Authors: - Brian Stacy - Lucas Kitzmüller - Xiaoyu Wang - Daniel Gerszon Mahler - Umar Serajuddin

Academic Data Use

Reproducibility package for "Missing Evidence: Tracking Academic Data Use around the World"

Overview

The code in this replication packages constructs the analysis files and tables and figures for Stacy, Kitzmüller, Wang, Mahler, and Serajuddin (2023) using R. One main file runs all of the code to generate the data and figures. The file is located in 02_programs/Data_Use_Academia_tables_figures.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 training set of articles, as well as raw data from the World Bank World Development Indicators, and other academic papers studying academic production (Das et al. (2013), Porteous (2020), National Science Board, National Science Foundation (2019)). A number of miscellaneous files are included as well that are used. The main file containing the 1 million classified articles is too large to host on github and is stored in a publicly accessible folder hosted using S3 from Amazon Web Services.

2. 02_programs contains the main replication file for the project, "Data_Use_Academia_tables_figures.Rmd". Execute this file to replicate the results. It also contains another file in ./02_programs/misc_scripts/wdi_pull.R. This file is used to pull the data from the WDI. 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 csv format. The most important is data_use_country_scores_annual.csv, which is primarily 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 or exported results for the user, but the results were not directly included in the paper.

Instructions to Replicators

- Clone the repository to your local machine.
- Before executing the Data_Use_Academia_tables_figures.Rmd file, users should set up the appropriate environment. The renv package helps maintain consistent package versions and dependencies, ensuring that users have the required libraries.
- Users should first ensure the renv package is installed. If it's not already present, it can be installed using install.packages("renv").
- Once installed, users should set up the environment by running the following commands:

```
renv::activate()
```

```
renv::restore()
```

- With the environment now properly set up, users can proceed, please run 02_programs/Data_Use_Academia_tables_figures.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.

License

Summary of Availability

- [X] 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	correlates_df.csv	01_raw_data/	TRUE	World Bank (2023). World Development Indicators.
S2ORC	results_completed_updated_20231003.fst	Hosted on S3 from AWS	TRUE	Lo, Kyle, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. "S2ORC: The Semantic Scholar Open Research Corpus." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 4969–83. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.447.

Data Description

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

1. results_completed_updated_20231003.fst
2. correlates_df.csv

results_completed_updated_20231003.fst

This file contains article metadata for the 1 million articles classified by the machine learning algorithm. The file is too large to host on github, so it is hosted on S3 from AWS. The file is in fst format, which is a compressed format that is faster to read than csv. The file is read into R in the code using the fst package. The file contains the following columns:

In addition to the metadat, two classifications are available in the dataset. This included `data_use`, which identifies whether or not the NLP model identified the article as using data. The second classification is `places`, which identifies the locations identified in the article. The places are then broken into columns with binary indicators of whether a specific country was identified.

Column.Name	Description
paper_id	numeric semantic scholar ID
title	Title of article
abstract	Abstract of article
year	Year of publication
doi	DOI of article
venue	Journal or other venue of publication
journal	Journal of publication
mag_field_of_study	Field of study of article
group_name	Field of study of article
outbound_citations	Number of outbound citations
inbound_citations	Number of inbound citations
data_use	Indicator of whether the article uses data
places	list of locations identified in the article

Column.Name	Description
countries	list of iso3c identified in the article
V1	Index
ATF	0/1 indicator of whether the article contains mentions of the country ATF (French Southern and Antarctic Territories)
...	...
AFG	0/1 indicator of whether the article contains mentions of the country AFG (Afghanistan)
...	...
ZWE	0/1 indicator of whether the article contains mentions of the country ZWE (Zimbabwe)
nf	0/1 indicator of whether the article contains mentions of the country nf (not found)

correlates_df.csv

The file contains indicator data from the World Bank World Development Indicators (WDI). A set of indicators were pulled from the World Bank API, which was current as of November 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.

Data was pulled using the wbstats package in R.

The following indicators were pulled from the WDI:

WDI Series Code	Description
SP.POP.TOTL	Population, total
NY.GDP.MKTP.PP.KD	GDP, PPP (current international $)
NY.GDP.PCAP.PP.KD	GDP per capita, PPP (current international $)

WDI Series Code	Description
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
IQ.SCI.OVRL	SCI Overall Score
NV.IND.MANF.ZS	Manufacturing, value added (% of GDP)
NV.AGR.TOTL.ZS	Agriculture, forestry, and fishing, value added (% of GDP)
NE.TRD.GNFS.ZS	Trade (% of GDP)
HD.HCI.OVRL	Human Capital Index (HCI) Score
HD.HCI.LAYS	Human Capital Index (HCI) Learning-Adjusted Years of School
SE.PRM.ENRR	School enrollment, primary (% gross)
BN.CAB.XOKA.GD.ZS	Current account balance (% of GDP)
CC.EST	Control of Corruption: Estimate
GE.EST	Government Effectiveness: Estimate
PV.EST	Political Stability and Absence of Violence\Terrorism: Estimate
RQ.EST	Regulatory Quality: Estimate
RL.EST	Rule of Law: Estimate

WDI Series Code	Description
VA.EST	Voice and Accountability: Estimate
BX.KLT.DINV.WD.GD.ZS	Foreign direct investment, net inflows (% of GDP)
SI.POV.DDAY	Poverty headcount ratio at $2.15 a day (2017 PPP) (% of population)
SI.POV.GINI	GINI index (World Bank estimate)
SE.TER.CUAT.MS.ZS	Educational attainment, at least Master's or equivalent, population 25+, total (%) (cumulative)
SE.TER.ENRR	School enrollment, tertiary (% gross)

Software

R version 4.3.1 (2022-06-23 ucrt) -- "Beagle Scouts" 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.