

README: Reproducibility Package for “Population Thresholds and Classification Stability: Evidence from the World Bank’s Small States Classification Rule”

Last updated: September 1, 2026

Corresponding author: Baran Han bhan3@worldbank.org

1. Purpose of this package

This reproducibility package contains the Stata code and input data needed to reproduce the tables and figures in the paper. The analysis uses World Development Indicators (WDI) data, archived WDI population vintages, and an author-compiled country classification file.

Run the package from the master file:
codes/0.master.do

Random seeds are set in the code (20260824 for the wild bootstrap, 20260825 for the power simulation), so all reported statistics, bootstrap critical values, p-values, and power figures reproduce exactly.

2. Computational requirements

The package is written for **Stata 18 for Windows 11**

The following Stata packages are required:

ssc install estout, replace

No internet connection is needed for the standard run. The cross-sectional analysis reads the frozen WDI extract included in the package (data/rawdata_wdi_frozen_2026-08-29.dta, downloaded via wbopendata on August 29, 2026). The download code is retained in 1.setup.do and executes only if the frozen file is absent; in that case wbopendata is required (ssc install wbopendata, replace) and results may differ from the paper, because the live WDI API is continuously revised.

3. Folder structure

The package should be organized as follows:

reproduction package/

README.docx

codes/

0.master.do

1.setup.do

2.xs_generate_baseline_regression.do

3.graphs.do

3b.hansen.do

3c.power_simulation.do

4. population_data_extraction.do

5.simulationcodes.do
6. revision prep.do
7. revision analysis.do
data/
country_categorization_distance.xlsx
rawdata_wdi_frozen_2026-08-29.dta
population/
WDIandGDF_excel_2010_07/WDI_GDF.xlsx
WDIandGDF_excel_2011_07/WDI_GDF.xlsx
WDIandGDF_excel_2012_07/WDI_GDF.xlsx
WDI_excel_2013_07/WDI.xlsx
WDI_excel_2014_07/WDI.xlsx
WDI_excel_2015_07/WDI.xlsx
WDI_excel_2016_07/WDI.xlsx
WDI_excel_2017_07/WDIEXCEL.xlsx
WDI_excel_2018_07_27/WDIEXCEL.xlsx
WDI_excel_2019_06_28/WDIEXCEL.xlsx
WDI_excel_2020_07_01/WDIEXCEL.xlsx
WDI_excel_2021_07_30/WDIEXCEL.xlsx
WDI_excel_2022_06_30/WDIEXCEL.xlsx
WDI_excel_2023_06_29/WDIEXCEL.xlsx
WDI_excel_2024_07_01/WDIEXCEL.xlsx
WDI_excel_2025_07_02/WDIEXCEL.xlsx
output/
log/

The output/ and log/ folders store files created by the code. One archived July snapshot per year, 2010-2025, is included; vintages 2010-2012 use the World Bank's older WDIandGDF archive naming. The archived snapshots are the source for the real-time vintage panel and are unchanged inputs. All sixteen snapshots are loaded by 6. revision prep.do, but the 2010 snapshot contributes no observations: reference years are restricted to 2010–2024 and each estimate must predate its vintage, so the earliest data-bearing vintage is 2011 and the analysis uses fifteen vintages (2011–2025).

The code also writes intermediate datasets into data/ and data/population/ (wdi_panel.dta, wdi_panel_clean.dta, distance.dta, small states definition vulnerability work_merged.dta and _merged2.dta, small_states_xs_2000_2024.dta, threshold_test_results.dta, power_results.dta, vintage_population_panel.dta and vintage_population_panel2.dta with CSV copies, revision_pairs.dta, and output/table3_raw.dta and output/table3_age1_raw.dta - intermediates for Table 3 and its robustness check). These are outputs of the code, not inputs.

4. Instructions for Replicators

1. Unzip the reproduction package.
2. Open codes/0.master.do.
3. Set the top-level directory in the first line:
4. Confirm that the required input files are in the data/ folder.
5. Run codes/0.master.do.

global dir "YOUR/PATH/reproduction package – to send"

The master file runs the code in this order:

- 1.setup.do
- 2.xs_generate_baseline_regression.do
- 3.graphs.do
- 3b.hansen.do
- 3c.power_simulation.do
- 4.population_data_extraction.do
- 5.simulationcodes.do
- 6.revision prep.do
- 7.revision analysis.do

Random seeds are set in 3b.hansen.do (20260824) and 3c.power_simulation.do (20260825), so bootstrap and simulation results reproduce exactly. The full run takes approximately 60-120 minutes depending on hardware; nearly all of the runtime is the wild bootstrap (3b) and the power simulation (3c).

5. A List of Exhibits

Paper item	Code file	Output
Table 1	2.xs_generate_baseline_regression.do	output/table_xs_baseline.csv
Table 2	5.simulationcodes.do	Printed in Stata and saved in log/log.smcl
Table 3	7.revision analysis.do	output/table3_crossing_probabilities.csv
Table 4	Not code generated	Descriptive table in the paper
Table 5	5.simulationcodes.do	Printed in Stata and saved in log/log.smcl
Figure 1	3.graphs.do	output/fig_lpoly_outcomes_lnpop.png
Figure 2	3b.hansen.do	Fig2_Hansen_F_profiles_combined.png
Figure 3	7.revision analysis.do	output/fig3_F_combined.png

Figure 4	7. revision analysis.do	output/fig4_size_dispersion.png
Figure 5	7. revision analysis.do	output/fig5_episodes_combined.png
Appendix Table A1	3b.hansen.do	output/table_threshold_test.csv
Appendix Figure A1	3.graphs.do	output/figA1_partial_residuals.png
Power footnote	3c.power_simulation.do	output/table_power_simulation.csv
Table 3 robustness	7. revision analysis.do	output/table3_crossing_probabilities_age1_robustness.csv

The simulation tables are currently printed to the Stata Results window and captured in the log file. If separate output files are required, export these tables from 5.simulationcodes.do to CSV, Excel, or text format.

5.1. Reproducing Table 5

Table 5 is based on the simulation output produced by codes/5.simulationcodes.do. The underlying results are printed to the Stata Results window and saved in log/log.smcl. The key output is:

```
list iso3 country_name switch_r1-switch_r4, noobs sep(0)
```

where switch_r1–switch_r4 report the number of Small State classification changes between 2019 and 2025 under each implementation rule.

Interpretation of the output:

- 0 indicates no change in classification during the sample period.
- 1 indicates one classification change (for example, Small State → non-Small State).
- 2 indicates two classification changes (Small State → non-Small State → Small State).

For the countries discussed in Table 5:

- **Equatorial Guinea (GNQ)** records one switch under all four rules, corresponding to declassification.
- **Trinidad and Tobago (TTO)** records two switches under Rules 1 and 4, corresponding to declassification followed by reclassification.
- **Kosovo and North Macedonia** have always had populations above 1.5 million throughout the sample period and have never entered the Small State classification; accordingly, they record zero switches under every rule.

- **Eswatini, Mauritius, and Timor-Leste** have always had populations below 1.5 million throughout the sample period and have remained Small States throughout the sample period; accordingly, they record zero switches under every rule.

To identify the year of each classification change reported in Table 5, run:

```
list iso3 vintage population r1 r2 r3 r4, sepby(iso3) noobs
```

The log file shows the specific years in which switches occurred.

Table 5 in the manuscript is a narrative summary of these simulation results. It is not generated as a formatted table by the code; rather, it translates the logged simulation output into the descriptive classification outcomes reported in the paper.

6. Data Availability Statement

This package uses publicly available World Bank data and one author-compiled country classification file. All data were accessed or downloaded on August 29, 2026.

Data source 1: World Development Indicators, current WDI data

File name:

data/rawdata_wdi_frozen_2026-08-29.dta [included in the package]

data/wdi_panel.dta [created by code]

data/wdi_panel_clean.dta [created by code]

Brief description:

The main macroeconomic and demographic variables come from the World Bank World Development Indicators for 2000–2024. The package includes a frozen extract of the raw download (rawdata_wdi_frozen_2026-08-29.dta), obtained via the Stata command `wbopendata` on August 29, 2026. The code runs from this frozen file; `wdi_panel.dta` and `wdi_panel_clean.dta` are intermediate datasets built from it. Because the live WDI database is continuously revised, re-downloading the indicators on a later date may return different values and change the results; the frozen extract is the canonical input for reproducing the paper.

Indicators used:

SP.POP.TOTL: Population

AG.LND.TOTL.K2: Land area

TX.VAL.FUEL.ZS.UN: Fuel exports (% of merchandise exports)

TT.PRI.MRCH.XD.WD: Net barter terms of trade index

NE.TRD.GNFS.ZS: Trade (% of GDP)

NY.GDP.MKTP.PP.KD: GDP, PPP, constant 2021 international dollars

NY.GDP.PCAP.PP.KD: GDP per capita, PPP, constant 2021 international dollars

Citation:

World Bank. World Development Indicators. Washington, DC: World Bank.

URL or access instructions:

<https://databank.worldbank.org/source/world-development-indicators>

Access or download date:

August 29, 2026.

Download steps:

No download is required. codes/1.setup.do loads the frozen extract included in the package. The download code is retained in 1.setup.do and runs only if the frozen file is absent. In that case the wbopendata package and an internet connection are required, and results may differ from the paper.

License or access restrictions:

The WDI data are publicly available from the World Bank under the CC BY 4.0 license and are subject to the World Bank data terms of use. Because the live WDI database is revised over time, the frozen extract included in the package is the canonical input for reproducing the paper's results.

Data source 2: World Development Indicators Archives, population vintages

File name:

data/population/WDIandGDF_excel_2010_07/WDI_GDF.xlsx

data/population/WDIandGDF_excel_2011_07/WDI_GDF.xlsx

data/population/WDIandGDF_excel_2012_07/WDI_GDF.xlsx

data/population/WDI_excel_2013_07/WDI.xlsx

data/population/WDI_excel_2014_07/WDI.xlsx

data/population/WDI_excel_2015_07/WDI.xlsx

data/population/WDI_excel_2016_07/WDI.xlsx

data/population/WDI_excel_2017_07/WDIEXCEL.xlsx

data/population/WDI_excel_2018_07_27/WDIEXCEL.xlsx

data/population/WDI_excel_2019_06_28/WDIEXCEL.xlsx

data/population/WDI_excel_2020_07_01/WDIEXCEL.xlsx

data/population/WDI_excel_2021_07_30/WDIEXCEL.xlsx

data/population/WDI_excel_2022_06_30/WDIEXCEL.xlsx

data/population/WDI_excel_2023_06_29/WDIEXCEL.xlsx

data/population/WDI_excel_2024_07_01/WDIEXCEL.xlsx

data/population/WDI_excel_2025_07_02/WDIEXCEL.xlsx

Brief description

Archived snapshots of the WDI database, one per year for 2010–2025, used to construct the real-time panel of population estimates (SP.POP.TOTL) across sixteen annual vintages. The workbook name inside each folder follows the archive's naming convention of the era: WDI_GDF.xlsx (2010–2012, when WDI was published jointly with Global Development Finance), WDI.xlsx (2013–2016), and WDIEXCEL.xlsx (2017 onward); codes/6. revision prep.do handles the name cascade automatically. The snapshots are fixed historical files and do not change. All sixteen snapshots are loaded by 6. revision prep.do, but the 2010 snapshot contributes no observations: reference years are restricted to 2010–2024 and each estimate must predate its vintage, so the earliest data-bearing vintage is 2011 and the analysis uses fifteen vintages (2011–2025).

Citation:

World Bank. World Development Indicators Archives. Washington, DC: World Bank.

URL or access instructions:

<https://datatopics.worldbank.org/world-development-indicators/wdi-archives.html>

Access or download date:

August 20, 2026.

Download steps:

Go to the World Bank WDI Archives page. Download the WDI Excel archive for each listed vintage. Unzip each archive and place the WDIEXCEL.xlsx file (or equivalent) in the corresponding folder under data/population/, using the folder names shown above. The script codes/4. population_data_extraction.do then imports the files and creates:

data/population/vintage_population_panel.dta

data/population/vintage_population_panel.csv

License or access restrictions:

The WDI archive files are publicly available from the World Bank and are subject to the World Bank data terms of use. Exact reproduction of the population-vintage analysis requires the archived WDI Excel files listed above.

Data source 3: Author-compiled country classification and distance file

File name:

data/country_categorization_distance.xlsx

Brief description:

This file contains country classification variables, sample restrictions, lending and income group variables, regional identifiers, and distance-from-markets variables used in the analysis. The distance-from-markets variable is extracted from Table 6 of the UNCTAD report cited below.

Citation:

UNCTAD. 2021. "The Least Developed Countries Report 2021: The Least Developed Countries in the Post-COVID World." Table 6. Geneva: United Nations Conference on Trade and Development.

URL or access instructions:

https://unctad.org/system/files/official-document/ser-rp-2021d10_en.pdf

Access or download date:

June 10, 2026.

Download steps:

The file `data/country_categorization_distance.xlsx` is included in the reproduction package. No separate download is required. The distance-from-markets variable can be checked against Table 6 of the UNCTAD report linked above.

License or access restrictions:

The file is author-compiled and is included for reproducibility. The underlying distance-from-markets information is from a publicly available UNCTAD report. Any third-party source materials remain subject to their original terms of use.

Datasets created by the code

`data/wdi_panel.dta`

`data/wdi_panel_clean.dta`

`data/distance.dta`

`data/small states definition vulnerability work_merged.dta`

`data/small states definition vulnerability work_merged2.dta`

`data/small_states_xs_2000_2024.dta`

`data/threshold_test_results.dta`

`data/power_results.dta`

`data/population/vintage_population_panel.dta`

`data/population/vintage_population_panel.csv`

`data/population/vintage_population_panel2.dta`

`data/population/vintage_population_panel2.csv`

`data/population/revision_pairs.dta`

`output/table3_raw.dta`

`output/table3_age1_raw.dta`

`output/table3_crossing_probabilities_age1_robustness.csv`