**Read Me**

This document provides instructions for replicating “Hazed and Confused: The Effect of Air Pollution on Dementia” by Kelly C. Bishop, Jonathan D. Ketcham, and Nicolai V. Kuminoff.

Replication data and code are contained in the archive BKK.zip.

**Software Requirements**: Stata (code was last run with version 16)

Matlab (code was last run with release 2018a)

1. **Data Availability**

Our estimation sample is constructed by merging air pollution data from the US Environmental Agency (EPA) with administrative records from the US Centers for Medicare and Medicaid Services (CMS) and Geolytics data on neighborhoods. The EPA data are contained within the BKK.zip archive. Our data use agreements with Geolytics and CMS prohibit us from posting their data. Instructions for obtaining each data set are provided below.

* 1. **EPA Data**

All EPA data used in our study are contained in the BKK.zip folder \data\EPA**.** Data on air pollution at monitoring stations were downloaded [here](https://aqs.epa.gov/aqsweb/airdata/download_files.html). Data on county attainment status were downloaded [here](https://www.epa.gov/green-book).

* 1. **Instructions for Obtaining CMS Identifiable Data**

The process to request CMS identifiable data is explained [here](https://resdac.org/cms-research-identifiable-request-process-timeline). The process begins by contacting the Research Data Assistance Center at 1-888-973-7322 or [resdac@umn.edu](mailto:resdac@umn.edu). The variables used in our study are listed below in section 2.

* 1. **Instructions for Obtaining Geolytics Data**

Geolytics Zip+4 data can be purchased [here](https://geolytics.com/zip-4-product) or by contacting Geolytics at 1-908-707-1505 or [questions@geolytics.com](mailto:questions@geolytics.com).

1. **Contents of the BKK.zip Archive**

Unzipping the archive yields a folder “BKK” with the following contents:

\data

\data\CMS

\data\EPA

\data\Geolytics

\results

BKK.do

pollution\_merge\_cms\_bal\_dd.m

pollution\_merge\_cms\_nobal\_dd.m

pollution\_merge\_zip5\_bal\_dd.m

pollution\_merge\_zip5\_nobal\_dd.m

1. **Replication Instructions**

Step 1: Follow the instructions in section 1 to obtain the CMS and Geolytics data and post them in the \BKK\data\CMS and \BKK\data\Geolytics sub-folders. Section 4 describes the necessary file contents.

Step 2: Set the address for the BKK folder on the following scrip files:

* line 6 of \BKK\BKK.do
* line 5 of \BKK\pollution\_merge\_cms\_bal\_dd.m
* line 5 of \BKK\pollution\_merge\_cms\_nobal\_dd.m
* line 5 of \BKK\pollution\_merge\_zip5\_bal\_dd.m
* line 5 of \BKK\pollution\_merge\_zip5\_nobal\_dd.m

Step 3: Run the BKK.do script file. It merges the CMS, EPA, and Geolytics data, reproduces all tables and figures in the paper and online appendix, and saves them in the \results folder. During the data construction process, BKK.do pauses and prompts the user to run the four Matlab script files referenced in Step 2.

1. **Contents of the folders: \BKK\data\CMS and \BKK\data\Geolytics**
   1. **\BKK\data\Geolytics**

Geolytics software should be used to export a file named “geolytics\_bg\_data.csv” that contains the following Geolytics variables: areakey, acsage65y0, acsttpop0, whnohsp0, race0, black0, hispan0, tt8gr0, tt12gr0, tths0, ttsmcol0, ttassoc0, ttbach0, ttgrad0, educat0, medhhin0, pcapin0, acsowochu0, acsrnochu0, acsvachu0, acstothu, mdyrblt0, medvalo0, aggvalo0, mdgrsrn0, zip4, lon, lat, block, blkgrp, state, and county.

* 1. **\BKK\data\CMS**

CMS data should be used to build the following files:

**analytic\_*year*.dta** for each *year* from 1999 through 2013. The file for each *year* should contain the following variables:

bene\_id unique beneficiary id code

rti\_race\_cd Research Triangle Institute (RTI) Race Code

bene\_birth\_dt Date of Birth

male =1 for males

death\_date Date of death

ma\_months months of Medicare Advantage enrollment

parta\_tm\_months months of enrollment in Medicare Part A

ami\_ever First occurrence date of Acute Myocardial Infarction

alzh\_ever First occurrence date of Alzheimer's Disease

alzh\_demen\_ever First occurrence date of Alzheimer's Disease and related dementias

atrial\_fib\_ever First occurrence date of Atrial Fibrillation

chronickidney\_ever First occurrence date of Chronic Kidney Disease

copd\_ever First occurrence date of Chronic Obstructive Pulmonary Disease

chf\_ever First occurrence date of Heart Failure

diabetes\_ever First occurrence date of Diabetes

glaucoma\_ever First occurrence date of Glaucoma

ischemicheart\_ever First occurrence date of Ischemic Heart Disease

depression\_ever First occurrence date of Depression

osteoporosis\_ever First occurrence date of Osteoporosis

ra\_oa\_ever First occurrence date of Rheumatoid Arthritis / Osteoarthritis

stroke\_tia\_ever First occurrence date of Stroke / Transient Ischemic Attack

cancer\_breast\_ever First occurrence date of Breast Cancer

cancer\_colorectal\_ever First occurrence date of Colorectal Cancer

cancer\_prostate\_ever First occurrence date of Prostate Cancer

cancer\_lung\_ever First occurrence date of Lung Cancer

cancer\_endometrial\_ever First occurrence date of Endometrial Cancer

asthma\_ever Asthma first occurrence date

hyperl\_ever Hyperlipidemia first occurrence date

hypert\_ever Hypertension first occurrence date

fibromyalgia\_ever Fibromyalgia first occurrence date

leuklymph\_ever Leukemias and Lymphomas first occurrence date

pvd\_ever Peripheral Vascular Disease first occurrence date

Note that “first occurrence date” in the above variable definitions refers to the first date on record since January 1, 1999 that the chronic condition appeared based on the Medicare chronic condition warehouse variable. The precise definitions for each chronic condition are available from CMS and ResDAC. The months of Medicare Advantage and Part A enrollment refer to the number of months during the calendar year that the beneficiary was enrolled in those programs. Finally, the files for 2006 through 2013 should additionally contain the following variable:

mbsf\_d\_rx\_claims\_months months of Medicare Part D enrollment

These datasets are built from the MBSF AB files, the MBSF D files, the MBSF chronic condition files and the MBSF other chronic condition files.

**firstrx.dta**. The file should contain the following variables:

bene\_id unique beneficiary id code

alzheimersinit first date that the beneficiary began taking one of these five drugs: donepezil, galantimine, rivastigmine, memantine, and donepezil and memantine in combination.

This dataset is built from the CMS Prescription Drug Event files.

**spend\_*year*.dta** for each *year* from 1999 through 2013. The file for each year should contain the following variables:

bene\_id unique beneficiary id code

spendnonrx\_total total gross spending on all services covered by Medicare Parts A and B for the calendar year

These datasets are built from the MBSF Cost and Use files.

**bene\_location\_history.dta**. The file should contain the following variables:

bene\_id unique beneficiary id code

zip4\_*year* zip+4 code of residential location as of January 1 of each calendar *year* from 1999 through 2013

state*year* state fips code of residential location as of January 1 of each calendar *year* from 1999 through 2013

county*year* county fips code of residential location as of January 1 of each calendar *year* from 1999 through 2013

CBSA*year* Census core business statistical area code of residential location as of January 1 of each calendar *year* from 1999 through 2013

startyear\_cms first year the beneficiary was enrolled in Medicare

This dataset is built from the MBSF AB files. The CBSA variables are merged in based on a CMS FIPS to SSA Crosswalk file and the Census 2013 FIPS to CBSA Crosswalk file.

**Data References**

Geolytics. 2022. “Zip+4 Product.” <https://geolytics.com/zip-4-product> Accessed October 9, 2022.

United States Census Bureau. 2022. “Core-Based Statistical Area (CBSA) to Federal Information Processing Series (FIPS) County Crosswalk”. <https://www.nber.org/research/data/census-core-based-statistical-area-cbsa-federal-information-processing-series-fips-county-crosswalk> Accessed October 9, 2022.

United States Environmental Protection Agency. 2022. “Pre-Generated Data Files.” <https://aqs.epa.gov/aqsweb/airdata/download_files.html> Accessed October 9, 2022.

United States Environmental Protection Agency. 2022. “Nonattainment Areas for Criteria Pollutants.” <https://www.epa.gov/green-book> Accessed October 9, 2022.

Centers for Medicare and Medicaid Services. 2022. “Master Beneficiary Summary File (MBSF) LDS.” <https://www.cms.gov/Research-Statistics-Data-and-Systems/Files-for-Order/LimitedDataSets/MBSF-LDS> Accessed October 9, 2022.

Centers for Medicare and Medicaid Services. 2022. “Part D Claims Data.” <https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovGenIn/PartDData> Accessed October 9, 2022.

Centers for Medicare and Medicaid Services. 2022. “SSA to Federal Information Processing Series (FIPS) State and County Crosswalk.” <https://www.nber.org/research/data/ssa-federal-information-processing-series-fips-state-and-county-crosswalk> Accessed October 9, 2022.