i. Region
        ii. Income

Official Use Only

iii. "SP.POP.TOTL", # Total population

iv. "SP.POP.GROW", # Population growth (annual %)

v. "SP.URB.TOTL", # Urban population

vi. "SP.RUR.TOTL", # Rural population

vii. "SP.POP.1564.TO", # Population ages 15-64, total

viii. "SP.POP.65UP.TO", # Population ages 65 and above, total

ix. "SP.POP.TOTL.FE.IN", # Female population

x. "SP.POP.TOTL.MA.IN" # Male population

c. Regions excluding high income were created by changing region to “high income” if income group is “high income”

d. Lagged population values were created because 2024 population was not available.

e. Data for the economy of Taiwan, China is not available via the API and was acquired from WDI team: wdi@worldbank.org

Statement about rights

I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package.

Instructions for Replicators

1. Execution_File.Rmd

a. Please specify file paths at top of document

b. All necessary packages that aren’t loaded should be loaded by first code chunk

c. The final code chunk is where the figures are created. You will need to specify the chapter for which you are making the figure here and where you would like to save the figure output.

i. You will have to do this once for each chapter, or write a loop.

Requirements

Software Requirements

R version 4.4.1

Memory and Runtime and Storage Requirements

Code runs in ~30 minutes using external server – 100GB, do not try to run on a normal computer.

Code Description

1. Execution_file.Rmd runs all code
2. Additional files:

a. Prereg_Masterdata_2024.R

i. This loads the raw data from Gallop, standardizes country names, creates years based off of waves, creates a cross-country, cross-year weight, among other things

b. Identifiers.R

i. For the 2024 data, this file creates identifiers that will be used to show whether a given respondent/country was asked certain question – this helps to avoid false "0s" when creating averages.

c. AD_14_17_21_22.R

i. "Account Definitions" – this file defines account ownership across the years and includes marginal additions (when someone answers "yes" to receiving certain payments into an account they are coded as having an account)

d. AD_2024.R

i. Like above, this is how account ownership in the latest wave is defined

e. TS_Summary.R – this file takes the cleaned Gallup data, creates binary variables for questions, and then creates country, region, and demographic-level averages.

2. Code that creates binary variables:

1. Findex_Questionnaire.R
2. ID_Questionnaire.R

3. Connectivity_Questionnaire.R
4. Constructed_Variables.R

a. This code goes a step beyond just translating the questionnaire into a binary and constructs variables using more than one input. For example, adults who don't have an account but have a phone.

5. Time_Series_Variables.R

a. This code creates the time-series variables that we use for our dataset. Over the years similar questions have different codes so we have to explicitly code the time series separately from the latest 2024 data.

b. Last chunk of Execution_file.Rmd will run the R scripts to make the figures

Folder Structure

1. Data

a. Cleaning

i. Prereg_Masterdata_2024.R
ii. Identifiers.R
iii. AD_14_17_21_22.R
iv. AD_2024.R
v. TS_Summary.R
vi. Findex_Questionnaire.R
vii. ID_Questionnaire.R
viii. Connectivity_Questionnaire.R
ix. Constructed_Variables.R
x. Time_Series_Variables.R

b. Figures (see table above)

c. Execution_file.Rmd