README – Data and Code for:

Example for the Reproducible Paper Template: Amazon Priority Municipalities

João Pedro Vieira

May 2023

> IMPORTANT DISCLAIMER: This is just an application of the Reproducible Paper Template. The analysis presented is just an illustration.
>
> IMPORTANT OBSERVATION: This README is written as if all data files were included in the replication package. However, because GitHub has a hard limit of 100 Mb for individual files, it was impossible to include all data files. If you want a complete version (including all data files), see the repository uploaded at Zenodo (Vieira 2023).

Overview

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

- `"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, and computational requirements. It ultimately explains how to replicate the analysis presented in the paper entirely.
- `"code`: a folder containing all scripts to clean, build, merge, and analyze the data
  - `"code/MASTERFILE.R"`: R script to run all scripts from data cleaning to generating the results;
  - `"code/setup.R"`: R script to install/load R packages and configure the initial setup. Uses `"groundhog"` to keep all packages version fixed at the specified date (2023-05-06);
  - `"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 samples 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 function used in multiple R scripts to export the processing time.
- `"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 samples for analysis and intermediate datasets:
    - "/muniLevel": folder with the samples of interest at the *municipality level*.
  - "data/analysis": folder with all regression outputs in "/regressions";
  - "data/_temp": folder to hold temporary files output (filled when running some .R scripts).

- "products": a folder with (this folder is generally omitted when sending a replication package to a journal):
  - "/aux_files": auxiliary files for the paper and slides output;
  - "/paper": a folder with the LaTeX paper template separated into the main paper and appendix;
  - "/aux_files": a folder with the R Markdown beamer slides template for a presentation and the preliminary results (only figures and tables).

- "references": folder with three BibTeX files to record all references for citation (literature: "references_literature.bib", data: "references_data.bib", and software: "references_software.bib") and one subfolder to store the references PDFs ("references_pdf").

- "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 "supportingStats.txt".

- "reproducible_paper_example2.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

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.

License for Data

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.

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

BRAZILIAN BIOMES (IBGE 2019)

- folder file path: "data/raw2clean/biomeDivision_ibge"
- Content: biomes perimeter (polygons data frame); Brazil (extent); 2019 (year of reference)
- source: Instituto Brasileiro de Geografia e Estatistica (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 unzipped the folder and moved the files to the "input" folder, then deleted the "Biomas_250mil" folders.
- provided: yes

BRAZILIAN MUNICIPALITIES (IBGE 2015)

- folder file path: "data/raw2clean/muniDivision_ibge"
- Content: municipal perimeter (polygons data frame); Brazil (extent); 2015 (year of reference)
- source: Instituto Brasileiro de Geografia e Estatistica (IBGE)
- original link: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/15774-malhas.html?=&t=downloads
- raw data downloaded on: SEP/16/2020
- web archive link (used for download): https://web.archive.org/web/20200916142056/ftp://geoftp.ibge.gov.br/organizacao_do_territorio/malhas_territoriais/malhas_municipais/municipio_2015/Brasil/BR/br_municipios.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 unzipped the files and moved them to the "input" folder, then deleted the "br_municipios" folders.
- provided: yes

PRIORITY LIST (MMA 2017)

- folder file path: "data/raw2clean/priorityList"
- Content: list of Legal Amazon priority municipalities with entry/exit dates and legal documentation (.pdf files); Brazilian Legal Amazon (extent); 2017 (year of reference)
- source: MMA (Ministerio do Meio Ambiente)

- original link: http://combateaodesmatamento.mma.gov.br/images/conteudo/lista_municipios_prioritarios_AML_2017.pdf
- raw data downloaded on: SEP/19/2020
- web archive link (used for download): https://web.archive.org/web/20200915211728/http://combateaodesmatamento.mma.gov.br/images/conteudo/lista_municipios_prioritarios_AML_2017.pdf
- raw data archived on: SEP/15/2020
- notes: used the web archive link to access the pdf file. Manually saved the file in the "input" folder
- provided: yes

PRODES AMAZON (LEGAL AMAZON LAND COVER AT MUNICIPALITY LEVEL) (INPE 2020)

- folder file path: "data/raw2clean/prodesAmazon_inpe"
- Content: land cover at the municipality level; Legal Amazon (extent), 2000-2019 (period of reference), yearly (frequency)
- source: Instituto Nacional de Pesquisas Espaciais (INPE)
- original link: http://www.dpi.inpe.br/prodesdigital/prodesmunicipal.php
- raw data downloaded on: OCT/24/2020
- web archive link (used for download):
- [2000] https://web.archive.org/web/20200915164422/http://www.dpi.inpe.br/prodesdigital/tabelatxt.php?ano=2000&estado=&ordem=DESMATAMENTO2000&type=tabela&output=txt
- [2001] https://web.archive.org/web/20200915164614/http://www.dpi.inpe.br/prodesdigital/tabelatxt.php?ano=2001&estado=&ordem=DESMATAMENTO2001&type=tabela&output=txt& 8 [2002] https://web.archive.org/web/20200915164920/http://www.dpi.inpe.br/prodesdigital/tabelatxt.php?ano=2002&estado=&ordem=DESMATAMENTO2002&type=tabela&output=txt
- [2003] https://web.archive.org/web/20200915165014/http://www.dpi.inpe.br/prodesdigital/tabelatxt.php?ano=2003&estado=&ordem=DESMATAMENTO2003&type=tabela&output=txt
- [2004] https://web.archive.org/web/20200915165038/http://www.dpi.inpe.br/prodesdigital/tabelatxt.php?ano=2004&estado=&ordem=DESMATAMENTO2004&type=tabela&output=txt
- [2005] https://web.archive.org/web/20200915165211/http://www.dpi.inpe.br/prodesdigital/tabelatxt.php?ano=2005&estado=&ordem=DESMATAMENTO2005&type=tabela&output=txt
- [2006] https://web.archive.org/web/20200915165249/http://www.dpi.inpe.br/prodesdigital/tabelatxt.php?ano=2006&estado=&ordem=DESMATAMENTO2006&type=tabela&output=txt
- [2007] https://web.archive.org/web/20200915165303/http://www.dpi.inpe.br/prodesdigital/tabelatxt.php?ano=2007&estado=&ordem=DESMATAMENTO2007&type=tabela&output=txt
- [2008] https://web.archive.org/web/20200915165337/http://www.dpi.inpe.br/prodesdigital/tabelatxt.php?ano=2008&estado=&ordem=DESMATAMENTO2008&type=tabela&output=txt
- [2009] https://web.archive.org/web/20200915165353/http://www.dpi.inpe.br/prodesdigital/tabelatxt.php?ano=2009&estado=&ordem=DESMATAMENTO2009&type=tabela&output=txt
- [2010] https://web.archive.org/web/20200915165427/http://www.dpi.inpe.br/prodesdigital/tabelatxt.php?ano=2010&estado=&ordem=DESMATAMENTO2010&type=tabela&output=txt
- [2011] https://web.archive.org/web/20200915165439/http://www.dpi.inpe.br/prodesdigital/tabelatxt.php?ano=2011&estado=&ordem=DESMATAMENTO2011&type=tabela&output=txt
- [2012] https://web.archive.org/web/20200915165504/http://www.dpi.inpe.br/prodesdigital/tabelatxt.php?ano=2012&estado=&ordem=DESMATAMENTO2012&type=tabela&output=txt
- [2013] https://web.archive.org/web/20200915165526/http://www.dpi.inpe.br/prodesdigital/tabelatxt.php?ano=2013&estado=&ordem=DESMATAMENTO2013&type=tabela&output=txt
- [2014] https://web.archive.org/web/20200915165551/http://www.dpi.inpe.br/prodesdigital/tabelatxt.php?ano=2014&estado=&ordem=DESMATAMENTO2014&type=tabela&output=txt
- [2015] https://web.archive.org/web/20200915165615/http://www.dpi.inpe.br/prodesdigital/tabelatxt.php?ano=2015&estado=&ordem=DESMATAMENTO2015&type=tabela&output=txt
- [2016] https://web.archive.org/web/20200915165626/http://www.dpi.inpe.br/prodesdigital/tabelatxt.php?ano=2016&estado=&ordem=DESMATAMENTO2016&type=tabela&output=txt

- [2017] https://web.archive.org/web/20201019163830/http://www.dpi.inpe.br/prodesdigital/tabelatxt.php?ano=2017&estado=&ordem=DESMATAMENTO2017&type=tabela&output=txt
- [2018] https://web.archive.org/web/20201019163935/http://www.dpi.inpe.br/prodesdigital/tabelatxt.php?ano=2018&estado=&ordem=DESMATAMENTO2018&type=tabela&output=txt
- [2019] https://web.archive.org/web/20201019164015/http://www.dpi.inpe.br/prodesdigital/tabelatxt.php?ano=2019&estado=&ordem=DESMATAMENTO2019&type=tabela&output=txt
- raw data archived on: SEP/15/2020 and OCT/19/2020
- notes: manually used the web archive link for each year to download the txt files and saved them in the "input" folder
- provided: yes

Computational requirements

Software Requirements

- 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)
    - did (version 2.1.2) (Callaway and Sant'Anna 2022)
    - kableExtra (version 1.3.4) (Zhu 2021)
    - modelsummary (version 1.4.0) (Arel-Bundock 2023)
    - tabulizer (version 0.2.2) (Leeper 2018)

Memory and Runtime Requirements

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

☒ <10 minutes

Details 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

- `"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/muniLevel/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).

License for Code

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

Instructions to Replicators

- (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.
  - `"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/` `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.
- If you want to re-compile the .tex and .Rmd files in the products folder you might need to do the following:
  - In Tools > Project Options… > Sweave > select `pdfLaTeX` for Typset LaTeX;
  - If there is already a LaTeX distribution installed but rendering .tex and .Rmd files is not working, a possible solution is to install the same tinytex version as used in the template `tinytex::install_tinytex(version = "2023.05")`;
  - If you encounter missing package errors when compiling to pdf a useful solution is to run `tinytex::parse_install(here::here("products/paper/main_paper.log"))` in R (substituting the file name and path with the relevant for your case).

List of tables and programs

The provided code reproduces:

☒ All numbers provided in text in the paper
☒ All tables and figures in the paper

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/supportingStats.R"` and saved in the `"results/stats/supportingStats.txt"` with page location and citation.

Acknowledgements

Adapted structure from Vilhuber et al. (2020).

References

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.

Arel-Bundock, Vincent. 2023. *Modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready*. https://vincentarelbundock.github.io/modelsummary/.

Callaway, Brantly, and Pedro H. C. Sant’Anna. 2022. *Did: Treatment Effects with Multiple Periods and Groups*. https://CRAN.R-project.org/package=did.

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. 2015. “Malhas Muncipais: Shapefile, 2015.” Instituto Brasileiro de Geografia e Estatística (IBGE), Ministério da Economia. Archived at: https://web.archive.org/web/20200916142056/ftp://geoftp.ibge.gov.br/organizacao_do_territorio/malhas_territoriais/malhas_municipais/municipio_2015/Brasil/BR/br_municipios.zip. Archived on: September 16, 2020.

———. 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.

INPE. 2020. “Projeto PRODES - Monitoramento Da Floresta Amazônica Brasileira Por Satélite: Desmatamento Nos Municípios, 2000-2019.” Coordenação-Geral de Observação da Terra (OBT), Instituto Nacional de Pesquisas Espaciais (INPE), Ministério da Ciência, Tecnologia e Inovação (MCTI). Available at: http://www.dpi.inpe.br/prodesdigital/prodesmunicipal.php. Accessed on: October 24, 2020.

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

Leeper, Thomas J. 2018. *Tabulizer: Bindings for Tabula PDF Table Extractor Library*. https://github.com/ropensci/tabulizer.

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

MMA. 2017. “Lista de Municípios Prioritários Da Amazônia: 2008-2017.” Ministério do Meio Ambiente (MMA). Archived at: https://web.archive.org/web/20200915211728/http://combateaodesmatamento.mma.gov.br/images/conteudo/lista_municipios_prioritarios_AML_2017.pdf. Archived on: September 15, 2020.

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.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with Kable and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.