Fine-particulate air pollution and behaviorally inclusive mortality impacts of China's winter heating policy, 2013-2018

Alberto Salvo, Qu Tang*, Jing Yang, Peng Yin, Maigeng Zhou

***Journal of Environmental Economics and Management*, forthcoming**

ReadMe file on data and code, prepared in June 2024

***Corresponding author**

The following files containing source (raw) data are described along with computer code below:

(i) monitors_1628.dta
(ii) monitor_dist.dta
(iii) town_dist.dta
(iv) county_dist.dta
(v) monitor_daily_2013_2014.dta
(vi) monitor_daily_2014_2018.dta
(vii) town_code_coordinates.dta
(viii) county_code_coordinates.dta
(ix) gsod_china_2013_2018.dta
(x) population_town_2010.dta
(xi) population_county_2013.dta to population_county_2018.dta
Note: As noted in the article's data availability statement, county-level population data are proprietary and can be accessed from the Chinese Center for Disease Control and Prevention (CDC).
(xii) QinHuaiBoundary_coordinates_every500m.dta
(xiii) AlternativeBoundary_coordinates_every500m.dta
(xiv) (Omitted) Files with geocoded deaths are proprietary and can be accessed from the CDC.

The user needs to manually change the working directory to the folder in which they will save the source data files.

There are three **script** files (Stata do-files). Do-files can be read as text files in any text editor; these files run in Stata version 16 (or a later version).

****

Script file **describe_PM_W_population.do** consists of several programs (functions). The do-file creates a temporary folder in the D drive and saves temporary files to this folder. In order to assist the user, the do-file is commented throughout.

Program *location_heat_dist_urban* uses source data files (i) to (iv) and prepares three datasets containing distances of air monitors, counties, and towns to the Qin-Huai and alternative

boundaries. The three generated datasets, monitor_heat_dist_urban.dta, county_code_heat_dist_urban.dta, and town_code_heat_dist_urban.dta, are saved to the temporary folder.

Program *pm_monitor_year_month* combines source data files (v) and (vi) and saves eight generated datasets pm_monitor_year_month.dta, pm_monitor_coordinates.dta, and active_monitors_[year].dta to the temporary folder. (Please refer to self-explanatory file title names and comments in the code for further details; for brevity, we omit these here.)

Program *pm_monitor_describe* uses pm_monitor_year_month to produce Fig. A2. (Programs *pm_town_describe* and *W_town_describe* similarly describe PM concentrations and weather at the township level.)

Program *pm_4within50kmTown_year_month* uses source data file (vii) and combines this with previously generated datasets to generate pm_4within50kmTown_year_month.dta, storing this file in the temporary folder.

Program *pm_4within50kmCounty_year_month* uses source data file (viii) and combines this with previously generated datasets to generate pm_4within50kmCounty_year_month.dta, storing this file in the temporary folder.

Program *W_station_year_month* uses source data file (ix) and saves eight generated datasets W_station_year_month.dta, W_station_coordinates.dta, and active_stations_[year].dta to the temporary folder.

Program *W_4within200kmTown_year_month* uses source data file (vii) and combines this with previously generated datasets to generate W_4within200kmTown_year_month.dta, storing this file in the temporary folder.

Program *W_4within200kmCounty_year_month* uses source data file (viii) and combines this with previously generated datasets to generate W_4within200kmCounty_year_month.dta, storing this file in the temporary folder.

Program *W_4within200kmTown_year_month* uses previously generated datasets to generate W_4within200kmMonitor_year_month.dta, storing this file in the temporary folder.

Program *pm_W_dist_combine* combines previously generated datasets to generate pm_W_dist_town_year_month.dta, pm_W_dist_county_year_month.dta, and pm_W_dist_monitor_year_month.dta, storing these files in the temporary folder.

Program *population_year* uses source data files (x) and (xi) and saves two generated datasets population_town_2010.dta and population_county_year.dta in the temporary folder. **Note**: Source data files (xi) can be accessed through the CDC.

Finally, there are three remaining programs in do-file **describe_PM_W_population.do**:
wind_station_date_hour,
monitor_boundary_bearing, and
wind_1within200kmMonitor_year.
These three programs prepare data used in the robustness tests reported in Table A15, "Local linear regression estimates of the pollution discontinuity: Robustness to accounting for wind blowing from across the policy boundary." Also see Fig. A10, "Constructing a measure of spatial atmospheric policy mixing for each location, used in the robustness test of Table A15." In addition to previously generated datasets and source data files (xii) and (xiii), these programs require 6.7 GB of data at the station-date-hour level to run, which we downloaded from NOAA and are not publishing. Please contact the corresponding author for these 6.7 GB of big data. As an alternative, we are publishing the datasets generated by these three programs, for the purpose of replicating Table A15:

(xv) windBlowsToB_monitor_year_month.dta
(xvi) active_windStations_2013.dta to active_windStations_2018.dta

****

A second script file, **pollution_analysis.do**, implements the pollution regression discontinuity analysis. The first lines of code (lines 36-81) prepare the estimation samples from the previously generated datasets, which are stored in the temporary folder. The user needs to select the local variables, e.g., analysisUnit "county", yearStart 2013, and yearEnd 2015 to generate the 2013-2015 pollution estimation sample at the county level, named county_2013_2015.dta. The estimation samples are then stored in the user's working directory. For convenience, as an alternative, we are publishing the estimation sample files (again, this is for convenience as these files, containing cross-sections of pollution concentrations, should be replicable by the user):

(xvii) (estimation) county_2013_2015.dta (**note**: we excluded county-level population)
(xviii) (estimation) county_2016_2018.dta (**note**: we excluded county-level population)
(xix) (estimation) town_2013_2015.dta
(xx) (estimation) town_2016_2018.dta
(xxi) (estimation) monitor_2013_2015.dta
(xxii) (estimation) monitor_2016_2018.dta

This script file replicates Tables A2 to A23 and the associated manuscript tables and descriptive figures and tables, specifically, manuscript Tables 2 and 3.

****

A third script file, **health_analysis.do**, largely follows the structure of pollution_analysis.do. It uses proprietary mortality data, which can be accessed through the CDC, merged with pollution, weather, and population data as prepared above. The script file replicates Tables 4, A24-A34, A35-A38, and A39-A46 as well as associated descriptive figures and tables.