Grantham Research Institute on Climate Change and the Environment

Disparities in pollution capitalization rates: the role of direct and systemic discrimination

Joshua Graff Zivin and Gregor Singer

January 2023

Grantham Research Institute on Climate Change and the Environment
Working Paper No. 392

ISSN 2515-5717 (Online)

The Grantham Foundation for the Protection of the Environment

LSE The London School of Economics and Political Science

The Grantham Research Institute on Climate Change and the Environment was established by the London School of Economics and Political Science in 2008 to bring together international expertise on economics, finance, geography, the environment, international development and political economy to create a world-leading centre for policy-relevant research and training. The Institute is funded by the Grantham Foundation for the Protection of the Environment and a number of other sources. It has 13 broad research areas:

1. Biodiversity
2. Climate change adaptation and resilience
3. Climate change governance, legislation and litigation
4. Climate, health and environment
5. Environmental behaviour
6. Environmental economic theory
7. Environmental policy evaluation
8. International climate politics
9. Science and impacts of climate change
10. Sustainable public and private finance
11. Sustainable natural resources
12. Transition to zero emissions growth
13. UK national and local climate policies

More information about the Grantham Research Institute is available at www.lse.ac.uk/GranthamInstitute

Suggested citation:
Graff Zivin J and Singer G (2023) *Disparities in pollution capitalization rates: the role of direct and systemic discrimination*. Grantham Research Institute on Climate Change and the Environment Working Paper 392. London: London School of Economics and Political Science

Disparities in Pollution Capitalization Rates: The Role of Direct & Systemic Discrimination

Joshua Graff Zivin* Gregor Singer*

UC San Diego & NBER LSE

January 27th, 2023

Abstract

We examine how exogenous changes in exposure to air pollution over the past two decades have altered the disparities in home values between Black and White homeowners. We find that air quality capitalization rates are significantly lower for Black h omeowners. In fact, they are so much lower that, despite secular reductions in the Black-White pollution exposure gap, disparities in housing values have increased during this period. An exploration of mechanisms suggests that roughly one-quarter of this difference is the result of direct discrimination while the remaining three-quarters can be attributed to systemic discrimination through differential access to complementary amenities.

Keywords: house prices, environmental justice, air pollution, race
JEL codes: Q53, R31, J15

*Joshua Graff Zivin: University of California, San Diego, CA 92093, E-mail: jgraffzivin@ucsd.edu. Gregor Singer: London School of Economics, London WC2A 2AE, UK, E-mail: g.a.singer@lse.ac.uk. We thank Akshaya Jha, Craig McIntosh and Anant Sudarshan for helpful discussion. We thank Zillow for providing data through the Zillow Transaction and Assessment Dataset (ZTRAX). More information on accessing the data can be found at http://www.zillow.com/ztrax. The results and opinions are those of the author and do not reflect the position of the Zillow Group. G.S. acknowledges support from the Grantham Research Institute on Climate Change and the Environment at the London School of Economics.

While racial segregation in the United States formally ended more than half a century ago, the existence of predominantly Black and White neighborhoods persists, and they continue to differ on a wide range of dimensions. One important dimension is pollution, where Black communities are disproportionately exposed to poor air quality relative to their White counterparts (Jbaily et al. 2022). Since the harms from air pollution, which include health as well as other human capital impairments (Graff Zivin & Neidell 2012), has been shown to capitalize into housing values (Chay & Greenstone 2005, Grainger 2012, Bento et al. 2015, Sager & Singer 2022) it may also contribute to the well-documented racial disparities in housing values across and within neighborhoods (Myers 2004, Faber & Ellen 2016, Bayer et al. 2017, Perry et al. 2018, Kermani & Wong 2021, Kahn 2021). In this paper, we examine this relationship directly by examining how changes in exposure to air pollution over the past two decades have altered the disparities in home values between Black and White homeowners.

We begin by noting that there are good reasons to be optimistic. The Clean Air Act Amendments and other secular trends have led to significant air quality improvements (Colmer et al. 2020), and those improvements were larger in black communities thereby reducing the black-white exposure gap (Currie et al. 2020, Sager & Singer 2022). Whether this also reduced disparities in housing values, however, depends not only on relative exposure, but also on whether this amenity capitalizes similarly across communities. To estimate this relationship, we rely on three distinct datasets that include detailed information on housing values, pollution exposure, and the racial composition of homeowners in the US at a fine level of spatial resolution. In particular, we utilize address-level housing characteristics and transaction data from Zillow, data on fine particulate matter (PM2.5) concentrations measured at a 1km-by-1km gridded scale, and the count of Black, Non-Hispanic White, and Other homeowners at the Census block level, which corresponds to approximately 14 owner-occupied households or 37 individuals in a given neighborhood.

Our study sample includes all households in residential homes for those Census blocks for which we have reliable data on both transaction level house prices and square footage in the period 2000-2019, comprising 92 million individuals in 33 million owner-occupied households in the contiguous US. Our core analysis is focused on homeowners for whom the implications of housing price changes are directly interpretable.[1] Our primary estimation strategy relies on a long-differences design that controls for time-invariant factors that differ across communities. Despite the richness of our data, there are two main challenges to estimation of pollution capitalization rates by racial groups.

[1]Increasing house values are generally good for homeowners and landlords, but bad for renters, as house appreciations are at least partially passed through to renters.

First, changes in pollution are correlated with changes in socioeconomics and other amenities that affect house prices. More economic activity, for example, is likely to increase both pollution and house prices. We overcome this challenge using a well-established instrumental variable strategy that exploits Clean Air Act rules that led to plausibly exogenous differential changes in air quality across counties (Chay & Greenstone 2005). We follow Sager & Singer (2022) to account for the bias in the first stage arising from differential time trends due to differences in pre-sample pollution, and allow for heterogeneous effects on nonattainment based on pre-sample pollution levels (Auffhammer et al. 2009, Bishop et al. 2018).

Second, sorting into neighborhoods based on preferences, income, or education is likely to be correlated with both changes in house prices and race, potentially biasing estimates of pollution capitalization rates by racial groups. Since we do not have reliable exogenous variation for changes in the spatial distribution of racial groups, we rely on the baseline distribution of homeowners by keeping household location fixed in 2000 to isolate all temporal variation coming from pollution changes. We employ two strategies to assess the robustness of this strategy: using time-invariant spatial distributions from different Census years and focusing on areas with little change in racial composition during our study period. To address remaining concerns that the interaction of pollution changes with the baseline distribution may be correlated with interactions with socioeconomic factors, we control for observed factors such as poverty, income, urbanicity or baseline pollution, all fully interacted with pollution changes and appropriately instrumented.

Our results show that a one unit decrease in $PM_{2.5}$ increases the price per square foot (PSQFT) of housing by 11.6%, a figure consistent with previous estimates (Sager & Singer 2022), but this average figure masks considerable heterogeneity across racial groups. While the Non-Hispanic White (NHW) pollution capitalization rate is 12.4%, the Black capitalization rate is only 7.6%, a difference of 63% in relative terms, and over 100% in absolute terms since the PSQFT levels are higher for NHW homeowners. Despite the larger decrease in $PM_{2.5}$ for Black homeowners (5.69 units) relative to NHW homeowners (4.92 units), the much lower Black capitalization rate per unit of cleaner air means that the Black-White housing-value gap actually increased as a result of those pollution reductions. To be clear, both groups still experience gains from cleaner air, but at differential rates. Indeed, if black homeowners had the same capitalization rate as their NHW counterparts, their house values would have been 28% higher by the end of 2019.

We probe the mechanisms underlying the results by distinguishing between systemic and direct discrimination (Bohren et al. 2022)). While the lines between the two can be blurry, systemic discrimination generally refers to discrimination that occurs at a societal level as a result of institutional

and cultural practices that unfairly privilege one group over another.[2] Since disadvantages can accumulate over time, these effects are pernicious and difficult to precisely measure. In this paper, we define systemic discrimination as one that arises due to differences in access to complementary amenities that impacts the capitalized value of clean air across communities (e.g. green outdoor spaces).[3] Our underlying assumption is that, conditional on socioeconomic factors, preferences for clean air do not vary across racial groups. We define direct discrimination as the racial differences in housing capitalization that remain even after accounting for differential access to those complementary amenities.[4]

Our main analysis is designed to capture the composite of disparities in pollution capitalization rates within Census blocks and across blocks. Across blocks, differences in amenities that are complementary to clean air may vary by racial groups driving disparities in capitalization rates. Within 200m-by-200m blocks, we assume that amenities are held constant, so disparities in pollution capitalization rates are more likely to stem from direct discrimination. Two pieces of evidence help us to arbitrate between the two mechanisms. First, we exploit data on the universe of renters by racial group to difference out the portion of complementary amenities that differ by racial groups of residents in a difference-in-differences of capitalization rates. Second, we exploit data on mortgages and seller names to additionally compare transactions strictly within Census blocks holding neighborhood amenities and racial composition constant, including a repeat sales design to account for unmeasured differences in housing quality that may be correlated with race and thus might also impact sales price. Both strategies suggest that approximately three quarters of our results are driven by systemic discrimination via complementary amenities across blocks, and one quarter from within-block direct discrimination.

This paper contributes to the growing literature on environmental justice (Banzhaf et al. 2019), and connects the literature on housing prices and racial groups (Akbar et al. 2020, Aaronson et al. 2021, Kahn 2021, Kermani & Wong 2021) with that on housing prices and pollution (Chay & Greenstone 2005, Grainger 2012, Currie et al. 2015, Bento et al. 2015, Bayer et al. 2016, Sager & Singer 2022). Our findings that the pollution capitalization rate differs by race provides novel insights into how the marginal effects of pollution exposure differ across the population, which is critical for understanding the distributional effects of air quality policies (Hsiang et al. 2019). Furthermore, our

[2] This is also closely related to the concept of statistical discrimination (Phelps 1972).

[3] As we show in robustness checks, these differences in complementary amenities persist even after controlling for socioeconomic factors, but we view differences without controlling for them more representative of systemic discrimination, since racial differences in factors such as income and education are often themselves the result of systemic discrimination.

[4] As an illustrative recent example, a Black couple filed a lawsuit after they received substantially higher valuations from an appraiser when a White colleague posed as the homeowner (US District Court 2022).

analysis of mechanisms is, to our knowledge, the first to unpack the relative roles of direct discrimination and systemic racial discrimination vis-a-vis neighborhood amenities. Our findings are consistent with recent evidence on discriminatory pathways, such as racial steering in the housing market Christensen & Timmins (2022, 2021), Christensen et al. (2022), racial disparities in mortgage lending and refinancing practices (Munnell et al. 1996, Charles & Hurst 2002, Ambrose et al. 2021, Bhutta et al. 2022), and lower offers for Black sellers in other marketplaces (List 2004, Doleac & Stein 2013).

I. Data and Descriptives

A. *House price data*

We use two databases from Zillow (2020) that allow us to calculate house prices at a fine spatial granularity. The first database are transactions (ZTransaction) sourced from county recorder's offices with information on transaction price, type of deed or date of sale. The second database contains hedonic information (ZAssessment), sourced from county assessor's offices, including square footage and geolocation. We combine both datasets to calculate price per square foot (PSQFT) for each transaction. This has the advantage that we can account for differential trends across racial groups in house sizes that would otherwise bias our estimates.[5] Importantly, we only use arm's length transactions and residential properties, dropping transactions such as refinancing or foreclosures, and use historic assessment data to reduce missing values of hedonic information. The construction of the PSQFT data at the transaction level along with descriptive statistics is discussed in detail in Appendix A.7. Since we use log PSQFT and state-by-year fixed effects for our analysis, our data are effectively deflated with state deflators. In robustness checks in Table A.9, we additionally predeflate all prices by a quarterly GDP deflator from FRED (2022) and obtain very similar results.

We map the geolocation of each transacted property to US Census blocks using the 2000 US Census boundaries. Census blocks are the finest administrative unit, and there are approximately 8 million Census blocks in the contiguous US with a population of 53 individuals in total or 37 in owner-occupied housing on average. This allows us to map housing price changes to an extremely fine spatial unit (see Appendix Figure A.2).

For our analysis using long differences, we use a base period of 2000-2003, and use the median PSQFT of all transacted properties within each Census block in that period. We do the same for our

[5] We only use property location and size as other hedonic information is often missing (details in Appendix A.7 and Table A.27).

end period 2016-19. The blocks for which we have data in both periods include 49% of the population that lives in owner occupied housing in the US. In a robustness check, we also impute missing PSQFT data using Census block *group* medians, expanding coverage to 68% of the population with similar results (Table A.9).

Figure 1c provides an overview of the spatial coverage of our PSQFT data in 2016-2019 where data in 2000-2003 is also available. The map aggregates Census block level data up to Census tract averages for better visualization across the entire US. This masks a large degree of spatial granularity within Census tracts. The maps in Appendix Figures A.2a and A.2b show the variation within Census tracts for New York State and a few counties around New York City, respectively. This illustrates that, even across a few city blocks, house prices can vary due to differences in amenities.

B. Census data on tenancy by race

We combine our Census block level PSQFT time series with data from the 2000 Census on population counts by tenancy and race of the householder. Following Currie et al. (2020), we use two racial groups, Black and Non-Hispanic White (NHW) Americans, and add a third group for Other Americans. For each racial group, we know how many homeowners and renters are in each Census block, allowing us to form six groups in total. For our main analysis, we focus on three homeowner groups: Black, NHW, and Other. We assign PSQFT data to each individual according to the race and tenancy of the householder based on the Census blocks where they lived in 2000. This switches off any changes induced by spatial sorting over time, an issue we discuss in more detail below.

The first set of rows in Table 1 shows the number of individuals in the contiguous US and the number of individuals in each of the six groups. The second set of rows shows the number of individuals for which we have PSQFT data in both our base and endline period, with relative shares of our reduced sample reassuringly close to the population-level figures.

C. Pollution data and Clean Air Act nonattainment areas

We use data on fine particulate matter concentrations ($PM_{2.5}$) at the 1km by 1km resolution from van Donkelaar et al. (2021), which is constructed by combining ground-based measurements, satellite images and chemical transport models. We map the $PM_{2.5}$ data into Census blocks using the closest pollution grid point to the Census block centroid. For analysis, we use the average $PM_{2.5}$ concentrations for 2000-03 and 2016-19, as well as pre-period average concentrations from 1998-1999.

To identify the effect of pollution on house prices, we make use of the 2005 Clean Air Act rules for $PM_{2.5}$, following Sager & Singer (2022) who provide a detailed account of this regulation. We use

the 208 counties from the (EPA 2005) that became regulated in 2005, because they did not meet the necessary threshold of 15 $\mu g/m^3$ for the three-year average of annual mean $PM_{2.5}$ concentrations. These counties were assigned into nonattainment, and were subject to stricter action to reach air pollution standards from the Environmental Protection Agency. Sager & Singer (2022) provide a detailed account of this regulation.

D. Descriptive statistics on housing and pollution disparities

Figure 1 shows how house prices in PSQFT evolved differently for Black, NHW and Other homeowners. Figure 1a shows the evolution of house prices in the raw data, aggregating across individuals for each group. Figure 1b partials out state-by-year and block fixed effects, and normalizes each series to start at zero in 2000. The housing crisis hit Black and Other homeowners particularly hard, consistent with Faber & Ellen (2016). While PSQFT recovered for areas with Other homeowners relative to areas with NHW homeowners, prices hardly recovered for areas with Black homeowners. This resulted in a widening of the gap of house prices between areas with NHW homeowners and areas with Black homeowners. In Appendix Figure A.1, we show that unlike areas with Black homeowners, prices in areas with Black renters rose much more.

Table 1 shows the average values for PSQFT in our base period, end period, as well as changes. Overall, PSQFT increased by $98 from $124 to $222, with considerably lower increases for Black homeowners than for either NHW or Other homeowners. Table 1 also shows that $PM_{2.5}$ concentrations decreased overall by 5μg, falling from 13μg to 8μg. Although the distribution of pollution changes is similar for Black and NHW homeowners, the pollution exposure gap narrowed from 1.2μg between Black and NHW homeowners to 0.4μg driven by larger decreases in pollution faced by Black homeowners, consistent with prior literature (Jbaily et al. 2022, Currie et al. 2020).

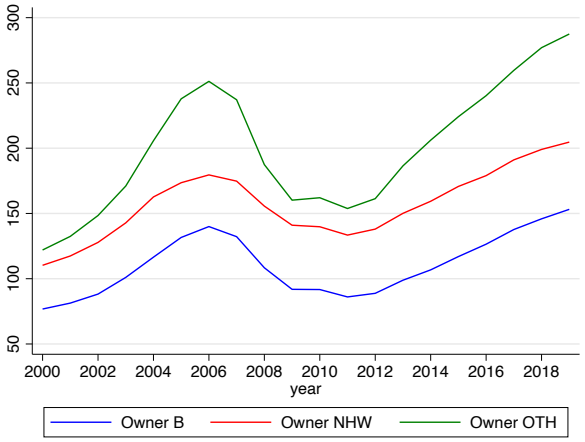
II. Empirical strategy

To formally explore how pollution reductions have affected the home value gap between areas with NHW homeowners and Black homeowners we run long difference regressions where Δ denotes the difference between 2016-19 and 2000-03:

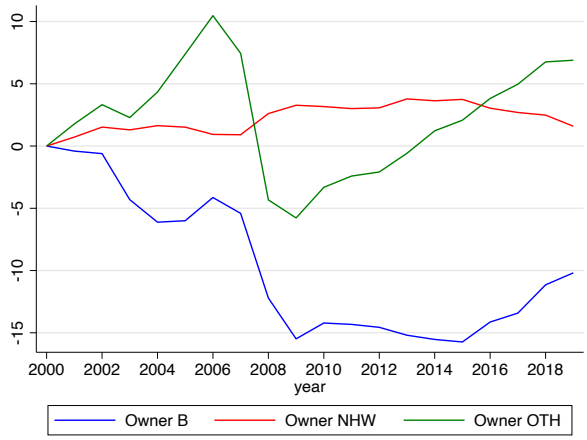
Table 1: Descriptive statistics: Counts and shares of tenancy by race groups, and PSQFT and PM2.5

	Total	Owner			Renter		
		Black	Non-Hisp. White	Other	Black	Non-Hisp. White	Other
US population (continental in 2000)							
Count	271859935	16289056	146176900	24399334	16643977	45081882	23268786
Share	1	0.06	0.54	0.09	0.06	0.17	0.09
US population with PSQFT data in 2000/03 and 2016/19 (continental in 2000)							
Count	130017550	8664097	68668877	14187858	7385241	18858379	12253098
Share	1	0.07	0.53	0.11	0.06	0.15	0.09
PSQFT in 2000/03							
Mean	123	84.8	123.1	142.5	85.4	131.4	136.5
SD	104.1	62.4	92.5	91.4	102.8	141.3	121.9
p5	28.7	15.2	35.5	46.2	12	27.2	33.3
p95	273.4	184.9	262.7	300.9	198.1	321	300.9
PSQFT in 2016/19							
Mean	219.7	142.2	203.3	274	167	246.4	294.1
SD	212.6	139.6	186.5	204.7	188.5	282	246.9
p5	35.7	10.9	52.3	65.9	9.3	35	49.4
p95	565.2	393.9	496.4	636.3	495	715	700.9
Δ PSQFT 2000/03 - 2016/19							
Mean	96.7	57.4	80.2	131.5	81.6	115	157.6
SD	163.5	106.6	137.5	151.4	162.8	221.2	208.1
p5	-21	-37.52	-18.32	-7.67	-34.4	-20.63	-11.16
p95	335.5	240.2	263.7	378.4	344.2	438	470
PM2.5 in 2000/03							
Mean	13	13.7	12.5	14.1	13.8	12.6	14.6
SD	3.5	2.7	3	4.7	2.9	3.3	4.6
p5	8.2	9.1	8	8.4	9.2	7.9	8.6
p95	20.5	17.4	16.7	23	20.4	19	22.8
PM2.5 in 2016/19							
Mean	7.9	8	7.6	8.9	8.1	7.8	9
SD	1.8	1.4	1.5	2.3	1.5	1.7	2.3
p5	5.6	6.2	5.4	6	6.2	5.4	6.1
p95	11.9	10.3	10.1	13.2	11.5	11.2	13.2
Δ PM2.5 2000/03 - 2016/19							
Mean	-5.09	-5.69	-4.92	-5.23	-5.67	-4.82	-5.56
SD	2.3	1.8	2.1	2.9	1.9	2.2	2.8
p5	-8.8	-7.72	-7.52	-10.68	-8.22	-7.87	-10.37
p95	-1.48	-2.05	-1.48	-1.35	-2.1	-1.32	-1.4

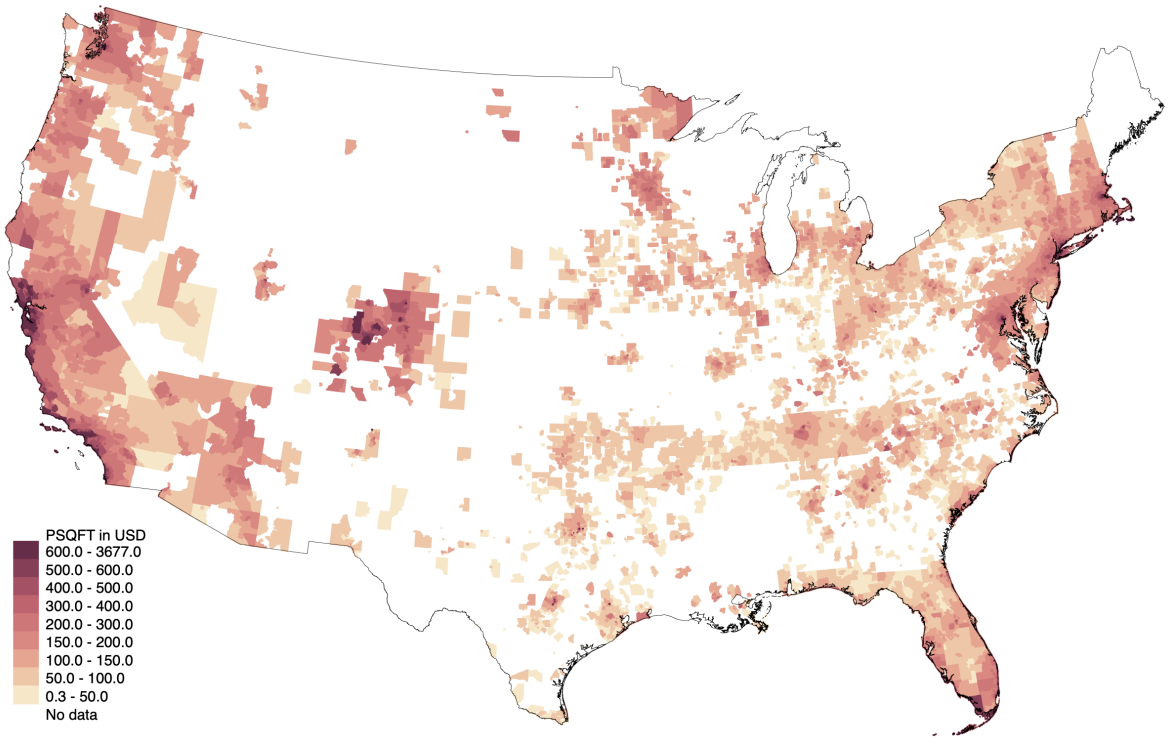
Notes: The table shows summary statistics for the indicated variables. Column "Total" contains overall statistics for the continental US, and the following six columns for each tenure (owner/renter) by race (Black, Non-Hispanic White, Other) group. Changes are indicated by Δ. Prices are in nominal terms. Appendix Table A.1 contains PSQFT summary statistics based on deflated prices (in 2012 $) using a quarterly GDP deflator.



(a) Unconditional PSQFT



(b) PSQFT with state-year FE, block FE and normalized



(c) Spatial distribution of price per square foot aggregated up to Census tracts (2016-2019)

Figure 1: Price per square foot (PSQFT): evolution by race and spatial distribution

Notes: Panel (a) shows the average price per square foot of transacted properties in blocks with homeowners by racial groups (in real 2012 US$). Racial groups are Non-Hispanic White (NHW), Black (B), and other groups (OTH). Panel (b) shows the average price per square foot after partialing out Census block fixed effects and state-by-year fixed effects. Additionally, every group series is normalized to zero in 2000. Therefore the graph shows the increase in price per square foot relative to 2000 for each respective racial group, net of all fixed effects. Since the panel of Census blocks is unbalanced, missing data in a Census block in one year would skew the aggregate away from the time-averaged level of that Census block. Census block fixed effects help address this by demeaning each Census block series. This issue would not arise in a balanced panel. Panel (c) maps the spatial distribution of price per square foot across the contiguous US in 2016-19. We aggregate Census block level information up to Census tract averages for visualization, and only use Census block for which data in 2000-03 is also available. There is little coverage in Texas, as the county recorder's office is not required to report transactions. The maps in Appendix Figures A.2a and A.2b show the variation within Census tracts for New York State and a few counties around New York City, respectively.

$$\Delta \log(\text{PSQFT}_i) = \alpha \Delta \text{PM}_i + \sum_j \left(\beta_j G_i^j + \gamma_j G_i^j \Delta \text{PM}_i \right) + \delta \boldsymbol{X}_i + \xi_i + \epsilon_i \quad (1)$$

The dependent variable $\Delta \log(\text{PSQFT}_i)$ is the long difference of the logged PSQFT in the area where individual i resided in 2000. We denote $\text{PM}_{2.5}$ concentrations by PM_i, and G_i^j is a dummy variable that takes value 1 if individual i is part of group j, where $j \in \{$Black homeowner, Other homeowner$\}$, with NHW homeowner the omitted group. $\boldsymbol{X}_i$ is a vector of controls. This includes baseline pollution and an indicator for whether individual i resides in an urban or rural block to allow for different trends in urban areas.

State fixed effects ξ_i account flexibly for state time trends. Long-differencing addresses unobserved time-invariant factors at the individual level that may be correlated with house prices and pollution, such as baseline amenities. Note that both our outcome and pollution are constructed at the Block level, and then assigned to individuals based on location.[6] This avoids the need to weight regressions by population and the thorny issue of which weights to choose (entire population, Black homeowners, NHW homeowners, Other homeowners), but requires clustering of standard errors, which we implement conservatively at the tract level.

An advantage of our framework, which relies on the construction of block level indices, is that we capture property values for the universe of homeowners by race in blocks with available transaction data, not merely transacted properties themselves, resulting in a more representative sample. We turn to transaction level analysis in our section on mechanisms for decomposing systemic across and direct within-block discrimination.

A. Sorting

Sorting during our sample period could present a measurement and estimation problem by introducing bias. Changes in racial population shares over time as well as changes in house prices are at least partially driven by sorting and as such likely endogenous.[7] To isolate the source of temporal variation that comes from pollution, rather than sorting, we keep the geographical distribution of individuals fixed at the 2000 Census. This limits the threat from bias due to unobservables that are correlated with changes in the spatial distribution of individuals. We use two strategies to as-

[6] We also capture within-block variation in the pollution capitalization rate, because a higher share of homeowners from one racial group will affect the probability of a transaction from a member of that group, which in turn affects estimated disparities in the presence of direct discrimination.

[7] Note that sorting in response to pollution changes is less problematic due to our instrument for pollution described below. Moreover, Table A.2 shows that changes in pollution (relative or absolute) are not significantly correlated with changes in the share of the block population that is Black, with a ten percent increase in pollution increasing the share of Black population by 0.001.

sess the robustness of this approach. First, we repeat our analysis using 2010 Census geographies and counts of homeowners instead of the 2000 Census, allowing for 10 years of sorting.[8] Second, we use the 2020 Census and calculate the change in population shares for each racial group at the block level to focus on those blocks where racial composition changed little. Both tests suggest that, if anything, our results may be too conservative and underestimate the disparity in pollution capitalization rates.

B. Instrumenting air pollution with regulatory nonattainment

Despite the use of fixed effects and long differencing, changes in air pollution are likely to be correlated with changes in unobserved amenities that also impact house prices. For example, changes in economic activity or infrastructure are likely to drive both, pollution and house prices. We therefore use the 2005 $PM_{2.5}$ Clean Air Act regulation that induced changes in pollution in nonattainment counties as an instrument. The identifying assumption is that the regulation only shifted pollution, and no other unobservables correlated with house prices. We follow Sager & Singer (2022) to address bias from underlying trends that differ by baseline pollution and attainment status by including pre-period pollution levels from 1998-99 (PMpre_i) as part of our controls $\boldsymbol{X}_i$.

As Auffhammer et al. (2009) show, nonattainment effects are often stronger in those parts of nonattainment areas that are initially more polluted. To allow for such heterogeneous effects, we additionally interact nonattainment status with pre-period pollution concentrations PMpre_i (see also Bishop et al. (2018)). We include instruments for each term that contains ΔPM_i in Equation 1, for example, for ΔPM_i, the first stage is:

$$\Delta\text{PM}_i = \theta_0 \Delta\text{NA}_i + \sum_j \left(\eta_j G_i^j + \theta_j G_i^j \Delta\text{NA}_i + \rho_j G_i^j \Delta\text{NA}_i \text{PMpre}_i \right) + \tau \boldsymbol{X}_i + \zeta_i + \mu_i \qquad (2)$$

Our set of instruments vary at the Census block level, but we allow for spatial correlation by clustering standard errors at the Census Tract level, which contains an average of 100 blocks.[9] Appendix Table A.3 shows the first stages for the three endogenous variables in our baseline specification. Reassuringly, the exogenous interaction between nonattainment and racial groups affect the corresponding endogenous interactions between change in $PM_{2.5}$ and racial groups.

[8] The detailed data for homeowners is not yet available for the 2020 Census.

[9] This is more conservative than clustering at the block group level in Bishop et al. (2018), who use a similar instrument by interacting nonattainment with historic pollution levels to examine the impacts of pollution on dementia.

III. Results

A. Disparities in pollution capitalization rates

Table 2 shows results from estimating versions of Equation 1. The first two columns omit interactions between pollution changes and race. The OLS results in Column 1 show that areas with Black homeowners had a significantly lower increase in house prices than those with NHW homeowners, and that areas with Other homeowners had slightly higher increases in house prices.[10] Air pollution decreases house prices, but the omitted variable bias is sizeable and positive, when comparing with Column 2. This is consistent with the notion that economic activity is accompanied by beneficial amenities that push up house prices, while simultaneously increasing pollution.

Column 2 shows our results when using the regulatory instruments. A one unit decrease in $PM_{2.5}$ increases house prices by 11.6%.[11] This corresponds to an overall elasticity of -0.87, broadly in line with Sager & Singer (2022) and Bento et al. (2015). Column 2 also shows that house prices in areas with Black homeowners decreased by 11.5% relative to areas with NHW homeowners. This confirms the trends in Figure 1 and Table 1, even after conditioning on fixed effects and controls.

For our main results, we next interact pollution with the indicators for different racial groups in Column 3 for OLS, and Column 4 using our set of instruments. The difference in coefficients between racial groups is statistically significant. The difference between Black and NHW homeowners is also economically significant. Column 4 implies that a one unit decrease in $PM_{2.5}$ increases house prices for Black homeowners by 4.4 percentage points less than for NHW homeowners.[12] This implies that racial house price disparities not only exist in levels, but also in house price changes resulting from plausibly exogenous changes in pollution levels.

Figure 2a visualizes the coefficients and differences in pollution capitalization rates using the estimated coefficients and covariance matrix from Column 4 in Table 2. While a one unit reduction in $PM_{2.5}$ increases house prices by 7.6% for Black homeowners, it increases them by 12.4% for NHW homeowners, a pollution capitalization rate that is 63% larger.[13] Figure 2b provides estimates in absolute terms using the race specific average house prices. For Black homeowners, a one unit decrease in pollution increases PSQFT by \$10.8, while the figure for NHW homeowners is more than double at \$25.2 PSQFT.

[10] We will interchangeably refer to house prices in areas with Black/NHW/Other homeowners and house prices for Black/NHW/Other homeowners, noting that the former is the more precise formulation.

[11] Since the outcome is in logs, the semi-elasticity is calculated as $\exp(0.11) - 1$, and the elasticity is calculated as $-0.11 * 7.9$ using the overall endline period mean of $PM_{2.5}$.

[12] The first stage is strong with a Kleibergen-Paap F statistic of around 199. Appendix Table A.3 shows the first stages for the three endogenous variables of Column 4.

[13] Appendix Table A.4 reports these estimates. All estimated effects in Figure 2a and 2b are statistically different from each other, as seen in Panel (b) of Appendix Table A.4.

Note that the bias in the OLS coefficients operates mainly through the coefficient on overall pollution, which changes substantially from Column 3 (-0.026) to Column 4 (-0.111). There is little, if any, omitted variable bias in the *differential* impact of pollution on house prices on the interaction coefficients, e.g. 0.039 vs. 0.041 for Black homeowners. This implies that the OLS bias that operates through a growth-pollution bundle and affects property prices is similar across racial groups. Therefore, even skeptics regarding our instrumental variables approach should be reassured that our findings on the *disparities* in the capitalization rate of pollution still hold.

To test whether complementary amenities are more unequally distributed in tracts that are more racially segregated, we split our sample by quartile-of-segregation within Census tracts. Figure 2d shows that the disparity in pollution capitalization rates is larger in tracts with more residential segregation, in line with more systemic discrimination in those areas.[14] Appendix A.3 presents more details on the construction of the segregation index and our regression results.

B. Robustness

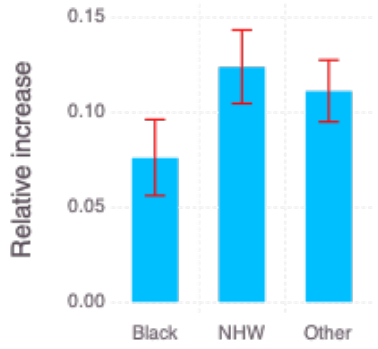
A major concern for the validity of our estimates is that the interaction between air quality improvements and racial groups may capture omitted interactions that are correlated with racial groups. Black communities tend to be poorer and more urban than NHW communities so our results may simply reflect the impact of these factors. Moreover, they also tend to be more polluted at baseline and experienced slightly larger pollution reductions, which could bias our estimates if capitalization effects are nonlinear. We directly address these concerns by adding baseline measures of wealth, income, urban share, baseline or changes in pollution, fully interacted with air quality improvements, and instrument the endogenous interaction with analogous interactions of nonattainment and baseline controls. In particular, we use data from Chetty et al. (2018) on baseline 2000 tract level share of households in poverty, mean household income (in levels and logs), and our data on urban and rural areas and baseline/changing pollution, and interact them in turn with pollution improvements. We visualize all results in Figure 2e, which are based on Appendix Tables as referenced in the Figure Note.

[14] Note that some segregation is required for complementary amenities to have bite, but it is a priori unclear whether disparities should be larger for more intermediate cases of segregation.

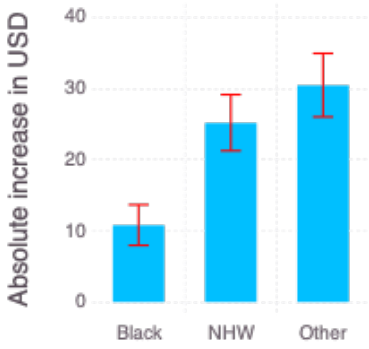
Table 2: The impact of pollution on house prices by race

	Change in (log) PSQFT					
	(1)	(2)	(3)	(4)	(5)	(6)
Black	-0.116***	-0.109***	0.114***	0.134***	0.115***	0.123***
	(0.006)	(0.006)	(0.010)	(0.018)	(0.012)	(0.019)
Other	0.031***	0.033***	0.010***	0.088***	0.029***	0.125***
	(0.002)	(0.002)	(0.003)	(0.008)	(0.006)	(0.010)
Change in PM2.5	-0.024***	-0.110***	-0.024***	-0.117***	-0.026***	-0.119***
	(0.001)	(0.009)	(0.001)	(0.009)	(0.001)	(0.008)
Change in PM2.5 * Black			0.041***	0.044***	0.030***	0.031***
			(0.002)	(0.003)	(0.002)	(0.003)
Change in PM2.5 * Other			-0.004***	0.011***	-0.010***	0.009***
			(0.001)	(0.002)	(0.001)	(0.002)
Owner					-0.010***	-0.013**
					(0.003)	(0.005)
Owner * Black					-0.006	0.009
					(0.010)	(0.015)
Owner * Other					-0.009	-0.025***
					(0.006)	(0.007)
Change in PM2.5 * Owner					0.003***	0.004***
					(0.001)	(0.001)
Change in PM2.5 * Owner * Black					0.010***	0.012***
					(0.002)	(0.003)
Change in PM2.5 * Owner * Other					0.010***	0.006***
					(0.001)	(0.001)
Urban share of block	-0.116***	-0.088***	-0.115***	-0.089***	-0.111***	-0.084***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)
Baseline PM2.5 (98-99)	-0.003***	-0.059***	-0.003***	-0.059***	0.003***	-0.054***
	(0.001)	(0.006)	(0.001)	(0.005)	(0.001)	(0.005)
State by year FE	Yes	Yes	Yes	Yes	Yes	Yes
Estimator	OLS	IV	OLS	IV	OLS	IV
N	91,520,832	91,520,832	91,520,832	91,520,832	130,017,550	130,017,550
R^2	0.194	0.175	0.196	0.176	0.204	0.186
First-stage F (KP)		528.000		198.944		130.368

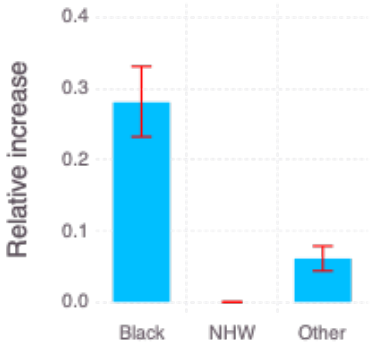
Notes: The table shows regression estimates from our long-differences approach at the individual level using OLS and IV as indicated. Columns 1-4 only use data on individuals in owner occupied housing. Columns 5-6 add renters. Standard errors are clustered at the Census tract level. *** Significant at the 1 percent level, ** significant at the 5 percent level, * significant at the 10 percent level.



(a) Effect of 1-unit decrease in PM2.5

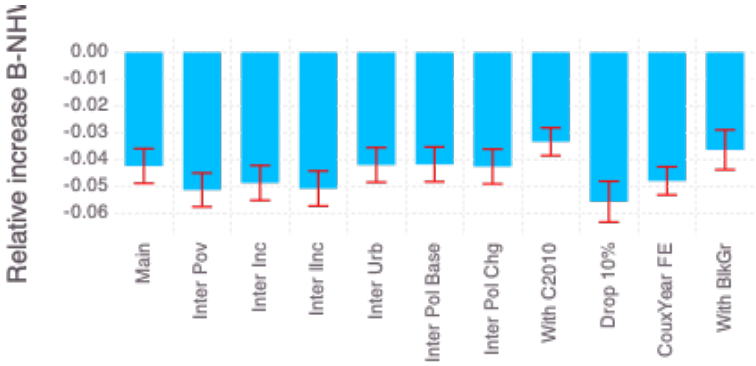


(b) Absolute effect of (a)

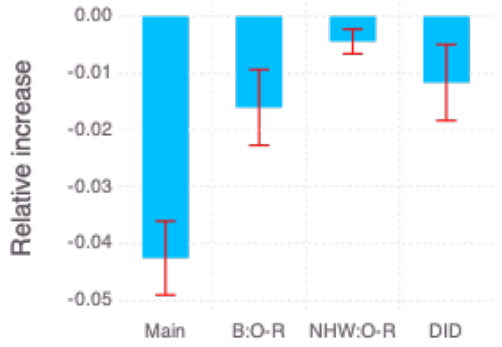


(c) Counterfactual

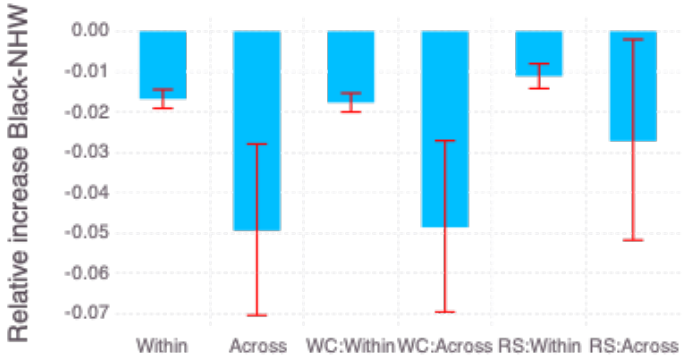
(d) By quartile of segregation



(e) Evidence on robustness



(f) Owners vs. renters and DID



(g) Transaction level: within and across disparities

Figure 2: Effects of air quality improvements on prices per square foot

Notes: The figures visualize the effects of our IV-based regression results by appropriately exponentiating coefficients or combinations of coefficients. Error bars represent 95% confidence intervals based on the respective joint covariance matrices. Black (or B where space is constrained), NHW and Other refer to our three racial groups. Panel (a) reports our main results (Tables 2 and A.4). All panels report the effect of a one unit decrease in PM2.5 on relative increases in house prices with two exceptions. Panel (b) translates the relative increase into absolute US$ increases using the respective house price levels of each racial group (Table A.4). Panel (c) reports relative increases, but from our counterfactual analysis based on NHW capitalization rates but the entire group-specific decreases in PM2.5 over our sample period (Table A.4). Panel (d) shows the disparity in capitalization rates between Black and NHW homeowners (B-NHW) when splitting our sample by quartile-of-segregation (Table A.5). Panel (e) shows the disparity for our main result and then ten robustness checks based on Tables A.6, A.7, A.8 and A.9. Panel (f) shows the disparity for our main result, then the first difference between owners and renters for Black (B:O-R) and NHW (NHW:O-R), and then our difference-in-differences (DID) in disparities (Tables A.10 and A.11). Panel (g) presents our results at the transaction level showing the within-block disparity based on seller race and across-block disparity based on block racial composition, as baseline, with a SQFT control interaction (WC) and as repeat sale analysis (RS) (Table A.12).

Figure 2e shows the difference in pollution capitalization rates between Black and NHW homeowners from a one unit decrease in PM2.5. The first bar shows our baseline effect from our main specification. The next six bars show the difference in capitalization rates when in turn controlling for fully interacted poverty, income, log income, urban share, baseline, or changing pollution. If anything, our estimated effects become slightly larger when introducing the interacted controls, suggesting that our effects by race are not simply capturing omitted socioeconomic characteristics. Tracts with higher poverty rates, lower income, lower urbanicity, lower baseline pollution and larger changes in pollution tend to have slightly higher pollution capitalization rates.

A second concern relates to sorting during our sample period, as discussed in Section II.. To probe whether concurrent sorting may be an issue, we use Census block geographies and counts of homeowners from the 2010 Census instead of the 2000 Census. The estimate "With C2010" in Figure 2e shows that results are very similar to our main specification. Figure 2e also shows that our estimates are similar, maybe even slightly larger, if we omit the 10% (and 20%) of blocks that experienced the largest changes in racial composition between 2000 and 2020 "Drop 10%".[15] Together, they suggest that our results are robust to sorting during our sample period and that, if anything, our baseline approach may be slightly too conservative.

Third, Figure 2e also shows that our results are robust to including county by year fixed effects, instead of state by year fixed effects ("CouxYear FE"). This provides additional reassurance on the validity of our instruments, as we are only exploiting the heterogeneous policy effects within nonattainment counties, rather than the policy effect across attainment and nonattainment counties. This also implies that there are no unobserved confounders that vary across counties and years, and that disparities in pollution capitalization rates are found even within counties.

Finally, our sample only includes homeowners in Census blocks that have house price transactions in the baseline and endline period. As a robustness check, we impute missing PSQFT data for Census blocks using Census block group medians (there are around 10 Census blocks per block group) to increase our sample size. The bar "With BlkGr" in Figure 2e shows that our estimated disparity in the pollution capitalization rate between Black and NHW homeowners is similar for this expanded sample.

[15] We rank observations by the absolute change in the population share for each racial group and omit all observations that are in the top 10% (or 20%) for any racial group. This restricts the sample to blocks with an absolute change in the Black population share to be less than 0.075, compared to an unrestricted 99th percentile at 0.44.

C. *Counterfactual analysis*

While air quality improvements helped to improve PSQFT for all groups, our striking result is that air quality improvements have actually widened the gap of house prices between Black and NHW homeowners despite the shrinking pollution exposure gap due to sizable differences in pollution capitalization rates. This can easily be seen as the reduction in pollution is only 16% greater for Black homeowners (Table 1), but the capitalization rate is 63% greater for NHW homeowners.

We next ask what counterfactual PSQFT would have prevailed if Black and Other homeowners had the same measured pollution reductions during our sample period, but experienced a capitalization rate at the level of NHW homeowners. Figure 2c shows that in this case house prices would be 28% higher for Black homeowners and 6% higher for Other homeowners. These figures translate into a $40 and $17 higher PSQFT in absolute terms for both groups, respectively (Panel (c) of Table A.4), or a counterfactual gain of $293 billion extrapolating to all Black homeowners. In turn, these figures imply a reduction in the Black-NHW homeowner gap from $38.3 to $21.1 PSQFT. In actuality, the gap increased to $61.1 PSQFT.

D. *Mechanisms: Complementary amenities and discrimination*

What drives the difference in pollution capitalization rates between Black and NHW homeowners? Our main analysis is designed to capture the composite of disparities in pollution capitalization rates both within Census blocks and across them. As noted in the introduction, differences across blocks, are likely to reflect differences in amenities that are complementary to clean air, such as green outdoor spaces, playgrounds, sports facilities, crime rates, walkability and other amenities that increase the desire to be outside that vary by racial groups, due to systemic discrimination, even after controlling for income. Within blocks, amenities and sorting are held constant, so disparities in pollution capitalization rates can be viewed as a form of direct discrimination. Two pieces of analysis help us to shed light on the role played by each of these mechanisms.

First, since there are a large number of unobserved amenities, we leverage data on the spatial distribution of landlords with Black and NHW renters, using the analogous 2000 Census data on block level renters by racial group. Comparing capitalization rates for homeowners and landlords with renters from the same racial group differences out complementary amenities that vary by racial groups of residents.[16] We use the analogous comparison from the other racial group to difference out any discrepancy between homeowners and landlords in a difference-in-differences of capitalization rates. The underlying assumption is that, conditional on controls and fixed ef-

[16]This also addresses possible differences in preferences for complementary amenities across groups.

fects and adjusting for average differences in complementary amenities between homeowners and renters across all groups, renters from the same group live in communities with similar complementary amenities. If differences in complementary amenities across racial groups drive all of the disparity in pollution capitalization rates, we would expect this estimate to be zero.

In Columns 5 and 6 of Table 2, we interact pollution with racial group by tenure (homeowner or renter), for OLS and IV respectively. For ease of interpretation, we illustrate the results from Column 6 in Figure 2f. It shows the pollution capitalization rates differenced between homeowners and renters for each racial group. Indeed, for NHW it is the case that capitalization rates for NHW homeowners (12.2%) and landlords that rent out to NHW renters (12.7%) are similar (see also Panel (a) and (b) of Table A.10). There is a much larger difference for Black homeowners (7.5%) and landlords with Black renters (9.2%). The difference-in-differences in capitalization rates is 1.2 ("DID" in Figure 2f, see also Table A.11), statistically significant, and equivalent to a 16% increase over the pollution capitalization rate of Black homeowners. This 16% represents approximately one-quarter of the overall 63% higher pollution capitalization rate for NHW relative to Black homeowners found earlier, suggesting that complementary amenities may explain roughly three-quarters of the disparity.

Second, we examine the role of within-block disparities more directly. We use data on mortgages from HMDA (2022) to retrieve information on the race of the seller at the transaction level, and use the first names and surnames of sellers in each transaction to predict the race of sellers for properties without mortgages.[17] We interact $PM_{2.5}$ with observed (or predicted) seller race as well as with baseline racial composition, and include Census block fixed effects (an area of approximately 200-by-200 meters) and several controls (state or county, urban population share of block, seller race, and racial composition of block) all interacted with year fixed effects. This design captures disparities within blocks through the pollution interaction with seller race and disparities across blocks through the interaction with block racial composition. Figure 2g shows that the within-disparity in the pollution capitalization rate is 1.7 and the across-disparity is 4.9 ("Trans", see also Table A.12).

The core assumption here is that conditional on all controls and fixed effects, there are no other property attributes that are correlated with race that also affect the pollution capitalization rate. Transforming our variable into logs helps to eliminate possible absolute differences in capitalization at different price points, and using PSQFT helps to rule out some nonlinearities in the pollution capitalization rate by property size that may correlate with race. To further minimize concerns

[17] We use Tzioumis (2018), Kaplan (2021), Xie (2022) to predict race according to first name, surname, or jointly. We only assign the race of seller if all three methods predict the same race. More details are presented in Appendix A.6.

about nonlinearities, we control for log square footage, arguably the single most important property attribute after conditioning on fixed effects, fully interacted with pollution and appropriately instrumented, and find very similar results. Finally, to fully rule out concerns that our results are being driven by unobserved property characteristics that correlate with race, we implement a repeat sales approach using property fixed effects. Despite this different approach and a considerable reduction in sample size, Figure 2g shows that both within and across-disparities in the pollution capitalization rate are reassuringly similar to our first strategy using renters, with a decomposition of one-quarter from direct discrimination and three-quarter from complementary amenities.

IV. Conclusion

The environmental justice movement has its roots in the Civil Rights Movement of the 1960s, but its prominence in national priority setting and policy making is much more recent. Indeed, in an effort to address "…the disproportionate health, environmental, and economic impacts that have been borne primarily by communities of color…" President Biden issued Executive Order 14008 aimed at providing 40 percent of the benefits from Federal investments in the environment to marginalized communities. Our analysis underscores the complexity of this effort. Despite improvements in the pollution exposure gap, Black homeowners in the US benefited substantially less from pollution reductions than NHW homeowners. These differential impacts have their roots in both direct and systemic sources of discrimination and highlight the need for research that moves beyond exposure analysis to better understand the marginal damages and benefits from that exposure.[18] They also underscore the inextricable link between various forms of inequality across communities such that policies designed to overcome environmental disparities must also address the unequal access to complementary amenities that help define the impacts of those disparities.

REFERENCES

Aaronson, D., Hartley, D. & Mazumder, B. (2021), 'The effects of the 1930s holc redlining maps', *American Economic Journal: Economic Policy* **13**(4), 355–92.

Akbar, P. A., Li, S., Shertzer, A. & Walsh, R. P. (2020), 'Racial segregation in housing markets and the erosion of black wealth', *NBER Working Paper* .

Ambrose, B. W., Conklin, J. N. & Lopez, L. A. (2021), 'Does borrower and broker race affect the cost of mortgage credit?', *The Review of Financial Studies* **34**(2), 790–826.

[18] While we study the distribution of housing capitalization changes resulting from pollution reductions, this insight is likely important for other realms of public policy such as health or education. (Graff Zivin et al. forthcoming), for example, find heterogeneous marginal pollution damages on health by vaccination status.

Auffhammer, M., Bento, A. M. & Lowe, S. E. (2009), 'Measuring the effects of the clean air act amendments on ambient concentrations: The critical importance of a spatially disaggregated analysis', *Journal of Environmental Economics and Management* **58(1)**, 15–26.

Banzhaf, S., Ma, L. & Timmins, C. (2019), 'Environmental justice: The economics of race, place, and pollution', *Journal of Economic Perspectives* **33**(1), 185–208.

Bayer, P., Casey, M., Ferreira, F. & McMillan, R. (2017), 'Racial and ethnic price differentials in the housing market', *Journal of Urban Economics* **102**, 91–105.

Bayer, P., McMillan, R., Murphy, A. & Timmins, C. (2016), 'A dynamic model of demand for houses and neighborhoods', *Econometrica* **84**(3), 893–942.

Bento, A., Freedman, M. & Lang, C. (2015), 'Who benefits from environmental regulation? evidence from the clean air act amendments', *Review of Economics and Statistics* **97**(3), 610–622.

Bhutta, N., Hizmo, A. & Ringo, D. (2022), 'How much does racial bias affect mortgage lending? evidence from human and algorithmic credit decisions', *FEDS Working Paper* .

Bishop, K. C., Ketcham, J. D. & Kuminoff, N. V. (2018), Hazed and confused: The effect of air pollution on dementia., NBER Working Paper 24970, National Bureau of Economic Research.

Bohren, J. A., Hull, P. & Imas, A. (2022), 'Systemic discrimination: Theory and measurement', *NBER working paper* .

Charles, K. K. & Hurst, E. (2002), 'The transition to home ownership and the black-white wealth gap', *Review of Economics and Statistics* **84**(2), 281–297.

Chay, K. Y. & Greenstone, M. (2005), 'Does air quality matter? evidence from the housing market', *Journal of Political Economy* **113(2)**, 376–424.

Chetty, R., Friedman, J. N., Hendren, N., Jones, M. R. & Porter, S. R. (2018), 'The opportunity atlas: Mapping the childhood roots of social mobility', *NBER working paper* **25147**, https://opportunityinsights.org/wp–content/uploads/2018/12/cty_covariates.dta (accessed on March 4th 2020).

Christensen, P., Sarmiento-Barbieri, I. & Timmins, C. (2022), 'Housing discrimination and the toxics exposure gap in the united states: Evidence from the rental market', *Review of Economics and Statistics* **104**(4), 807–818.

Christensen, P. & Timmins, C. (2021), 'The damages and distortions from discrimination in the rental housing market', *NBER Working Paper* .

Christensen, P. & Timmins, C. (2022), 'Sorting or steering: The effects of housing discrimination on neighborhood choice', *Journal of Political Economy* **130(8)**.

Colmer, J., Hardman, I., Shimshack, J. & Voorheis, J. (2020), 'Disparities in pm2. 5 air pollution in the united states', *Science* **369**(6503), 575–578.

Currie, J., Davis, L., Greenstone, M. & Walker, R. (2015), 'Environmental health risks and housing values: evidence from 1,600 toxic plant openings and closings', *American Economic Review* **105**(2), 678–709.

Currie, J., Voorheis, J. & Walker, R. (2020), What caused racial disparities in particulate exposure to fall? new evidence from the clean air act and satellite-based measures of air quality, NBER Working Paper 24970, National Bureau of Economic Research.

Doleac, J. L. & Stein, L. C. (2013), 'The visible hand: Race and online market outcomes', *The Economic Journal* **123**(572), F469–F492.

EPA (2005), 'Air quality designations and classifications for the fine particles (pm2.5) national ambient air quality standards; final rule', *Federal Register* **70**(3).

Faber, J. W. & Ellen, I. G. (2016), 'Race and the housing cycle: Differences in home equity trends among long-term homeowners', *Housing Policy Debate* **26**(3), 456–473.

FRED (2022), *Gross Domestic Product: Implicit Price Deflator*, Federal Reserve Bank of St Louis. https://fred.stlouisfed.org/series/GDPDEF (accessed on August 4th 2022).

Garg, T., Gennaioli, C., Lovo, S. & Singer, G. (2022), 'Can competition reduce conflict?', *CEGA Working Paper Series* **WPS-197**.

Graff Zivin, J. & Neidell, M. (2012), 'The impact of pollution on worker productivity', *American Economic Review* **102**(7), 3652–73.

Graff Zivin, J., Neidell, M. J., Sanders, N. J. & Singer, G. (forthcoming), 'When externalities collide: Influenza and pollution', *American Economic Journal: Applied Economics* .

Grainger, C. A. (2012), 'The distributional effects of pollution regulations: Do renters fully pay for cleaner air?', *Journal of Public Economics* **96**, 840–852.

HMDA (2022), 'Loan/application register code sheets', *Home Mortgage Disclosure Act* .

Hsiang, S., Oliva, P. & Walker, R. (2019), 'The distribution of environmental damages', *Review of Environmental Economics and Policy* **13**(1), 83–103.

Jbaily, A., Zhou, X., Liu, J., Lee, T.-H., Kamareddine, L., Verguet, S. & Dominici, F. (2022), 'Air pollution exposure disparities across us population and income groups', *Nature* **601**(7892), 228–233.

Kahn, M. E. (2021), 'Racial and ethnic differences in the financial returns to home purchases from 2007 to 2020', *NBER Working Paper* .

Kaplan, J. (2021), *predictrace: Predict the Race and Gender of a Given Name Using Census and Social Security Administration Data*. R package version 2.0.0.
URL: *https://CRAN.R-project.org/package=predictrace*

Kermani, A. & Wong, F. (2021), 'Racial disparities in housing returns', *NBER Working Paper* .

List, J. A. (2004), 'The nature and extent of discrimination in the marketplace: Evidence from the field', *The Quarterly Journal of Economics* **119**(1), 49–89.

Munnell, A. H., Tootell, G. M., Browne, L. E. & McEneaney, J. (1996), 'Mortgage lending in boston: Interpreting hmda data', *The American Economic Review* **86**(1), 25.

Myers, C. K. (2004), 'Discrimination and neighborhood effects: Understanding racial differentials in us housing prices', *Journal of urban economics* **56**(2), 279–302.

Perry, A., Rothwell, J. & Harshbarger, D. (2018), 'The devaluation of assets in black neighborhoods', *Library Catalog: www.brookings.edu* .

Phelps, E. S. (1972), 'The statistical theory of racism and sexism', *The american economic review* **62**(4), 659–661.

Reardon, S. F. & Firebaugh, G. (2002), 'Measures of multigroup segregation', *Sociological methodology* **32**(1), 33–67.

Sager, L. & Singer, G. (2022), 'Clean identification? the effects of the clean air act on air pollution, exposure disparities and house prices', *Grantham Research Institute Working Paper* **376**.

Tzioumis, K. (2018), 'Demographic aspects of first names', *Scientific data* **5**(1), 1–9.

US District Court (2022), 'Connolly et al v. lanham et al', *Complaint and Jury Demand* **1:2022cv02048**.

van Donkelaar, A., Hammer, M. S., Bindle, L., Brauer, M., Brook, J. R., Garay, M. J., Hsu, N. C., Kalashnikova, O. V., Kahn, R. A., Lee, C. et al. (2021), 'Monthly global estimates of fine particulate matter and their uncertainty', *Environmental Science & Technology* **55**(22), 15287–15300.

Xie, F. (2022), 'rethnicity: An r package for predicting ethnicity from names', *SoftwareX* **17**, 100965.

Zillow (2020), *Transaction and Assessment Dataset (ZTRAX)*, Zillow. http://www.zillow.com/ztrax (accessed on November 20th 2020).

APPENDIX FOR ONLINE PUBLICATION

Disparities in Pollution Capitalization Rates

by Josh Graff Zivin[1,*], Gregor Singer[2,*]

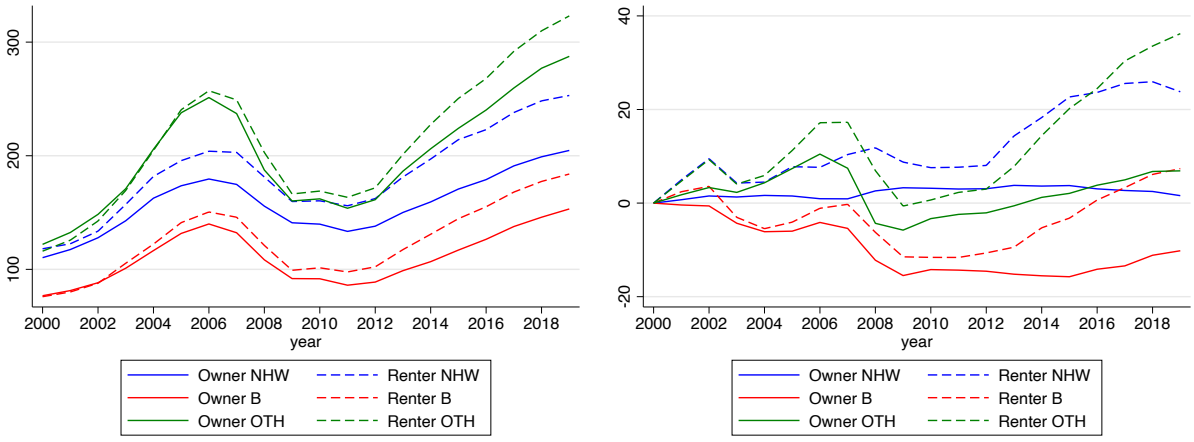
[1]UCSD and NBER

[2]London School of Economics

[*]Correspondence to: Josh Graff Zivin: jgraffzivin@ucsd.edu, Gregor Singer: g.a.singer@lse.ac.uk

Sections A.1 to A.6 show additional results for our main paper on page A-1 to A-14. The longer Section A.7 (page A-15 to A-54) provides details and descriptive statistics for constructing our price per square foot (PSQFT) variable from the Zillow data.

A.1 Additional Descriptive Statistics & Graphs

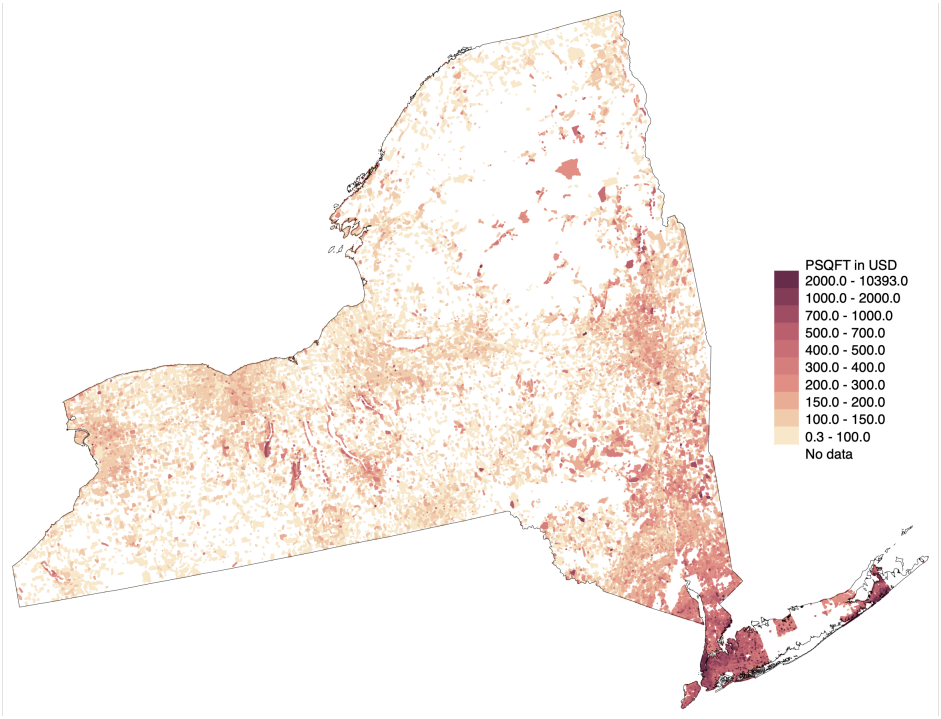


(a) Unconditional

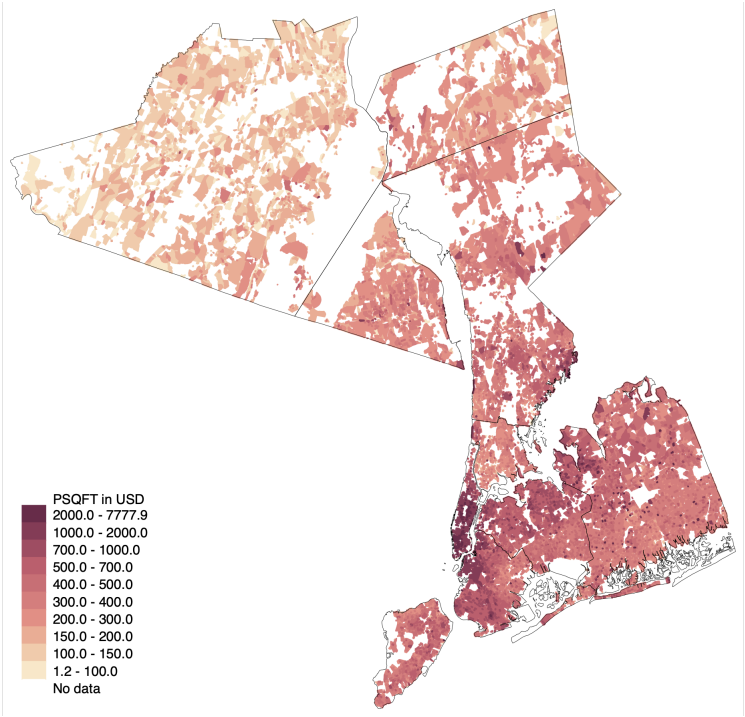
(b) Conditional on state-year FE and normalized

Figure A.1: Price per square foot: unconditional and conditional evolution by tenancy and race

Notes: Panel (a) shows the average price per square foot of transacted properties in blocks with homeowners by tenancy-race group (in real 2012 US$). Tenancy groups are owner and renter and racial groups are non-Hispanic White (NHW), Black (B), and other groups (OTH). Panel (b) shows the average price per square foot after partialing out Census block fixed effects and state-by-year fixed effects. Additionally, every tenancy-race series is normalized to zero in 2000. Therefore the graph shows the increase in price per square foot relative to 2000 for each respective tenancy-race group, net of all fixed effects.



(a) New York State



(b) Selected nine counties around New York City

Figure A.2: Spatial distribution of price per square foot in New York State

Notes: Both panels show the spatial distribution of PSQFT across Census blocks. Panel (a) shows New York State. Panel (b) shows the counties of Bronx, Kings, Nassau, New York, Orange, Putnam, Queens, Richmond, Rockland, and Westchester within New York State

Table A.1: Descriptive statistics: PSQFT deflated by quarterly GDP deflator

	Total	Owner			Renter		
		Black	Non-Hisp. White	Other	Black	Non-Hisp. White	Other
PSQFT in 2000/03							
Mean	152.7	105.4	153	177.2	105.7	162.8	169.4
SD	123	76.9	110	110.8	120.8	162.4	144.7
p5	35.4	18.7	44	57.5	14.7	33.7	41
p95	338.9	229.3	326.1	373.3	245.1	398	372.5
PSQFT in 2016/19							
Mean	201.8	130.5	186.7	251.6	153.7	226.6	270.4
SD	196.3	128.4	171.7	188.1	174.4	261.8	228.6
p5	32.8	10.1	48	60.5	8.6	32.2	45.6
p95	519.2	361.3	455.9	583.6	455	657	643.6
Δ PSQFT 2000/03 - 2016/19							
Mean	49.1	25.2	33.7	74.4	48	63.7	101.1
SD	142.9	93.3	117.5	128.4	150.8	195.2	190.8
p5	-53.47	-59.48	-55.09	-43.96	-55.15	-52.07	-44.13
p95	240.5	173.5	175.7	271.3	271.4	318.6	365.9

Notes: The table shows summary statistics. It replicates part of Table 1, but additionally deflates all prices by the quarterly GDP deflator from FRED (2022).

Table A.2: Change in neighborhood composition and pollution changes

	(1)	(2)	(3)	(4)
Change in PM2.5	.00065	.001		
	(.00063)	(.00074)		
Log change in PM2.5			.0092	.012
			(.0073)	(.0099)
N	6731080	3118264	6731080	3118264

Notes: The table shows regressions where the dependent variable is the change in the share of Black people in a given Census block between 2000 and 2020. All regressions include state fixed effects. The independent variable is based on the change from 2000 to 2020 of block level pollution concentrations. Column (2) and (4) exclude Census blocks with zero change in the dependent variable. Standard errors in parentheses are clustered at the county level.

A.2 Additional Results – Main Analysis

Table A.3: First stage regressions

	Change in PM2.5	Change in PM2.5 * Black	Change in PM2.5 * Other
	(1)	(2)	(3)
Nonattainment	0.356***	-0.361***	-0.553***
	(0.115)	(0.024)	(0.025)
Nonattainment * Black	0.875***	4.295***	0.138***
	(0.083)	(0.099)	(0.014)
Nonattainment * Other	1.433***	-0.032***	5.702***
	(0.102)	(0.009)	(0.130)
Nonattainment * Baseline PM2.5	-0.061***	0.035***	0.051***
	(0.008)	(0.002)	(0.002)
Nonattainment * Baseline PM2.5 * Black	-0.056***	-0.433***	-0.011***
	(0.006)	(0.006)	(0.001)
Nonattainment * Baseline PM2.5 * Other	-0.122***	0.006***	-0.617***
	(0.007)	(0.001)	(0.009)
Black	0.062***	-4.175***	0.038***
	(0.009)	(0.029)	(0.003)
Other	0.285***	-0.048***	-2.900***
	(0.009)	(0.002)	(0.016)
Urban share of block	0.312***	0.004*	0.018***
	(0.010)	(0.002)	(0.003)
Baseline PM2.5 (98-99)	-0.538***	-0.045***	-0.075***
	(0.005)	(0.002)	(0.002)
State by year FE	Yes	Yes	Yes
Estimator	OLS	OLS	OLS
N	91,520,832	91,520,832	91,520,832
R^2	0.861	0.956	0.938
First stage F (KP)	198.94	198.94	198.94

Notes: The table shows the first stage regressions with the indicated endogenous variable as dependent variable. The first stage regressions correspond to Column 4 in Table 2. Standard errors are clustered at the Census tract level. *** Significant at the 1 percent level, ** significant at the 5 percent level, * significant at the 10 percent level.

Table A.4: Interpretation of coefficients and counterfactuals from Table 2 Column (4).

	Black	Non-Hisp. White	Other
Panel (a): PSQFT increase from one point decrease in PM2.5			
Relative	0.076***	0.124***	0.111***
Absolute	10.803***	25.176***	30.454***
Panel (b): Difference in relative capitalization rates from Panel (a) in percentage points			
Black		-4.4***	-3.3***
Non-Hisp. White			1.1***
Panel (c): Counterfactual increase with NHW rate but actual decrease in PM2.5			
Relative	0.281***	0.0	0.061***
Absolute	39.917***	0.0	16.764***

Notes: The table shows appropriately exponentiated (combinations of) coefficients based on the estimated coefficients and covariance matrix from Column 4 in Table 2, using only individuals in owner-occupied homes. *** Significant at the 1 percent level, ** significant at the 5 percent level, * significant at the 10 percent level.

A.3 Additional Results – Segregation

To construct our measure of residential segregation, we use information on racial composition at the Census block level and calculate an index of segregation at the Census tract level that measures how uniformly residents of different races are mixing across Census blocks within Census tracts, following Reardon & Firebaugh (2002), Garg et al. (2022). The index is low if all blocks contain a similar proportion of racial groups, and the index is high if members of racial groups live in different blocks. Figure A.3 maps the calculated segregation index across the US for which price per square foot data in 2000-03 and 2016-19 is also available. Table A.5 reports regression results when we split our sample by quartile-of-segregation.

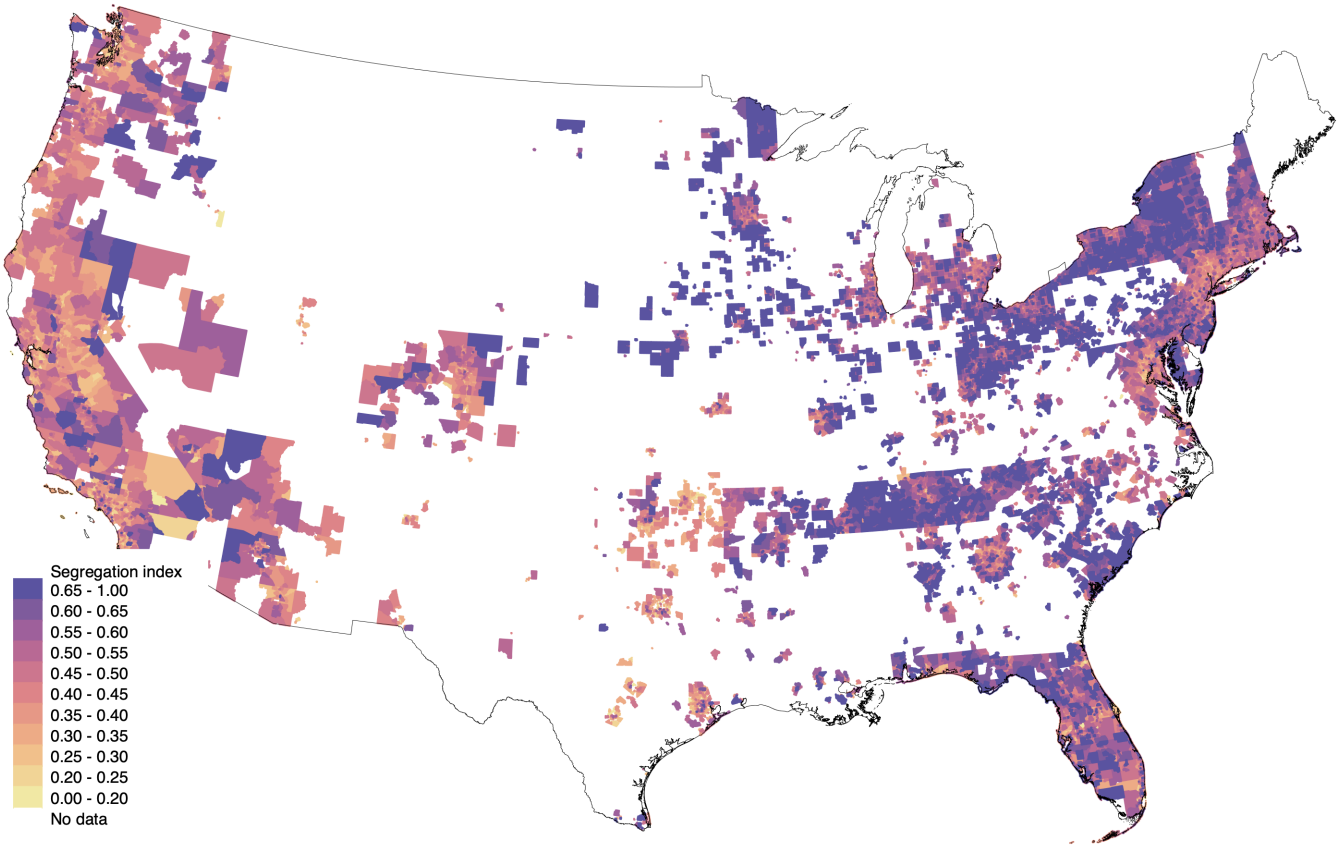


Figure A.3: Spatial distribution of Census tract segregation in 2000

Notes: The map the spatial distribution of segregation within Census tracts across the contiguous US in 2000. We only use Census tracts for which price per square foot data for Census blocks in 2000-03 and 2016-19 is also available.

Table A.5: Splitting sample by quartiles of racial segregation

	Change in (log) PSQFT			
	(1)	(2)	(3)	(4)
Black	0.006	0.099***	0.191***	0.264***
	(0.014)	(0.032)	(0.048)	(0.058)
Other	0.049***	0.074***	0.108***	0.033
	(0.006)	(0.014)	(0.028)	(0.040)
Change in PM2.5	-0.071***	-0.096***	-0.147***	-0.221***
	(0.007)	(0.015)	(0.030)	(0.055)
Change in PM2.5 * Black	0.004	0.031***	0.062***	0.081***
	(0.003)	(0.006)	(0.009)	(0.011)
Change in PM2.5 * Other	0.003***	0.010***	0.021***	0.001
	(0.001)	(0.003)	(0.006)	(0.008)
Urban share of block	-0.008	-0.056***	-0.073***	-0.090***
	(0.022)	(0.015)	(0.012)	(0.015)
Baseline PM2.5 (98-99)	-0.039***	-0.043***	-0.064***	-0.114***
	(0.005)	(0.009)	(0.015)	(0.027)
State by year FE	Yes	Yes	Yes	Yes
Estimator	IV	IV	IV	IV
N	22,878,314	22,880,613	22,878,710	22,882,885
R^2	0.266	0.246	0.180	0.090
First-stage F (KP)	107.363	64.465	33.813	17.200

Notes: The table shows regression estimates from our long-differences approach at the individual level using IV, and splitting our sample by quartile-of-segregation. The Columns 1-4 indicate the quartile. Standard errors are clustered at the Census tract level. *** Significant at the 1 percent level, ** significant at the 5 percent level, * significant at the 10 percent level.

A.4 Additional Results – Robustness

Table A.6: Fully interacting and instrumenting baseline controls with PM2.5: poverty and income

	Change in (log) PSQFT					
	Inter. w. poverty		Inter. w. income		Inter. w. log income	
	(1)	(2)	(3)	(4)	(5)	(6)
Black	0.116***	0.161***	0.115***	0.165***	0.115***	0.174***
	(0.010)	(0.018)	(0.010)	(0.019)	(0.010)	(0.019)
Other	0.017***	0.130***	0.010***	0.133***	0.011***	0.148***
	(0.003)	(0.009)	(0.003)	(0.010)	(0.003)	(0.010)
Change in PM2.5	-0.001	-0.083***	-0.037***	-0.272***	-0.199***	-1.784***
	(0.003)	(0.009)	(0.003)	(0.018)	(0.036)	(0.139)
Change in PM2.5 * Black	0.045***	0.053***	0.041***	0.050***	0.041***	0.052***
	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.004)
Change in PM2.5 * Other	-0.001**	0.021***	-0.004***	0.021***	-0.004***	0.024***
	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)
Change in PM2.5 * Share Poor 2000	-0.184***	-0.408***				
	(0.016)	(0.033)				
Change in PM2.5 * HH inc. 2000			0.001***	0.011***		
			(0.000)	(0.001)		
Change in PM2.5 * Log HH inc. 2000					0.015***	0.140***
					(0.003)	(0.011)
Share Poor 2000	-0.280***	-1.351***				
	(0.080)	(0.166)				
HH inc. 2000			0.007***	0.052***		
			(0.002)	(0.005)		
Log HH inc. 2000					0.059***	0.643***
					(0.016)	(0.056)
Urban share of block	-0.115***	-0.085***	-0.116***	-0.074***	-0.114***	-0.068***
	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)	(0.006)
Baseline PM2.5 (98-99)	-0.006***	-0.074***	-0.003**	-0.094***	-0.003***	-0.105***
	(0.001)	(0.006)	(0.001)	(0.006)	(0.001)	(0.007)
State by year FE	Yes	Yes	Yes	Yes	Yes	Yes
Estimator	OLS	IV	OLS	IV	OLS	IV
N	91,520,832	91,520,832	91,520,832	91,520,832	91,520,832	91,520,832
R^2	0.199	0.168	0.196	0.140	0.196	0.123
First-stage F (KP)		12.645		3.098		3.187

Notes: The table shows regression estimates from our long-differences approach at the individual level using OLS and IV as indicated. Compared to our main Table 2, we additionally interact the indicated variable in the column heading with pollution (appropriately instrumented). Standard errors are clustered at the Census tract level. *** Significant at the 1 percent level, ** significant at the 5 percent level, * significant at the 10 percent level.

Table A.7: Fully interacting and instrumenting baseline controls with PM2.5: urban share and initial pollution

	Change in (log) PSQFT					
	Inter. w. urban share		Inter. w. base. pollution		Inter. w. Δ pollution	
	(1)	(2)	(3)	(4)	(5)	(6)
Black	0.109***	0.132***	0.116***	0.133***	0.115***	0.137***
	(0.010)	(0.018)	(0.010)	(0.018)	(0.010)	(0.018)
Other	0.006*	0.085***	0.013***	0.094***	0.014***	0.092***
	(0.003)	(0.008)	(0.003)	(0.009)	(0.003)	(0.008)
Change in PM2.5	-0.039***	-0.125***	-0.015***	-0.180***	-0.011***	-0.181***
	(0.002)	(0.012)	(0.003)	(0.016)	(0.003)	(0.015)
Change in PM2.5 * Black	0.040***	0.043***	0.041***	0.043***	0.041***	0.044***
	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)
Change in PM2.5 * Other	-0.005***	0.011***	-0.003***	0.012***	-0.003***	0.012***
	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)
Change in PM2.5 * Urban share	0.016***	0.010**				
	(0.002)	(0.005)				
Change in PM2.5 * Baseline PM2.5 (98-99)			-0.001***	0.003***		
			(0.000)	(0.000)		
Change in PM2.5 * Change in PM2.5					0.001***	-0.004***
					(0.000)	(0.001)
Urban share of block	-0.039***	-0.042	-0.115***	-0.083***	-0.115***	-0.084***
	(0.011)	(0.026)	(0.005)	(0.006)	(0.005)	(0.006)
Baseline PM2.5 (98-99)	-0.003***	-0.058***	-0.005***	-0.063***	-0.002**	-0.073***
	(0.001)	(0.006)	(0.001)	(0.006)	(0.001)	(0.006)
State by year FE	Yes	Yes	Yes	Yes	Yes	Yes
Estimator	OLS	IV	OLS	IV	OLS	IV
N	91,520,832	91,520,832	91,520,832	91,520,832	91,520,832	91,520,832
R^2	0.196	0.177	0.196	0.163	0.196	0.164
First-stage F (KP)		5.073		8.725		0.310

Notes: The table shows regression estimates from our long-differences approach at the individual level using OLS and IV as indicated. Compared to our main Table 2, we additionally interact the indicated variable in the column heading with pollution (appropriately instrumented). Standard errors are clustered at the Census tract level. *** Significant at the 1 percent level, ** significant at the 5 percent level, * significant at the 10 percent level.

Table A.8: Using 2010 Census homeowners, or dropping top 10% or 20% of sorters

	Change in (log) PSQFT					
	Using 2010 Census		Drop top 10% of sorters		Drop top 20% of sorters	
	(1)	(2)	(3)	(4)	(5)	(6)
Black	0.083***	0.081***	0.077***	0.102***	0.086***	0.114***
	(0.008)	(0.015)	(0.010)	(0.022)	(0.011)	(0.024)
Other	-0.007**	0.065***	0.012***	0.077***	0.015***	0.087***
	(0.003)	(0.008)	(0.003)	(0.007)	(0.004)	(0.008)
Change in PM2.5	-0.027***	-0.116***	-0.024***	-0.093***	-0.023***	-0.098***
	(0.001)	(0.008)	(0.001)	(0.008)	(0.001)	(0.008)
Change in PM2.5 * Black	0.035***	0.034***	0.054***	0.058***	0.057***	0.061***
	(0.002)	(0.003)	(0.002)	(0.004)	(0.002)	(0.004)
Change in PM2.5 * Other	-0.004***	0.010***	-0.003***	0.010***	-0.003***	0.012***
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)
Urban share of block	-0.081***	-0.063***	-0.094***	-0.076***	-0.097***	-0.077***
	(0.011)	(0.009)	(0.005)	(0.005)	(0.005)	(0.005)
Baseline PM2.5 (98-99)	-0.006***	-0.059***	-0.001	-0.043***	0.001	-0.045***
	(0.001)	(0.005)	(0.001)	(0.005)	(0.001)	(0.005)
State by year FE	Yes	Yes	Yes	Yes	Yes	Yes
Estimator	OLS	IV	OLS	IV	OLS	IV
N	91,994,114	91,994,114	66,131,721	66,131,721	60,036,573	60,036,573
R^2	0.164	0.145	0.187	0.173	0.188	0.173
First-stage F (KP)		236.547		183.460		176.239

Notes: The table shows regression estimates from our long-differences approach at the individual level using OLS and IV as indicated. The column headings indicate the differences to our main Table 2 to test robustness. Standard errors are clustered at the Census tract level. *** Significant at the 1 percent level, ** significant at the 5 percent level, * significant at the 10 percent level.

Table A.9: County by year fixed effects, block group averages for missing outcomes, or predeflated prices

	Change in (log) PSQFT					
	County by year FE		Using block groups		Predeflated	
	(1)	(2)	(3)	(4)	(5)	(6)
Black	0.060***	0.147***	0.070***	0.085***	0.114***	0.133***
	(0.008)	(0.014)	(0.013)	(0.021)	(0.010)	(0.018)
Other	-0.015***	-0.000	-0.017**	0.043***	0.010***	0.085***
	(0.002)	(0.003)	(0.007)	(0.013)	(0.003)	(0.008)
Change in PM2.5	-0.008***	-0.034***	-0.025***	-0.103***	-0.024***	-0.113***
	(0.002)	(0.006)	(0.001)	(0.013)	(0.001)	(0.009)
Change in PM2.5 * Black	0.034***	0.049***	0.035***	0.037***	0.040***	0.043***
	(0.002)	(0.003)	(0.002)	(0.004)	(0.002)	(0.003)
Change in PM2.5 * Other	-0.001***	0.001*	-0.009***	0.003	-0.004***	0.011***
	(0.000)	(0.001)	(0.001)	(0.003)	(0.001)	(0.002)
Urban share of block	-0.139***	-0.133***	-0.136***	-0.113***	-0.120***	-0.095***
	(0.004)	(0.004)	(0.005)	(0.006)	(0.005)	(0.005)
Baseline PM2.5 (98-99)	0.006***	-0.008**	-0.007***	-0.053***	-0.003**	-0.056***
	(0.001)	(0.003)	(0.001)	(0.008)	(0.001)	(0.005)
County by year FE	Yes	Yes				
State by year FE			Yes	Yes	Yes	Yes
Estimator	OLS	IV	OLS	IV	OLS	IV
N	91,520,828	91,520,828	126,813,895	126,813,895	91,536,527	91,536,527
R^2	0.279	0.278	0.157	0.145	0.197	0.178
First-stage F (KP)		226.344		222.106		198.957

Notes: The table shows regression estimates from our long-differences approach at the individual level using OLS and IV as indicated. The column headings indicate the differences to our main Table 2 to test robustness. There are slightly more observations when we predeflate because we drop fewer outliers. Standard errors are clustered at the Census tract level. *** Significant at the 1 percent level, ** significant at the 5 percent level, * significant at the 10 percent level.

A.5 Additional Results – With Data on Renters

Table A.10: Interpretation of coefficients and counterfactuals from Table 2 Column (6).

	Owner & Black	Renter & Black	Owner & NHW	Renter & NHW	Owner & Other	Renter & Other
Panel (a): PSQFT increase from one point decrease in PM2.5						
Relative	0.075***	0.092***	0.122***	0.127***	0.105***	0.117***
Absolute	10.595***	15.375***	24.785***	31.267***	28.737***	34.285***
Panel (b): Difference in relative capitalization rates from Panel (a) in percentage points						
*Owner * Black*		-1.6***	-4.4***	-4.9***	-2.8***	-3.9***
*Renter * Black*			-2.7***	-3.2***	-1.2***	-2.2***
*Owner * Non-Hisp. White*				-0.4***	1.5***	0.5**
*Renter * Non-Hisp. White*					2.0***	0.9***
*Owner * Other*						-1.1***
Panel (c): Counterfactual increase with NHW renter rate but actual decrease in PM2.5						
Relative	0.311***	0.195***	0.022***	0.0	0.109***	0.053***
Absolute	44.21***	32.559***	4.498***	0.0	29.817***	15.474***

Notes: The table shows appropriately exponentiated (combinations of) coefficients based on the estimated coefficients and covariance matrix from Column 6 in Table 2, using individuals in owner-occupied homes as well as renters. *** Significant at the 1 percent level, ** significant at the 5 percent level, * significant at the 10 percent level.

Table A.11: Difference-in-differences in pollution capitalization rates

	Non-Hisp. White	Other
Black	-1.2***	-0.6
Non-Hisp. White		0.6***

Notes: The table shows our difference-in-differences in pollution capitalization rates between racial groups after differencing owners minus renters from the same racial group. The estimate is appropriately exponentiated and based on the estimated coefficients and covariance matrix from Column 6 in Table 2, using individuals in owner-occupied homes as well as renters. *** Significant at the 1 percent level, ** significant at the 5 percent level, * significant at the 10 percent level.

A.6 Additional Results – Transaction level

We obtain information on the race of the seller at the transaction level through two ways. First, we use data based on the Home Mortgage Disclosure Act (HMDA 2022) that requires financial

institutions to report data on mortgages including information on race. This data only provides information for buyers or refinances, but not for sellers. To match the HMDA data to sellers we first identify all transactions in the Zillow data where the seller name matches the buyer name of a previous transaction on the same property. We then use the buyers/refinancing of these previous transactions to obtain the race of the sellers. To match to HMDA loan level information, we use information that is contained in both datasets, specifically, the amount of the loan, Census tract, year, and name of the financial institution that provided the loan. We only keep instances with unique matches for these four variables.[19]

Since HMDA data only provides us with seller race for those cases where sellers previously used mortgages and where we find unique matches, we complement this information by predicting race based on the name of sellers. We use three algorithms, one for first names, one for surnames, and one for first and surnames jointly. Prediction for first names is based on Tzioumis (2018), who use confidential HMDA data and two further confidential lender data sets from different years, specifically designed to create a correspondence between first names and race (e.g. surnames are often shared between partners). Prediction for surnames are based on Census data providing probabilities of race by surnames (Kaplan 2021). Joint prediction using both first names and surnames is based on (Xie 2022), who uses a neural network and Florida voter registration data with a focus on minority groups. We classify the race of a seller if all three methods predict the same race.[20] With both methods, we are able to obtain seller race information for around 22 million transactions (around one-third of all transactions).

Table A.12 reports our transaction level results that we visualize in the main paper by taking the appropriate exponential of the coefficients. For the smaller sample were we have both observed and predicted information on race, we can run separate regressions using only observed or only predicted race. Reassuringly, the coefficients from these regressions are very similar, but insignificant and much noisier due to the smaller sample cut. Recall that this selects a sample of homesellers that previously had a mortgage, otherwise they would not show up in the HMDA data.

[19] We use fuzzy string matching to match names with a high threshold for matches, and verified that results are similar with exact matching.

[20] We verified that we get similar results when additionally restricting classification to those cases where all three methods predict the same race with a very high probability.

Table A.12: Using transaction level data with observed or predicted race of seller

	(log) PSQFT					
	Main trans. result		With PSQFT control (WC)		Repeat sales (RS)	
	(1)	(2)	(3)	(4)	(5)	(6)
PM2.5	-0.018***	-0.096***	0.109***	-0.089***	-0.017***	-0.095***
	(0.001)	(0.008)	(0.003)	(0.019)	(0.001)	(0.008)
PM2.5 * Black seller	0.012***	0.017***	0.012***	0.018***	0.007***	0.011***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
PM2.5 * Other seller	0.001***	0.004***	0.001***	0.004***	0.001***	0.003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
PM2.5 * Black share	0.041***	0.051***	0.037***	0.050***	0.022***	0.028**
	(0.004)	(0.011)	(0.004)	(0.011)	(0.005)	(0.013)
PM2.5 * Other share	-0.018***	-0.003	-0.020***	0.003	-0.016***	0.000
	(0.002)	(0.005)	(0.002)	(0.005)	(0.002)	(0.005)
PM2.5 * log(SQFT)			-0.017***	-0.003		
			(0.000)	(0.002)		
log(SQFT)			-0.268***	-0.405***		
			(0.005)	(0.021)		
Census block FE	Yes	Yes	Yes	Yes		
Property FE					Yes	Yes
State by year FE	Yes	Yes	Yes	Yes	Yes	Yes
Black seller by year FE	Yes	Yes	Yes	Yes	Yes	Yes
Other seller by year FE	Yes	Yes	Yes	Yes	Yes	Yes
Baseline PM2.5 (98-99) by year slopes	Yes	Yes	Yes	Yes	Yes	Yes
Urban share by year slopes	Yes	Yes	Yes	Yes	Yes	Yes
Black share by year slopes	Yes	Yes	Yes	Yes	Yes	Yes
Other share by year slopes	Yes	Yes	Yes	Yes	Yes	Yes
Estimator	OLS	IV	OLS	IV	OLS	IV
N	22,243,376	22,243,376	22,243,376	22,243,376	9,175,231	9,175,231
R^2	0.727	0.726	0.750	0.747	0.878	0.877
First-stage F (KP)		102.193		13.527		95.264

Notes: The table shows regression estimates from our transaction level approach using OLS and IV as indicated. Columns 1-2 present our main transaction level results. Columns 3-4 add log of PSQFT fully interacted with pollution (and appropriately instrumented). Columns 5-6 use a repeat sales design by including property level fixed effects. Standard errors are clustered at the Census tract level. *** Significant at the 1 percent level, ** significant at the 5 percent level, * significant at the 10 percent level.

A.7 Details of constructing price per square foot data from Zillow

This section documents how we construct the transaction level prices per square foot. This has three building blocks which we discuss in turn: (A) identifying arm's length transactions for residential properties from the raw data, (B) identifying a property's location, and (C) identifying the property's area in square footage. We provide some descriptive statistics in part (D). Overall, we have both square footage and coordinates for 44,799,731 (84.3%) of our arm's length properties and for 80,544,782 (86.9%) of our arm's length transactions. We trim the bottom and top 0.01% transactions in terms of prices per square foot from the final Zillow data. We then aggregate to Census block prices for the years 2000-2019 as described in the main paper, and drop a small amount of observations for which the Census block price per square foot is lower than 0.05 or higher than 20 times the mean of the respective Census blocks prices across years 2000-2019.

A. Identification of residential arm's length transactions

The Zillow data contains a large number of transactions which are not arm's length housing transactions for residential property. These often have missing or zero prices, are foreclosures, intra-family transfers, pure loans, or refinancing transactions. This section documents how we identify arm's length transactions for residential properties. The raw transaction data contains 460.8 million observations which we reduce to 92.6 million transactions that are defined as arm's length, which are in turn based on 53.2 million properties.[21] The following sections document how we identify our set of residential arm's length transactions

1. Missing or low sales price (71.3% of total)

As a first step, we remove transactions with a missing or low sales price. This amounts to removing 328.7 transaction bringing the count down to 132.1 million. This helps to address several other issues as well (as e.g. refinancing transactions are likely to have a missing sales price). We choose a threshold for low sales price of ≤ 1000. Figure A.4 shows a histogram of non-zero sales price transactions up to 10000USD. There is a drop-off in density after 1000USD and 5000USD.

Out of all transactions (incl. Texas where prices are typically not recorded), 70% have a missing sales price, and 1.3% of transactions have a sales price of zero or $\leq$ 1000USD, as shown in the last row of Tables A.5, A.14 or A.15, which break up statistics by year, state or property type. Of the

[21] This is the data downloaded from Zillow on April 7th 2020. This excludes observations with a transaction date before 1990. Since one transaction can contain multiple housing units, the number of units transacted in the raw data is slightly higher at 485 million.

1.3% of the last category (≤ 1000USD), most prices are near zero. We drop all transactions with a sales price of ≤ 1000 or missing.

The remainder of this section provides some descriptive statistics for the three categories of sales price (missing, ≤ 1000USD and > 1000USD) to build confidence that the dropped transactions are not affecting the sample in any peculiar way.

First, Figure A.5 reports the temporal distribution of transactions within the three groups. With the exception of 2019, and the lower number of missing values in the 90s, they have similar trends. It is reassuring that the share of missing sales prices does not fluctuate greatly over time, and more importantly, that the share of transactions with price > 1000USD is fairly stable over time.

Second, Table A.14 shows the coverage of transaction across states in the last column, and the share of sales price categories within states in the first three columns. Some more populous states have fewer observations of nonmissing prices than less populous states, likely due to their disclosure policies (e.g. Texas).

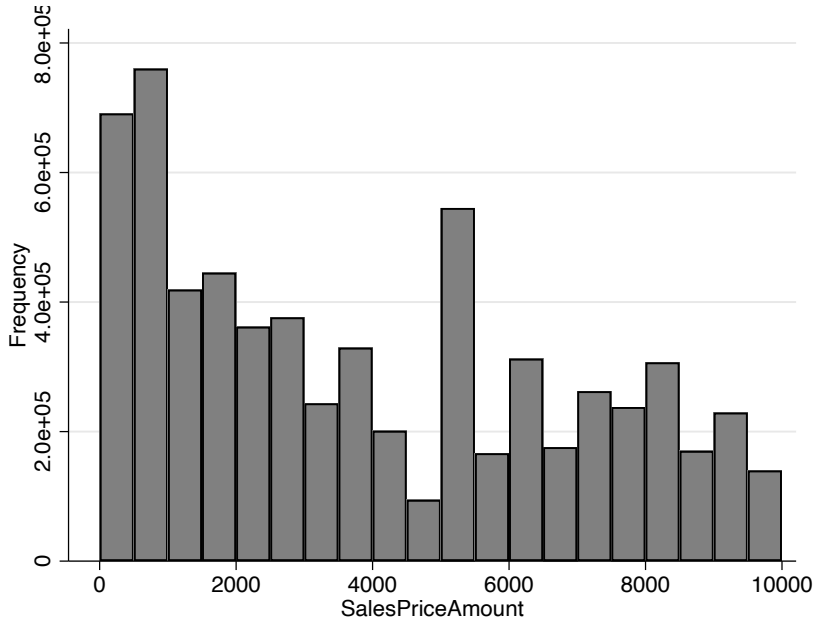
Third, Table A.15 shows the distribution of property types within these categories. Relatively speaking, there are more condos, more single family residence, and fewer missing property types for the "regular" category of > 1000USD.

Fourth, Table A.16 shows the distribution across data type code within the three sales price categories. Most missing sales prices are associated with M - Mortgages or F - Foreclosures. More detailed analysis by document types (available upon request) shows that within the regular D - Deed Transfers, the missing sales prices are often associated with intrafamily transfers (INTR) or gift deeds (GFDE). This shows that removing the transactions with missing and low sales prices already goes a long way in removing mortgages, gifts or intrafamily transfers, for example.

Fifth, loan types in Table A.17 are mostly missing (95% for sales price > 1000USD). Out of the transactions with missing sales prices, 11% are HELOCs. These are home equity lines of credit which can be used to purchase other goods, or consolidate debts such as credit card debts. A later step will remove all loan types.

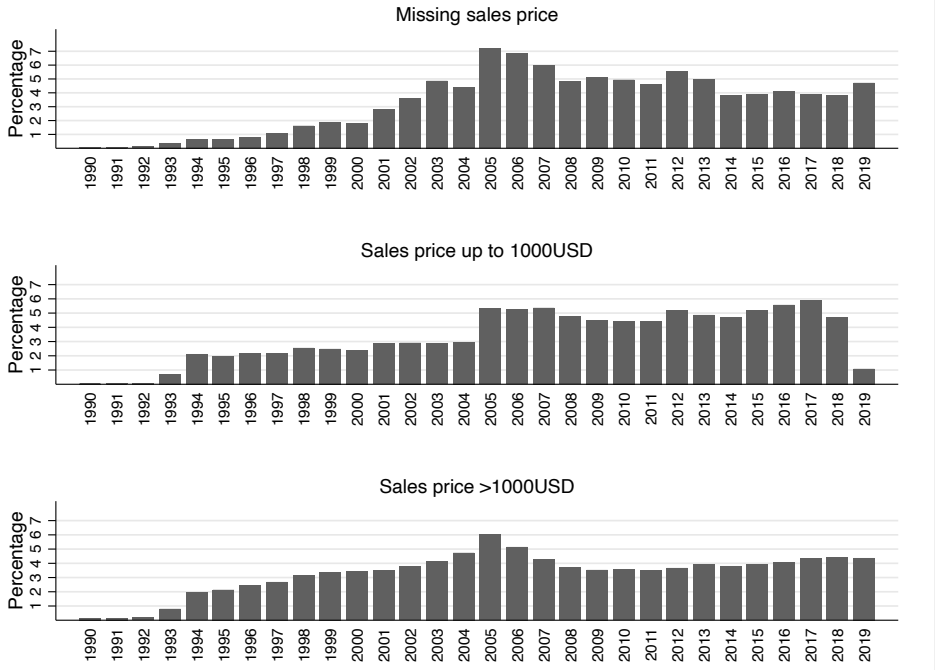
Sixth, Table A.18 shows the reported source of the sales price data. For the > 1000USD category, the primary sources are the reported sales price on the document (RD) or computed from the transfer tax (CF and CR). For around 25% of the transactions in this category there is no source for the sales price reported.

Figure A.4: Histogram of low positive sales prices



Notes: The figure plots the histogram of sales prices *excluding* zero and up to 10000USD. Based on all states and available years.

Figure A.5: Temporal distribution for missing or low sales price



Notes: The figure plots the annual shares of transactions within the three groups of sales price types (missing, low, regular). Transactions before 1990 and after 2019 are ignored. For a corresponding table, see Table A.13.

Table A.13: Distribution of years for missing or low sales prices

Years	Missing	$\leq$ 1000 USD	$>$ 1000 USD	All
1990	0.07	0.03	0.14	0.09
1991	0.07	0.04	0.15	0.09
1992	0.12	0.07	0.18	0.14
1993	0.38	0.72	0.81	0.51
1994	0.67	2.13	1.96	1.06
1995	0.63	1.97	2.16	1.08
1996	0.78	2.19	2.49	1.29
1997	1.11	2.16	2.72	1.58
1998	1.64	2.56	3.19	2.10
1999	1.89	2.49	3.43	2.34
2000	1.83	2.41	3.47	2.30
2001	2.85	2.89	3.54	3.05
2002	3.65	2.92	3.81	3.68
2003	4.91	2.88	4.21	4.68
2004	4.44	2.96	4.79	4.52
2005	7.26	5.33	6.11	6.90
2006	6.90	5.27	5.21	6.39
2007	6.05	5.39	4.31	5.54
2008	4.89	4.79	3.75	4.56
2009	5.18	4.52	3.58	4.71
2010	4.99	4.46	3.62	4.59
2011	4.68	4.45	3.51	4.34
2012	5.59	5.22	3.72	5.05
2013	5.01	4.89	3.96	4.71
2014	3.83	4.74	3.85	3.85
2015	3.94	5.18	3.95	3.96
2016	4.13	5.59	4.11	4.14
2017	3.92	5.93	4.40	4.08
2018	3.87	4.73	4.45	4.04
2019	4.76	1.09	4.42	4.62
Total (count)	322,621,078	6,081,035	132,120,117	460,822,230
Total (percentage)	70.0	1.3	28.7	100.0

Notes: The cells contain the percentage of a annual counts in the transaction data across rows within each column. The two bottom rows contain the total count of transactions as well as the percentages across rows. Based on all states.

Table A.14: Share of missing or low sales prices by state

State	Missing	$\leq$ 1000 USD	$>$ 1000 USD	State %
Alabama	66.6	3.7	29.7	1.1
Alaska	81.9	14.5	3.6	0.2
Arizona	65.4	0.5	34.1	3.4
Arkansas	66.3	2.9	30.7	0.9
California	77.6	0.5	22	17.0
Colorado	71.6	0.4	28	2.8
Connecticut	66.7	0	33.3	1.3
Delaware	65.8	3.5	30.6	0.2
District Of Columbia	57.6	0	42.3	0.1
Florida	57.2	0.6	42.3	9.0
Georgia	64.3	0.3	35.5	3.4
Hawaii	65.4	1.7	32.9	0.4
Idaho	84	14.1	1.9	0.6
Illinois	72.8	0.3	26.9	4.3
Indiana	92	0.5	7.5	1.5
Iowa	67.3	1.4	31.3	0.9
Kansas	96.5	0.1	3.3	0.5
Kentucky	63.2	0.4	36.4	0.8
Louisiana	54.7	3.3	42	0.6
Maine	98.7	0	1.3	0.4
Maryland	62.4	0.2	37.4	1.9
Massachusetts	71.7	0.1	28.2	2.8
Michigan	67.6	1	31.4	3.0
Minnesota	59	0.5	40.5	1.3
Mississippi	92.4	6.2	1.4	0.6
Missouri	86.8	3.3	9.9	1.7
Montana	97	1.6	1.3	0.3
Nebraska	62.9	4.2	32.9	0.4
Nevada	63.6	0.2	36.1	1.4
New Hampshire	54.6	0	45.4	0.4
New Jersey	65	0.5	34.5	2.4
New Mexico	96.6	0.6	2.8	0.5
New York	62.6	0.7	36.7	3.7
North Carolina	64.5	1.2	34.3	2.8
North Dakota	78.9	2.7	18.4	0.2
Ohio	65.8	0.4	33.8	4.0
Oklahoma	60.7	5.4	33.9	1.0
Oregon	63.9	0.2	35.8	1.2
Pennsylvania	63.8	0.8	35.4	2.9
Rhode Island	69.2	0.1	30.7	0.4
South Carolina	58.5	2.3	39.2	1.4
South Dakota	79.2	1.6	19.1	0.1
Tennessee	60	0.6	39.5	2.6
Texas	92.1	3.6	4.3	6.3
Utah	77.2	21.2	1.6	1.1
Vermont	43.4	0.7	55.9	0.2
Virginia	56.3	0.5	43.2	1.6
Washington	64.4	0.6	35	2.2
West Virginia	69.1	1.5	29.4	0.3
Wisconsin	66.9	1.6	31.5	1.5
Wyoming	96.9	1.2	1.9	0.1
Total	70	1.3	28.7	100.0

Notes: The cells in the first three columns contain the percentage for the three categories within each state. The last column contains the state shares of observations across rows. Based on all years.

Table A.15: Distribution of property types for missing or low sales prices

Property type	Missing	$\leq$ 1000 USD	$>$ 1000 USD	All
AG – Agricultural	0.1	0.2	0.2	0.1
AP – Apartment Building	0.1	0.0	0.1	0.1
CD – Condominium	5.1	3.3	8.3	6.0
CI – Commercial & Industrial	0.0	0.0	0.1	0.0
CM – Commercial	0.4	0.2	0.9	0.5
CP – Cooperative	0.0	0.0	0.2	0.1
EX – Exempt	0.0	0.0	0.0	0.0
GV – Government	0.0	0.0	0.0	0.0
IM – Improved Land	0.0	0.0	0.1	0.0
IN – Industrial	0.0	0.0	0.1	0.1
MB – Mobile Home	0.1	0.1	0.4	0.2
MF – Multi-Family Dwelling (2-4 Units)	2.3	1.0	2.4	2.3
MH – Manufactured Home	0.1	0.1	0.1	0.1
MX – Mixed Use	0.0	0.0	0.0	0.0
NW – New Construction	0.0	0.0	0.0	0.0
PD – Planned Unit Development	5.8	6.0	6.4	6.0
RC – Recreational	0.0	0.0	0.1	0.0
RR – Residential	1.3	0.8	2.7	1.7
SR – Single Family Residence	20.1	11.5	27.5	22.1
UL – Unimproved Land/Lot	0.6	2.5	2.6	1.2
VL – Vacant Land/Lot	0.0	0.0	0.0	0.0
Missing	64.0	74.1	47.7	59.5
Total (count)	322,621,078	6,081,035	132,120,117	460,822,230
Total (percentage)	70.0	1.3	28.7	100.0

Notes: The cells contain the percentage of the property type counts in the transaction data across rows within each column. The two bottom rows contain the total count of transactions as well as the percentages across rows. Based on all states and all available time periods.

Table A.16: Distribution of data type for missing or low sales prices

Data type code	Missing	$\leq$ 1000 USD	$>$ 1000 USD	All
S – Assessment Historical Sales	0.1	0.3	0.5	0.2
E – Declaration of Easement; Documents	0.0	0.0	0.0	0.0
D – Deed Transfer	28.3	75.4	53.2	36.0
H – Deed with Concurrent Mortgage	7.2	22.4	45.9	18.5
X – Deeds that consummate a Lot Line Ad	0.0	0.0	0.0	0.0
F – Foreclosure	8.3	0.0	0.0	5.8
U – Hawaii - Deed Transfer	0.2	0.4	0.2	0.2
J – Hawaii - Deed with Concurrent Mortg	0.0	0.1	0.3	0.1
V – Hawaii - Stand Alone Mortgage	0.1	0.0	0.0	0.0
M – Mortgage	55.2	1.2	0.0	38.6
P – Stand Alone Purchase Money Mortgage	0.7	0.2	0.0	0.5
Total (count)	322,621,078	6,081,035	132,120,117	460,822,230
Total (percentage)	70.0	1.3	28.7	100.0

Notes: The cells contain the percentage of the data type counts in the transaction data across rows within each column. The two bottom rows contain the total count of transactions as well as the percentages across rows. Based on all states and all available time periods.

Table A.17: Distribution of loan types for missing or low sales prices

Loan type	Missing	$\leq$ 1000 USD	$>$ 1000 USD	All
AC – Agricultural/Commercial	0.0	0.0	0.0	0.0
AS – Assumption	0.0	0.1	0.1	0.1
BL – Balloon	0.3	0.1	0.4	0.3
CE – Closed-end Mortgage or Closed End w	0.0	0.0	0.0	0.0
CM – Commercial	1.8	0.4	1.2	1.6
CT – Commercial Construction Loan	0.0	0.0	0.0	0.0
CS – Construction Loan	0.8	0.4	0.4	0.7
CC – Construction Loan Credit Line	0.0	0.0	0.0	0.0
CL – Credit Line (HELOC)	10.0	0.4	0.2	7.1
DP – Down Payment Assistance Loan	0.0	0.0	0.0	0.0
FO – Farm Ownership Loan	0.0	0.0	0.0	0.0
FE – First Lien Home Equity Loan	0.0	0.0	0.0	0.0
FM – First Mortgage (First Lien Deed of	0.6	0.1	0.3	0.5
HE – Home Equity Loan	0.0	0.0	0.0	0.0
LC – Land Contract (Contract/Agreement o	0.0	0.0	0.0	0.0
EB – Loan Amount $10-$99 Billion	0.0	0.0	0.0	0.0
EX – Loan Amount $1-9 Billion	0.0	0.0	0.0	0.0
MD – Loan Modification	0.0	0.0	0.0	0.0
NP – Non-Purchase Money (no other loan t	0.0	0.0	0.0	0.0
FA – Open End Mortgage or Open End with	1.6	0.3	0.4	1.2
PM – Purchase Money (no other loan type	0.5	0.2	0.6	0.5
RE – Refinance	1.3	0.9	0.0	0.9
RM – Reverse Mortgage (HECM)	0.2	0.0	0.0	0.2
RD – Rural Development Loan	0.0	0.0	0.0	0.0
SM – Second (Subordinate) Mortgage	2.8	0.1	0.1	2.0
SE – Second Lien Home Equity Loan	0.0	0.0	0.0	0.0
SL – Seller Take Back	0.3	1.3	1.4	0.6
TR – Trade	0.0	0.0	0.0	0.0
Missing	79.6	95.8	94.9	84.2
Total (count)	322,621,078	6,081,035	132,120,117	460,822,230
Total (percentage)	70.0	1.3	28.7	100.0

Notes: The cells contain the percentage of the loan type counts in the transaction data across rows within each column. The two bottom rows contain the total count of transactions as well as the percentages across rows. Based on all states and all available time periods.

Table A.18: Distribution of sales price origin for missing or low sales prices

Sales price origin	Missing	$\leq$ 1000 USD	$>$ 1000 USD	All
AF – Full Amount from Assessment File	0.0	0.3	0.4	0.1
AV – Price from recorded Affidavit of Va	0.0	0.4	4.9	1.4
BL – Sales Price amount or Transfer Tax	0.0	0.0	0.0	0.0
CF – Full Consideration - Computed from	0.0	0.9	16.1	4.6
CM – Comparable Market Value	0.0	0.0	0.0	0.0
CN – Unknown if Computed amount from Tra	0.0	0.0	0.2	0.1
CP – Partial Consideration - Computed fr	0.0	0.2	0.1	0.0
CR – Full Consideration - Computed from	0.0	1.2	12.6	3.6
CS – Cash Sale	0.0	0.0	0.0	0.0
CU – Unknown if Computed Price from Tran	0.0	1.2	1.5	0.5
DL – Unpaid Balance/Debt or Delinquent A	0.0	0.0	0.0	0.0
EX – Exchange	0.0	0.0	0.0	0.0
GT – No Consideration - Gift	0.0	0.0	0.0	0.0
HB – Highest Bid Amount	0.0	0.1	0.1	0.0
LN – Liens exceed value or assumption of	0.0	0.0	0.0	0.0
MP – Sales Price manually computed from	0.0	0.1	0.0	0.0
NO – No Consideration or calculable Tran	6.8	11.2	0.1	4.9
NP – Sales Price Not Public Record	0.1	0.2	0.0	0.1
QS – Qualified Sale (Assessor)	0.0	0.0	0.0	0.0
RA – Redemption Amount	0.0	0.7	0.0	0.0
RD – Sales Price/Amount as reported on d	0.1	7.5	38.4	11.2
ST – Sold for Taxes	0.0	0.5	0.0	0.0
UN – Unable to calculate Sales Price fro	0.1	0.2	0.0	0.0
WA – Sales Price computed using current	0.0	0.0	0.1	0.0
Missing	92.8	75.4	25.3	73.2
Total (count)	322,621,078	6,081,035	132,120,117	460,822,230
Total (percentage)	70.0	1.3	28.7	100.0

Notes: The cells contain the percentage of the sale price origin counts in the transaction data across rows within each column. The two bottom rows contain the total count of transactions as well as the percentages across rows. Based on all states and all available time periods.

Table A.19: Foreclosure document types

Doc. type code	Document type	Data type	% of dropped
AFNS	Affidavit of Publication of Notice of Sale	F	0
ASCP	Assignment of Certificate of Purchase	F	0
ASCS	Assignment of Certificate of Sale	F	0
CFPR	Certificate of Purchase (Public Trustee's Certificate of Purchase)	F	0
CNDF	Cancellation of Notice of Default	F	0
CNLP	Cancellation of Lis Pendens	F	0
CNSL	Cancellation of Notice of Sale (Trustee, Sheriff or Forecl. Sale)	F	0
FASD	Foreclosure Auction - Status Sold	F	0
JGFC	Judgment of Foreclosure	F	0
NFCM	Newly Filed Complaint/Petition to Foreclose	F	0
NTDF	Notice of Default	F	0
NTLP	Notice of Lis Pendens	F	0
NTSL	Notice of Sale (Trustee Sale, Sheriff Sale or Foreclosure Sale)	F	0
ORDS	Order of Dismissal	F	0
PAFC	Power of Attorney to Foreclose Mortgage	F	0
PCDF	Partial Cancellation of Notice of Default	F	0
PCLP	Partial Cancellation of Lis Pendens	F	0
SLCH	Sale Change (Rescheduled or Postponed)	F	0
SLCN	Sale Cancelled	F	0
CFSL	Certificate of Sale	D	0
CMDE	Commissioner's Deed	D	1.4
COCA	Court Order/Action	D	1.5
DELU	Deed in Lieu of Foreclosure	D	1.7
DESL	Deed Under Power of Sale	D	0
DSSL	Distress Sale	D	0
ESDE	Estoppel Deed	D	0
FCDE	Foreclosure Deed/Certificate	D	11.1
RCDE	Receiver's Deed	D	0.1
RDDE	Redemption Deed	D	0.9
RDQC	Redemption Quit Claim Deed	D	0
SFSL	Sheriff's Certificate of Sale	D	0
SHDE	Sheriff's Deed	D	22.3
SHTX	Sheriff's Tax Deed	D	0
TRFC	Trustee's Deed (foreclosure sale transfer)	D	56.8
TXDE	Tax Deed	D	4.2

Notes: The upper part of the correspondence is directly based on Ztrax documentation. The bottom part is keyed manually. The last columns shows each deed's percentage share of total removed observations in this step. The upper part has all missing sales values so has already been removed in the previous step.

2. Foreclosures and distress sales (2.0% of total)

As a second step, all remaining identified foreclosures are removed. This includes all *data* types F, which also removes all *document* types listed in the upper part of Table A.19. Since all of these observations have a missing or low sales prices, no additional observations are removed. However, the lower part of Table A.19 manually classifies document types as foreclosure. Few of them are not foreclosures in a strict legal sense, but practically very close to it (RCDE for example). The Bargain and Sale Deed (BSDE) is one of the main deed types in Nevada and is therefore retained. BSDE deeds make up 69% of transactions with a sales price $> USD1000$ in Nevada, compared to 1.3% in all states.

Removing the transactions with a document type classified in the bottom half of the table removes an additional 9.3 million transactions bringing the count down to 122.8 million. The last column in Table A.19 shows the percentage of different foreclosure deeds that are removed.

Table A.20: Intra-family and gift document types

Doc. type code	Document type	Data type
AFDL	Affidavit - Death of Life Tenant (termination of life interest))	D
AFDR	Affidavit - Death of Trustee/Successor Trustee	D
AFDT	Affidavit - Death of Joint Tenant	D
AFSJ	Affidavit - Surviving Joint Tenant	D
AFSS	Affidavit - Surviving Spouse	D
AFSV	Affidavit - Survivorship	D
AFTD	Affidavit - Transfer on Death	D
BFDE	Beneficiary Deed	D
EXDE	Executor's Deed/Executrix's Deed	D
GDDE	Guardian's Deed	D
GFDE	Gift Deed	D
GFGR	Gift Grant Deed	D
GFWD	Gift Warranty Deed	D
INTR	Intrafamily Transfer & Dissolut	D
JTDE	Joint Tenancy Deed	D
SVDE	Survivorship Deed	D
SVWD	Survivorship Warranty Deed	D
TFDD	Transfer on Death Deed	D

Notes: The table lists the document type codes that are removed.

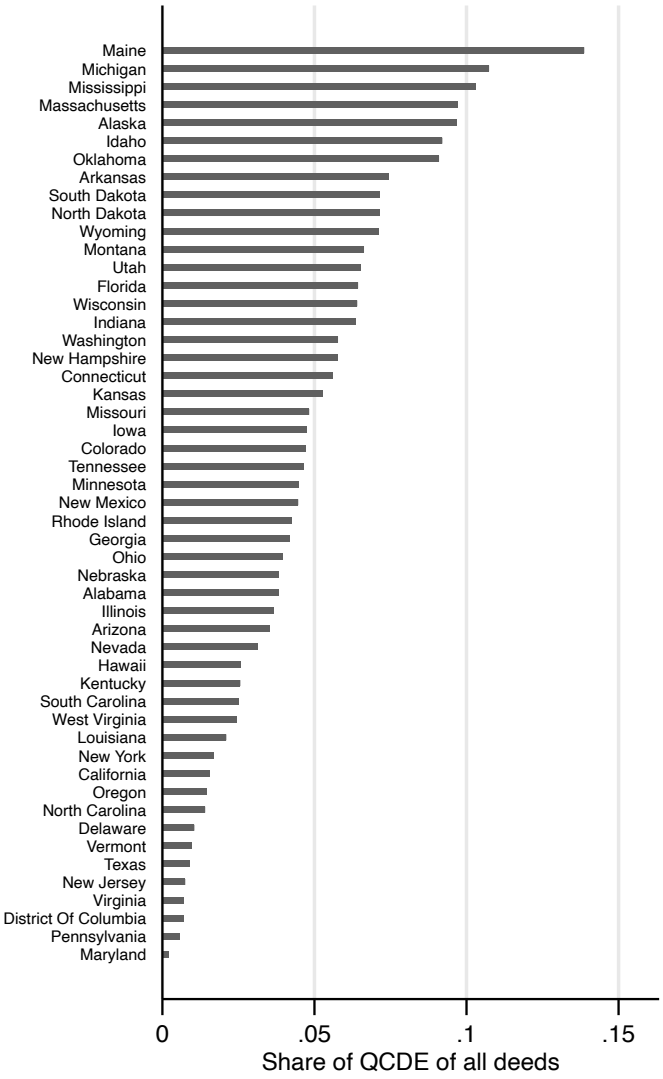
3. Intra-family and gift transfers (0.7% of total)

The third step is to remove intra-family and gift transactions. There is an intra-family flag coded by the Zillow team, which predominately corresponds to the INTR document type. Having removed these observations, Table A.20 lists the document types that were in addition manually identified as intra-family and gift transfers and also removed.[22] In total, 3.4 million intra-family and gift transfer transactions are removed, bringing the count down to 119.5 million transactions.[23]

[22]There is an additional code in the data types (GT: No Consideration - Gift), which does not remove any additional observations, however.

[23]Quitclaim deeds are around 3 million transactions. Some may be used for intra-family transfer, but not necessarily, so they are retained in the data. See Figure A.6 for an overview of the shares of Quitclaim deeds within each state.

Figure A.6: Share of Quitclaim deeds within states



Notes: This figure is based on the 129.4 million transactions with a sales price $>$ 1000USD.

4. Credit lines, refinancing and pure mortgages (1.4% of total)

While Zillow defaults deed transfer documents to DEED, pure loan documents default to MTGE. All document types MTGE are removed (only 9 at this stage). The default value for loan types is empty. There are 6.5 million observations with recorded loan types, mainly commercial loans and seller take back loans. All transactions with a recorded loan type are removed, which includes refinancing transactions and new credit lines (e.g. HELOCs). In total this brings the observations down to 113.0 million transactions.[24]

[24] It is possible that some of these loan types are actual transactions, although it is unlikely. One way to further refine this could be to use the information on buyers and sellers.

Table A.21: Retained and removed property types

Retained property types	Removed property types
AP – Apartment Building	AG – Agricultural
CD – Condominium	CI – Commercial & Industrial
MF – Multi-Family Dwelling (2-4 Units)	CM – Commercial
MH – Manufactured Home	CP – Cooperative
MX – Mixed Use	EX – Exempt
NW – New Construction	GV – Government
PD – Planned Unit Development	IM – Improved Land
RR – Residential	IN – Industrial
SR – Single Family Residence	MB – Mobile Home
	RC – Recreational
	UL – Unimproved Land/Lot
	VL – Vacant Land/Lot

Notes: The table lists the retained and removed property types. Missing property types are also retained.

5. Non residential property types (1.2% of total)

We next remove non-residential property types. Table A.21 lists the retained and removed property types. Transactions with missing property types (45% of the remaining transactions) are also retained. Most of the non-missing property types are single family residences. This step removes 5.4 million transactions bringing the count to 107.6 million transactions.

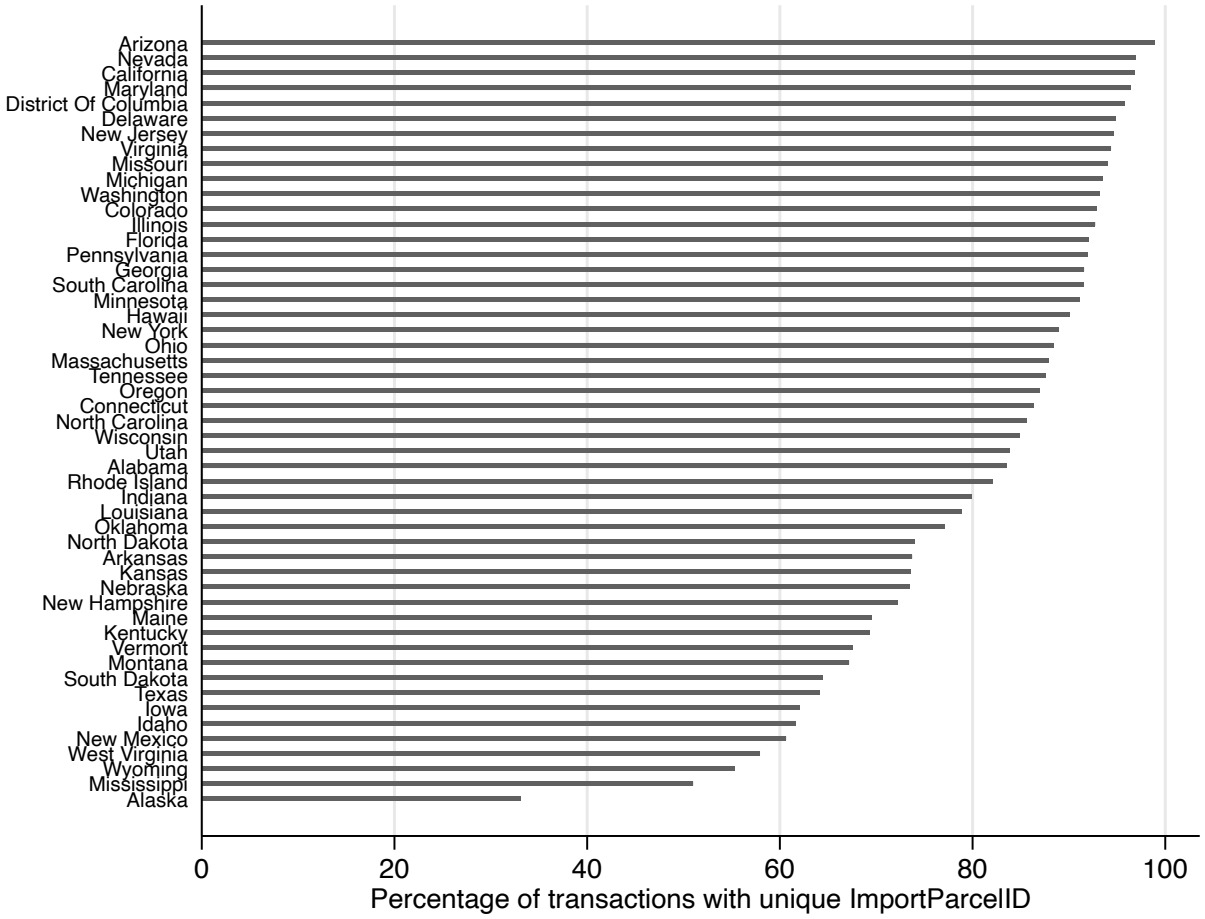
6. Multiple properties per transaction and missing panel ids (2.3% of total)

A particular transaction can contain multiple units. All transactions with multiple units are dropped as the sales price cannot be assigned to a particular property. This first step removes 2.3 million transactions (0.5% of total). The third column in Table A.22 shows the percentage of transactions with multiple ImportParcelIDs (the panel id) within each state, based on the cleaned data from the previous steps.

We also remove the transactions with missing ImportParcelIDs, which constitutes 8.4 million transactions (1.8% of the total). The first column in Table A.22 shows the percentage of transactions with missing ImportParcelIDs (the panel id) within each state, based on the cleaned data from the previous steps.

The second column of Table A.22, the percentage of retained transactions with unique ImportParcelIDs, are plotted in Figure A.7. For a handful of small states, the availability of ImportParcelIDs is less than or near 50%, driven by the missing ImportParcelIDs. In total, this step removes 10.7 million observations bringing the count to 96.8 transactions.

Figure A.7: Percentage of unique ImportParcelIDs transactions by state



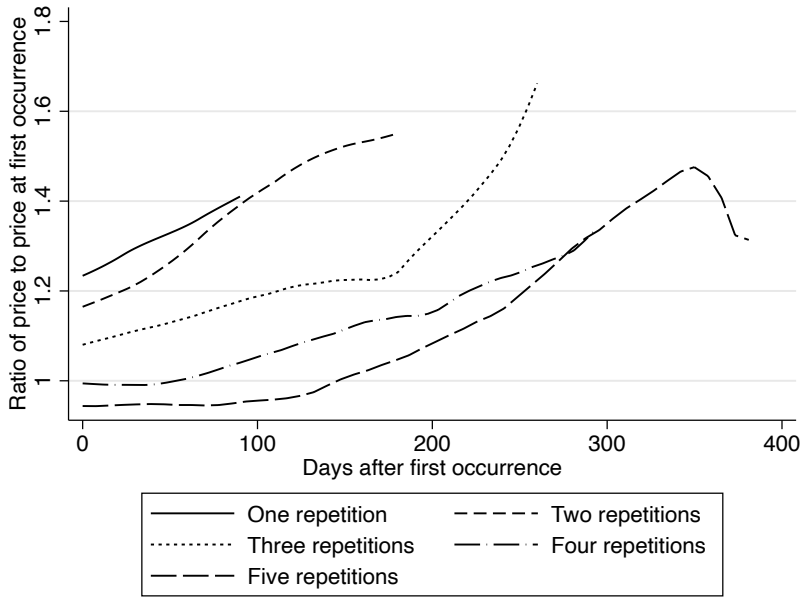
Notes: The figure shows the percentage of transactions that has a single non-missing ImportParcelID, as opposed to multiple units being transacted or units transacted with missing ImportParcelIDs. The percentages are as share of the transactions left after the previous cleaning steps.

Table A.22: Shares of missing or multiple units per transaction by state

State	Missing ImpParcID	Unique ImpParcID	Multiple ImpParcID	All
Alabama	16.1	83.5	0.4	100.0
Alaska	66.8	33.1	0.1	100.0
Arizona	0.5	98.9	0.6	100.0
Arkansas	26.0	73.7	0.2	100.0
California	2.0	96.8	1.1	100.0
Colorado	7.0	92.9	0.1	100.0
Connecticut	13.6	86.4	0.0	100.0
Delaware	3.2	94.8	2.0	100.0
District Of Columbia	4.0	95.8	0.2	100.0
Florida	7.2	92.0	0.8	100.0
Georgia	7.5	91.6	1.0	100.0
Hawaii	8.4	90.1	1.5	100.0
Idaho	37.3	61.6	1.1	100.0
Illinois	3.6	92.7	3.7	100.0
Indiana	10.9	79.9	9.1	100.0
Iowa	37.5	62.0	0.4	100.0
Kansas	26.1	73.6	0.3	100.0
Kentucky	30.2	69.3	0.5	100.0
Louisiana	21.0	78.8	0.1	100.0
Maine	30.4	69.6	0.1	100.0
Maryland	1.1	96.4	2.5	100.0
Massachusetts	12.0	87.9	0.1	100.0
Michigan	4.7	93.5	1.9	100.0
Minnesota	6.0	91.1	2.9	100.0
Mississippi	48.8	51.0	0.2	100.0
Missouri	5.5	94.0	0.5	100.0
Montana	32.4	67.2	0.4	100.0
Nebraska	25.4	73.5	1.1	100.0
Nevada	2.5	96.9	0.6	100.0
New Hampshire	27.7	72.2	0.0	100.0
New Jersey	4.7	94.7	0.6	100.0
New Mexico	39.0	60.6	0.4	100.0
New York	10.4	89.0	0.6	100.0
North Carolina	12.1	85.6	2.3	100.0
North Dakota	24.2	74.0	1.8	100.0
Ohio	5.7	88.4	5.9	100.0
Oklahoma	22.8	77.1	0.2	100.0
Oregon	11.8	86.9	1.2	100.0
Pennsylvania	6.5	91.9	1.6	100.0
Rhode Island	17.8	82.1	0.1	100.0
South Carolina	5.3	91.5	3.2	100.0
South Dakota	35.5	64.4	0.0	100.0
Tennessee	11.2	87.6	1.2	100.0
Texas	35.5	64.2	0.3	100.0
Utah	14.5	83.9	1.6	100.0
Vermont	32.4	67.6	0.0	100.0
Virginia	5.0	94.4	0.7	100.0
Washington	4.4	93.2	2.4	100.0
West Virginia	41.5	57.9	0.6	100.0
Wisconsin	11.8	84.9	3.3	100.0
Wyoming	44.4	55.3	0.3	100.0

Notes: The cells contain the percentage of transactions with missing, unique or multiple ImportParcelIDs within a state. These shares are based on the cleaned data from the previous steps, e.g. do not contain transactions with missing sales price.

Figure A.8: Repeated sales within a 90 days rolling window and sales price



Notes: The figure plots the smoothed average ratio of the sales price to the sales price at the first occurrence of a spell of repeated sales. The average ratio is the exponentiated average of the log ratios to account for the non-linear scale of ratios. One particular property can have multiple spells of repeated sales. A transaction belongs to a spell of repeated sales if the previous transaction was up to 90 days ago. It is a rolling window, so a spell can extend beyond 90 days from the first transaction if there are multiple transactions in a spell. The figure plots five separate graphs by the number of repeated sales in a spell, e.g. "one repetition" indicates a spell of two transactions, and can therefore only be up to 90 days. Plotted is a kernel smoother with a triangular kernel and a bandwidth of 30 days. The graphs show that the later transactions in a spell of transactions are on average higher than the previous transactions. The graphs for more than five repetitions look similar and become more noisy due to fewer spells with highly repetitive sales.

7. Repeated sales (0.7% of total)

Next, a subset of transactions within a short period of time are removed. Specifically, if there are multiple transactions of the same property within a 90 day rolling window, only the last of these transactions is retained.[25] The last transaction of a spell of repeated sales is typically higher than the removed previous sales as Figure A.8 shows. Overall the last transaction in a spell is larger than the first in around 80% of spells. Table A.23 shows how the repeated sales that are to be removed are distributed across states. Florida contains the most repeated sales. This step removes 3.4 million transactions bringing the total count to 93.5 million transactions.

[25]Since the window is rolling, if there are multiple repeated sales a spell of repeated sales can extend beyond 90 days, and the last observation of the entire spell is retained. A property can have multiple spells through time.

Table A.23: Distribution of repeated sales across states

State	Percentage
Alabama	1.2
Alaska	0.0
Arizona	3.9
Arkansas	0.6
California	8.3
Colorado	2.6
Connecticut	1.0
Delaware	0.2
District Of Columbia	0.1
Florida	15.9
Georgia	5.7
Hawaii	0.6
Idaho	0.0
Illinois	2.5
Indiana	0.4
Iowa	0.5
Kansas	0.0
Kentucky	0.9
Louisiana	0.5
Maine	0.0
Maryland	5.4
Massachusetts	3.0
Michigan	3.0
Minnesota	2.3
Mississippi	0.0
Missouri	0.6
Montana	0.0
Nebraska	0.3
Nevada	1.9
New Hampshire	0.4
New Jersey	3.8
New Mexico	0.0
New York	6.3
North Carolina	2.6
North Dakota	0.1
Ohio	4.5
Oklahoma	0.9
Oregon	1.4
Pennsylvania	2.4
Rhode Island	0.3
South Carolina	3.7
South Dakota	0.0
Tennessee	6.6
Texas	0.1
Utah	0.0
Vermont	0.1
Virginia	1.4
Washington	1.6
West Virginia	0.1
Wisconsin	1.8
Wyoming	0.0

Notes: The cells contain the share of a state in the number of repeated sales to be dropped as percentages.

8. Multiple unit properties (0.2% of total)

Finally, after merging the properties to the assessment data, there are a few properties with multiple units per unique property ID. These are for example two individual apartments treated as one. Since it is not clear how to aggregate hedonic variables across these, they are dropped. This removes 0.8 million transactions bringing the total count to 92.6 million.

B. Identifying property locations

We next describe how we define property locations. The 92,639,072 transactions that are defined as arm's length above are based on 53,164,562 properties.

The property geolocation and address is provided in the transaction, assessment and historical assessment tables. To evaluate the quality of the provided latitudes/longitudes and addresses, we have drawn a sample of 10,000 properties and geocoded the provided addresses with ESRI based on the provided address. For 95%, the newly geocoded location is less than 160 meters away from the original lats/lons, and for 99% it is less than 1400 meters away. One discrepancy, for example, arises in rural areas, where the geocoded ESRI coordinates are at the street entrance of the property, while the Zillow coordinates are sometimes on the property itself. In the few cases with a large distance between original lats/lons and the geocoded ones, the original lats/lons are closer to a third set of coordinates derived from Google Maps. The ESRI coordinates are slightly closer to the lats/lons from the transaction tables than to those in the assessment tables for the 3.1% when they do not match exactly. Furthermore, in the cases where the zip code from the transaction and assessment tables disagrees (0.8% of times), the transaction zip code matches the Google Maps zip code much more frequently (85%).[26]

We construct the set of lats/lons, zip codes and street addresses in five steps. First we take the lats/lons from the transaction tables, which are available in 97.5% of the cases (we do the same steps for zip codes and addresses).[27] Second, we complement missing ones from the assessment tables which adds 0.4 percentage points to the lats/lons. As a third step, we complement the missing values with the historic information, preferring the most recent non-missing values which adds another 0.7 percentage points to the lats/lons. Of the 53,164,562 properties with arm's length transactions, there are non-missing coordinates for 98.6% (52,443,223), non-missing zip codes for 99.8% (53,066,792), non-missing addresses for 97.3% (51,737,628).

As a fourth step, we ensure the quality of the existing lats/lons by calculating the distance to the official TIGER county boundaries. If the counties in the Zillow data match the TIGER counties (distance is zero) they pass our quality test. The existing lats/lons also pass the quality test if the distance to the matching counties is less than 1km. Manual inspection shows that the shape files at the county boundaries can be imprecise (i.e. in the case of a winding road at the border), and that the lats/lons are actually in the correct county. For the lats/lons that do not pass our quality test, we set them to missing and pass them to the next geocoding step. This adds 47,720 properties to

[26] The disagreement between assessment and transaction coordinates and zips is scattered across all states and years.

[27] For multiple addresses per property for different transactions, we keep the longer street addresses, after cleaning upper/lower cases and spaces.

the 721,339 properites with missing coordinates. In total, for the 769,059 properties with missing coordinates, we have 51.8% (398,378) with non-missing address and zip code, 3.9% and 38.5% with only address and zip respectively, and 5.8% without address or zip code.

To ensure a high quality of geocoding, we only geocode the properties with existing addresses and zip codes in the fifth step using ESRI Streetmap Premium.[28] A few of the geocoded properties have non-matching geocoded counties and original Zillow counties. We only use the geocoded coordinates for matching counties and where the ESRI score is high (¿80%), which is 91.2% of the 398,378 proprieties. With reverse geocoding, we retrieve missing addresses and zip codes from existing coordinates. We set the location of 912 properties to missing where the reverse geocoded counties do not match existing Zillow counties.

The final share of properties with non-missing coordinates is 99.0% (52,612,606), corresponding to 92,006,045 transactions. The share of properties with non-missing addresses is 98.7%, and the share with non-missing zip codes is 99.8%.

[28] We feed in the addresses and county names as the county identification should be the most reliable data because the raw data is obtained from the individual counties.

C. *Identifying square footage of the property*

We next identify the size of the property in square footage and link it to our 92,006,045 transactions based on our 52,612,606 properties for that we also have coordinates from the previous section.[29] Due to the last step of identifying the arm's length transactions, all properties are single unit properties.[30]

There are several different types of building areas that define the size of the property. Some refer to total areas such as "Living Building Area" (BAL), "Gross Building Area" (BAG) or "Total Building Area" (BAT), and others refer to parts, such as "Balcony/Overhang", "Basement", "Porch". The coverage on the total areas is much better than on the individual parts. Each property can have multiple building area types, referring e.g. to the balcony area and the total area. According to Zillow, the "Living Building Area" is usually taken as the property area. While it has the lowest number of missing observations of all types of areas, it is still only available for 66.2% of arm's length properties in the assessment tables as Table A.24 shows.

Before proceeding, we ask whether the missing data comes from particular counties or states. We calculate the share of properties with non-missing "Living Building Area" (BAL) information both within counties or within states. Figure A.9 plots the histogram of the shares of non-missing BAL information within counties weighted by the number of properties in a county. There are many counties that do not report the BAL for any property, so there seems to be little selection within counties. This is the main driver for the missing information in Table A.24. Figure A.10 shows that there are some states (e.g. Illinois) in which less than 40% of properties have information on the BAL.

We next supplement the 66.2% of nonmissing observations of BAL. As a first step we complement this data with information from historical versions of the assessment tables. This increases the share of non-missing "Living Building Area" to 73.6% as shown in Table A.24.[31]

[29] Around 0.1% of these cannot be matched to the assessment tables. These missing properties are missing across states and years, and are not just concentrated in recent years. The ca. 50 million properties are a third of the 150 million properties in the raw assessment tables. For the other 100 million properties, there are no arm's length transactions recorded.

[30] There are sales prices available in the assessment tables as well, but it is recommended to avoid them, as Zillow notes: "Generally, you can think of the data in ZAssessment tables as data sourced ultimately from county's assessor's offices and ZTransaction tables as data ultimately sourced from legal recordings processed by each county recorder's offices. These are usually two separate agencies in the county administration. The Assessor's office tracks many things, like property attributes, completely independently from the County Recorder's office. However, when the County Assessor reports sale prices on homes (the SalesPriceAmount variable in the ZAssessment tables), this is data that the county assessor's office has taken from the recorder's office and blended into their data set before they sent it to us. Some counties will do this to use the most recent sales prices in their assessment amount models. That being said, we've found that the transaction data we get through assessors tends to be marginal and not always up to date, so when available, use the transaction data reported in the ZTransaction tables."

[31] The most recent available historical information is used for each individual property and area type to replace missing

As a second step, we further impute the missing values of BAL by taking the other total area types into account reported in Table A.24 – Total, Base, Finished, and Gross Building Area. The bottom part of Table A.24 reports the share of properties where we have none, one, or multiple area types reported. Importantly, for 84.4% of properties, we have at least one area type reported (100%-15.6%). We therefore impute the missing BALs, by taking one of the other codes adjusted by the median ratio between BAL and the other code.[32] We therefore recover square footage for 84.4% of the properties, corresponding to 80,618,103 of our arm's length transactions. Overall, we have both square footage and coordinates for 44,799,731 (84.3%) of our arm's length properties and 80,544,782 (86.9%) of our arm's length transactions.

Table A.24: Coverage of building area codes

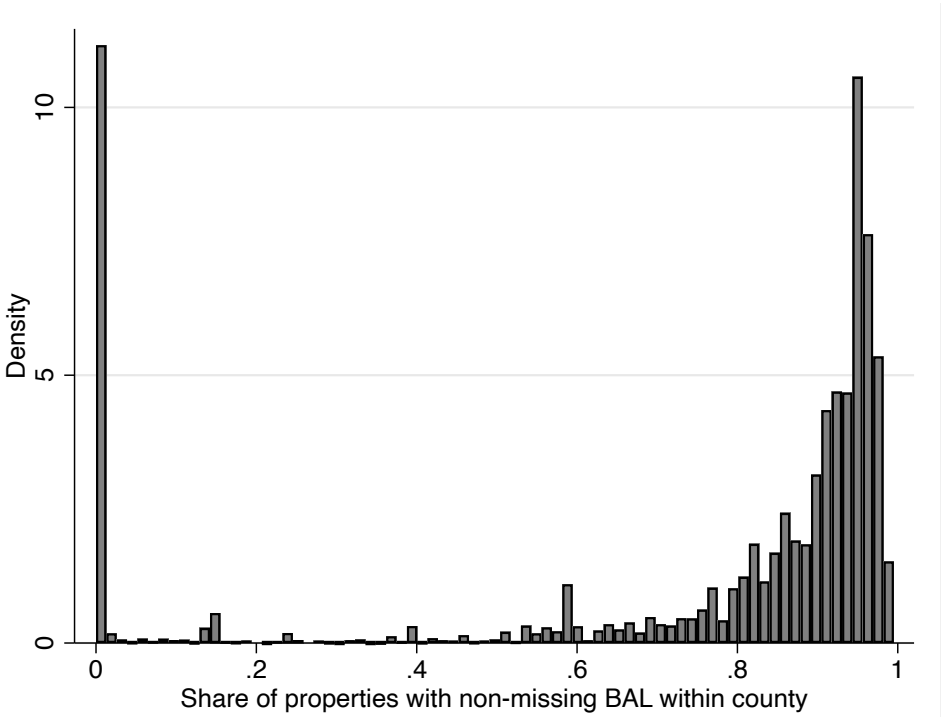
Building area type	% available	% available incl. hist. asmt.
BAT (Total Building Area)	15.5%	27.6
BAL (Living Building Area)	66.2%	73.6
BAB (Base Building Area)	16.6%	26.3
BAF (Finished Building Area)	3.4%	5.6
BAG (Gross Building Area)	8.9%	10.6
None available	19.3%	15.6%
One available	55.1%	40.0%
Two available	21.9%	32.7%
Three available	3.2%	9.5%
Four available	0.6%	2.6%
Five available	0.0%	0.0%

Notes: The table shows the percentage of available values for different types of building areas. The bottom part shows the percentage where none, one etc of the different codes above are available. The left column is based on the assessment tables, and the right column complements missing values from the historical assessment tables where available.

values. Table A.24 reports the availability of the six most common building area codes that refer to some version of the total area.

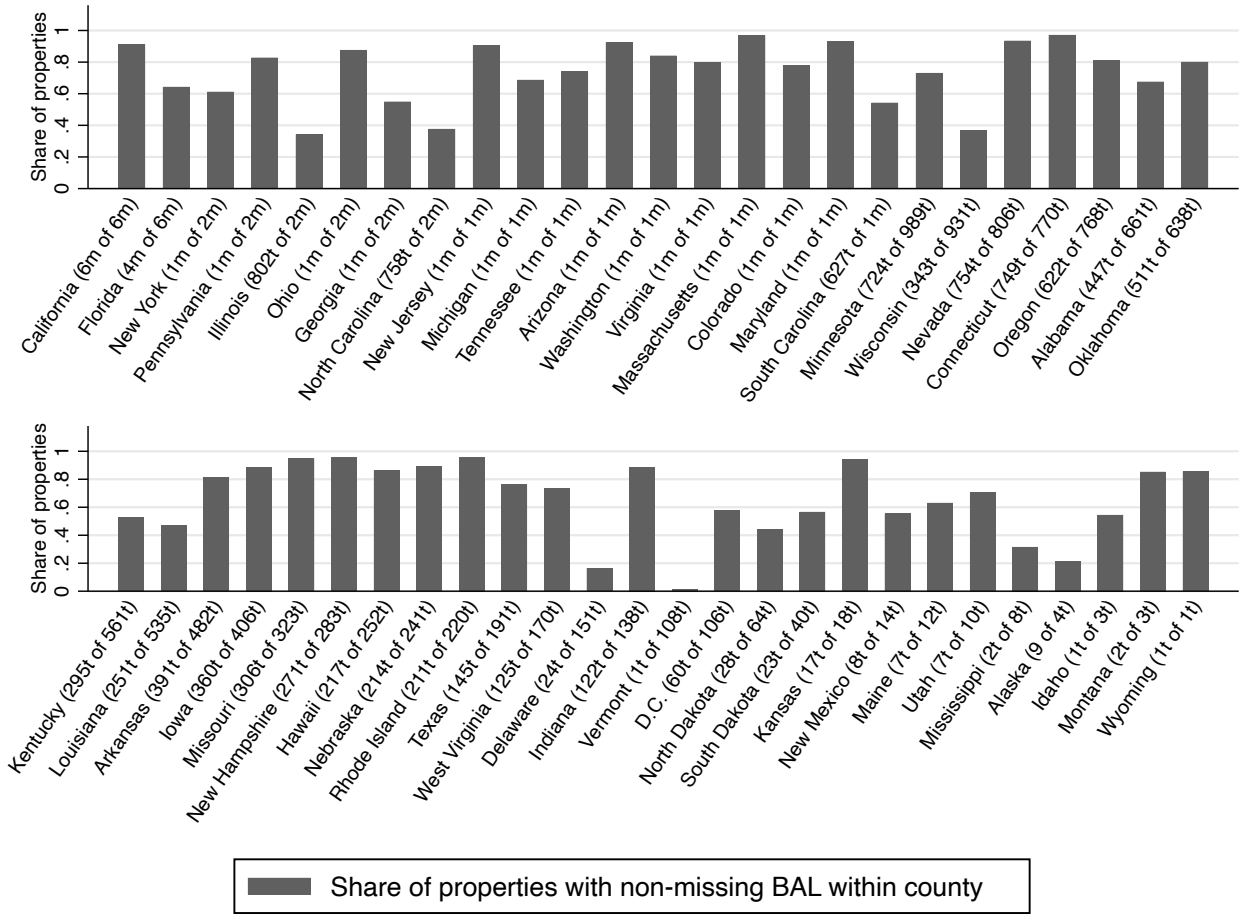
[32] We use the other codes sequentially in the following order: BAT, BAG, BAF, BAB. The median and interquartile ranges of the ratios are unity except for BAG, where the median is 1.2.

Figure A.9: Histogram of shares of non-missing living building area (BAL) within counties



Notes: The figure shows a histogram of the shares of properties with non-missing living building area (BAL) information within counties. The histogram is weighted by the number of properties within a county.

Figure A.10: Shares of non-missing living building area (BAL) within states



Notes: The figure shows the shares of properties with non-missing living building area (BAL) information within states. The states are ordered by the total number of properties in the sample. The number in parentheses indicates how many of the properties have non-missing BAL information within states.

D. Further Descriptives for Zillow data

This section provides further descriptive statistics for the Zillow data. The first two sections 1. and 2. provide statistics by state on identified arm's length transactions in Section A. above. Section 3. shows how often we observe repeat sales. Section 4. shows descriptives on the missing information in further property hedonics to show why we solely rely on price per square foot.

1. Share and number of retained and removed transactions by states

Table A.25 ranks the 50 states and D.C. by the number of retained transactions in Section A. above. Figure A.11 show the shares of retained and removed transactions by state and the sequential steps described above. Since most of the removed transactions are due to missing sales price in the first step, Figure A.12 shows the share of transactions removed only for the other steps.

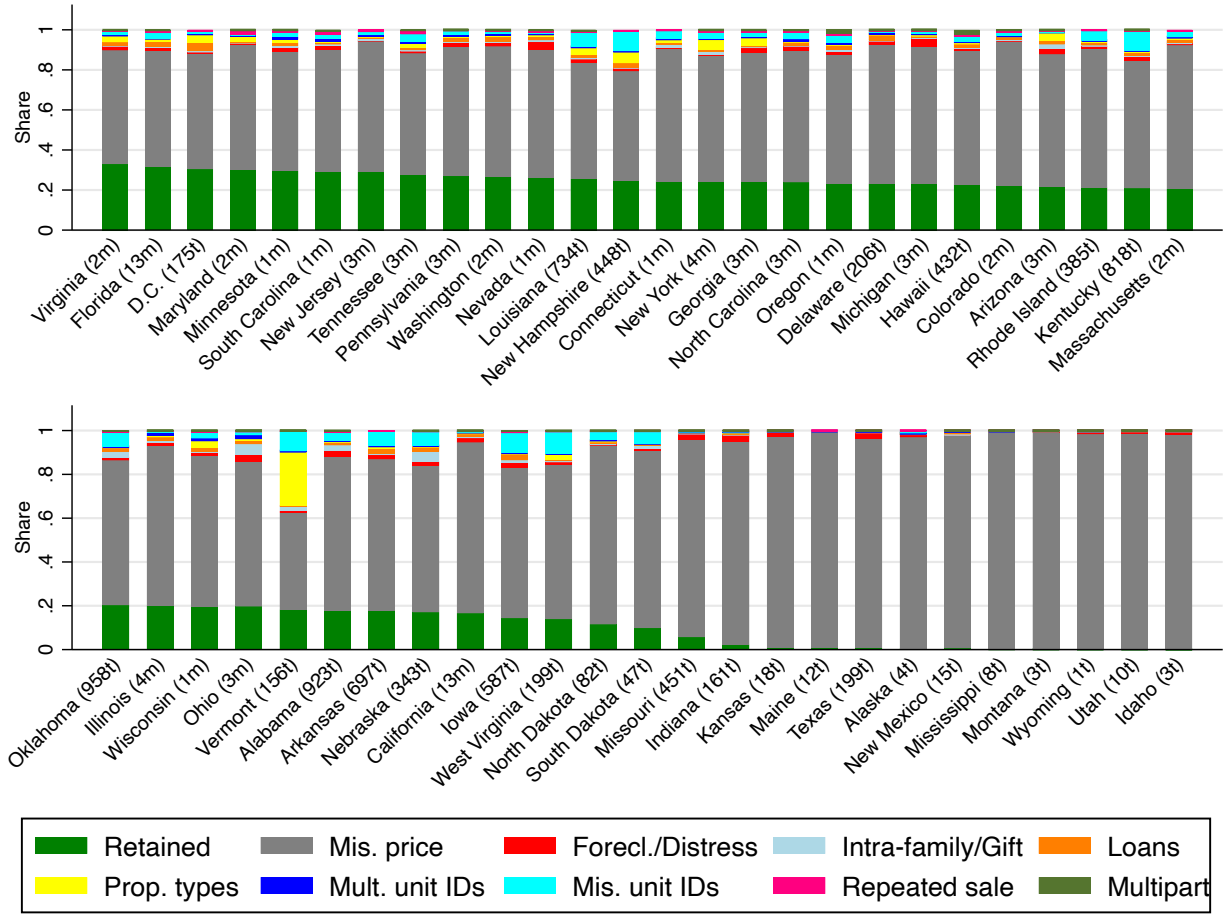
Vermont warrants a closer inspection of removed transactions. It has a high number of transactions removed due their property type. This is because at this stage of the cleaning process, Vermont has less than 2% of transactions with missing property types, compared to 48% in other states (see also Table A.15). Since transactions with missing property type are retained, there is a higher chance for a property to be removed due to its non-missing but non-residential property type in Vermont. Therefore the share of removed transactions due to property types is naturally larger in Vermont.

Table A.25: Number of retained transactions by state

No.	State	Ret. trans.	%	No.	State	Ret. trans.	%
1	Florida	13.1m	31.8%	27	Louisiana	734t	25.5%
2	California	13.0m	16.6%	28	Arkansas	697t	17.6%
3	New York	4.13m	24.1%	29	Iowa	587t	14.4%
4	Illinois	4.01m	20.1%	30	Missouri	451t	5.7%
5	Georgia	3.80m	24.1%	31	New Hampshire	448t	24.7%
6	Ohio	3.63m	19.7%	32	Hawaii	432t	22.7%
7	Pennsylvania	3.62m	27%	33	Rhode Island	385t	21.3%
8	Arizona	3.44m	21.9%	34	Nebraska	343t	17.1%
9	Tennessee	3.33m	27.8%	35	Delaware	206t	23.1%
10	New Jersey	3.27m	29%	36	West Virginia	199t	13.9%
11	Michigan	3.17m	23%	37	Texas	199t	0.7%
12	North Carolina	3.08m	23.9%	38	D.C.	175t	30.4%
13	Colorado	2.86m	22.4%	39	Indiana	161t	2.3%
14	Washington	2.75m	26.7%	40	Vermont	156t	18.3%
15	Massachusetts	2.67m	20.5%	41	North Dakota	82t	11.5%
16	Maryland	2.58m	30.1%	42	South Dakota	47t	9.8%
17	Virginia	2.46m	33%	43	Kansas	18t	0.8%
18	South Carolina	1.94m	29.3%	44	New Mexico	15t	0.6%
19	Nevada	1.72m	26.2%	45	Maine	12t	0.7%
20	Minnesota	1.70m	29.5%	46	Utah	10t	0.2%
21	Connecticut	1.40m	24.3%	47	Mississippi	8t	0.3%
22	Wisconsin	1.36m	19.9%	48	Alaska	4t	0.6%
23	Oregon	1.33m	23.3%	49	Idaho	3t	0.1%
24	Oklahoma	958t	20.3%	50	Montana	3t	0.3%
25	Alabama	923t	17.8%	51	Wyoming	1t	0.2%
26	Kentucky	818t	20.9%				

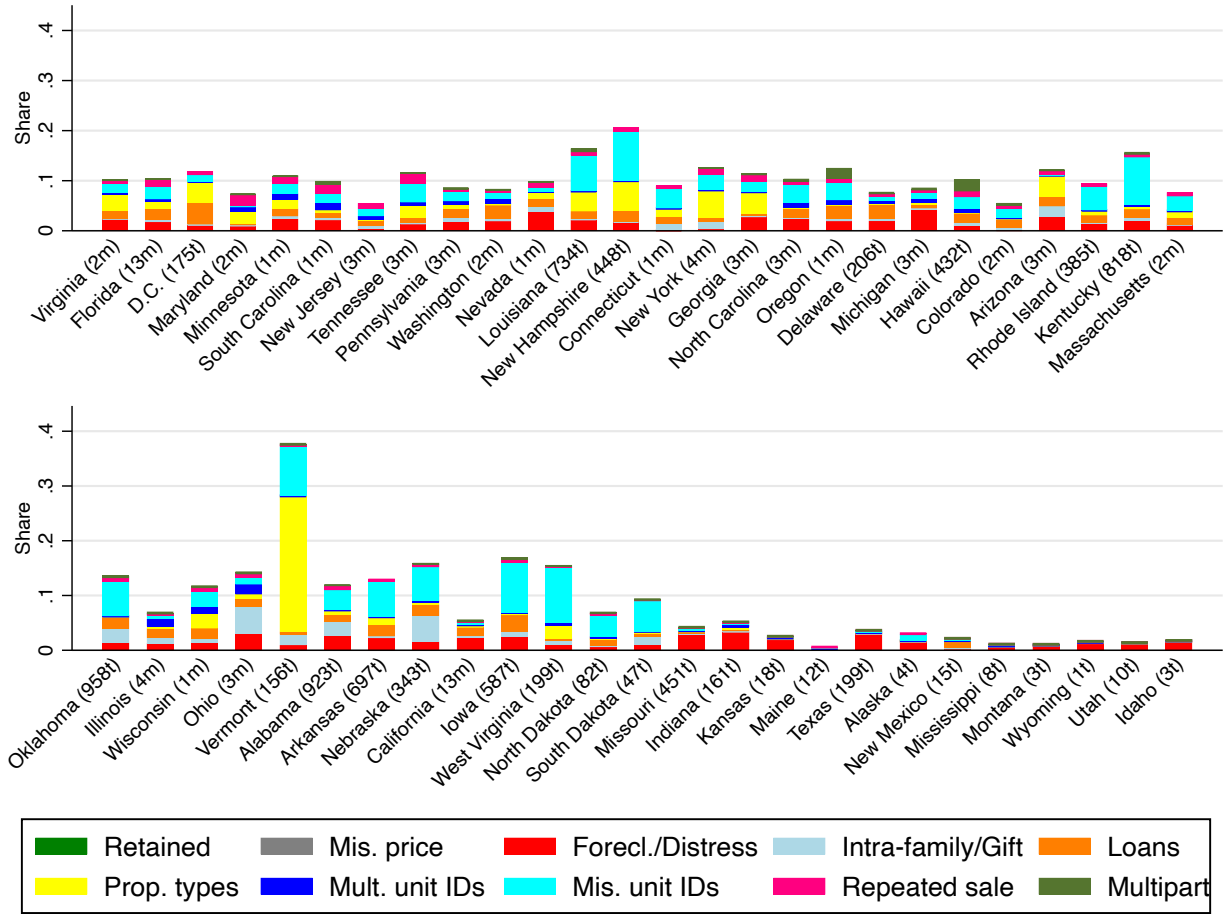
Notes: The table shows the number of retained transactions by state. The percentage is the share of retained transactions by state.

Figure A.11: Shares of retained and removed transactions by state and total number of transactions



Notes: The figure shows the shares of transactions retained and removed by the sequential steps described in the text. The states are ordered by the share of retained transactions. Total retained transactions by state are indicated in parentheses after state names.

Figure A.12: Shares of removed transactions by state and total number of transactions



Notes: The figure shows the shares of transactions removed by the sequential steps described in the text. The states are ordered by the share of retained transactions in Figure A.11. Total retained transactions by state are indicated in parentheses after state names.

2. Transaction coverage across time by states

Figures A.13, A.14 and A.15 plot the number of arm's length transactions per year by state from Section A. above. Some states have good coverage in the mid-90s already (e.g. California, Colorado, New Jersey), but generally, coverage starts to be good from 2000. This is one reason why we focus our analysis to begin in 2000. The other reason is the availablity of pollution data. Most states have a decline in transactions during the housing crisis 2006/2007 in the lead up to the financial crisis. Most states that have implausible spikes in certain years (e.g. Indiana in 2014 or Utah in 2011) also have a very low share of retained observations (2.4% and 0.2% respectively).

Figure A.13: Temporal transaction coverage by state (1/3)

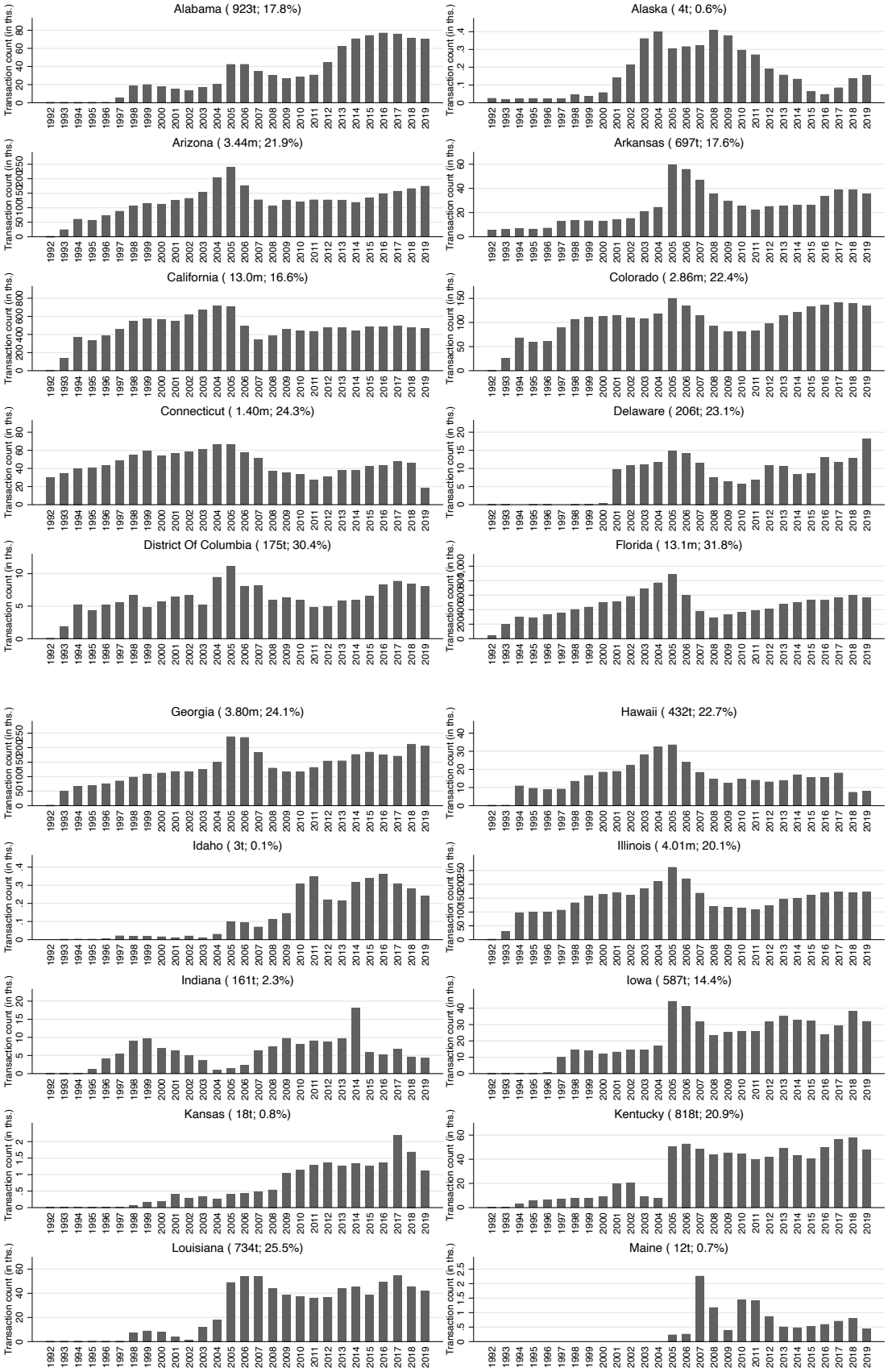


Figure A.14: Temporal transaction coverage by state (2/3)

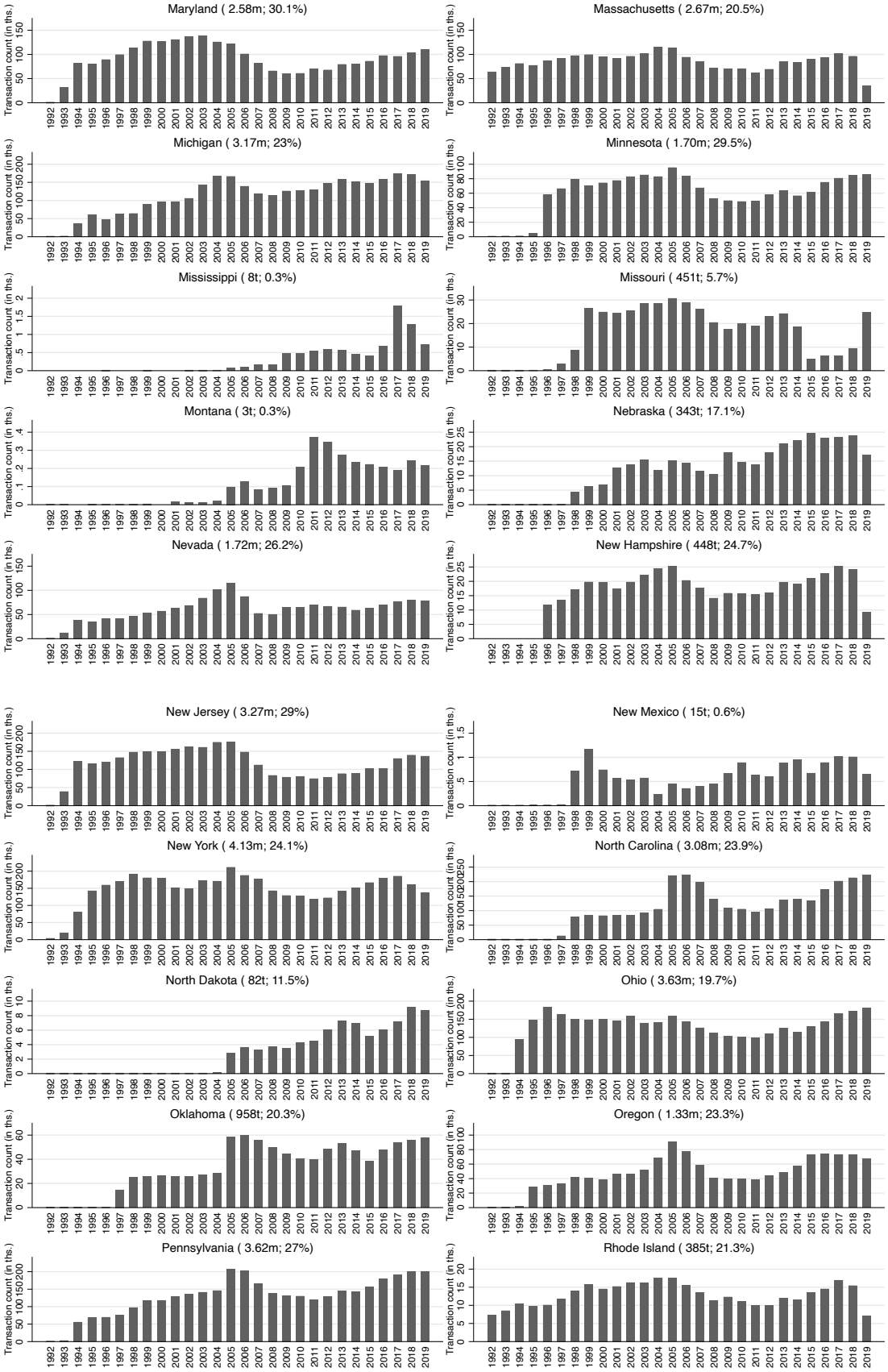
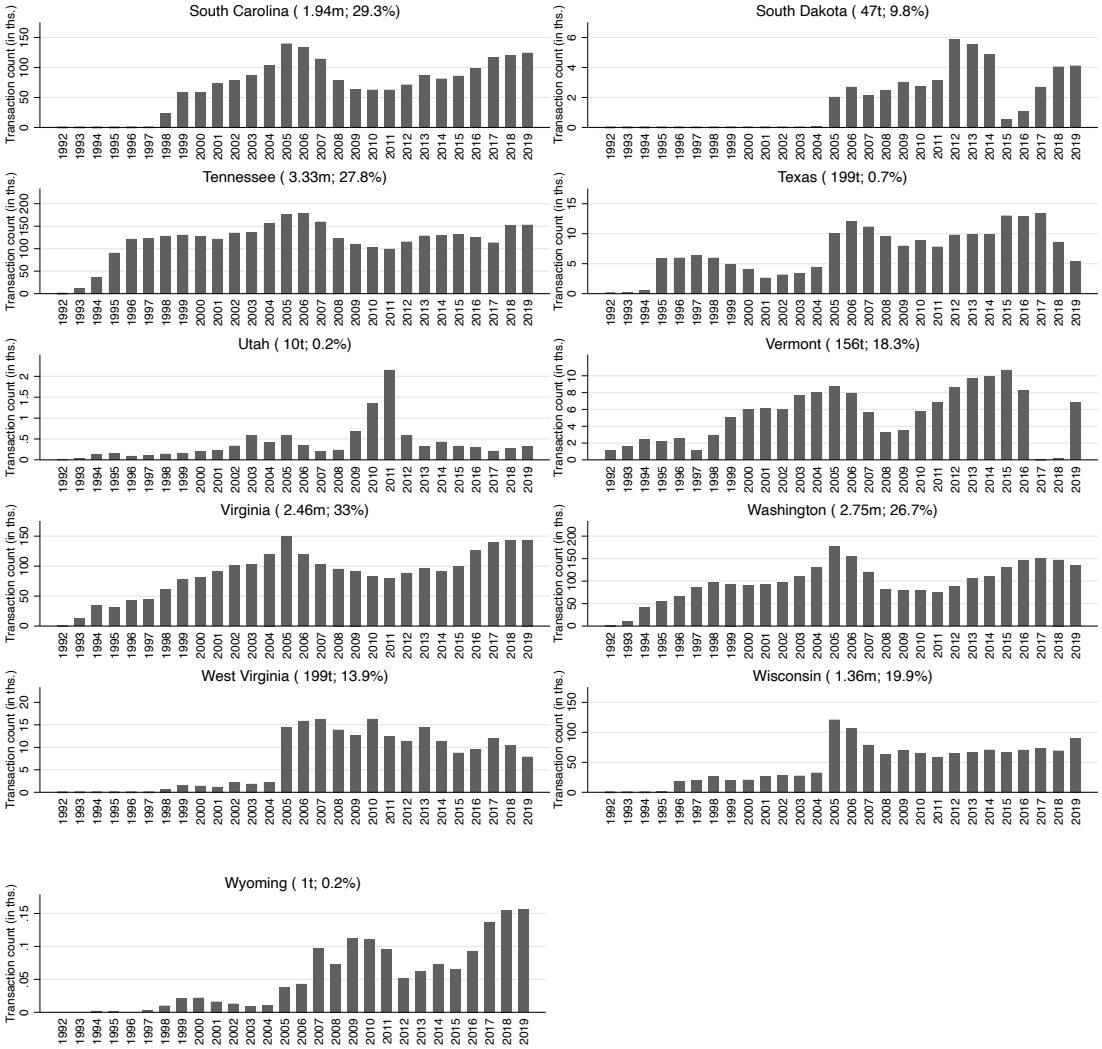


Figure A.15: Temporal transaction coverage by state (3/3)



Notes: The figures show the number of transaction by year by state. The total number of transactions, as well as the percentage of the raw data that is retained by state are in parentheses.

3. Distribution of sales frequencies and periodicity

Table A.26 shows how frequently individual properties have transacted during the sample period. 54.4% of properties have only been sold once, 26.9% twice, 12.0% three times, 4.6% four times, and roughly 2% at least five times.[33] We therefore do not rely on repeat sales as this would result in a much smaller spatial coverage.

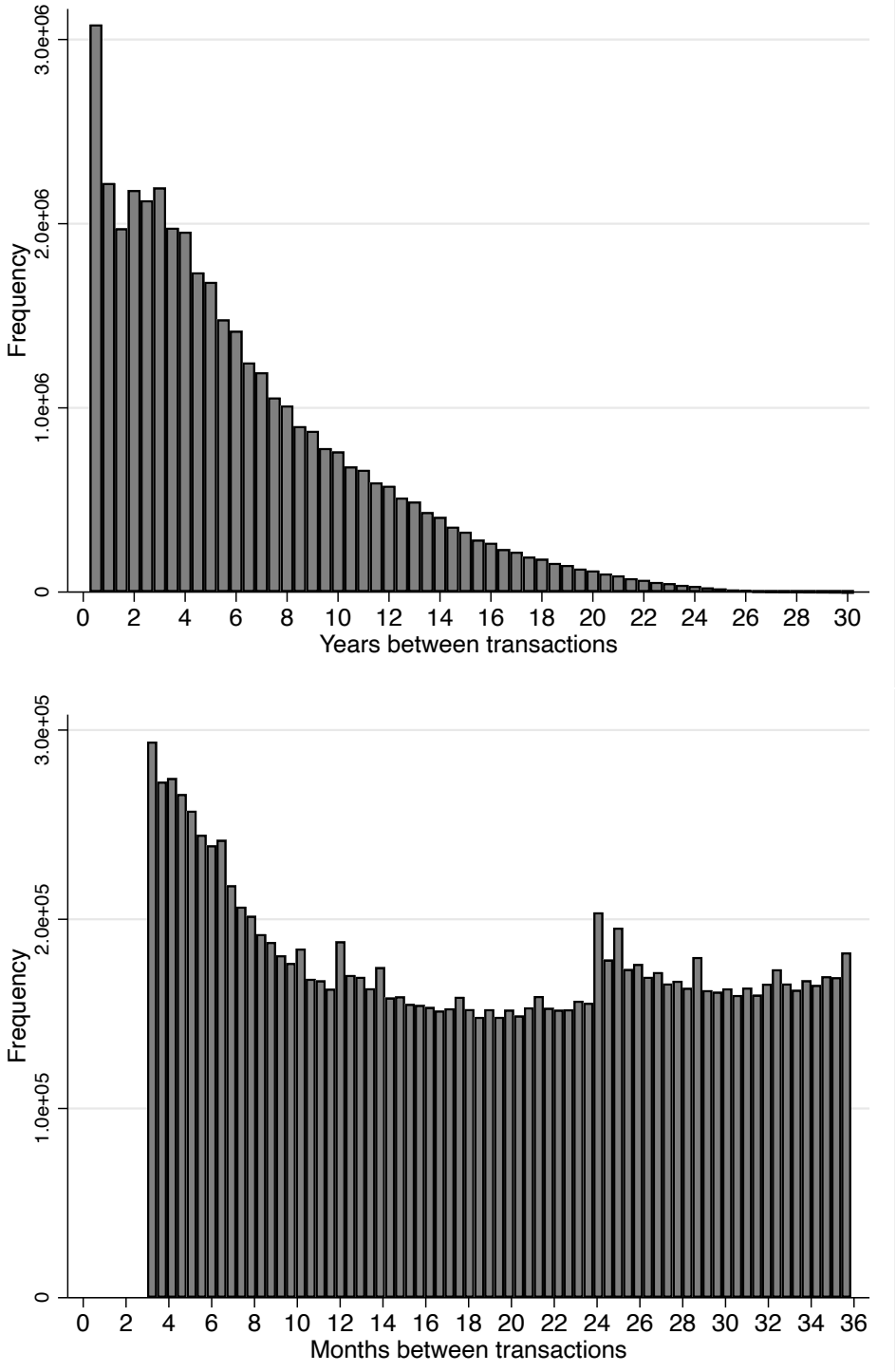
Table A.26: Transaction frequency

Transaction count	Num. of properties	Percentage
1	28902770	54.36%
2	14272318	26.85%
3	6375629	11.99%
4	2465689	4.64%
5	820823	1.54%
6	238692	0.45%
7	63904	0.12%
8	16477	0.03%
9	4530	0.01%
10	1554	0.00%
11	686	0.00%
12	384	0.00%
13	254	0.00%
14	201	0.00%
15	155	0.00%
16	108	0.00%
17	99	0.00%
18	55	0.00%
19	47	0.00%
20	38	0.00%
21	32	0.00%
22	26	0.00%
23	15	0.00%
24	13	0.00%
25	11	0.00%
26	10	0.00%
27	14	0.00%
28	11	0.00%
29	5	0.00%
30	2	0.00%
31	3	0.00%
32	2	0.00%
33	3	0.00%
34	1	0.00%
36	1	0.00%

Notes: The table shows the number and share of properties by how frequently they were sold in the transaction data.

[33] For those properties that have transacted at least twice, we can calculate the distribution of the duration between transactions for the same property. Figure A.16 shows the histogram of this duration with different truncations. Of those properties that transact at least twice during the sample period, more than 50% of transactions are within 5 years. The bottom histogram truncates the distribution at three years, to show that the distribution is reasonably smooth in the first three years. There are no observations for the first 90 days as these transactions are not defined as arm's length and have been removed.

Figure A.16: Transaction periodicity (truncated at 30 years and truncated at 2 years)



Notes: Both figures show a histogram of the periods between transactions of the same property. The top histogram is truncated at 30 years. There are 0.04% of transactions with a longer periodicity. The maximum duration between two transactions is 125 years. The bottom histogram is truncated at 3 years.

4. Further building characteristics

Finally, while there are additional hedonic variables in the data, we show that coverage is insufficient to include them as additional characteristics in our analysis.

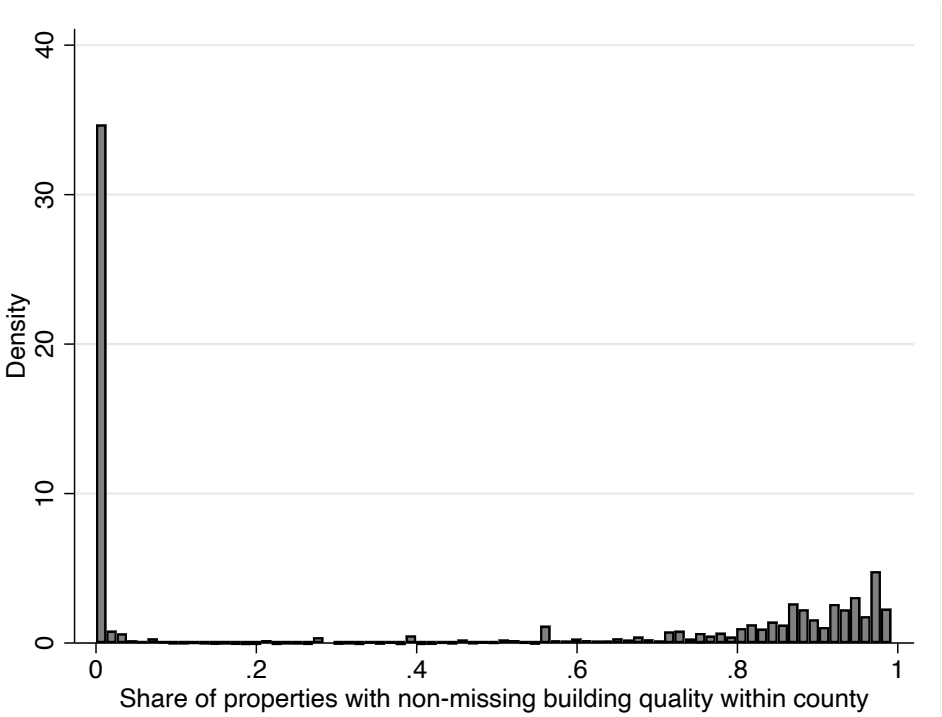
First, the number of bedrooms is positive for only around 65% of properties in our final data. The number of bathrooms only positive for around 70%. Second, Table A.27 shows the coverage for further building characteristics, such as building quality, building foundation, AC, heating, roof cover, story type (e.g. attic or basement), building condition, elevator presence and the number of stories in the property. Missing values in the assessment data have been complemented by most recent non-missing values from the historical assessment data where available.

The information on the building characteristics have initially been reported by the county offices. Information on heating is most widely available (63%). Air conditioning information is available for around half of the properties.[34] Building quality and building conditions are also only available for around half of the properties.

Is the missing data more widespread in particular counties or states? We first calculate the share of properties with non-missing building quality information both within counties or within states. Figure A.17 plots the histogram of the shares of non-missing building quality information within counties. There are many counties that do not report the building quality for any property. This is driving the missings in Table A.27 and the pattern is similar for the other building characteristics. Figure A.18 shows that there are some states (e.g. Massachusetts) in which no property has information on building quality.

[34] For air conditioning and heating, it is plausible that at least some of the missing values correspond to no AC/heating.

Figure A.17: Histogram of shares of non-missing building quality information within counties



Notes: The figure shows a histogram of the shares of properties with non-missing building quality information within counties. The histogram is weighted by the number of properties within a county.

Table A.27: Coverage of further building characteristics (in %)

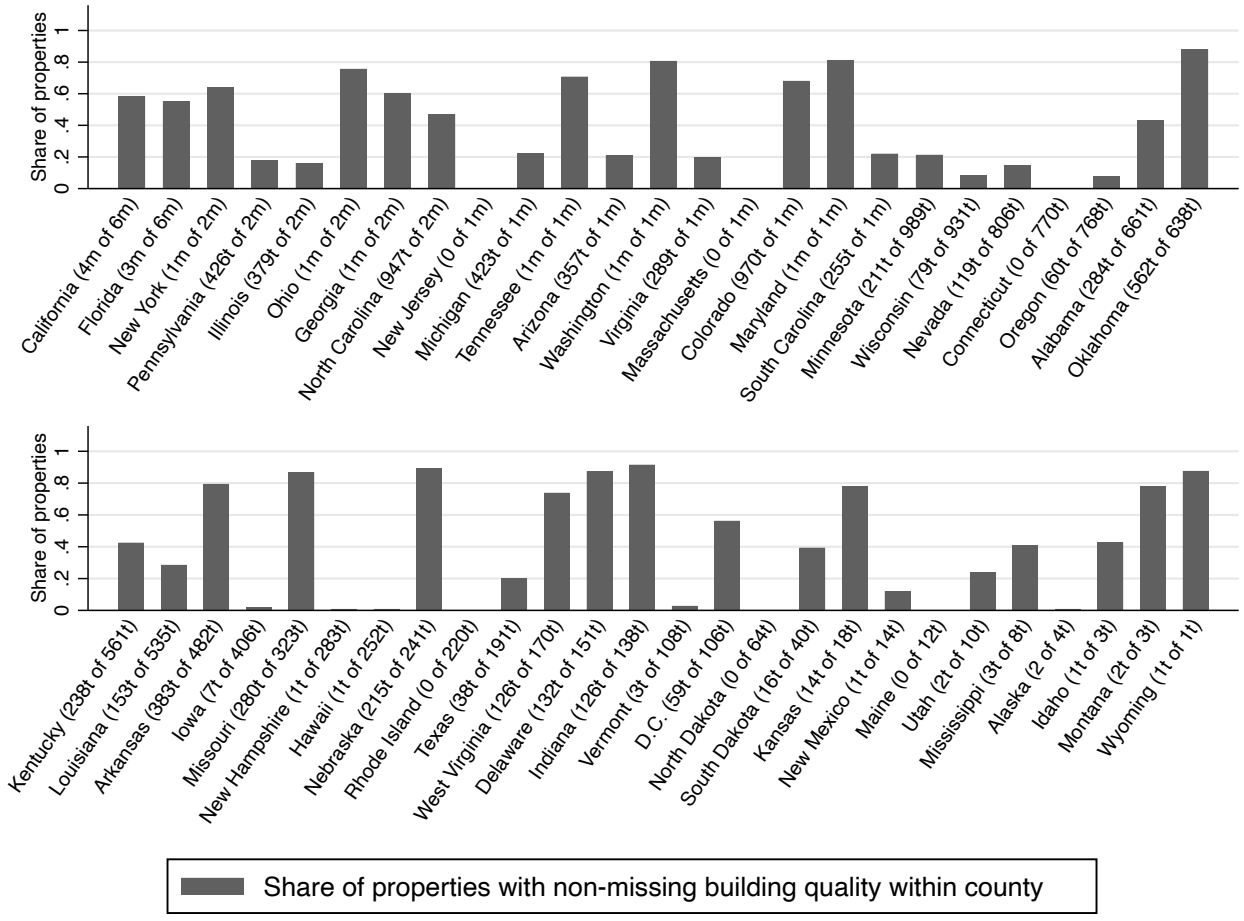
Building quality		**Building foundation**		**AC type**	
Missing	56.15	*Missing*	78.85	*Missing*	53.8
A	0.73	Concrete Block	1.8	Central	21.35
A+	0.28	Concrete	3.96	Chilled Water	0.03
A-	0.29	Crawl Space/Raised	3	Evaporative Cooler	0.68
B	5.39	Crossed Walls	0.04	Geo Thermal	0.01
B+	1.14	Earth/Soil	0	None	2.21
B-	1.31	Footing	2.61	Other	0.32
C	21.59	Masonry	1.01	Packaged AC Unit	0.84
C+	4.73	Mud Sill	0	Partial	0
C-	3.14	Other	1.03	Refrigeration	2.57
D	3.14	Pilings	0.07	Ventilation	0
D+	1.56	Piers	1.15	Wall Unit	0.1
D-	0.25	Raised Foundation	0.01	Window Unit	0.12
E	0.25	Retaining Wall	0.13	Yes	17.98
E+	0.04	Slab	5.89		
E-	0.02	Stone	0.19		
		Wood	0.27		

Heating		**Roof cover**		**Story type**	
Missing	40.02	*Missing*	58.4	*Missing*	65.1
Baseboard	1.03	Aluminum	0.04	”Attic & Basement”	1.86
Central	13.31	Asphalt	11.83	Attic	0.83
Coal	0.01	Asbestos	0.09	”Bi-Level with Attic & Basement”	0
Convection	0.1	Bermuda	0.01	Bi-Level	2.44
Electric	2.4	Built Up	0.91	Bi-Level with Attic	0.01
Forced air	19.56	Concrete	2.43	Bi-Level with Basement	0.16
Floor/Wall	1.38	Composition Shingle	12.52	Basement	22.22
Gas	2.17	Fiberglass	0.29	Level with Attic	0
Geo Thermal	0.01	Gravel/rock	0.27	Level	0.02
Gravity	0.1	Gypsum	0	Multi-Level	0.58
Heat Pump	3.24	Metal	1.51	”Split Level with Attic & Basement”	0
Hot Water	3.82	Masonite/cement shake	0.02	”Single Level with Attic & Basement”	0
None	0.96	Other	0.74	Split Entry with Attic	0
Oil	0.12	Roll Composition	0.28	Split Entry with Basement	0.08
Other	2.17	Shingle	6.53	Split Foyer with Attic	0
Propane	0.02	Slate	0.21	Split Foyer with Basement	0.09
Partial	0	Steel	0.04	Single Level with Attic	0.01
Radiant	0.38	Tar and gravel	0.33	Single Level with Basement	0.01
Steam	0.33	Tile	2.31	Single Level	4.5
Solar	0.09	Urethane	0.01	Split Level with Attic	0.02
Space/Suspended	0.13	Wood	0.28	Split Level with Basement	0.5
Vent	0.12	Wood shake/shingle	0.95	Split Entry	0.13
Wood Burning	0.03			Split Foyer	0.41
Yes	8.47			Split Level	0.88
Zone	0.03			”Tri-level with Attic & Basement”	0
				Tri-level with Attic	0
				Tri-level with Basement	0.02
				Tri-level	0.12

Building condition		**Elevator**		**No. of stories**	
Missing	49.91	*Missing*	98.74	*Missing*	24.83
1 - Excellent	1.57	Escalator	0	0 - Zero	0.46
2 - Good	8.8	No	0.94	1 - One	42.39
3 - Average	35.17	Yes	0.32	2 - Two	29.22
4 - Fair	3.54	y	0	3 - Three	2.17
5 - Poor	0.71			Four to ten	0.54
6 - Unsound	0.3			More than ten	0.39

Notes: The tables show the distribution of further building characteristics. Missing characteristics have been complemented with values from the historical assessment tables where available.

Figure A.18: Shares of non-missing building quality information within states



Notes: The figure shows the shares of properties with non-missing building quality information within states. The states are ordered by the total number of properties in the sample. The number in parentheses indicates how many of the properties have non-missing building quality information within states.