Global poverty estimation using private and public sector big data sources

2024-02-16

This package contains code to replicate analysis of the paper. This package incorporates datasets from public domains. Due to redistribution restrictions on some datasets, users must download these directly from the original sources. There are comprehensive instructions for downloading these datasets in the package's README file and the above data entries. This includes URLs, necessary access protocols, and a data folder structure essential for the package's functionality once the appropriate data files are placed. The `_main.R` script runs or points to all code that needs to be run for the analysis. `_main.R` is initially set up so that it loads the analysis-ready datasets, runs analysis, and creates all figures and tables; it skips creating the analysis-ready datasets from the raw data, which is a more cumbersome process.

However, because of data sharing restrictions some files need to be downloaded directly by the user. By changing a parameter, `_main.R` will also run scripts to process raw data to create the analysis-ready datasets. However, (1) raw data from a number of sources must be manually downloaded and (2) a few scripts are run in Stata and Python; `_main.R` points to these scripts, but they must be manually opened and run. All the instructions are provided below.

This replication package is divided into two parts:

1. Create Analysis-Ready Datasets from Raw Data. Contains instructions for manually downloading raw data, and steps for running code. While the `_main.R` script helps to automate running relevant scripts, code from other software (Stata and Python) must be manually run.
2. Replicating analysis, starting from analysis-ready datasets. Once de analysis-ready datasets are produced, the `_main.R` script can be set to easily run all analysis code.

Create Analysis-Ready Datasets from Raw Data

Download raw data

The data directory present in the package contains some of the analysis-ready data files, as well as folders where raw data must be placed. The `Data` folder contains sub-folders for each dataset. The sub-folders generally contain a `RawData` folder for raw data and a `FinalData` folder for data processed by code. In many cases, code is used to automatically download data. However, the following datasets need to be manually downloaded:

1. **Global DHS Data:** Download data from the DHS website to be put in `Data/DHS/RawData`; this directory contains folders that indicate which datasets need to be downloaded. For example, 2020 data for Kenya for the "HR" (Household

Recode) dataset should be placed here: `/KE/KE_2020_MIS_03292022_2054_82518/KEHR81DT`.

2. **Nigeria DHS Data:** The paper includes specific analysis for Nigeria. Following a similar process as above, data should be placed in `Data/DHS_nga_policy_experiment/RawData`.
3. **LSMS Data:** Download LSMS data from the World Bank Microdata catalogue. `Data/LSMS/RawData/individual_files` includes a folder for each country. Within each country folder, there is a `README Files to Download.md` file which lists the individual datasets that need to be downloaded into the folder.
4. **Harmonized Nighttime Lights:** Download data from here and place in `Data/DMSPOLS_VIIRS_Harmonized/RawData`.
5. **ESA Land Cover Data:** Download data from here and place in `Data/Globcover/RawData`.

- For 1992 to 2015 data, put the `ESACCI-LC-L4-LCCS-Map-300m-P1Y-1992_2015-v2.0.7.tif` file in the `/1992_2015_data` folder
- For 2016 to 2018 data, (1) put the `.nc` files in the `2016_2018` folder, then (2) use this script to convert .nc files to .tif files.

6. **OpenStreetMap Data:** Download data from Geofabrik. To find data for a specific country, (1) click the continent the country is in, (2) click the name of the country, (3) click "raw directory index", (4) and find the relevant date to download; the file that ends in `shp.zip` should be downloaded. Download the file and unzip it. Place the file in the relevant folder within `Data/OSM/RawData`; this folder contains subfolders for each country and year where OpenStreetMap data needs to be downloaded. For example, the data downloaded and unzipped from `kenya-210101-free.shp.zip` should be placed in `Data/OSM/RawData/kenya-210101-free.shp`
7. **Sentinel 5P Pollution Data:** Run this code in the Google Earth Engine code editor, and put the data in `Data/Sentinel 5P Pollution/RawData`

Setup and run code

1. Download this package.
2. In `_main.R`, change `dropbox_dir` to point to the data folder and `github_dir` to point to the github repo.
3. Create a folder in Google Drive, mount Google Drive to your computer, and change `gdrive_dir` to point to this folder. Code to download satellite imagery from Google Earth Engine requires a Google Drive folder; data from GEE is exported to Google Drive.
4. In `_main.R`, ensure that `RUN_DATA_CREATION_CODE` is set to `TRUE` and `RUN_ANALYSIS_CODE` is set to `FALSE`. When `RUN_DATA_CREATION_CODE` is `TRUE`, code from the following sub-folders are run:

- `00_download_gadm`: Downloads GADM data that is used in cleaning survey data.
- `01_clean_dhs`: Cleans DHS survey data.

- `01_clean_dhs_nga_experiment`: Cleans DHS survey data for Nigeria, using four rounds of data (used for **Application: Estimating Wealth in Different Years** section of paper)
- `01_clean_lsms`: Cleans LSMS survey data.
- `02_get_process_ancillary_data`: Extracts and process data around survey locations, such as from satellites, OpenStreetMaps, and Facebook Marketing data.
- `03_merge_ancillary_data_with_survey`: Merges ancillary data (satellite, OSM, Facebook data) extracted in previous step to survey data; creates cleaned, analysis-ready datasets.

5. Run the code. Running the `_main.R` script will run all R files. However, instead of running the `_main.R` script, we recommend running files one-by-one as scripts in Python and Stata need to be run as well; the `_main.R` script notes when these need to be run, but does not call these scripts (eg, indicating `*[RUN USING PYTHON]*`). Within Stata and Python scripts, follow directions for how these should be set up (eg, variables need to be changed to point to the data folder).
6. Re-run the code when setting the `SURVEY_NAME` to (1) `DHS_nga_policy_experiment` and (2) `LSMS`. By default, the `SURVEY_NAME` parameter is set to `DHS`, to process data for `DHS` data. However, the `SURVEY_NAME` parameter (set in line 215) needs to be changed and the code re-run.

Replicating analysis, starting from analysis-ready datasets

Steps

1. Once you have the analysis-ready datasets, after ensuring you have all the needed datasets and have run the code to clean them as detailed above.
2. In `_main.R`, change `dropbox_dir` to point to the data folder and `github_dir` to point to the github repo.
3. Run `_main.R`. Ensure that `RUN_ANALYSIS_CODE` is set to `TRUE`. When `RUN_ANALYSIS_CODE` is set to `TRUE`, the `_main.R` script runs all code in the following sub-folders:

- `DataWork/04_poverty_estimation`: Implements machine learning models and appends results.
- `DataWork/05_figures_tables_global`: Produces figures and tables for paper. Figures and tables are exported to `Paper Tables and Figures` `Paper Tables and Figures/main.tex` compiles all the tables and figures for the main text together, and `Paper Tables and Figures/supplementary_materials.tex` compiles all the tables and figures for the supplementary information/appendix document.

Parameters in main script

`_main.R`: Main script that runs all code for project.

At the beginning of the `_main.R` script, three parameters are set at the beginning.

1. `RUN_CODE`: If `TRUE`, runs other scripts (eg, creating figures and tables). If `FALSE`, just loads packages and sets filepaths.

2. `DELETE_ML_RESULTS`: A large number of machine learning models are implemented for the analysis (ie, a separate model for each country for each set of features, etc). After each model is implemented, results are exported (eg, predicted values from the model). The script: DataWork/04_poverty_estimation/01_pov_estimation.R) that implements the machine learning analysis checks which models have already been implemented by checking the results files. Only models that have not yet been implemented are implemented. Consequently, by default, the code will see all machine learning results and skip running machine learning models. By setting `DELETE_ML_RESULTS` to `TRUE`, machine learning results will be deleted, and machine learning models will be re-implemented. *NOTE:* All machine learning models can take over 15 hours to run.
3. `EXPORT_TXT_REPORT_CODE_DURATION`: If set to `TRUE`, a .txt file will be exported that indicates how long the code took to run. The main script produces all figures and tables for the paper, with one minor exception; the main script does not produce the figure with example daytime satellite images. This script: DataWork/02_get_process_ancillary_data/CNN Features Predict NTL/example_images.ipynb) produces the figure, but the figure requires satellite data to be downloaded, which can be done using the script: DataWork/02_get_process_ancillary_data/CNN Features Predict NTL/01_create_ntlgroup_tfrecord_name_ntlharmon.R. All other figures and tables are generated based on the cleaned datasets and subsequent analysis.