Data Sources and Results for Housing Gotham: The 21st Century So Far (Part I)

Jason Barr, Rutgers University-Newark

Sean Franklin, Sia Partners

September 27, 2021

Additional Maps

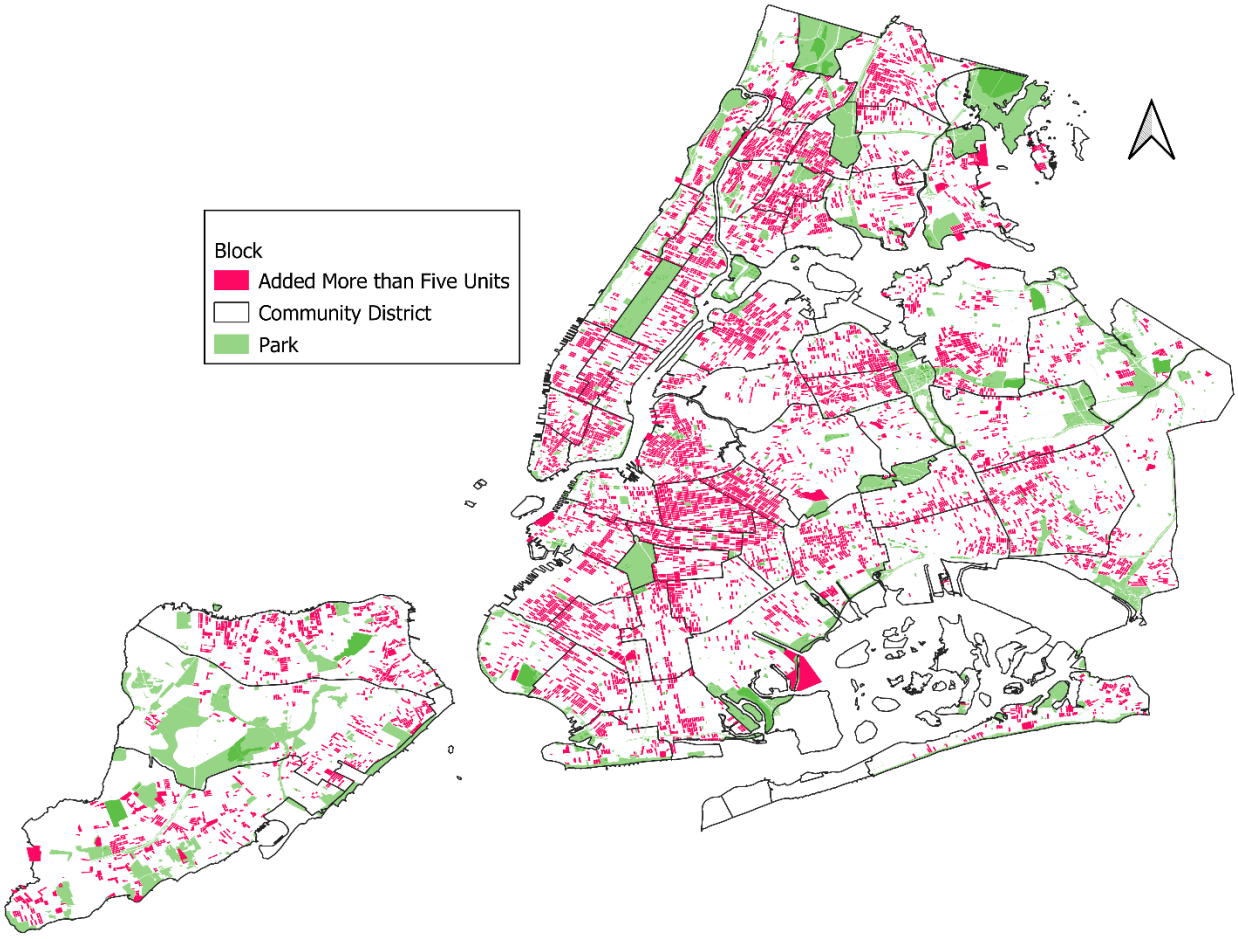

Figure 1: Blocks that added six or more units between 2002 and 2020. Source: NYC PLUTO Files.

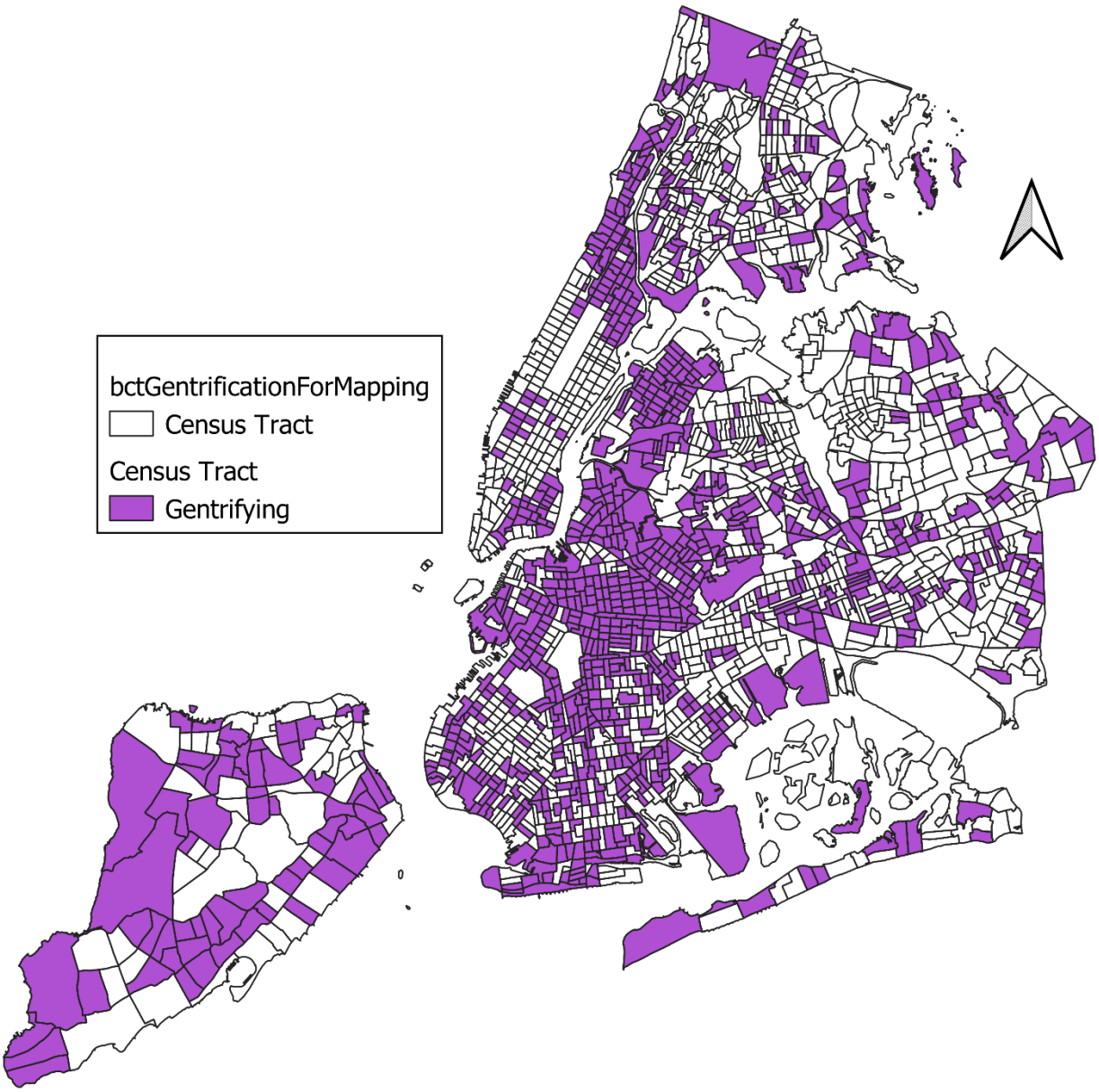

Figure 2: Gentrifying Neighborhoods. Purple census tracts were in bottom 50 percentile for average household income in 2002 but were in top 50 percentile for growth in college education shares across NYC from 2000 to 2019. Census tract data from US Census for 2000 and American Community Survey for 2019.

Data Sources & Preparation

Block Level Data

The main data sets are the PLUTO files from 2002 and 2020 v.2, which gives information for every tax lot in New York City. For each block we used the following variables from 2002: total residential units, the number of properties that are in a landmark district, number of buildings, the (unweighted) average maximum allowable FAR, and a dummy variable equal to one if there are no residential units on the block, 0 otherwise.

We also included a variable from 2017, which was the number of buildings with at least one rent stabilized unit. In the end, we did not discuss the variable in the blog. The coefficient was positive. We ran some instrumental variable regressions, which made the coefficient negative, but we decided to leave full exploration of this variable for future work (and its inclusion did not alter the other results). However, we briefly discuss the results of one IV regression below.

From the 2020 file, we obtained for each block the total number of residential units. We then created the change in the total units over the 18-year period. Furthermore, for the regressions, we converted several variables to logs. For the variables that had some zeros, we created *ln(1+variable)*.

The regressions also include building type fixed effects, and Community District or Borough fixed effects. For more information on the PLUTO variables see the PLUTO Data Dictionary.

Census Tract Data

To test the effect of gentrification, we created a second database at the Census Tract level using the 2010 census tract demarcations. Using the PLUTO files we created the same variables as above, but at the census tract level.

For demographic variables we used 2000 census data and we used 2019 ACS data (one year). We collected data for 2000 and 2019 population, the average household income in the census tract in the year 2000 or 2019, respectively. This was calculated by dividing aggregate 12-month household income by the total number of households. Last, we obtained the fraction of the census tract population that was classified as white-non-Hispanic.

For college share, we took the percentage of people over 25 in the census tract with a bachelor's degree or higher, in the year 2000 and 2019, respectively. This was calculated by dividing the number of people 25 or older with a bachelor's, master's, doctorate, or professional degree by the total number of people 25 or older.

Lastly, using a GIS subway map, we also calculated the number of subway stations in each census tract, as an additional control variable.

Note that for both the block-level and the census tract level data sets we omitted from the data set those block or census tracts that had residential unit changes in either the top one or bottom one

percentile. Some of these big changes might have been due to typos or mistakes in the PLUTO file. Additionally, if they were, in fact, accurate changes, then they also might represent large policy interventions which would not be representative of the housing market more broadly.

Descriptive Statistics & Graphs

Variable	Mean	Std. dev.	Min	Max
# Residential units, 2020	112.9	208.6	0	11302
# Residential units, 2002	103.4	202.6	0	11114
Change in res. units (2020-2002)	9.5	30.2	-47	286
# Properties landmarked, 2002	0.6	4.5	0	100
Avg. max. allowable FAR, 2002	1.5	1.6	0	15
# Buildings, 2002	29.5	22.5	1	525
No residential units dummy, 2002	0.1	0.3	0	1
# Rent stabilized buildings, 2017	1.3	3.4	0	68
Dist. to Empire State Bldg. (miles)	9.3	4.0	0.06659	22.0303

Figure 3: Block Level Descriptive Statistics. Note: Observations were removed from data set if change in residential units were in top or bottom one percentile, respectively.

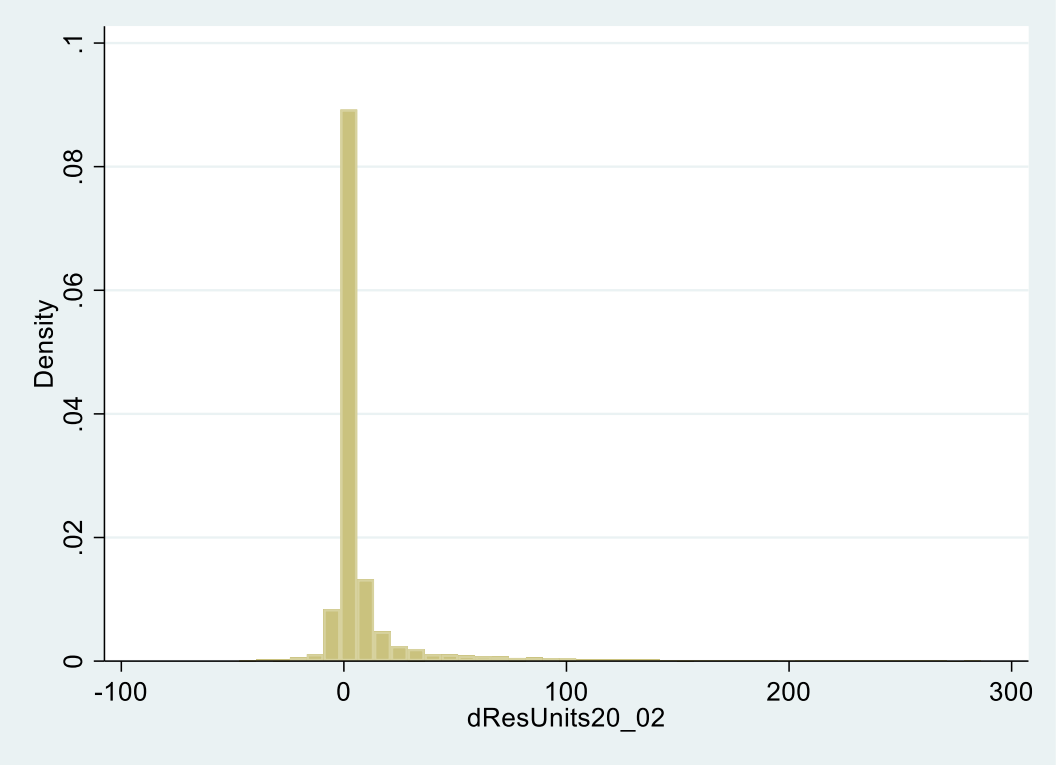

Figure 4: Histogram for Block Residential Units Changes from 2002 to 2020. Note: Observations were removed from data set if change in residential units were in top or bottom one percentile, respectively.

Variable	Mean	Std. dev.	Min	Max
# Res. units, 2020	1600.4	1145.8	0	11327
# Res. units, 2002	1445.1	1096.6	0	11133
Change in res. Units (2020-2002)	155.3	313.9	-504	2744
Gentrification Dummy	0.5	0.5	0	1
Avg. HH Income, 2000	53,401	30,144	0	487,375.6
# Subway stations, 2002	0.2	0.5	0	6
# Properties landmarked, 2002	8.3	40.8	0	436
% Residential buildings, 2002	91.5	15.3	0	100
Avg. total building area (sq. ft.)	2,247,665	2,351,969	0	3.13E+07
Avg. max. allowable FAR	2.0	1.8	0	14.14433

Figure 5: Census Tract Descriptive Statistics. Note: Observations were removed from data set if change in residential units were in top or bottom one percentile, respectively.

Table 1: Regression Results at the Block Level. Dep. Var.: *ln(1+ Res Units 2020)*.

Variable	(1)	(2)	(3)	(4)	(5)
ln(1+ Res. Units 2002)	0.964***	0.947***	0.926***	0.932***	0.944***
	(136.59)	(108.42)	(94.44)	(106.09)	(113.01)
# Landmarked Properties		-0.0023***	-0.0028***	-0.0039***	-0.00243*
		(3.01)	(3.66)	(2.69)	(1.88)
ln(Avg. Max Allowable FAR)		0.125***	0.142***	0.127***	0.122***
		(7.81)	(6.87)	(6.86)	(6.61)
# Landmarked Property x No Residential Units Dummy		0.00364**	0.00350**	0.00328**	0.00409**
		(2.37)	(2.23)	(2.01)	(2.57)
Num. Rent Stabilized Buildings		0.00156	0.000893	0.000171	0.00493***
		(1.17)	(0.74)	(0.14)	(3.57)
Dist. to the Empire State Building		-0.00925	-0.00328	0.00148	-0.0008
		(1.66)	(0.58)	(0.20)	(0.14)
ln(Longitude Coordinate)		1.639			
		(1.59)			
ln(Latitude Coordinate)		-0.147			
		(0.92)			
Bronx Dummy		0.128***	0.0963**		
		(2.89)	(2.64)		
Brooklyn Dummy		0.111**	0.0939**		
		(2.00)	(2.07)		
Queens Dummy		0.0600*	0.0815**		
		(1.71)	(2.39)		
Staten Island Dummy		0.236**	0.117**		
		(2.57)	(2.63)		
ln(1+# Buildings)			0.0549***	0.0616***	0.0719***
			(7.01)	(6.76)	(5.14)
# Single Family Houses					-0.0014***
					(3.13)
# Two Family Houses					-0.0014***
					(3.01)
# Walk Up Apartment Buildings					-0.0056***
					(5.30)
# Elevator Apartment Buildings					-0.0072***
					(3.03)
# Warehouse Buildings					-0.0166*
					(1.81)
# Factory Buildings					0.00797
					(1.14)
# Garages or Gas Stations					0.00912**
					(2.29)
# Hotel Buildings					0.022

					(0.48)
# Hospital or Health Facilities					0.0136
					(1.44)
# Theatre Buildings					0.0214
					(0.36)
# Retail Buildings					0.00239
					(1.13)
# Loft Buildings					0.0221**
					(2.42)
# Houses of Worship					0.0140***
					(2.65)
# Asylums and Homes					-0.0194
					(1.64)
# Office Buildings					0.00292
					(0.39)
# Public Assembly and Cultural Buildings					-0.0138
					(0.66)
# Outdoor Rec. Facilities					-0.00447
					(0.49)
# Condo Buildings					-0.0101
					(1.55)
# Multiple Use Residences					-0.0048***
					(3.45)
# Transportation Facilities					-0.0539**
					(2.29)
# Utility Facilities					-0.0158*
					(1.68)
# Vacant Lots					0.0131***
					(3.68)
# Educational Facilities					-0.013
					(1.53)
# Certain Govt. Facilities					-0.00935
					(0.47)
# Misc. Structures					-0.0053
					(0.67)
Constant	0.247***	-20.59	0.141**	0.145**	0.119**
	(6.34)	(1.54)	(2.26)	(2.02)	(2.17)
N	28254	27935	27935	27935	27935
R-sq	0.928	0.927	0.928	0.929	0.931
adj. R-sq	0.928	0.927	0.928	0.929	0.931
AIC	36030.8	34708	34560	33792	33298.5
BIC	36047.3	34815.1	34658.9	33849.6	33562.1
ND Fes	NO	CD	Boro	Boro	CD

Note: All RHS Vars as of 2002. T-statistics below estimates, clustered by community district. *** p<0.01, ** p<0.05, * p<0.1.

Table 2: Regression Results at the Block Level. Dep. Var.: $\Delta ln(1+ Res\ Units)$.

Variable	(1)	(2)	(3)	(4)	(5)
# Landmarked Properties		-0.00329***	-0.00296***	-0.00411**	-0.00183
		(3.42)	(3.28)	(2.53)	(1.31)
ln(Avg. Max Allowable FAR)		0.120***	0.106***	0.0884***	0.0951***
		(6.50)	(4.81)	(5.04)	(5.49)
# Landmarked Property x No Res. Units Dummy		0.00611***	0.00566***	0.00489***	0.00531***
		(4.21)	(3.88)	(3.05)	(3.32)
Num. Rent Stabilized Buildings		-0.00826***	-0.00680***	-0.00628***	0.00349***
		(5.11)	(4.68)	(4.13)	(2.71)
Dist. to the Empire State Building		-0.00989	-0.00494	0.000279	-0.00108
		(1.43)	(0.76)	(0.03)	(0.17)
ln(Longitude Coordinate)		1.663			
		(1.32)			
ln(Latitude Coordinate)		-0.134			
		(0.66)			
Bronx Dummy		0.116**	0.109**		
		(2.21)	(2.41)		
Brooklyn Dummy		0.105	0.125**		
		(1.54)	(2.23)		
Queens Dummy		0.0707	0.109**		
		(1.57)	(2.54)		
Staten Island Dummy		0.269**	0.168***		
		(2.31)	(3.06)		
ln(1+# Buildings)			-0.0248**	-0.0103*	0.0095
			(2.41)	(1.80)	(0.97)
# Single Family Houses					-0.00116***
					(3.17)
# Two Family Houses					-0.00161***
					(3.39)
# Walk Up Apartment Buildings					-0.00631***
					(5.74)
# Elevator Apartment Buildings					-0.0223***
					(4.71)
# Warehouse Buildings					-0.00909
					(0.94)
# Factory Buildings					0.0149*
					(1.94)
# Garages or Gas Stations					0.0119***

					(2.86)
# Hotel Buildings					0.0257
					(0.57)
# Hospital or Health Facilities					0.014
					(1.52)
# Theatre Buildings					0.0276
					(0.47)
# Retail Buildings					0.00458**
					(2.09)
# Loft Buildings					0.0248**
					(2.39)
# Houses of Worship					0.0156***
					(3.02)
# Asylums and Homes					-0.0231*
					(1.92)
# Office Buildings					0.0076
					(1.02)
# Public Assembly and Cultural Buildings					-0.00814
					(0.42)
# Outdoor Rec. Facilities					-0.00488
					(0.56)
# Condo Buildings					-0.0218***
					(3.11)
# Multiple Use Residences					-0.00532***
					(3.38)
# Transportation Facilities					-0.0299
					(1.25)
# Utility Facilities					-0.00249
					(0.29)
# Vacant Lots					0.0165***
					(4.90)
# Educational Facilities					-0.0056
					(0.70)
# Certain Govt. Facilities					-0.00194
					(0.10)
# Misc. Structures					-0.00194
					(0.25)
Constant	0.115***	-21.27	0.120*	0.145*	0.111**
	(7.84)	(1.30)	(1.70)	(1.71)	(2.07)
N	28254	27935	27935	27935	27935
R-sq	0	0.031	0.032	0.064	0.093
adj. R-sq	0	0.031	0.031	0.061	0.089
AIC	36544.5	35515.5	35490.4	34549.8	33714.4

BIC	36552.7	35614.4	35581	34599.3	33969.8
ND Fes	NO	CD	Boro	Boro	CD

Note. All RHS Vars as of 2002. T-statistics below estimates, clustered by community district. Note this set of regressions is the same as Table 1 but the Dep. Var is Change in Units from 2002 to 2020. *** p<0.01, ** p<0.05, * p<0.1.

Table 3: Regression Results at the Census Tract Level. Dep. Var.: (1+ Res Units 2002).

Variable	(1)	(2)	(3)	(4)	(5)
ln(1+Res Units 2002)	0.946***	0.937***	0.904***	0.832***	0.802***
	(37.49)	(31.24)	(27.68)	(16.72)	(14.20)
Gentrification Dummy		0.0481**	0.0556***	0.0611***	0.0359**
		(2.20)	(2.74)	(3.50)	(2.10)
avg inc (25-50th ptile) 2000			-0.0337	-0.0282	-0.0036
			(1.30)	(1.27)	(0.13)
avg inc (50-75th ptile) 2000			-0.089***	-0.0401	-0.00464
			(3.11)	(1.51)	(0.15)
avg inc (top 75th ptile) 2000			-0.0823**	-0.0508	-0.0296
			(2.02)	(1.50)	(0.65)
# Subway stations				0.0276	0.0226
				-1.05	-0.85
# Landmarked properties				0.00222	0.00246
				(1.42)	(1.55)
# Landmarked x % Residential properties				-2.5E-05	-0.0000304*
				(1.51)	(1.75)
ln(1+Total building area)				0.135**	0.167**
				(2.47)	(2.55)
ln(Avg. max allowable FAR)				0.121***	0.106***
				(4.44)	(3.93)
Constant	0.488***	0.529**	0.815***	-0.703	-0.945
	(2.67)	(2.59)	(3.51)	(-1.37)	(-1.51)
N	2117	2117	2089	2078	2078
R-sq	0.915	0.934	0.891	0.882	0.907
adj. R-sq	0.915	0.932	0.891	0.881	0.904
AIC	1888	1339.7	1694.1	1402.1	892.9
BIC	1899.3	1351	1728	1464.1	949.3
ND Fes	No	CD	Boro	Boro	CD

RHS building variable as of 2002. Income and Demographics using 2000 data (and 2019 to calculate gentrification dummy). *** p<0.01, ** p<0.05, * p<0.1.

Discussion

While we did not discuss the effects of rent stabilization for the housing stock in the blog post, we did test for its effects. The OLS Regression (Tables 1 & 2) yield a positive effect, which is contrary to expectations. More stabilized units would suggest lower building turnover and therefore a lower probability of teardowns (or combining units).

To this end we ran an IV regression, where the variable, # of rent stabilized units (as of 2017) was instrumented by some geographical variables, in particular the first stage regression included the following instruments, *DistESB_miles_2020, xcoodBlock2002, ycoodBlock2002,*

lnAvgLotsize, where DistESB is the distance of the center of the block to the Empire State Building, *xcoord* and *ycoord* are longitude and latitude values of blook centrod (measured in feet), and *lnAvgLotSize* is the log of the average lot size for the block.

Rent Stabilized IV Regression

Here is the coefficient results for the second stage regression, showing only the effect of rent stabilization on *# ln(1+units 2020)* and the rest of the coefficient estimates are omitted.

lnResUnits_2020	Coefficient	SE	z	P>z	[95% conf.	interval]
numRentStablized	-.0330449	.0094435	-3.50	0.000	-.0515539	-.014536

The Overid tests suggest reasonably valid instruments.

```
. estat overid

  Tests of overidentifying restrictions:

  Sargan (score) chi2(3) =  5.83184  (p = 0.1201)
  Basmann chi2(3)        =  5.81631  (p = 0.1209)
```

The tests for endogeneity suggest that the variable is in fact endogenous.

```
. estat endog

  Tests of endogeneity
  H0: Variables are exogenous

  Durbin (score) chi2(1)          =  14.5391  (p = 0.0001)
  Wu-Hausman F(1,27787)           =   14.506  (p = 0.0001)
```

First stage regressions suggest strong instruments

```
. estat first

  First-stage regression summary statistics
  ------------------------------------------------------------------------------
               |             Adjusted      Partial
      Variable |   R-sq.       R-sq.        R-sq.    F(4,27785)   Prob > F
  -------------+----------------------------------------------------------------
  numRentSta~d |  0.4507      0.4491       0.0123       86.6478    0.0000
  ------------------------------------------------------------------------------


  Minimum eigenvalue statistic = 86.6478

  Critical Values                       # of endogenous regressors:    1
  H0: Instruments are weak              # of excluded instruments:     4
  ---------------------------------------------------------------------
                                      |    5%     10%     20%     30%
  2SLS relative bias                  |  16.85   10.27    6.71    5.34
  ------------------------------------+--------------------------------
                                      |   10%     15%     20%     25%
  2SLS size of nominal 5% Wald test   |  24.58   13.96   10.26    8.31
  LIML size of nominal 5% Wald test   |   5.44    3.87    3.30    2.98
  ---------------------------------------------------------------------
```

Gentrification

Definitions

We thought about two issues related to gentrification. One was definition and the second was endogeneity. For definitions, we tried two versions. One, as described in the blog post, is that a neighborhood was classified as "gentrifying" if it was in the bottom 50% for average household income, and if its growth in the share of college education residents was in the top 50%. A second version was if the neighborhood was in the bottom quartile, but its college education share was in the top quartile. We used a version of definitions seen in recent literature.

We discussed the first definition in the blog post and found that it was positive and statistically significant. The second version was found to be positive but not statistically significant.

Here are the coefficient estimates for the second gentrification variable (rest of regression results omitted)

	(1) lnResUn~2020	(2) lnResUn~2020	(3) lnResUn~2020	(4) lnResUn~2020
gentrifcat~2	0.0825	0.0745	0.0926	0.0402
	(1.05)	(1.13)	(1.64)	(0.62)

Endogeneity

There is some concern that our definition of gentrification is endogenous. For this reason, we played around with instruments. In short, our findings: IV yield positive and larger coefficients, thought significance varies. Also, our instruments are somewhat weak. In the end, we decided to only discuss the OLS results. Here IV regressions (same as Table 1 above) (and some dummy variables have been omitted).

	(1) lnResUn~2020	(2) lnResUn~2020	(3) lnResUn~2020
gentrifcat~1	0.472	0.222	0.492
	(1.35)	(1.00)	(1.46)
lnResUn~2002	0.785***	0.819***	0.784***
	(15.00)	(12.13)	(15.56)
numHDPr~2002	0.00376	0.00339	0.00382
	(1.50)	(1.62)	(1.54)
numHD_ResPct	-0.0000402	-0.0000381*	-0.0000406
	(-1.60)	(-1.76)	(-1.63)
lntotal~2002	0.199***	0.111*	0.200***
	(3.42)	(1.82)	(3.71)
lnAvgMaxFar	0.100***	0.106***	0.0993***
	(2.87)	(3.79)	(2.74)
_cons	-2.449***	2.210***	-2.468***
	(-3.26)	(2.96)	(-3.55)

N	2078	2060	2078
R-sq	0.871	0.896	0.868
adj. R-sq	0.866	0.892	0.863
CD Fes	Yes	Yes	Yes
IVTEST (p-val.)	0.0790	0.0186	4.16492
ENDOG	0.0256	0.1738	0.2307
FIRST (Min Eig. Stat)	3.09994	3.19104	0.0068

t statistics in parentheses (clustered by CD) * p<0.10, ** p<0.05, *** p<0.01. Sargan OVERID test. Durban endog test.

Instruments for regression (1): (gentrifcation1 = xcoodBlockCT2020 ycoodBlockCT2020 lnDistESB numstation)

Instruments for regression (2): (gentrifcation1 = xcoodBlockCT2020 ycoodBlockCT2020 white2000_pct numstation)

Instruments for regression (3): (gentrifcation1 = xcoodBlockCT2020 ycoodBlockCT2020 AvgFloorsCT2002 numstation)