README – Data and Code for: “Reproducible Paper Example”

João Pedro Vieira

03/03/2021

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 not possible. If you want to see a complete version (including all data files) see the repository uploaded at Zenodo (PENDING - LINK TO ZENODO).

Overview

The code in this replication package cleans the raw data for each data source (4 sources), extracts the relevant information specifically for the project, combines it, and generates the results using R. A master file run all of the code to generate the data for the PENDING figures and PENDING tables in the paper. The replicator should expect the code to run for about PENDING hours divided into 35 minutes in the cleaning part, 45 minutes in the extraction/merge part, and PENDING in the final analysis part. 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.

Files Structure

1. `"README.md"`: a markdown file used to generate `"README.pdf"` and `"README.html"`.
2. `"README.pdf"`: This document. It provides the necessary information about the structure of the replication folder, data sources and access, computational requirements, and ultimately explains how to fully replicate the analysis presented in the paper. Also available in html format `"README.html"` (best for reading).
3. `"code`: a folder containing all scripts to clean, build, merge, and analyze the data

    a. `"code/raw2clean"`: R scripts that clean the data on input and save on the output for each dataset;
    b. `"code/projectSpecific"`: R scripts that select/combine/construct the information relevant for this project and creates the sample for analysis;
    c. `"code/analysis"`: R scripts to generate the results presented in the paper;
    d. `"code/_functions"`: auxiliary folder with custom R functions used in multiple R scripts.

4. `"data"`: a folder containing data in a variety of formats: raw, cleaned, intermediate, final datasets for analysis, and analysis outputs

    a. `"data/raw2clean"`: one folder for each dataset with the following structure:
        - `"input"`: raw datasets;
        - `"/output"`: cleaned dataset;
        - `"documentation"`: with at least two files:
            - `"_metadata.txt"` text file that describes the data, provides access instructions, and an example of citation following the AEA guidelines;

    - "codebook_datasetName.txt" text file with summary statistics and variables description;

  b. "data/projectSpecific": sample of interest, intermediate datasets with the variables of interest, and merged sample for analysis;
  c. "data/analysis": analysis outputs, including all figures and tables presented in the paper;
  d. "data/_temp": temporary files output (to be filled when running some .R scripts).

5. "reproducible_paper_example.Rproj": R project to automatically adjust file path references. Always open RStudio from this file when running any R script.

6. "renv": a folder to be filled with an isolated library containing all R packages and dependencies with the correct versions. It also contains:

   a. "active.R": a script to activate the renv structure when using the project for the first time
   b. "settings.dcf": a file to store the settings used in the renv project

7. "renv.lock": a file containing the specifications of R and its dependencies necessary for renv to restore them to a new computer.

8. ".Rprofile": a file that will be automatically sourced whenever you open the RProject ("reproducible_paper_example.Rproj") and that will source "active.R" script to guarantee that the renv structure is being used.

9. "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, 2021). 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 Division Data on Brazilian Biomes Division were downloaded from the Instituto Brasileiro de Geografia e Estatística (IBGE) (IBGE, 2019). Data can be directly downloaded from https://web.archive.org/web/20200916173523/ftp://geoftp.ibge.gov.br/informacoes_ambientais/estudos_ambientais/biomas/vetores/Biomas_250mil.zip. The link will download a zip file containing multiple files that compose the shapefile (.shp, .prj, .shx, .sbn, .xml, .cpg). The zip file was manually unzipped and its files were moved to the `"raw2clean/biomeDivision_ibge/input"` folder. A copy of the data is provided as part of this archive. The data are in the public domain.

Datafiles: multiple files with pattern `"lm_bioma_250"`

Codebook: `"raw2clean/biomeDivision_ibge/documentation/codebook_biomeDivision.txt"`

Metadata: `"raw2clean/biomeDivision_ibge/documentation/_metadata.txt"`

Brazilian Municipality Division Data on Brazilian Municipality Division were downloaded from the Instituto Brasileiro de Geografia e Estatística (IBGE) (IBGE, 2015). Data can be directly downloaded from 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. The link will download a zip file containing multiple files that compose the shapefile (.shp, .prj, .shx, .dbf, .cpg). The zip file was manually unzipped and its files were moved to the `"raw2clean/muniDivision2015/input"` folder. A copy of the data is provided as part of this archive. The data are in the public domain.

Datafiles: multiple files with pattern `"BRMUE250GC_SIR"`

Codebook: `"raw2clean/muniDivision2015/documentation/codebook_muniDivision2015.txt"`

Metadata: `"raw2clean/muniDivision2015/documentation/_metadata.txt"`

Amazon Priority Municipalities Data on Amazon Priority Municipalities were downloaded from the Ministério do Meio Ambiente (MMA) (MMA, 2017). Data can be directly downloaded from https://web.archive.org/web/20200915211728/http://combateaodesmatamento.mma.gov.br/images/conteudo/lista_municipios_prioritarios_AML_2017.pdf. The link will open a pdf file. The pdf file was manually saved in the `"raw2clean/priorityMuniAmazon/input"` folder. A copy of the data is provided as part of this archive. The data are in the public domain.

Datafiles: `"lista_municipios_prioritarios_AML_2017.pdf"`

Codebook: `"raw2clean/priorityMuniAmazon/documentation/codebook_priorityMuniAmazon.txt"`

Metadata: `"raw2clean/priorityMuniAmazon/documentation/_metadata.txt"`

Legal Amazon PRODES Deforestation Municipality-Level Data on Legal Amazon PRODES Deforestation Municipality-Level were downloaded from the Instituto Nacional de Pesquisas Espaciais (INPE) (MMA, 2001-2020). Data can be directly downloaded for each year using the following links: 2000; 2001; 2002; 2003; 2004; 2005; 2006; 2007; 2008; 2009; 2010; 2011; 2012; 2013; 2014; 2015; 2016; 2017; 2018; 2019

The links will download txt files. The txt files were saved in the `"raw2clean/prodesDeforestationAmazonMuni/input"` folder. A copy of the data is provided as part of this archive. The data are in the public domain.

Datafiles: 20 files with pattern `"DesmatamentoMunicipiosYYYY.txt"` with YYYY going from 2001 through 2019

Codebook: `"raw2clean/prodesDeforestationAmazonMuni/documentation/codebook_prodesDeforestationAmazonMuni.`

Metadata: `"raw2clean/prodesDeforestationAmazonMuni/documentation/_metadata.txt"`

Computational requirements

Software Requirements

- R 4.0.2

  - the file "`renv.lock`" has all R packages and dependencies version used in the project.
  - the file "`reproducible_paper_example.Rproj`" will guarantee that the working directory is set to the root of the project (always open RStudio using this file).

Memory and Runtime Requirements

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

☒ 1-8 hours

Details The code was last run on a **4-core Laptop; Intel Core i7-855U CPU @ 1.80 GHz processor; 16GB RAM; Windows 10 Home**.

Total disk size (expected) to be consumed by the project considering everything (including intermediate dataset, libraries, etc.) in an uncompressed format is approximately 1GB (~450 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_muniLevel.R`" will construct the base sample, extract the information from each dataset relevant for this paper, construct the variables of interest, merge them with the base sample, and generate the sample for analysis in multiple formats: cross-section, spatial, panel (4 scripts).
  - "`code/analysis/masterfile_analysis.R`" will generate all tables and figures (PENDING scripts).

License for Code

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

Instructions to Replicators

- Download replication package.
- Open RStudio using "`reproducible_paper_example.Rproj`" to set the working directory to the project root.
- Run "`renv::restore`" in the R console. Write "`y`" in the R console to answer the question "`Do you want to proceed? [y/N]:`".
- Run "`code/_MASTERFILE.R`" to run all R scripts in sequence.

Details

- "`reproducible_paper_example.Rproj`": will bootstrap renv package the first time it is opened.
- "`renv::restore`": will install all the necessary R packages and dependencies with the specified versions.

List of tables and programs

PENDING - will be completed in the next version after including the analysis examples.

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.

Figure/Table #	Program	Line Number	Output file	Note
Table 1	02_analysis/table1.do		summarystats.csv	
Table 2	02_analysis/table2and3.do	15	table2.csv	
Table 3	02_analysis/table2and3.do	145	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

Acknowledgements

Adapted from Villhuber et al (2020).

References

Instituto Brasileiro de Geografia e Estatística (IBGE). 2015. “Malhas Muncipais: shapefile, 2015.” Instituto Brasileiro de Geografia e Estatística, Ministério da Economia. 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 (accessed via Archive.org September 16, 2020).

Instituto Brasileiro de Geografia e Estatística (IBGE). 2019. “Biomas do Brasil: shapefile, 2019.” Instituto Brasileiro de Geografia e Estatística, Ministério da Economia. https://web.archive.org/web/20200916173523/ftp://geoftp.ibge.gov.br/informacoes_ambientais/estudos_ambientais/biomas/vetores/Biomas_250mil.zip (accessed via Archive.org September 16, 2020).

Instituto Nacional de Pesquisas Espaciais (INPE). 2001-2020. “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). http://www.dpi.inpe.br/prodesdigital/prodesmunicipal.php (accessed October 24, 2020).

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

Vilhuber, L., Connolly, M., Koren, M., Llull, J., and Morrow, P.. 2020. “A template README for social science replication packages (Version v1.0.0)”. Zenodo. http://doi.org/10.5281/zenodo.4319999 (accessed March 2, 2021)