

Code Documentation

February 4, 2023

1 Constructing the Master Assessments

1. Start with the raw delimited current assessments file. Read this in chunks into R using `currentAssess_construct.R` from the `Construct_CoreLogic` module. and select handful of variables that we need (see the file for a list).
2. Do some cleaning in R. Any variables (i.e. `fireplaceindicator`) that are indicators but are stored as strings are coded as 0-1 dummies. Each property is matched to a 2020 definition block group if either parcel or block group coordinates are nonmissing from the main dataset. About 80 percent of the data have geocodes and are in the main sample geography (2020 block groups whose centroids lie in a 2013 definition MSA). Data are outputted in `.dta` format in considerably smaller chunks. This roughly corresponds to the fraction of US households in our sample geography.
3. Use Stata to take these chunks and append them, as well as assign encoded value labels, using `compress_append_currentAssessments.do`. This is `currentAssess_master.dta` in `DataV2/CoreLogic/Output`.

2 Constructing the Master Transactions

1. Start with the raw delimited transactions files. Read this in chunks into R using `transactions_construct.R` and select handful of variables that we need (see the file for a list). Indicators are cleaned to be 0-1 as well. Note that there may be transactions that do not correspond to our sample geography. Transactions outside of the sample geography will be dropped when matched to the assessment files above.
2. I select only transactions that are recorded at arms length. If the corresponding code is missing, I do not select it. Moreover, we only take transactions after 2015 to have a purely contemporary measure of prices. Transactions with missing dates are dropped.

3. Using the corresponding `compress_append*.do` file, we arrive at a relatively smaller dataset of transactions to be matched to the current assessments. Value labels are assigned in this do file, as well.

3 Constructing Zoning Districts

1. With the assessment data matched to block groups, we can start by constructing statistics to arrive at these zoning districts. Start with `collapse_stats_forClustering.do`. This file does a few things:
 - (a) Assigns one zoning code and a municipality (either CoreLogic or Geocoded municipalities) to a block group by looking at the property level zoning codes and municipalities and taking the mode of them within each block group. This mode ignores missing data. For block groups not lying within a particular definition of municipality, it is assigned its county as its municipality; zoning in unincorporated areas are generally handled at the county level.
 - i. Approximately x % of the block groups in the sample have an assigned zoning code.
 - (b) Constructs some basic statistics for clustering, such as the mean and mode of the lot size distributions by structure type, and the share of properties in residential units. There are many arbitrary choices over what we should cluster on here. Things that are calculated here may not show up in the actual set of clustering variables. The output is in `CoreLogic/output/blkgrpstats_forClustering.dta`.
2. Next, move on to the actual clustering algorithm, `cluster_ZoningDistricts.R` in `Construct_Regulation`. It takes the previous file as input. Below is a detailed account of the entire clustering algorithm:
 - (a)

4 Validating Zoning Districts

5 Constructing the Facts