Explanation of code / data

This folder includes several folders and subfolders that are explained in more detail below

1) codeMATLAB: our analysis was done using STATA. We used MATLAB to plot some figures. This folder includes the code, where file names correspond to the figure number.

2) codeSTATA: this folder contains the code for our analysis. It contains several subfolders that are numbered in the order in which the code was executed. For each subfolder there is a STATA file that calls all subroutines in a subfolder (e.g., 1_readinData.do is STATA code that calls subroutines in the folder 1_readinData). Please change the local `rootDir' to whatever directory you extracted the data into on your machine.

2a) 1_readinData contains subroutines to read in the data. To preserve space, the raw data is NOT included, but the data sources are given in the paper. The folder provides the code we used import, clean, and aggregate the data. The full panel of the compiled data is produced in part 2c below and included among the files.

2b) 2_matchDataSources does spatial matching between zip codes, airports, and pollution monitors. The raw data are omitted again as the matches are included in the full merged data (part 2c below).

2c) 3_constructPanel constructs the full panel used in the analysis based on the data from part 2a and the spatial matching from part 2b. The full panels are saved in the folder dataSTATA (part 3 below) as files 3_fulPanel_*.

2d) 4_regression links airport pollution in California to taxi time in California, which in turn is instrumented by taxi time at Eastern airports.

2e) 5_constructPanel constructs predicted taxi time at California that is caused by network delays (first stage of part 2d). The predicted taxi time is saved in the dataSTATA folder (part 3 below) and used as instrument in part 2f below.

2f) 6_regression has the regression code linking health outcomes in California to pollution, which is instrumented by predicted taxi time from part 2e.

2g) 7_zipCodeChecks does additional sensitivity checks by zip code. The baseline assigns people to the zip code where they reside. This part splits people into subgroups as we have both the hospital and the residence zip code, i.e., whether hospital and residence are in the same or different zip codes. See Table A10.

2h) 8_tables has the code to construct the tables. Tables (except the ones needing raw sickness data) can be recreated as the saved regression results are available in the folder resultSTATA.

2i) 9_figures has the code to construct figures that are drawn using STATA and not MATLAB (for those see part 1 above). All tables can be recreated as the saved regression results are available in the folder resultSTATA.

3) dataSTATA contains STATA data sets. The numbering corresponds to the code in part 2 above. The full panels that are constructed using the code in folder 3_constructPanel are saved as files 3_fulPanel_*. For privacy reasons that hospital sickness counts are not included but can be obtained here: https://cphs.keyusa.net
Our health data set is constructed from two data sources about hospital visits in the State of California for the years 2005-2007.

(1) Inpatient Hospital Discharge Data (patients that stay at least one night in the hospital)
(2) Emergency Room Data (patients that visit the emergency room and get discharged the same day. If they end up being admitted they only show up in the inpatient hospital discharge data).

We utilize the following data fields for our analysis:

(i) zip code of the patients residence
(ii) zip code of the hospital
(iii) admission date
(iv) age of patient (in years)
(v) primary as well as all secondary diagnosis codes

4) dataTEXT contains data in text format that can be read by MATLAB to construct the figures.

5) documents includes subfolders where the tables and figures are saved.

6) resultSTATA contains regression results that are used in the figures and tables. Additional tests can be run on these STATA results as we are not able to post the hospital discharge data.