README (Deposit Number) – Data and Code for:

Paper Title

Author 1 Author 2

Month YEAR

Template README and Guidance

INSTRUCTIONS: This README suggests structure and content that have been approved by various journals, see Endorsers. It is available as Markdown/txt, Word, LaTeX, and PDF. In practice, there are many variations and complications, and authors should feel free to adapt to their needs. All instructions can (should) be removed from the final README (in Markdown, remove lines starting with `> INSTRUCTIONS`). Please ensure that a PDF is submitted in addition to the chosen native format.

DISCLAIMER: The original template was adapted to fit the structure of this paper template. Also, additional examples were added, mostly, from an example application of the template (Vieira 2023).

Overview

INSTRUCTIONS: The typical README in social science journals serves the purpose of guiding a reader through the available material and a route to replicating the results in the research paper. Start by providing a brief overview of the available material and a brief guide as to how to proceed from beginning to end.

Original Example: The code in this replication package constructs the analysis file from the three data sources (Ruggles et al., 2018; Inglehart et al., 2019; BEA, 2016) using Stata and Julia. Two master files run all of the code to generate the data for the 15 figures and 3 tables in the paper. The replicator should expect the code to run for about 14 hours.

Additional Example: The code in this replication package cleans the raw data for each data source (4 sources), constructs the samples for analysis, and generates the results using R. A master file runs all of the code to generate the data for the two figures and one table in the paper. The replicator should expect the code to run for about 2 minutes, divided into 0.5 minutes in the cleaning, 1 minute in the construction, and 0.5 minutes in the analysis. A specific time for all individual scripts is reported in CSV files with the prefix `"_timeProcessing_"` in each code folder. All the data is provided, from raw to final, including intermediate, so it is possible to skip any individual script.

Description of Files Structure

INSTRUCTIONS: short description of only the folders and files that should be included in the final replication package based on the project structure.

DISCLAIMER: Section not present in the original AEA template.

- "README.md": This document. A markdown file to generate "README.pdf" and "README.html" (best for reading). It provides the necessary information about the structure of the replication folder, data sources, data access, computational requirements, and ultimately explains how to fully replicate the analysis presented in the paper.
- "code": a folder containing all scripts to clean, construct, and analyze the data
  - "code/_MASTERFILE.R": R script to run all scripts from data cleaning to generating the final results;
  - "code/setup.R": R script to install/load R packages and configure the initial setup. Uses "groundhog" to keep all packages version fixes at the specified date (YYYY-MM-DD);
  - "code/raw2clean": R scripts that clean the data on input and save on the output for each dataset;
  - "code/projectSpecific": R scripts that construct the sample(s) for analysis;
  - "code/analysis": R scripts to generate the results presented in the paper (statistics, figures, and tables);
  - "code/_functions": auxiliary folder with custom R functions used in multiple R scripts.
- "data": a folder containing data in a variety of formats: raw, cleaned, intermediate, final datasets for analysis, and analysis outputs
  - "data/raw2clean": one folder for each dataset with the following structure:
    - "/input": folder with raw datasets;
    - "/output": folder with the cleaned dataset;
    - "/documentation": folder with at least two files:
      - "_metadata.txt" text file that describes the data and provides access instructions;
      - "codebook_datasetName.txt" text file with summary statistics and variables description.
  - "data/projectSpecific": folder with the sample(s) of interest, intermediate datasets with the variables of interest, and merged sample(s) for analysis:
    - "/prepData": folder with intermediate datasets with the variables of interest;
    - "/unitLevel": folder with the sample(s) of interest at the *unit level*.
  - "data/analysis": folder with all regression outputs in "/regressions";
  - "data/_temp": folder to hold temporary files output (filled when running some .R scripts).
- "references": folder with three BibTeX files to record all references for citation (literature: "references_literature.bib", data: "references_data.bib", and software: "references_lsoftware.bib")
- "results": folder with the main results used in the paper
  - "figures": folder with all figures. The figures of the main paper are listed in "figures.tex". The figures of the appendix are listed in "figures_appendix.tex";
  - "tables": folder with all tables. The tables of the main paper are listed in "tables.tex". The tables of the appendix are listed in "tables_appendix.tex";
  - "stats": folder with the log output from the R script that calculates all the statistics cited in the text "stats_inText.txt".
- "name_proj.Rproj": R project to automatically adjust file path references. Always open RStudio from this file when running any R script.
- "LICENSE.txt": a text file with a dual-license setup.

Data Availability and Provenance Statements

INSTRUCTIONS: Every README should contain a description of the origin (provenance), location, and accessibility (data availability) of the data used in the article. These descriptions are generally referred to as "Data Availability Statements" (DAS). However, in some cases, there is no external data used.

☐ This paper does not involve analysis of external data (i.e., no data are used or the only data are generated by the authors via simulation in their code).

If box above is checked and if no simulated/synthetic data files are provided by the authors, please skip directly to the section on Computational Requirements. Otherwise, continue.

INSTRUCTIONS: - When the authors are **secondary data users** (they did not generate the data), the provenance and DAS coincide, and should describe the condition under which (a) the current authors (b) any future users might access the data. - When the data were generated (by the authors) in the course of conducting (lab or field) **experiments**, or were collected as part of **surveys**, then the description of the provenance should describe the data generating process, i.e., survey or experimental procedures: - Experiments: complete sets of experimental instructions, questionnaires, stimuli for all conditions, potentially screenshots, scripts for experimenters or research assistants, as well as for subject eligibility criteria (e.g., selection criteria, exclusions), recruitment waves, demographics of subject pool used. - For lab experiments specifically, a description of any pilot sessions/studies, and computer programs, configuration files, or scripts used to run the experiment. - For surveys, the whole questionnaire (code or images/PDF) including survey logic if not linear, interviewer instructions, enumeration lists, sample selection criteria.

The information should describe ALL data used, regardless of whether they are provided as part of the replication archive or not, and regardless of size or scope. For instance, if using GDP deflators, the source of the deflators (e.g., at the national statistical office) should also be listed here. If any of this information has been provided in a pre-registration, then a link to that registration may (partially) suffice.

DAS can be complex and varied. Examples are provided here, and below.

Importantly, if providing the data as part of the replication package, authors should be clear about whether they have the **rights** to distribute the data. Data may be subject to distribution restrictions due to sensitivity, IRB, proprietary clauses in the data use agreement, etc.

NOTE: DAS do not replace Data Citations (see Guidance). Rather, they augment them. Depending on journal requirements and to some extent stylistic considerations, data citations should appear in the main article, in an appendix, or in the README. However, data citations only provide information **where** to find the data, not **how to access** that data. Thus, DAS augment data citations by going into additional detail that allow a researcher to assess cost, complexity, and availability over time of the data used by the original author.

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.

(Optional, but recommended) License for Data

INSTRUCTIONS: Most data repositories provide for a default license, but do not impose a specific license. Authors should actively select a license. This should be provided in a LICENSE.txt

file, separately from the README, possibly combined with the license for any code. Some data may be subject to inherited license requirements, i.e., the data provider may allow for redistribution only if the data is licensed under specific rules - authors should check with their data providers. For instance, a data use license might require that users - the current author, but also any subsequent users - cite the data provider. Licensing can be complex. Some non-legal guidance may be found here.

The data is licensed under a Creative Commons Attribution 4.0 International Public License. See LICENSE.txt for details.

Summary of Availability

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

Additional Example: The data used to support the findings of this study comes from multiple data sources; all of them are publicly available online and have been deposited in a Zenodo repository (Vieira 2023). Each raw dataset is listed and described in more detail below. Access to download from the original source is guaranteed by providing a persistent link, using the Save a Page feature from Archive.org, pointing directly to the data download.

Details on each Data Source

INSTRUCTIONS: For each data source, list the file that contains data from that source here; if providing combined/derived data files, list them separately after the DAS. For each data source or file, as appropriate,

- Describe the format (open formats preferred, but some software-specific formats OK if open-source readers available): `.dta`, `.xlsx`, `.csv`, `netCDF`, etc.
- Provide a data dictionary, either as part of the archive (list the file name), or at a URL (list the URL). Some formats are self-describing *if* they have the requisite information (e.g., `.dta` should have both variable and value labels).

Original Example for public use data collected by the authors

The [DATA TYPE] data used to support the findings of this study have been deposited in the [NAME] repository ([DOI or OTHER PERSISTENT IDENTIFIER]). [1]. The data were collected by the authors and are available under a Creative Commons Non-commercial license.

Original Example for public use data sourced from elsewhere and provided

Data on National Income and Product Accounts (NIPA) were downloaded from the U.S. Bureau of Economic Analysis (BEA, 2016). We use Table 30. Data can be downloaded from https://apps.bea.gov/regional/downloadzip.cfm, under "Personal Income (State and Local)", select CAINC30: Economic Profile by County, then download. Data can also be directly downloaded using https://apps.bea.gov/regional/zip/CAINC30.zip. A copy of the data is provided as part of this archive. The data are in the public domain.

Datafile: `CAINC30__ALL_AREAS_1969_2018.csv`

Original Example for public use data with required registration and provided extract

> The paper uses IPUMS Terra data (Ruggles et al, 2018). IPUMS-Terra does not allow for redistribution, except for the purpose of replication archives. Permissions as per https://terra.ipums.org/citation have been obtained, and are documented within the "data/IPUMS-terra" folder. > Note: the reference to "Ruggles et al, 2018" would be resolved in the Reference section of this README, **and** in the main manuscript.

Datafile: `data/raw/ipums_terra_2018.dta`

Original Example for free use data with required registration, extract not provided

> The paper uses data from the World Values Survey Wave 6 (Inglehart et al, 2019). Data is subject to a redistribution restriction, but can be freely downloaded from http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp. Choose `WV6_Data_Stata_v20180912`, fill out the registration form, including a brief description of the project, and agree to the conditions of use. Note: "the data files themselves are not redistributed" and other conditions. Save the file in the directory `data/raw`.
>
> > Note: the reference to "Inglehart et al, 2018" would be resolved in the Reference section of this README, **and** in the main manuscript.

Datafile: `data/raw/WV6_Data_Stata_v20180912.dta` (not provided)

Original Example for confidential data

> INSTRUCTIONS: Citing and describing confidential data, in particular when it does not have a regular distribution channel or online landing page, can be tricky. A citation can be crafted (see guidance), and the DAS should describe how to access, whom to contact (including the role of the particular person, should that person retire), and other relevant information, such as required citizenship status or cost.
>
> The data for this project (DESE, 2019) are confidential, but may be obtained with Data Use Agreements with the Massachusetts Department of Elementary and Secondary Education (DESE). Researchers interested in access to the data may contact [NAME] at [EMAIL], also see www.doe.mass.edu/research/contact.html. It can take some months to negotiate data use agreements and gain access to the data. The author will assist with any reasonable replication attempts for two years following publication.

Original Example for confidential Census Bureau data

> All the results in the paper use confidential microdata from the U.S. Census Bureau. To gain access to the Census microdata, follow the directions here on how to write a proposal for access to the data via a Federal Statistical Research Data Center: https://www.census.gov/ces/rdcresearch/howtoapply.html. You must request the following datasets in your proposal: 1. Longitudinal Business Database (LBD), 2002 and 2007 2. Foreign Trade Database – Import (IMP), 2002 and 2007 [. . .]

(adapted from Fort (2016))

Original Example for preliminary code during the editorial process

> Code for data cleaning and analysis is provided as part of the replication package. It is available at https://dropbox.com/link/to/code/XYZ123ABC for review. It will be uploaded to the [JOURNAL REPOSITORY] once the paper has been conditionally accepted.

Additional Examples

BRAZILIAN BIOMES (IBGE 2019)

- folder file path: `"data/raw2clean/administrative/territorial_ibge/brazil"`
- content: biomes perimeter (polygons data frame); Brazil (extent); 2019 (year of reference)
- source: Brazilian Institute for Geography and Statistics (IBGE)
- original link: https://www.ibge.gov.br/geociencias/informacoes-ambientais/15842-biomas.html?=&t=sobre
- raw data downloaded on: SEP/16/2020
- web archive link (used for download): https://web.archive.org/web/20200916173523/ftp://geoftp.ibge.gov.br/informacoes_ambientais/estudos_ambientais/biomas/vetores/Biomas_250mil.zip
- raw data archived on: SEP/16/2020
- CRS: LongLat (coordinate system); SIRGAS2000; not projected (EPSG: 4674)

- notes: downloaded zip file containing multiple files that compose the shapefile (.shp, .prj, .shx, etc), using the web archive link. Manually unziped the folder and moved the files to "input" folder, then deleted the "Biomas_250mil" folders.
- provided: yes

Dataset list

> INSTRUCTIONS: In some cases, authors will provide one dataset (file) per data source, and the code to combine them. In others, in particular when data access might be restrictive, the replication package may only include derived/analysis data. Every file should be described. This can be provided as an Excel/CSV table, or in the table below.

Data file	Source	Notes	Provided
`data/raw/lbd.dta`	LBD	Confidential	No
`data/raw/terra.dta`	IPUMS Terra	As per terms of use	Yes
`data/derived/regression_input.dta`	As listed	Combines multiple data sources, serves as input for Tables 2, 3 and Figure 5.	Yes

Computational requirements

> INSTRUCTIONS: In general, the specific computer code used to generate the results in the article will be within the repository that also contains this README. However, other computational requirements - shared libraries or code packages, required software, specific computing hardware - may be important, and is always useful, for the goal of replication. Some example text follows.

> INSTRUCTIONS: We strongly suggest providing setup scripts that install/set up the environment. Sample scripts for Stata, R, Python, Julia are easy to set up and implement.

Software Requirements

> INSTRUCTIONS: List all of the software requirements, up to and including any operating system requirements, for the entire set of code. It is suggested to distribute most dependencies together with the replication package if allowed, in particular if sourced from unversioned code repositories, Github repos, and personal webpages. In all cases, list the version *you* used.

Original Example: - Stata (code was last run with version 15) - `estout` (as of 2018-05-12) - `rdrobust` (as of 2019-01-05) - the program" `0_setup.do`" will install all dependencies locally, and should be run once. - Python 3.6.4 - `pandas` 0.24.2 - `numpy` 1.16.4 - the file "`requirements.txt`" lists these dependencies, please run" `pip install -r requirements.txt`" as the first step. See https://pip.readthedocs.io/en/1.1/requirements.html for further instructions on using the "`requirements.txt`" file. - Intel Fortran Compiler version 20200104 - Matlab (code was run with Matlab Release 2018a) - R 3.4.3 - `tidyr` (0.8.3) - `rdrobust` (0.99.4) - the file" `0_setup.R`" will install all dependencies (latest version), and should be run once prior to running other programs.

Portions of the code use bash scripting, which may require Linux.

Portions of the code use Powershell scripting, which may require Windows 10 or higher.

Additional Example:

- R (code was last run with version 4.3.0 (2023-04-21 ucrt))
  - the file `"code/setup.R"` will install/load R packages and configure the initial setup. It uses the R package `"groundhog"` (version 3.1.0) to keep all package versions fixed at the specified date (2023-05-06). It also uses `knitr::write_bib` to record all R packages as software citations in a BibTeX file `"references/references_software.bib"`. It is automatically sourced within any .R script in the project.
  - the file `"reproducible_paper_example2.Rproj"` will guarantee that the working directory is set to the root of the project (always open RStudio using this file).
  - List of R packages:
    - groundhog (version 3.1.0) (Simonsohn and Gruson 2023)
    - conflicted (version 1.2.0) (Wickham 2023a)
    - Hmisc (version 5.0-1) (Harrell 2023)
    - sjlabelled (version 1.2.0) (Lüdecke 2022)
    - tidyverse (version 2.0.0) (Wickham 2023b)
    - sf (version 1.0-12) (Pebesma 2023)
    - rmarkdown (version 2.21) (Allaire et al. 2023)
    - tictoc (version 1.2) (Izrailev 2023)
    - here (version 1.0.1) (Müller 2020)
    - tinytex (version 0.45) (Xie 2023)
    - janitor (version 2.2.0) (Firke 2023)

Memory and Runtime Requirements

> INSTRUCTIONS: Memory and compute-time requirements may also be relevant or even critical. Some example text follows. It may be useful to break this out by Table/Figure/section of processing. For instance, some estimation routines might run for weeks, but data prep and creating figures might only take a few minutes.

Summary Approximate time needed to reproduce the analyses on a standard (CURRENT YEAR) desktop machine:

☐ <10 minutes
☐ 10-60 minutes
☐ 1-8 hours
☐ 8-24 hours
☐ 1-3 days
☐ 3-14 days
☐ > 14 days
☐ Not feasible to run on a desktop machine, as described below.

Details

INSTRUCTIONS: Identifiying hardware and OS can be obtained through a variety of ways: Some of these details can be found as follows:

- (Windows) by right-clicking on "This PC" in File Explorer and choosing "Properties"
- (Mac) Apple-menu > "About this Mac"
- (Linux) see code in tools/linux-system-info.sh‘

Original Example:

The code was last run on a **4-core Intel-based laptop with MacOS version 10.14.4**.

Portions of the code were last run on a **32-core Intel server with 1024 GB of RAM, 12 TB of fast local storage**. Computation took 734 hours.

Portions of the code were last run on a **12-node AWS R3 cluster, consuming 20,000 core-hours**.

Additional Example:

The code was last run on an **8-core Desktop; Intel Core i7-2600 CPU @ 3.40 GHz processor; 32GB RAM; Windows 10 Pro**.

The total disk size (expected) to be consumed by the project considering everything (including intermediate dataset, libraries, etc.) in an uncompressed format is approximately 541MB (~634 files).

Description of programs/code

INSTRUCTIONS: Give a high-level overview of the program files and their purpose. Remove redundant/ obsolete files from the Replication archive.

Original Example:

- Programs in `programs/01_dataprep` will extract and reformat all datasets referenced above. The file `programs/01_dataprep/master.do` will run them all.
- Programs in `programs/02_analysis` generate all tables and figures in the main body of the article. The program `programs/02_analysis/master.do` will run them all. Each program called from `master.do` identifies the table or figure it creates (e.g., `05_table5.do`). Output files are called appropriate names (`table5.tex`, `figure12.png`) and should be easy to correlate with the manuscript.
- Programs in `programs/03_appendix` will generate all tables and figures in the online appendix. The program `programs/03_appendix/master-appendix.do` will run them all.
- Ado files have been stored in `programs/ado` and the `master.do` files set the ADO directories appropriately.
- The program `programs/00_setup.do` will populate the `programs/ado` directory with updated ado packages, but for purposes of exact reproduction, this is not needed. The file `programs/00_setup.log` identifies the versions as they were last updated.

- The program `programs/config.do` contains parameters used by all programs, including a random seed. Note that the random seed is set once for each of the two sequences (in `02_analysis` and `03_appendix`). If running in any order other than the one outlined below, your results may differ.

Additional Example:

- `"code/_MASTERFILE.R"` will run individual master files for each folder:
  - `"code/raw2clean/_masterfile_raw2clean.R"` will run one R script to clean each input dataset (4 scripts).
  - `"code/projectSpecific/_masterfile_projectSpecific.R"` will construct the base sample, extract the information from each dataset relevant to this paper, construct the variables of interest, merge them with the base sample, and generate the sample for analysis in multiple formats: panel and spatial (4 scripts).
  - `"code/analysis/_masterfile_analysis.R"` will run the regressions and generate all supporting statistics, tables, and figures (5 scripts).

(Optional, but recommended) License for Code

> INSTRUCTIONS: Most journal repositories provide for a default license, but do not impose a specific license. Authors should actively select a license. This should be provided in a LICENSE.txt file, separately from the README, possibly combined with the license for any data provided. Some code may be subject to inherited license requirements, i.e., the original code author may allow for redistribution only if the code is licensed under specific rules - authors should check with their sources. For instance, some code authors require that their article describing the econometrics of the package be cited. Licensing can be complex. Some non-legal guidance may be found here.

The code is licensed under a Modified BSD License. See LICENSE.txt for details.

Instructions to Replicators

> INSTRUCTIONS: The first two sections ensure that the data and software necessary to conduct the replication have been collected. This section then describes a human-readable instruction to conduct the replication. This may be simple, or may involve many complicated steps. It should be a simple list, no excess prose. Strict linear sequence. If more than 4-5 manual steps, please wrap a master program/Makefile around them, in logical sequences. Examples follow.

Original Example:

- Edit `programs/config.do` to adjust the default path
- Run `programs/00_setup.do` once on a new system to set up the working environment.
- Download the data files referenced above. Each should be stored in the prepared subdirectories of `data/`, in the format that you download them in. Do not unzip. Scripts are provided in each directory to download the public-use files. Confidential data files requested as part of your FSRDC project will appear in the `/data` folder. No further action is needed on the replicator's part.
- Run `programs/01_master.do` to run all steps in sequence.

Additional Example:

- (Only in the first time) Download the replication package.

- (Only in the first time) Download R 4.3.0 (strongly recommended).
- Open RStudio using `"reproducible_paper_example2.Rproj"` to set the working directory to the project root.
- (Only in the first time) Run `"code/setup.R"` to install all the necessary R packages with the same version as when it was last run.
  - Strongly recommended to run line by line to answer possible prompts from `groundhog`;
  - `"groundhog"` might give the following message `"IMPORTANT. R does not have a personal library to save packages to. The default location for it is: 'C:\Users\username\AppData\Local/R 1) Type 'create' to create that directory 2) Otherwise type 'stop'"`. Answer with `create` in the console to proceed;
  - Package `tabulizer` might require installing Java 64-bits (https://stackoverflow.com/questions/17376939/problems-when-trying-to-load-a-package-in-r-due-to-rjava)
  - In some cases Rtools might be necessary (https://groundhogr.com/rtools/);
  - In some cases re-running the script might solve possible installation issues.
- Run `"code/_MASTERFILE.R"` to run all R scripts in sequence.
  - Skipping individual R programs will not prevent others from running correctly because all intermediate datasets are available. However, you should manually adjust the folder-specific master files to remove the scripts you do not want to run.

Details

- `programs/00_setup.do`: will create all output directories, install needed ado packages.
  - If wishing to update the ado packages used by this archive, change the parameter `update_ado` to `yes`. However, this is not needed to successfully reproduce the manuscript tables.
- `programs/01_dataprep`:
  - These programs were last run at various times in 2018.
  - Order does not matter, all programs can be run in parallel, if needed.
  - A `programs/01_dataprep/master.do` will run them all in sequence, which should take about 2 hours.
- `programs/02_analysis/master.do`.
  - If running programs individually, note that ORDER IS IMPORTANT.
  - The programs were last run top to bottom on July 4, 2019.
- `programs/03_appendix/master-appendix.do`. The programs were last run top to bottom on July 4, 2019.
- Figure 1: The figure can be reproduced using the data provided in the folder "2_data/data_map", and ArcGIS Desktop (Version 10.7.1) by following these (manual) instructions:
  - Create a new map document in ArcGIS ArcMap, browse to the folder "2_data/data_map" in the "Catalog", with files "provinceborders.shp", "lakes.shp", and "cities.shp".
  - Drop the files listed above onto the new map, creating three separate layers. Order them with "lakes" in the top layer and "cities" in the bottom layer.
  - Right-click on the cities file, in properties choose the variable "health"... (more details)

Additional Example:

- `"code/_functions/_setup.R"`: will be the first script sourced by `"code/_MASTERFILE.R"`, and will install the `"checkpoint"` package, create the project library, and populate it with all needed packages with the correct versions.

  - Use of R version "3.6.1" is highly recommended to guarantee a successful installation of the necessary packages. It can be downloaded here for Windows.
  - If you have more than one R version and need to select the 3.6.1 you can follow these instructions.
  - If you are trying to use a different R version you will need to manually change the line 40 of the script `"code/_functions/_setup.R"` from `"checkpoint::checkpoint(R.version = "3.6.1", snapshotDate = "2019-09-03", checkpointLocation = getwd())"` to `"checkpoint::checkpoint(R.version = "numberOfYourVersion", snapshotDate = "2019-09-03", checkpointLocation = getwd())"`. However, this option is not recommended. For example, we tested it using R version "4.0.2" and it gave an error in the installation of the first package (`"cleangeo"`).

- Skipping one individual R program will not prevent others from running correctly, because all intermediate datasets are available. There are only two exceptions that require an additional manual step:

  - If you want to skip `"code/raw2clean/geography/weather_CPCG/geography_br_weather_CPCG_raw2clean.R"` (time of processing: 6 days), unzip `"data/raw2clean/geography/weather_CPCG/output.zip"` extracting the files to `"data/raw2clean/geography/weather_CPCG"` in order to fill the existing and empty folder `"data/raw2clean/geography/weather_CPCG/output"` without creating redundant directory levels.
  - If you want to skip `"code/raw2clean/geography/weather_ncepDoeReanalysis/geo_br_weather_reanalysis_raw2clean.R"` (time of processing: 1 hour), unzip `"data/raw2clean/geography/weather_ncepDoeReanalysis/output.zip"` extracting the files to `"data/raw2clean/geography/weather_ncepDoeReanalysis"` in order to fill the existing and empty folder `"data/raw2clean/geography/weather_ncepDoeReanalysis/output"` without creating redundant directory levels.
  - The two output folders above and the input folder mentioned in the instructions had to be zipped due to the limit in the number of files allowed in the openICPSR deposit (1,000 files). We opted for the zip solution because the number of files was very concentrated in these 3 folders (~24,000 files), so even though it added an additional manual step it also allowed us to preserve a similar workflow structure for all data sources.

- Skipping individual do-files is not recommended, there are some dependencies across do-files and the setup is only done in the `"code/_MASTERFILE.do"`.

- If running R programs individually note that sometimes ORDER IS IMPORTANT.

List of tables and programs

INSTRUCTIONS: Your programs should clearly identify the tables and figures as they appear in the manuscript, by number. Sometimes, this may be obvious, e.g. a program called "`table1.do`" generates a file called `table1.png`. Sometimes, mnemonics are used, and a mapping is necessary. In all circumstances, provide a list of tables and figures, identifying the program (and possibly the line number) where a figure is created.

NOTE: If the public repository is incomplete, because not all data can be provided, as described in the data section, then the list of tables should clearly indicate which tables, figures, and in-text numbers can be reproduced with the public material provided.

The provided code reproduces:

☐ All numbers provided in text in the paper
☐ All tables and figures in the paper
☐ Selected tables and figures in the paper, as explained and justified below.

Original Example:

Figure/Table #	Program	Line Number	Output file	Note
Table 1	02_analysis/table1.do		summarystats.csv	
Table 2	02_analysis/table2and3.do	35	table2.csv	
Table 3	02_analysis/table2and3.do	345	table3.csv	
Figure 1	n.a. (no data)			Source: Herodus (2011)
Figure 2	02_analysis/fig2.do		figure2.png	
Figure 3	02_analysis/fig3.do		figure-robustness.png	Requires confidential data

Additional Example:

Figure/Table	Script in "code/analysis/"	Output in "results/"
Table 1	tab1_summaryStat.R	tables/tab1_summaryStat.tex
Figure 1	fig1_eventStudyBalanced.R	figures/fig1_eventStudyBalanced.png
Figure A.1	figA1_eventStudyUnbalanced.R	figures/figA1_eventStudyUnbalanced.png

The numbers provided in the text in the paper are generated in `"code/analysis/stats_inText.R"` and saved in the `"results/stats/stats_inText.txt"` with page location and citation.

Acknowledgements

Adapted structure from Vilhuber et al. (2020). Used additional examples from Vieira (2023).

References

> INSTRUCTIONS: As in any scientific manuscript, you should have proper references. For instance, in this sample README, we cited "Vilhuber et al. (2020)" and "Assunção, Gandour, and Rocha (2023)". The reference should thus be listed here, in the style of your journal:

Allaire, JJ, Yihui Xie, Christophe Dervieux, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, et al. 2023. *Rmarkdown: Dynamic Documents for r.* https://CRAN.R-project.org/package=rmarkdown.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data.* https://CRAN.R-project.org/package=janitor.

Harrell, Frank E, Jr. 2023. *Hmisc: Harrell Miscellaneous.* https://hbiostat.org/R/Hmisc/.

IBGE. 2019. "Biomas Do Brasil: Shapefile, 2019." Instituto Brasileiro de Geografia e Estatística (IBGE), Ministério da Economia. Archived at: https://web.archive.org/web/20200916173523/ftp://geoftp.ibge.gov.br/informacoes_ambientais/estudos_ambientais/biomas/vetores/Biomas_250mil.zip. Archived on: September 16, 2020.

Izrailev, Sergei. 2023. *Tictoc: Functions for Timing r Scripts, as Well as Implementations of "Stack" and "StackList" Structures.* https://github.com/jabiru/tictoc.

Lüdecke, Daniel. 2022. *Sjlabelled: Labelled Data Utility Functions.* https://strengejacke.github.io/sjlabelled/.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files.* https://CRAN.R-project.org/package=here.

Pebesma, Edzer. 2023. *Sf: Simple Features for r.* https://CRAN.R-project.org/package=sf.

Simonsohn, Uri, and Hugo Gruson. 2023. *Groundhog: Version-Control for CRAN, GitHub, and GitLab Packages.* https://CRAN.R-project.org/package=groundhog.

Vieira, João Pedro. 2023. "Reproducible Paper Example." Zenodo. Available at: https://doi.org/10.5281/zenodo.7971743. Accessed on: May 25, 2023.

Vilhuber, Lars, Marie Connolly, Miklós Koren, Joan Llull, and Peter Morrow. 2020. "A template README for social science replication packages (v1.0.0)." Zenodo. Available at: 10.5281/zenodo.4319999. Accessed on: May 19, 2023.

Wickham, Hadley. 2023a. *Conflicted: An Alternative Conflict Resolution Strategy*. https://CRAN.R-project.org/package=conflicted.

———. 2023b. *Tidyverse: Easily Install and Load the Tidyverse*. https://CRAN.R-project.org/package=tidyverse.

Xie, Yihui. 2023. *Tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents*. https://github.com/rstudio/tinytex.