1 Overview

This README file contains an overview on how to replicate the results in "JUE Insight: Air Quality and Student Performance in the U.S." (Gilraine and Zheng, 2024).

The "Replication Package" folder consists of the following four subfolders:

1. Raw Data: contains several subfolders for each raw data source.
2. Derived Data: contains derived data files that combine multiple sources of raw data.
3. Results: contains program files for working with the derived data files and creating the results.
4. Output: contains outputted tables and figures.

2 Instructions

These instructions explain how to go through the replication process without downloading the raw data files, which are very large data files and require a considerable amount of time to process.

1. In "Programs" open the folder "PrepareData".
2. Follow the instructions in "README.pdf".
3. Go back to the main directory and open the folder "Results".
4. Follow the instructions in "README.pdf".

For those who would like to start directly from the Raw Data: complete the following step below before running steps 1 through 4 above.

- Open the "Raw Data" folder. For each of the folders numbered 1 through 6, open it and follow the instructions in the README. Note these folders are zipped, so will need to be unzipped. Also note that due to size constraints at the Dataverse, the zip files for 3 data files ("Derived Data\Table 2 Data," "Raw Data\Pollution Data," and "Raw Data\Weather Data") had to be split into parts; parts are numbered and denoted with the "_split" suffix. Then open the "Raw Data\Prepare Data" folder and run the R script "1_masterprep.R" (the README in "Raw Data\Prepare Data" describes this script and the programs that it runs).

3 Data Availability

All data files are provided except for any data file that contains data from the Stanford Education Data Archive which we cannot post as their data use agreement states that the data are non-transferable. Their data, however, are publicly-available at https://edopportunity.org/get-the-data/ along with terms of agreement. Researchers can therefore download the Raw Data directly from SEDA and place these data in the "Raw Data\6. SEDA DATA\" directory and run the R script "Raw Data\Prepare Data\1_masterprep.R" to create each missing derived data dataset. Note that researchers do *not* need to download and process the very large "Raw Data\Pollution Data.zip" and "Raw Data\Weather Data.zip" zip archives to create the below derived data sets.

The following is a list of datafiles that are not provided:

1. Raw Data (direct from SEDA)
    - Raw Data/6. SEDA DATA/seda_geodist_long_cs_4.1.csv
    - Raw Data/6. SEDA DATA/seda_cov_geodist_long_4.1.csv
    - Raw Data/6. SEDA DATA/seda_shapefiles_2019_4.0
2. Derived Data (containing SEDA data). All these files are created by the R script: "Raw Data\Prepare Data\1_masterprep.R"
    - Derived Data/df_avg_noinst.csv
    - Derived Data/dfbartik2001_40km.csv
    - Derived Data/dfbartik2005_40km.csv
    - Derived Data/dfbartik2005_60km.csv
    - Derived Data/dfbartik2005_40kmanalysis_annual.csv
    - Derived Data/dfbartik2005_60kmanalysis_annual.csv
    - Derived Data/dfbartik2001_40kmanalysis_annual.csv
    - Derived Data/df_bwg.csv