Data Sources and Statistical Results for "Is Decay Contagious? Building Code Violations across New York City"

Jason M. Barr
August 15, 2025

Data Sources

Class C Violations: New York posts the complete list of building violations over the last decade, downloadable from https://data.cityofnewyork.us/Housing-Development/Housing-Maintenance-Code-Violations/wvxf-dwi5/about_data. For each 2020 census tract, I added up the total number of violations in 2017-18 and from 2023-24.

Census Tracts: Shapefile for NYC downloadable here: https://www.nyc.gov/content/planning/pages/resources/datasets/census-tracts

Demographics and Income Variables: For each Census Tract, the data are from the American Community Survey, five-year averages ending in 2023.

- % with Bachelor's Degree or higher: ACSST5Y2023.S1501
- % Below the Poverty Line: ACSST5Y2023.S1701
- Median Family Income: ACSST5Y2023.S1903
- Race and Ethnicity: ACSDP5Y2023.DP05
- Population: ACSST5Y2023.S1701

Building Characteristics and Land Use: Data are from the 2023 PLUTO file, downloadable from https://www.nyc.gov/content/planning/pages/resources/datasets/mappluto-pluto-change.

All variables (except borough dummies and distance to the Empire State Building) are calculated at the census tract level. These variables are: Total residential lot area, number of residential units, and average maximum allowable Floor Area Ratio (FAR). Distance to the Empire State Building is as the crow flies from the centroid of each census tract (in degrees).

Descriptive Statistics

Variable	Obs	Mean	Std. dev.	Min	Max
# Class C Viols, 23-24	2,159	244.6369	372.5788	0	3445
# Class C Viols, 17-18	2,159	90.65262	143.5209	0	1416
BA or +	2,159	38.89402	21.38058	2.9	96.1
Med. Fam. Inc.	2,159	106,177	55,838	18,299	250,000
Population	2,159	3857.925	1899.649	200	16684
Black(%)	2,159	23.25831	26.99946	0	95.6
White(%)	2,159	30.97957	27.77831	0	99.9
Hispanic(%)	2,159	26.88374	21.83048	0	94.6
Asian(%)	2,159	15.35804	17.404	0	91.2
# Res. Units	2,159	1651.252	1203.848	0	20976
Total Lot Area	2,159	2355711	4212569	204745	8.41e+07

MN	2,159	.1315424	.3380709	0	1
BX	2,159	.1523854	.3594773	0	1
BK	2,159	.3483094	.476545	0	1
QN	2,159	.3131079	.4638653	0	1
SI	2,159	.0546549	.2273581	0	1
Dist. to ESB (deg.)	2,159	.1302244	.0614295	.0028015	.3529803
Avg Allow Max FAR	2,159	1.984724	1.752981	0	10

Regression Results

Table 1: OLS Regressions. Dep. Var.: ln(1+Violations, 2023-24)

	(1)	(2)	(3)
lnViols17_18	0.920***	0.851***	0.802***
	(30.19)	(26.38)	(16.80)
baplus		0.00936	0.00517
		(1.64)	(1.41)
lnMedianInc_fam		-0.376	-0.220
		(-1.78)	(-1.60)
lnPop		0.241	0.0874
		(2.12)	(0.89)
white_pct		-0.00763*	-0.00206
		(-2.26)	(-0.62)
black_pct		-0.00814**	-0.00303
		(-3.47)	(-2.05)
asian_pct		-0.00462	-0.000656
		(-1.56)	(-0.23)
medIncMwx		-0.168	-0.285**
		(-1.89)	(-3.36)
lnResUnits			0.353***
			(6.38)
lnTotalLotArea			-0.191*
			(-2.57)
boro2			0.904***
			(9.40)
boro3			0.416**
			(3.06)
boro4			0.665**
			(4.43)
boro5			0.392
			(1.55)
distESB			-0.577
			(-0.63)
avg_Allow_ResFAR			0.0337
			(1.25)
_cons	1.238***	3.963	3.092
	(14.16)	(1.82)	(2.07)
N	2159	2159	2159
R-sq	0.751	0.769	0.792
adj. R-sq	0.751	0.768	0.790
AIC	5555.3	5397.9	5175.6
BIC	5566.6	5420.6	5198.4

t statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Not variables with zeros are logged using ln(1+ var. Standard errors clustered at the borough level.

Table 2: OLS Regressions. Dep. Var.: ln(1+Violations, 2023-24), additional specifications

	(1)	(2)	(3)	(4)
lnViols17_18	0.824***	0.804***	0.823***	0.807***
	(23.63)	(19.86)	(25.48)	(20.71)
baplus	0.00831**	0.00543	0.00405**	0.00413
	(3.96)	(1.89)	(3.12)	(2.03)
lnMedianInc_fam	-0.331**	-0.226	-0.343*	-0.227
	(-3.59)	(-1.75)	(-2.73)	(-1.59)
poverty_pct	0.000275		-0.00263*	
	(0.12)		(-2.33)	
lnPop	0.385*	0.0904	0.0394	0.0886
	(2.65)	(0.90)	(0.49)	(0.86)
lnTotalLotArea	-0.145	-0.208*		-0.195*
	(-1.88)	(-2.52)		(-2.48)
boro2	0.905***	0.863***	0.984***	0.921***
	(7.34)	(9.49)	(6.29)	(7.77)
boro3	0.315**	0.336**	0.454*	0.401**
	(3.01)	(3.58)	(2.73)	(3.16)
boro4	0.654**	0.587***	0.729**	0.686**
	(4.16)	(5.32)	(3.45)	(4.21)
boro5	0.365	0.332	0.320	0.381
	(1.54)	(1.43)	(1.31)	(1.48)
distESB	-1.764	-0.798	-1.653	-0.662
	(-1.90)	(-0.86)	(-1.44)	(-0.64)
lnResUnits		0.366***	0.296***	0.348***
		(5.79)	(8.83)	(6.29)
white_pct		-0.00178		
		(-0.97)		
black_pct		-0.00261***		-0.00191
		(-7.44)		(-1.89)
medIncMwx		-0.260**		-0.285**
		(-3.80)		(-3.52)
asian_pct			0.00213	
			(1.34)	
avg_Allow_ResFAR			0.0518	0.0344
			(1.46)	(1.26)
_cons	3.731**	3.387*	2.461	3.185
	(4.59)	(2.55)	(2.12)	(2.05)
N	2159	2159	2159	2159
R-sq	0.782	0.791	0.787	0.791
adj. R-sq	0.781	0.790	0.786	0.790
AIC	5271.7	5179.3	5221.5	5177.9
BIC	5294.4	5202.0	5244.2	5200.6

t statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Note: Since median family income by CT is capped at $250,000, a dummy variable was included for those CTs at the max as an additional control variable. Standard errors clustered at the borough level.

A Note on Estimation

I include a spatial error term in the spatial autoregression (SAR) and estimate a spatial error coefficient (λ). According to Esri, "[SAR] is similar to the ordinary least squares regression formula, in which a dependent variable (y) is predicted by a set of explanatory variables (x) and coefficients (β). However, the residual term (u) is modeled by a different regression equation. This second regression predicts the residual using a spatial autoregressive parameter λ (lambda) and a spatial weights matrix (W), along with its own residual term (ε).

The lambda parameter quantifies the strength of the spatial dependence in the error term and measures how much one location's error term influences the error terms of its neighbors. The SEM works by filtering out spatial autocorrelation from each of the variables in the model and performing a regression on the spatially filtered variables. As a result, the coefficient estimates are not as affected by the spatial autocorrelation in each variable." Note, however, that to be conservative, I don't include the estimate of λ in the discussion of the "contagion" effect. But given the positive estimate, it suggests that "shocks" to surrounding communities will continue to positively impact the number of Class C violations over time in each CT.

Table 3: SAR Regressions. Dep. Var.:ln(1+Violations, 2023-24)

	(1)	(2)	(3)	(4)	(5)
lnViols17_18	0.777***	0.758***	0.756***	0.752***	0.747***
	(49.05)	(31.68)	(31.16)	(31.85)	(31.24)
lnPop	0.0949*	0.0929	0.100*	0.0863	0.0683
	(1.80)	(1.64)	(1.79)	(1.60)	(1.19)
white_one_pct	0.000581	-0.000194	-0.000619		
	(0.74)	(-0.17)	(-0.51)		
lnResUnits	0.396***	0.392***	0.378***	0.372***	0.411***
	(10.00)	(8.21)	(8.11)	(8.41)	(8.57)
boro2	0.823***	0.812***	0.891***	0.836***	0.699***
	(11.34)	(9.69)	(9.33)	(8.71)	(8.12)
boro3	0.313***	0.294***	0.339***	0.277***	0.209***
	(5.13)	(3.87)	(4.27)	(3.62)	(2.64)
boro4	0.599***	0.544***	0.615***	0.566***	0.443***
	(8.84)	(6.06)	(6.63)	(6.09)	(4.74)
boro5	0.313***	0.260*	0.369**	0.309**	0.143
	(2.95)	(1.76)	(2.31)	(2.00)	(0.94)
lnTotalLotArea	-0.241***	-0.234***	-0.211***	-0.197***	-0.219***
	(-7.93)	(-6.03)	(-5.13)	(-5.00)	(-5.83)
avg_yearbult		-0.000157	-0.000126		
		(-0.92)	(-0.73)		
distESB			-1.003	-0.977	
			(-1.47)	(-1.46)	
baplus				0.00356*	
				(1.91)	
lnMedianInc_fam				-0.182***	-0.0910
				(-2.58)	(-1.38)

poverty_pct				-0.000216	-0.000350
				(-0.09)	(-0.14)
medIncMwx				-0.242***	-0.192**
				(-2.84)	(-2.29)
_cons	0.786*	1.149*	0.928	2.649***	1.909**
	(1.79)	(1.84)	(1.44)	(2.78)	(2.00)
W					
lnViols23_24	0.0777***	0.0766***	0.0718***	0.0731***	0.0701***
	(5.60)	(5.69)	(5.33)	(5.51)	(4.15)
e.lnViols23_24		0.344***	0.350***	0.326***	0.346***
		(8.36)	(8.53)	(7.84)	(8.26)
voils_17_18_ct					0.000219
					(0.83)
N	2159	2159	2159	2159	2159

t statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01These regressions include spatial right-hand side variables. The weight matrix here is of the "contiguity" variety, where a weight matrix entry is one if it borders a CT, zero otherwise. The weight matrix is normalized so all rows sum to 1. Regression (1) has a spatial AR variable, (2) includes a spatial error term, as do (3) and (4), but with different control variables. Equation (5) also includes a "spatial lag" of building violations in 2017-18 (but are not significant. Standard errors are robust, and the estimation method was Gs2sls.

Table 4: SAR Regressions. Dep. Var.: ln(1+Violations, 2023-24)

	(1)	(2)	(3)	(4)	(5)
lnViols17_18	0.795***	0.782***	0.780***	0.776***	0.772***
	(51.73)	(33.32)	(32.84)	(33.65)	(33.18)
lnPop	0.103*	0.115**	0.124**	0.110**	0.0875
	(1.93)	(2.03)	(2.20)	(2.07)	(1.51)
white_pct	0.000607	-0.000242	-0.000680		
	(0.76)	(-0.22)	(-0.58)		
lnResUnits	0.399***	0.387***	0.372***	0.360***	0.406***
	(9.99)	(7.87)	(7.72)	(8.09)	(8.19)
boro2	0.810***	0.790***	0.879***	0.842***	0.679***
	(10.84)	(9.33)	(9.07)	(8.69)	(7.89)
boro3	0.293***	0.255***	0.305***	0.255***	0.154**
	(4.68)	(3.35)	(3.83)	(3.31)	(1.97)
boro4	0.579***	0.543***	0.623***	0.587***	0.417***
	(8.42)	(6.10)	(6.70)	(6.36)	(4.60)
boro5	0.275**	0.223	0.344**	0.303*	0.0921
	(2.56)	(1.49)	(2.13)	(1.94)	(0.61)
lnTotalLotArea	-0.243***	-0.237***	-0.214***	-0.198***	-0.221***
	(-7.90)	(-6.23)	(-5.33)	(-5.12)	(-5.95)
avg_yearbult		-0.000198	-0.000167		
		(-1.21)	(-1.02)		
distESB			-1.073	-0.918	
			(-1.58)	(-1.38)	
baplus				0.00520***	
				(2.78)	

lnMedianInc_fam				-0.225***	-0.103
				(-3.19)	(-1.59)
poverty_pct				-0.000275	-0.000718
				(-0.11)	(-0.29)
medIncMwx				-0.238***	-0.186**
				(-2.75)	(-2.19)
_cons	0.800*	1.164*	0.945	2.981***	1.976**
	(1.78)	(1.91)	(1.51)	(3.17)	(2.12)
W2					
lnViols23_24	0.0529***	0.0559***	0.0502***	0.0557***	0.0609***
	(3.73)	(3.70)	(3.31)	(3.74)	(3.07)
e.lnViols23_24		0.438***	0.445***	0.400***	0.420***
		(8.76)	(8.98)	(7.88)	(8.20)
viols_17_18_ct					-0.0000985
					(-0.27)
N	2159	2159	2159	2159	2159

t statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01. These regressions include spatial right-hand side variables. The weight matrix here is of the "contiguity" variety, where a weight matrix entry is one if it borders a CT or if it borders a CT that borders a CT (i.e., 2-degree contiguity), zero otherwise. The weight matrix is normalized so that all rows sum to 1. Regression (1) has a spatial AR variable, (2) includes a spatial error term, as do (3) and (4), but with different control variables. Equation (5) also includes a "spatial lag" of building violations in 2017-18 (but they are not significant). Standard errors are robust, and the estimation method was Gs2sls.

Table 5: SAR Regressions. Dep. Var.: ln(1+Violations, 2023-24)

	(1)	(2)
lnViols17_18	0.748***	0.750***
	(30.20)	(31.21)
baplus	0.00319*	
	(1.67)	
lnMedianInc_fam	-0.187***	-0.112*
	(-2.67)	(-1.69)
poverty_pct	-0.0000149	-0.000329
	(-0.01)	(-0.13)
lnPop	0.0844	0.0614
	(1.60)	(1.11)
avg_Floors	-0.00947	
	(-0.71)	
lnResUnits	0.349***	0.378***
	(7.93)	(8.25)
boro2	0.828***	0.689***
	(7.38)	(7.08)
boro3	0.237**	0.175*
	(2.41)	(1.92)
boro4	0.511***	0.371***
	(4.49)	(3.68)
boro5	0.256*	0.0813
	(1.66)	(0.59)

distESB	-1.416**	
	(-2.07)	
lnTotalLotArea	-0.0972**	-0.117***
	(-2.14)	(-2.58)
medIncMwx	-0.188**	-0.151*
	(-2.17)	(-1.78)
_cons	1.626	1.009
	(1.57)	(0.97)
Widist1		
lnViols23_24	0.117***	0.124***
	(4.62)	(3.73)
e.lnViols23_24	0.681***	0.696***
	(9.85)	(10.36)
viols_17_18_ct		0.0000796
		(0.17)
N	2159	2159

t statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01. These regressions include spatial right-hand side variables. The weight matrix here is of the "contiguity" variety, where a weight matrix entry is the weighted distance from the centroid (w=1/dist) if within a half mile of the CT, zero otherwise. The weight matrix is normalized so that all rows sum to 1. Regression (1) has a spatial AR variable and a spatial error term, (2) includes a "spatial lag" of building violations in 2017-18 (but they are not significant). Standard errors are robust, and the estimation method was Gs2sls.