JASA ACS Reproducibility Initiative - Author Contributions Checklist Form

The purpose of the Author Contributions Checklist (ACC) Form is to document the artifacts associated with a manuscript (e.g., code and data supporting the computational findings), and describe how to reproduce the findings. The final version of this document will be included as online supplemental material with the published paper and referenced in the abstract.

As of Sept. 1, 2016, the ACC Form must be included with all submissions to JASA ACS.

This document is the template that will be provided to authors; please replace the (non-bold) text below that provides guidance on how to fill out each item with the actual information for your manuscript.

Data

Abstract (Mandatory)
We have compiled a dataset consisting of health data from Medicare records, pollution exposure data from the US Environmental Protection Agency and other public and private sources, and confounder data from the US census. The Medicare data contain records for each beneficiary in the US in the years under study (2000 and 2001), which include the beneficiary's date of death and the cause and date of any hospitalizations as well as limited personal information including zipcode of residence. As described in the paper, we have factual and counterfactual air pollution data on grids of varying resolutions for the entire continental US. Confounder data are obtained from the census at the zipcode level. The Medicare and pollution data are aggregated to the zipcode level, to obtain counts of each health event of interest in each zipcode (and the size of the Medicare population in each zipcode to create health event rates as needed) and area-weighted averages of the gridded pollutants in each zipcode. The health, pollution, and confounder data are then linked by zipcode for analysis.

Availability (Mandatory)
Our data are compiled from numerous sources. Most of the data are confidential and cannot be made public; however, we provide information about both how to access the data that are public and how to request access to data that are private.

1. Medicare data: These contain highly sensitive health information, and are only available to researchers with qualifying private servers. Access can be requested via application to the Centers for Medicare and Medicaid Services (see https://www.resdac.org/research-identifiable-files-rif-requests).

2. Air pollution data: Our factual pollution data come from two different sources. Factual particulate matter (PM2.5) data are publicly available and can be downloaded here: http://fizz.phys.dal.ca/~atmos/martin/?page_id=140.
   Factual ozone (O3) data are not publicly available. Access can be requested via correspondence with the authors of the following paper:

Di, Q., Rowland, S., Koutrakis, P. and Schwartz, J., 2017. A hybrid model for spatially and temporally resolved ozone exposures in the continental United States. *Journal of the Air & Waste Management Association*, *67*(1), pp.39-52.

Counterfactual data for both pollutants are obtained from the US Environmental Protection Agency (EPA). The data are considered internal, but access can be requested via the EPA website (see https://www.epa.gov/clean-air-act-overview/forms/contact-us-about-clean-air-act).

3. Confounder data: Confounder data are publicly available from the US census, and they can be downloaded from the US Census Bureau website (https://www.census.gov/data.html).

4. Synthetic data: We provide synthetic zipcode level data that closely match some of the key features of the Medicare and air pollution data while maintaining confidentiality of the data sources used in the real data analysis.

Description (Mandatory if data available)
Because our real data cannot be made public, here we describe the structure of the synthetic data submitted to demonstrate the methods and code. The structure of our real datasets is analogous. Separate datasets are created for each year under study. In the year-specific datasets, each row corresponds to a zipcode and each column is a variable representing some feature of the zipcodes. The columns in the dataset are as follows:

- id=Zipcode ID
- mort=Count of mortality events
- dementia=Count of dementia events
- cvd=Count of CVD events
- pmWith=Factual PM2.5
- ozWith=Factual O3
- pmNo=Counterfactual PM2.5
- ozNo=Counterfactual O3
- X1-X5=Confounders

These synthetic data are generated with a portion of the code described below and are exported into an .RData file

Optional Information (complete as necessary)
None to report

Code

Abstract (Mandatory)

We provide documented code to (1) reproduce the main results in Section 5 of the manuscript on a synthetic dataset and (2) reproduce the simulation results shown in Section 4 of the manuscript, with sufficient computing time. The code should allow users to become familiar with the required inputs of the models, how the data should be structured, and what type of output is obtained for a real data analysis.

Description (Mandatory)

Code is written in R and is publicly available at https://github.com/rachelnethery/AQregulation. In order to run the code, the following R packages must be installed:

- mvtnorm
- dplyr
- splines
- BayesTree
- xtable
- ggplot2

For both the 'real' data analysis and the simulations, the code splits the data into the following objects for processing:

- N: sample size (i.e. number of zipcodes)
- Y: Vector of count of health events in each zipcode
- X: Matrix of confounders
- Eobs: Matrix of observed/factual pollutants
- Ecf: Matrix of counterfactual pollutants

The code then executes our matching procedure, BART and Poisson regression model on these data, as described in the manuscript. For the 'real' data analysis, the point estimates and confidence/credible intervals for each method and each health outcome are output to a .txt file. This .txt file is then read back into R to create the figures/tables in the main results. For the simulations, results are a data frame for each tested method, containing point estimates and confidence/credible intervals for each simulated dataset. These data frames, along with the true parameter values, are output to .RData files. (These results can then be used to create Table 2, if the simulations are run for all simulation type/tuning parameter combinations. See below for more detail.)

Optional Information (complete as necessary)

- If a cluster can be accessed, the simulations can be modified to run in parallel to dramatically decrease computing time.

Instructions for Use

Reproducibility (Mandatory)

1. 'Real' data analysis: Code in subfolder 'real_data'
    a. What is to be reproduced: Table 2 and Figure 2 in the manuscript, containing the main results for the paper (but for the synthetic dataset instead).
    b. How to reproduce analyses: The master script, 'realdata_master.R', does the following: (1) generates and exports the synthetic data to an .RData file; (2) reads in the data, conducts the analyses, and exports the results to a .txt file; (3) reads in results to create Table 2, which is saved in a file named 'table_2.txt'; and (4) reads in the results to create Figure 2, which is saved in a file named 'figure_2.pdf'. If all the files/scripts are placed in a folder with the same structure as the github repository, then the master script can be run without any changes to produce these results. The files 'table_2_verify.txt' and 'figure_2_verify.pdf' can be used to verify that the user has obtained the same results that we obtained for the synthetic data.
    c. Expected run-time of the workflow: Approximately 2 hours.

2. Simulations: Code in subfolder 'sims'
    a. What is to be reproduced: With sufficient computing time, this code could be used to reproduce all simulation results shown in Section 4 of the manuscript. In practice, unless a cluster can be accessed to run the simulations in parallel, reproducing all simulations is likely not computationally feasible. Using the instructions at the top of the code, users can set parameter values to select the simulation type, tuning parameters, and number of simulations they wish to test. By default we have set the parameters to create and analyze a single simulated dataset of type S-1 (described in the manuscript).
    b. How to reproduce analyses: The master script, 'simulation_master.R', generates the data and then runs each of the competing methods on the dataset. Data frames containing the results are exported into an .RData file. If all the files are placed in a folder with the same structure as the github repository, then the master script can be run without any changes.
    c. Expected run-time of the workflow: Approximately 30 mins for a single simulated dataset.

Replication (Optional)

Nothing to report

Notes

Code is also included with the submission to allow for blinded reproduction of the results.