**Code directory**

Had I started working on the processing computer at the beginning, I’d probably have done everything there, but at the beginning of the project I was still working on Longleaf for things involving large files or long processes. Eventually I moved a lot of the things that had originally been done on Longleaf to processing computer, and this document shows the location of the final code version.

**Run on longleaf:**

od\_coordinate\_prep.R - generates NConly\_2018\_unique\_od\_filtered\_coordinates.csv

for use in routing – csv of all unique block group centroid – placekey location pairs in 2018 data.

**Run on personal computer:**

OSM.Rmd – helper script to semi-manually identify the OSM ids corresponding to locations of each expansion

osrm\_run\_large.jl – generates many .gpkgs (NC\_2018\_unique\_od\_filtered\_paths\_xxx.gpkg ) of routes for all OD pairs in NConly\_2018\_unique\_od\_filtered\_coordinates.csv

**Run on processing computer:**

ncdot\_data.Rmd – reading and processing NC DOT provided expansion data to spatial data, using OSM shapefiles

poi\_filtering.R – copied from longleaf but this version is now the master version. This filters the destinations in poi\_info to just the NAICS categories I’m interested in, and creates the poi\_subset\_placekeys\_naics\_borderlabel.csv

combine\_months.R – takes all of the od\_home monthly safegraph files (eg NC\_monthlypatterns\_od\_home\_2018-01.csv) and combines them into one file. Only includes placekeys in poi\_subset\_placekeys.csv (generated by poi\_filtering.R) and only includes origins from within NC. Combined file is od\_home\_combined\_filtered.csv

add\_distances\_odhomedata.R – adds distances between origins and destinations to od\_home\_combined\_filtered.csv and generates od\_home\_combined\_filtered\_dists.csv

label\_od\_habitual.R and label\_od.R (deprecated) – takes all\_expansions\_ncdot\_data.shp and joins with each of the many route files (NC\_2018\_unique\_od\_filtered\_paths\_xxx.gpkg) to label, at the destination level, whether a poi is “served” by a road by checking for overlap and then doing a nearest-neighbor relabel of the unlabeled or FALSE points. Habitual version uses only od pairs that appear in 6/12 months. Produces dest\_level\_labeling/placekey\_habitual\_labels\_nn.gpkg

alldata\_expansion.R – takes od\_home\_combined\_filtered\_dists.csv and removes near-border destinations (using poi\_subset\_placekeys\_naics\_borderlabel.csv), then combines with placekey\_habitual\_labels\_nn.gpkg to get segment presence labels. Then adds expanded\_ columns based on presence columns and expansion dates from all\_expansions\_ncdot\_data.shp. Saves result to alldata\_nozeros.csv

refined\_control.R – takes alldata\_nozeros.csv and restricts to only those destinations within the “influence area” of the segments – that is, restricts which points can be used as ‘controls’. For each road segment, finds buffer distance at which 95% of the “trues” are within the buffer area. Then restricts all data to just placekeys within that distance. This restricted dataset is called alldata\_nozeros\_refinedcontrol.csv and has 27,822,484 rows, from 6,331,118 distinct placekey-cbg pairs. Groups to destination level and finds avg dist, total count, and quantiles – this is written to alldata\_nozeros\_refinedcontrol\_destgrouped.csv.

aadt.R – find Annual Average Daily Traffic from AADT stations near each expanded roadway.

Synth.Rmd – the synthetic control analysis! Reads alldata\_nozeros\_refinedcontrol\_destgrouped.csv and adds contract info. Selects destinations present in different # of months of data. Runs microsynth on 12 treatments (segments) for two different sets of destinations (60-month and 57-month). Makes figures from all synthetic control runs.

figures.Rmd – additional figures not generated in other scripts.

**Partial list of deprecated code, not included:**

reshape\_labeled\_data.R and reshape\_labeled\_data\_func.R (deprecated) – FLAG: poorly named. Takes all non-border destinations from od\_home\_combined\_filtered.csv and tests filling in 0s in a sample of 1000, then adding expanded T/F columns based on placekey\_habitual\_labels\_nn.gpkg and dates from all\_expansions\_ncdot\_data.shp. Needs to be rewritten/scrapped because I have a more memory-efficient solution for the expansion-date labeling in agg\_model.R now

agg\_model.R – uses alldata\_nozeros.csv to fit a model where data are aggregated to destination level. Deprecated, used fixed effects model

disagg\_model.R – uses alldata\_wzeros\_refinedcontrol.sqlite3 and samples 10,000 destinations to fit a model where data are kept at the origin-destination pair level. Deprecated, used fixed effects model

model\_run\_test.R – holder for testing out model runs, will be split into scripts

figures.R – generally makes figures, in progress

expansion\_data\_prep.R – adds information about expansions (e.g. roadway class, construction time?, distance to border) for use in model

fill\_zeros.R – wacky order, but takes alldata\_nozeros\_refinedcontrol.csv, keeps just placekey, visitor\_home\_cbgs, date\_range\_start, and count columns, and fills in 0 counts for placekey-visitor\_home\_cbgs pairs that exist in the data but are missing for a particular date – one half of the data at a time. Then recreates the work in alldata\_expansion.R – adds presence labels and expansion labels – to each half, and writes to disk: halfdata \_wzeros.csv and otherhalfdata\_wzeros.csv. Then recreates work from add\_distances\_odhomedata.R by adding previously-calculated OD distances to each half-data set, and saves to disk again (overwriting halfdata \_wzeros.csv and otherhalfdata\_wzeros.csv).

Between fill\_zeros and future steps I used cat on the command line to combine the two halves (halfdata \_wzeros.csv and otherhalfdata\_wzeros.csv ) into one giant file - alldata\_wzeros\_refinedcontrol.csv

Fill\_zeros\_sample.R – turns out the massive dataset produced by fill\_zeros is way too big to analyze, so this does all the same steps as fill\_zeros but just to a sample of 9000 destinations (about a third of the refined-control set). Produces for\_analysis/sample9000\_wzeros\_refinedcontrol.csv

import.sql - turns giant giant csv (alldata\_wzeros\_refinedcontrol.csv) into sqlite reference to be able to load/query parts of the data in other scripts. Run from commandline:

*sqlite3 alldata\_wzeros\_refinedcontrol.sqlite3 < ../code/import.sql*

generates alldata\_wzeros\_refinedcontrol.sqlite3

synth\_bg.Rmd - abandoned attempt to do block-group level matching rather than destination-level