**FHFA STAFF WORKING PAPER SERIES**

# The Price of Residential Land for Counties, ZIP Codes, and Census Tracts in the United States

Morris A. Davis
William D. Larson
Stephen D. Oliner
Jessica Shui

October, 2020 (revised)
January, 2019 (original)

Working Paper 19-01

FEDERAL HOUSING FINANCE AGENCY
Division of Housing Mission & Goals
Office of Policy Analysis & Research
400 7th Street SW
Washington, DC 20219, USA



Please address correspondence to William Larson (william.larson@fhfa.gov).

**The Price of Residential Land for Counties, ZIP Codes, and Census Tracts in the United States**

Morris A. Davis, William D. Larson, Stephen D. Oliner, Jessica Shui
FHFA Staff Working Paper 19-01
October, 2020 (revised)
January, 2019 (original)

## Abstract

Data from millions of appraisals in 2012-2019 are used to estimate residential land prices, the share of house value attributable to land, and related statistics down to the census-tract level for areas that include the vast majority of U.S. population and single-family housing. The results confirm predictions about land prices from canonical urban models. Over 2012-2019, we show that land prices rose faster than house prices in large metro areas, boosting the land share of house value, while the land share fell in smaller metros. The data are available for download at https://www.fhfa.gov/papers/wp1901.aspx.[1]

**Keywords:** land prices · land leverage · price gradient · standard urban model · price dynamics

**JEL Classification: R14, R21, R32**

[1] These indices are works in progress and all data, tables, figures, and other results in this working paper are subject to change.

Morris A. Davis
Department of Finance and Economics
Rutgers Business School
Rutgers University
1 Washington Park Room 627
Newark, NJ 07102
mdavis@business.rutgers.edu

William D. Larson
Federal Housing Finance Agency
Office of Policy Analysis & Research
400 7th Street SW
Washington, DC 20219, USA
william.larson@fhfa.gov

Stephen D. Oliner
American Enterprise Institute for Public Policy Research
1789 Massachusetts Avenue, NW
Washington, DC 20036
stephen.oliner@aei.org

Jessica Shui
Federal Housing Finance Agency
Office of Policy Analysis & Research
400 7th Street SW
Washington, DC 20219, USA
jessica.shui@fhfa.gov

# 1 Introduction

Researchers have taken to describing a single-family house as a physical structure occupying some land: See Bostic, Longhofer, and Redfearn (2007), Davis and Heathcote (2007) and Davis and Palumbo (2008), for example. Because housing structures are infrequently renovated and construction costs change relatively slowly from year to year, rapid change in the value of housing typically occurs when the underlying land is appreciating or depreciating. For this reason, the housing boom and bust of 1998-2012 has been described as a land boom and bust by Davis, Oliner, Pinto, and Bokka (2017) and others.

Although the importance of studying and monitoring the price of land in residential use is now well understood, few studies have produced data on land prices at a relatively fine level of geography. Broadly speaking, researchers have used one of two methods to estimate the price of residential land. Both of these methods require data that have, until recently, been hard to acquire. The first method uses data from sales of vacant or near-vacant land. Three examples of the first method are Haughwout, Orr, and Bedoll (2008), Nichols, Oliner, and Mulhall (2013) and Albouy, Ehrlich, and Shin (2018). These authors all use data from the Costar Group, Inc. Haughwout, Orr, and Bedoll (2008) estimate the price of land inside the New York metro area; Nichols, Oliner, and Mulhall (2013) produce price indexes for land for 23 metro areas; and Albouy, Ehrlich, and Shin (2018) estimate the average value of urban land in nearly all metropolitan areas in the United States.

The second method measures the price of land as the difference between house value and the replacement cost of the structure on the land. Davis and Palumbo (2008) apply this method to data from the American Housing Survey to generate the average price of land for 46 metro areas. Davis, Oliner, Pinto, and Bokka (2017) use proprietary data on house prices and construction costs from a number of sources to generate the level of land prices and changes in land prices at the ZIP code level for the Washington, DC metropolitan area.[2]

In this paper, we use a huge database of home appraisals to produce annual panel data

[2]See Nichols, Oliner, and Mulhall (2013) and Davis, Oliner, Pinto, and Bokka (2017) for additional references in this literature. Note that a third possible method to generate land prices would be to use a hedonic regression to separate house value into its land and structure components. The difficulty, however, is that land value and structure value tend to be correlated across properties, which hinders the identification. In a prominent example of the limitations of this approach, Diewert, Haan, and Hendriks (2015) had to impose restrictions based on external structure cost data to obtain satisfactory results.

for the price of land in single-family residential use for 960 counties, 7,742 ZIP codes, and 10,515 census tracts from 2012 through 2019. These estimates of land prices cover 85% of the U.S. population and 83% of all single-family homes.[3] To our knowledge, ours is the first study to produce these estimates at a fine geography for nearly the entirety of the United States.[4] Our source data are from the Uniform Residential Appraisal Report submissions to the Government Sponsored Enterprises (GSEs), Fannie Mae and Freddie Mac. The reports are required by the GSEs before they purchase or securitize a mortgage. These data contain approximately 21.4 million unique appraisals for single-family homes submitted between 2012 and 2019.[5]

Our estimates of land values using this data set are based on "cost-approach" appraisals; we set land value equal to the appraised value of the house less an estimate of depreciated replacement cost of the housing structure.[6] A common concern about this residual method of estimating land values is that it assumes that the sum of the replacement cost of the housing structure and the value of the land (if it were vacant) is equal to the market value of housing. We show this assumption does not hold when a housing structure has become functionally obsolete and is due to be torn down or extensively remodeled. To address this issue, we calibrate a simple option model for tearing down and rebuilding a house. Simulations of the calibrated model suggest that the value of housing is well approximated as the sum of the replacement cost of the structure and the market value of the land if vacant for about the first 20 years of the life of the structure. As a conservative application of this result, we only use appraisals for relatively new homes with an effective age of no more than 15 years. We also eliminate appraisals that are nearly equal to or seem anchored to a public tax-assessor value due to concerns about the accuracy of assessed values as a measure of market value. Finally,

[3]Additionally, by aggregating county-level values using the March 2020 CBSA delineation file available at https://www.census.gov/geographies/reference-files/time-series/demo/metro-micro/delineation-files.html, we create panel estimates for 509 Core-Based Statistical Areas (CBSAs), all 50 states and the District of Columbia, and the United States as a whole. Note that we report pooled cross-section estimates of land prices for many more localities, including 915 CBSAs, 2,378 counties, 18,322 ZIP codes, and 53,144 census tracts.

[4]Albouy, Ehrlich, and Shin (2018) produce estimates for nearly all Primary Metropolitan Statistical Areas (PMSAs) in the United States, but they do not report on or make available data for any finer level of geography. On average, they observe 212 direct land sales per PMSA.

[5]Our data do not include condominiums but may contain a small number of 2-4 unit homes. The appraisals we use were submitted using Fannie Mae Form 1004. Appraisers are supposed to use Fannie Mae Form 1025 for 2-4 unit properties, although our understanding is that occasionally appraisers mistakenly use Form 1004 for these properties.

[6]We exclude data on vacant land sales due to the difficulty of controlling properly for differences in the characteristics of the vacant land, for example if water and sewer lines are in place.

because the quality of our land value estimates depends on having an accurate measure of structure value, we only use appraisals where the main reported source of cost information is Marshall & Swift, owned by CoreLogic, or R.S. Means, owned by Gordion, two widely used sources for construction costs. After applying all filters, our working sample contains 6.7 million appraisals, 31% of the original sample.

We generate two sets of estimates of land value, "as-is" and "standardized." Our as-is estimates report the value of land per-acre, without any adjustments or corrections. Our standardized estimates report the price of land per quarter-acre, roughly the median sized lot in our data, after adjusting for the fact that the price of land per acre tends to fall as acreage increases, the so-called "plattage effect." We generate the standardized estimates using a two-step method. First, we use the procedure of Davis, Oliner, Pinto, and Bokka (2017) to adjust for the effect of lot size on land prices and compute the price of land per quarter acre for each assessed property in our working sample. Then, we use a procedure called Kriging, as described by Basu and Thibodeau (1998), to interpolate this standardized price-per-quarter-acre of land to lots under all remaining single-family housing units in a given geography (county, ZIP code, or census tract).[7] Using a 20% hold-out sample, we show in our data that Kriging offers a lower root mean square error in interpolating land prices than some other commonly used methods of spatial interpolation. In the online appendix to this paper (see Davis, Larson, Oliner, and Shui (2020)), we analytically derive a land-price gradient from a simple, calibrated urban model and show that when we simulate data from that model, Kriging delivers the correct land-price gradient from model-simulated data. The procedure we use to compute the as-is estimates is exactly the same as with the standardized estimates, except we add one step at the end to undo the correction for plattage effects.

The primary goal of the paper is to generate land values and indices covering the United States at a fine geography for use by researchers and policy-makers. To that end, the aggregated land-price data we generate in this paper are available for download at the web site of the Federal Housing Finance Agency, at `https://www.fhfa.gov/papers/wp1901.aspx`. For each county, ZIP code, and census tract where we compute land prices, and for broader geographies built up from the county data, we post a variety of useful statistics, including land shares of total property value and both as-is and standardized land prices.

[7] We obtain our universe of single-family housing units in a given geography from the assessor data licensed from CoreLogic. These data contain nearly all parcels for all land use types in the counties for which CoreLogic has acquired data.

In addition, we use the data to develop and confirm important stylized facts about land prices in the United States. First, as shown by Albouy, Ehrlich, and Shin (2018) and others, the level of land prices at the center of a metro area varies greatly. For example, the price of land at the center of metro areas with more than 2 million single-family housing units is more than 25 times greater than the price of land at the center of metro areas with less than 500 thousand housing units. Second, the rate at which land prices decline from the city center also varies across metro areas. Finally, the price of land covaries with certain variables in accordance with predictions of classic models of urban economics. In particular, measured at the average price per acre in a ZIP code, residential land prices are negatively correlated with lot size and are positively correlated with the size of the housing structure on a lot.

## 2 Option Model of Housing Teardowns

In this section, we build a simple model for when a land owner should optimally tear down his or her house and rebuild. We use the model to develop a rule-of-thumb to determine the oldest existing homes for which we can derive unbiased estimates of land value from a cost-approach value decomposition.[8]

In the model, the land owner owns property with a building of size $S$ on a lot of size $L$. The lot size is fixed in perpetuity, but the building size can be changed. This property earns rents of

$$q^H S^{1-\phi} L^{\phi} \tag{1}$$

where $q^H$ is the rental price per unit of housing service provided and the Cobb-Douglas aggregate of structures and land, $S^{1-\phi} L^{\phi}$, is the number of units of housing services.

Each period, the land owner must decide whether or not to demolish the building and rebuild on vacant land, or to let the building depreciate some and revisit the choice next period. Since $L$ is fixed in perpetuity, we can summarize the decision problem of the land owner as one over a choice of $S$ only. Denote $V(S)$ as the value of owning a property with a building of size $S$ and similarly let $V(0)$ denote the value of the property as vacant land. When the

[8] Although we are the first to develop such a rule-of-thumb, many other papers have studied the option value of development, for example Titman (1985), Clapp, Eichholtz, and Lindenthal (2013) and McMillen and O'Sullivan (2013).

land is not vacant and a structure of size $S$ exists on the property, the land owner chooses either to let the property sit as is and collect rents, or to knock the property down and make the land vacant. This problem has the expression

$$V(S) = \max\left\{q^H S^{1-\phi} L^{\phi} + \beta V(S(1-\delta)), V(0)\right\} \tag{2}$$

$\beta$ is the factor by which the land owner discounts the future and $\delta$ is the rate at which the housing structure depreciates. The first term in the max operator is the value of the property with the structure left intact. This term includes the discounted value of owning a property with a structure of size $S(1-\delta)$ next period. The second term is the value of the property when it is made vacant, assuming there are no demolition costs that must be paid to clear the land of the structure.

Denote $p^S$ as the price per unit of newly-built structure. When the land is vacant, the land owner chooses to build the optimally-sized structure to maximize the value of the land. The choice determines the rents earned this period, plus the discounted value of the property in the future after accounting for depreciation of the structure, less the cost of building the structure. This choice satisfies:

$$V(0) = \max_S \left\{ q^H S^{1-\phi} L^{\phi} - p^S S + \beta V(S(1-\delta)) \right\} \tag{3}$$

The solution to this model can be characterized by two variables: $\bar{S}$, the size of the structure that is built when the land is vacant,[9] and $\underline{S}$, the smallest structure that exists (i.e. a structure of any smaller size is demolished). Using the relationship $\underline{S} = \bar{S}(1-\delta)^T$, where $T$ represents the maximum age of any housing structure, $T = \left(\log \underline{S} - \log \bar{S}\right)\left[\log(1-\delta)\right]^{-1}$.

We solve this model and calibrate it as follows. We set the discount rate to $\beta = 0.90$; the annual depreciation rate to $\delta = 0.023$ (Harding, Rosenthal, and Sirmans, 2007); land's share of value to $\phi = 0.30$; the price per unit of structure, a normalization, to $p^S = 1$; and then we find the level of rent $q$ such that the value of housing for a newly built optimally-sized structure is $V\left(\bar{S}\right) = 100$. This calibration delivers simulation results that the oldest house is 80 years old; the smallest (most depreciated) house size at the time of demolition $\underline{S} = 10.8$ with the value at that house size equal to the value of the underlying land $V(0) = V(\underline{S}) = 30$;

[9] This is the argmax of equation (3).

and the optimal house built on vacant land is $\bar{S} = 70$.

The top panel of Figure 1 shows how housing value (blue dashed line) and the replacement cost of structures (red dot-dash line) change with the building age in this model. The replacement cost of structures declines at a constant rate from 70 when newly built to 10.8 after 80 years, at which point the structure is torn down. The value of the vacant land is always 30 (solid black line). The value of housing declines gradually over time from 100 when newly built to 30 after 80 years. When the structure is torn down, the value of housing (30) is less than the sum of the value of land and the replacement cost of the structure (40.8 = 30 + 10.8).

Using this calibrated model, we investigate the age of the structure at which house value is no longer well-approximated as the sum of the value of the vacant land and the cost of the depreciated housing structure. In the bottom panel of Figure 1 we compute a "land share" of house value two different ways, and determine the age of the housing structure at which these two methods stop producing similar results. The first (correct) method, the solid black line, computes land's share of value as the ratio of the value of vacant land to the value of housing, $V(0) \;/\; V(S)$. This method shows that land's share of value increases monotonically from 30% for newly built homes to 100% for homes about to be torn down. The second method, the red dashed line, computes land value residually as house value less the replacement cost of structures, $V(S) - S$. This is meant to approximate how land is measured residually when given an appraised value of housing, $V(S)$, and an estimate of the depreciated reconstruction cost of the housing structure, $S$. Land's share of housing is this residually-measured land value divided by house value. This panel shows that this measure of land's share of housing ranges from 30% for newly constructed homes to only about 60% for homes about to be torn down. The value of housing at the point of teardown is 30, entirely equal to land value. A residually-measured estimate of land value at the point of teardown would be biased down and only equal to $30 - 10.8 = 19.2$. The figure shows that the two methods produce nearly identical estimates of land share of value for structures that are younger than 20 years old.[10] This result guides restrictions to our sample of data that we describe later.

[10] At 20 years old, the correctly-measured land share is 41.6% and the residually-measured land share is 39.2%.

Of course, different models will produce different results. The point of this section is not to write down the most realistic model of land ownership and teardowns. Rather, it is to gain intuition about the value of the option to tear down a house and how that option affects residually measured estimates of land's share of house value. Our general results should be robust to any model where an optimal teardown occurs decades after a house is built. The reason is as follows: When a house is relatively newly built, the expected date of a teardown is so far away in the future that the option of tearing down the house has little value. Since this option has little value and depreciation is low, the value of housing is well approximated as the sum of the replacement cost of the structure and the value of the vacant land.

# 3 Data

In each mortgage appraisal, there are typically three separate approaches to estimating the value of the underlying property. The first is the sales comparison (or "comps") approach, by which an appraised value is generated based on recent comparable transaction prices. A second "income" approach estimates the value of the property as the discounted flow of imputed rental income. Finally, the "cost" approach attempts to separately estimate the cost of the components of the property, the land and the structure, and assumes the estimated value of the property is the sum. We use cost-approach appraisals in our analysis.

The data on cost-approach appraisals are from Uniform Residential Appraisal Report submissions collected by the GSEs. After data cleaning, including the removal of duplicate and/or resubmitted appraisals, we have approximately 21.4 million unique cost-approach appraisal records submitted between 2012 and 2019.[11] The Uniform Appraisal Dataset is not exclusive to mortgages purchased by the GSEs. Even though the appraisals are posted on a GSE platform, in 2013-2014, only about 2/3rds of the appraisals were for GSE loans; the other 1/3rd were either for loans packaged into Ginnie Mae mortgage-backed securities (mainly FHA/VA loans) or were for loans held by lenders on their books. The latter set of loans is not constrained to have amounts below the applicable conforming loan limit and

[11] In our pre-cleaning of the data to arrive at this count, we exclude appraisal records with 1) lots smaller than 500 square feet or larger than 2 acres; 2) missing property value or property value less than $10,000; 3) cost-approach-estimated site value missing or less than $200; 4) land-price-per-acre smaller than $200; 5) site value greater than cost-approach-estimated property value; 6) missing depreciation information or depreciation at least three times greater than the contract price or the appraised value; 7) land share of property value less than 1% or greater than 99%; 8) structure-land area ratio equal less than .01 or greater than 4; 9) construction date before 1850 or after 2019.

indeed the dataset includes homes with appraised values far above the applicable loan limit.[12]

We wish to include in our sample only those cost-approach appraisals where the land component is an unbiased estimate of the market value of land. This likely occurs when three conditions hold: First, the option value of redevelopment is low; second, appraisals are not anchored to tax assessments, which are likely to be biased for reasons we discuss later; and third, the estimates of construction cost are from a known provider, either Marshall & Swift or R.S. Means. We now discuss how we adjust our sample to ensure these conditions hold.

To address any biases arising from the option value of redevelopment, we use the *effective age* variable in the appraisal dataset to eliminate observations from our working sample. Appraisers compute effective age as follows:

$$\text{Economic Life} \times \left( \frac{\text{New Replacement Cost} - \text{Depreciated Structure Value}}{\text{New Replacement Cost}} \right) \tag{4}$$

For example, for a structure with an assumed economic life of 80 years, a depreciated structure value of \$100,000, and a replacement cost of the structure as new of \$150,000, the effective age would be $80 \times (50/150) = 26.7$ years.[13]

The results from our calibrated model of section 2 suggest that the residual method produces accurate estimates of the value of land for structures less than 20 years old. Given an assumed constant depreciation rate of 2.3% per year (Harding, Rosenthal, and Sirmans, 2007) and a maximum economic life of a given house of 80 years, the calibrated model implies the maximum effective age from equation (4) for reliable cost-based appraisals of land should be

[12] Lenders have multiple reasons for posting appraisals to this platform without selling the mortgage to the GSEs. For example, many smaller lenders do not have the infrastructure to store their own appraisals, so they outsource that capability to this platform.

[13] The U.S. Bureau of Economic Analysis uses a service life of 80 years for new 1-4 unit residential structures; see Katz and Herman (1997). We have been asked if appraisers consider obsolescence when computing effective age, for example would the effective age of a "shrink-wrapped" house built 50 years ago be nearly zero, because the house was perfectly preserved, or would the effective age be considerably higher to take into account improvements in technology and changes in design demanded by modern households in newly constructed housing. Based on our interpretation of the guidance offered by authorities in the field to appraisers, we believe appraisers should correct for economic obsolescence, suggesting a shrink-wrapped house from 1970 would not be assigned a near-zero effective age. The key is that appraisers are supposed to account for the utility of the structure in addition to its condition when determining effective age. See, for example, Dzierbicki (2014) and Siebers (2016) for details.

approximately

$$80 \times \left[1 - (1 - 0.023)^{20}\right] = 29.8 \text{ years} \tag{5}$$

In our analysis, we conservatively restrict our data to appraisals reporting an effective age to half of this value, 15 years or less, corresponding to an age of 8.9 years in the model of section 2.[14] This filter eliminates 32.2% of the working sample, reducing it from 21.4 million observations to 14.5 million observations.[15]

With respect to the second potential bias in the appraisal data set, for a variety of reasons appraisers sometimes anchor their reported appraisal to existing estimates of value such as the contract price or tax assessments. The anchoring of appraisals to contract prices does not seem problematic for our purposes as contract prices should reflect market values. However, if an appraiser anchors to a tax assessment this may impart a significant downward bias to estimates of the value of land. Lutz, Molloy, and Shan (2011) and others show that tax assessments are often biased relative to transaction values for two reasons. First, the rate of change of tax assessments is sometimes capped to prevent these assessments from rising too quickly in rapidly-appreciating housing markets. Second, assessments can be right-censored near the market value: assessments less than or equal to the market value are unchallenged by the property owner, but assessments significantly greater than the market value can be challenged by the property owner and potentially adjusted down to the market value.

To illustrate the size and nature of potential biases, the top panel of Figure 2 shows a histogram of the ratio of appraised value to tax-assessed value for all housing units for which we have an appraisal and a tax assessment.[16] The top panel shows the ratio for total property value, inclusive of land and structures, while the bottom panel shows the ratio for the land component alone. Both panels clearly show that appraised values tend to be higher than tax assessments. In addition, the spike at the value of 1.0 in the bottom panel shows that a sizeable mass of appraisers anchor their appraisals of land value to the tax-assessed value of

[14] At 8.9 years, the correctly-measured land share in the model is 34.8%, trivially different from the residually-estimated land share of 34.3%. We obtain 8.9 years as the solution for $x$ in $80\left[1 - (1 - 0.023)^{x}\right] = 15$.

[15] At the request of a referee, we experimented with a higher depreciation rate in the model of 3.5% per year. At this rate of depreciation, the optimal maximum age of the structure decreases to 55 years, and an effective age of 15 years (also) corresponds to a model age of 8.9 years. At 3.5% depreciation, at 8.9 years the correctly measured land share of the house is 37.5% and the residually estimated land share is 36.4%.

[16] Not all properties in the appraisal database have a tax assessment.

land. To remove this potential source of bias from our results, we exclude from our working sample any property with an appraised value within 2% of the assessed value for either the land component or total property value. We also remove any property for which the appraisal for land or total value is within 0.5% of a 5 percentage point notch (5%, 10%, and so on) above or below the assessment, which may indicate a softer form of anchoring. This reduces our working sample from 14.5 million to 11.5 million observations.[17]

Our last filter removes construction cost estimates that are not based on a well-known source with a market incentive to produce accurate estimates. The two such sources in our data are the commercial firms Marshall & Swift and R.S. Means.[18] This filter removes an additional 4.8 million appraisals from the remaining sample. Our final sample includes 6.7 million valid cost-approach appraisals – 31% of the original sample – from which we estimate land values.[19]

[17]For most counties, we have tax assessment data for 2012, 2013, 2017, 2018 and 2019; for some counties we have fewer years. In all cases, we assign values for missing years using a straight-line interpolation for both house value and land value. If an imputed value lies within 5% of the value recorded in the appraisal data in the year of the imputation, the appraisal is dropped from our working sample. Although we have removed what we believe is obvious anchoring, there may still be some residual anchoring in the data we have not identified.

[18]The other listed sources of construction costs are "Local Market Data," "None," "Internet/Software," "Cost Handbook," and a few other sources that comprise less than 1% of appraisals. For appraisals that list more than one source for construction costs, we include the appraisal in our sample as long as Marshall & Swift or R.S. Means is one of the listed sources.

[19]As a final thought on the suitability of our sample to estimate land prices, some readers may be concerned about the difficulty of parsing housing values separately into land and structures components in declining areas with no construction activity based on intuition from Glaeser and Gyourko (2005). Of the 960 counties in our panel data set, all but five had some residential building permits in every year and each of those five have some building permits in at least one year. Thus, our county-level results should be largely immune to this concern, though there likely are some ZIP codes and census tracts where a lack of construction could affect the land price estimates.

# 4 Computing Land Values and Land Shares

## 4.1. Overview

Our working sample, while quite large, covers only a fraction of all single-family properties. In this section, we describe how we estimate land values for *all* single-family properties.[20] We produce two estimates of land value for every parcel. In the first, our standardized estimates, we compute the value of land as if every parcel was exactly 0.25 acres. Standardizing lot size corrects for the fact that the price-per-acre of land tends to decrease with acreage, all else equal, a relationship known as the plattage effect. In the second set of estimates, we estimate the value of land for each parcel without correcting for plattage effects. We call these our as-is estimates, which we report on a per-acre basis. The public-use data files posted on the FHFA website provides both measures of land value. For both measures, we report the average value across all single-family parcels by county, ZIP code, and census tract, along with aggregations of the county-level data to CBSAs, states, and the U.S. as a whole.[21]

## 4.2. Standardized Estimates of Land Value

In this section, we estimate the value of land for every parcel in a given geography as if the parcel was exactly one-quarter acre. We start by converting the value of land for parcels in our working sample to a quarter-acre equivalent. This is more complicated than taking the value of land on (say) a one acre parcel and dividing by four. Instead, we use a regression procedure similar to that described by Davis, Oliner, Pinto, and Bokka (2017). For each county, we pool the data in all years. Then, county by county, we regress the log of the value of land on the log of lot size, including ZIP-code fixed effects and year dummies as regression controls. That is, denoting $Y_i^L$ as the log of the value of land and $X_i^L$ as the log of lot size (measured in acres), we run the regression

$$Y_i^L \;=\; b_0 \;+\; X_i^L b_L \;+\; Z_i b_Z \;+\; \nu_i^L \tag{6}$$

[20] We have been asked why we interpolate the data in our working sample to produce estimates of land value for all parcels when we only report averages. The intuition behind our procedure is that the relatively young homes in our working sample can be geographically clustered, and by interpolating land values to all parcels (and then computing averages) we undo any effects of this clustering. This approach can be viewed as an imputation index similar to Hill and Melser (2008). Later on, we show using a hold-out-sample analysis that our estimates of land value are more accurate than simple averaging of the data in the working sample, providing some validation of the intuition.

[21] When ZIP codes span multiple counties, the reported value is the average of values in each represented county, weighted by the share of single-family housing stock.

where $Z_i$ are the ZIP-code fixed effects and year dummies and $\nu_i^L$ is the error term.[22] For each parcel in our working sample, we compute the predicted log price per quarter acre as the observed price plus our (county-specific) estimate of $b_L$ times the difference of the log of a quarter-acre and the log of the actual size of the lot, i.e. we compute

$$\widetilde{Y}_i^L \;=\; Y_i^L \;+\; \left(\log 0.25 - X_i^L\right) b_L \tag{7}$$

where $\widetilde{Y}_i^L$ is the predicted log price of parcel $i$ if it were 0.25 acres instead of its actual size.

Next, we merge our working sample of data with the 2017 vintage of single-family parcels from assessor data licensed from CoreLogic. These data contain the near-census of parcels in the counties for which it has acquired rights to the data.[23] After this merge, we assign a standardized, quarter-acre log land value to each single-family property in the assessor data that is not included in our working sample using an interpolation procedure called Kriging. After Kriging, we convert predicted log standardized land values to levels.

The Kriging procedure is commonly used in the hard sciences. For our purposes, we need a method that can be used in urban areas with steep and varying gradients over short distances, and in rural areas with relatively flat gradients and geographically sparse transactions. We also need a method that is computationally manageable to loop over thousands of areas that include millions of parcels. Kriging satisfies these requirements.[24] To our knowledge, there have been no studies in the land-price literature that have evaluated the relative accuracy of different spatial imputation methods. Later on, we compare the accuracy of Kriging to a number of alternative interpolation procedures in a 20% holdout sample. In this sample, Kriging produces more accurate estimates.

In the online appendix, we discuss the Kriging procedure in detail. Here we provide a brief

[22] The county-level estimates of $b_L$ are reported in the dataset for the paper posted on the FHFA website. The dataset also includes the estimated parameters from similar regressions used in the estimation of land shares, as discussed in section 4.4.

[23] We maintain this fixed sample of properties in order to avoid composition bias, effectively turning the sequence of average land values we report in our panel data into a Laspeyres measure.

[24] Basu and Thibodeau (1998) conduct an analysis of spatial autocorrelation in housing prices by comparing predictions from hedonic models to models with spatially autocorrelated errors. They find that traditional hedonic models are more accurate when unexplained price variation is spatially uncorrelated; otherwise, Kriging is more accurate. Our method of standardizing-then-Kriging land prices is analogous to their approach. This technique is also referred to as "Regression Kriging."

summary. Before we do so, the key takeaway is that Kriging, like other estimators, uses a weighted-average of $n$ nearest neighbors to generate predicted land prices. What makes Kriging different is its algorithm to generate the weights.

Derivation of those weights proceeds in five steps. The first step involves calculating pair-wise differences in values between each pair in the sample within a certain distance range, approximately 6.9 miles (0.1 degrees in coordinate distance) in our case.[25] The next step establishes 15 bins of distances and computes the average "semivariance," defined as half of the squared difference in land values, of all the points in each distance bin. The third step fits a 3-parameter curve that preserves monotonicity to this set of 15 binned averages. This is referred to as a "variogram;" we estimate one variogram per county per year. The fourth step applies this curve to estimate covariances between values in an unsampled location and a number of nearby sampled locations – we choose 20 nearby neighbors. The fifth step uses these fitted covariances to construct the weights on the nearby sampled locations.[26]

Figure 3 shows a heat map of the level of land values for the city of Washington, DC and provides a look at how Kriging interpolates land value. Panel (a) shows the standardized land values in our working sample and panel (b) shows how the Kriging interpolation algorithm interpolates these values to the entire geography of Washington, DC, including areas with few observations. As can be seen from the legend of panel (b), interpolated standardized land values vary widely, from less than \$150 thousand per quarter-acre to more than \$750 thousand per quarter-acre. Panel (c) shows the average value of the standardized land prices by ZIP code and panel (d) shows the average value for as-is land prices (we describe how we compute as-is land prices later).[27] In both cases, the mean is computed using the estimated land prices for properties in the working sample and the interpolated land prices for all other properties. Interestingly, the as-is land prices in panel (d) are uniformly higher than the standardized prices in panel (c). This occurs because lot sizes in Washington, DC are generally much smaller than a quarter-acre, so standardizing to the much larger

[25] Our data are partitioned by county (and by year in the panel). This implies the Kriging procedure only considers pairwise points where both points are in the same county. The 6.9 mile cutoff will not bind in any county where the maximum distance between two locations in that county is less than 6.9 miles.

[26] Due to the fact that Kriged values are in logs, this step also involves calculating the prediction error variance which we use to convert the log of land prices to levels.

[27] Panel (b) shows that it is possible to estimate a land price for every location in the city but panels (c) and (d) illustrate that we do not report land values for some ZIP codes in Washington, DC. We do not report land prices for a ZIP code or tract if there are fewer than 10 land-price observations in our working sample for that geography.

quarter-acre lot size reduces the value of land per acre.[28] Maps like this help to illustrate the value that households place on location-specific attributes such as school quality, access to transportation and other natural and man-made amenities.

To evaluate the accuracy of the Kriging procedure, we omit a randomly selected 20% of our working sample and determine the root mean-square error (RMSE) of the Kriging interpolation procedure for that omitted 20%. In every year, we compute the RMSE for each county in this hold-out sample. Across counties, the median RMSE is about 40%. This might seem high, but it is in fact well within our priors. To understand why, denote true land, structure and house value for a given parcel as $p_L^* L$, $p_S^* S$ and $p_H^* H$ such that

$$p_L^* L \;=\; p_H^* H \;-\; p_S^* S \tag{8}$$

where $p_L^*$, $p_S^*$ and $p_H^*$ are the price per unit of land, structures and housing, respectively, and $L$, $S$ and $H$ are the quantities. Suppose that the value of structures is not measured with error but housing is measured with multiplicative error $e$ such that observed house value $p_H^o H$ is equal to $p_H^* H\,(1+e)$. We can then write an expression for the percentage deviation of observed land value $p_L^o L = p_H^o H - p_S^* S$ from the truth as

$$\frac{p_L^o L - p_L^* L}{p_L^* L} \;=\; \left(\frac{p_H^* H}{p_L^* L}\right) e \tag{9}$$

Equation (9) says that the percentage measurement error in land values is equal to the inverse of land's share of house value times the percentage measurement error in house values. If the standard deviation of measurement error in house prices is 15%, as suggested by Case and Shiller (1989), and land's share of house value is 40% – approximately our estimate for the aggregate United States over 2012-2019 – then the standard deviation of measurement error in land prices would be 37.5%, which is close to what we observe in the hold-out sample.

We also use our hold-out sample to evaluate the accuracy of the Kriging procedure relative to three other commonly-used spatial interpolation procedures: Null (the average across all properties in a defined geography), Nearest Neighbor (the average across nearby properties in that geography), and Inverse-Distance Weights. In every year, Kriging has the lowest RMSE. Additionally, we check to see that the Kriging procedure can replicate the gradient

[28] Restated, due to plattage effects the value of a single quarter-acre lot is much less than the value of, say, four lots of 1/16 acre each.

of land values that endogenously arise from a simple rendition of the standard monocentric city model that we compute analytically. We simulate two data sets from this model, one in which land values are measured perfectly and another in which land values are measured with error and then check the extent to which Kriging can replicate the analytic gradient of land prices. In the data set in which land values are perfectly measured, Kriging nearly exactly replicates the analytic gradient. In the data set in which land values are measured with error, Kriging produces relatively small average errors. More details are available in the online appendix.

### 4.3. As-Is Estimates of Land Value

In some applications it is inappropriate to control for plattage effects. For example, standard urban models with optimizing households, i.e. Alonso (1964), Mills (1967), Muth (1969), and Brueckner (1987), predict that lot sizes will be larger in areas where land is plentiful and smaller where land is scarce. Such differences in optimal lot sizes contribute to observed differences in actual land prices per-acre across a metro area. Testing the predictions of these models using land values standardized to a quarter-acre lot may be inappropriate. Additionally, many accounting exercises require actual, as opposed to standardized, land values including tabulations used for national accounting or local tax assessments. For these reasons, we compute land prices per acre without adjusting for plattage effects, i.e. our as-is estimates of land value. To do this, we simply undo the correction for plattage effects for all parcels.[29] In our data files, we report as-is estimates of land value on a per-acre basis alongside the average lot size of single-family homes in the assessor dataset. These can be multiplied to yield estimates of the average value of land for single-family residential lots in each area.

### 4.4. Land Shares

While the price of land is useful in itself, the share of house value attributable to land is also of interest to researchers and policy-makers. For example, as has been remarked by Davis and Heathcote (2007), Davis, Oliner, Pinto, and Bokka (2017) and others, rapid changes in the share of housing attributable to the value of land are indicative of positive shocks to the demand for housing.

For each housing unit in an area, we estimate the value of land and the value of housing.

[29] We set the log of the as-is estimate equal to the log of the standardized estimate plus $b_L$ times the difference of the log of the size of the lot (in acres) and the log of a quarter-acre, i.e. referring to equation (7) for each parcel we compute the as-is estimate by adding $b_L \left(X_i^L - \log 0.25\right)$ to the standardized estimate.

The land share we report is the average value of land across housing units divided by the average value of housing across those same housing units. For each single-family housing unit in the assessor data, we use the as-is estimates of the actual value of land. Our procedure to compute the value of housing for all the single-family units in the assessor data is more complicated and the remainder of this section provides details. But the procedure resembles the one used to estimate land values: estimate the standardized home value for homes with admissible appraisals, use Kriging to interpolate values to all homes in the area, then undo the standardization to create as-is values for all homes.

To start, for *all* of the housing units in the appraisal data and not just the relatively new units in our working sample, we regress the log of appraised house value ($Y_i^H$) on the log of effective age ($X_i^A$), the log of lot size ($X_i^L$) and the log of structure size ($X_i^S$) with intercept $\beta_0$ and coefficients $\beta_A$, $\beta_L$ and $\beta_S$ and error term $v_i^H$.[30]

$$Y_i^H \;=\; \beta_0 \;+\; X_i^A \beta_A \;+\; X_i^L \beta_L \;+\; X_i^S \beta_S \;+\; v_i^H \tag{10}$$

We estimate this model for each county. Given the coefficient estimates in equation (10), we compute the expected log value for each house in the appraisal data if the effective age was (counterfactually) 15 years, the lot size was 0.25 acres and the structure size was 2,000 square feet. Call this standardized house value $\widetilde{Y}_i^H$, computed as

$$Y_i^H + \left(\log 15 \text{ years} - X_i^A\right)\beta_A + \left(\log 0.25 \text{ acres} - X_i^L\right)\beta_L + \left(\log 2{,}000 \text{ sq ft} - X_i^S\right)\beta_S$$

We use Kriging to interpolate the standardized log house value $\widetilde{Y}_i^H$ to all parcels in the assessor data.

In the last step, we use information about the lot size and structure size in the assessor data to convert the estimate of standardized log house value to an un-standardized estimate of house value (also referred to as "as-is"), one that correctly takes on-board the age and size of the structure. This step is complicated for two reasons: (1) Effective age is not reported in the assessor data and (2) the lot size and structure size reported in the assessor data does not necessarily equal the lot and structure sizes reported in the appraisal data.

Starting with the second issue, we run regressions of $X_i^L$ and $X_i^S$ in the appraisal data set

[30] We also include ZIP-code and year fixed effects.

on the same variables in the assessor data, call them $\chi_i^L$ and $\chi_i^S$, at the county level

$$\begin{aligned} X_i^L &= a^L + \chi_i^L \gamma^L + u_i^L \\ X_i^S &= a^S + \chi_i^S \gamma^S + u_i^S \end{aligned}$$

where the $u$ terms are the errors in the regression. We then compute for each value in the assessor data the predicted values of lot size and square footage

$$\begin{aligned} \widetilde{\chi}_i^L &= a^L + \chi_i^L \gamma^L \\ \widetilde{\chi}_i^S &= a^S + \chi_i^S \gamma^S \end{aligned}$$

Next, using the appraisal data we run county-specific regressions of effective age on square footage, lot size, and ZIP-code fixed effects of the form

$$X_i^A = \delta_0 + X_i^L \delta_L + X_i^S \delta_S + \nu_i^A \qquad (11)$$

We predict effective age in the assessor data, $\widetilde{\chi}_i^A$, using the coefficients from the regression in equation (11) and the corrected values of lot size and square footage

$$\widetilde{\chi}_i^A = \delta_0 + \widetilde{\chi}_i^L \delta_L + \widetilde{\chi}_i^S \delta_S$$

Finally, using parameters estimated in equation (10), we compute log house value for each parcel in the assessor data that is not in the appraisal data as

$$\widetilde{Y}_i + \left(\widetilde{\chi}_i^A - \log 15 \text{ years}\right)\beta_A + \left(\widetilde{\chi}_i^L - \log 0.25 \text{ acres}\right)\beta_L + \left(\widetilde{\chi}_i^S - \log 2,000 \text{ sq ft}\right)\beta_S$$

In the final step, we convert the log of predicted house values to levels.[31]

# 5 Results

We report results for a given county only if our working sample includes at least 50 observations within that county in the relevant time period. When we report results for ZIP codes and census tracts, we require at least 50 observations within the county and at least 10 observations within the appropriate geography, also in the relevant time period. These minimum required sample sizes generate two data sets. In the first, our "pooled cross-section," the working sample combines all data from 2012 through 2019 in the appraisal database and cen-

[31] As we do with land prices (see footnote 26), we use the prediction-error variance that we compute during the Kriging procedure to convert the log of predicted house prices to levels.

ters the estimates using 2015 prices.[32] Pooling increases the number of geographic areas that satisfy the minimum required sample sizes. Our pooled cross-section includes data for 2,378 counties, 18,322 ZIP codes, and 53,144 census tracts. In the second dataset, we do not pool data by year and simply report annual estimates for a balanced panel. Given our minimum data requirements, we report land prices for 960 counties, 7,742 ZIP codes, and 10,515 census tracts each year from 2012 through 2019. The annual and pooled data sets cover 83% and 98% of the U.S. population residing in the 50 states plus the District of Columbia, respectively, and 85% and 98% of the single-family housing units. The coverage difference between the annual and pooled datasets is fairly small because well-populated areas are included in both. After constructing county-level statistics, CBSA, state, and national statistics are calculated by aggregating over counties for which land values are available, weighted by the single-family housing stock. We report national data and state data for all 50 states plus the District of Columbia in both the pooled and panel datasets. We are also able to calculate values for 915 CBSAs for the pooled sample and 509 in the balanced panel.[33]

Table 1 shows some basic statistics from our as-is estimates of land value and land shares. The top panel reports estimates from the pooled cross-section of counties and the bottom panel reports estimates from the annual panel data of counties (pooled to cover 2012-2019). The statistics reported in Table 1 are generated by applying an equal weight to each county. The data show an enormous range in the average price per acre of land in single-family, residential use from just over $11 thousand per county at the 1st percentile of the pooled cross-section data to nearly $1.45 million at the 99th percentile. The land shares from this data set also show huge variation, from 7.7% at the 1st percentile to 54.3% at the 99th percentile. The panel data display similar variation, but the price of land and land shares are uniformly higher, as this data set drops many rural counties with cheap land and low land shares. Returning to the pooled cross-section data, the average price of land per acre is $152 thousand and the median price per acre is $53 thousand, indicating significant right-skewness. The standard deviation of the average price per acre is $1.26 million, 8.3 times the average value.

In the remainder of this section, we present other stylized facts related to the land-price and land-share data. For expositional purposes, we present pooled estimates first in order

[32] We adjust data to 2015 prices by using the appropriate year dummy variables in equation (6).

[33] See the online appendix for a comparison of our results to those in Davis and Palumbo (2008) for the metropolitan areas covered by their study.

to validate our data in terms of known cross-sectional relationships between land prices and other variables. We then proceed to the annual panel, where we present several new findings. Overall, there are five main categories of stylized facts that we present: 1) land-price gradients and levels; 2) parallel information for land shares; 3) spatial variation in housing-structure density; 4) barriers to building housing and urban decline; and 5) changes over time in land values and land shares. For these stylized facts, we use as-is land prices, as these reflect the variation in lot sizes due, at least in part, to optimizing decisions by builders and households.

## 5.1. Land Prices: Gradients and Levels

The traditional monocentric city model predicts land prices fall with distance to the Central Business District (CBD) because households are willing to pay less per unit of housing as commuting costs rise. Since the marginal cost of an additional unit of structures is roughly constant within the city, the solution to the zero-profit condition for housing producers requires variation in the price of land. Therefore, the negative house price gradient translates to a negative gradient for land prices.

Figure 4 illustrates the relation between land prices and proximity to the CBD for all metro areas. Panel (a) shows the relationship for all ZIP codes with centroids within 25 miles of a central ZIP-code centroid after separating metropolitan areas into three size groups: Those with more than 2 million housing units (solid blue line), those with 500 thousand to 2 million units (dashed red line) and those with less than 500 thousand units (dotted orange line). The three groupings each show a downward sloping relationship, consistent with the monocentric city model, but the level and rate of change of land prices with respect to distance from the CBD depends on the size of the metro area. This panel shows that in the largest metro areas, land prices are about 3 times higher than in the middle grouping and about 25 times higher than in the smallest metro areas. Panel (b) shows how land prices changed over 2012-2019 for each of the three groups as a function of distance to the CBD. This panel shows that land prices of the closest-in locations increased more rapidly than locations farther out for the largest metro areas, thus steepening the land-price gradient. In contrast, land prices increased by about the same rate at all distances from the CBD for the medium-sized and smallest metro areas, indicating little or no change to the slopes of the land-price gradients over 2012-2019.

An open question is the extent to which land supply restrictions, as typically measured by researchers, affect the level and patterns of land pricing in a metro area. Supply restrictions may lower some land values, as these restrictions tend to reduce or eliminate the types of structures that can be built on the land. However, as discussed by Saks (2008), Saiz (2010) and Davidoff (2016), whether land-supply restrictions boost or reduce land prices is an empirical question. Panel (c) of Figure 4 shows land-price gradients by metro area size group and the top and bottom halves of regulatory burden as measured by the Wharton Residential Land Use Regulation Index from Gyourko, Saiz, and Summers (2008); and panel (d) shows the land price gradients by topographic interruptions as measured by Saiz (2010). These panels show supply restrictions are positively correlated with the levels of land prices but have little correlation with the slopes of the gradients.

### 5.2. The Land Share Gradient

Figure 5 highlights some of our results using data on land shares across ZIP codes. The structure of Figure 5 is the same as Figure 4. Comparing the results in the two figures, we draw three conclusions. First, similar to land values, land shares vary widely across geography. Second, land shares also fall with distance to the CBD, though the rate of decline is generally modest. Some decline is to be expected as our model predicts that land's share of home value rises with age until the option to rebuild is exercised, and in most cities older homes are located closer to the CBD than newer homes.[34] Third, greater regulation and topographical interruption are associated with higher land shares. Also note that changes to land shares over time with respect to distance to the CBD, shown in panel (b), closely reflect changes to land prices discussed in panel (b) of Figure 4. Locations close to the CBD in the most populated metro areas experienced the greatest increases in land shares, consistent with the fast growth in land prices in these locations; land shares increased relatively uniformly at all distances to the CBD in the middle-grouping of metros; and, land shares decreased relatively uniformly at all distances to the CBD for the smallest metro areas, a consequence of relatively slow growth of land prices everywhere in those areas.

### 5.3. Land Prices, Structure Density, and the Reservation Use

Economists typically model the production of housing as a function of structures and land inputs. Earlier, we showed that the price of land tends to rise as distance to the CBD

[34] The mix of older and newer homes helps explain the steep decline in the land share for very large cities with low regulation, shown in panel (c) of Figure 5. The only metro areas in this group are Dallas and Houston. In both areas, the number of newer homes rises sharply beyond 10 miles from the CBD, which boosts the structure share and reduces the land share. The other groups of metro areas in panel (c) have much more limited increases in newer homes with distance from the CBD.

declines. As the price of land rises, with no change in structure cost, the market equilibrium should feature less use of land per housing unit. We can test this prediction directly. The assessor data contain information on both interior square feet and the lot size. We use these two variables to construct the ratio of the interior square footage to the lot size, known as the "floor-area ratio" (FAR), for single-family homes in all areas where we estimate land value.

Panel (a) of Figure 6 graphs FARs for single-family housing against land prices for the counties in our data. At a low land price per acre, structure density is low, with the cheapest land containing FARs at about 0.05 (e.g., a house with about 2,200 interior square feet on a one-acre lot) while the most expensive land contains FARs above 1 (a 3 story, 2,400-square-foot row house on a 2,400-square-foot lot has an FAR of 1). Panel (b) demonstrates that on average, lot sizes decrease with land prices. Both panels show that builders and households economize on land when its price is high.

Panel (c) shows the relation of the average price of agricultural land by county, as measured by the U.S. Department of Agriculture, for counties with some land in agricultural use, and the average value of land in residential use in those counties. Models of urban economics predict that these values should be linked, as land at the edge of urban areas can be used for either agricultural or residential use. Not surprisingly, panel (c) shows a strong positive relationship.

### 5.4. Barriers to Building Housing and Urban Decline

Figure 7 shows the correlation across CBSAs of as-is land prices per acre, left column, and land shares of house value, right column, against covariates of housing-supply elasticities. Each dot in every graph represents a CBSA from the pooled cross-section data. The top and middle panels display, respectively, the correlation with the Wharton regulatory index of Gyourko, Saiz, and Summers (2008) and with topographic interruptions as measured by Saiz (2010).[35] These panels show that, on average, the price of land and the share of housing value attributable to land increase across CBSAs with both regulatory burden and topographic difficulty in building housing. The bottom panel shows the correlation of land prices and land shares with the fraction of the housing stock with a value less than replacement cost in

[35] Amior and Halket (2014) first compared measures of regulation to estimates of land shares from Davis and Palumbo (2008). In their Figure 7, they show that land scarcity and the land share of house value are positively correlated at the MSA level.

1990 ("Urban Decline"), as first documented by Glaeser and Gyourko (2005). The bottom panel shows that metropolitan areas that are in relative decline, as evidenced by a larger fraction of the housing stock that is below replacement cost, also have relatively low land prices and low land shares.

### 5.5. Changes in Land Shares

In this section, we further investigate changes in land shares between 2012 and 2019, a topic we touched on when discussing panel (b) of Figure 5. Panel (a) of Figure 8 plots a histogram of all the ZIP-level changes and panel (b) plots changes in the land share as a function of the land share in 2012. Both panels show that land shares changed little, on average, during the housing recovery but that there was wide variation in the change across ZIP codes. Panel (c) shows changes as a function of the initial land share in 2012, for metropolitan areas with more than 2 million housing units (solid blue line), with 500 thousand to 2 million housing units (dashed red line), and with less than 500 thousand housing units (dot-dash orange line). This panel shows that the share of house value attributable to land tended to increase in metro areas with at least 500 thousand units and declined in smaller metro areas, as the solid blue and dashed red lines are everywhere above 0 and the dot-dash orange line is everywhere below it. In the smallest metro areas, the largest declines occurred in ZIP codes with high initial land shares.

Although not apparent from the results in Figures 5 or 8, we estimate that the land share for the United States in the aggregate increased from 38.2% in 2012 to 40.9% in 2019. This nearly 3 percentage point increase was driven by a rapid increase in land prices at the national level of 6.5% per year that outstripped the national rise in house prices. Figure 9 reconciles the relatively balanced changes in land shares shown in Figure 8 with the substantial increase at the national level. In both the top and bottom panels of Figure 9, we sort counties into 25 bins based on the number of single-family housing units over 2013-2017. We use counties rather than ZIPs because counties are the most granular units that we aggregate to higher-level geographies. Each bin represents the experience of about 90 counties. The blue triangles in each bin show the aggregate value of land in single-family, residential use in 2012 for all the counties included in each bin. These triangles represent the weight for each bin in the national aggregate.

The red bars in the top panel of Figure 9 show the change in as-is land prices in each bin from 2012 to 2019 and the red bars in the bottom panel show the change in the land share over the same period. When taken together, the two panels tell an interesting story. The top panel shows that the largest counties experienced a much faster rise in land prices than did smaller counties, and importantly, that the largest counties heavily influence what happens to land prices for the U.S. as a whole. The right-most bin alone, which is dominated by counties in California, has greater aggregate land value than the bottom 23 bins combined, and more than four times the land value of the bottom 18 bins combined. This explains why the increase in land prices at the national level over 2012-2019 – 6.47% per year on average, as shown by the black dashed line – outpaced the much slower rise in most of the county bins. For these smaller counties, the bottom panel shows that the relatively modest growth in land prices translated to a decline in the land share of home value over 2012-2019. In contrast, five of the top seven bins experienced an increase in land shares, including very large increases in the highest and third-highest bins. Thus, even though land's share of home value was flat or falling over 2012-2019 in most counties in the United States, the large rise in the land share in the most populous counties drove up the national share.

# 6 Conclusion

Although it is widely recognized that booms and busts in house prices largely reflect booms and busts in underlying land prices, until recently limited data was available to study land prices. We help fill this gap by using a very large data set of appraisals to generate annual panel data from 2012 through 2019 of the average price of land used in single-family homes for 960 counties, 7,742 ZIP codes, and 10,515 census tracts. We also calculate pooled cross-sectional estimates of land prices for more than twice as many counties and ZIP codes and five times as many census tracts. Overall, we document a number of properties of the level and growth rate of land prices that are generally consistent with predictions of traditional models of urban economics. We expect that future researchers will use the data we generate to build on our results, and current and future policy-makers will monitor these data to better understand emerging risks in housing markets.

# References

Albouy, D., G. Ehrlich, and M. Shin (2018): "Metropolitan Land Values," *Review of Economics and Statistics*, 100(3), 101–120.

Alonso, W. (1964): *Location and land use: Toward a general theory of land rent*. Harvard University Press.

Amior, M., and J. Halket (2014): "Do households use home-ownership to insure themselves? Evidence across U.S. cities," *Quantitative Economics*, 5(3), 631–674.

Basu, S., and T. G. Thibodeau (1998): "Analysis of spatial autocorrelation in house prices," *Journal of Real Estate Finance and Economics*, 17(1), 61–85.

Bostic, R. W., S. D. Longhofer, and C. L. Redfearn (2007): "Land leverage: Decomposing home price dynamics," *Real Estate Economics*, 35(2), 183–208.

Brueckner, J. K. (1987): "The structure of urban equilibria: A unified treatment of the Muth - Mills model," *Handbook of Regional and Urban Economics*, 2, 821–845.

Case, K. E., and R. J. Shiller (1989): "The Efficiency of the Market for Single-Family Homes," *American Economic Review*, 79(1), 125–137.

Clapp, J. M., P. Eichholtz, and T. Lindenthal (2013): "Real option value over a housing market cycle," *Regional Science and Urban Economics*, 43(6), 862–874.

Davidoff, T. (2016): "Supply constraints are not valid instrumental variables for home prices because they are correlated with many demand factors," *Critical Finance Review*, 5(2), 177–206.

Davis, M. A., and J. Heathcote (2007): "The price and quantity of residential land in the United States," *Journal of Monetary Economics*, 54(8), 2595–2620.

Davis, M. A., W. D. Larson, S. D. Oliner, and J. Shui (2020): "Online Appendix for The Price of Residential Land for Counties, ZIP Codes, and Census Tracts in the United States," Working Paper.

Davis, M. A., S. D. Oliner, E. J. Pinto, and S. Bokka (2017): "Residential land values in the Washington, DC metro area: New insights from big data," *Regional Science and Urban Economics*, 66, 224–246.

Davis, M. A., and F. Ortalo-Magné (2011): "Household Expenditures, Wages, Rents," *Review of Economic Dynamics*, 14(2), 248–261.

Davis, M. A., and M. G. Palumbo (2008): "The price of residential land in large U.S. cities," *Journal of Urban Economics*, 63(1), 352–384.

Diewert, W. E., J. d. Haan, and R. Hendriks (2015): "Hedonic regressions and the decomposition of a house price index into land and structure components," *Econometric Reviews*, 34(1-2), 106–126.

Dzierbicki, D. J. (2014): "Developing Effective Age Estimates," http://www.appraisers.org/docs/default-source/discipline_rp/dzierbicki-development-of-effective-age-estimates.pdf?sfvrsn=0, American Society of Appraisers.

Glaeser, E. L., and J. Gyourko (2005): "Urban decline and durable housing," *Journal of Political Economy*, 113(2), 345–375.

Gyourko, J., A. Saiz, and A. Summers (2008): "A new measure of the local regulatory environment for housing markets: The Wharton Residential Land Use Regulatory Index," *Urban Studies*, 45(3), 693–729.

Harding, J. P., S. S. Rosenthal, and C. Sirmans (2007): "Depreciation of housing capital, maintenance, and house price inflation: Estimates from a repeat sales model," *Journal of Urban Economics*, 61(2), 193–217.

Haughwout, A., J. Orr, and D. Bedoll (2008): "The price of land in the New York metropolitan area," *Federal Reserve Bank of New York: Current Issues in Economics and Finance*, 14(3).

Hengl, T. (2007): *A Practical Guide to Geostatistical Mapping of Environmental Variables.* Luxembourg: Office for Official Publications of the European Communities.

Hill, R. J., and D. Melser (2008): "Hedonic imputation and the price index problem: an application to housing," *Economic Inquiry*, 46(4), 593–609.

Katz, A. J., and S. W. Herman (1997): "Improved estimates of fixed reproducible tangible wealth," *Survey of Current Business*, 77(5), 69–92.

Lutz, B., R. Molloy, and H. Shan (2011): "The housing crisis and state and local government tax revenue: Five channels," *Regional Science and Urban Economics*, 41(4), 306–319.

McMillen, D., and A. O'Sullivan (2013): "Option value and the price of teardown properties," *Journal of Urban Economics*, 74, 71–82.

Mills, E. S. (1967): "An Aggregative Model of Resource Allocation in a Metropolitan Area," *American Economic Review*, 57(2), 197–210.

Muth, R. F. (1969): *Cities and housing; the spatial pattern of urban residential land use.* University of Chicago Press.

Nichols, J. B., S. D. Oliner, and M. R. Mulhall (2013): "Swings in commercial and residential land prices in the United States," *Journal of Urban Economics*, 73(1), 57–76.

Saiz, A. (2010): "The Geographic Determinants of Housing Supply," *The Quarterly Journal of Economics*, 125(3), 1253–1296.

Saks, R. E. (2008): "Job Creation and Housing Construction: Constraints on Metropolitan Area Employment Growth," *Journal of Urban Economics*, 64(1), 178–195.

Sherman, M. (2011): *Spatial statistics and spatio-temporal data: covariance functions and directional properties*. John Wiley & Sons.

Siebers, J. (2016): "Effective Age Versus Actual Age - An Important Appraisal Question," https://www.corelogic.com/blog/2016/11/effective-age-versus-actual-age.aspx, CoreLogic Insights Blog.

Titman, S. (1985): "Urban land prices under uncertainty," *American Economic Review*, 75(3), 505–514.

# A Spatial Interpolation Methods

This section describes various interpolation methods we compare in the paper. Our land price estimation approach begins by dividing the universe of parcels $N$ within a geography into those with values that are observed (i.e. included in our working data) and those with values that are not observed. For convenience, below we denote the number of observed values in a geography as $N^s$. All land prices we report are calculated as the simple average of the price estimate (or actual value, when available) of each individual parcel within the county or county subaggregate (i.e. census tract). Accordingly, differences between land values within a given geography arise due the method used to estimate prices for parcels with unobserved values.

At the end of the day, we estimate the value of land when land value is not directly observed as a weighted average of land values of a subset of parcels within the county where land prices are directly observed. The estimated price for parcel $i$ is calculated as the average of the $n$ nearest (by proximity) neighbor parcels, indexed by $j$, with weights $\lambda_{i,j}$.[36] For a particular method, the estimated price for parcel $i$ is:

$$\hat{p}_i = \sum_{j=1}^{n} \lambda_{i,j} p_{i,j} \tag{12}$$

The weights are assumed to sum to unity, or $\sum \lambda_{i,j} = 1$.

## 1.1. Spatial Statistics

Spatial statistical methods do not consider spatial relations in outcomes, only proximity. Here, we discuss three spatial statistics that are commonly used to interpolate prices spatially. The general form of each of these statistics is below, where $h$ is the distance between the location to be imputed, $i$, and another nearby sampled location, $j$, that is one of the $n$ nearest locations. The exponent $c$ gives the degree of decay in the weight that is due to distance between the parcels.

[36] So $p_{i,1}$ is the closest observed price to location $i$; $p_{i,2}$ is the second closest observed price to location $i$; and so on.

$$\lambda_{i,j} = \frac{1}{h_{i,j}^c} \left( \sum_{j=1}^{n} \frac{1}{h_{i,j}^c} \right)^{-1} \tag{13}$$

**Null Estimator**

The null estimator ("Null") sets $n$=$N^s$ and $c = 0$, giving $\lambda_{i,j} = 1/n$. This gives the estimate of an individual parcel as the sample average.

**Nearest Neighbor**

The nearest neighbor estimator ("NN") also sets $c = 0$, giving $\lambda_{i,j} = 1/n$. But $n$ is typically set to the 5 to 25 nearest observed prices. This gives the estimate of a parcel as the sample average of nearby parcels.

**Inverse-distance weights**

The inverse-distance weight estimator ("IDW") sets $n$ generally within the range of 5 to 25 nearest observed prices as with the NN estimator. The calculation of $\lambda$ then involves assuming an exponent $c$. This exponent is commonly set equal to $c = 2$, giving a relation between points that declines with the square of the distance. In this case, $\lambda_{i,j} = \frac{1}{h_{i,j}^2} \left( \sum_{j=1}^{n} \frac{1}{h_{i,j}^2} \right)^{-1}$

## 1.2. Geostatistics

In addition to spatial statistics, we include a single geostatistical estimator: ordinary Kriging. Ordinary Kriging is our preferred Kriging estimator because we eliminate known variation across parcels by way of standardization. In effect, we are performing a two-step regression-Kriging estimator, which can also be interpreted as estimating a rudimentary hedonic and then Kriging the residual.

As with the nearest neighbor estimator, $n$ nearest neighbors are weighted and summed to generate predicted prices. The key difference with Kriging is $\lambda$ is estimated based on the strength of the observed relationships between observations of different proximities within the sample such that it is the best linear unbiased predictor.

Calculation of the weights proceeds in four steps. The first step involves calculating pairwise differences in values between each pair in the sample within a certain distance range. The next step collapses and bins the semivariances (half of the squared differences) into averages by the distance between the points. The third step fits a curve, referred to as a "variogram," to this set of binned averages. The fourth step uses this curve to estimate semivariances between any two points and then construct the weights.[37]. Once the weights are known, predictions and prediction error variances can be readily calculated for any location.

As an illustration of this procedure, we present kriging steps for land prices in Washington, DC, pooled between 2012 and 2019. There are about 110,000 parcels and 16,000 sampled standardized land prices. To start, differences and semivariances $\gamma$ are calculated for each pair of points in $N$.[38]

The results from the first two kriging steps are shown in Figure 10. The hollow circles represent semivariance averages within each of the 15 distance bins. Distances are reported as miles but are in terms of latitude/longitude degrees in the programs.[39]

In step three, a functional form for the relationship between the semivariances and the distances (the hollow circles) is assumed and fit to the data. The fitted curve, as shown by the blue line, is typically upward sloping, indicating the greater the distance the higher the variance. The spherical functional form has three parameters, $a_0$, $a_1$, and $r$ that we estimate

$$\gamma(h; a_0, a_1, r) = \begin{cases} a_0 + a_1 \left( \frac{3}{2}\left(\frac{h}{r}\right) - \frac{1}{2}\left(\frac{h}{r}\right)^3 \right), & 0 < h < r \\ a_0 + a_1, & h \geq r \end{cases} \tag{14}$$

The three parameters combine to give the "sill" which is the value to which the variogram asymptotically approaches as the distance between points approaches infinity, or $a_0 + a_1$; the "nugget" which is the value of the variogram when distance approaches zero, or $a_0$, and the "range," $r$, which is the value of $h$ when the variogram reaches the sill. Figure 10 shows a fitted spherical functional form for Washington, DC, with $\hat{a}_0 = 0.06$, $\hat{a}_1 = 0.55$ and $\hat{r} = 9.18$ miles. This function is used to estimate the semivariance between any two points given a

[37] For a more in-depth overview of kriging, see Hengl (2007) or Sherman (2011).

[38] The semivariance for prices at two points $i$ and $j$ is half of the squared difference, $\frac{1}{2}(p_i - p_j)^2$. Isotropy (i.e. the direction between the points does not affect the strength of the relationship) is a standard assumption, which we make here in order to express proximity using a single variable, $h$.

[39] In other words, we take the square root of the sum of the squared differences between two sets of coordinates. Since our distances are generally small, we use this simplified distance measure as a proxy for the actual distance, which varies due to changes in latitude.

distance between them.[40]

In the fourth step, define a linear system that gives the best linear predictor of the weights $\lambda$ for an unsampled location $i$. $\gamma$ is an $n \times 1$ vector of estimated semivariances between $i$ and its $n$ nearest points, indexed by $j$ (Sherman, 2011). $\mathbf{\Gamma}$ is an $n \times n$ matrix of semivariances between these $n$ nearest points. These matrices are augmented in the standard fashion with a Lagrange multiplier and column/row vectors of ones and a zero to normalize the weights to sum to one. These give the vector of weights $\lambda_i$, an $n \times 1$ vector.

$$\begin{bmatrix} \lambda_{\mathbf{i}} \\ \mathcal{L} \end{bmatrix} = \begin{bmatrix} \mathbf{\Gamma}_{\mathbf{i}} & \mathbf{1}_n \\ \mathbf{1'}_n & 0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma_{\mathbf{i}} \\ 1 \end{bmatrix} \tag{15}$$

Then, the predicted value for location $i$ is $\hat{p}_i = \lambda_{\mathbf{i}}' p_{i,j}$ and the prediction error variance is $\hat{V}_{p_i} = \lambda_{\mathbf{i}}' \gamma_i$. The prediction error variance is relevant because it is used in the Jensen's inequality correction for log prices, $\hat{P}_i = \exp(\hat{p}_i + \frac{1}{2}\hat{V}_{p_i})$.

## 1.3. Comparison

We compare the fit of each spatial interpolation method by comparing actual (log) standardized land prices for a 20% hold-out sample to predicted (log) land prices estimated using an 80% training sample. We consider a number of values for the number of nearest neighbors and the overall distance boundary considered.

Table A.1 reports, by year, median RMSEs across counties for Kriging and the other spatial interpolation methods we consider for the hold-out sample. The table shows that kriging provides the greatest interpolation accuracy in every sample year. Table A.2 provides a bit more color for these results in terms of the mean, median, and standard deviation of RMSEs across the 2,378 counties in the pooled sample. The table shows that each of the kriging estimates at the various parameterizations shown yields similar average fit, with means RMSEs ranging from 0.393 to 0.395. We interpret this finding as indicating that nether the boundary nor the number of nearest neighbors considered tends to affect the estimates at the ranges considered. Overall, we interpret these findings as lending support to our decision to use the 20 nearest neighbors and a 6.9 mile boundary in our county-specific

[40] Other functional forms are common, especially the exponential function, $\gamma(h; a_0, a_1, r) = a_0 + a_1(1 - \exp(-\frac{h}{r}))$. We consider the exponential function as well but the spherical functional form has a slightly better fit, on average.

kriging procedure.

We also evaluate the accuracy of the Kriging procedure for our application by seeing if it can recover land values that endogenously arise from a simple rendition of the standard monocentric city model that we can compute analytically. We simulate two data sets from this model, one in which land values are measured perfectly and one in which land values are measured with error and then check the extent to which Kriging can replicate the analytic gradient of land prices. In the data set in which land values are perfectly measured, Kriging nearly exactly replicates the analytic gradient. In the data set in which land values are measured with error, Kriging produces relatively small average errors. This is described in the next section.

# B Monte Carlo Simulation of Standard Urban Model

Assume that a city lies on a featureless plane with a region called the central business district (CBD) at its center. This district provides all employment and because commuting is costly in a way we specify precisely, households wish to live near the CBD. Spatial equilibrium requires identical households to have identical utility in all locations in the city. As we show, this implies that households consume less housing at a higher price near the CBD. Additionally, the housing production function implies housing is produced with high density and structure intensity near the CBD where land prices are high, and with low density near the edge of the city.

To be precise, assume a person consuming $c$ units of consumption and $h$ units of housing receives utility of

$$(1-\alpha)\ln c + \alpha \ln h \tag{16}$$

If a person lives distance $d$ from the city center, their wage after commuting is $w\,(1-td)$ where $t$ is the percentage of income that must be paid to commute for each unit of distance $d$. Denote the rental cost per unit of housing $d$ units from the city center as $q_d^h$. A person living $d$ units from the city center faces the budget constraint of:

$$w\,(1-td) \quad = \quad c + q_d^h h \tag{17}$$

A person choosing to live in location $d$ units away from the CBD maximizes utility (16) subject to the budget constraint (17) by choosing optimal consumption $c_d$ and housing $h_d$ of

$$\begin{aligned} c_d &= (1-\alpha)\, w\,(1-td) && (18) \\ q_d^h h_d &= \alpha w\,(1-td) && (19) \end{aligned}$$

This means maximized utility at distance $d$ from the center can be written as $\mathcal{U}_d$

$$\begin{aligned} \mathcal{U}_d &= (1-\alpha)\ln\left[(1-\alpha)\, w\,(1-td)\right] + \alpha \ln\left[\alpha w\,(1-td)\,/q_d^h\right] \\ &= \kappa_u + \ln w + \ln(1-td) - \alpha \ln q_d^h && (20) \end{aligned}$$

where $\kappa_u$ is a constant equal to $\alpha \ln \alpha + (1-\alpha)\ln(1-\alpha)$. In equilibrium, we assume all locations have to provide the same utility, for example location $d$ and $d'$ must satisfy

$$\mathcal{U}_d = \mathcal{U}_{d'}$$

Then from equation (20) this implies

$$\begin{aligned} \ln(1-td) - \alpha \ln q_d &= \ln(1-td') - \alpha \ln q_{d'}^h \\ \frac{q_{d'}^h}{q_d^h} &= \left(\frac{1-td'}{1-td}\right)^{\frac{1}{\alpha}} && (21) \end{aligned}$$

Equation (21) governs the rate at which housing rental prices per unit change with distance from the CBD, roughly $t/\alpha$ percent per unit of $d$.

Note that we can also work out how the quantity of housing changes as a function of distance to the CBD. We start by using the definition of utility and substituting in optimal consumption but keeping housing

$$\mathcal{U}_d = (1-\alpha)\ln\left[(1-\alpha)\, w\,(1-td)\right] + \alpha \ln h_d \qquad (22)$$

Once we impose $\mathcal{U}_d = \mathcal{U}_{d'}$, this gives us

$$\begin{aligned} (1-\alpha)\ln(1-td) + \alpha \ln h_d &= (1-\alpha)\ln(1-td') + \alpha \ln h_{d'} \\ \rightarrow \frac{h_{d'}}{h_d} &= \left(\frac{1-td'}{1-td}\right)^{-\frac{1-\alpha}{\alpha}} && (23) \end{aligned}$$

Now that we have worked out how housing quantities $h$ and prices per unit $q_d^h$ vary from the city center, we can also work out how the quantities and prices of land and structures change with distance. Temporarily suppressing the distance subscripts, assume competitive builders build housing using land $l$ and structures $s$ according to a CES production function

$$h = \left[(1-\theta)\, s^{\rho} + \theta l^{\rho}\right]^{\frac{1}{\rho}} \tag{24}$$

with $\rho \in (-\infty, 1]$. Assume each unit of housing generates revenue of $q^h$; further, assume each unit of land costs $q^l$ and each unit of structure costs 1. Builders maximize

$$q^h \left[(1-\theta)\, s^{\rho} + \theta l^{\rho}\right]^{\frac{1}{\rho}} - s - q^l l \tag{25}$$

The first-order conditions for optimal structures are

$$1 = q^h h^{1-p} (1-\theta)\, s^{p-1} \tag{26}$$

$$\rightarrow s = \left[q^h (1-\theta)\right]^{\frac{1}{1-\rho}} h \tag{27}$$

This implies that once we know $q^h$ and $h$, we also know $s$. Note that because we know $s$, we also know $q^l l = q^h h - s$. Now consider the first-order condition for optimal land:

$$q^l l = q^h h^{1-p} \theta l^{p} \tag{28}$$

and thus

$$l = \left[\frac{q^l l}{q^h h^{1-\rho} \theta}\right]^{\frac{1}{\rho}} \tag{29}$$

Given a set of parameters, we can compute how quantities and prices and expenditures on housing, structures and land change with distance from a CBD. For a rough calibration, we set $\alpha = 0.25$ based on the median housing budget shares of renters as documented by Davis and Ortalo-Magné (2011). The other parameters we set to match some approximate features of a city. We set $t = 0.02$ such that people 10 miles from the CBD consume about double the housing than people at the CBD but spend 20% less.[41] We jointly set $\theta = 0.90$ and $\rho = -2.0$ such that land's share of value rises from about 15% 10 miles from the CBD to about 55%

[41] This is referring to single-family homes.

at the CBD. We normalize the price per unit of housing to 1 at the CBD and normalize the quantity of housing consumed at the CBD such that the total expenditure at the CBD is for a \$1 million house. As noted earlier, we assume the price per unit of housing structure is 1.0 everywhere in the metro area. Table A.3 shows prices, quantities and expenditures on housing, structures and land as well as land's share of house value, the quantity of land once we normalize the size of a single-family plot at the CBD to 0.25 acres, and land price per acre.

We simulate two data sets based on the calibration of this model. In both data sets, we draw 100 observations for houses in neighborhoods uniformly between 0 and 3.5 miles from the CBD; 200 observations for houses in neighborhoods uniformly between 3.5 and 7.5 miles from the CBD; and 300 observations for houses in neighborhoods uniformly between 7.5 miles and 10 miles from the CBD. In the first data set we assume no quantities or prices are measured with error. This enables us to see the accuracy of the Kriging procedure with regards to this application in an ideal environment.

In the second data set, we allow for i.i.d. measurement error in both the value of housing and the value of structures.[42] This simulation gives us some intuition for how the Kriging procedure performs under conditions where land is imperfectly measured. Denote $\widetilde{q_d^h h_d}$ as observed housing value and $\widetilde{s}$ as observed structures costs. $\widetilde{q_d^h h_d}$ and $\widetilde{s}$ are determined as

$$\begin{aligned} \widetilde{q_d^h h_d} &= q_d^h h_d \left(1 + e_d^h\right) & e_d^h &\sim U[-0.10, 0.10] \\ \widetilde{s} &= s\left(1 + e_d^s\right) & e_d^s &\sim U[-0.10, 0.10] \end{aligned}$$

with $e_d^h$ and $e_d^s$ drawn independently. We then compute observed land value residually,

$$\begin{aligned} \widetilde{q_d^l l_d} &= \widetilde{q_d^h h_d} - \widetilde{s} \\ &= q_d^h h_d - s \quad + \quad \left(q_d^h h_d e_d^h - s e_d^s\right) \end{aligned}$$

Denote the term in parentheses as $e_d^l$, measurement error in land value. Even though the standard deviation of $e_d^h$ and $e_d^s$ are relatively small (5.7 percent each), the standard deviation of measurement error as a percent of true land value, measured as $e_d^l / \left(q_d^l l_d\right)$, in this second data set is much larger, 27 percent. The measurement error for land value is magnified because land value is residually measured and accounts for a relatively small fraction of

[42]This measurement error can be thought of as a deviation from model-determined prices.

home value (as we discuss earlier in the paper).

Table A.4 compares the estimates from Kriging for land price per acre by distance to CBD to the numbers we compute analytically in Table A.3 for the simulated data set measured without error (data set 1) and the simulated data set with measurement error (data set 2) for 0 to 9 miles to the CBD. When the data are measured with error, the Kriging results are less accurate, though the average error across the distance bins is relatively small at 4.2%.

# C Comparison of CBSA Results to Davis and Palumbo (2008)

Davis and Palumbo (2008) created the first dataset of publicly available land prices and land shares for residential land in a large number of large metropolitan areas. The data for the published article span the time period 1984-2004, but the authors have routinely updated the data and the publicly-available data are now available through 2018.[43] In this Appendix we compare land prices and land shares for the CBSAs and time period in which our data overlap with Davis and Palumbo's data. To facilitate this comparison, in the Davis and Palumbo data we merge Anaheim with Los Angeles, Oakland with San Francisco, and Dallas with Fort Worth to reduce the number of metro areas from 46 to 43 MSAs. Our data and the Davis and Palumbo data overlap for 2012-2018, so we compute average land prices and land shares over this period in each data set and compare the averages.

Broadly speaking, the data align very well. The correlation is 0.95 for the level of land prices and 0.87 for the land shares. However, variation in land-price levels and land shares at the metro area level is less extreme in our data than in the Davis and Palumbo data. To see this, in Figure 11 we show scatter plots of log land prices (top panel) and land shares (bottom panel) along with regression lines. The graph comparing log land prices shows that, in the cross section of metro areas, our log land price increases by 0.73 percent for each one percent increase in the Davis and Palumbo data. Similarly, the graph comparing land shares shows that in the cross section of metro areas, our estimates of land share increases by 0.55 percentage point for each percentage point increase in the Davis and Palumbo data.

Ultimately, we believe the data in our paper are more accurate for three reasons: First, our sample sizes are orders of magnitude larger. For example, the maximum metro-area sample

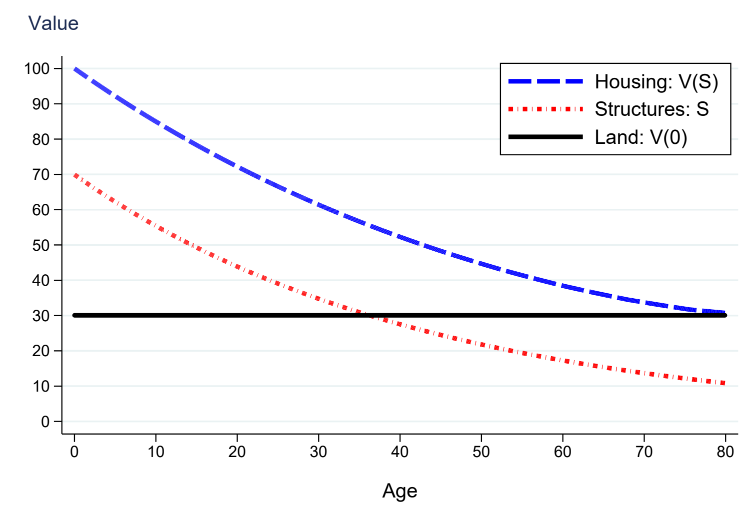
[43]These data are currently available at *https://www.aei.org/historical-land-price-indicators/*.

size in the Davis and Palumbo data is 2,513 in Salt Lake City (Table 2 of that paper). Second, Davis and Palumbo include all observations, including old homes. We drop old homes from the sample to avoid the problem of using construction cost as a proxy for structure value for these homes. Finally, the geography of the specific counties that constitute a given metro area is unclear in Davis and Palumbo. They use whatever data appear in the Metropolitan-AHS survey files, and the documentation of those files is not clear. In our paper, we know exactly the counties comprising each metro area.[44]

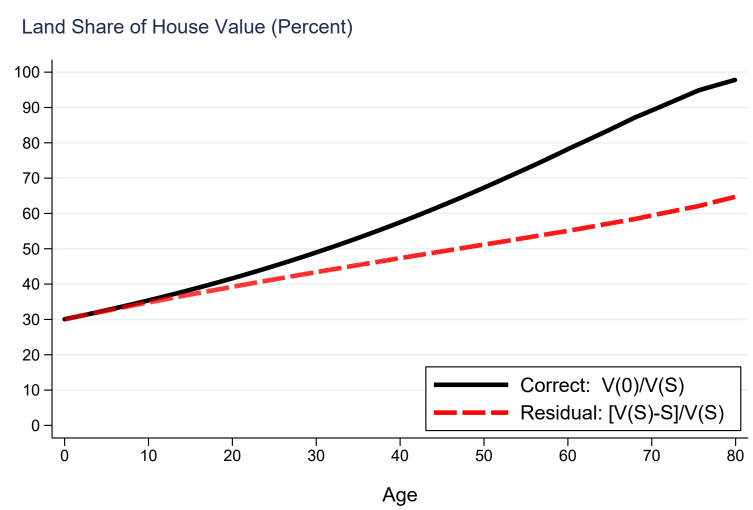
[44]For these reasons, we suspect the statistically significant intercepts and slope coefficients less than 1.0 when we regress our data on the Davis and Palumbo data arise from attenuation bias due to measurement error in their data.

Figure 1: Predictions of Teardown Model of Housing

(a) House Value, Structure Cost, Vacant Land by Age



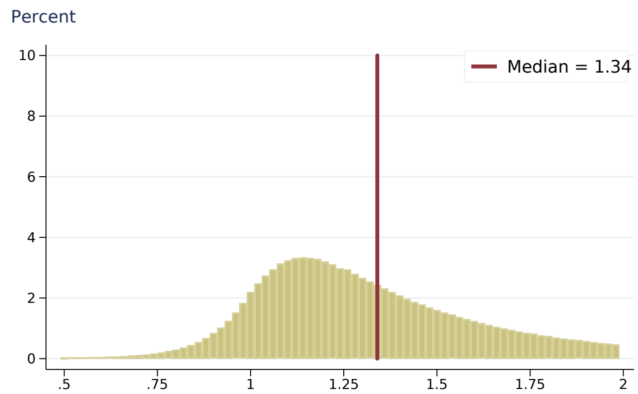

(b) Correctly- and Residually-Measured Land Share of House Value



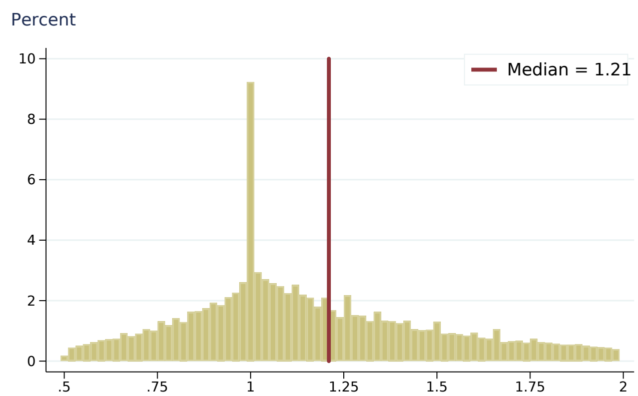

Notes: The top panel shows the model-predicted price of housing (blue dash), structures (red dot-dash) and land (black line) as a function of age. The bottom panel shows land's share of home value measured correctly (black line) and measured residually (red dash). The correct value of land's share is the value of land divided by the value of the house. The residual estimate of land's share is equal to the value of the house less the replacement cost of structure, all divided by the value of the house.

Figure 2: Appraisal Anchoring to Assessed Land and Property Values

(a) Property Value Ratios




(b) Land Value Ratios

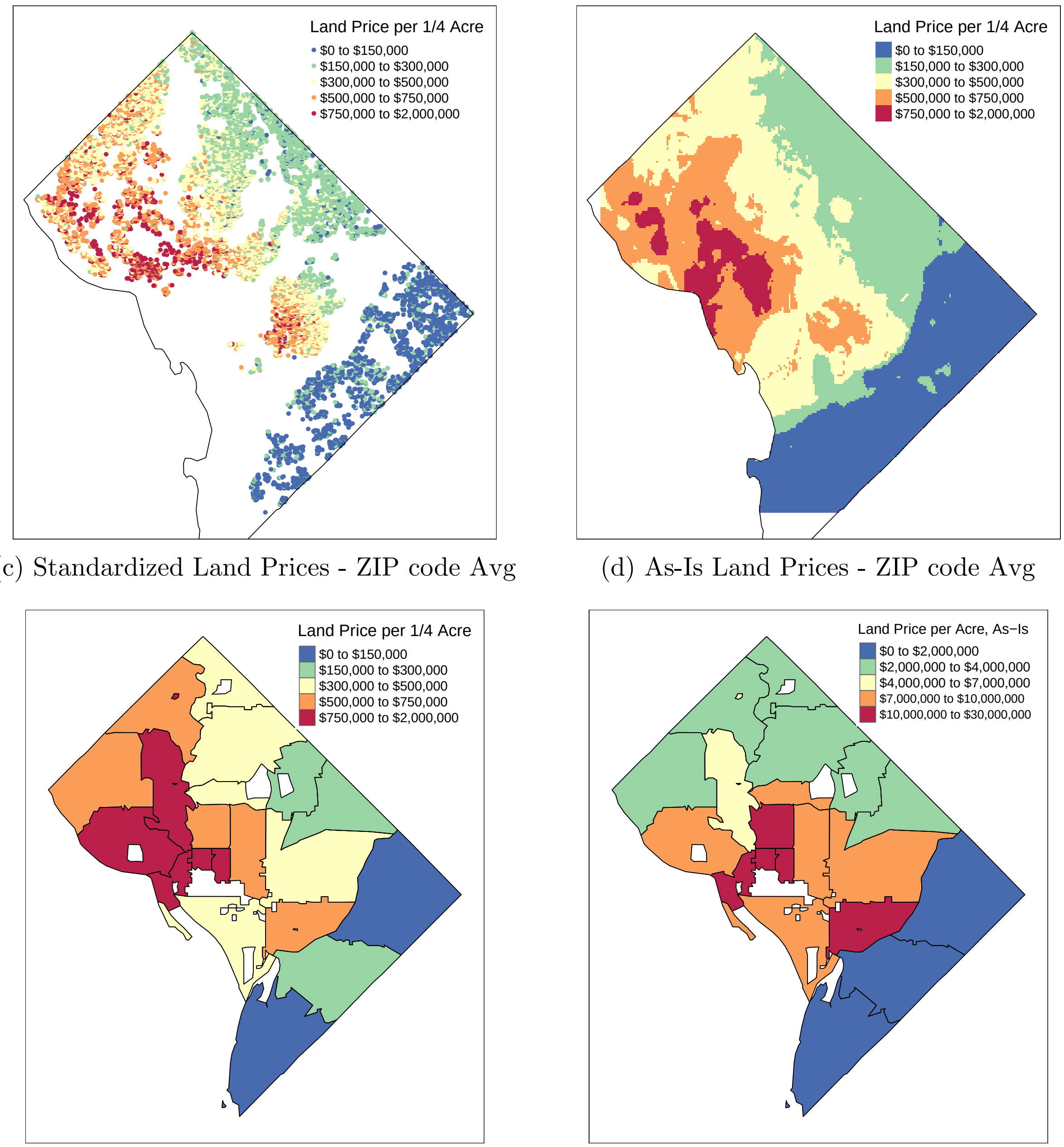



Notes: This figure presents the ratio of appraised values to assessed values for the entire property (top panel) and for the land component (bottom) using all properties in the pooled cross-section data set that have both an appraisal and a tax assessment. Appraisals are defined as anchored to assessments using methods described in the text. Values less than 0.5 and more than 2 are omitted from the display of the histogram but are included in the calculation of median values. Each red bar shows the median.

Figure 3: Washington, DC Kriging Example

(a) Standardized Land Prices - Working Sample (b) Standardized Land Prices - Kriged Surface

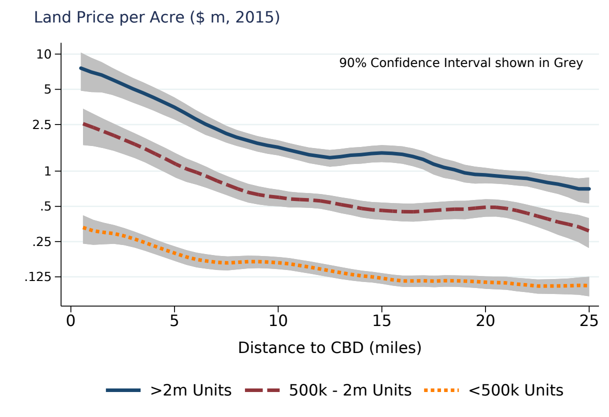
(c) Standardized Land Prices - ZIP code Avg (d) As-Is Land Prices - ZIP code Avg



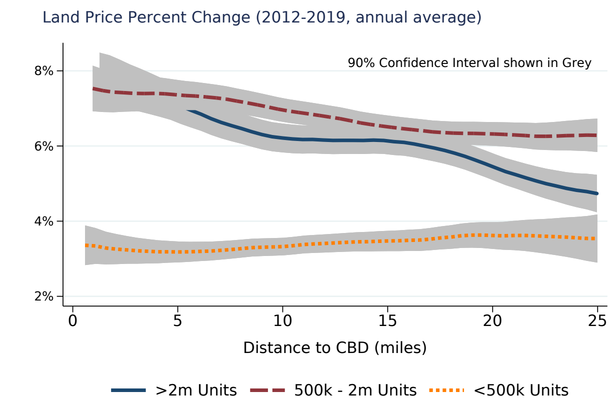

Notes: This figure shows the average level of land prices by ZIP code for the District of Columbia in the pooled cross-section data set over the period 2012-2019 with base year 2015 prices.

Figure 4: Land Price Gradients

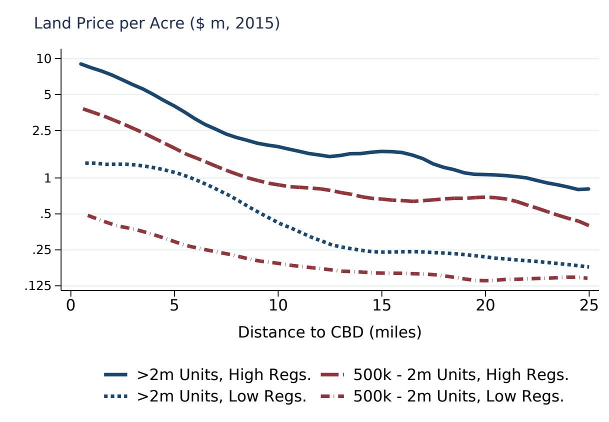
(a) Large vs. Small Metropolitan Areas



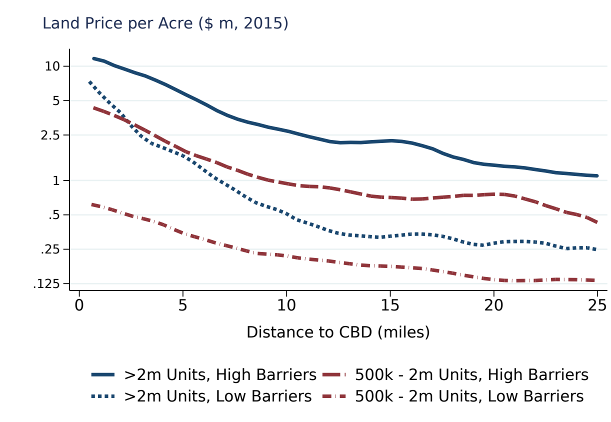

(b) Large vs. Small: Changes over Time




(c) High vs. Low Regulation

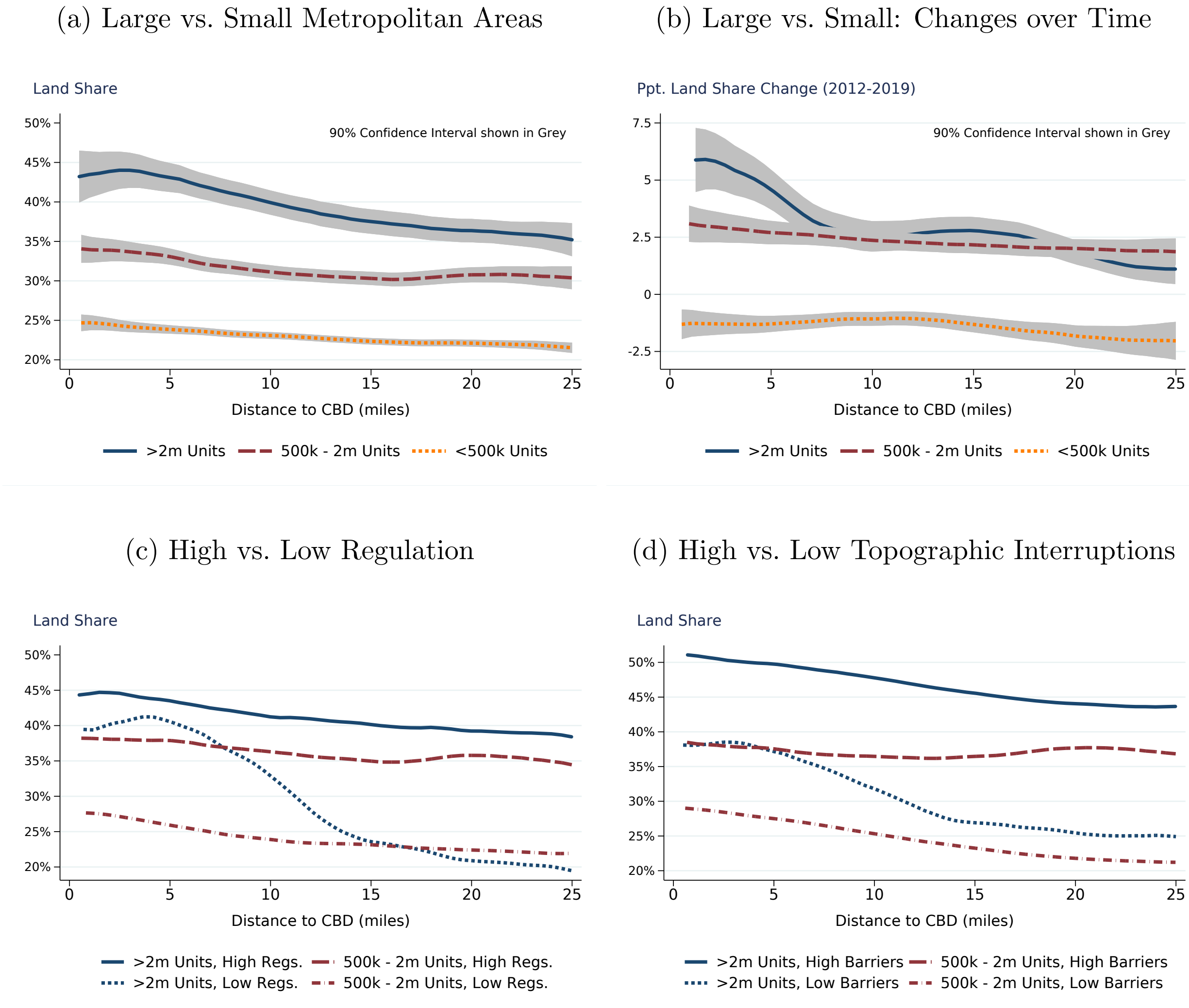



(d) High vs. Low Topographic Interruptions




Notes: Each panel includes all ZIP codes with centroids within 25 miles of an identified central ZIP code centroid. Panels (a), (c), and (d) are based on the pooled cross-section data set, while panel (b) shows the change from 2012 to 2019 in the annual panel data set. Each estimated line is fit using a local polynomial smoother.

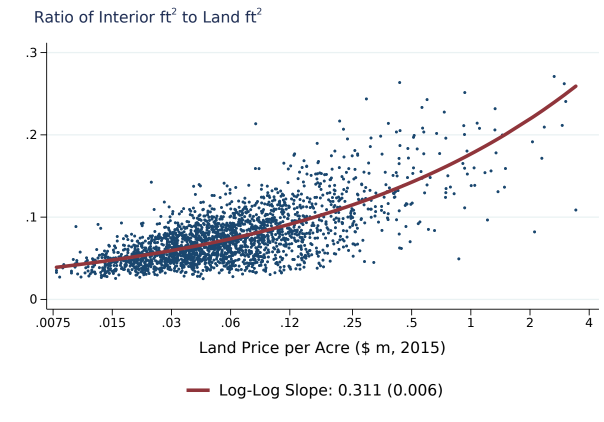
Figure 5: Land Share of Home Value Gradients



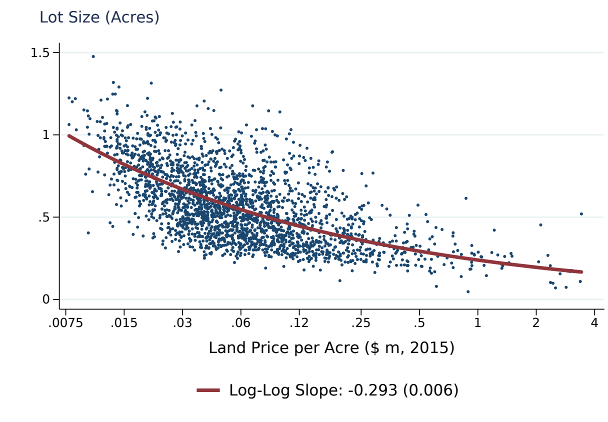

Notes: Each panel includes all ZIP codes with centroids within 25 miles of an identified central ZIP code centroid. Panels (a), (c), and (d) are based on the pooled cross-section data set, while panel (b) shows the change from 2012 to 2019 in the annual panel data set. Each estimated line is fit using a local polynomial smoother. Land values and house values are expressed in 2015 dollars before computing the land shares shown in panels (a), (c) and (d).

Figure 6: County Land Prices and Covariates

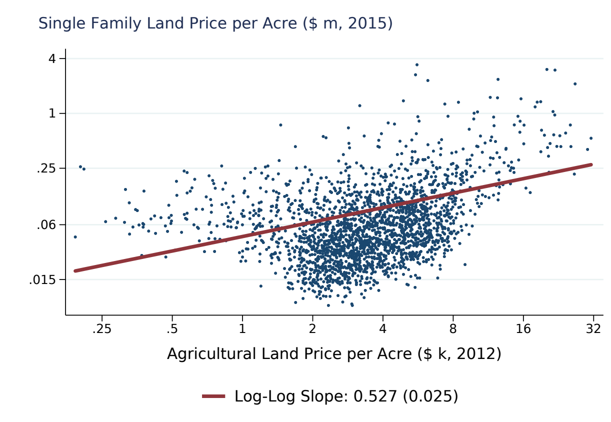
(a) Floor-Area Ratios and Land Prices




(b) Lot Size and Land Prices



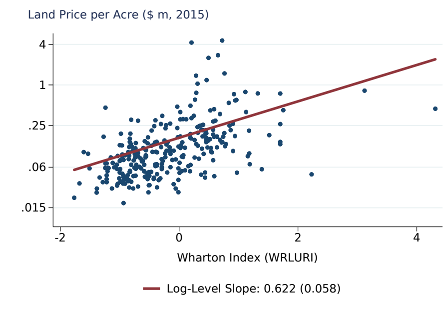

(c) Agricultural and Single-Family Land Prices



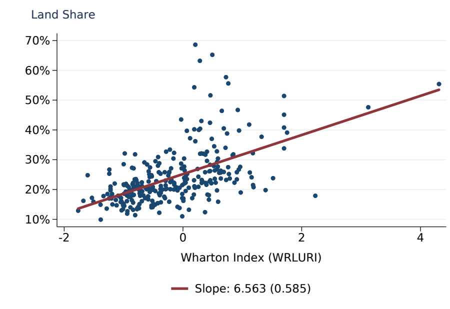

Notes: Panels (a) and (b) include all counties in the pooled cross-section data set, while panel (c) includes only the counties with agricultural land. Slope estimates are based on the equation $\ln Y = a + b \ln X + e$, with the fit line based on $\hat{Y} = \exp(\hat{a} + \hat{b} \ln X + 0.5\hat{\sigma}_e^2)$ using data from the entire sample. The value for $\hat{b}$ is presented in the legend, with the standard error in parenthesis. Panel (a) omits 14 observations, panel (b) omits 7 observations and panel (c) omits 16 observations to conserve on white space.

Figure 7: CBSA Land Prices, Shares, and Covariates

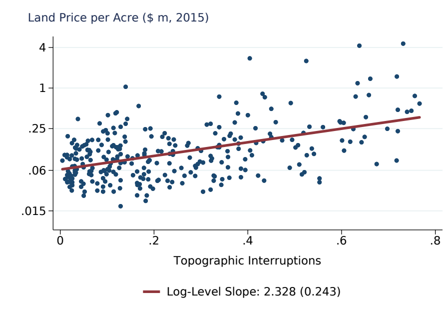
(a) Regulation - Land Prices



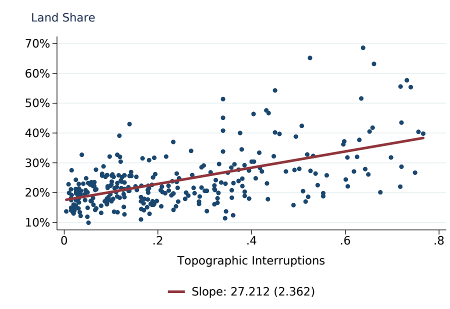

(b) Regulation - Land Shares



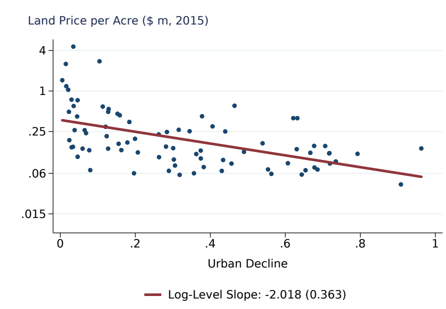

(c) Topographic Interruptions - Land Prices



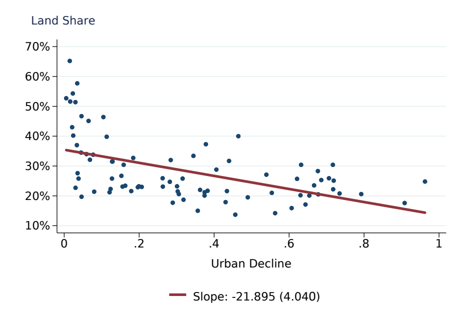

(d) Topographic Interruptions - Land Shares




(e) Urban Decline - Land Prices



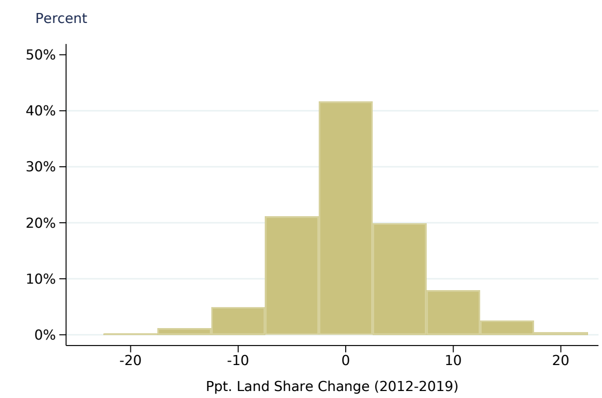

(f) Urban Decline - Land Shares



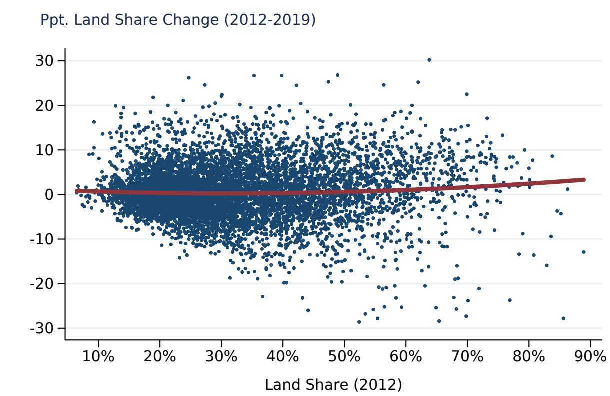

Notes: Each panel is based on the pooled cross-section data set, using all CBSAs with data for the covariate shown. Land values and house values are expressed in 2015 dollars before computing the land shares shown in panels (b), (d) and (f). For land prices, slope estimates are based on the equation $\ln Y = a + bX + e$, with the fit line based on $\hat{Y} = \exp(\hat{a} + \hat{b}X + 0.5\hat{\sigma}_e^2)$; for land shares, slope estimates are based on the equation $Y = a + bX + e$, with fit line equal to $\hat{Y} = \hat{a} + \hat{b}X$. The value for $\hat{b}$ is presented in the legend, with the standard error in parenthesis. Measures of regulation are from Gyourko, Saiz, and Summers (2008); topographic interruptions are from Saiz (2010); urban decline is from Glaeser and Gyourko (2005).

Figure 8: Changes in Land Share of Home Value by ZIP Code, 2012-2019

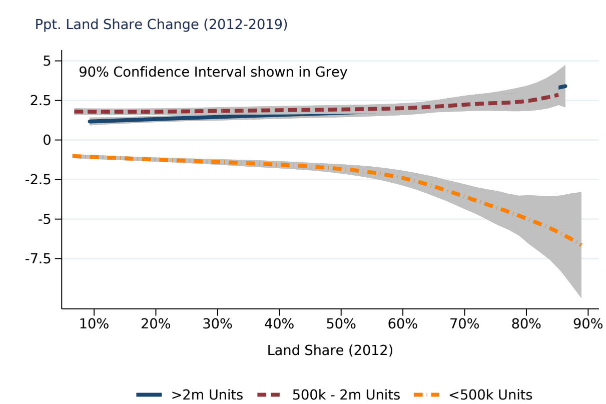
(a) Histogram




(b) By Initial Land Share



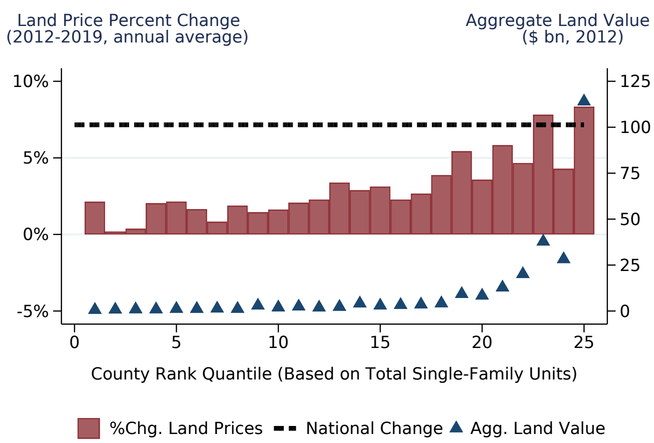

(c) By Metro Area Size and Initial Land Share



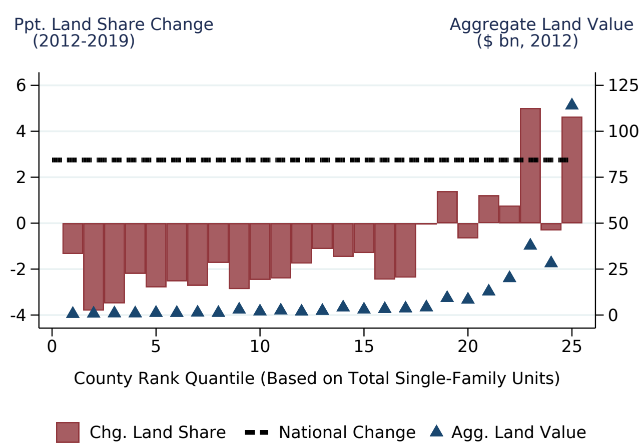

Notes: The panels in this figure show changes in average land share by ZIP code between 2012 and 2019 using the annual panel data set for those years. The solid line in panel (b) shows a quadratic fit through the individual data points. The lines shown in panel (c) are fit using a local polynomial smoother. Panel (a) omits observations with changes in the land share of more than 22.5 percentage points in absolute value.

Figure 9: Changes in Land Prices and Land Shares by County-Based Quantiles of Housing Units, 2012-2019

(a) Changes in Land Prices




(b) Changes in Land Shares




Notes: Counties in the annual panel data set are sorted into 25 bins based on the number of single-family housing units in 2013-2017, as measured in the American Community Survey for those years (5-year sample). In both panels, the blue triangles show the aggregate value of land in 2012 for all the counties in each bin. In the top panel, the red bars show annualized growth in the price of land for each bin and the black dashed line shows growth of the annualized price of land in the aggregate United States between 2012 and 2019, 6.47% per year. In the bottom panel, the red bars show the change in the land share for each bin and the black dashed line shows the change in the land share of the aggregate United States between 2012 and 2019, 2.75 percentage points.

Table 1: Land Statistics
As-Is Estimates of Land Value Reported Per Acre

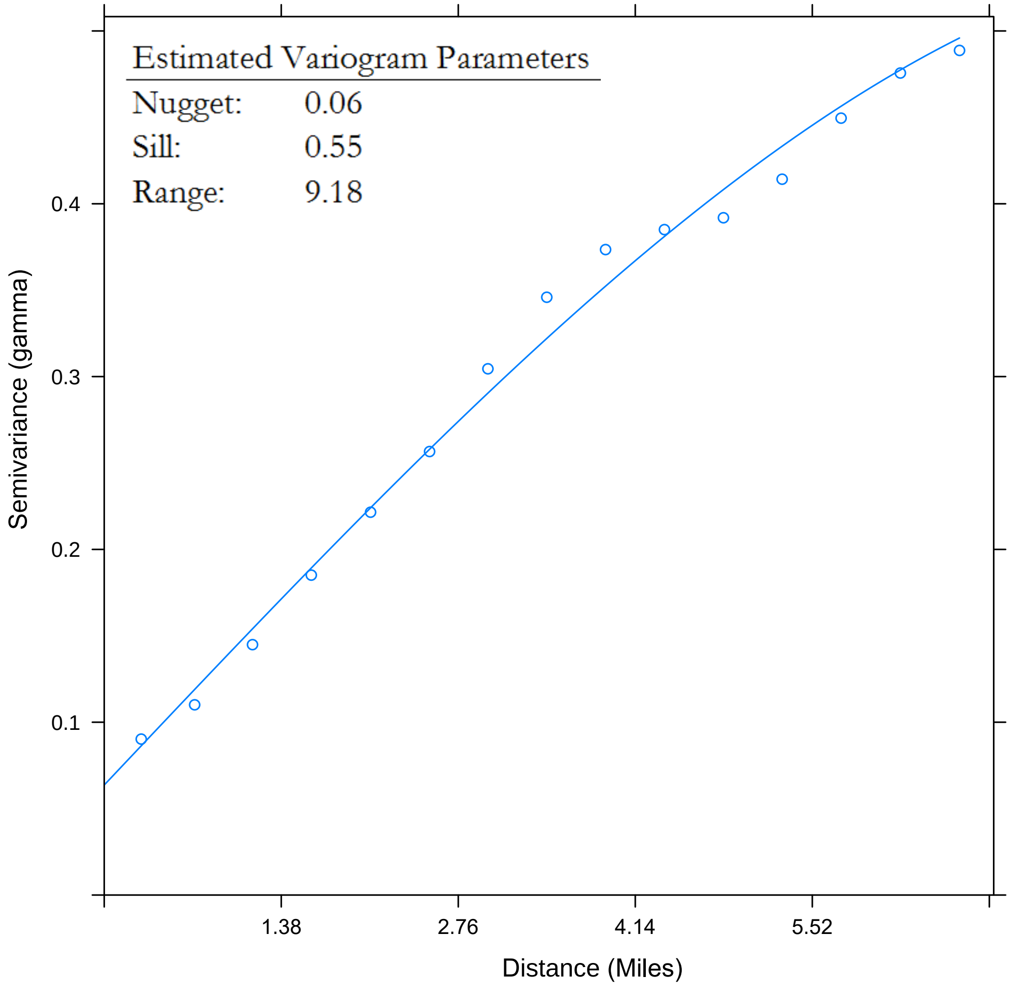
Pooled Cross Section (2,378 Counties):

| Variable | 1st | 10th | 25th | 50th | 75th | 90th | 99th | Avg. | Std. Dev. |
|---|---|---|---|---|---|---|---|---|---|
| Land Value | $11,300 | $19,900 | $30,900 | $52,950 | $98,500 | $204,100 | $1,447,300 | $151,944 | $1,258,006 |
| Land Share | 7.7% | 11.3% | 14.1% | 18.2% | 24.6% | 33.6% | 54.3% | 20.6% | 9.6% |

Annual Panel, Pooled (960 Counties):

| Variable | 1st | 10th | 25th | 50th | 75th | 90th | 99th | Avg. | Std. Dev. |
|---|---|---|---|---|---|---|---|---|---|
| Land Value | $22,200 | $39,300 | $59,100 | $101,250 | $190,400 | $395,400 | $3,119,500 | $241,651 | $755,451 |
| Land Share | 11.2% | 15.7% | 19.0% | 24.1% | 31.5% | 41.5% | 62.0% | 26.6% | 10.7% |

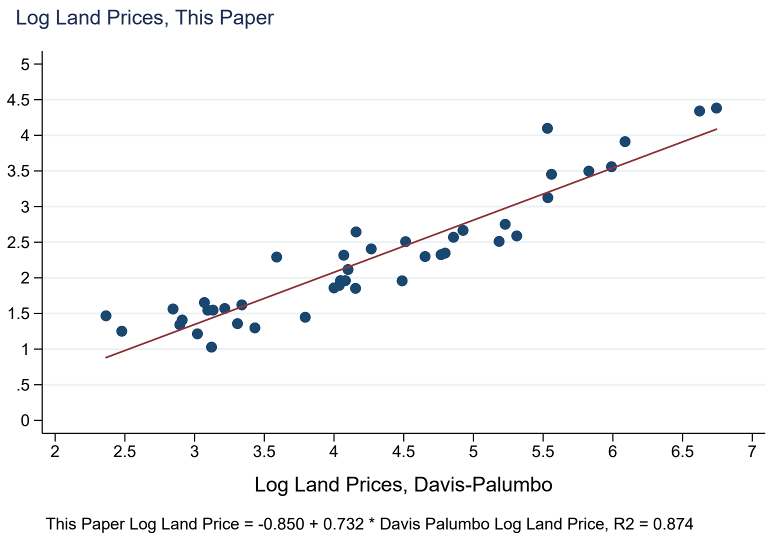
Figure A.1: Variogram for Washington, DC



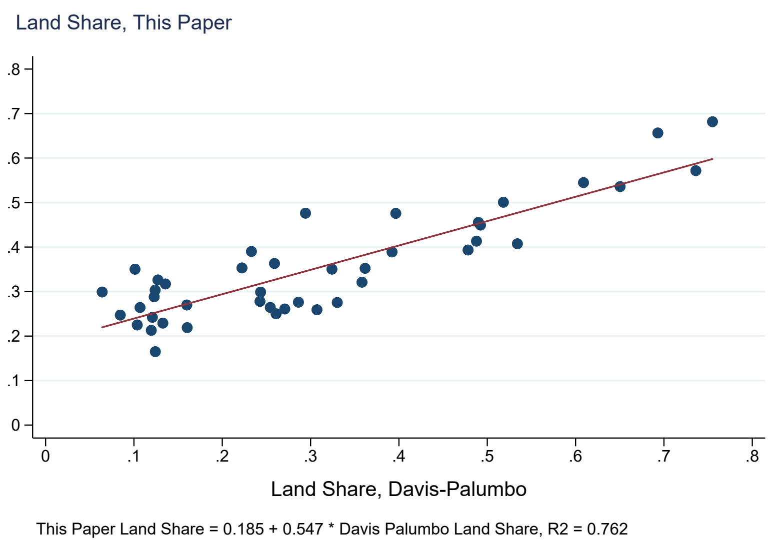

Notes: This figure presents the binned semivariances (hollow circles) and fitted variogram (blue line) for the District of Columbia. The variogram here is estimated for all parcel pairs extending to 6.9 miles.

Figure A.2: Comparison of CBSA-Level Land Prices and Shares to Davis and Palumbo (2008)

(a) Land Prices




(b) Land Shares




Notes: Panel (a) compares the log of land prices from Davis and Palumbo (2008) to the log of land prices from this paper for 43 CBSAs over the 2012-2018 period. For both data sets, the log of the average value of land prices over 2012-2018 is reported. Panel (b) compares average land shares over 2012-2018 for the same CBSAs.

Table A.1: Interpolation RMSE (20% hold-out sample)

| | Kriging | IDW | NN | Null-Tract | Null-ZIP Code | Null-County | Hold-Out Obs |
|---|---|---|---|---|---|---|---|
| 2012 | 0.394 | 0.420 | 0.430 | 0.440 | 0.444 | 0.513 | 99,395 |
| 2013 | 0.389 | 0.403 | 0.421 | 0.430 | 0.442 | 0.522 | 147,759 |
| 2014 | 0.387 | 0.408 | 0.418 | 0.424 | 0.438 | 0.513 | 126,714 |
| 2015 | 0.383 | 0.399 | 0.414 | 0.424 | 0.431 | 0.520 | 166,654 |
| 2016 | 0.379 | 0.395 | 0.410 | 0.419 | 0.430 | 0.506 | 201,222 |
| 2017 | 0.386 | 0.401 | 0.418 | 0.425 | 0.437 | 0.513 | 169,185 |
| 2018 | 0.385 | 0.403 | 0.421 | 0.424 | 0.434 | 0.519 | 142,231 |
| 2019 | 0.381 | 0.395 | 0.399 | 0.412 | 0.429 | 0.504 | 167,683 |

Notes: Interpolation RMSE calculated as follows. 1) Estimate an interpolated estimate for each hold-out parcel for each year. 2) Calculate an RMSE for each county for each year. 3) Calculate the median RMSE across counties (reported in table). IDW = inverse-distance weights, NN = nearest neighbor.

Table A.2: Interpolation RMSE (20% hold-out sample), alternative parameterizations

| Estimator | Mean | Median | SD |
|---|---|---|---|
| Null - County Average | 0.514 | 0.576 | 0.228 |
| NN - 20 NN | 0.418 | 0.453 | 0.167 |
| IDW - 20 NN | 0.420 | 0.462 | 0.181 |
| Kriging - 10 NN, 6.9 Mile Boundary | 0.395 | 0.431 | 0.162 |
| Kriging - 20 NN, 6.9 Mile Boundary | 0.395 | 0.429 | 0.162 |
| Kriging - 30 NN, 6.9 Mile Boundary | 0.393 | 0.430 | 0.163 |
| Kriging - 20 NN, 3.4 Mile Boundary | 0.394 | 0.430 | 0.164 |
| Kriging - 20 NN, 10.4 Mile Boundary | 0.395 | 0.430 | 0.163 |

Notes: Sample is the pooled cross-section (2,378 counties). Interpolation RMSE calculated as follows. 1) Estimate an interpolated estimate for each hold-out parcel. 2) calculate an RMSE for each county for each year. 3) Calculate the median/mean/SD RMSE across counties (reported in table).

Table A.3: Predictions of Calibrated Urban Model

| $d$ | $q_d^h$ | $h_d$ | $q_d^h h_d$ | $s$ | $q_d^l$ | $l_d$ | $q_d^l l_d$ | land share | $l_d$ (acres) | $q_d^l$ per acre |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1.000 | 1,000,000 | $1,000,000 | $464,159 | 0.413 | 1,295,995 | $535,841 | 54% | 0.25 | $2,143,364 |
| 1 | 0.922 | 1,062,482 | $980,000 | $480,054 | 0.354 | 1,411,219 | $499,946 | 51% | 0.27 | $1,836,505 |
| 2 | 0.849 | 1,130,281 | $960,000 | $496,838 | 0.300 | 1,543,749 | $463,162 | 48% | 0.30 | $1,555,320 |
| 3 | 0.781 | 1,203,972 | $940,000 | $514,581 | 0.251 | 1,697,826 | $425,419 | 45% | 0.33 | $1,298,934 |
| 4 | 0.716 | 1,284,211 | $920,000 | $533,360 | 0.206 | 1,879,309 | $386,640 | 42% | 0.36 | $1,066,527 |
| 5 | 0.656 | 1,371,742 | $900,000 | $553,260 | 0.165 | 2,096,587 | $346,740 | 39% | 0.40 | $857,343 |
| 6 | 0.600 | 1,467,412 | $880,000 | $574,375 | 0.129 | 2,362,219 | $305,625 | 35% | 0.46 | $670,706 |
| 7 | 0.547 | 1,572,189 | $860,000 | $596,810 | 0.098 | 2,696,126 | $263,190 | 31% | 0.52 | $506,049 |
| 8 | 0.498 | 1,687,183 | $840,000 | $620,680 | 0.070 | 3,132,449 | $219,320 | 26% | 0.60 | $362,959 |
| 9 | 0.452 | 1,813,671 | $820,000 | $646,116 | 0.047 | 3,736,426 | $173,884 | 21% | 0.72 | $241,250 |
| 10 | 0.410 | 1,953,125 | $800,000 | $673,261 | 0.027 | 4,655,227 | $126,739 | 16% | 0.90 | $141,134 |

This table reports model-generated values of the price-per-unit of housing $q_d^h$, housing quantities $h_d$, the value of housing $q_d^h h_d$, the value of structures $s$, the price-per-unit of land $q_d^l$, the quantity of land $l_d$, the value of land $q_d^l l_d$, land's share of house value, the quantity of land in acres, and the price per land per-acre as a function of the distance to CBD $d$.

Table A.4: Land price per acre, as predicted by the model and estimated via Kriging from two data sets

| | | From Kriging Procedure | | | |
|---|---|---|---|---|---|
| | | Data Set 1 | | Data Set 2 | |
| Distance | From Model | Predicted | % Error | Predicted | % Error |
| 0 | $2,143,364 | $2,137,352 | 0.28% | $2,107,724 | 1.70% |
| 1 | $1,836,505 | $1,836,186 | 0.02% | $1,762,402 | 4.00% |
| 2 | $1,555,320 | $1,555,104 | 0.01% | $1,575,406 | -1.30% |
| 3 | $1,298,934 | $1,298,750 | 0.01% | $1,276,002 | 1.80% |
| 4 | $1,066,527 | $1,066,499 | 0.00% | $998,657 | 6.40% |
| 5 | $857,343 | $857,259 | 0.01% | $878,682 | -2.50% |
| 6 | $670,706 | $670,688 | 0.00% | $641,687 | 4.30% |
| 7 | $506,049 | $506,015 | 0.01% | $450,198 | 11.00% |
| 8 | $362,959 | $362,945 | 0.00% | $364,878 | -0.50% |
| 9 | $241,250 | $241,259 | 0.00% | $200,874 | 16.70% |
| Mean | | | 0.03% | | 4.16% |

This table reports model-generated price per-acre of land and the kriging-based estimates of that price when the model is simulated without measurement error (columns marked Data Set 1) and when the model is simulated with measurement error (columns marked Data Set 2).