Readme File for the Replication Package of "Exports, Labor Markets, and the Environment: Evidence from Brazil"

Carlos Góes   Otavio Conceição   Gabriel Lara Ibarra   Gladys Lopez-Acevedo
IFC   *World Bank*   *World Bank*   *World Bank*

July 24, 2025

1   Overview

This document is a guide for reproducing the results presented in the paper "Exports, Labor Markets, and the Environment: Evidence from Brazil". It presents a step-by-step procedure to ensure accurate reproduction.

2   General folder structure

This package contains the following subfolders:

- `../code/`: contains Stata do-files and Python scripts used to pre-process, produce figures, and replicate regressions.

    - `../code/%00_base.do`: sets globals, calls the following subroutines:

        * `../code/%01_stata_preprocessing.do`: runs all python pre-processing files, produces some Figures, calls subroutines in folder `../code/%stata_preprocess/`
        * `../code/%02_python_preprocessing.py`: runs all python pre-processing files, calls subroutines in folder `../code/%python/`
        * `../code/%03_stata_main.do`: runs all the main regressions, calls subroutines in folder `../code/%stata_mun/`

- `../data/`: contains required datasets for the analysis; some of these are not publicly available, and we detail availability and details in the next section.

- `../figs/`: contains figures that the replication file outputs.

- `../logs/`: contains log-files that the replication file outputs.

- `../results/`: contains result tables and files that the replication file generates.

- `../temp/`: temporary files created during the estimation process.

In general, if you have access to all the data required to run analysis, you would run the master do-file named `../code/%00_base.do` to replicate the results of the paper. Ensure that the global "root" - which defines the main directory - is updated before executing the master do-file.

3 Data sources and pre-processing

This section describes how to have access to each piece of the raw data required for reproduction. Some of the data we use in our paper requires pre-processing.

- **CNAE and ISIC Concordances**: CNAE (*Classificação Nacional de Atividades Econômicas*) is the official economic activity statistical code in Brazil and is published by the Brazilian national statistical office (IBGE). There exist multiple versions of the statistical classification: CNAE 95, CNAE 1.0, CNAE 2.0 (which are used in RAIS microdata), CNAE Domiciliar 1.0 and CNAE Domiciliar 2.0 (which are used in censuses microdata). In this paper, we used available concordances to transform other versions into CNAE95.

    - CNAE 2.0 to CNAE 1.0. [link]
    - CNAE 1.0 to CNAE95. [link]

    We also used official concordances to map CNAE95 to ISIC 3.0, which will be our baseline industry classification in this paper. Mapping economic activities to industries allows us to join the trade dataset to economic activities.

    - CNAE X ISIC/CIIU 3.0 [link]

- **RAIS - Relação Anual de Informações Sociais**: RAIS is an annual census of formal workers administered by the Brazilian Ministry of Labor, containing detailed information about nearly the universe of formal employees in the country. Employers are required to submit information about their employees to RAIS every year and face penalties for non-compliance with submission deadlines, ensuring high accuracy of reported information. We observe employees[1] and firm[2] unique tax identifiers and utilize municipality and industry codes of firms and wages and demographic characteristics of workers to aggregate data at the region-industry-year level.

    The RAIS datasets are non-public and were obtained through a formal data agreement between the Brazilian Ministry of Labor and the World Bank. There are two main differences between the RAIS data sets used in the paper and the public microdata: i) we transformed the RAIS' original classification of economic activities from the different years between 2000 and 2020 to CNAE 95 (to further combine RAIS data with the environmental taxonomy of economic activities in Brazil) and ii) conducted additional data cleaning and harmonization procedures. The reproducibility package does not include the RAIS data set used to generate the paper results.

---

[1] Natural Persons Registry, *Cadastro de Pessoas Físicas*.
[2] National Registry of Legal Entities, *Cadastro Nacional de Pessoas Jurídicas*.

- **Brazilian Population Censuses 2000 and 2010**: The paper uses the Brazilian Population Census public microdata from the latest two waves available at the time this version of the paper was prepared: 2000 and 2010. The microdata corresponds to a sample of 10 percent of the universe of households visited by census surveyors and is published by IBGE. The reproducibility package does not include these data sets, but they can be downloaded using the links below.

    - 2000 Brazilian Population Census microdata [link]
    - 2010 Brazilian Population Census microdata [link]

- **UNCOMTRADE Data**: The United Nations Comtrade database aggregates detailed global annual and monthly trade statistics by HS6 product-line, source country, destination country, and period. We use the dollar value of bilateral trade recorded at the year-source-destination-HS6 level. We use UNCOMTRADE's bulk download API to bulk download a yearly file {YEAR}.csv that contains every observation in the database for every reporting country for a given $YEAR \in \{2000, \cdots, 2020\}$.

- **HS to ISIC concordances**: UN COMTRADE data are published at the HS 6-digit product level. We use the concordances from the United Nations Statistics Division to harmonize the HS vintages and map them to ISIC rev. 3 3-digit industry codes. First, we harmonize the HS6 vintages using concordances from the United Nations Statistics Division. Then we map them to ISIC using concordances from WITS. The processing of these concordances will be part of the routine, so there is no pre-processing required. In the next section we will refer back to these files.

    - Complete Mapping HS 1996 to HS 2022, United Nations Statistics Division [link]
    - HS2007 (H3) to ISIC Rev. 3, World Integrated Trade Solution [link]

- **Ministry of Industry and Commerce's (MDIC) Customs Data**: Brazilian trade data comes from customs records. The Ministry of Industry and Commerce of Brazil (MDIC) publishes customs records at the municipality level. Location of imports and exports are recorded based on the address reported by the importing/exporting firm. Municipal exports are published at the harmonized system (HS) 4-digit product level and 3-digit ISIC industry level. We observe flows quantities and Free on Board (FOB) nominal dollar values and use total FOB dollar values as a metric of municipal exports for a given $YEAR \in \{2000, \cdots, 2020\}$. Data are publicly available [link].

- **Environmental taxonomy of economic activities in Brazil**: This data set is public and published by the Brazilian Federation of Banks (FEBRABAN). Since we transformed the original classification of economic activities of RAIS and censuses microdata from different years to CNAE95, we also transformed the FEBRABAN classification (CNAE 2.0) into CNAE95 so that we can have all data sets with the same classification.

    - FEBRABAN taxonomy [link]

- **PNAD 2001-2009 and 2011**: PNAD was the main national household survey to track socioeconomic indicators in Brazil until the emergence of PNAD-C in 2012. The raw files were download from IBGE's website and we used the Stata package called Datazoom the import the raw PNAD files into Stata. To facilitate, we included the PNAD Stata files in the reproducibility package in `../data/PNAD`.

    - PNAD 2001-2011 raw files [link]

- **PNAD-C 2012-2020**: PNAD-C microdata is public and published by IBGE. We used harmonized PNAD-C data sets for Brazil prepared by SEDLAC (World Bank and CEDLAS), which are data sets of restricted use and intended for internal World Bank use only. The reproducibility package does not include the SEDLAC-harmonized PNAD-C Stata files. In principle, however, the results can be replicated with the public version of PNAD-C microdata published by IBGE using PNAD-C microdata for the first visit for 2012-2019, and for the fifth visit for 2020:

    - PNAD-C raw files for 2012-2019 (first visit) [link]
    - PNAD-C raw file for 2020 (fifth visit) [link]

- **CPI deflators**: To transform the monetary variables into real terms, we use the deflators of the official inflation indicator in Brazil (IPCA) and which are published by IBGE.

- **Brazilian National Statistical Office (IBGE) population estimates**: We use IBGE's official numbers for the population of each municipality for the years of 2000 [link] and 2010 [link].

4 Routines Description

4.1 Pre-processing Stata routines

The file `../code/%01_stata_preprocessing.do` executes the pre-processing Stata routines. The do-files executed by the file are:

1. `../code/%stata_preprocess/1_IPCA_deflator.do`: this do-file imports the file `{ipca_202502SerieHist}.xls`, which contains the CPI deflators used to deflate monetary variables. The original file contains the CPI index numbers since the 1990s, and the do-file consider the average deflators at the yearly level for the year of 2010. It outputs:
   `..temp/yearly_adjustment_factors_IPCA_at_2010_prices.dta`.

2. `../code/%stata_preprocess/2_CNAE_concordances.do`: this do-file imports the IBGE-provided concordances between different versions of the national classification of economic activities (CNAE), and the CNAE95 structure. A detailed description of how we dealt with the concordances is described in Appendix B of the paper.

3. `../code/%stata_preprocess/3_Census_preparation.do`: this do-file imports the 2000 and 2010 Brazilian Population Censuses microdata. It takes advantage of the concordances between different version of CNAE produced in Step 4.1.2 to transform the censuses microdata's original classification to CNAE95 (recall that the 2000 Census uses CNAE Domiciliar 1.0, while the 2010 Census uses CNAE Domiciliar 2.0). It outputs:
   `..temp/overall_formal_and_informal_2000_and_2010_census_municipality_CNAE95.dta`

4. `../code/%stata_preprocess/4_FEBRABAN_taxonomy.do`: this do-file imports the FEBRABAN taxonomy of economic activities and transform its original classification (CNAE 2.0 at the 5-digit industry level) into CNAE95 at the 5-digit industry level by leveraging the concordances produced in Step 4.1.2. It outputs:
   `..temp/CNAE95_at_the_5_digit_with_green_classification_FEBRABAN.dta`.

5. `../code/%stata_preprocess/5_RAIS_preparation.do`: this do-file processes the yearly RAIS files from 2000 to 2020 by (i) converting the monetary variables into real terms using the deflators from Step 1 of this section, (ii) transforming the CNAE classification into CNAE95, and (iii) calculating the number of workers and average wages of these workers in each municipality (with disaggregations by schooling and gender) and year. It outputs:
   `..temp/RAIS_data_2000_2020_by_municipality_and_CNAE95_5digits.dta`.

6. `../code/%stata_preprocess/6_PNAD_and_PNAD-C.do`: this do-file leverages the PNAD and PNAD-C Stata files to calculate the unemployment rate by region in Brazil between 2001 and 2020 (except for 2010). We saved these numbers in a separate Excel spreadsheet, stored as
   `../data/ibge/employment_data.xlsx`.

4.2 Python routines

The file `../code/%02_python_preprocessing.py` executes all necessary Python routines. The scripts executed by the master file are:

1. `../code/%python/comtrade/comtrade_global_exports.py`: this file processes a file `{YEAR}.csv`, which contains the dollar value of bilateral trade recorded at the year-source-destination-HS6 level, downloaded using UNCOMTRADE's bulk download API. We cannot redistribute the raw data. A user who downloads the entire yearly dataset should save it as `../data/comtrade/bulk/{YEAR}.csv` for $YEAR \in \{2000, \cdots, 2020\}$. Using those data, this python script will collapse global exports for all countries other than Brazil, per year, by HS6 product line.

2. `../code/%python/trade-processing/newTradeProcessing.py`: this file processes a file `{YEAR}.csv`, which contains the dollar value of bilateral trade recorded at the year-source-destination-HS6 level, downloaded using UNCOMTRADE's bulk download API. We cannot redistribute the raw data. A user who downloads the entire yearly dataset should save it as `../data/comtrade/bulk/{YEAR}.csv` for $YEAR \in \{2000, \cdots, 2020\}$. Using those data, this python script will collapse global exports for all countries other than Brazil, per year, by HS6 product line.

3. `../code/%python/trade-processing/summaryStats.py`: uses consolidated exports from the previous file, concordances from NCM to ISIC and dollar deflators to consolidate summary statistics: exports at the 1-digit and 2-digit industry level (ISIC).

4. `../code/%python/figs-stylized-facts/chartTimeSeries.py`: uses the files outputted from the previous script to produce Figure 1.

5. `../code/%python/mun-trade/exports-map.py`: uses data on municipal level exports (processed in some of the previous files) and data on municipal population from IBGE, combined with official shapefiles, to produce Figure 2.

6. `../code/%python/figs-stylized-facts/chartUnemployment.py`: uses data on state unemployment levels (stored in a separate Excel spreadsheet) generated in Step 4.1.6 to produce Figure 3.

7. `../code/%python/gren-map/green-map.py`: combines the data on economic activities with RAIS data to create the number and share of employment for environmentally risky and environmentally sustainable activities at a given Census year (2000, 2010). Density plots are in Figure 6. It then combines it with combined with official shapefiles, to produce Figure 4.

4.3 Main Stata routines

The file `../code/%03_stata_main.do` executes all the main Stata routines. The do-files executed by the file are:

1. `../code/%stata_mun/d00_prepare_comtrade.do`: this file starts from the files produced by step 4.2.1, stored as `../data/comtrade/{YEAR}-hs6.csv`. It then (a) uses the official UNSD concordances to harmonize the different vintages of HS codes; and (b) map harmonized concordances to ISIC rev 3. The output file is a panel of global exports at the 3-digit ISIC industry. It outputs: `../data/temp/panel-isic3-3d.dta`.

2. `../code/%stata_mun/d01_prepare_rais.do`: this file processes the individual level microdata from RAIS and collapses it at the industry-municipality-year level. It starts from a panel of individual-level observations, pre-processed in step 4.1.5, transforms the data by summing over individual data for employment, gender, schooling levels, and reported wages, and averages out wages for municipalities. The output file is a panel of labor market indicators at the 3-digit ISIC industry $\times$ municipality. It outputs: `../data/temp/rais-panel-processed-mun.dta`.

3. `../code/%stata_mun/d02_instrument_gdp.do`: this file creates the alternative instrument, based on trade partners GDP growth, per equation (3). It calculates (municipality) labor shares and (industry) export destination shares from the two files above plus GDP growth from WDI. The outcome is a panel at the municipality level with exposure to foreign demand shocks. It outputs: `../data/temp/mun-trade-panel-inst.dta`.

4. `../code/%stata_mun/d03_instrument_comtrade.do`: this file creates the baseline instrument, based on global exports growth at the industry level, per equation (1). It calculates (municipality) labor shares and (industry) exports from steps 1 and 2 in this section. It also merges data with the alternative instrument above. It outputs: `../data/temp/mun-trade-panel-processed.dta`.

5. `../code/%stata_mun/d04_tradable_group.do`: this file creates a panel municipal headcounts for employment in tradable employment. It also merges data with the alternative instrument above. It outputs: `../data/temp/mun-trade-panel-processed.dta`.

6. `../code/%stata_mun/d05_microregion_exp.do`: this file reads data on municipal exports, generated through Python scripts on step 4.2.2, merges it with our labor market dataset created above, and calculates growth rates. It outputs: `../data/temp/mun-trade-panel-processed-exp.dta`.

7. `../code/%stata_mun/d06_green_decomp.do`: this file reads the taxonomy of economic activities file, constructed on step 4.1.4, merges it with RAIS, and creates a municipal headcount of employees by "environmentally sustainable" and "environmentally risky" activities for a given year. It outputs: `../data/temp/mun-master-green.dta`.

8. `../code/%stata_mun/d07_lhs_const.do`: this file will construct the left-hand side lag/lead variables necessary to estimate the local projections defined by equation (2). It also labels all of the variables appropriately. It outputs: `../data/temp/mun-master-dataset.dta`.

9. `../code/%stata_mun/d08_prepare_census.do`: this file processes the individual level microdata from Decennial Census and collapses it at the industry-municipality-year level. It starts from

a cross-sections of individual-level observations for $year \in \{2000, 2010\}$, pre-processed in step 4.1.3, transforms the data by summing over individual data for employment, gender, schooling levels, and reported wages, and averages out wages for municipalities. The output file is a cross-section of decennial growth rates for labor market indicators at the 3-digit ISIC industry $\times$ municipality. It outputs: `../data/temp/mun-census-growth-rates.dta`.

10. `../code/%stata_mun/d09_census_instrument_gdp.do`: this file creates the alternative instrument for the long-difference specification (5). It calculates (municipality) labor shares and (industry) export destination shares plus GDP growth from WDI. The output file is a cross-section of decennial exposure to foreign demand shocks. It outputs:
    `../data/temp/mun-census-microregion-trade-panel-processed.dta`.

11. `../code/%stata_mun/d10_instrument_comtrade.do`: this file creates the baseline instrument. It calculates (municipality) labor shares and (industry) exports from steps 1 and 2 in this section. It also merges data with the alternative instrument above. The output file is a cross-section of decennial exposure to foreign demand shocks. It outputs:
    `../data/temp/mun-census-microregion-trade-panel-processed.dta`.

12. `../code/%stata_mun/d11_census_exp.do`: this file reads data on municipal exports for $year \in \{2000, 2010\}$ and calculate growth rates, generated through Python scripts on step 4.2.2, calculates growth rates, and merges it with our labor market and instruments dataset created above. It outputs:
    `../data/temp/mun-master-dataset-census.dta`.

13. `../code/%stata_mun/r01_inst_validation.do`: this file performs validation checks on the different instruments. It runs the first-stage regressions and produces Figure 5. It also compares the instruments and produces Figures A.3, A.4, and A.5 in the Appendix.

14. `../code/%stata_mun/r02_irfs_comtrade_ptrend.do`: this file runs the local projection regressions with the baseline instrument.

15. `../code/%stata_mun/r03_irfs_gdp_ptrend.do`: this file runs the local projection regressions with the alternative instrument.

16. `../code/%stata_mun/r04_census_reg.do`: this file runs the regressions with Census data and produces Table 1. It outputs the file `../results/mun-census.tex`, whose content should be automatically updated after running this file.

17. `../code/%stata_mun/g01_plot_irfs_gdp.do`,
    `../code/%stata_mun/g02_plot_irfs_comtrade.do`, and
    `../code/%stata_mun/g03_plot_irfs_comtrade_combined.do`: these files plot charts.

5 List of exhibits

Exhibit name	Do-file/script used
Figure 1	`../code/%python/figs-stylized-facts/chartTimeSeries.py`
Figure 2	`../code/%python/mun-trade/exports-map.py`
Figure 3	`../code/%python/figs-stylized-facts/chartUnemployment.py`
Figure 4	`../code/%python/gren-map/green-map.py`
Figure 5	`../code/%stata_mun/r01_inst_validation.do`
Figure 6	`../code/%python/gren-map/green-map.py`
Figure 7	`../code/%stata_mun/g02_plot_irfs_comtrade.do`
Figure 8	`../code/%stata_mun/g02_plot_irfs_comtrade.do`
Figure 9	`../code/%stata_mun/g03_plot_irfs_comtrade_combined.do`
Figure 10	`../code/%stata_mun/g03_plot_irfs_comtrade_combined.do`
Table 1	`../code/%stata_mun/r04_census_reg.do`
Appendix Figure 1	`../code/%stata_mun/g03_plot_irfs_comtrade_combined.do`
Appendix Figure 2	`../code/%stata_mun/g03_plot_irfs_comtrade_combined.do`
Appendix Figure 3	`../code/%stata_mun/r01_inst_validation.do`
Appendix Figure 4	`../code/%stata_mun/r01_inst_validation.do`
Appendix Figure 5	`../code/%stata_mun/r01_inst_validation.do`
Appendix Figure 6	`../code/%stata_mun/g02_plot_irfs_comtrade.do`
Appendix Figure 7	`../code/%stata_mun/g02_plot_irfs_comtrade.do`

Exhibit name	File name
Figure 1	timeSeriesReal.pdf
Figure 2	map_VL_FOBrpc_2000.pdf ; map_VL_FOBrpc_2010.pdf
Figure 3	unemployment.pdf
Figure 4	**Panel A:** map_emp_alto_risco_shr_2000.pdf ; map_emp_alto_risco_shr_2010.pdf **Panel B:** map_emp_economia_verde_shr_2000.pdf ; map_emp_economia_verde_shr_2010.pdf
Figure 5	mun-first-stage-comtrade.pdf
Figure 6	**Panel A**: emp_alto_risco_shr.pdf ; emp_economia_verde_shr.pdf **Panel B**: emp_alto_risco_shr_f_2000.pdf ; emp_economia_verde_shr_f_2000.pdf
Figure 7	mirf_giv_comtrade_employment.pdf
Figure 8	mirf_giv_comtrade_w.pdf
Figure 9	mgiv_comtrade_green.pdf
Figure 10	mgiv_comtrade_green2.pdf
Table 1	mun-census.tex
Appendix Figure 1	mgiv_comtrade_college.pdf
Appendix Figure 2	mgiv_comtrade_male_female.pdf
Appendix Figure 3	mun-first-stage.pdf
Appendix Figure 4	mun-iv-hist.pdf
Appendix Figure 5	mun-first-stage-ivs.pdf
Appendix Figure 6	mirf_giv_employment.pdf
Appendix Figure 7	mirf_giv_w.pdf