README: Can Pollution Markets Work in Developing Countries? Experimental Evidence from India

1 Overview

This is the replication package for "Can Pollution Markets Work in Developing Countries? Experimental Evidence from India" by Michael Greenstone, Rohini Pande, Nicholas Ryan, and Anant Sudarshan.

The package contains code and data files necessary to replicate the analysis presented in the paper. The package uses Stata, R, and MATLAB. Seven main files run all of the code to generate the tables and figures in the paper. `RunReplication.sh` is a shell script that calls these seven files. The replicator should expect the code to run for 10 to 15 minutes on a laptop.

The raw input data is stored in 00RawData, code in 01Code, intermediate and cleaned data in 02DataPipeline, and final exhibits in 03Output. The 04Wrapper folder contains a LaTeX file called `all_exhibits.tex` which compiles all exhibits into a single document. The 05Logs folder contains logs generated by running `RunReplication.sh`. The 06Surveys folder includes the survey instruments.

02DataPipeline, 03Output, and 05Logs folders can all be deleted and can then be exactly reproduced by running `RunReplication.sh`. The only remaining file is `dir_struct.txt` which `RunReplication.sh` uses to recreate the appropriate file directory structure if 02DataPipeline and 03Output are deleted prior to running the replication. Section 2 displays the directory tree, section 3 includes the replication instructions, section 4 describes the data, section 5 includes the hardware and software specifications, and section 6 lists the script that generates each exhibit.

2 Directory Tree

```
root/
├── 00RawData/
│   ├── baseline/
│   ├── cems/
│   ├── emissions/
│   ├── phone_survey/
│   └── trading/
├── 01Code/
│   ├── emissions/
│   │   └── imputation/
│   ├── installers/
│   ├── model/
│   │   ├── cleaning/
│   │   └── matlab/
│   │       ├── @etsCounter/
│   │       ├── @etsData/
│   │       ├── @etsEstim/
│   │       ├── @etsModel/
│   │       ├── cost_benefit_calculation/
│   │       └── exhibits/
│   ├── phone_survey/
│   └── trading/
│       └── construct_datasets/
├── 02DataPipeline/
│   ├── baseline/
│   ├── cems/
│   ├── emissions/
│   ├── phone_survey/
│   └── trading/
├── 03Output/
│   ├── figures/
│   └── tables/
├── 04Wrapper/
├── 05Logs/
└── 06Surveys/
    ├── Baseline/
    ├── Endline/
    └── Phone Survey/
```

3 Instructions to Replicators

The results from the paper can be generated in two ways:

- Option 1. Use `RunReplication.sh` to call all scripts.
- Option 2. Run 9 scripts.

Option 1 is recommended for replicators who already have access to a command line and are able to call software from it. We provide help pages that might be useful to get this set up, but it can be tricky. Option 2 is more straightforward and will suffice to generate all results.

3.1 Option 1: Use `RunReplication.sh` to call all scripts

3.1.1 Software requirements

- Stata 17 or higher with terminal utility ["stata" should be on PATH]. Please find more information about the PATH variable for Windows and for Mac. This help page provides guidance for locating the Stata executable for Mac. This help page provides guidance for locating the Stata executable for Windows.
- R and RStudio ["Rscript" should be on PATH]. This help page provides guidance.
- MATLAB 2019a or higher ["matlab" should be on PATH]. This help page provides guidance.
- Bash or another POSIX compliant Unix shell [Mac/Linux users can use the default shell; Windows users can use the Linux subsystem]. This help page provides guidance.

Note: It can be tricky to get this configured, so if you are having trouble go with Option 2. Windows users may need to install GIT to run `RunReplication.sh`.

3.1.2 Instructions

To run the replication from top to bottom, go to: AnonymizedETSReplication/ and execute the shell script `RunReplication.sh` by running `source RunReplication.sh` in the shell.

`RunReplication.sh` calls the following 7 scripts to replicate the project:

1. Emissions (Stata): `01Code/emissions/00MainStata.do`
2. Emissions (R): `01Code/emissions/00MainR.R`
3. Trading (Stata): `01Code/trading/00MainStata.do`
4. Phone survey (Stata): `01Code/phone_survey/00MainStata.do`
5. Model data prep (Stata): `01Code/model/cleaning/00MainStata.do`
6. Model (MATLAB): `/01Code/model/matlab/main_caller.m`
7. Trading (R): `01Code/trading/00MainR.R`

These are all all run by default, but if you want to exclude scripts from a specific language set the corresponding `Run_LANGUAGE` flag to 0 in `RunReplication.sh`. For example, if you wanted to skip all R scripts, you would replace `Run_R=1` with `Run_R=0`. If you want to run the entire project from scratch set `Clean_build=0`. Note that compilation will fail if you run the project from scratch, but try to skip scripts written in a particular coding language.

In order for these scripts to execute properly, it is necessary to have the following R and Stata packages installed with the package versions matching those listed in table 1.

- **Stata**: 'ftools', 'regdhfe', 'xtoverid', 'estout', 'mvsumm', 'grstyle', 'listtex', and 'ivreg2'
- **R**: 'tidyverse', 'haven', 'readxl', 'janitor', 'kableExtra', 'viridis', 'ggrepel', 'ggpattern', 'modelsummary', 'fixest', 'modelr', 'cowplot', 'gridExtra', and 'grid'

These are all all installed by default in `RunReplication.sh`, but if you don't want these packages from a specific language to be installed, set the `Install_LANGUAGE_packages` flag to 0 in `RunReplication.sh`. For example, if you wanted to prevent R packages from being installed, you would replace `Install_R_packages=1` with `Install_R_packages=0`.

3.2 Option 2: Run 9 scripts

`RunReplication.sh` allows users to run the full replication package with one click, but if you are having difficulty configuring your software environment to run it, you can also replicate the project by hand by doing the following:

1. Run Stata package installer: `01Code/installers/install_Stata_packages.do`
2. Run R package installer: `01Code/installers/install_R_packages.R`
3. Run each project script:
   (a) Emissions (Stata): `01Code/emissions/00MainStata.do`
   (b) Emissions (R): `01Code/emissions/00MainR.R`
   (c) Trading (Stata): `01Code/trading/00MainStata.do`
   (d) Phone survey (Stata): `01Code/phone_survey/00MainStata.do`
   (e) Model data prep (Stata): `01Code/model/cleaning/00MainStata.do`
   (f) Model (MATLAB): `/01Code/model/matlab/main_caller.m`
   (g) Trading (R): `01Code/trading/00MainR.R`

Note: Scripts must be run in order. When running each script, the working directory must be set to AnonymizedETSReplication/.

4 Data Availability and Provenance

The data were generated through surveys, plant trading data, and plant emissions measured between 2018 and 2021, as part of a field experiment. For more information on the different data sources and how they were obtained, see "Appendix B: Data" and "Appendix C: Pollution Monitoring" in the paper.

4.1 Data sets

Raw data are stored in 00RawData. They are included as base data sets and are not generated by running the replication code. These data sets were generated through surveys, plant trades, and emissions recorded from CEMS devices. They were merged and compiled to produce the data sets listed below. Data sets are listed based on the directory within 00RawData.

4.2 baseline/

- BaselineCovariates_318i.dta: Baseline plant covariates including boiler characteristics, abatement equipment, and pre-experiment revenue and costs. These data were collected using the baseline survey in December 2018.

- Master ETS In-sample - Plants (318i).dta: Treatment assignment and indicators for whether the plant was incompletely treated or removed from the sample.

4.3 cems/

- CEMS Panel Monthly (clean, PO).dta: Plants' CEMS purchase dates and monthly PM emissions. This data set is only used as a robustness check to test the effect of CEMS installation on plant emissions (Table C3: Effects of CEMS Installation on Plant Emissions).

4.4 emissions/

- Historical Factors Preliminary Cleaned (Raw).dta: Raw CEMS calibration data. This data set is used primarily to assess if CEMS readings are consistent with manual sampling (Figure F4: Simultaneous CEMS and sampling comparison).
- CEMS_2018-12-31 to 2021-04-03_StackDay_Balanced_HistCFs_Master337s_3sample2trunc.dta: Daily CEMS data at the stack level. This is the primary data input used to construct plant emissions.
- master-data_flow-data_stack-level_surat-sample_20-04-2022_v3.xlsx: Stack-level flow data from CEMS devices. This is used to compute potential maximum pollution measures for plants.

4.5 phone survey/

- Phone Survey All Covariates (Plant).dta: Endline plant information such as boiler characteristics, abatement equipment, and revenue and costs. These data were collected using the endline survey in 2020-2021.

4.6 trading/

- NeMLinventory_Sept2019-Mar2021.csv: Daily plant permit inventories from NeML.
- NeMLordertrade_Sept2019-Mar2021.csv*: Plant bid data from NeML.

5 Hardware & Software Specifications

5.1 Hardware requirements

The code does not require any special hardware. A laptop is sufficient to replicate all tables and figures.

5.2 Software requirements

- Stata 17 or higher.
- R and RStudio.
- MATLAB 2019a or higher.

Table 1: Versions Used

Language	Package	Version
R		4.3.1
R	‘tidyverse‘	2.0.0
R	‘purrr‘	1.0.2
R	‘haven‘	2.5.4
R	‘readxl‘	1.4.3
R	‘janitor‘	2.2.0
R	‘kableExtra‘	1.4.0
R	‘viridis‘	0.6.5
R	‘ggrepel‘	0.9.5
R	‘ggpattern‘	1.1.1
R	‘modelsummary‘	2.1.1
R	‘fixest‘	0.12.0
R	‘modelr‘	0.1.11
R	‘cowplot‘	1.1.1
R	‘gridExtra‘	2.3
R	‘grid‘	4.3.1
Stata		18
Stata	‘ftools‘	2.49.1
Stata	‘reghdfe‘	6.12.3
Stata	‘xtoverid‘	2.1.8
Stata	‘estout‘	3.31
Stata	‘mvsumm‘	1.1.1
Stata	‘grstyle‘	1.1.1
Stata	‘listtex‘	9/25/09
Stata	‘ivreg2‘	4.1.11
MATLAB		R2023a

5.3 Controlled randomness

MATLAB RandStreams are created on:

- Line 565 of `01Code/model/matlab/@etsCounter/etsCounter.m`
- Line 86 of `01Code/model/matlab/@etsModel/etsModel.m`
- Line 131 of `01Code/model/matlab/@etsModel/etsModel.m`

5.4 Memory, Runtime, and Storage Requirements

- Approximate time needed to reproduce the analyses on a standard 2023 laptop machine: 15 minutes.
- Approximate storage space needed: 200 MB.
- The code was last run on a 10-core Apple Silicon-based laptop with MacOS version 14.5 with 500Gb of storage and 32 GB of memory. Computation took 10 minutes.

6 List of Exhibits

Table 2: Tables and Figures

Figure/Table #	Program	Called from	Note
Figure 1	NA	NA	Requires confidential geographic data that would compromise plants anonymity
Figure 2	trading/plot_time_series.do	trading/00MainStata.do	
Figure 3	trading/plot_histogram_permit_consumption.do	trading/00MainStata.do	
Figure 4	trading/plot_histogram_permit_consumption.do	trading/00MainStata.do	
Figure 5	emissions/plot_emissions_time_series.do	emissions/00MainStata.do	
Figure 6	trading/plotting_all_bids.R	trading/00MainR.R	
Figure 7	trading/trading_plant_panels.r	trading/00MainR.R	
Figure 8	model/code/exhibits/plotFitTimeSeriesPrice.m	model/matlab/main_new.m	
Figure 9	trading/trading_plant_panels.r	trading/00MainR.R	
Figure 10	model/code/exhibits/plotFitTimeSeriesPrice.m	model/matlab/main_new.m	
Figure B1	emissions/plot_emissions_distribution.do	emissions/00MainStata.do	

Figure B2	trading/plot_histogram_number_bids.do	trading/00MainStata.do	
Figure C1	emissions/plot_intra_day_data_availability.R	emissions/00MainR.R	
Figure C2	emissions/plot_emissions_time_series.do	emissions/00MainStata.do	
Figure C3	emissions/plot_emissions_time_series.do	emissions/00MainStata.do	
Figure C4	emissions/compare_imputation_rules.do	emissions/00MainStata.do	
Figure E1	NA	NA	Tikz picture
Figure E2	model/code/exhibits/plotFitTimeSeriesPrice.m	model/matlab/main_new.m	
Figure E3	model/matlab/exhibits/plotFitHistEmissions.m	model/matlab/main_new.m	
Figure F1	trading/make_elasticity_coefplot.do	trading/00MainStata.do	
Figure F2	trading/covid_net_demand_graphs.R	trading/00MainR.R	
Figure F3	trading/plot_histogram_permit_consumption.do	trading/00MainStata.do	
Figure F4	emissions/plot_cems_manualsampling.R	trading/00MainR.R	
Table 1	emissions/make_balance_table_new.do	emissions/00MainStata.do	
Table 2	trading/compute_summary_statistics.do	trading/00MainStata.do	
Table 3	emissions/make_regression_table.do	emissions/00MainStata.do	
Table 4	trading/make_elasticity_table_hetero.do	trading/00MainStata.do	
Table 5	phone_survey/create_regression_tables_KW.do	phone_survey/00MainStata.do	
Table 6	model/matlab/cost_benefit_calculation/benefit_analysis_calculation.m	model/matlab/main_new.m	

Table A1	NA	NA	Administrative details
Table A2	NA	NA	Administrative details
Table A3	NA	NA	Administrative details
Table A4	emissions/make_balance_table_new.do	emissions/00MainStata.do	
Table B1	trading/compute_summary_statistics.do	trading/00MainStata.do	
Table C1	emissions/compare_imputation_rules.do	emissions/00MainStata.do	
Table C2	NA	NA	Administrative details
Table C3	emissions/make_effect_of_cems_table.do	emissions/00MainStata.do	
Table C4	emissions/make_missing_data_analyses.R	emissions/00MainR.R	
Table D1	model/matlab/exhibits/tabulateCounterResults.m	model/matlab/main_new.m	
Table E1	model/matlab/exhibits/tabulateCounterResults.m	model/matlab/main_new.m	
Table F1	phone_survey/create_cost_tables.do	phone_survey/00MainStata.do	
Table F2	phone_survey/create_regression_tables_KW.do	phone_survey/00MainStata.do	
Table F3	emissions/make_balance_table_reporting_split.do	emissions/00MainStata.do	
Table F4	emissions/make_missingdata_abatecost_reg.R	emissions/00MainR.R	
Table F5	emissions/make_regression_table.do	emissions/00MainStata.do	
Table F6	emissions/make_regression_table_covid_inter.do	emissions/00MainStata.do	
Table F7	model/matlab/exhibits/tabulateCounterResults.m	model/matlab/main_new.m	