Master Readme File for Replication Files for "The Microgeography of Housing Supply"
Nathaniel Baum-Snow and Lu Han, *Journal of Political Economy*

This document explains how to use this replication package.

The data construction was performed mostly in Stata, with one small component done in R. All of the data construction code runs in Stata 18 and also works in versions of Stata back to Stata 14. The analysis was all performed in Stata 18. This code is backward compatible to Stata 16.

The code was run on a Linux cluster with 32 cores, though most parts used at most 12 cores. The data construction can be fully replicated from start to finish with this archive except for the component that uses the Zillow micro data on transactions and assessments. All relevant code and source data, except Zillow micro data, is included in the replication package. Zillow data sets were not included to comply with our data use agreement with Zillow. Please see the Data Appendix of the paper for a complete explanation of the components of the Zillow data. The directory structure is fully maintained but with some empty directories and missing files.

The data construction code would take several weeks to run from end to end. The analysis portion of the replication code takes about 5 days to run end to end. We include all results that appear in tables and figures in the replication package. Most tabular results have been reformatted from the replication files for clarity of exposition.

The replication package incorporates a directory structure with some zip files and some native files. Dataverse automatically unzips uploaded zip files. To keep it from doing this, I have replaced the "zip" ending on all zip archives to "zipx". To use these, you must download the "zipx" files of interest, change the file extension to "zip" and unzip. If you only wish to look at the original code creating results in the paper, you can download only analyze_replication.zipx. If you wish to work with the main data set used to generate results, you should also download data_construct/combined.zipx.

Below is a description of all of the files in the replication archive. Because the largest upload size is 2.5GB, zip files are split across directories.

Master_Readme_Replication.pdf (this file)

analyze_replication.zipx (unzip to create the analyze_replication directory with included files and subdirectories for creating results in the paper)

data_construct/ (directory with all data construction)

- master_housing_supply.do (Runs all data construction code from start to finish. Look in this code to find what portion of the data construction takes place in each subdirectory.)
- combined.zipx (unzip to create combined directory with included files and subdirectories)
- crosswalks.zipx (unzip to create crosswalks directory with included files and subdirectories)
- market_access/
  - do.zipx (dofiles calculating market access variables)

data/
commutes/
gen/ (files left unzipped as many are quite large)
source/ (files left unzipped as many are quite large)
bartik.zipx
ctpp00-cleaned.zipx
ctpp90-cleaned.zipx
lodes_sources.zipx
market_access.zipx
unix_source_data.zipx

nlcdb01.zipx (unzip to create nlcdb01 directory with included files and subdirectories)
The 2001 and 2011 national land cover database image files in /sourcedata/landcover have been removed because they are too large for Dataverse at over 2.5GB each.

other_data.zipx (unzip to create other_data directory with included files and subdirectories)

zillow.zipx (unzip to create zillow directory with included files and subdirectories)

To fully replicate the data construction, one only needs to run
./data_construct/master_housing_supply.do
Though because of the lack of Zillow micro data in the replication package, Part 4 would not run properly.
To fully replicate the analysis and results, one only needs to run
./analyze_replication/master-hsupply.do

All output for tables and figures in the paper are included as part of the replication package.
These two master dofiles call all other dofiles and (one R script) needed to generate the data and results.