**R Script Order for EVs and Where to Charge Them (V2-OSM)**

*Last edited: 6/3/2025*

1. Run “Cleaning EPA UST Data V2-OSM.R”.
   1. This file cleans the EPA’s facilities and tanks data and merges them into one file.
2. Run “Cleaning OSM Data.R”.
   1. This file converts OpenStreetMap’s data from geoJSON to csv and RData and performs some basic cleaning.
3. Run “Cleaning DOE Charger Data.R”.
   1. This file cleans the raw DOE public charger data, assigns geographic variables (e.g., county name, county FIPS, MSA name, MSA ID, and census tract code) to the data, and creates survival time variables.
      1. Assigns geovariables from both the 2020 and 2010 Census data, as in “Assigning Geovariables to Gas Stations.R”.

Run all steps in “Part 1 of Do File Order for EVs and Where to Charge Them”, then proceed as below:

1. Run “Identifying Gas Stations in UST Data.R”.
   1. This file uses the set of UST facility neighbors to OSM stations identified in the prior .do files to identify gas stations in the ‘mergedust’ dataset of all UST facilities in the U.S.
      1. It also uses the DOE charger data’s facility\_type variable to identify gas stations in the UST data (for chargers with facility\_type==”GAS\_STATION” or facility\_type==”CONVENIENCE\_STORE”).
      2. It also uses major station brand name (e.g., Wawa, Chevron, Cumberland Farms) to identify gas stations.
   2. This file also prunes gas stations that are located at supermarket chains (since their incentives likely differ from stations that only exist as fuel and convenience goods sales locations).
2. Run “Assigning Geovariables to Gas Stations.R”.
   1. This file assigns geovariables (e.g., MSA, county, census tract) to each observation in the (UST) gas station data.
      1. From both the 2020 and 2010 Census data.

Run all steps in “Part 2 of Do File Order for EVs and Where to Charge Them”, then manually identify what matches are actually gas stations (barring more detailed data on gas station characteristics from, say, Google Maps’ API, this is the best-available approach to accurately identifying chargers that are at actual gas stations (and not across the street for example) and/or identifying stations that actually have chargers); manually identifying the actual gas stations matched to chargers in the appropriate Do file is what produces the file “charger nearest actual station”. Then proceed with the following:

1. Run “Merging EPA UST and DOE Data V2.R”.
   1. This file merges the cleaned EPA UST data on gas stations (which have been identified using OSM data) with the cleaned DOE public charger data in order to *identify the filling stations in the DOE public charger data*.
   2. This file creates a single dataset of all gas stations w/charging and all gas stations w/o charging.
2. Run “Spatially Exploring Merged Station Charger Data.R”.
   1. This file produces point and choropleth plots of adapted/unadapted station locations.
3. Run “Cleaning ZIP-County Crosswalk Data.R”.
   1. This file cleans the quarterly ZIP-county crosswalk data and combines all data into one file.
   2. That is, this file produces a time series of ZIP-county relationships over time (as ZIP codes change over time (quarterly)).
4. Run “Cleaning EV Registration Data.R”.
   1. This file (currently) creates county-level annual totals of EVs in California; later on, it will do so for all states in the data.
      1. Note that zip code level totals are possible, but going with county level for now.
5. Run “Cleaning Socioeconomic Data.R”.
   1. This file cleans and combines socioeconomic data.
      1. Currently only using median household income from the ACS. This data is annual and spans 2010-2023.
6. Run “Cleaning Population Data.R”.
   1. This file cleans and combines population data from the ACS’ 5-Year estimates.
   2. This data is annual and spans 2010-2023.
7. Run “Constructing Competition Variables.R”.
   1. *This file creates the initial form of the analysis data ‘analysis’.*
   2. This file creates region (e.g., tract) level cumulative counts of competitors (other stations) and participants (firms besides stations that have chargers).
   3. This file also creates a panel data form for each station’s data (i.e., a separate observation for each month a station has been in existence).
   4. The working definition of a market at this stage is region-month (to account for dynamics), specifically tract-month as of 9/26/2022.
8. Run “Adding Income, Population, and EV Registrations to Analysis.R”.
   1. This file merges the cleaned EV registration, population, socioeconomic data into the analysis data created in “Constructing Competition Variables.R”.
   2. This file also assigns 0s to all stations w/NA for n\_evs variable. This is because after merging in the EV registration data, only stations in states with EV registration data available are kept in the analysis dataset. Thus, any remaining stations w/NA for n\_evs are not in counties/ZIPs that have EVs (at least in certain periods) and thus should have 0 for n\_evs.
9. Run “Assigning Retail Electric Service Utilities.R”.
   1. This file uses the HIFLD shapefile of retail electric service territories to assign each station (in the analysis dataset) its utility name(s) and utility ID(s).
10. Run “Assigning Max Demand Charges.R”.
    1. This file uses NREL data on demand charges at the utility-level to identify each utility’s maximum demand charge and assign it to stations based on their assigned utility in the previous step/file.
11. Run “Calculating Station Distance to Highways.R”.
    1. This file uses the U.S. Census Bureau’s TIGER shapefile to map stations and highways in the CONUS and calculate each station’s distance from a highway (specifically interstate and U.S. highways).
12. Run “Formatting Analysis Data for Survival.R”.
    1. This file creates new time and charger variables in the analysis dataset and reshapes the dataset to be used in duration/survival models.
       1. Both standard Cox proportional hazards models and discrete time proportional hazards models (e.g., complementary log-log models).
    2. This file produces two new analysis datasets: survanalysis\_all (single common adaptation risk start date) and survanalysis\_type (charger type- and level-varying adaptation risk start date; that is, each station’s start date for being at risk of adaptation varies with the type of charger they installed, if they’ve adapted. Non-adapted stations receive the Level 2 start date).
       1. Survanalysis\_all is the preferred dataset because regardless of the type or level of charging that stations chose to install, all stations experienced access to some form of EV charging when Level 2 charging became available in January 2010.
    3. In each dataset, this file also calculates a number of EVs per capita variable.