Replication Material for "Neighborhoods, Perceived Inequality, and Preferences for Redistribution: Evidence from Barcelona"

By Gerard Domènech-Arumí

December 2024

1. Accessing the Spanish Cadaster

The Spanish Cadastre cadaster data can be accessed from the following link: https://www.sedecatastro.gob.es/Accesos/SECAccDescargaDatos.aspx [last accessed December 5, 2024]

The user must have a Spanish Digital ID (e.g., Cl@ve) to download the data. Numerical data can be downloaded by clicking on "*Descarga de información alfanumérica por provincial (formato CAT)*" and shapefiles can be downloaded by clicking on "*Descarga de cartografia vectorial por provincial (formato Shapefile)*". The system will require a valid Spanish Digital ID to proceed.

To download the data for the Barcelona province, the user should select "BARCELONA" and choose "Urbana" as an option, and then click "*Descarga Fichero*" (see screenshot below) [December 5, 2024]. The process is the same for the numerical data and shapefiles.

Note: this will download the most updated version of the cadaster data. At the time of writing this document (December 2024), that was October 2024. Typically, the Cadastre is updated twice per year.

It is not possible download recover earlier versions of the cadastral data, which means that one cannot download the exact file I used for the paper (I downloaded the data in 2020). Also, at that time, it was not possible to download the data for an entire province. The data had to be downloaded one municipality at a time. Because of that, the do-file cleaning codes provided and discussed below should be adapted to the new data format in order for them to run.

2. Accessing the Catalan Tax Authority (ATC) Data

Unlike the Cadastre, the transactions data from the Catalan Tax Authority (ATC) is not publicly available. To request access, the user should get in touch directly with the agency. Their website is https://atc.gencat.cat/ca/inici [last accessed December 5, 2024]. They can be contacted at consultes@atc.cat [last contact February 2020].

Requests are approved on a case-by-case basis, at the discretion of the agency.

3. Replication: structure of the main directory

To successfully replicate all results, the user should specify their main directory in "master.do" and create a series of folders and subfolders in that location. These are:

prog: to store the do-files.

orig: to store the raw data. This directory should have nine subfolders, one for each original data source. These are: "**catastro**" (Spanish Cadaster), "**atc**" (transactions), "**survey**" (Survey data), "**aj_barcelona**" (Barcelona city council), "**mfomento**" (*Ministerio de Fomento*), "**tmb**" (Barcelona Transportation Authority), "**generalitat**" (Catalan Government), "**ine**" (Spanish National Statistical Office), "**elections**" (Elections data from the *Ministerio del Interior*), "**idescat**" (Catalan National Statistical Office).

data: to store generated datasets. This folder should contain two subfolders named "**int**" and "**temp**". "**int**" should have another subfolder: "**lng**".

tables: to store all the tables.

figs: to store all the figures. This folder should contain two subfolders named "**gph**" and "**maps**".

4. Replication: do-files and R scripts

Note:
1- The user should have approximately 50 GB of free storage to process the raw data.
2- Some programs need to be run in R. Stata can run R Scripts with the command "rscript," which can be installed by running the following line of code:

net install rscript, from("https://raw.githubusercontent.com/reifjulian/rscript/master") replace

3- The preamble of master.do defines the main directory (global "dir"). To run Stata do-files successfully, users must update that path to their home directory. This only needs to be done once in the master do-file.
4- R Scripts always start by defining the main directory in the preamble (variable "dir"). To run R Scripts successfully, users must update that path to their home directory, which has to be done in each script.

master.do: Replicates all the tables and figures. It runs all the programs described below.

I- Data Cleaning Programs:

Catastro [CONFIDENTIAL]
clean_catastro.do: clean catastro data. Files generated:
(1) **int/08_900.dta**: cleaned catastro data for Barcelona.
(2) cleaned data for other municipalities (to generate Appendix Table A5).
(3) **int/CITY_clus.dta**: These smaller files with the suffix "_clus" (i.e., 08_900_clus) contain the smallest subset of variables required to compute the Local Neighborhood Gini index in R Script "ineq_lng.R."

catastro_match_census.R: add census tract, neighborhood, and coordinates to Barcelona parcels. R Script. Files generated:
(1) **int/catastro_census_neigh_08_900.dta**: parcels matched to census units.

catastro_agglevel_des.do: generate new variables from the Barcelona catastro at different levels of aggregation. Files generated:
(1) **int/catastro_csec_chars.dta**: Catastro summary characteristics at census tract level 2009-19.
(2) **int/catastro_plot_chars.dta**: Catrastro summary characteristics at parcel level 2009-19.
(3) **int/catastro_blng_expan.dta**: Catastro summary characteristics of new buildings in years 1999-2019.

ATC [CONFIDENTIAL]:
atc_itp.R: clean ATC-ITP data. Files generated:
(1) **int/ATC-ITP_2009_19.dta**: cleaned ATC data with parcel identifier.

Other Data Sources: Ajuntament de Barcelona/AMB:
ajbcn_houseprice.do: Ajuntament de Barcelona/ AMB: house prices. Files generated:
(1) **int/ajbcn_houseprice_cdis.dta**: price per square meter of used dwellings at district level 2001-19.
(2) **int/ajbcn_houseprice_barri.dta**: price per square meter of used dwellings at neighborhood level 2013-19.

bcn_metro_schools_parks_hosp.do: Geolocate Barcelona Metro Stations, Schools, Parks, and Hospitals. Files generated:
(1) **int/metro_stations.dta**: geolocated metro stations.
(2) **int/bcn_schools.dta**: geolocated schools.
(3) **int/bcn_parks.dta**: geolocated parks.
(4) **int/bcn_hospitals.dta**: geolocated hospitals.

dist_metro_schools_parks_hosp_tract.R: Compute distance to metro stations, schools and parks. R Script. Note: slow code (10+ hours). Files generated:
(1) **int/distance_stations_tract.dta**: Distance between stations and centroid of all Barcelona's census tracts.
(2) **int/distance_schools_tract**: Distance between schools and centroid of all Barcelona's census tracts.
(3) **int/distance_parks_tract**: Distance between parks and centroid of all Barcelona's census tracts.

distance_trans.do: identify closest public transit stations. Files generated:
(1) **int/distance_trans.dta**: Distance to the closest public transit stations in each tract.

Other Data Sources: INE:
census_2011.do: Clean 2011 Census. Files generated:
(1) **int/census2011_csec.dta**: cleaned 2011 census data at tract level.
(2) **int/census2011_cdis.dta**: cleaned 2011 census data at district level.

ine_atlas_renta.do: INE opportunity atlas. Files generated:
(1) **int/ine_atlas_renta_csec.dta**: INE atlas renta at census tract level 2015-17.
(2) **int/ine_atlas_renta_cdis.dta**: INE atlas renta at district level 2015-17.
(3) **int/ine_atlas_renta_cmun.dta**: INE atlas renta at municipality level 2015-17.

ine_ecv.do: INE-ECV national income distribution. Files generated:
(1) **int/income_pctiles.dta**: Data on income distribution.

ine_registry_ctract_bcn.do: clean INE municipal registry data (demographic information: population by age groups, nationality, country of origin). Files generated:
(1) **int/reg_csec_YYYY.dta**: registry data in year YYYY at census tract level (YYYY=2010-2020).
(2) **int/reg_csec.dta**: pooled registry data (2010-2020) at census tract level.

Other Data Sources: Ministerio del Interior
elections_bcn.do: clean 2015/19 national election results for Barcelona at census tract level. Files generated:
(1) **int/elections_bcn.dta**: 2015/19 national election results at the census tract level.

Other Data Sources: Ministerio de Fomento
mfomento_rentals.do: clean rental price data at census tract level from the *Ministerio de Fomento* (2015-18). Files generated:
(1) **int/rentals_csec.dta**: rental price data at census tract level.

II- Predict Dwelling Prices and Measure Inequality:

prepare_predict_price_bcn.do: Merge cadaster, property registry, and other intermediate datasets into one single dataset to be used to run the random forest model. Files generated:
(1) **int/ATC_predict.dta**: cleaned ATC data to train random forest model.
(2) **temp/catdata_predict_YYYY.dta**: dwellings data from the cadaster with prices to impute (year YYYY, with YYYY=2009-2019).

predict_price_bcn.R: Estimate dwelling prices using a random forest. R Script. Files generated:
(1) **int/predict_price_ranger.RData**: Random forest prediction model.
(2) **int/PredPriceForest_08_900.dta**: Price predictions for Barcelona 2009-19.
(3) **int/PredPriceForest_ATC.dta**: Price predictions vs ATC data for Barcelona 2009-19.

ineq_lng.R: Compute LNG in BCN and other large Spanish cities. R Script. Note: extremely slow code (24+ hours). That code has two clearly delimited parts. The first part computes LNG estimates in a given city-year for a specific r (e.g., 100 meters). It does so for multiple cities and multiple r's. The second part appends all the intermediate datasets for a given city in one data frame. Due to the computing power and memory required to run this code, I used the Boston University Shared Computing Cluster (BU SCC) to run the first part of the code and my personal computer to run the second part. The code provided in this replication package is set to generate the value LNG for Barcelona. To compute the space LNG in that city, the user should uncomment line 58 in the R Script "TYPE <- SPACE". Files generated:
(1) **int/RelIneqWBuffYears_CITY.dta**: LNG for main Spanish cities in 2019.
(2) **int/RelIneqWValBuffYears_08_900.dta:** LNG for BCN 2015-19.

lng_data_combine.do: Combine intermediate datasets generated in R. Files generated:
(1) **lng_bcn.dta**: LNG Value/Space 2015-19 in BCN.

inequality_bcn.do: generate summary measures of inequality at diff aggregations in BCN. Files generated:
(1) **int/Ineq_08_900_XXX.dta**: Space/Value Inequality in Barcelona 2009-19 at municipality, district, neighborhood, Zip, and census tract level. XXX (=CMUN/CDIS/BARRI/ZIP/CSEC)
(2) **int/LNG_08_900_XXX.dta**: Space/Value LNG in Barcelona 2015-19 at municipality, district, neighborhood, and census tract level. XXX (=CMUN/CDIS/BARRI /CSEC)
(3) **int/IneqChng_08_900_PlotCode.dta**: Change in LNG at parcel level 2015-19.
(4) **int/IneqChng_08_900_CUSEC.dta**: Change in LNG at census tract level 2015-19.

III- Clean survey data:

survey_cleaning.do: Clean survey data. Files generated:
(1) **survey_netquest.dta**: cleaned survey data.

survey_matchaddress.R: match addresses from survey to catastro using fuzzy string matching. The program finds the best address match within a ZIP code in the Spanish Cadaster. R Script. Files generated:
(1) **int/survey_match.xlsx**: survey address matches with fuzzy string matching

[Not a program] Manually check and update results of fuzzy string matching from the previous code. Note: this step is not reproducible. Files generated;
(1) **int/survey_match_final.xlsx**: matched survey addresses to cadaster.

address_match_final.do: convert xlsx file to stata format. Files generated:
(1) **int/address_match.dta**: dataset with confirmed matched survey addresses to cadaster.

IV- Prepare dataset for the analysis:

neigh_shocks.R: Calculate distances between survey respondents and closest new buildings. R Script. Note: slow code (4+ hours). Files generated:
(1) **int/Survey_Dist_NewApt.dta**: Distance from respondents to closest new buildings.
(2) **int/CSEC_Dist_NewApt.dta**: Distance from centroids of census tracts to closest new buildings.

blng_shock_csec.do: stata data building shocks csec. Files generated:
(1) **analysis_dataset_aggcsec.dta:** analysis dataset agg csec (elections data).

gen_analysis_dataset.do: Combine all intermediate datasets. Generate main dataset for analysis.
(1) **analysis_dataset.dta**: Analysis dataset.

V- Analysis:

analysis.do: Generate all figures and tables, except maps. Note: the specification curve (only the figure) was generated using Stata/MP 14.0.

maps.R: Generate maps (Figures 1, 7, and A3). R Script.

5. Software and operating system used in the paper

Software: Stata/SE 18.0, and R version 4.3.1
Operating system: macOS Sonoma version 14.6.1.