Read Me

This document facilitates replication of data formation and analysis in *Caution Drivers! Children Present: Traffic, Pollution, and Infant Health*. We describe below how to construct our final dataset and where to find relevant raw data.

1 Replication Process

All do-files assume the user has a folder called "KMS_REPLICATION" in their home folder, denoted by the use of the tilde (~) in all local file references within do-files. All do-files should be placed in said "KMS_REPLICATION" folder. A user interested in replicating results from raw data should place all files in the relevant folders and run the do-file "master_dofile.do", which calls a number of additional do-files all available on the Dataverse website.

1.1 Necessary Folders

The do-file "building_folders.do", available on the RESTAT Dataverse, generates all necessary folders for the do-files to run. Users should run this first, then place all raw data in the folders noted in each data description.

Do-files store all data in the folder "data", which contains several subfolders. Regression output, including summary tables, save to the folder "regs". Graphs save to the folder "graphs". Log files save to the folder "log_files".

2 Raw data

2.1 Traffic Data

Traffic data are available on the Caltrans Performance Measurement System (PeMS) website,

`http://pems.dot.ca.gov`

as of June 1, 2015. We accessed data via the PeMS Data Clearinghouse, and use "Station Hour" datasets for each relevant month and region (see main text). For use in the replication do-files, we assume all files follow the current PeMS naming convention:

d*region*_text_station_hour_*year*_*month*.txt.gz.

For example, data from region 3 in January of 2002 would have the file name:

d03_text_station_hour_2002_01.tx.gz,

where the suffix "gz" indicates files compressed using the unix "gzip" command.

Save raw traffic data files here:

"~/KMS_REPLICATION/data/traffic_data/text_data/d*region*_data".

2.2 Weather Data

Weather data are available on the National Oceanic and Atmospheric Administration (NOAA) Global Surface Summary of the Day website,

`https://data.noaa.gov/dataset/global-surface-summary-of-the-day-gsod`

as of June 1, 2015. We accessed data via the NOAA ftp server,

`ftp://ftp.ncdc.noaa.gov/pub/data/gsod/`.

For use in the replication do-files, we assume all files follow the current NOAA naming convention:

usaf-*ban*-*year*.op.gz,

where *usaf* and *wban* are NOAA station identifier numbers.

Save raw weather data files here:

"~/KMS_REPLICATION/data/weather_data".

2.3 Natality/Mortality Data

Natality and infant mortality data are from the Vital Statistics natality and mortality data files — use is restricted due to geographic identifiers including mother zip code. We obtained restricted use data through a request with the California Department of Public Health (CDPH).

Save raw natality and mortality data files here:

"~/KMS_REPLICATION/data/birth_data/raw_data/*year*"

where *year* indicates the relevant year of data. Dictionary files for importing the data (using the file naming convention provided by the CDPH) are available on the RESTAT Dataverse website, with the file extension ".dict". Place dictionary files here:

"~/KMS_REPLICATION/data/dicts".

2.4 Emissions Data

Emissions data are available from the California Air Resources Board (CARB) website,

`http://www.arb.ca.gov/aqd/aqdcd/aqdcddld.htm`

as of June 1, 2015. We use the daily gas comma-delimited file for carbon monoxide and PM mass comma-delimited (daily, hourly) values for PM10.

Save raw emissions data files here:

"~/KMS_REPLICATION/data/emissions_data".

2.5 Location Data

To conduct analysis at the zip code level, we calculate distance between weather, emissions, and traffic sensors and zip code centroids. We gathered data on the location of each of the relevant geographic locations, and make these data available on the RESTAT Dataverse.

Save provided location data files here:

"~/KMS_REPLICATION/data/location_data"

We obtain these data from the following sources.

- Weather location data from

  `ftp://ftp.ncdc.noaa.gov/pub/data/inventories/ISH-HISTORY.TXT`,

  saved as file "station_list.csv".

- Traffic monitor location data from PeMS website, saved as files "d*region*_stations_2007" for each relevant region.

- Emissions location data from file "location.dbf.exe" on the CARB website, saved as file "pollution_location.csv".

- Zip code centroid information from the Census ZCTA5 files, available at

  `https://www.census.gov/geo/maps-data/data/gazetteer2000.html`,

  saved as file "zip_dict.dct" in the "dicts" folder.

In addition, the file "ziptomsa.csv", available on the RESTAT dataverse, provides a crosswalk between zip codes and Metropolitan Statistical Areas for those interested in matching zip codes to larger regions. Place this file in the "location_data" folder path above. The do-file "building_hazard_model.do" calls this file for matching purposes.

2.6 Further Data for Download

Birth data are restricted use, so we cannot provide a copy of our full data set for download. Instead, we provide a limited data file, "weekly_pollution_weather_traffic_KMS.dta," which contains week-level data for all variables other than natality/mortality information (traffic, pollution, and weather). These are the data we merge with our birth data in "building_hazard_model.do."

3 Relevant Do-Files by Task

Folder generation:

building_folders.do

Distances:

distances_to_zips.do

Traffic data files:

importing_hourly_traffic.do

hourly_to_weekly_traffic.do

zip_level_traffic.do

Weather data files

importing_daily_weather.do

daily_to_weekly_weather.do

zip_level_weather.do

Pollution data files:

importing_daily_pollution.do

daily_to_weekly_pollution.do

zip_level_pollution.do

Birth data files:

importing_daily_birth_data.do

Combining all files and building analysis data files:

combine_pollution_weather_traffic.do

building_hazard_model.do

Data analysis:

table_birthsum.do

table_pollutionsum.do

reduced_form.do

regressions_collapsed_OLS.do

regressions_collapsed_IV.do

regressions_collapsed_IV_weather_order.do

Monte Carlo analysis:

fakepollution.do

KMS_mcloop.do

4 Software used

Stata/SE 13.1

Please note some do-files call the unix shell commands "gunzip" and "gzip", which uncompress and compress large data sets and require Stata to call programs outside of Stata itself.