**Code Review Memo**

**Project title:** Wildfires and DME in older adults who use electricity-dependent medical equipment

**Primary analyst**: Heather M

**Code reviewer:** Ben Steiger

**Date of code review: May/x/2022**

**Code location and structure:**

* **All Scripts:** …\GitHub\wildfires\_and\_DME\data\_processing\code\_pm\_analyses **and** …\GitHub\wildfires\_and\_DME\data\_processing\code\_proximity\_analyses. Data are respectively in …\data\_pm\_analyses and …\data\_proximity\_analyses. Raw data are in …\raw\_data.
* **Structure: wildfires\_and\_DME.Rproj**
  + **Everything can be run in the R project with the here() package**
* **NOTE: the pathnames are confused between \ and /. This shouldn’t matter if you use the here package!**

**Summary of scripts to review (in order):**

This project has two data processing pipelines in it, to make two separate analytic datasets for two separate analyses.

Code PM Analyses: (all data here that’s written gets written into the data\_pm\_analyses folder).

1. 01\_get\_zips\_temp\_data.R gets ZCTA codes in the Kaiser study area and saves them as a csv in the data folder for use in the next two scripts. It uses the Kaiser data, which are in their own folder, in the raw data folder as DMEdatasets20200929172326.
2. 02\_calculate\_temp\_day\_zcta.R processes temperature data, and gets the mean temperature for each zcta on each day between Jan 1st 2016 and Dec 31st 2019. It uses PRISM data in the raw data folder, and outputs the temperature data into the data folder.
3. 03\_get\_acs.R gets American Community Survey variables relevant to the analyses using the tidycensus package, and tidies them and writes the tidied variables in the data folder.
4. 04\_pm\_dat.R This script merges all previous data into a full dataset. It will merge the 5 Kaiser outcome files into one dataset containing all outcomes (from raw\_data), add temperature data for all zips to outcome data (from data\_pm\_analyses), add PM exposure information from corrected PM files (INFO IN KAISER IS WRONG.), create an offset variable for population exposed in each ZCTA grouping using population counts of each ZCTA from Kaiser, add ACS covariates by ZCTA. Then, it will save the dataset to data\_pm\_analyses.
5. 05\_add\_groupings\_an\_dat.R adds spatial groupings to the previous dataset so that we can aggregate the outcome (visit counts) over these higher-level spatial groupings. It writes an analytic dataset.
6. 06\_add\_weekly\_pm.R aggregates grouping-level PM to the weekly level, for some sensitivity analyses.

Code proximity analyses: (all data written gets written to the data\_proximity\_analyses folder).

1. 01\_get\_zips\_temp\_data.R gets ZCTA codes in the Kaiser study area and saves them as a csv in the data folder for use in the next two scripts. It uses the Kaiser data, which are in their own folder, in the raw data folder as DMEdatasets20200929172326. (Same as previous.)
2. 02\_calculate\_temp\_day\_zcta.R processes temperature data, and gets the mean temperature for each zcta on each day between Jan 1st 2016 and Dec 31st 2019. It uses PRISM data in the raw data folder, and outputs the temperature data into the data folder.
3. 03\_find\_exposed\_zctas.R will take wildfire evacuation boundary data, Kaiser data on wildfire boundary proximity, and the census California ZCTA shapefile, and mark certain ZCTAs as exposed to wildfire if they are within 10km of an evacuation zone or within 20km of a wildfire boundary.
4. 04\_find\_exposed\_control\_ZCTAs.R This script finds all ZCTAs exposed to fires (within 20 km of a fire boundary) during our study period in California, and writes a list of these ZCTAs along with the fire ignition date to the data\_proximity\_analyses folder. It’s meant to remove ZCTAs that were not exposed to either the Woolsey or Getty fires, but were exposed to other fires, and thus should not be in the control population.
5. 05\_an\_dat.R creates analytic dataset for proximity to wildfires models. This script is designed to merge all previous data into four analytic datasets enabling a negative binomial analysis with 'wildfire proximity and evacuation as the exposures. It will: merge the 5 Kaiser outcome files into one dataset containing all outcomes (from raw\_data), add temperature data for all zips to outcome data (from data\_proximity\_analyses), create and add a variable called weekyears to help with time series add wildfire exposure information, both for exposed regions and for removing exposed regions from controls, from data\_proximity\_analyses, create an offset variable for population exposed in each week using population counts of each ZCTA from Kaiser. Then, it will save the analytic dataset to data\_proximity\_analyses.

**Detailed code summaries for review:**

**PM Analyses:**

1. **01\_get\_zips\_temp\_data.R**

* **Overview:** Takes Kaiser DME data from raw\_data folder and outputs a list of ZCTAs in the Kaiser study area.
* **Inputs: …/wildfires\_and\_DME/data\_processing/raw\_data/DMEdatasets20200929172326/dme\_anydisease\_A\_09282020.csv**
* **Outputs: …/wildfires\_and\_DME/data\_processing/data\_pm\_analyses/zcta.csv**
* **Notes:**
  + We only need to reference one DME dataset because they’re all the same, other than the outcome/visit type recorded in the data.

1. **02\_calculate\_temp\_day\_zcta.R**

* **Overview: DO NOT RUN THIS SCRIPT.** This script loads a raster file of PRISM temperature data for particular year, month and day, loads shapefile of ZCTAs, isolates each zcta's temperature data in raster via the shapefile and, then finds the area-weighted mean over the ZCTA. It outputs mean temperature data by day and ZCTA for that year
* **Inputs:** PRISM temperature data raster files, downloaded from the PRISM website. **All files in:**

1. **…/wildfires\_and\_DME/data\_processing/raw\_data/PRISM2016,**
2. **…/PRISM2017**
3. **…/PRISM2018**
4. **…/PRISM2019, and**
5. **…/wildfires\_and\_DME/data\_processing/raw\_data/zcta\_shapefile**

* **Outputs:** Two ED-tract crosswalks, one area-weighted and one not.

1. **…/wildfires\_and\_DME/data\_processing/data\_pm\_analyses/weighted\_area\_raster\_zip\_tmean\_daily\_2016.rds**
2. **/Users/heathermcbrien/Documents/research project folders/wildfires\_and\_DME/data\_processing/data\_pm\_analyses/weighted\_area\_raster\_zip\_tmean\_daily\_2017.rds**
3. **/Users/heathermcbrien/Documents/research project folders/wildfires\_and\_DME/data\_processing/data\_pm\_analyses/weighted\_area\_raster\_zip\_tmean\_daily\_2018.rds**
4. **/Users/heathermcbrien/Documents/research project folders/wildfires\_and\_DME/data\_processing/data\_pm\_analyses/weighted\_area\_raster\_zip\_tmean\_daily\_2019.rds**

* **Notes:**
  + This script is a loop adapted from Robbie Parks’ code.

1. **03\_get\_acs\_vars.R**

* **Overview:** This script will use tidycensus to get ACS data from the 2015-2019 ACS on the following covariates:
  + population < 5 years old, 5-18 years, 18-64
  + population density
  + % male
  + home ownership
* **Inputs: No inputs; uses the tidycensus package to pull covariates. May need to use your own API key in order to make this work. I haven’t included mine in the script bc I have it autoloaded into my own R environment.**
* **Outputs:**

1. **…wildfires\_and\_DME/data\_processing/data\_pm\_analyses/acs\_covars.csv**

* **Notes:**
  + Important to verify that all the variables are named correctly and they have reasonable values for what they are, because that’s tricky.

**4. 04\_pm\_dat.R**

* **Overview:** This script merges all previous data into a full dataset. It will merge the 5 Kaiser outcome files into one dataset containing all outcomes (from raw\_data), add temperature data for all zips to outcome data (from data\_pm\_analyses), add PM exposure information from corrected PM files (INFO IN KAISER IS WRONG.), create an offset variable for population exposed in each ZCTA grouping using population counts of each ZCTA from Kaiser, add ACS covariates by ZCTA. Then, it will save the dataset to data\_pm\_analyses.
* **Inputs: I won’t list inputs because there are too many. They are listed above and all included at the beginning of the script.**
* **Outputs:**

1. **…\data\_pm\_analyses\** **daily\_with\_acs\_pm\_temp.csv**

* **Notes:**
  + Final dataset should be at the day-zcta level. Many joins; important to verify that joins are performed correctly, as are summaries/summarise functions.

**5. 05\_add\_groupings\_an\_dat.R**

* **Overview:** This script adds groupings to the ZCTAs in the analytic dataset so that we have some geographic units which are a bit bigger than the zctas themselves, and summarize existing variables over those groupings. It saves the resulting analytic dataset in the data\_pm\_analyses folder.
* **Inputs:**

1. **…\data\_pm\_analyses\** **daily\_with\_acs\_pm\_temp.csv**
2. **…\raw\_data\** **zcta\_shapefile\tl\_2019\_us\_zcta510.shp**

* **Outputs:**

**1. …\raw\_data\data\_pm\_analyses\an\_dat\_daily.csv**

* **Notes:**
  + The strategy for grouping ZCTAs is to use the ZCTA code numbers (90001, 90002, etc.) to form groups. Here is what we wrote about it:
  + “We created higher-level groupings of ZCTAs using the numerical ZCTA codes. We used a bespoke method, and then tested the resulting spatial groupings to make sure that ZCTAs grouped together had similar exposure measurements, to guard against exposure misclassification. We grouped ZCTAs together if all their numerical codes differed by 1 in sequence. For example, codes 90001-90008 and 90011-90014 were in the study area. We grouped codes 90001 - 90008 together, as they are all sequentially 1 digit apart, while 90011-90014 formed a second grouping. This method resulted in groupings of ZCTAs that were all adjacent, since similar codes tend to be geographically close.”
  + Good to check that this is working as intended.

1. **06\_add\_weekly\_pm.R**

* **Overview:** This script creates mean weekly PM measurements by ZCTA grouping. It takes in the grouped PM dataset, and outputs mean weekly PM measurements. Should be relatively simple.
* **Inputs:**
  + **…\data\_pm\_analyses\an\_dat\_daily.csv**
* **Outputs:**
  + **…\data\_pm\_analyses\an\_dat\_weekly.csv**
* **Notes:**
  + Should be relatively simple.

**Proximity Analyses:**

1. **01\_get\_zips\_temp\_data.R**

* **Overview:** Takes Kaiser DME data from raw\_data folder and outputs a list of ZCTAs in the Kaiser study area.
* **Inputs: …/wildfires\_and\_DME/data\_processing/raw\_data/DMEdatasets20200929172326/dme\_anydisease\_A\_09282020.csv**
* **Outputs: …/wildfires\_and\_DME/data\_processing/data\_pm\_analyses/zcta.csv**
* **Notes:**
  + We only need to reference one DME dataset because they’re all the same, other than the outcome/visit type recorded in the data.
* SAME AS PREVIOUS – should just be copy-pasted!

1. **02\_calculate\_temp\_day\_zcta.R**

* **Overview:** This script loads a raster file of PRISM temperature data for particular year, month and day, loads shapefile of ZCTAs, isolates each zcta's temperature data in raster via the shapefile and, then finds the area-weighted mean over the ZCTA. It outputs mean temperature data by day and ZCTA for that year
* **Inputs:** PRISM temperature data raster files, downloaded from the PRISM website. **All files in:**

1. **…/wildfires\_and\_DME/data\_processing/raw\_data/PRISM2016,**
2. **…/PRISM2017**
3. **…/PRISM2018**
4. **…/PRISM2019, and**
5. **…/wildfires\_and\_DME/data\_processing/raw\_data/zcta\_shapefile**

* **Outputs:** Two ED-tract crosswalks, one area-weighted and one not.

1. **…/wildfires\_and\_DME/data\_processing/data\_pm\_analyses/weighted\_area\_raster\_zip\_tmean\_daily\_2016.rds**
2. **/Users/heathermcbrien/Documents/research project folders/wildfires\_and\_DME/data\_processing/data\_pm\_analyses/weighted\_area\_raster\_zip\_tmean\_daily\_2017.rds**
3. **/Users/heathermcbrien/Documents/research project folders/wildfires\_and\_DME/data\_processing/data\_pm\_analyses/weighted\_area\_raster\_zip\_tmean\_daily\_2018.rds**
4. **/Users/heathermcbrien/Documents/research project folders/wildfires\_and\_DME/data\_processing/data\_pm\_analyses/weighted\_area\_raster\_zip\_tmean\_daily\_2019.rds**

* **Notes:**
  + This script is a loop adapted from Robbie Parks’ code.
* SAME AS PREVIOUS – should just be copy-pasted!

1. **03\_find\_exposed\_zctas.R**

* **Overview:** This script will take wildfire evacuation boundary data, Kaiser data on wildfire boundary proximity, and the census California ZCTA shapefile, and mark certain ZCTAs as exposed to wildfire if they are within 10km of an evacuation zone or within 20km of a wildfire boundary, and output a csv with the exposure information.
* **Inputs:**

1. **…\raw\_data\** **zcta\_shapefile\tl\_2019\_us\_zcta510.shp – zcta shapefile**
2. **…\data\_proximity\_analyses\woolsey\_evac\_boundary\woolsey\_evac\_boundary.shp – Woolsey evac boundary**
3. **…\data\_proximity\_analyses\getty\_evac\_boundary\getty\_evac\_boundary.shp – Getty evac boundary**
4. **…/wildfires\_and\_DME/data\_processing/raw\_data/DMEdatasets20200929172326/dme\_anydisease\_A\_09282020.csv – DME data – only need one file because we just need fire proximity information in all files (which are the same)**

* **Outputs:**

1. **…/data\_proximity\_analyses/proximity\_exposure.csv**

* **Notes:**
  + Pretty straightforward.

**4. 04\_find\_all\_exposed\_control\_ZCTAs.R**

* **Overview:** This script identifies ZCTAs that have been exposed to other fires, other than the Getty and Woolsey, and flags them as exposed after the start date of that fire. It takes the wildfire data from all of California during the study period, and uses the ignition date and fire boundaries. It intersects the boundaries with the ZCTA shapefile, and then assigns a 1 to exposed ZCTAs after the ignition date. It writes a file of these exposure data.
* **Inputs:**

1. **…/calfire\_disasters\_2016\_2020.rdata - calfire data**
2. **…/wildfires\_and\_DME/data\_processing/data\_proximity\_analyses/zcta.csv**

* **Outputs:**

1. **…\data\_proximity\_analyses\** **exposed\_controls.csv**

* **Notes:**
  + The key of this script is the control assignment – need to check that it’s working properly.

**5. 05\_an\_dat.R**

* **Overview:** This script is designed to merge all previous data into an analytic dataset.
  + It will:
  + merge the 5 Kaiser outcome files into one dataset containing all outcomes (from raw\_data)
  + add temperature data for all zips to outcome data (from data\_proximity\_analyses)
  + create and add a variable called weekyears to help with time series
  + add wildfire exposure information, both for exposed regions and for removing exposed regions from controls, from data\_proximity\_analyses
  + create an offset variable for population exposed in each week using population counts of each ZCTA from Kaiser

Then, it will save the analytic dataset to data\_proximity\_analyses.

* **Inputs: A lot, see script.**
* **Outputs:**

**1. …\raw\_data\data\_proximity\_analyses\an\_dat.csv**

* **Notes:**
  + Good to make sure joins are working properly.