

For Online Publication

A Appendix Tables

Table A.1: Sample breakdown over time, by state

State	Census Region	Blocks, Time Avg
California	West	38029
Florida	South	31389
Texas	South	20969
Pennsylvania	Northeast	19504
New York	Northeast	16167
New Jersey	Northeast	15718
Illinois	Midwest	14822
Massachusetts	Northeast	14269
Ohio	Midwest	11604
Virginia	South	10885
Washington	West	8934
Michigan	Midwest	8254
Missouri	Midwest	8237
Maryland	South	7500
Arizona	West	7318
Connecticut	Northeast	6755
Tennessee	South	5780
Georgia	South	5232
Minnesota	Midwest	4769
Indiana	Midwest	4731
North Carolina	South	4614
Oklahoma	South	3595
Colorado	West	3131
Utah	West	2951
Rhode Island	Northeast	2875
South Carolina	South	2830
Alabama	South	2793
Kansas	Midwest	2595
Wisconsin	Midwest	2182
Oregon	West	1897
District of Columbia	South	1300
Nevada	West	1159
Delaware	South	1152
Maine	Northeast	1023
Mississippi	South	939
Nebraska	Midwest	805
Idaho	West	801
New Mexico	West	773
Iowa	Midwest	711
New Hampshire	Northeast	705
Hawaii	West	450
Kentucky	South	377

Table A.2: Summary Statistics for Overlap for Historic Boundaries

Statistic		Range of MLS in subsample				
		Total	1–6000	6000–11000	11000–22000	22000+
HOLC Class A (sq. ft.)	Mean	3.0	3.1	3.5	2.0	0.1
	St. Dev.	16.9	17.3	18.3	14.0	3.8
	N	54,945	18,793	24,063	9,974	2,115
HOLC B–C	Mean	5.4	9.5	4.4	1.1	0.2
	St. Dev.	22.6	29.3	20.6	10.2	4.9
	N	54,945	18,793	24,063	9,974	2,115
HOLC Class D	Mean	3.7	7.5	2.4	0.4	0.1
	St. Dev.	18.8	26.4	15.2	6.4	3.1
	N	54,945	18,793	24,063	9,974	2,115
Rail tracks	Mean	7.1	10.3	6.0	5.0	2.9
	St. Dev.	25.6	30.4	23.7	21.8	16.7
	N	77,960	24,757	35,312	14,794	3,097
Elem. school catchment area	Mean	27.1	33.9	27.0	19.0	13.3
	St. Dev.	44.5	47.3	44.4	39.2	34.0
	N	78,178	24,825	35,387	14,850	3,116
Middle school catchment area	Mean	9.9	11.4	10.0	8.3	4.2
	St. Dev.	29.8	31.8	30.0	27.5	20.1
	N	78,178	24,825	35,387	14,850	3,116
High school catchment area	Mean	7.5	8.8	7.9	5.2	4.0
	St. Dev.	26.3	28.3	27.0	22.1	19.6
	N	78,178	24,825	35,387	14,850	3,116

Notes: This summary table plots statistics on the rate of overlap within a 350 meter wide buffer, between historical boundaries within cities that could affect neighborhood change. Calculation of overlap closely follows the procedure used to construct 2, but applied over the full national sample used in analysis. All data are expressed in percent units of the overall sample.

Sources: Calculations from CoreLogic records; the MAPC Zoning Atlas (); digitized HOLC “redlining maps” (); national roads and railroads data from the Census Bureau; and the SABINS sample of school catchment areas ().

Table A.3: Descriptive statistics for areas bound by MLS impact clusters

		Impact Cluster Name			
		High-impact	Moderate-impact	Low-impact	Large lot
Pct of MLS < 6K sq. ft.	Share	0.8%	6.7%	36.8%	0.8%
Pct of MLS ≥ 1/2 acre	Share	2.2%	2.7%	3.7%	91.9%
MLS in incorporated places	Share	64%	73.6%	84%	33.4%
Density differences with comparison areas (sqft)	Median IQR	7826.5 [5900, 10900]	3350 [2400, 4920.6]	1095 [720, 1749]	17864 [7470, 34216.8]
Treated area distance from CBD (km)	Median IQR	22.4 [13.2, 37.2]	22.7 [11.9, 37.5]	19.2 [9.6, 32.5]	32.6 [20.7, 46.5]
2020 population of adopting jurisdiction	Median IQR	58455.5 [23183, 196100]	63633 [24968, 200733]	126090 [44253, 440646]	20692 [10421, 41629.2]
Share of natl sample		14%	41%	43%	2%

↳ *Notes:* Based on the impact clusters identified in Section 4.2, this Table presents selected summary statistics for blocks treated by minimum lot size border segments in each cluster. The first three rows list the share of blocks next to border segments that are both in a specific impact cluster and follow a certain characteristic. The third row is defined for MLS requirements in the 37 states where unincorporated county land can be zoned. The following rows present the median and interquartile ranges (IQR) of three different variables. Density differences are defined over non-multifamily units in the comparison blocks, net of the MLS requirement.

Sources: Calculations from 1980–2020 NHGIS Tables (Manson et al. (2024)) and CoreLogic Tax Records.

Table A.4: Outcomes and Covariates Around Lot Size Borders

Statistic		Sample moments	
		Treated	Comparison
Median year built	Mean	1,966	1,961
	St. Dev.	21	27
	N	8,705	26,785
Distance from CBD (km)	Mean	26.9	25.5
	St. Dev.	19.0	19.6
	N	10,319	38,583
Population of area, 2010	Mean	159.2	381.3
	St. Dev.	222.9	512.7
	N	3,884	7,474
% Black in block, 1980	Mean	9.7	11.2
	St. Dev.	25.4	26.9
	N	3,275	16,657
% minority in block, 1980	Mean	16.8	18.9
	St. Dev.	29.3	30.6
	N	3,275	16,657
% rental units, 1980	Mean	23.1	33.8
	St. Dev.	28.0	29.9
	N	2,951	15,185
Impact Segment		High-impact (14% of sample)	

Notes: This summary table plots statistics on the predetermined character of residential development for two types of blocks. Blocks determined to be in a minimum lot size district following the detection procedure in Section 2 is in the “Treated group.” Blocks in adjacent areas developed at a specified elevated density compared to the treated areas is in the “Comparison group.” Results are plotted for two urban context-specific samples, each representing a subset of all blocks in the analysis sample. The definitions of those context-specific samples are given in Section 4.2. The level of observation is Census block based on 2010 boundaries. The population variable is observed at the level of treatment area, which is the union of all blocks identified to be surrounding a regulatory boundary segment and where development is restricted by the lot size regulation.

Sources: Calculations from 1980–2020 NHGIS Tables (Manson et al. (2024)) and CoreLogic Tax Records.

Table A.5: Outcomes and Covariates Around Lot Size Borders

Statistic		Sample moments		Sample moments	
		Treated	Comparison	Treated	Comparison
Median year built	Mean	1,961	1,959	1,956	1,956
	St. Dev.	21	24	23	26
	N	31,405	86,311	36,392	74,434
Distance from CBD (km)	Mean	27.0	25.6	23.8	23.1
	St. Dev.	19.6	19.8	19.0	18.5
	N	36,556	117,390	44,804	104,474
Population of area, 2010	Mean	168.1	358.9	255.1	389.3
	St. Dev.	244.0	518.8	369.2	812.4
	N	12,719	22,502	11,956	20,350
% Black in block, 1980	Mean	8.3	10.7	13.2	17.1
	St. Dev.	23.2	26.7	29.1	33.0
	N	13,888	57,035	20,443	54,199
% minority in block, 1980	Mean	15.8	18.6	24.7	29.0
	St. Dev.	27.4	30.3	33.5	36.2
	N	13,888	57,035	20,443	54,199
% rental units, 1980	Mean	22.8	29.1	28.5	35.3
	St. Dev.	27.8	29.5	29.2	30.3
	N	12,795	53,103	19,352	50,980
Impact Segment		Moderate-impact (41% of sample)		Low-impact (43% of sample)	

Notes: This summary table plots statistics on the predetermined character of residential development for two types of blocks. Blocks determined to be in a minimum lot size district following the detection procedure in Section 2 is in the “Treated group.” Blocks in adjacent areas developed at a specified elevated density compared to the treated areas is in the “Comparison group.” Results are plotted for two urban context-specific samples, each representing a subset of all blocks in the analysis sample. The definitions of those context-specific samples are given in Section 4.2. The level of observation is Census block based on 2010 boundaries. The population variable is observed at the level of treatment area, which is the union of all blocks identified to be surrounding a regulatory boundary segment and where development is restricted by the lot size regulation.

Sources: Calculations from 1980–2020 NHGIS Tables (Manson et al. (2024)) and CoreLogic Tax Records.

Table A.6: Pooled effects of MLS requirements, Causal Forest model

BD Estimates	Block Nonwhite Shares		Block Black Shares	
	(1)	(2)	(1)	(2)
2020 Data	-0.003 (0.002)	-0.007*** (0.002)	-0.011*** (0.002)	-0.011*** (0.001)
2010 Data	-0.019*** (0.002)	-0.024*** (0.002)	-0.014*** (0.002)	-0.015*** (0.002)
1980 Data	-0.000 (0.003)	-0.010*** (0.003)	-0.000 (0.003)	-0.005** (0.002)
Jurisdiction FE	X		X	
Border FE		X		X
Total N	259714		259714	

Significance levels: * = 10%; ** = 5%; *** = 1%.

Notes: This table presents outputs using the border discontinuity research design to estimate causal effects of MLS requirements on neighborhood resident shares by race and ethnicity. Over Census blocks b in year t , we run the regression:

$$\text{Share}_{bt} = \alpha_{j(b)t} + \beta^t \mathbf{1}[\text{Dist}_b \geq 0] + \eta_-^t \text{Dist}_b + \eta_+^t \text{Dist}_b \cdot \mathbf{1}[\text{Dist}_b \geq 0] + \varepsilon_{bt},$$

over the people of color (POC) share at the block level Share_{bt} as well just for Black residents, $\text{Share}_{bt}^{\text{Black}}$. Results are presented for two sets of fixed effects, and for a sample including blocks with centroid distances of 400 meters or less from the nearest border segment detected using the procedure in Section 2. Standard errors are calculated clustering at the border segment-year level.

Sources: Calculations from 1980–2020 NHGIS Tables (Manson et al. (2024)) and CoreLogic Tax Records.

Table A.7: Neighborhood Composition Effects of Alternative Minimum Lot Size Segment

2020 Outcomes	Causal Forest		Border Discontinuity	
	(1)	(2)	(1)	(2)
People of Color Share	-0.001 (0.003)	-0.007*** (0.003)	0.002 (0.006)	-0.003 (0.004)
Black American Share	-0.010*** (0.002)	-0.013*** (0.002)	-0.002 (0.005)	-0.008*** (0.002)
Hispanic American Share	0.002 (0.002)	-0.001 (0.002)	0.006 (0.005)	0.001 (0.003)
Asian American Share	0.001 (0.001)	0.001 (0.001)	-0.005** (0.002)	-0.001 (0.002)
Impact Segment	Moderate-Impact (41% of sample)			
Jurisdiction FE	X		X	
Border FE		X		X
Total N	28520		95513	

Significance levels: * = 10%; ** = 5%; *** = 1%.

Notes: This table presents outputs of border discontinuity designs over Census blocks b in year t for racial minority m ,

$$\text{Share}_{bt}^m = \alpha_{j(b)t} + \beta^t \mathbf{1}[Dist_b \geq 0] + \eta_-^t Dist_b + \eta_+^t Dist_b \cdot \mathbf{1}[Dist_b \geq 0] + \varepsilon_{bt},$$

where shares are taken over all residents who are not non-Hispanic white, as well as just for Black Americans. For each outcome, two models are estimated with different fixed effects specifications. All data surrounding the regulatory segments detected using the procedure in Section 2 are used. The border discontinuity specification is parametric, as we do not drop observations and fit linear functions on both sides of the border discontinuity. Standard errors are calculated clustering at the county-year level.

Sources: Calculations from 1980–2020 NHGIS Tables (Manson et al. (2024)) and CoreLogic Tax Records.

Table A.8: Comparison of Estimated Effects With Literature

Paper	Policy intervention	Census year	Point estimate	95% confidence interval
Sood and E-S	Minneapolis racial covenants	2020	-0.015	[−0.04, 0.01]
Cui and Been	High-impact cluster estimate	2010	-0.029	[−0.041, −0.017]
Cui and Been	Moderate-impact cluster estimate	2010	-0.011	[−0.016, −0.006]
AHM	HOLC maps: gap btwn C-B zones	2010	-0.0074	[−0.023, 0.008]
M&S	Stacked jurisdiction borders	2010	-0.04	[−0.042, 0.038]
Resseger	Stacked SF zones in Boston	2010	-0.0087	[−0.013, −0.004]
Sood and E-S	Minneapolis racial covenants	1980	-0.004	[−0.008, −0.00]
Cui and Been	High-impact cluster estimate	1980	-0.012	[−0.030, 0.006]
Cui and Been	Moderate-impact cluster estimate	1980	-0.004	[−0.011, 0.003]
AHM	HOLC maps: gap btwn C-B zones	1980	-0.049	[−0.126, 0.028]
M&S	Stacked jurisdiction borders	1990	-0.035	[−0.03892, −0.0311]

Notes: Following the discussion in Section 5.4, this Table puts our preferred estimates of racial disparities around lot sizes in 2010 with other border discontinuity estimates of policies in the literature. We report our more conservative results for different impact clusters, based on border discontinuity designs. When abbreviated, “AHM” refers to Aaronson, Hartley and Mazumder (2021) and “M&S” refers to Monarrez and Schönholzer (2023).

Table A.9: Neighborhood Composition Effects, Without Preexisting Segments

2020 Outcomes	Border Discontinuity	
	(1)	(2)
People of Color Share	0.008 (0.012)	-0.018* (0.010)
Black American Share	-0.018* (0.009)	-0.027*** (0.008)
Impact Segment	High-Impact (14% of sample)	
Jurisdiction FE	X	
Border FE	X	
Total N	17220	

Significance levels: * = 10%; ** = 5%; *** = 1%.

Notes: This table presents outputs of border discontinuity designs over Census blocks b in year t for racial minority m ,

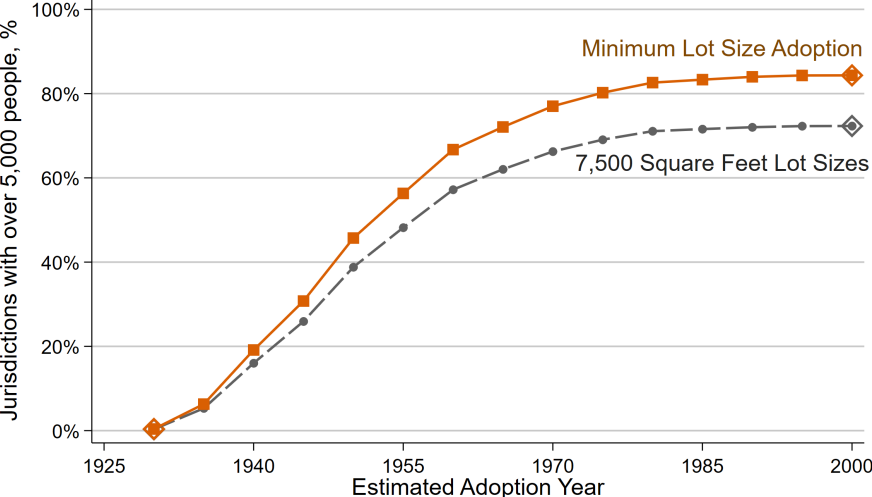
$$\text{Share}_{bt}^m = \alpha_{j(b)t} + \beta^t \mathbf{1}[Dist_b \geq 0] + \eta_-^t Dist_b + \eta_+^t Dist_b \cdot \mathbf{1}[Dist_b \geq 0] + \varepsilon_{bt},$$

where shares are taken over all residents who are not non-Hispanic white, as well as just for Black Americans. For each outcome, two models are estimated with different fixed effects specifications. All data surrounding the regulatory segments detected using the procedure in Section 2 are used. The border discontinuity specification is parametric, as we do not drop observations and fit linear functions on both sides of the border discontinuity. Standard errors are calculated clustering at the county-year level.

Sources: Calculations from 1980–2020 NHGIS Tables (Manson et al. (2024)) and CoreLogic Tax Records.

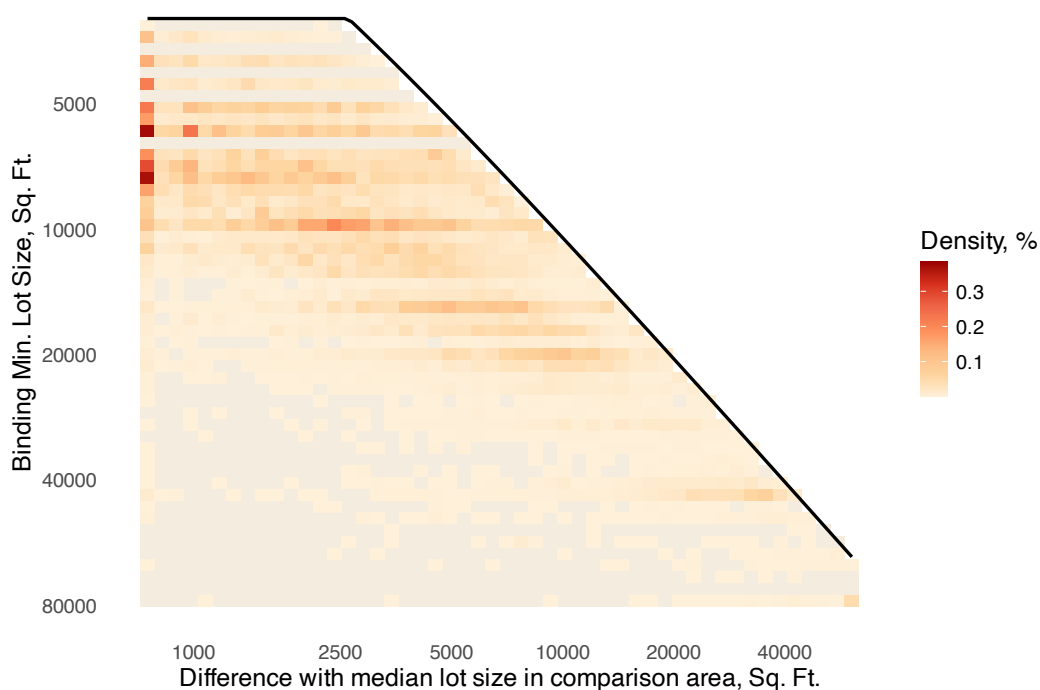
B Appendix Exhibits

Figure B.1: Measured adoption of U.S. density zoning dynamics



Notes: This Figure visualizes estimates of minimum lot size adoption from Cui (2024) among U.S. jurisdictions with a population over 5,000. The time series plots both adoption of any minimum lot sizes, along with initial adoption of a zone with lot sizes over 7,500 square feet.
Sources: Calculations from CoreLogic Tax Records.

Figure B.2: Relative sizes of stratified samples

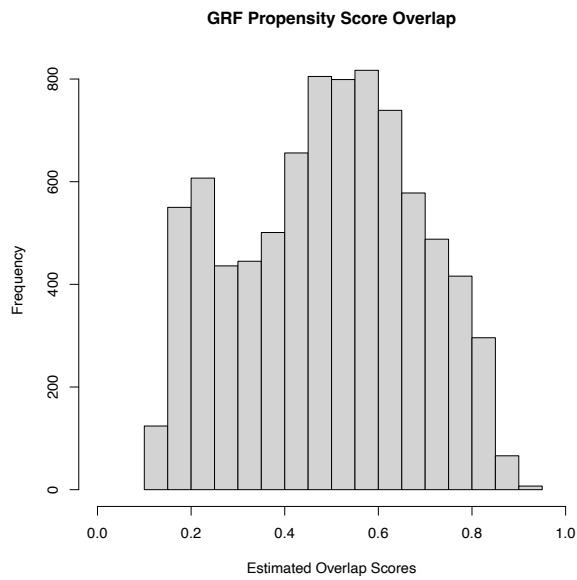


Notes: This figure presents a heatmap plotting a histogram of detected lot size discontinuities, used in our analysis sample. Each cell in the heatmap plots the share of all blocks used in the 2010 Census data satisfying two criteria. First, the sample includes blocks belonging to a minimum lot size district whose values are in one of four ranges on the Y axis. Second, the surrounding urban context of the lot size regulated blocks have higher densities that differ within the ranges on the X axis. When cells span multiple X axis columns, that means the underlying sample includes comparison blocks in the union of the ranges. Only discontinuities detected within jurisdiction boundaries, using the algorithm of Section 2, are included.

Sources: Calculations from 2020 NHGIS Tables (Manson et al. (2024)) and CoreLogic Tax Records.

Figure B.3: Additional tests of research design assumptions

(a) Propensity score overlap for causal forest design



(b) No observation manipulation for border discontinuity design

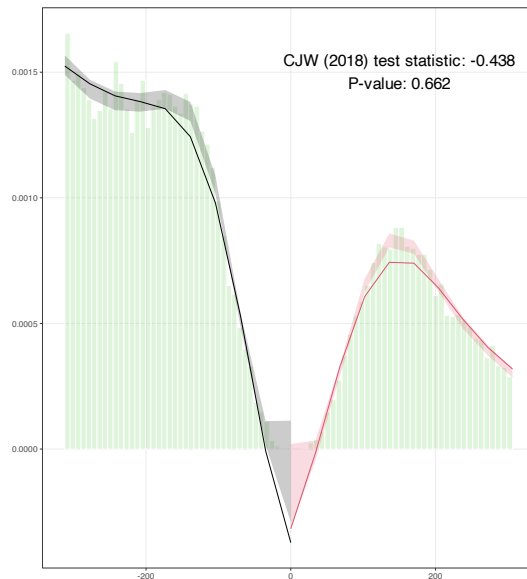
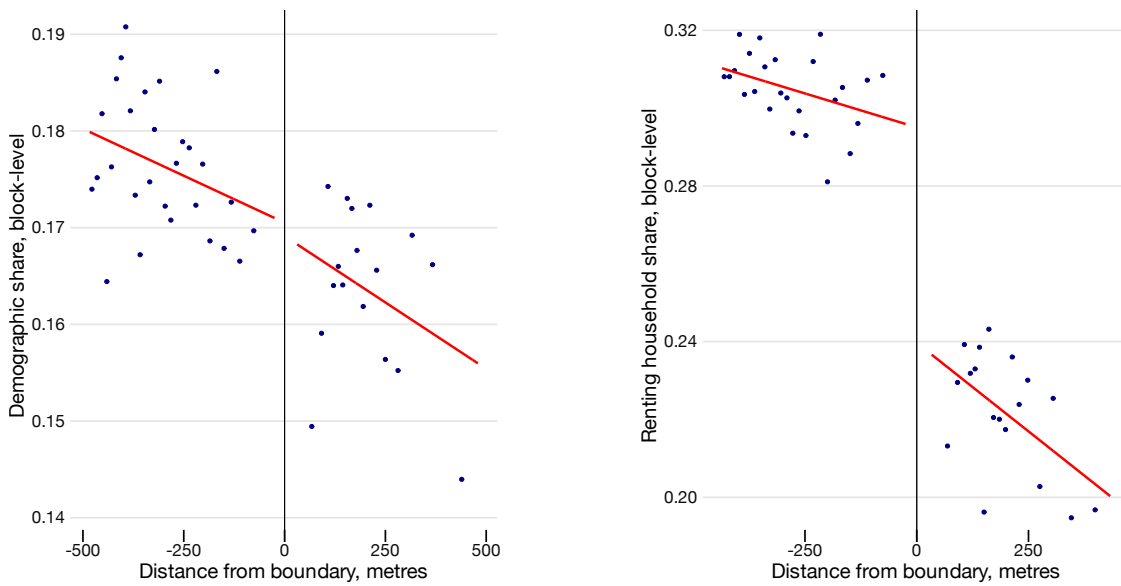


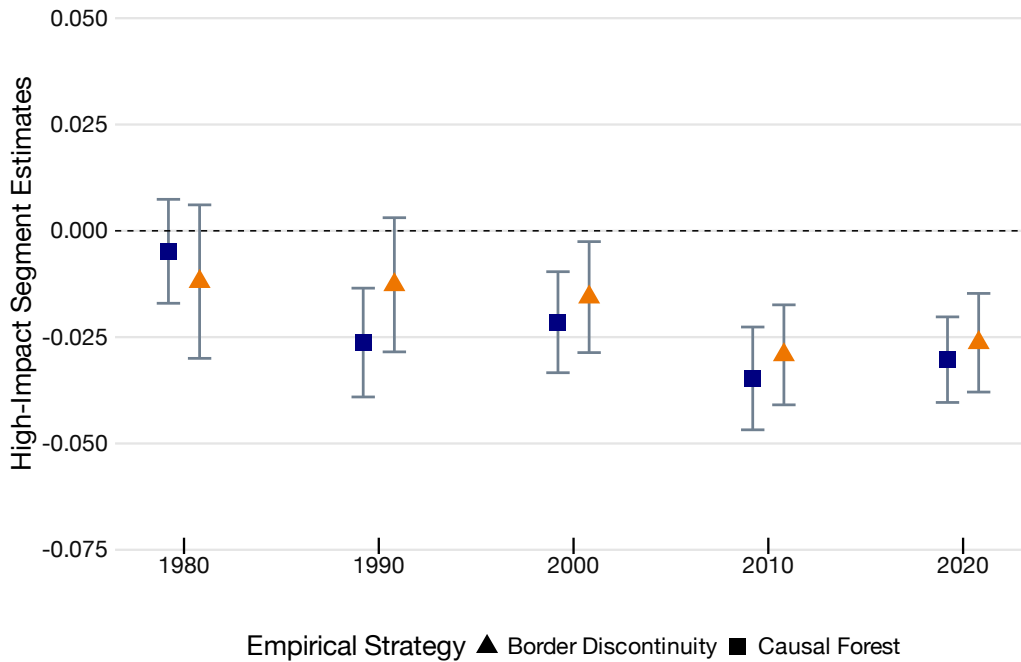
Figure B.4: Border discontinuities over 1980 Census data

(a) Outcome: Block-level POC shares (b) Outcome: Block-level rented housing unit shares



Notes: Sources: Calculations from 2020 NHGIS Tables (Manson et al. (2024)) and CoreLogic Tax Records.

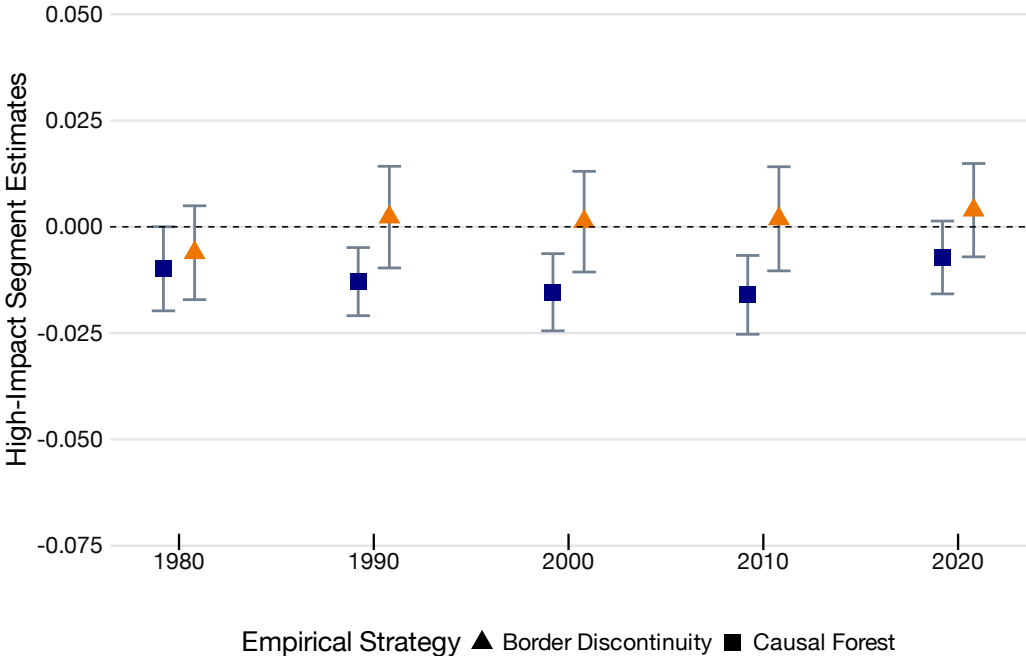
Figure B.5: Dynamic composition effects, Black resident share as outcome



Notes: This figure presents border discontinuity effects for the block-level people of color share $Share^m$, using data from 1980 to 2020. Effects are estimated over the high-impact sample defined in Section 4.2. The sample is limited to blocks around straight segments of lot size borders, detected through the procedure in Section 2. We present effects estimated from two models, with different identifying assumptions listed in 4.3: a causal forest design controlling for selection on observables, and a border discontinuity design controlling for unobservable divergence in neighborhoods further away from the discontinuity. 95% confidence intervals are presented with standard errors clustered at the discontinuity level.

Sources: Calculations from 1980–2020 NHGIS Tables (Manson et al. (2024)) and CoreLogic Tax Records.

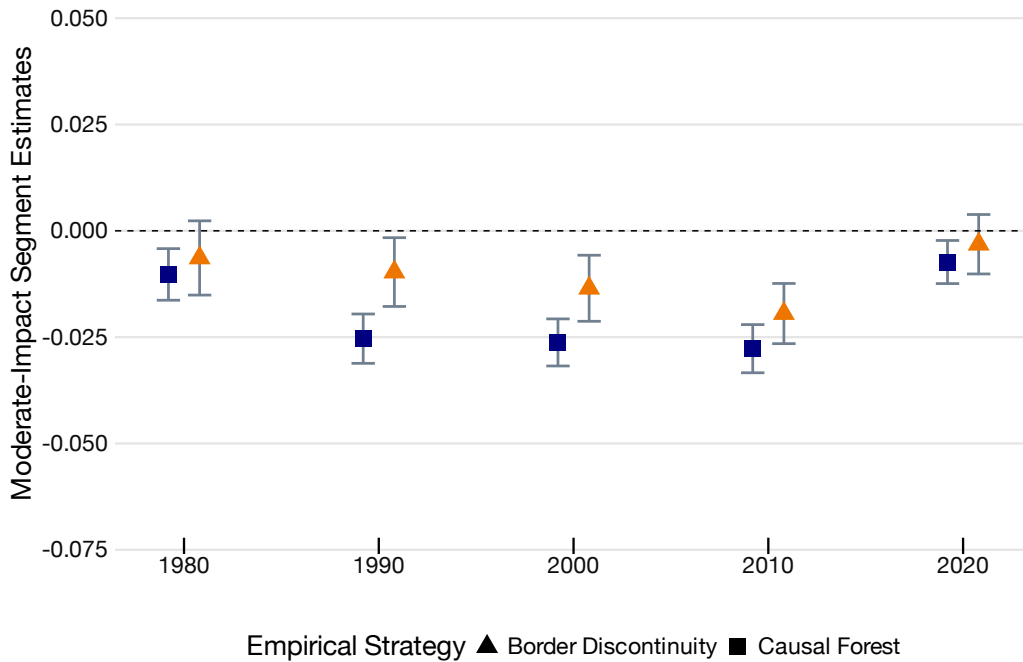
Figure B.6: Dynamic composition effects, Hispanic resident share as outcome



Notes: This figure presents border discontinuity effects for the block-level people of color share $Share^m$, using data from 1980 to 2020. Effects are estimated over the high-impact sample defined in Section 4.2. The sample is limited to blocks around straight segments of lot size borders, detected through the procedure in Section 2. We present effects estimated from two models, with different identifying assumptions listed in 4.3: a causal forest design controlling for selection on observables, and a border discontinuity design controlling for unobservable divergence in neighborhoods further away from the discontinuity. 95% confidence intervals are presented with standard errors clustered at the discontinuity level.

Sources: Calculations from 1980–2020 NHGIS Tables (Manson et al. (2024)) and CoreLogic Tax Records.

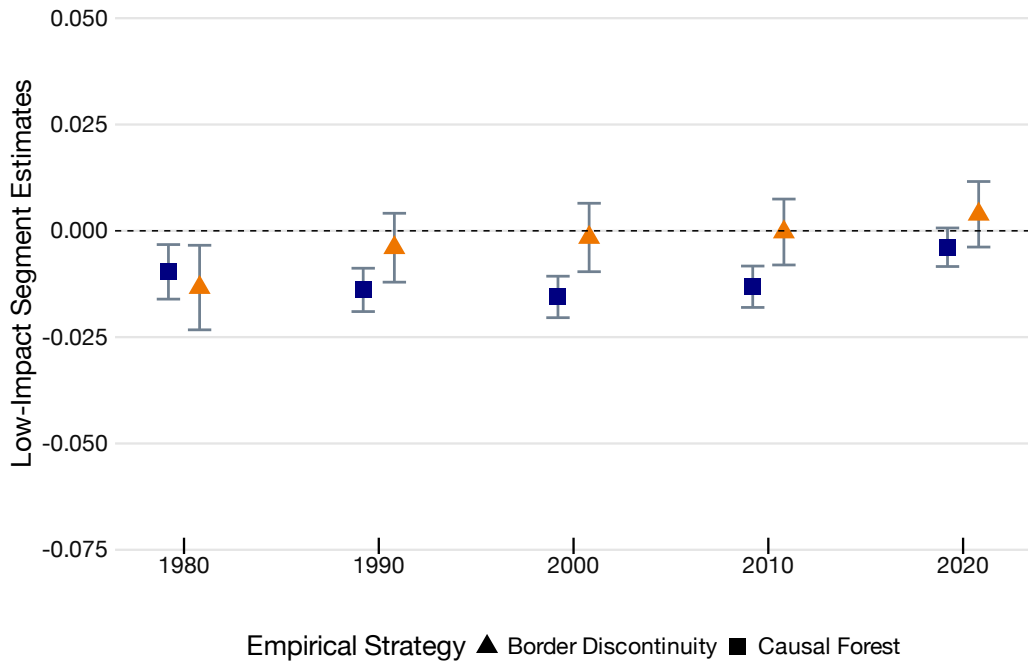
Figure B.7: Dynamic composition effects, moderate-impact segment MLS



Notes: This figure presents border discontinuity effects for the block-level people of color share $Share^m$, using data from 1980 to 2020. Effects are estimated over the high-impact sample defined in Section 4.2. The sample is limited to blocks around straight segments of lot size borders, detected through the procedure in Section 2. We present effects estimated from two models, with different identifying assumptions listed in 4.3: a causal forest design controlling for selection on observables, and a border discontinuity design controlling for unobservable divergence in neighborhoods further away from the discontinuity. 95% confidence intervals are presented with standard errors clustered at the discontinuity level.

Sources: Calculations from 1980–2020 NHGIS Tables (Manson et al. (2024)) and CoreLogic Tax Records.

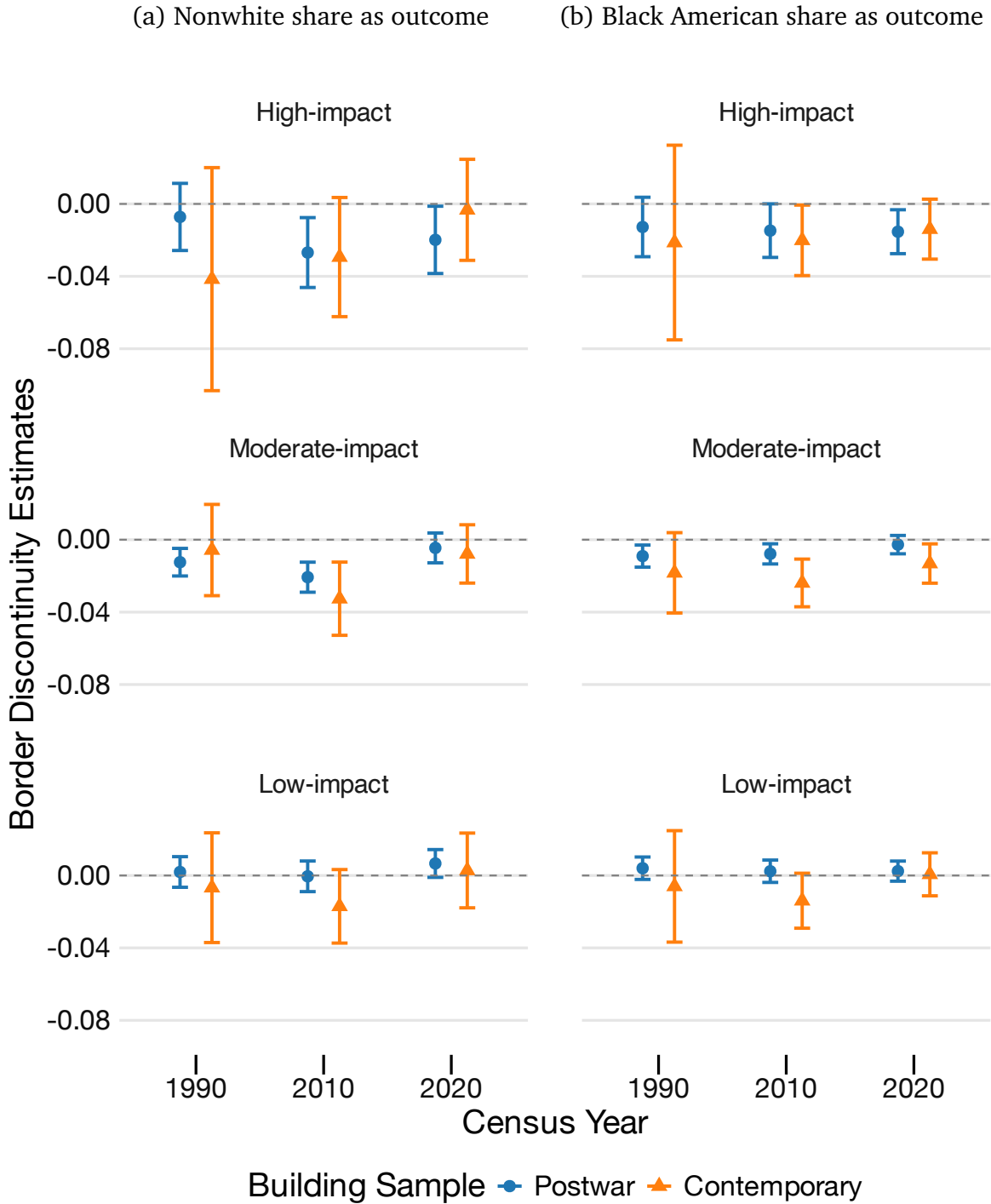
Figure B.8: Dynamic composition effects, low-impact segment MLS



Notes: This figure presents border discontinuity effects for the block-level people of color share $Share^m$, using data from 1980 to 2020. Effects are estimated over the high-impact sample defined in Section 4.2. The sample is limited to blocks around straight segments of lot size borders, detected through the procedure in Section 2. We present effects estimated from two models, with different identifying assumptions listed in 4.3: a causal forest design controlling for selection on observables, and a border discontinuity design controlling for unobservable divergence in neighborhoods further away from the discontinuity. 95% confidence intervals are presented with standard errors clustered at the discontinuity level.

Sources: Calculations from 1980–2020 NHGIS Tables (Manson et al. (2024)) and CoreLogic Tax Records.

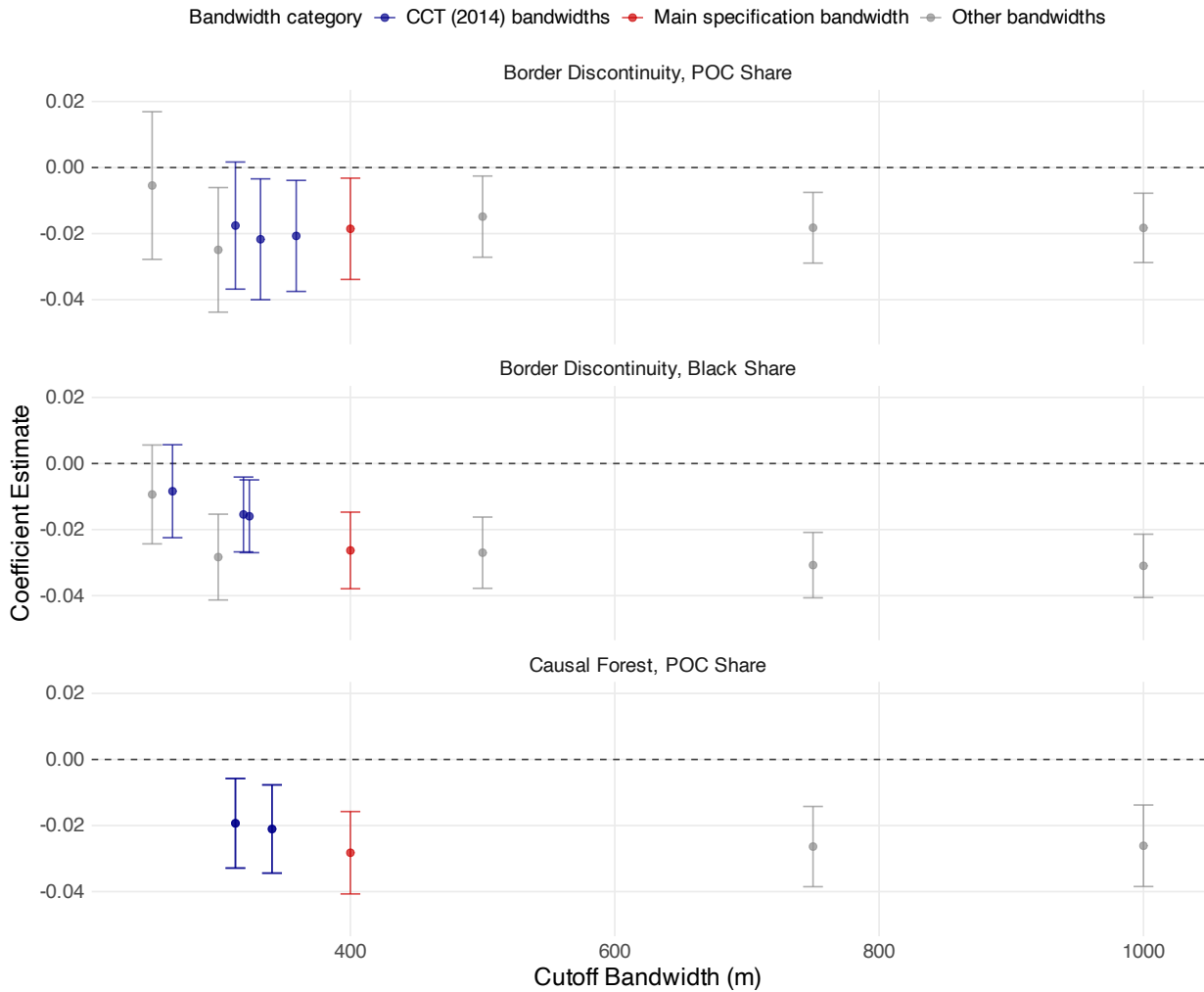
Figure B.9: Diversity effects around minimum lot sizes, robustness to housing vintage



Notes: This figure presents border discontinuity effects on the 2020 Census block-level Black share, estimated across stratified samples of lot size segments. Effects are estimated from the fixed effects model described in Section 5.6, with standard errors clustered at the county level. The samples reflect distinct urban contexts around straight segments of lot size borders, detected through the procedure in Section 2. In addition, the samples are split by the median year built of properties in the area around the segments: the *contemporary* sample includes only areas all built after 1980. Section 4.2 gives the exact definitions of the samples.

Sources: Calculations from 1990–2020 NHGIS Tables (Manson et al. (2024)) and CoreLogic Tax Records.

Figure B.10: Diversity effects around minimum lot sizes, robustness to bandwidth



Notes: This figure presents border discontinuity effects on the 2020 Census block-level Black share, estimated across stratified samples of lot size segments. Effects are estimated from the fixed effects model described in Section 5.6, with standard errors clustered at the county level. The samples reflect distinct urban contexts around straight segments of lot size borders, detected through the procedure in Section 2. In addition, the samples are split by the median year built of properties in the area around the segments: the *contemporary sample* includes only areas all built after 1980. Section 5.6 gives the exact definitions of the samples.

Sources: Calculations from 1990–2020 NHGIS Tables (Manson et al. (2024)) and CoreLogic Tax Records.

C Details of automated procedure for MLS border segments

Detailed workflow of automated procedure. At a high level, the procedure is first trained on jurisdictions in the Boston metropolitan area, where the MAPC Atlas provide us with detailed zoning borders. We cycle through a grid of parameter estimates, of which there are 9 in total. 2 of the 9 relate to the post-processing pruning of lot size discontinuities flagged to be far away from the MAPC zoning borders, and are reestimated for every combination of the 7 other hyperparameters we test.

Figure C is a schematic of how the code proceeds during the training stage, and then how it scales up to many MSAs while storing the parameters trained from Boston data. Appendix Table C.1, the full table of parameters expanding on Table 1, references additional parameters that are defined in detail in the rest of the section.

Additional details on identifying interior tiles with lot size discontinuities. We employ two asymmetric rules to filter interior tiles that contain a lot size discontinuity. First, corresponding to a bunching bin $\underline{\ell}$ in a jurisdiction, there needs to be \underline{N} lots whose sizes are between the range $[\underline{\ell}, M \cdot \underline{\ell}]$. Then, there needs to be at least $0.4 \cdot \underline{N}$ lots whose sizes are below $\underline{\ell}$. The latter criterion should then be thought of as an additional check, ensuring that we are not in a tile where every lot is of size $\underline{\ell}$ or higher: such tiles are uninformative about the shape of the MLS border segment.

A common enough edge case is that in the same interior tile, there was detection of a linear border segment for multiple MLS requirements in that jurisdiction. The multiple detection could correspond to short zoning district borders that do not stretch over the entire tile, but without further adjustment this leads to overlap issues with regions. Where classified treated areas intersect, a block could be identified as belonging to two separate minimum lot size districts at once.

Later in this appendix, we discuss a general approach based on a K-nearest neighbor method to assign areas to only one district with a binding MLS requirement. A simpler step we do before that is to check if any treated areas of one MLS requirement is a strict subset of another. If this is the case and the subsetted MLS requirement is also not too great a size from the requirement with more broad coverage, the subsetted MLS requirement is removed. In this way, we aim to remove false positives that correspond to development bunching at values that are not necessarily induced by a MLS regulation.

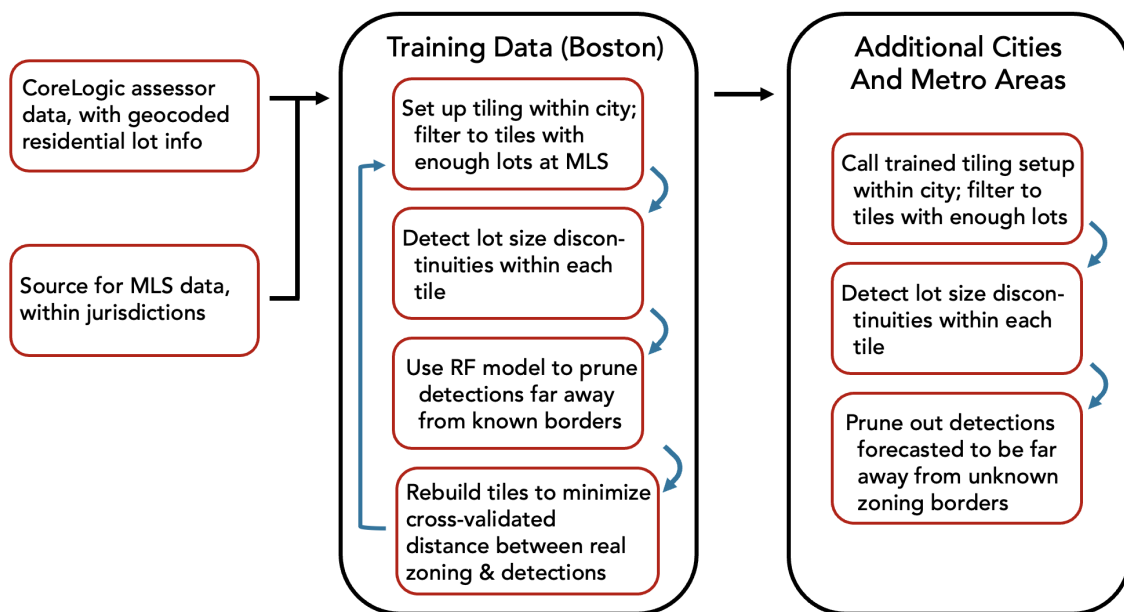
Matching lots to the zoning jurisdiction in place when developed. Even within a single interior tile of a jurisdiction, some lots may have been developed under one jurisdiction's MLS requirements and the rest under another. This is the case if the tile includes development that was later incorporated to be in a jurisdiction, or annexed to be within a nearby jurisdiction's borders. Some developments would have been built while the land was in the unincorporated section of a county. In rare cases, jurisdiction mergers can occur where a larger jurisdiction consolidates land that used to be under more fragmented jurisdictions.

The Cui (2024) data, for each jurisdiction as of 2010, also includes information on the year of incorporation. If the jurisdiction has annexed land following 1980, we also have the requisite shapefile data that demonstrates if the land was annexed at a later date.

If a tile covers a part of the jurisdiction with homes built before and after incorporation,

or homes in annexed land that are built before or after 1980, we run the border segment detection exercise twice on that tile. The homes built before the relevant date (incorporation year or before 198) are treated separately by checking if a lot size discontinuity holds following one of the bunching bins ever detected for the wider unincorporated county the homes were in. If two separate border segments are outputted through this process for the same interior tile, the border segment we select is based on whichever of the two employed a larger sample of housing units. This reflects suggestive historical records on how, after incorporation, newly incorporated jurisdictions might have reused the zoning district borders set up by a prior county zoning commission (see also Gallagher, Shertzer and Twinam (2024))

Figure C.1: Schematic of procedure training and national scaling



Local support vector machine problem. In each of the interior cells, we take all geocoded lots within the cell’s borders. Based on the minimum lot size we want to measure, the lots are classified into two classes $Y \in \{1, -1\}$. We want to find a linear boundary over two variables, the longitude and latitude, which is the separating hyperplane for if a lot belongs in either class. The optimization problem is to choose the linear boundary such that the lots are correctly classified as much as possible.

The objective function solved by the SVM classifier is pick the support vector of the boundary (w, b) and misclassification distances ξ , then solve

$$\begin{aligned} \min_{w,b,\xi} & \frac{1}{2} w^T w + C \sum_{i=1}^n \xi_i \\ \text{s.t.} & y_i(w^T x_i + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0. \end{aligned}$$

When minimizing the objective, the SVM classifier is trading off accurate prediction — in which case $\text{sgn}(y_i(w^T x_i + b)) \geq 0$ and can be normalized to be ≥ 1 — and minimizing deviations in distance between misclassified lots and the boundary. The strength of this tradeoff is governed by the hyperparameter C .

We implement the SVM procedure using the `scikit-learn` package in Python, which speeds up the procedure by solving the dual problem to the above primal problem. The dual problem involves minimizing a quadratic form subject to linear constraints.

K-nearest neighbor procedure for multiple predicted segments. A common enough edge case is that in the same interior tile, there was detection of a linear border segment for multiple MLS requirements in that jurisdiction. The multiple detection could correspond to short zoning district borders that do not stretch over the entire tile, but without further adjustment this leads to overlap issues with regions. Where classified treated areas intersect, a block could be identified as belonging to two separate minimum lot size districts at once.

For these tiles, we refine boundaries using a K-nearest neighbor algorithm. Figure C walks through this exercise for a sample cell and geocoded property data within it. For this interior cell, the SVM procedure detected regulatory border segments for three separate lot sizes. The implied lot size areas detected are shown on the left graphic, where overlapping between the detected areas is evident.

To set up the KNN algorithm, we classify all properties in the cell into $N + 1$ classes, where N is the number of detected lot sizes. The first class is all properties lower than the least detected lot size, visualized as white diamonds in the figure. Then, indexing the detected lot sizes in ascending order $\underline{\ell}_1, \dots, \underline{\ell}_N$, properties are classified as class $k = 2, \dots, N$ if their lot sizes are in the interval $[\underline{\ell}_{k-1}, \underline{\ell}_k]$. All lots above $\underline{\ell}_N$ are classified as class $N + 1$.

With the target classes set up, we run the KNN algorithm by predicting class at interior points in the cell based on the nearest k_{mult} lots. We only run this prediction at a certain resolution, which is determined by the radius parameter r_{mult} . We standardize distances within the cell, so r_{mult} is expressed in standardized units: the larger it is, the wider the distance is between points at which we predict the MLS district class under which the area falls.

The figure on the right side of Figure C shows the output of the KNN procedure for the cell. First, note that lots below any detected lot size are now in their own class. Then, the regions where each lot size minimum is predicted to apply are the union of predictions at local points, expanded by taking a square buffer around each.

As the example shows, the output lot boundaries may look more irregular or even disjoint. To ensure robustness of the KNN procedure, we apply the same filter as the SVM procedure: KNN predictions with a misclassification rate above M_{err} are dropped as well from the analysis sample.

Details of random forest pruning step. Information specific to Boston’s MAPC Zoning Atlas can, nevertheless, help us to detect predicted lot size discontinuity border segments that do not actually match zoning borders. Those border segments could have alternatively been generated by, for example, a subdivision where lots all follow a common lot size standard that also coincides with a lot size requirement in another zoning district.

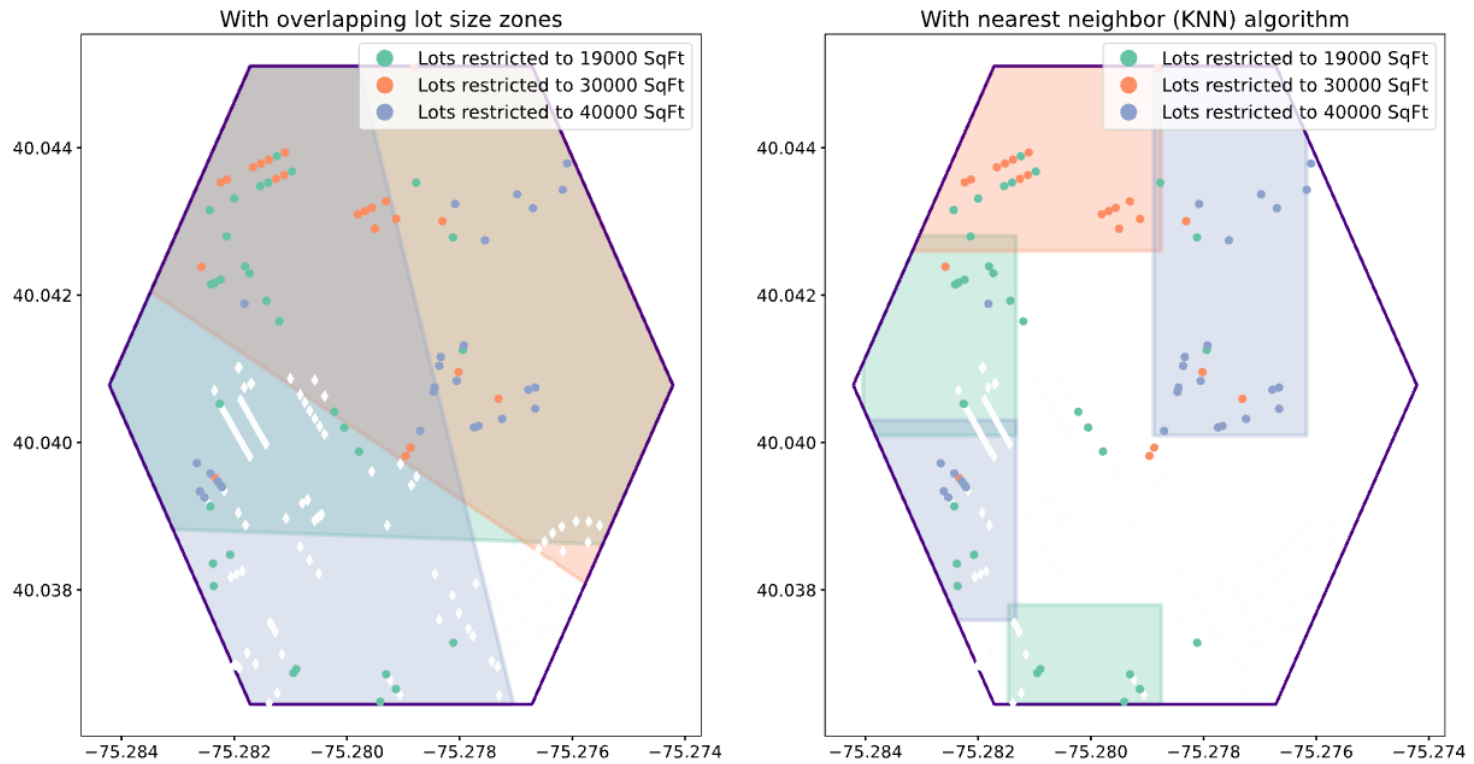
When first fit on MAPC data as part of hyperparameter tuning, the “post-processing step” employs a supervised random forest approach to identify and filter outlier border segments far away from any known zoning borders. Whether a lot size discontinuity is an “outlier” is defined geometrically: the binary outcome equals 1 if the closest distance of the lot size discontinuity to any real zoning border segment is greater than $d_{threshold} = \sqrt{3} \cdot R$, where R is the interior tile side length hyperparameter. This threshold is equal to the diameter of a single interior tile.

To predict whether the lot size outliers is most likely to exceed that outlier threshold, we use a set of features that track the underlying MLS requirement’s characteristics, as well as its relative positioning among all requirements passed in the jurisdiction. Out of a total of 14 characteristics, we include tile-level characteristics like median year built of homes on the treatment and comparison sides; the Cui (2024) estimated MLS adoption year for the jurisdiction, to capture relative age in the development compared to the jurisdiction’s zoning history; counts of properties used for border detection; the cardinal ranking of the local MLS required lot size compared to all requirements in that jurisdiction; and the ratio of the MLS requirement size on the treated side to the median lot size on the denser comparison side. By flexible modeling these variables in a random forest, we can capture some broad patterns where zoning behavior involving the most extreme requirements in a within-zoning jurisdiction is less likely to be associated with an actual zoning border.

When training on the MAPC data, the non-pruning hyperparameters are first used to detect the potential set of lot size discontinuity border segments. Then, the random forest is trained under a 10-fold cross-validation stage, using group-aware splitting clustered at the jurisdiction level to prevent data leakage between training and test sets. For the random forest model, the hyperparameter search space includes: the number of estimators $N_{estimators} \in \{25, 50, 80\}$ and maximum tree depth $D_{max} \in \{12, 15, 20\}$. The number of features to consider at each split $m_{features}$, follows a logarithmic rule $\log_2(p)$ where p is the total number of features. The output MAE after this post-processing step, for each fold of validating jurisdictions, is one realization of the training objective Equation 1 in the paper.

On average, the post-processing pruning steps keep 75-80% of the fuller set of lot size discontinuity border segments. Then, since all predictive features are functions of the CoreLogic assessor data or of the Cui (2024) data on bunching around MLS requirements, the weights from the MAE-optimizing random forest model are stored and reused to be applied nationally.

Figure C.2: Illustration of additional KNN procedure



Sources: Calculations from CoreLogic Tax Records.

Table C.1: Full parameter table for automated border detection method

Parameter name	Symbol	Value
Hexagonal tile radius, meters	R	300
Count of lots at bunched sizes	\underline{N}	8
Bunching range factor	\underline{M}	1.25
Misclassification rate threshold	\overline{M}_{err}	0.35
Neighboring lots for KNN extension	k_{mult}	5
Misclassification weight	C	1
KNN prediction radius	r_{mult}	1.05
Number of separate trees in RF	$N_{estimators}$	80
Maximum tree depth of RF	D_{max}	20

Notes: This Table lists the key parameters used in different stages of the automated detection procedure for lot size discontinuities, as detailed in Section 2. Except for dimensionless parameters, parameter units are given in the leftmost column.