README
Replication Package
Environmental Regulations and Air Pollution in India: A Reexamination
Author: Olexiy Kyrychenko
26.02.2025

1. OVERVIEW

This replication package contains the data and code necessary to reproduce the results in *Environmental Regulations and Air Pollution in India: A Reexamination*. It includes all analyses from the paper, along with the tables and figures from the Online Appendix.

2. DATA SOURCES AND DESCRIPTION

Statement about Rights

The author of the manuscript certifies that they have legitimate access to and permission to use the data utilized in this study.

Summary of Availability

The main data necessary to replicate the analysis are publicly available and included in the replication package. Some raw data files, such as the village- and town-level shapefiles (see below), are not part of the replication archive. However, this does not affect replicability, as the resulting city-level shapefile used in the computations is included. The sources and access details are provided below. A copy of the data is included in the replication package in its original format and, where possible, converted to .csv. All data are stored in the folder *Data*.

2.1 Greenstone and Hanna (2014) Dataset

Greenstone and Hanna (2014) dataset is publicly available as part of their replication package. It contains air pollution and regulatory data used in their original analysis.

- *Source:* American Economic Review
- *Access:* Download the replication package from OpenICPSR Project 112693
- *File Path:* 112693-V1/Greenstone-and-Hanna--2014--Replication-Files/Data/Air-Data/Final-Data/combined.dta
- *Data format:* Stata's .dta and .csv
- *Coverage:* 1987–2007
- *Observations:* 2940
- *Replication Package Inclusion:* Yes, under *Data/Intermediate_Data/GH_Data*

2.2 NASA's MERRA-2 Air Pollution Data

The MERRA-2 dataset provides global gridded air pollution data, including monthly estimates of $PM_{2.5}$ components and SO_2 concentrations at a spatial resolution of 0.5° x 0.625°. $PM_{2.5}$ concentrations were calculated using $PM_{2.5}$ components following Buchard et al. (2016).

- *Source:* NASA GES DISC
- *Product name:* M2TMNXAER: Monthly mean, time-averaged, single-level aerosol diagnostics (V5.12.4)
- *Data format:* NetCDF (.nc)
- *Variables used:*
  BCSMASS – black carbon
  DUSMASS25 – dust
  OCSMASS – organic carbon
  SO2SMASS – SO_2 (sulfur dioxide)
  SO4SMASS – SO_4 (sulfate)
  SSSMASS25 – sea salt
- *Coverage:* 1987–2007
- *Download Date*: December 2020
- *Replication Package Inclusion:* Yes, under *Data/Raw_Data/Air_Pollution/MERRA-2_air_pollution_data_vars/Raw_data*

2.3 EDGAR Air Pollution Emissions Data

The EDGAR database provides estimates of global anthropogenic emissions of $PM_{2.5}$ and SO_2 on a 0.1° x 0.1° grid and was used for the Appendix Figure 1.

- *Source:* EDGAR
- *Data format:* NetCDF (.nc)
- *Variable used:*
  emissions – level of $PM_{2.5}$ and SO_2 emissions in tonnes
- *Coverage:* 1975–2022
- *Download Date*: May–August 2024
- *Replication Package Inclusion:* Yes, under *Data/Raw_Data/Emissions*

2.4 Shapefiles for India and City-Level Boundaries

Two types of shapefiles were used:

- *India-Level Shapefile*
  - *Source:* GADM
  - *Data format:* ESRI's shapefile (.shp)
  - *Use:* Provides country-level administrative boundaries of India
  - *Replication Package Inclusion:* Yes, under Data/Raw_Data/Shapefiles
- *City-Level Shapefiles*
  - *Source:* Manually created from ML InfoMap's 2011 digital maps

  - *Data format:* ESRI's shapefile (.shp)
  - *Use:* Used to map MERRA-2 air pollution at the city level
  - *Raw data:* Village- and town-level ML InfoMap shapefiles were accessed from the Princeton University Digital Maps & Geospatial Data Library during a research visit in 2018
  - *Replication Package Inclusion:* The constructed city-level shapefile is included under *Data/Raw_Data/Shapefiles*, but the original ML InfoMap shapefiles are not

2.5 Final Dataset

The final dataset was constructed by merging Greenstone and Hanna (2014) dataset with alternative air pollution measures derived from the NASA MERRA-2 reanalysis data product.

- *Data format:* Stata's .dta and .csv
- *Variables used:*
  - *Identifiers:* state, district, city, year – geographic and time identifiers.
  - *Regulatory ground-based air pollution measures:*
    - e_spm_mean – mean suspended particulate matter (SPM) concentration from regulatory ground-based monitoring data, originally used in GH.
    - e_GH_pm_mean – mean $PM_{2.5}$ concentration converted from GH's SPM using SPM-PM_{10}-$PM_{2.5}$ ratios.
    - e_so2_mean – mean sulfur dioxide (SO_2) concentration from regulatory ground-based monitoring data, originally used in GH.
  - *MERRA-2 air pollution measures:*
    - e_pm_asGH_mean – mean $PM_{2.5}$ concentration adjusted to match GH's sample of cities.
    - e_pm_mean – mean $PM_{2.5}$ concentration with the maximum possible number of city-by-year observations.
    - e_so2_asGH_mean – mean SO_2 concentration adjusted to match GH's sample of cities.
    - e_so2n_mean – mean SO_2 concentration with the maximum possible number of city-by-year observations.
  - *Urban characteristics:*
    - lit_urban – urban literacy rates.
    - pop_urban – urban population.
  - *MERRA-2 $PM_{2.5}$ components:*
    - bc – black carbon concentration.
    - du – dust concentration.
    - oc – organic carbon concentration.
    - so4 – sulfate concentration.
    - ss – sea salt concentration.
- *Coverage:* 1987–2007
- *Observations:* 2940
- *Replication Package Inclusion:* Yes, under *Data/Final_Data*.

The original GH dataset includes many additional variables not used in the analysis. For further details, refer to Greenstone and Hanna (2014) and their replication archive.

3. CODE DESCRIPTION

The code necessary to replicate the analysis is organized into three main categories: ArcPy codes for ESRI ArcGIS, .r scripts for R, and .do files for Stata. All code files are organized within the *Codes* folder, with ArcPy scripts in *ArcPY_Codes*, R scripts in *R_Codes*, and Stata do-files in *Do_Files*.

3.1 ArcPy Codes

The ArcPy codes are used to compute alternative GIS-based measures for $PM_{2.5}$ and SO_2 air pollution concentrations at the city level. These codes utilize the MERRA-2 gridded datasets and city-level shapefile as primary sources. The full data processing procedure is documented in Online Appendix 7 of the manuscript.

1-BC_netCDF_to_raster_&_clip.py
1-DU_netCDF_to_raster_&_clip.py
1-OC_netCDF_to_raster_&_clip.py
1-SO2_netCDF_to_raster_&_clip.py
1-SO4_netCDF_to_raster_&_clip.py
1-SS_netCDF_to_raster_&_clip.py

- *Purpose:* Converts monthly NetCDF files to raster format and clips them to the extent of India for various pollutants.
- *Pollutants:* BC, DU, OC, SO4, SS, SO2
- *Years:* 1987-2007
- *Output:* Clipped monthly raster files for each pollutant.

2-annual_means.py

- *Purpose:* Calculates annual mean raster values for each pollutant.
- *Pollutants:* BC, DU, OC, SO4, SS, SO2
- *Years:* 1987-2007
- *Output:* Annual mean raster files for each pollutant.

3-multiply_PM_elements.py

- *Purpose:* Multiplies the annual mean raster values for OC and SO_4 by their respective coefficients as part of the $PM_{2.5}$ calculation.
- *Formula:* PM2.5 = Dust2.5 + SS2.5 + BC + 1.4**OC* + *1.375**SO4
- *Years:* 1987-2007
- *Output:* Multiplied raster files for OC and SO_4.

4-summation_PM_elements.py

- *Purpose:* Sums the raster values of the $PM_{2.5}$ components to compute the annual $PM_{2.5}$ concentration.

- *Formula:* PM2.5 = Dust2.5 + SS2.5 + BC + 1.4**OC + 1.375**SO4
- *Years:* 1987-2007
- *Output:* Annual $PM_{2.5}$ raster files.

5-BC_projection_to_CSV.py
5-DU_projection_to_CSV.py
5-OC_projection_to_CSV.py
5-PM_projection_to_CSV.py
5-SO2_projection_to_CSV.py
5-SO4_projection_to_CSV.py
5-SS_projection_to_CSV.py

- *Purpose:* These scripts represent the final step in city-level data creation. They project pollutant-specific raster files into an appropriate projected coordinate system, resample them, and extract zonal statistics (annual average air pollutant concentrations within city boundaries), saving the results in .csv format.
- *Pollutants:* BC, DU, OC, PM, SO2, SO4, SS
- *Years:* 1987-2007
- *Output:* .csv files containing zonal statistics for each pollutant. The files are stored in *Data/Raw_Data/Air_Pollution/MERRA-2_air_pollution_data_vars/5-table_to_csv*

6-Figure1_PM.py
7-Figure1_SO2.py

- *Purpose:* These scripts generate raster files representing the average concentrations of $PM_{2.5}$ and SO_2 across all years at the cell level. The process involves calculating the mean raster values, projecting them into the appropriate coordinate system, resampling, and clipping them. The resulting raster files are used to create Figure 1, which visualizes the average concentrations of $PM_{2.5}$ and SO_2.
- *Years:* 1987-2007
- *Output:* Clipped raster files for Figure 1.

The runtime for the ArcPy codes is approximately *100* minutes.

3.2 R Codes

EDGAR_Emissions_PM2.5.R
EDGAR_Emissions_SO2.R

- *Purpose:* These scripts process EDGAR emissions data to generate city-level annual emissions for $PM_{2.5}$ and SO_2. The steps involved in processing the emissions data are similar to those used in creating MERRA-2 $PM_{2.5}$ and SO_2 concentration levels. The resulting data is used to create Appendix Figure 1, which visualizes the annual emissions of $PM_{2.5}$ and SO_2.
- *Years:* 1975-2022
- *Output:* .csv files containing city-level annual emissions data for $PM_{2.5}$ and SO2, used for Appendix Figure 1. The files are stored in *Data/Intermediate_Data/Emissions_Data*

The runtime for the R codes is approximately *14* minutes.

3.3 Stata Codes

The main analysis is conducted using Stata codes. The .do files fall into two categories: data preparation and data analysis (tables and figures). The *Preparing_data* folder contains codes for preparing the final dataset and is located in *Do_Files/Preparing_data*. The *Tables_Figures* folder contains codes for all figures and tables in the manuscript and Online Appendix.

3.3.1 Codes to prepare data

Convert_GIS_to_Stata_AP_data.do

- *Purpose:* Prepare GIS-based MERRA-2 air pollution data for merging with the original Greenstone and Hanna (2014) dataset.
- *Output:* Processed GIS-based air pollution data ready for merging

Merge_GH_and_GIS_AP_data.do

- *Purpose:* Compile the final dataset by merging the processed GIS-based air pollution data with the original Greenstone and Hanna (2014) dataset.
- *Output:* Final merged dataset for analysis stored in *Data/Final_Data*.

3.3.2 Codes for data analysis

AppFigure1.do

- *Purpose:* Generate Appendix Figure 1, which plots the trends in $PM_{2.5}$ and SO_2 emissions from 1975 to 2022. This figure is constructed using data from the Emissions Database for Global Atmospheric Research (EDGAR).
- *Years:* 1975-2022
- *Output:* Appendix Figure 1 visualizing emission trends.

AppFigure2.do

- *Purpose:* Generate Appendix Figure 2, which visualizes the trends in $PM_{2.5}$ components over the sample period. This figure provides insights into the composition of particulate matter over time.
- *Years:* 1987-2007
- *Output:* Appendix Figure 2 showing $PM_{2.5}$ component trends.

AppTable1.do

- *Purpose:* Generate Appendix Table 1, which provides summary statistics for air pollution. This table offers a statistical overview of the air pollution data used in the analysis.
- *Output:* Appendix Table 1 with summary statistics.

AppTable2.do

- *Purpose:* Generate Appendix Table 2, which reports the estimated coefficients for SCAP and CAT policies on particulate air pollution. This table compares Greenstone and Hanna's (GH) original SPM with $PM_{2.5}$ converted from GH's SPM, based on GH's original code for Table 3.
- *Output:* Appendix Table 2 with policy coefficients.

Figure2.do

- *Purpose:* Generate Figure 2, which compares air pollution trends from Greenstone and Hanna (GH) using ground-based CPCB data with alternative outcomes based on MERRA-2 reanalysis data. This figure highlights the differences between observed and reanalysis data.
- *Years:* 1987-2007
- *Output:* Figure 2 comparing air pollution trends.

Table1.do

- *Purpose:* Generate Table 1, which reports the main results of the paper. This table presents the estimated coefficients for SCAP and CAT policies on air pollution, comparing outcomes from ground-based CPCB data with those from MERRA-2 reanalysis data. The code builds on GH's original code for Table 3.
- *Output:* Table 1 with main results.

The runtime for the .do files is approximately *1* minute.

Note that Figure 1 was created manually using ESRI ArcGIS from the raster datasets generated by the ArcPy codes described above.

4. COMPUTATIONAL AND SOFTWARE REQUIREMENTS

The computational environment includes a combination of Windows and macOS systems, utilizing virtualization for compatibility with ArcGIS.

1. ArcPy Codes:

- *Software:* ArcGIS for Desktop version 10.3.1
- *Operating System:* Windows 10 Home
- *Hardware:* Intel(R) Core(TM) i7-4870HQ CPU @ 2.50GHz, 4 GB memory, 64-bit Operating System, x64-based processor
- *Virtualization:* Parallels Desktop 13 for Mac Home Edition version 13.3.2 (43368)
- *Host Machine:* Apple MacBook Pro (Retina, 15-inch, Mid 2014) with 2.5 GHz Intel Core i7, 16 GB memory, and macOS Mojave version 10.14.6

2. R Codes:

- *Software:* R version 4.3.0 GUI 1.79 Big Sur ARM build (8225), RStudio version 2023.06.2+561
- *Operating System:* macOS Sonoma version 14.6.1
- *Hardware:* Apple MacBook Pro (16-inch, 2021) with Apple M1 Pro chip, 16 GB memory

3. Stata Codes:

- *Software:* Stata/SE 18.5 for Mac (Apple Silicon)

- *Operating System:* macOS Sonoma version 14.6.1
- *Hardware:* Apple MacBook Pro (16-inch, 2021) with Apple M1 Pro chip, 16 GB memory

The approximate total time needed to reproduce the analyses is *2* hours.

5. REPLICATION INSTRUCTIONS

1. Extract the Replication Package:

- Decompress (unzip) the file JAE_MS_14996_Replication_package.zip.
- Save the unzipped folder to a convenient location.
- Download PM2.5.zip and SO2.zip, and save them to *Data/Raw_Data/Emissions*.
- Extract both zip files within this directory. This will create two folders, PM2.5 and SO2, containing EDGAR air pollution emissions data.
- The folder structure is essential for replication. Do not rename any subfolders or modify the file organization.

2. Construction of MERRA-2 Air Pollution Data

- Requirements**:**
  - Ensure access to ArcGIS Desktop on a Windows PC.
  - Before running the codes, update all file paths in the ArcPy codes.

- Updating Paths in ArcPy codes**:**
  - Locate the lines with paths in the codes.
  - Replace the original path (e.g.,):

inNetCDF1 = "C:\\GIS\\2-India_GIS\\Paper_2\\MERRA-2_air_pollution_data_vars\\Raw_data\\MERRA2_100.tavgM_2d_aer_Nx."

with your own path, ensuring it follows this structure:

"...\\JAE_MS_14996_Replication_package\\Data\\Raw_Data\\Air_Pollution

followed by:

\\MERRA-2_air_pollution_data_vars\\Raw_data\\MERRA2_100.tavgM_2d_aer_Nx."

- Running ArcPy Codes**:**
  - Execute all codes sequentially from the folder *Codes/ArcPY_Codes.*
  - The output .csv files containing MERRA-2 air pollution data will be saved in:

\\MERRA-2_air_pollution_data_vars\\5-table_to_csv\\

- Alternative to Running ArcPy Codes**:**

  - If ArcGIS Desktop is unavailable, pre-processed results are provided.
  - The folder:

...\\JAE_MS_14996_Replication_package\\Data\\Raw_Data\\Air_Pollution\\MERRA-2_air_pollution_data_vars\\

contains the most recent outputs of all ArcPy codes. This allows proceeding directly to the next steps of the replication without running ArcPy.

3. Construction of EDGAR Air Pollution Emissions Data

- Requirements
  - Ensure R is installed on your computer.
  - Before running the scripts, update the working directory in both .r script files (line 5).

- Running R Scripts
  - Navigate to the folder *Codes/R_Codes.*
  - Execute all .r scripts sequentially.
  - The output .csv files containing EDGAR air pollution emissions data will be saved in *Data/Intermediate_Data/Emissions_Data*

4. Creation of the Final Dataset and Main Analysis

- Replicators have two options:
  - Run the master file to execute the entire process automatically.
  - Run each .do file separately for more control over individual steps.

- Setting the Working Directory
  - In both cases, update the path in the following line(s):

```
global mypath ".../JAE_MS_14996_Replication_package/"
```

replace "..." with the actual path to your root directory. This global macro sets the working directory for all .do files.

- Option 1: Running the Master File
  - Update global mypath in line 31 of JAE_MS_14996_master.do.
  - Run JAE_MS_14996_master.do, which will automatically execute all data preparation and analysis .do files in the correct sequence.

- Option 2: Running .do Files Independently
  - Update mypath within the specific .do file you intend to run.
  - First, execute the codes in *Codes/Do_Files/Preparing_data*
  - Then, run the codes in *Codes/Do_Files/Tables_Figures*

- Output Locations
  - Results will be saved in *JAE_MS_14996_Replication_package/Output*
  - Log files will be stored in *JAE_MS_14996_Replication_package/Log*

6. REFERENCES

Greenstone, M., & Hanna, R. (2014). Environmental regulations, air and water pollution, and infant mortality in India. *American Economic Review*, *104*(10), 3038-3072.

Buchard, V., Da Silva, A. M., Randles, C. A., Colarco, P., Ferrare, R., Hair, J., ... & Winker, D. (2016). Evaluation of the surface PM2. 5 in Version 1 of the NASA MERRA Aerosol Reanalysis over the United States. *Atmospheric Environment*, *125*, 100-111.