

README - Replication package

Algorithms and Bureaucrats: Evidence from Tax Audit Selection in Senegal

October 17, 2025

Pierre Bachas Anne Brockmeyer Alipio Ferreira Basirou Sarr

Contact information

- Roldan Enamorado (Team Member/Consultant): renamoradoirias@worldbank.org
- Anne Brockmeyer (Author): abrockmeyer@worldbank.org

Contents

1. [Overview](#)
2. [Data Availability](#)
3. [Instructions for Replicators](#)
4. [List of Exhibits](#)
5. [Requirements](#)
6. [Code Description](#)
7. [Folder Structure](#)

Overview

This README accompanies the replication package for “Algorithms and Bureaucrats: Evidence from Tax Audit Selection in Senegal” (Bachas, Brockmeyer, Ferreira, Sarr). It explains how to reproduce all empirical exhibits (tables and figures) using confidential administrative records and surveys. The workflow is: (i) construct audit: firm: year datasets; (ii) run main and robustness specifications; and (iii) export manuscript ready LaTeX outputs. This document contains a brief description of the process to generate the outputs, instructions to carry out the reproduction of the tables and figures, computational requirements and a visual representation of the folder structure to guide the replicators.

Verification materials are shared via Dropbox. The top level folder `replication_package` contains all code and documentation required to run the pipeline. Its main components are:

- Four “Code” subfolders
- “Intermediate data” (temporary harmonized files) and “Working data” (analysis ready datasets)
- “Output” (all tables and figures)
- “Documentation” (this README and an Excel Readme appendix with the Data Availability Statement and a script: `input: output crosswalk`)

- A copy of the manuscript in pdf format.

Data processing has three layers. First, Stata prepares the data (cleaning, harmonization, and merging DGID administrative tax records, audit reports, and survey data) to build an audit: firm: year panel. Second, analysis is conducted in Stata (econometric estimations) and in R (machine learning via random forest). Third, LaTeX exports are produced. A separate “Code Additional Figures” subfolder contains stand alone Stata programs for exhibits not directly generated by the econometric and ML pipeline.

Raw confidential data are stored in the project Dropbox workspaces “Senegal tax audits” and “Infractions saisie.” These sources (taxpayer registry, declarations, audit assignments and outcomes, and survey microdata) are transformed into harmonized intermediate files and merged into the master analysis dataset `datasetforanalysis.dta`, which feeds all downstream steps (with an additional preparation step for the R analysis). Principal Code Subdirectories are set in line with data analysis workflow; a brief description is made as follows:

1. Code Prepare Tax Data: Imports and cleans self declarations and DGID registries; normalizes identifiers; constructs firm: year panels and core variables (lags, transformations).
2. Code Prepare Analysis Data: Integrates audit program records and survey data; merges with outputs from Step 1; produces `datasetforanalysis.dta` for analysis.
3. Code Analysis Data: Stata estimation scripts for descriptive and econometric results; automated generation of tables and figures. Includes an “R analysis” subfolder implementing the ML algorithm and related outputs.
4. Code Additional Figures: Stand-alone Stata code for exhibits not produced by the main econometric and ML pipeline (e.g., Comparison of algorithm gains with in-code manually inputted estimates found in the literature review, aggregate revenue statistics of Senegal using the PFR database from the WB).

Data Availability

Because of confidentiality agreements with the Senegal Tax Authority and the sensitive nature of the data, the underlying microdata cannot be made publicly available. For dataset descriptions, sources, and access dates, see the Excel file, sheet “Raw Datasets: DAS”.

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

Data Sources

Main data source for this project

- Tax declarations and administrative data of third-party reports, audit process, risk scores, and general characteristics of taxpayers recorded by the Senegal tax administration (DGID).

Additional Information

- Primary data was collected by the researchers by implementing surveys on a sample of audited taxpayers and inspector's involved in audit programs.

Auxiliary data

- Aggregated tax revenue statistics for Senegal were obtained via the PFR database from the World Bank's Knowledge Product Resource Center.
- Firm census data provided by the National Agency of Statistics and Demography of Senegal.
- Google Maps data was extracted to obtain the distances of the taxpayers address to the DGID office.

Please refer to the excel readme file for more information on the complete lists of files used for this project.

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.
- ☐ I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package. Appropriate permissions are documented in the LICENSE.txt file.

Instructions for Replicators

This replication package provides the complete set of datasets and code required to reproduce the paper's results. Please first download local copies of the Dropbox folders Senegal Tax Audits and Infraction-saisie; these contain the datasets called by the scripts. The codebase is organized by workflow stage, with each folder containing a master script that executes the relevant data-preparation and analysis tasks. To obtain the manuscript outputs, execute the master scripts sequentially in the following order:

Step I: Run dofiles in Code Prepare Tax Data

1. Open do file Master Tax.do in subfolder: Dropbox\Senegal tax audits\Analysis all data\replication_package\Code Prepare Tax Data
2. Uncomment lines 28 to 33 and replace with your OS username in line 29 where it says YOUR COMPUTER'S USERNAME.
3. Add the path to the local directory where "Senegal Tax Audits" is stored.
4. Run the dofile, which will run the following sub-codes:

- 0 Read ANSD data.do
- 1 Read tax data.do
- 2 Read TP, customs and treasury data.do
- 3 Annexes TVA.do
- 4 Read Repertoires.do
- 5 Read Past audits.do
- 6 Collapse and merge datasets.do
- 7 Cleaning and sanity checks.do

Step II: Run dofiles in Code Prepare Analysis Data ¹

1. Open do file MASTER Prepare Analysis.do in subfolder: Dropbox\Senegal tax audits\Analysis all data\replication_package\Code Prepare Analysis Data
2. Uncomment lines 28 to 34 and replace with your OS username in line 29 where it says YOUR COMPUTER'S USERNAME.
3. Replace line 30 with the path to “Senegal tax audits” folder.
4. Replace line 31 with the path to the local Dropbox directory.
5. Run the dofile, which will run the following sub-codes:

- 1 Selection.do
- 2 read saisie data.do
- 3 read suivi data.do
- 4 merge selection and audits.do
- 5 Tax declarations data.do
- 6 Inspector survey.do
- 7 taxpayer survey.do
- 8 Make final dataset for analysis.do

Step III: Run dofiles and R scripts in Code Analyse Data

A. Descriptive Stats and Econometric Analysis in Stata

1. Open do file MASTER Analysis.do in subfolder: Dropbox\Senegal tax audits\Analysis all data\replication_package\Code Analyse Data
 2. Uncomment lines 27 to 31 and replace with your OS username in line 28 where it says YOUR COMPUTER'S USERNAME.
 3. Replace line 29 with the path to “Senegal tax audits” folder.
 4. Run the dofile, which will run the following sub-codes:
- 1 Summary Statistics.do
 - 2 Regressions main results.do
 - 3 Analysis taxpayer survey.do

¹ The do-file 8 Make final dataset for analysis.do also calls the Jupyter notebook “Search coordinates and distances of firms.ipynb” to programmatically retrieve firm coordinates and compute distances using Google Maps. Because these outputs can vary by access date, this step is omitted from the replication run. Instead, the Excel file produced at the time of the study is included in the replication package and is loaded by the do-file.

- 4 Prediction exercise.do
- 5 Deterrence.do

B. Machine Learning estimation in R

1. Open do-file Construct_Data_RF_For_All_Logs.do in subfolder: Dropbox\Senegal tax audits\Analysis all data\replication_package\Code Analyse Data\
2. Uncomment lines 27 to 31 and replace with your OS username in line 28 where it says YOUR COMPUTER USERNAME.
3. Replace line 29 with the path to “Senegal tax audits” folder
4. Run the dofile: this will create datasetforanalysis_predictionexercise_for_all_logs_exercise.dta used as dataframe in R.
5. Open R script 00_Main.R in subfolder: Dropbox\Senegal tax audits\Analysis all data\replication_package\Code Analyse Data\Replication R\
6. Edit the 17 of the script by replacing the path of setwd() with your path to the “replication_package” folder.
7. Run the R script: this step will generate all the outputs derived from the Machine Learning estimations and store them in the “Output” folder.

Generating Additional Figures

1. Figure G.1: Comparison of Optimized Results With Literature:

- Open do file Senegal_Lit_Graph.do in subfolder: Dropbox\Senegal tax audits\Analysis all data\replication_package\Code Additional Figures
- Uncomment lines 27 to 31 and replace with your OS username in line 28 where it says YOUR COMPUTER USERNAME.
- Replace line 29 with the path to “Senegal tax audits” folder and run the do file.

2. Table A.1 Senegal’s Tax Revenue Composition in Comparison

1. This table was generated using the excel PFR (Public Finance Review) Data Visualization Tool from the World Bank, version from October 30, 2024.
2. First, open the document and go to the page “Countries” and select “Senegal” in cell G9.
3. Second, select “Standard Groups in cell D30
4. Go to tab “Review” in toolbar and change click on Unprotect Sheet. Enter password “PFR”
5. Change cell M40 and input “High Income”
6. Press F9 to run macros and update the document with new changes
7. Go to sheet “2_1”
8. Change E93 Initial year = 2021
9. Change H93 Final year = 2022
10. Change K93 Calculation=Last Available Figure
11. Press F9 to run macros and update the document with new changes
12. Data will be located in table “Total tax revenues and social contribution located O71:R87.

List of Exhibits

The outputs generated to include in the manuscript are saved in the subfolder “Output” of the main folder “replication_package”. A list of the filenames generated and their reference to the actual table/figure names in the manuscript can be found in the Excel readme by referring to sheets “

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

Requirements

Computational Requirements

The whole code runs successfully in around 45 – 55 minutes when using a laptop HP Pavilion 14-dv2xxx with the following specifications:

- Processor: 12th Gen Intel(R) Core(TM) i5-1235U (1.30 GHz)
- Installed RAM: 16.0 GB (15.7 GB usable)
- System type: 64-bit operating system, x64-based processor
- Operating system: Windows 11 Home

Software Requirements

The codes require having properly installed the following statistical packages

- **Stata version 17**
- **R version 4.4.3** ran through RStudio 2025.09.0+387 "Cucumberleaf Sunflower".

Memory and Storage Requirements

Since the root directory is approximately 113 GB, you will need a Dropbox plan with sufficient quota and a computer with adequate local storage to synchronize the data and execute the scripts.

Code Description

Please refer to the [Excel Readme](#) file for a list of inputs and outputs per do file or script used in the replication package.

Folder Structure

