in— Authors: - Kathleen Beegle - Umar Serajuddin - Brian Stacy - Divyanshi Wadhwa —

Missing SDG Gender Indicators

Reproducibility Package for "Missing SDG Gender Indicators" paper

Overview

The code in this replication packages constructs the analysis files and tables and figures for Beegle, Serajuddin, Stacy, and Wadhwa (2023) using R. One main file runs all of the code to generate the data and figures. The file is located in 02_programs/Missing-SDG-Gender-Indicators.Rmd. The replicator should expect the code to run for around 20-30 minutes.

Directory Structure

1. 01_raw_data contains the raw data for the project for each indicator. This folder contains the raw data from the UN SDG database used in the analysis, as well as raw data from the World Bank World Development Indicators. A number of miscellaneous files are included as well that are used.
2. 02_programs contains the main replication file for the project, "Missing-SDG-Gender-Indicators.Rmd". Execute this file to replicate the results. It also contains another file in ./02_programs/misc_scripts/SDG_data_pull.Rm This file is used to pull the data from the UN SDG database. It is not necessary to run this file, as the data is already included in the repository. However, it is included for transparency purposes. If this file is executed, the replication code will no longer replicate, as the data will be overwritten.
3. 03_output_data. This folder contains a number of final output files in either csv or stata .dta format. The most important is SPI_Gender_UNSD_data_5yr.csv, which is used to generate the tables and figures in the paper. The other files are produced in the course of the data production, but are not used in the paper. Some of them were used as sensitivity checks, but the results were not included in the paper.

Instructions to Replicators

- Clone the repository to your local machine.
- Please run 02_programs/Missing-SDG-Gender-Indicators.Rmd to generate the data and figures. This file will run all of the code to generate the data and figures. The replicator should expect the code to run for around 20-30 minutes.
- There should be no need to change the working directory. The code should run as is, because the code is using the here package in R, which automatically handles file paths on local machines. Make sure the .here file is included when you clone the repository.
- This repository contains several files from the R package "renv". The renv package helps manage specific package versions used to produce the results in this repository. Because package version conflicts can make code that runs on one system not run on another system, it is important to have a list of the specific package versions used and a workflow for accessing these specific packages. The renv package provides this. In order to use renv, see the renv documentation here (https://rstudio.github.io/renv/articles/renv.html). In general, the renv::restore() command should install all packages found in the renv.lock file in this repository, so that version conflicts do not cause errors.

License

Summary of Availability

☒ All data **are** publicly available.
☐ Some data **cannot be made** publicly available.
☐ **No data can be made** publicly available.

Data Sources

Data.Name	Data.Files	Location	Provided	Citation
World Development Indicators	WDIEXCEL.xlsx	01_raw_data/	TRUE	World Bank (2023). World Development Indicators.
UN SDG Database	IT_CEN_MGTN.csv; IT_MOB_OWN.csv; SD_MDP_MUHC.csv;...	01_raw_data/sdg_data/	TRUE	UN Statistics Division (2023). UN Global SDG Database.

Data Description

There are four main files used in the analysis. Each file will be discussed in turn.

1. SPI_Gender_UNSD_data_5yr.csv
2. Gender_SDGs.csv
3. Class.xlsx
4. WDIEXCEL.xlsx

SPI_Gender_UNSD_data_5yr.csv This file contains 9 columns. It is available in ./03_output_data/, because it is produced in the Missing-SDG-Gender-Indicators.Rmd, based on raw data files from the UN Global SDG Database. It is generated by looping through each of the SDG files, which each represent data for one SDG indicator (or subindicator), appending the results together, and then checking whether SDG indicator values are available for each country over a specific period. ./02_programs/misc_scripts/SDG_data_pull.Rmd was used to compile the raw SDG indicator data from the UN. Data was last updated in March 2022.

Each row represents a country-year-SDG goal combination. The SDG goal combinations are the 17 SDG goals, plus groups of indicators based on tier (tier 1, tier 2). See here for a description of the SDG tiers.

Column.Name	Description
iso3c	3 letter ISO country code
date	Year of observation
country	Country name
region	Region of country
income_level	Income level of country
goal	SDG goal number or tier of indicators (Tier 1 indicators, Tier 2, etc)
ind_quality	**Key measure of availability of data**. This contains the fraction of SDG indicators with a non-missing value over a 5 year period for a country.
ind_number	Indicator number
goal_count	Number of indicators in goal

Methodology:

1. Download the latest SDG indicator data from UN Stats (https://unstats.un.org/sdgs/indicators/en/#) using their API
2. Transform the data so that for each indicator we can create a score documenting whether a value exists for the country in a year, whether the value is based on country data, country data adjusted, estimated, or modelled data according the UN Stats metadata.
3. Combine the resulting data into a single set of indicators by calculating the average across the gender SDGs.

Below is a paraphrased description from the UN stats webpage (https://unstats.un.org/sdgs/indicators/indicators-list/):

The global indicator framework for Sustainable Development Goals was developed by the Inter-Agency and Expert Group on SDG Indicators (IAEG-SDGs) and agreed upon at the 48th session of the United Nations Statistical Commission held in March 2017.

The global indicator framework includes 231 unique indicators. Please note that the total number of indicators listed in the global indicator framework of SDG indicators is 247. However, twelve indicators repeat under two or three different targets.

For each value of the indicator, the responsible international agency has been requested to indicate whether the national data were adjusted, estimated, modelled or are the result of global monitoring. The "nature" of the data in the SDG database is determined as follows:

- Country data (C): Produced and disseminated by the country (including data adjusted by the country to meet international standards);
- Country data adjusted (CA): Produced and provided by the country, but adjusted by the international agency for international comparability to comply with internationally agreed standards, definitions and classifications;
- Estimated (E): Estimated based on national data, such as surveys or administrative records, or other sources but on the same variable being estimated, produced by the international agency when country data for some year(s) is not available, when multiple sources exist, or when there are data quality issues;
- Modelled (M): Modelled by the agency on the basis of other covariates when there is a complete lack of data on the variable being estimated;
- Global monitoring data (G): Produced on a regular basis by the designated agency for global monitoring, based on country data. There is no corresponding figure at the country level.

For each indicator, we will produce a value for each country with the following coding scheme:

- **1 Point**: Indicator exists and the value is based on the **country**, **country data adjusted**, or **estimated or Global Monitoring** data
- **0 Points**: Indicator **based on modeled data or does not exists**

We give countries no credit for modeled data, because the country did not produce indicators in a form that was directly usable for reporting on an SDG indicator.

When we average over all indicators in a goal to get a score, we compute a 5 year moving average to avoid year to year variability in reporting for SDGs. The overall score for an SDG is then the 5 year average of the percentage of indicator values based on **country**, **country data adjusted**, or **estimated or Global Monitoring** data that were available for the SDG.

Gender_SDGs.csv This file contains codes for the 50 SDG indicators we consider in the paper. Note that there are 68 rows in the data file, because several SDG indicators have sub-indicators that are stored as separate files in the UN Global SDG Database. The file has 5 columns:

SDG Description Code Gender_Breakdown ccode

Column.Name	Description
SDG	SDG indicator number
Description	Description of SDG indicator
Code	UN Global SDG Indicator database series code
Gender Breakdown	Indicator of whether the series requires a gender breakdown, or is specific to women
ccode	A special code used to aggregate to our 50 SDG indicators

Class.xlsx This file contains the World Bank country classifications for regions and income groups used in the analysis. They were current as of June 2023. However, because the classifications are updated periodically, the classifications may have changed since the data were downloaded. A description of the classifications are available here.

You can download the classifications used here. Version 20 was used.

WDIEXCEL.xlsx The file contains indicator data from the World Bank World Development Indicators (WDI). The file contains the bulk download, which was current as of June 2023. However, because the WDI is updated periodically, the data may have changed since the data were downloaded. A description of the WDI is available here.

The WDIECXEL.xlsx file contains data for all indicators, but just a subset of the indicators were used. These include:

WDI Series Code	Description
SP.POP.TOTL	Population, total
IQ.SPI.OVRL	SPI Overall Score
IQ.SPI.PIL1	SPI Pillar 1: Data Use
IQ.SPI.PIL2	SPI Pillar 2: Data Services
IQ.SPI.PIL3	SPI Pillar 3: Data Products
IQ.SPI.PIL4	SPI Pillar 4: Data Sources
IQ.SPI.PIL5	SPI Pillar 5: Data Infrastructure
NY.GDP.PCAP.PP.CD	GDP per capita, PPP (current international)||$SG.LAW.INDX$|$WomenBusinessandtheLawIndexScore(scale1-$ 100)||$NY.GDP.PCAP.KD$|$GDPpercapita(constant2015US)$
HD.HCI.OVRL	Human Capital Index (HCI) (Scale 0-1)
GE.EST	World Governance Indicators, Government Effectiveness Estimate

You can download the WDI data used here. Version 20 was used.

Software

R version 4.2.1 (2022-06-23 ucrt) – "Funny Looking Kid" was used for data production and to produce the tables and figures.

This repository contains several files from the R package "renv". The renv package helps manage specific package versions used to produce the results in this repository. Because package version conflicts can make code that runs on one system not run on another system, it is important to have a list of the specific package versions used and a workflow for accessing these specific packages. The renv package provides this. In order to use renv, see the renv documentation here (https://rstudio.github.io/renv/articles/renv.html). In general,

the renv::restore() command should install all packages found in the renv.lock file in this repository, so that version conflicts do not cause errors.