Documentation for Alexander and Schwandt (2022) “The Impact of Car Pollution on Infant and Child Health: Evidence from Emissions Cheating”

Introduction

This document explains the organization of the replication package, the statistical programs used, and the code that builds each dataset used in Alexander and Schwandt (2022) “The Impact of Car Pollution on Infant and Child Health: Evidence from Emissions Cheating”.

Replication package organization

The replication package is organized with all the programs in the top level of the file structure, and the raw data sets organized in subfolders. Cleaned and intermediate datasets as well as figures and tables are saved into the main folder with the do files. This document details how to use each of the do files to clean and analyze the data.

For many of the program files there are also versions with the same name but the suffix “_cz” appended to the name. These are versions of the main programs that replicate the analysis at the commuting zone level instead of the county level. Similarly, there are files with “_year” instead of “_month” in the file name; these are for analyses in the appendix at the annual level. As these alternate analyses are only used in the appendix, we have placed these programs in a separate subfolder for clarity. However, the paths are set from the main folder, so these files should be moved to the same folder as the main programs if they are being used. For brevity, we have not listed out each of these analogous files in what follows.

For proprietary data files, the descriptions of the datasets in the following pages detail where the data can be obtained. However, the data files are not included in the replication package and some programs will reference these files.

For public use files, the descriptions of the datasets in the following pages detail where the data can be obtained. We have also uploaded raw datasets to replication website when we are able, as noted in the data availability statement. The vehicle registration, sales, natality, mortality, and hospitalization data are proprietary data purchased for use in this research and/or restricted-use microdata obtained for this research. These files could be applied for or purchased by other researchers.

Statistical software and packages

Data cleaning and all analysis except for map creation was done in Stata; maps were created in ArcPro.

The following stata packages need to be installed. Each can be installed by executing the command in stata: “ssc install `x’”, where x is: carryforward, spmap, shp2dta, mif2dta, binscatter, grc1leg, estout, reghdfe.

Main data construction overview

This document explains the code that builds each dataset used in our Volkswagen/Chrysler cheating diesel analysis. The main analysis dataset (“month_combined.dta”) contains:

1. County-level annual vehicle registration data (Proprietary data from IHS Markit)
2. County-level monthly vehicle sales data (Proprietary data from Experian Autocount; aggregated from individual-level vehicle sales data)
3. County-level monthly EPA $PM_{2.5}$, PM_{10}, ozone, CO, and NO_2 data (Environmental Protection Agency)
4. County-level 1997 $PM_{2.5}$ NAAQS nonattainment status (Environmental Protection Agency)
5. County-level 1997 ozone NAAQS nonattainment status (Environmental Protection Agency)

6. County-level monthly PM$_{2.5}$ satellite pollution data (Environmental Protection Agency /Stanford)
7. County-level monthly natality and mortality data (Restricted-use microdata from the Center for Disease Control's Vital Statistics program; aggregated from individual-level data)
8. County-level quarterly Emergency Department visit data (Restricted-use microdata from the Healthcare Cost Utilization Project's State Emergency Department Databases program; aggregated from individual-level ED visit data)
9. County-level annual unemployment rate data (Bureau of Labor Statistics)
10. County-level annual median income and poverty rate data (Census Bureau's Small Area Income and Poverty Estimates program)
11. County-level annual population data (Census Bureau)
12. County-level monthly temperature and rainfall data (Oregon State University's PRISM weather program)
13. County-level yearly mortgage data (Home Mortgage Disclosure Act)
14. County-level yearly vehicle miles data (Federal Highway Administration)

Data Availability Statement

1. County-level annual vehicle registration data, *Vehicles in Operation data* (Proprietary from IHS Markit)[1]

The raw files are a custom slice of the "Vehicles in Operation" registration data from IHS Markit. Our data consist of counts of vehicle registrations by make/model/model year for each county in the U.S., along with counts of the gas equivalents and the total number of gas and diesel vehicles. These data are proprietary and can be purchased from IHS Markit. For more information on the dataset and how to purchase, see their website: https://ihsmarkit.com/products/automotive-market-data-analysis.html

2. County-level monthly vehicle sales data, *Experian Autocount data* (Proprietary data from Experian Autocount; aggregated from individual-level vehicle sales data)[2]

Individual-level vehicle sales data for every vehicle sale in the U.S. each month, broken down by make/model/model year and new/used status. Does not include fuel type. These data are proprietary, and access can be purchased from Experian: https://www.experian.com/automotive/autocount

3. County-level monthly EPA PM$_{2.5}$, PM$_{10}$, ozone, CO, and NO$_2$ data, *AirData Download Files – Daily Summary Data* (Environmental Protection Agency Air Quality System)[3]

The raw data files consist of daily pollution reading from the EPA's national network of air quality monitors. This data is publicly available, and the most recent data can be downloaded from the EPA's website at https://aqs.epa.gov/aqsweb/airdata/download_files.html#Daily
Raw files also available as a part of our replication package.

4. County-level 1997 PM$_{2.5}$ NAAQS nonattainment status (Environmental Protection Agency)[4]

Attainment designations were published in a document released by the EPA entitled "Air Quality Designations and Classifications for the Fine Particles (PM$_{2.5}$) National Ambient Air Quality Standards." at https://www.federalregister.gov/documents/2005/01/05/05-1/air-quality-designations-and-classifications-for-the-fine-particles-pm25-national-ambient-air. Also available as a part of our replication package.

5. County-level 1997 ozone NAAQS nonattainment status (Environmental Protection Agency)[5]
   Attainment designations were published in a document released by the EPA entitled “Air Quality Designations and Classifications for the 8-Hour Ozone National Ambient Air Quality Standards; Early Action Compact Areas with Deferred Effective Dates.” at https://www.federalregister.gov/documents/2004/04/30/04-9152/air-quality-designations-and-classifications-for-the-8-hour-ozone-national-ambient-air-quality. Also available as a part of our replication package.

6. County-level monthly $PM_{2.5}$ satellite pollution data (Environmental Protection Agency)[6]
   Dataset received from Wes Austin, at the Environmental Protection Agency's National Center for Environmental Economics (https://wes-austin.com/research/), based on the monthly raw data available at ftp://stetson.phys.dal.ca/Aaron/V4NA02/Monthly/ASCII/PM25/

7. County-level monthly natality and mortality data, *Restricted-Use Vital Statistics Data – Deaths, and Restricted-Use Vital Statistics Data – Births* (National Center for Health Statistics, Center for Disease Control’s Vital Statistics program; aggregated from individual-level data)[7,8]
   Applications for the restricted microdata can be submitted at https://www.naphsis.org/research-requests

8. County-level quarterly Emergency Department visit data, *State Emergency Department Databases* (Agency for Healthcare Research and Quality, restricted-use microdata from the Healthcare Cost Utilization Project’s State Emergency Department Databases program; aggregated from individual-level ED visit data)[9]
   The data can be purchased from the HCUP website at https://www.hcup-us.ahrq.gov/tech_assist/centdist.jsp

9. County-level annual unemployment rate data (Bureau of Labor Statistics)[10]
   Data available at: https://www.bls.gov/lau/. Raw files also available as a part of our replication package.

10. County-level annual median income and poverty rate data, *Small Area Income and Poverty Estimates* (U.S. Census Bureau’s Small Area Income and Poverty Estimates program)[11]
    Data available at: https://www.census.gov/programs-surveys/saipe.html. Raw files also available as a part of our replication package.

11. County-level annual population data, *County Population by Characteristics* (U.S. Census Bureau)[12]
    Data available at: https://www.socialexplorer.com/explore-maps. Raw files also available as a part of our replication package.

12. County-level monthly temperature and rainfall data (Oregon State University’s PRISM weather program) [13]
    Data available at: https://prism.oregonstate.edu/

13. County-level yearly mortgage data, *Home Mortgage Disclosure Act data* (Consumer Financial Protection Bureau, Home Mortgage Disclosure Act)[14]

    Data available at: https://www.consumerfinance.gov/data-research/hmda/historic-data/

14. County-level yearly vehicle miles data, *Highway Statistics Series* (Federal Highway Administration, Office of Highway Policy Information)[15]

    Data available at: https://www.bts.gov/content/us-vehicle-miles. Raw files also available as a part of our replication package.

Data Citations

[1]IHS Markit (2015) "Vehicles in Operation (VIO)," Vehicle Registration Data.

[2]Experian (2015) "Experian Autocount data," County-level Monthly Vehicle Sales Data.

[3]EPA Air Quality System (2021), "AirData Download Files – Daily Summary Data," (accessed November 1, 2021). Environmental Protection Agency.

[4]Environmental Protection Agency (2005), "Air Quality Designations and Classifications for the Fine Particles ($PM_{2.5}$) National Ambient Air Quality Standards," Federal Registrar.

[5]Environmental Protection Agency (2004), "Air Quality Designations and Classifications for the 8-Hour Ozone National Ambient Air Quality Standards," Federal Registrar.

[6]Wes Austin (2019), "County-level monthly $PM_{2.5}$ satellite pollution data," the Environmental Protection Agency's National Center for Environmental Economics.

[7]NCHS mortality (2007-2016), "Restricted-Use Vital Statistics Data – Deaths," (accessed November 1, 2021). National Center for Health Statistics, Centers for Disease Control and Prevention.

[8]NCHS natality (2007-2016), "Restricted-Use Vital Statistics Data – Births," (accessed November 1, 2021). National Center for Health Statistics, Centers for Disease Control and Prevention.

[9]HCUP State Emergency Department Databases (SEDD) (2007-2015), (accessed November 1, 2021). Healthcare Cost and Utilization Project (HCUP). Agency for Healthcare Research and Quality, Rockville, MD.

[10]Bureau of Labor Statistics (2016) "Local Area Unemployment Statistics," (accessed November 1, 2021).

[11]U.S. Census Bureau SAIPE (2007-2015), "Small Area Income and Poverty Estimates," (accessed November 1, 2021).

[12]U.S. Census Bureau, County Population by Characteristics (2010-2019), (accessed November 1, 2021).

[13]PRISM Climate Group, "PRISM Climate Data," Oregon State University, (accessed November 1, 2021).

[14]Consumer Financial Protection Bureau, "Home Mortgage Disclosure Act data," (accessed November

1, 2021).

[15] Federal Highway Administration, Office of Highway Policy Information, "Highway Statistics Serie," (accessed November 1, 2021).

County-level vehicle registration data (IHS Markit)

- **Do-files:** "ihs_revision.do"
- **Years covered:** 2007, 2011, 2015, 2017, and 2018 ("rolled back" to cover 2000-2015)
- **Description:**
    - This do-file reads in the raw IHS year-end vehicle registration data and cleans it. We have year-end vehicle registration "snapshots" from 2007, 2011, 2015, 2017, and 2018.
    - Our IHS Markit data contains diesel vehicle registrations by make/model/year for each county in the U.S., along with counts of the gas equivalents and the total number of gas and diesel vehicles.
    - In particular, this do-file:
        - Flags the VW and Chrysler vehicles affected by the diesel emissions scandals, and rolls back the 2015 registration data at the make/model level to construct a yearly county-level time series of affected vehicle registrations from 2007 to 2015. (The first non-zero values appear in 2008, since the affected vehicles weren't released until MY 2009.)
        - For each affected make/model (i.e. Volkswagen Jetta), we used our Autocount vehicle sales data to compute the fraction of new versions of that make/model that were purchased in the model year, relative to the year before. (We calculate these shares in the do-file "aff_sharenew.do")
        - We use these sales shares to assign affected make/model/years in the 2015 IHS data an initial year of purchase and create a full time series of affected diesel registrations by taking cumulative sums. The simple example below illustrates how this works in practice:

"Rolling-back" example for County A

Assume County A has two types of affected make/model/years registered in 2015, such that the year-end 2015 IHS vehicle registration data for County A looks like this:

County	Make	Model	Vehicle Year	Registration Count, YE 2015
A	Volkswagen	Jetta	2011	100
A	Volkswagen	Golf	2014	100

From the Autocount sales data, we know that 81% of new Volkswagen Jettas were purchased in the same year as the model year and 76% of new Volkswagen Golfs were purchased in the same year as the model year. So, applying these shares to the 2015 IHS registration data gives us:

County	Make	Model	Vehicle Year	Registration Count, YE 2015	Number Purchased in Model Year	Number Purchased in Year Before Model Year
A	Volkswagen	Jetta	2011	100	81	19
A	Volkswagen	Golf	2014	100	76	24

Reshaping this data gives me:

County	Make	Model	Vehicle Year	Year Purchased	Vehicle Count
A	Volkswagen	Jetta	2011	2010	19
A	Volkswagen	Jetta	2011	2011	81
A	Volkswagen	Golf	2014	2013	24
A	Volkswagen	Golf	2014	2014	76

Assuming vehicles don’t move between counties, we can take cumulative sums to create a county-level annual time series of affected vehicles from 2000 to 2015:

County	Year	Number of affected vehicles
A	2007	0
A	2008	0
A	2009	0
A	2010	19
A	2011	100 (19+81)
A	2012	100 (19+81)
A	2013	124 (19+81+24)
A	2014	200 (19+81+24+76)
A	2015	200 (19+81+24+76)

- We applied a similar procedure to construct a time series of all diesels (not just affected diesels). We used the Autocount data to calculate the share of new diesels purchased in the same year as the model year (“all_sharenew.do”), and we used this share to assign all diesels an initial year of purchase. Then, we took cumulative sums to create a yearly time series of all diesel cars in each county in the U.S.
- Since we only have detailed vehicle information for diesels, we could not use this technique to create a time series of all cars. Instead, we simply linearly interpolated the missing years of data.
- We also create two alternate versions of this rolled-back vehicle registration time series by (1) starting the rollback in 2011 (instead of 2015), and (2) combining the 2015 and 2011 rolled-back times series into a single variable.
- Complication: A fraction of the most recent model year that shows up in a given IHS year-end total (i.e. MY 2016 in IHS 2015) will be coded as being sold in the model year, whereas they would have to be on the road in the IHS year. We recode these by capping the year of sale at the IHS year, i.e. MY 2016 cars in IHS 2015 are coded as 2015.
- Do file with _cz in name does the same exercise, but at the commuting zone level.

County-level monthly vehicle sales data (Experian Autocount)

- **Do-files:** “autocount.do”
- **Years covered:** 2008-2018
- **Description:**
  - This dataset provides individual-level vehicle sales data for every vehicle sale in the U.S. each month, broken down by make/model/year and new/used status. The Experian Autocount data does not differentiate between gas and diesel sales.
  - The do-file “autocount.do” attaches county codes to all of the vehicle sales using a county code-to-county name crosswalk file. Finally, it flags the affected make/model/years and collapses the data to the county-month level to generate a county-level monthly time series of vehicle sales that covers 2008 to 2018.

County-level monthly EPA PM$_{2.5}$, PM$_{10}$, ozone, CO, and NO$_2$ data

- **Do-files:** “epa_polluation_daily_revision.do”, “diesel_pollution_cleaning_r08_2020.do”, “diesel_pollution_cleaning_r08_2020_monitor_list.do”
- **Years covered:** 2000-2019
- **Description:**
  - The raw data files consist of daily pollution reading from the EPA’s national network of air quality monitors. This data is publicly available, and the most recent data can be downloaded from the EPA’s website.
  - We analyze five “criteria pollutants”: (1) PM$_{2.5}$ (parameter code 88101), (2) PM$_{10}$ (parameter code 81102), (3) CO (parameter code 42101), (4) ozone (parameter code 44201), and (5) NO$_2$ (parameter code 42602).
  - The “epa_pollution_daily_revision.do” file reads in the raw daily pollution Excel data files and cleans the county code variables.
  - Next, you need to run “Diesel_pollution_cleaning_r08_2020.do”. This do-file constructs the pollution dataset that forms the backbone of our analysis. For each pollutant, the code:
    - Drops monitoring sites located at airports and pollution monitoring days during which an “exceptional” weather event (i.e., a wildfire) occurred.
    - Since it’s possible for a single monitoring location to have more than one monitoring instrument, we averaged all pollution readings at the same monitor on the same day and collapsed the data to the county-day-monitor level.
    - Then, we found the longest-running monitor in each county and dropped all other monitors in that county. This creates a “balanced panel” of monitors, with a maximum of one observation per county-day. We also created alternative versions of the dataset where we keep the 2, 3, 4, or 5 longest-running monitors in each county, as long as (1) the 2nd-, 3rd-, 4th-, or 5th-longest-running monitor was open for at least half the number of days as the longest-running monitor in that county, and (2) we had data from all 2, 3, 4, or 5 monitors on that day.
    - Once we created a dataset of monitors, we used the EPA’s 2016 cut-points to create a daily Air Quality Index (AQI) measure for each county, and then we took the average PM$_{2.5}$, ozone, CO, and NO$_2$ concentration across all of the days

in each county-month and collapsed the data to create a monthly county-level time series of air quality and pollutant concentrations.

County-level 1997 $PM_{2.5}$ NAAQS nonattainment status (EPA)

- **Do-files:** "naaqs_pm25_attain.do"
- **Years covered:** N/A
- **Description:**
  - The do-file "naaqas_pm25_attain.do" assigns counties attainment/non-attainment status for the 1997 $PM_{2.5}$ NAAQS.
  - Each county in the U.S. was designated as in or out of attainment with the 1997 $PM_{2.5}$ NAAQS, and these designations were published in a document released by the EPA entitled "Air Quality Designations and Classifications for the Fine Particles ($PM_{2.5}$) National Ambient Air Quality Standards."
  - Technically, this document doesn't differentiate between "attainment" and "unclassifiable" counties ("unclassifiable" counties are those with insufficient data), but we consider all counties assigned the designation "attainment/unclassifiable" to be in attainment with the 1997 $PM_{2.5}$ NAAQS.
  - While most designations are made at the county level, the EPA allows regulators to classify only part of a county as being out of compliance with the 1997 $PM_{2.5}$ NAAQS. For the purposes of this analysis, we considered the entire county to be a nonattainment county if any part of it was classified as nonattainment.

County-level 1997 ozone NAAQS nonattainment status (EPA)

- **Do-files:** "naaqs_ozone_attain.do"
- **Years covered:** N/A
- **Description:**
  - The do-file "naaqs_ozone_attain.do" assigns counties attainment/non-attainment status for the 1997 8-hour ozone NAAQS.
  - Each county in the U.S. was designated as in or out of attainment with the 1997 ozone NAAQS, and these designations were published in a document released by the EPA entitled "Air Quality Designations and Classifications for the 8-Hour Ozone National Ambient Air Quality Standards; Early Action Compact Areas with Deferred Effective Dates."
  - This code is analogous to the code written to designate counties as in or out of compliance with the 1997 $PM_{2.5}$ NAAQS.
  - The key difference between the EPA's enforcement of 1997 $PM_{2.5}$ and ozone air quality standards is that the 1997 ozone NAAQS allowed counties to submit proposals to become "early action compact areas" (EACA). Counties that requested EACA status were required to submit accelerated ozone NAAQS compliance plans before the national deadline, but as a reward for their efforts, these counties were able to defer any nonattainment designations until 2008. We considered all EACAs that were assigned a "nonattainment/deferred" status to be nonattainment counties.

County-level 1997 monthly $PM_{2.5}$ satellite pollution data (Stanford/EPA)

- **Do-files:** “pm25_satellite.do”
- **Years covered:** 2000-2016
- **Description:**
  - This do-file cleans the county codes in the raw $PM_{2.5}$ satellite data. We received this dataset from Wes Austin, a graduate student at Stanford.

County-level monthly temperature and rainfall data (OSU’s PRISM program)

- **Do-files:** “county_weather.do”
- **Years covered:** 2000-2016
- **Description:**
  - This do-file constructs a monthly county-level time series of rainfall and temperature data. All of the weather statistics were calculated from daily data (i.e., mean precipitation in a month is the average daily rainfall over all of the days in the month).
  - This dataset was originally obtained by Dave Keiser (University of Massachusetts-Amherst) for a project on wastewater systems and health with Bhash Mazumder.

County-level annual median income and poverty rate data (Census Bureau’s SAIPE program)

- **Do-files:** “saipe_r09_2020.do”
- **Years covered:** 2000-2018
- **Description:**
  - This do-file uses data from the SAIPE program to construct yearly county-level estimates of (1) median household income, (2) the fraction of people living in poverty, and (3) the fraction of people ages 0-17 living in poverty. Median income figures were converted to 2010 dollars.
  - We pulled the SAIPE data directly from the Census Bureau’s website.

County-level annual population data (Census Bureau)

- **Do-files:** “county_pop_r_09_2020.do”
- **Years covered:** 2000-2018
- **Description:**
  - This do-file uses Census (2000 and 2010), intercensal (2000-2009), and post-censal (2011-2018) population estimates to construct a yearly time series of county population that contains: (1) Total population, (2) population 0-4, (3) population 5-14, (4) population 35-44, (5) population 45-64, (6) population 65-79, (7) population 80+, (8) black population, (9) white population, and (10) Hispanic population.

  - The 2000-2010 population data is pulled from the Census Bureau's website, and the 2011-2018 population data is from here from the Census Bureau website.[1]

County-level annual unemployment rate data (Bureau of Labor Statistics)

- **Do-files:** "unemprate_r09_2020.do"
- **Years covered:** 2000-2018
- **Description:**
  - This do-file reads in annual county-level unemployment rate data from 2000-2018. We downloaded these county-level unemployment rates from the Bureau of Labor Statistics' website.

County-level monthly natality and mortality data (Center for Disease Control's Vital Statistics program)

- **Do-files:** "Infix_Natality_2000_2018_r12_2020.do", "Natality_Cleaning_PART1_r12_2020.do", and "Natality_Cleaning_PART2_r10_2020.do"; "diesel_clean_VitalStats_infant_mortality.do"
- **Years covered:** 2000-2018
- **Description:**
  - We obtained restricted versions of the individual-level natality and mortality data from the CDC that contain county identifiers.
  - "Infix_Natality_2000_2016.do" reads in the raw .txt birth and death certificate files and create the variables needed for our analysis. (Warning: The birth certificates changed in 2003.)
  - "Natality_Cleaning_PART1.do" collapses the individual-level natality data by county/conception month.
  - "Natality_Cleaning_PART2.do" merges the collapsed natality data with county-level population estimates to construct birth rate variables.
  - "diesel_clean_VitalStats_infant_mortality.do" does the above steps, for infant mortality instead of natality, in a single file.
  - Versions of files with_cz collapse the data at the commuting zone level instead of the county level.

County-level quarterly Emergency Department visit data (Healthcare Cost Utilization Project's State Emergency Department Databases program)

- **Do-files:** "sedd_icd9_r08_2020.do" and "sedd_icd10_r08_2020.do"
- **Years covered:** 2005-2017
- **Description:**

[1] County population by sex/age-Table title is "Annual County and Resident Population Estimates by Selected Age Groups and Sex: April 1, 2010 to July 1, 2019 (CC-EST2019-AGESEX)"
County population by race- Table title is "Annual County Resident Population Estimates by Age, Sex, Race, and Hispanic Origin: April 1, 2010 to July 1, 2019 (CC-EST2019-ALLDATA)"

- The SEDD files contain individual-level data on all ED visits that did not result in a hospital admission for Arizona, Florida, Kentucky, New Jersey, and Rhode Island. The Arizona, Florida, and New Jersey data covers 2005-2017, the Rhode Island data covers 2007-2017, and the Kentucky data covers 2008-2017.
- The SEDD data lists the principal reason for the visit (i.e, an asthma attack), but the codes used to identify these visits switches from ICD-9 codes to ICD-10 codes in Q4 2015. The program "sedd_icd9_r09_2020.do" cleans the 2005-2014 SEDD data, flags asthma-related visits using ICD-9 codes, and collapses the data to the county-quarter level. Analogously, the program "sedd_icd10_r09_2020.do" cleans the 2015-2017 SEDD data, flags asthma-related visits using ICD-9 codes for Q1-Q3 2015 and ICD-10 codes for Q4 2015-2017, and collapses the data to the county-quarter level. The last section of "sedd_icd10_r09_2020.do" appends the 2005-2014 and 2015-2017 data together to create a quarterly county-level dataset of ED visits from 2005-2017.

County-level yearly mortgage data (public HMDA)

- **Do-files:** "1_hmda_aggregates_county.do", "2_hmda_county_sub_alt.do", "3_hmda_county_collapse.do", and "4_hmda_county_append.do"
- **Years covered:** 2000-2015
- **Raw data:** This data was queried from RADAR. To pull the public HMDA data, we used the following SQL query:

  SELECT year,action_type,agency_code,applicant_ethnicity,applicant_income,applicant_sex,census_tract, county_code,lien_status,loan_amount,loan_purpose,loan_type,msa_md,msa_md_census, occupancy,preapproval,property_type,purchaser_type,respondent_id,sequence_number, state_code FROM hmda.view_lar_hmda WHERE (view_lar_hmda.year=[year])

- **Description:**
  - The raw public HMDA data is at the loan level (i.e., there's one observation per loan application). The do-files listed above collapse the data to the county-year level, and all loan totals are adjusted to 2010 dollars.

County-level yearly vehicle miles data (FHA)

- **Do-files:** "fha_r_09_2020.do"
- **Years covered:** 2000-2018
- **Description:**
  - Constructs highway miles per capita at the county-year level.

Combining the data

- To combine all of the above datasets, run the do-file **"combine_r12_22_2020.do"**
- This do-file generates the datasets **"month_combined.dta"** and **"year_combined.dta"**. These datasets form the backbone of our analysis, and they are at the county-month and county-year levels, respectively.

- Note: we also create a commuting zone version of the combined analysis datasets, using do-files and datasets denoted with “_cz.”

Running the main analysis:

- **Main do-file:** “diesel_main_analysis_FINAL.do”
- **Secondary do-files run from the main do-file:** “diesel_clean_month_combined_FINAL.do”, “diesel_clean_month_combined_cz_FINAL.do” “diesel_clean_year_combined_FINAL.do”
- **Description:**
    - diesel_main_analysis_FINAL.do first runs the cleaning files, then runs the main analysis and constructs all the tables and figures in the paper, with a few exceptions: maps (Figures 1 and A.3), the literature review figures and tables (Tables A.9-A.11), and tables documenting EPA cut points, the MSRP of cheating models, and the details of the cheating scandal (Tables A.18-A.20).

Maps:

We created the maps in the paper in ArcPro.

- **Main analysis file:** “diesel_maps_arcpro_revision.aprx”
- **Description:** creates maps (both for main text and appendix figures).