METROPOLITAN LAND VALUES

David Albouy, Gabriel Ehrlich, and Minchul Shin*

Abstract—We estimate the first cross-sectional index of transaction-based land values for every U.S. metropolitan area. The index accounts for geographic selection and incorporates novel shrinkage methods using a prior belief based on urban economic theory. Land values at the city center increase with city size, as do land-value gradients; both are highly variable across cities. Urban land values are estimated at more than two times GDP in 2006. These estimates are higher and less volatile than estimates from residual (total – structure) methods. Five urban agglomerations account for 48% of all urban land value in the United States.

I. Introduction

WE estimate the first index of land values across U.S. metropolitan areas that is based on directly observed market transactions and cross-sectionally comparable. Standard economic theory (e.g., Roback, 1982; Brueckner, 1983; Albouy, 2016) suggests that this index captures differences in the combined value of household amenities, employment, and building opportunities, ignoring cross-metro externalities. Urban land values have been central to questions of wealth, income, and taxation since the seminal works of Ricardo (1821) and George (1884).

Unfortunately, market data on land values have been notoriously piecemeal and subject to numerous measurement challenges. Flow of funds (FOF) accounts of the Federal Reserve stopped publishing series for land value in 1995 because of accuracy concerns made plain by negative values inferred for land. We aim to overcome these challenges using a large national data set of market transactions of land from the CoStar COMPS database and an econometric model informed by urban theory. Our estimates of urban land values prove to be higher and more stable than values implied by the FOF.

Our indices of both central and average land values have intuitive properties. While they vary considerably, the indices increase with city area, providing nuanced support for the monocentric city model of Alonso (1964), Mills (1967), and Muth (1969). The highest central land values are found in New York, Chicago, Washington, San Francisco, and Los Angeles. In these cities, central values are 21 times higher than peripheral values 10 miles away, although across all cities, the (unweighted) average ratio of central to peripheral values is only 4. Over their entire urban areas, New York, Jersey City, Honolulu, San Francisco, and Los Angeles–Long Beach have the highest average values, which are 82 times higher than those in the lowest five cities. Values in 2009 averaged \$373,000 per acre, down from \$624,000 in 2006, as the total value of urban land fell from \$28 million to \$18 trillion, or from 2.2 to 1.3 times GDP.

II. Description of Transactions Data and Urban Land Area

Our primary data source is the CoStar COMPS database, with land transaction prices recorded between 2005 and 2010.[1] CoStar provides fields containing the price, lot size, address, and a proposed use for each property. We exclude transactions CoStar has marked as non-arm's-length, without complete information, that feature a structure, are over 60 miles from the city center, or are less than \$100 per acre. The remaining data set contains 68,756 observed land sales.[2]

The "cities" we examine correspond to 1999 OMB definitions of Metropolitan Statistical Areas (MSAs). Some MSAs, known as consolidated MSAs (CMSAs) are divided into constituent primary MSAs, which we treat as separate cities. In 2000, all MSAs accounted for 80% of the U.S. population. Because MSAs consist of counties, which often contain a large amount of agricultural land, we consider only land that is part of an urban area by 2000 Census definitions. The main requirement is that the area consists of contiguous block groups with a population density of over 1,000 residents per square mile (1.56 per acre), with a total population of over 2,500.

We take city centers to be the city hall or mayor's office of each city. Many MSA names contain multiple cities (e.g., Minneapolis–St. Paul). We address this by considering each named city as having its own center. Land parcels within the MSA are assigned to the city center closest in Euclidean

Received for publication August 31, 2015. Revision accepted for publication June 26, 2017. Editor: Bryan S. Graham.

* Albouy: University of Illinois and NBER; Ehrlich: University of Michigan; Shin: University of Illinois.

We thank Henry Munneke, Nancy Wallace, and participants at seminars at the AREUEA Annual Meetings (Chicago), Ben-Gurion University, Brown University, the Federal Reserve Bank of New York, the Housing-Urban-Labor-Macro Conference (Atlanta), Hunter College, the NBER Public Economics Program Meeting, the New York University Furman Center, the University of British Columbia, the University of California, the University of Connecticut, the University of Georgia, the University of Illinois, the University of Michigan, the University of Rochester, the University of Toronto, the Urban Economics Association Annual Meetings (Denver), and Western Michigan University for their help and advice. We especially thank Morris Davis, Andrew Haughwout, and Matthew Turner for sharing data, or information about data, with us. Nicolas Bottan provided outstanding research assistance. The National Science Foundation (grant SES-0922340) generously provided financial assistance.

A supplemental appendix is available online at http://www.mitpressjournals.org/doi/suppl/10.1162/rest_a_00710.

[1] The CoStar Group claims to have the commercial real estate industry's largest research organization. The COMPS database provided by CoStar University is not publicly available but can be accessed for free by academics. The data include transaction details for all types of commercial real estate. We use every sale CoStar considers "land." Recently, a small literature has used this data for analyses within metro areas. Haughwout, Orr, and Bedoll (2008) demonstrate the data's extensive coverage and construct a land price index for 1999 to 2006 within the New York metro area. Kok, Monkkonen, and Quigley (2014) document land sales within the San Francisco Bay Area and relate sales prices to topographical, demographic, and regulatory features. Nichols, Oliner, and Mulhall (2013) construct a panel of land-value indices for 23 metros from the 1990s to 2009. These indices are for use over time and are not comparable across metros.

[2] Appendix A provides additional detail on the data treatment.

The Review of Economics and Statistics, July 2018, 100(3): 454–466

doi:10.1162/rest_a_00710

distance. In such cases, our central values average the named centers.

Online appendix figure A.1 displays the geographic pattern of land sales for four CMSAs: New York, Los Angeles, Chicago, and Houston. The figure shows that land sales are well dispersed throughout the metro areas, with sales activity more frequent near city centers.[3]

III. Econometric Methods

There are two major obstacles to constructing a cross-metropolitan land value index from observed transactions data. First, observed transactions are not a random sample of all parcels in a city. Second, we observe few sales in many smaller metro areas, reducing the reliability of the estimates. Our econometric methods try to overcome both of these obstacles.

A. *Regression Model of Land Values over Space and Time*

Following the monocentric city model, we take each city j as having a fixed center, with coordinates $\mathbf{z}_j^c$. Land values, r, vary according to a city-specific polynomial in the distance metric, $D(\mathbf{z}_{ij}, \mathbf{z}_j^c)$, between plot i's coordinates $\mathbf{z}_{ij}$ and the center. City center values α_{jt} may vary by year, t; coefficients δ_{jk}, which determine the shape of the value-distance gradient, are held constant over time due to limited sample sizes:

$$\ln r_{ijt} = \sum_{t=2005}^{2010} \alpha_{jt} + \sum_{k=1}^{K} \delta_{jk} \left[D(\mathbf{z}_{ij}, \mathbf{z}_j^c)\right]^k + X_{ijt}\beta + e_{ijt},$$
$$e_{ijt} \sim \text{ i.i.d. } N(0, \sigma_e^2). \tag{1}$$

Controls X_{ijt} include proposed use and lot size. The idiosyncratic error term, e_{ijt}, follows an independent and identically distributed normal distribution.[4]

Figure 1a shows estimated first-order and fourth-order polynomials for the Houston MSA, along with the underlying transaction prices. Both polynomials slope downward with distance, but the fourth-order polynomial reveals a subtler distance function.

B. *Shrinkage Estimation and Its Target "Meta-City"*

To deal with limited sample sizes, we develop a hierarchical model. It "shrinks" metro-level estimates toward a national average function. This function target depends on each city's urban area, A_j. We begin by decomposing the central value α_{jt} into two components, $\alpha_{jt} = \alpha_j + \alpha_{jt}^\star$, where $\alpha_{j2005}^\star$ is normalized to 0. The time-varying component follows the prior $\alpha_{jt}^\star \sim N(\tau_t, \sigma_t^2)$. Vectorizing the distance coefficients $\boldsymbol{\delta_j} = \begin{bmatrix}\delta_{j1} & \delta_{j2} & \cdots & \delta_{jK}\end{bmatrix}'$, time-invariant cross-sectional priors are modeled as

$$\begin{bmatrix}\alpha_j \\ \boldsymbol{\delta_j}\end{bmatrix} = \begin{bmatrix}a_0 & a_1 \\ \mathbf{d_0} & \mathbf{d_1}\end{bmatrix}\begin{bmatrix}1 \\ \ln A_j\end{bmatrix} + \begin{bmatrix}e_{\alpha,j} \\ \mathbf{e}_{\boldsymbol{\delta},\boldsymbol{j}}\end{bmatrix},$$
$$\begin{bmatrix}e_{\alpha,j} \\ \mathbf{e}_{\boldsymbol{\delta},\boldsymbol{j}}\end{bmatrix} \sim \text{ i.i.d. } N\left(\begin{bmatrix}\mathbf{0} \\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\Sigma_{\alpha\alpha} & \boldsymbol{\Sigma_{\alpha\delta}} \\ \boldsymbol{\Sigma_{\delta\alpha}} & \boldsymbol{\Sigma_{\delta\delta}}\end{bmatrix}\right). \tag{2}$$

This technique essentially constructs a "metacity" described by the parameters $a_0, a_1, \boldsymbol{\delta_0}$, and $\boldsymbol{\delta_1}$. The metacity provides the land rent gradient typical of a city with area A_j. This area adjustment is important as larger cities typically have higher central land values. These land values descend and dovetail with agricultural (or other nonurban) values at different rates from the center than in smaller cities. The model allows for a full covariance matrix between the random components of the intercept and distance coefficients, $e_{\alpha,j}$ and $\mathbf{e}_{\boldsymbol{\delta},\boldsymbol{j}}$.

When all other parameters are known and $\alpha_{jt}^\star = 0$, the best linear unbiased predictor (BLUP) for $[\alpha_j, \boldsymbol{\delta_j'}]'$ is a weighted average between their prior mean and conventional metro-level (fixed effect) estimates, $[\widehat{\alpha}_j, \widehat{\boldsymbol{\delta}}_{\boldsymbol{j}}]'$:

$$\begin{bmatrix}\widetilde{\alpha}_j \\ \widetilde{\boldsymbol{\delta}}_{\boldsymbol{j}}\end{bmatrix} = \mathbf{W_j}\begin{bmatrix}a_0 & a_1 \\ \mathbf{d_0} & \mathbf{d_1}\end{bmatrix}\begin{bmatrix}1 \\ \ln A_j\end{bmatrix} + (\mathbf{I} - \mathbf{W_j})\begin{bmatrix}\widehat{\alpha}_j \\ \widehat{\boldsymbol{\delta}}_{\boldsymbol{j}}\end{bmatrix}, \tag{3}$$

where the weighting matrix $\mathbf{W_j}$ accounts for the amount of shrinkage in city j. This shrinkage term falls with the number of observations in city j and rises with the uncertainty in the prior ($\Sigma_{\alpha\alpha}, \boldsymbol{\Sigma_{\delta\alpha}}, \boldsymbol{\Sigma_{\delta\delta}}$) and the idiosyncratic error term (σ_e^2).

The second component in the intercept, α_{jt}^*, captures the city-specific time trend. By similar logic, we shrink the MSA-level time trend toward the national-level time trend, τ_t, where the degree of the heterogeneity in MSA-level time trends is allowed to change over time through σ_t^2.

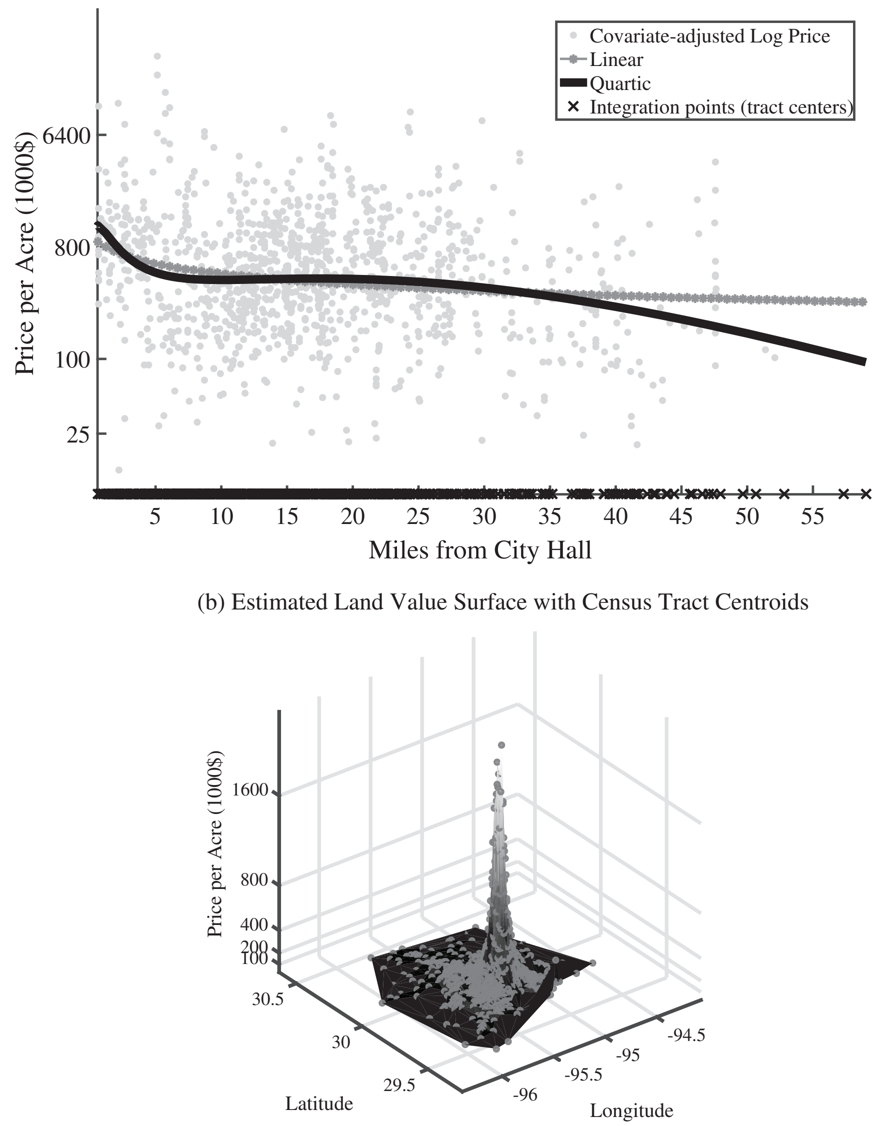
[3] Land transactions are not randomly distributed over space. Yet as Haughwout, Orr, and Bedoll (2008) comment on the New York data, "Overall, vacant land transactions occurred throughout the region, with a heavy concentration in the most densely developed areas." As Nichols, Oliner, and Mulhall (2013) discuss, it is impossible to correct for all types of selection bias without observing transaction prices for unsold lots, a logical contradiction. Fortunately, the literature has generally found selection bias to be minor for land and commercial real estate prices. Colwell and Munneke (1997), studying land prices in Cook County, Illinois, report, "The estimates with the selection variable and those without are surprisingly consistent for each land use." Studying the office market in Phoenix, Munneke, and Barrett (2000) find that "the price indices generated after correcting for sample-selection bias do not appear significantly different from those that do not consider selectivity bias." In their construction of metro price indices, Munneke and Barrett (2001) report, "Little selection bias is found in the estimates." Finally, Fisher, Geltner, and Pollakowski (2007), in their study of commercial real estate properties, state, "Sample selection bias does not appear to be an issue with our annual model specification." Nevertheless, we correct for selection bias on observables in section III.

[4] We define $D(\mathbf{z}_{ij}, \mathbf{z}_j^c) = \ln\left(1 + \|\mathbf{z}_{ij} - \mathbf{z}_j^c\|\right)$, using Euclidean distances in miles. Adding 1 in the logarithm argument creates two desirable features. First, it dampens the effect of small changes in distance very close to the city center. Second, it makes D operate as a distance metric, so that the α_{jt} coefficients may be interpreted as (finite) log land values at the city center. Since the true gradient may vary along rays with different angles from the center, this serves largely as an averaging technique, used for comparisons across cities. Some cities have land rent gradients that decline monotonically from the center all the way to their agricultural fringe. Others see a dip in central city values that rise again for the inner suburbs, before declining again at the fringe.

FIGURE 1.—EXAMPLE OF LAND VALUE GRADIENT ESTIMATES FOR THE HOUSTON METRO AREA

(a) Estimated Distance Polynomial with $D = \ln(1 + \text{mileage})$

(b) Estimated Land Value Surface with Census Tract Centroids



Our empirical model is then completed by specifying the joint distribution of error terms, controls, and the prior. We assume that observed control variables are random and strictly exogenous. That is, for each city j, the error term vector $e_j = \{\{e_{ijt}\}_{i=1}^{n_j}\}_{t=2005}^{2010}$ is uncorrelated with the control vector $\{\{D(\mathbf{z_{ij}}, \mathbf{z_j}^c), \ldots, D(\mathbf{z_{ij}}, \mathbf{z_j}^c)^K, \{X'_{ijt}\}_{t=2005}^{2010}\}_{i=1}^{n_j}, \ln A_j\}$ and the random component of the coefficient vector $\{e_{\alpha,j}, \boldsymbol{e}_{\boldsymbol{\delta},\boldsymbol{j}}', \alpha^{\star}_{j2006}, \ldots, \alpha^{\star}_{j2010}\}$. In addition, the random component of the coefficient vector is uncorrelated with the control vector a priori.

In practice, to estimate the BLUP for the random intercept and gradient parameters, the unknown fixed parameters $(\beta, a_0, a_1, \mathbf{d_0}, \mathbf{d_1})$ and variance parameters $(\sigma^2, \Sigma_{\alpha\alpha}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha\delta}}, \boldsymbol{\Sigma}_{\boldsymbol{\delta\delta}})$ must also be estimated. To do this, we adopt an empirical Bayes-type approach in which these parameters are found by maximizing the marginal likelihood with a flat improper prior. Then we obtain estimates for $[\alpha_{jt}, \boldsymbol{\delta}_j']'$ by substituting these estimates into the posterior mean formula as if the fixed and variance parameters were known. Appendix B describes the shrinkage procedure in much greater detail.

C. *Integrating Land Values over the Urban Area*

We use the estimated land value functions to compute average land values over each city's urban area in each year. For each census tract l in city j in year t, we calculate the predicted land value $\hat{r}_{ljt}$ at the tract centroid. The predicted value is based on the expected characteristics X (planned use and lot size) of the tract, conditional on the city, distance

TABLE 1.—ECONOMETRIC MODEL CROSS-VALIDATION RESULTS

	Model Specification						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A: Three observations per city-year							
Mean squared error	1.640	1.143	0.939	0.938	0.936	0.936	0.935
Bias	−0.004	0.013	0.016	0.013	0.013	0.013	0.013
Variance	1.586	1.105	0.910	0.909	0.907	0.906	0.905
B: Thirty observations per city-year							
Mean squared error	1.449	0.912	0.904	0.902	0.898	0.897	0.896
Bias	−0.004	−0.003	0.001	0.000	0.001	0.001	0.000
Variance	1.441	0.907	0.899	0.898	0.893	0.892	0.891
Shrunken?	No	No	Yes	Yes	Yes	Yes	Yes
Polynomial order: Distance	0	1	1	2	3	4	4
Polynomial order: Lot size	0	1	1	1	1	1	3

Out-of-sample cross-validation exercise described in detail in appendix B. Column 1 shows results of a naive model that is the simple average of values per acre. Columns 2 through 7 contain controls for all covariates in appendix table A1. Panel A shows results for an exercise in which three observations per city-year are combined with all out-of-city data to predict remaining land values in city. Panel B shows results for an exercise in which thirty observations per city-year are combined with all out-of-city data to predict remaining land values in city. Out-of-sample predictions in both panels were conducted in 58 cities that had at least 50 observations per year for at least two years.

from the center and coast, and observed transaction data. We then assign that average value to the entire tract.[5] This value is then multiplied by the area of each tract A_{jl}, excluding any nonurban block groups. The total value of land in city j is then $R_{jt} = \sum_l A_{jl}\hat{r}_{ljt}$, and the average value is $r_{jt} = R_{jt}/A_j$. In other words, total land values in city j are the volume of the estimated land value "cone," while the average land value is the cone's average height. Figure 1b displays the estimated cone for the Houston MSA, with the small dots representing Census tract centers. Very high land values at the city center are clearly visible in the figure, which also shows slightly elevated values for the Census tracts near the coastline.

The estimated "meta-city" parameters allow us to impute land values for metros with no observations, in which case $\boldsymbol{W_j} = \boldsymbol{I}$. Tract values are imputed based on typical intercepts and gradients for cities of size A_j in year t, using their position relative to the closest city center and coastlines.

D. *Model Selection and Cross-Validation*

The cross-validation exercise summarized in table 1 assesses the performance of several econometric specifications, as detailed in appendix B. The exercise fixes a number of MSAs and retains a few observations per year. It then uses those few observations and the model estimates from other MSAs to predict the values of the nonretained observations. The mean squared error (MSE) between the predicted price and the actual price of these nonretained observations is used to assess the model. Results in panel A retain three observations per city-year; panel B retains thirty.

The first specification, in column 1, is of a naive model that takes the (geometric) average value per acre of all sales by metro. It establishes a baseline for other models to improve on. The second column shows the results from a simple version of model (1), with only linear city-specific terms in distance ($K = 1$), as well as city-time specific intercepts, measures of coastal proximity, controls for proposed use, and a linear term in log lot size. This basic econometric model lowers the mean squared error (MSE) over the naive model substantially by reducing the variance of the estimates. The third specification applies the empirical Bayes' shrinkage technique according to the prior (2), allowing both intercepts and gradients to be random. As expected, this produces a substantial improvement by further reducing the variance. As seen by lower prediction errors, both the monocentric regression model and the shrinkage technique help overcome the obstacles of small samples and nonrandom locations.

The rest of the table considers what are minor improvements. The fourth through sixth columns contain additional distance polynomials to the model in equation (3). Allowing for a more flexible distance gradient reduces the MSE only moderately. The final column includes a cubic polynomial in log lot size, which also slightly improves the prediction. As further terms produce no noticeable improvement, we take the model from column 7 with Bayesian shrinkage, a quartic polynomial in distance, and a cubic polynomial in log lot size as our preferred specification.

IV. Cross-Sectional Results

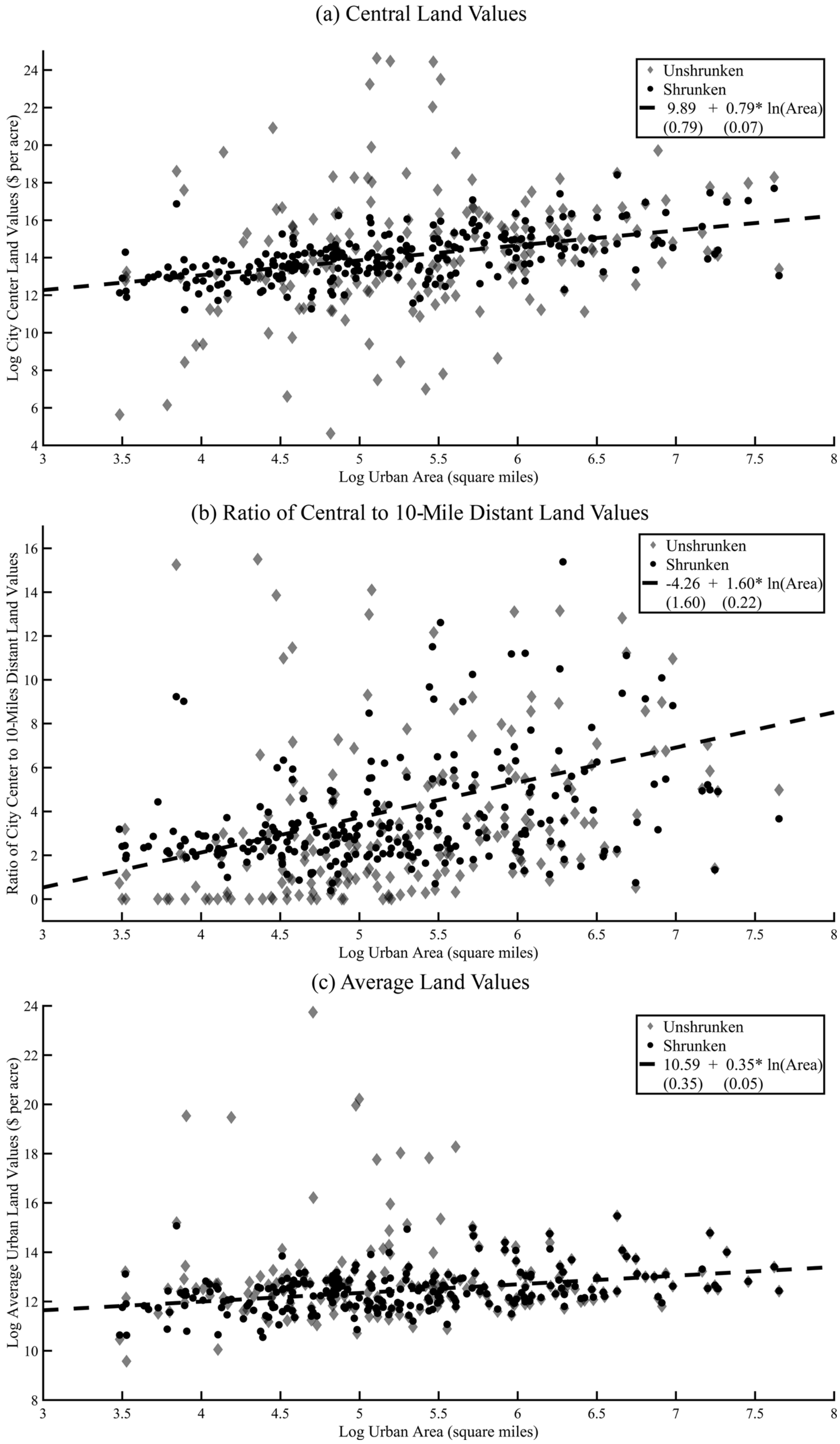
A. *Patterns in the Data*

Figures 2a to 2c plot estimated central land values, the ratio of those values to values 10 miles from downtown, and average land values, each against the urban area of the metro area.[6] The gray dots represent the unshrunken estimates; the dark dots, the shrunken estimates. The vertical distances between the two display how much the Bayesian approach

[5] We include only tract centers within 60 miles of the city center. To obtain the predicted characteristics X, we build and estimate a model for characteristics X that is a similar but simplifed version of the hierarchical model used for the land prices. The procedure and required assumptions for the land value prediction at the tract centroid is discussed at length in section 3 in appendix B.

[6] We take land values one-half mile from the point defined at the center as our measure of central land values.

FIGURE 2.—ESTIMATION RESULTS FOR ALL METRO AREAS



shrinks the estimates. Larger cities, which feature more observations, experience less shrinkage, as the additional observations make the prior less important.

The dashed upward-sloping line of best fit in figure 2a reflects the tendency of larger cities to have more expensive central land. A 10% increase in a city's footprint implies an 8% increase in the central land value. The upward-sloping fitted line in figure 2b reveals that land values in larger cities are much higher centrally than values 10 miles away. For the smallest cities, the gradient is offen nearly flat. In large cities, the ratio is much larger but highly variable, even after shrinkage. Together, these two patterns lead to the weaker, but still positive, correlation between city size and average urban land values in figure 2c. These empirical results are generally supportive of a monocentric city with convex rent gradients. Theoretically, these gradients steepen toward the center as firms and households sort according to how their bid per acre varies with distance. Furthermore, agents substitute away from using land in consumption and production as it rises in price.[7]

While our main interest is estimating land values and their cross-sectional differences by MSA, it is worth briefly describing the estimated coefficients on the model covariates, presented in appendix table A1. The most important predictor of log value per acre is log lot size, which enters the regression model as a cubic polynomial. The estimates imply that price per acre is declining in lot size over the size range. This is a standard result called the plattage effect, described by Colwell and Sirmans (1993) in this REVIEW as "a well-known empirical regularity." It is often ascribed to costs of subdividing land parcels, arising from both infrastructure requirements, zoning laws, and bureaucratic costs.[8]

Most of the planned use regressors have statistically and economically significant associations with land values. Retail, apartment, mixed-use, and medical proposed uses have substantially higher values, while commercial, industrial, and multifamily uses have lower values. Lots with no planned use or a planned use of "hold for development" or "hold for investment" also have lower values. Not surprisingly, within-metro land values rise with coastal proximity.

[7] Combes, Duranton, and Gobillon (2016) also find that land-rent gradients are steeper in large French cities than in small ones.

[8] With such costs, large lots may contain more land than is optimal for their intended use. For instance, a lot may have more land than is needed to build an apartment building but cannot be subdivided into two lots on which to build two apartment buildings. In that case, the price per acre of the large lot will be lower than if it contained the optimal amount of land for its intended use.

We have also computed our land value index using total parcel prices on the left-hand-side variable in order to circumvent possible problems with division bias. If lot size is measured with error, then the coefficient estimates are subject to biases. In our log price per acre specification, classical measurement error in log size biases the first coefficient toward -1. To check on the robustness of our fit, we recompute the land value index based on the log of total prices instead. The fitted land values are virtually identical, and the correlation between the two land indices is essentially 1. Essentially all that changes is the nature of the shrinkage estimation.

B. A Cross-Metropolitan Land Value Index

Table 2 presents urban land value estimates for selected metro areas.[9] The first two columns show the name of each MSA and its rank out of 324 according to the estimated average land value in our preferred model specification in column 7 of table 1. Next are the urban (not total) areas of each metro and the number of observed land sales. The fifth column presents average values from the naive model. Column 6 reports estimated central land values[10] using the preferred model, and column 7 presents estimated average values across the urban area. Column 8 reports the estimated ratio of central values to those 10 miles away. The last column provides the total value of urban land by metro, which is totaled at the bottom of the table.

The numbers in columns 5 and 7 contrast the role of the model-based estimator over the naive one. While the two are positively correlated with a coefficient of 0.86, the standard deviation of the naive estimates is 3.2 times higher than that of the model-based estimates. For instance, New York has the highest naively estimated value per acre, \$26 million. Pittsfield, Massachusetts, has the lowest naively estimated values, \$17,000. In general, MSAs with high naively estimated values benefit from favorable covariates, such as small lot sizes.

Overall, the estimates cover 76,581 square miles of urban land. The total estimated value of this land is \$25,025 billion on average over the sample period. The average value of urban land was \$511,000 per acre, with an unweighted standard deviation of \$519,000 across metro areas. This average implies a cost of roughly \$100,000 for a typical fifth-acre residential lot, or \$2,000 for a typical parking spot.

The highest central land values are found in New York, at a whopping \$123 million per acre. The remaining top five are Chicago, Washington, DC, San Francisco, and Los Angeles–Long Beach, with values between \$17 million and \$38 million. With the exception of tightly regulated Washington, all of these central areas are known for their towering skylines—see Ahlfeldt and McMillen (forthcoming) for an application that ties land values to building heights in Chicago.

The New York PMSA has the highest average values as well, \$5.3 million per acre, even after averaging in several counties in addition to New York County (Manhattan). The next three highest averages are found in quality locations with smaller land areas. For instance, Jersey City, a valuable strip of 47 square miles with great views of Manhattan, is second, with an average value of \$3.3 million per acre. Honolulu, which takes third place by a hair, is loaded with scenic views, miles of coastline, and a desirable climate. San Francisco, which completes the almost three-way tie for second, is famous for similar natural amenities, as well as a booming business environment. In fifth place, Los Angeles–Long Beach has average values of \$2.7 million per acre over its

[9] Estimates for all MSAs in the sample are available in table A.3 of the online appendix.

[10] We take estimated values ½ mile from downtown as our estimate of city center land values.

TABLE 2.—SELECTED METROPOLITAN LAND VALUE INDICES, 2005–2010

				Land Values ($000s/Acre)				
Rank (1)	Metropolitan Area Name (2)	Total Urban Area (square miles) (3)	Number of Land Sales (4)	Naive Model (5)	Central (6)	Urban Average (7)	Ratio of Central to 10-Mile Values (8)	Total Urban Land Value ($ billions) (9)
1	New York, NY	749	1,603	26,139	123,335	5,264	22.3	2,524.4
2	Jersey City, NJ	47	43	7,667	9,554	3,305	8.8	98.8
3	Honolulu, HI	198	56	4,357	16,256	3,290	7.0	416.3
4	San Francisco, CA	300	152	8,722	25,446	3,239	9.3	622.8
5	Los Angeles–Long Beach, CA	1,359	1,760	3,709	16,801	2,675	5.5	2,326.8
6	Orange County, CA	494	233	3,163	3,208	2,595	1.3	820.5
7	San Jose, CA	305	217	2,580	3,552	2,347	1.6	458.3
8	Miami, FL	372	1,233	3,052	4,478	1,794	3.2	427.5
9	Stamford-Norwalk, CT	179	19	2,753	2,740	1,505	3.2	172.4
10	Bergen-Passaic, NJ	316	79	1,957	4,145	1,423	3.7	287.7
16	Washington, DC-MD-VA-WV	1,458	1,840	3,548	36,913	1,214	32.6	1,133.0
22	Las Vegas, NV-AZ	317	2,553	1,193	1,841	849	2.4	172.4
26	Chicago, IL	2,035	3,511	1,455	37,632	663	35.1	863.3
27	Boston, MA-NH	1,295	122	1,243	8,457	600	9.8	497.5
32	Denver, CO	536	2,015	828	7,586	539	18.6	185.1
52	Phoenix-Mesa, AZ	897	5,946	370	3,529	452	8.4	259.4
99	Dallas, TX	1,057	811	454	2,774	305	10.1	206.4
118	Houston, TX	1,341	1,143	423	2,813	272	9.4	233.1
120	Detroit, MI	1,426	679	456	2,321	270	6.6	246.6
130	Atlanta, GA	2,105	5,229	402	1,750	251	5.5	338.6
227	Pittsburgh, PA	1,003	240	433	1,772	156	10.6	100.0
322	Glens Falls, NY	33	21	46	65	45	2.6	0.9
323	Jackson, MI	57	8	49	74	38	3.0	1.4
324	Jamestown, NY	46	10	43	63	30	2.1	0.9
	Total United States	76,581	68,756	–	–	–	–	25,024.8
	Simple average, United States	235	212	591	1,672	344	3.7	76.8
	Simple SD across metros	304	592	1,660	7,472	519	3.6	226.6
	Weighted average, United States	–	739	1,052	5,068	511	6.5	244
	Weighted SD across metros	–	1,214	2,701	13,850	715	7.2	430.9

MSAs are ranked by average urban land values. Land value data from CoStar COMPS database for years 2005 to 2010. The naive model is a simple average of observed prices per acre. The estimator allows land values to depend on quartic polynomial in log distance from city center plus 1 mile, with random coefficients. City center land values are for ½ mile from downtown, and mile 10 land values are for 10 miles from downtown. Weighted statistics for United States are weighted by total metropolitan urban area. Standard deviations are unweighted. See online appendix table A2 for complete list of MSAs. Averages and standard deviations for the United States do not include MSAs for which there were no observed land sales.

extended area of 1,359 square miles. Note that L.A. is the most populace PMSA in our sample, but is only sixth in total land area.

The top ten cities in terms of average values are all on or near saltwater coasts. Average land values are more moderate in the Midwest and South: Chicago has an average value of $663,000 per acre, and Pittsburgh has an average of $156,000. Dallas, Houston, and Atlanta have averages values roughly in the $250,000 to $300,000 per acre range. The lowest values, at less than $50,000 per acre, are found in small cities of Glens Falls, New York; Jackson, MI; and Jamestown, New York, at less than $50,000 per acre.

Although the estimated rank correlation between central and average land values is 0.85, the ratio of central values to those 10 miles away varies considerably. The weighted (unweighted) average is 3.7 (6.5), with a standard deviation of 3.6 (7.2). Chicago, with its circumscribed Loop District, has the highest ratio, 35.1, followed by Washington, DC, with its political hub, at 32.6. The tenth percentile ratio of central to 10 miles distant values is 1.6. San Jose, California, and Orange County, California, are the most valuable cities beneath that threshold, reflecting their decentralized urban structures.

The New York PMSA has the greatest total land value of any metro, at roughly $2.5 trillion.[11] The Los Angeles–Long Beach PMSA is not far behind, with a total value of $2.3 trillion. When cities are aggregated to the CMSA level, the top five for total urban land values are New York, Los Angeles, San Francisco, Washington, DC, and Chicago, which together account for 48% of the value of all urban land in the United States.

C. *Comparing Transaction- and Residual-Based Estimates*

A common approach to measure land values is to treat them as the residual difference between a property's entire value and the estimated value of its structure.[12] A caveat of this method is that it equates the market value of a structure with its replacement value, neglecting adjustment costs in building and irreversibilities in investment (Glaeser &

[11] Barr, Smith, and Kulkarni (2016) estimate a geometric average value of $991 billion for the island of Manhattan alone (less than 23 square miles) during that time. Therefore, we consider our estimates of New York land values, while high in absolute terms, to be within reason.

[12] Case (2007) explains how to use FOF data to impute land values in this way, using the replacement cost of housing structures.

Gyourko, 2005). When the market value of structures falls below replacement costs, the residual method assigns the entire decrease to land values. The residual method can even infer negative value to land, as Davis and Heathcote (2007) do for residential land in 1940. Larson (2015) show that the Flow of Funds approach implied the value of land in the corporate business sector in 2009 was worth negative $178 billion (Bureau of Economic Analysis, 2013). Yet it seems unlikely that there was no "buyer" in 2009 who would have been willing accept $178 billion to take a long position on all the corporate land in the United States.

Davis and Palumbo (2008; henceforth, DP) use the residual method to estimate an index of land values across 46 metros. Despite the differences in measurement and intended coverage, we attempt to compare our index to theirs.[13] To compare acres and lots, we estimate average residential lot acreage by metro and divide the DP numbers by this acreage. To aggregate the DP values, we multiply their estimated value per lot by the number of housing units in urbanized Census block groups in the year 2000, counting rental units as having half the land as an owned unit, which roughly reflects national averages. This aggregation method avoids estimating acreages but misses nonresidential land.[14] Online appendix table A.3 contains the estimates for the 45 MSAs in both samples, which are plotted in figure 3.

Our transactions-based estimates imply higher land values than the residual-based estimates: $722,000 versus $392,000 per acre. Across metros, the correlation coefficient between the two is 0.72. The aggregated DP and transaction numbers are more strongly correlated, with a coefficient of 0.95. Figure 3a contrasts the average values per acre, and figure 3b contrasts the aggregate land values for each city. Recall these are for all urban land in our transaction index and for residential land only in the DP index. Our transaction index is higher than the DP index for nearly every city.

Looking at individual cities, both indices imply average land values over $3 million per acre for San Francisco and values near $60,000 for Charlotte. But for New York our transaction index implies urban land values of $5.3 million per acre versus $835,000 for the DP estimates. For Oklahoma City, our index is $161,000 per acre, while the DP index implies $24,000 per acre. These differences may arise from the differences in the types of land considered: our index includes highly valuable central and commercial land. Nevertheless, our data sources and estimation technique seem to play large roles. Furthermore, the value of transactional land should reflect available building opportunities, good or bad, while built-on land reflects the structure that is permitted de facto.[15]

Over time, our transaction index implies smaller price movements than the DP index within cities over the boom-and-bust cycle in our data. This is seen in figure 3c, which plots the estimated difference between the minimum and maximum annual estimated average land values within each city, expressed as a percentage of the maximum value. The average coefficient of variation of land values within the 45 cities according to our index was 0.24 versus 0.44 in the DP estimates. The greater volatility of the residual method is also seen in the time series for aggregate U.S. land values, which we consider below.

V. Aggregate Urban Land Values over Time

In this section, we sum our urban land values across metros to calculate annual aggregate urban land values for the United States.[16] Table 3 presents these totals.

Over our sample period, average values peaked in 2006 at $624,000 per acre, an increase of 8% from 2005. Average values then fell to near their 2005 levels in 2007, before declining precipitously. By 2009, the average value was roughly $373,000 per acre, 65% of its 2005 level. The ratio of aggregate urban land values to GDP declined considerably as well. The ratio was 2.1 to 2.2 in 2005 and 2006 before declining to reach a value 1.28 by 2010.

For comparison, we construct a series for aggregate U.S. land values using the residual method based on FOF data (now the Financial Accounts of the United States). We sum the total value of real estate at market value held by nonfinancial noncorporate businesses, nonfinancial corporate businesses, and households and nonprofit organizations to arrive at the total market value of privately held real estate. We then subtract the current-cost net stock of private structures to arrive at a residual-based value for land. In 2006, the estimated value of real estate was $43.3 trillion, while structures were valued at $26.3 trillion, implying that the total value of land was $16.9 trillion. Our transactions-based estimate, in contrast, is $30.4 trillion, nearly 80% higher, signifying that urban land is an even more important asset in the U.S. economy.

In addition to the methodological differences, the totals may differ because they cover different land. Our estimates

[13] Their index is purely residential, for owner-occupiers only, and is estimated by lot. Our transaction index is for all urban land (including commercial and industrial), for owners and renters, and is estimated by acre.

[14] We divide by average lot size, since DP report an arithmetic average of land value. This may introduce significant measurement error in some numbers. Using medians or geometric averages produces substantially higher average values per acre. The DP (2008) index is quarterly; we take geometric averages to arrive at annual and whole-sample values. Matching our MSAs to their cities is typically straightforward using the name of the principal city. We do not match their estimates for Santa Ana to the Orange County, California, MSA, because we lack lot size information.

[15] As Davis and Heathcote (2007) note, the residual method attaches "the label 'land' to anything that makes a house worth more than the cost of putting up a new structure of similar size and quality on a vacant lot." Thus, the residual method will attribute higher costs stemming from inefficiencies in factor usage (e.g., geographic and regulatory constraints that hinder building) to higher land values. In a follow-up paper, Albouy and Ehrlich (2016) use differences in the value of housing prices from land and structure costs to measure the costs imposed by such constraints. See Glaeser and Gyourko (2017) for a related but more reduced-form approach that assumes land is a fixed fraction of housing costs.

[16] Our sample includes observations from 324 of the 331 MSAs and PMSAs in the 1999 OMB definitions. The combined imputed land value for the seven metros with no data is $61 billion, less than 0.25% of our aggregate number.

FIGURE 3.—COMPARISON OF TRANSACTIONS-BASED INDEX TO RESIDUAL-BASED INDEX

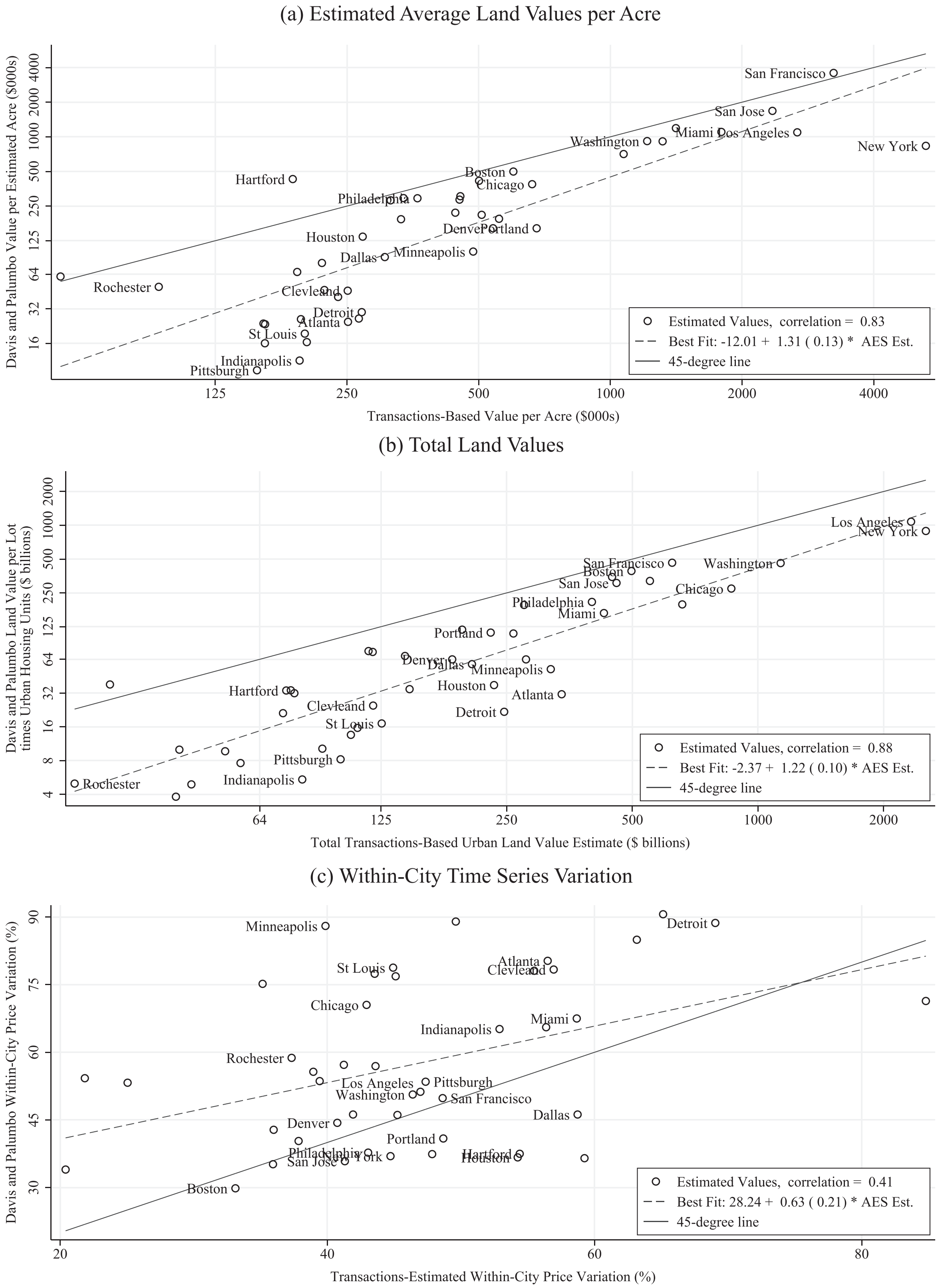


TABLE 3.—URBAN LAND VALUES IN THE UNITED STATES, 2005–2010

Year	Average Urban Land Value per Acre($)	Total Urban Land Value ($ billions)	Average Urban Land Value per Acre (Index, 2005 = 100)	Nominal GDP ($ billions)	Ratio of Total Urban Land Value to GDP	S&P CoreLogic Case-Shiller U.S. National HPI (normalized to 2005 = 100)	Total Urban Land Value—Residual Method ($ billion)
2005	$577,336	$28,117	100.0	$13,094	2.15	100.0	$16,758
2006	623,950	30,387	108.1	13,856	2.19	106.8	16,931
2007	584,682	28,475	101.3	14,478	1.97	104.8	16,001
2008	513,413	25,004	88.9	14,719	1.70	95.5	9,569
2009	372,819	18,157	64.6	14,419	1.26	86.5	5,767
2010	392,683	19,124	68.0	14,964	1.28	84.2	6,234

Land-value data from CoStar COMPS database for 2005 to 2010. Residual method calculates total real estate holdings at market value of nonfinancial businesses, households, and nonprofit organizations from Financial Accounts of the United States (formerly known as the flow of funds) and subtracts current-cost net stock of private structures from National Income and Product Accounts.

are based on total metro urban areas, including public lands for roads, parks, and civic buildings. Assuming that the public owns urban land worth 40% of the total value, only $18.2 trillion of land would be owned privately, which is much closer to the FOF numbers. On the other hand, the FOF numbers include land outside metro-urban areas, which we exclude.

Land values calculated from the FOF fell even more dramatically than our series, down to only $5.8 trillion in 2009, as opposed to $14.4 trillion. The peak-to-trough decline in the transactions-based index was 40%, substantially less than the 66% decline in the FOF.

Finally, we consider how land values compare with housing prices. The final column of table 3 reports the S&P CoreLogic Case-Shiller U.S. National House Price Index, normalized to have value 100 in 2005. Overall, land values appear to have led house prices slightly and were substantially more volatile than house prices over the sample period. This result is consistent with the Bostic, Longhofer, and Redfearn (2007) land leverage hypothesis that housing should have less volatile values than land.

VI. Conclusion

Our analysis combines insights from the economic monocentric city model with empirical Bayesian methods to produce novel and plausible estimates of land values, even in metros with relatively thin data. These methods might easily be applied to estimate other citywide measures, such as wages or property prices. Relative to residual approaches, our method suggests that urban land values may be higher, less volatile, and less likely to be negative. Furthermore, the model sheds light on the enormous differences in land values both across and within cities, with high central values providing indirect support for monocentric cities, albeit with heterogeneous value gradients.

We hope that the measures we provide may form the basis of reliable estimates of aggregate land wealth. With additional data, future modeling could be enriched to incorporate greater spatial structure and modifications for observed land uses. The cross-sectional index should also prove useful to researchers examining differences in amenities and real-estate costs across metro areas.

REFERENCES

Ahlfeldt, Gabriel M., and Dan McMillen, "Tall Buildings and Land Values: Height and Construction Cost Elasticities in Chicago, 1870–2010," this REVIEW (forthcoming).

Albouy, David, "What Are Cities Worth? Land Rents, Local Productivity, and the Total Value of Amenities," this REVIEW 98 (2016), 477–487.

Albouy, David, and Gabriel Ehrlich, "Housing Productivity and the Social Cost of Land Use Restrictions," NBER working paper 18110 (2016).

Alonso, William, *Location and Land Use: Toward a General Theory of Land Rent* (Cambridge, MA: Harvard University Press, 1964).

Barr, Jason, Fred Smith, and Sayali Kulkarni, What's Manhattan Worth? A Land Values Index from 1950 to 2014 (New Brunswick, NJ: Rutgers University Press, 2016).

Bostic, Raphael W., Stanley D. Longhofer, and Christian L. Redfearn, "Land Leverage: Decomposing Home Price Dynamics," *Real Estate Economics* 35 (2007), 183–208.

Brueckner, Jan K., "Property Value Maximization and Public-Sector Efficiency," *Journal of Urban Economics* 14 (1983), 1–15.

Bureau of Economic Analysis, "Relation of BEA's Current-Cost Net Stock of Private Structures to the Corresponding Items in the Federal Reserve Board's Financial Accounts of the United States," technical report (2013). https://www.bea.gov/national/pdf/st13.pdf.

Case, Karl, "The Value of Land in the United States: 1975–2005," in G. K. Ingram and Y.-H. Hong, eds., *Land Policies and Their Outcomes* (Cambridge, MA: Lincoln Institute of Land Policy, 2007).

Colwell, Peter F., and Henry J. Munneke, "The Structure of Urban Land Prices," *Journal of Urban Economics* 41 (1997), 321–336.

Colwell, Peter F., and Clemon F. Sirmans, "A Comment on Zoning, Returns to Scale, and the Value of Undeveloped Land," this REVIEW 75 (1993), 783–786.

Combes, Pierre-Philippe, Gilles Duranton, and Duranton Gobillon, "The Costs of Agglomeration: House and Land Prices in French Cities," University of Pennsylvania mimeograph (2016).

Davis, Morris A., and Jonathan Heathcote, "The Price and Quantity of Residential Land in the United States," *Journal of Monetary Economics* 54 (2007), 2595–2620.

Davis, Morris A., and Michael G. Palumbo, "The Price of Residential Land in Large US Cities," *Journal of Urban Economics* 63 (2008), 352–384.

Fisher, Jeff, David Geltner, and Henry Pollakowski, "A Quarterly Transactions-Based Index of Institutional Real Estate Investment Performance and Movements in Supply and Demand," *Journal of Real Estate Finance and Economics* 34 (2007), 5–33.

George, Henry, *Progress and Poverty: An Inquiry into the Cause of Industrial Depressions, and of Increase of Want with Increase of Wealth, the Remedy* (London: W. Reeves, 1884).

Glaeser, Edward L., and Joseph Gyourko, "Urban Decline and Durable Housing," *Journal of Political Economy* 113 (2005), 345–375.

——— "The Economic Implications of Housing Supply," *Zell/Lurie* working paper 802 (2017).

Harville, D. A., "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," *Journal of the American Statistical Association* 72 (1977), 320–338.

Haughwout, Andrew, James Orr, and David Bedoll, "The Price of Land in the New York Metropolitan Area," *Federal Reserve Bank of New York Current Issues in Economics and Finance* (2008).

Kok, Nils, Paavo Monkkonen, and John M. Quigley, "Land Use Regulations and the Value of Land and Housing: An Intra-Metropolitan Analysis," *Journal of Urban Economics* 81 (2014), 136–148.

Laird, Nan M., and Thomas A. Louis, "Empirical Bayes Ranking Methods," *Journal of Education Statistics* 14 (1989), 29–46.

Larson, William, "New Estimates of Value of Land of the United States," Bureau of Economic Analysis technical report (2015).

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

Munneke, Henry, and Slade Barrett, "An Empirical Study of Sample-Selection Bias in Indices of Commercial Real Estate," *Journal of Real Estate Finance and Economics* 21 (2000), 45–64.

——— "A Metropolitan Transaction-Based Commercial Price Index: A Time-Varying Parameter Approach," *Real Estate Economics* 29 (2001), 55–84.

Muth, Richard F., *Cities and Housing: The Spatial Pattern of Urban Residential Land Use* (Chicago: University of Chicago Press, 1969).

Nichols, Joseph B., Stephen D. Oliner, and Michael R. Mulhall, "Swings in Commercial and Residential Land Prices in the United States," *Journal of Urban Economics* 73 (2013), 57–76.

Ozimek, Adam, and Daniel Miles, "Stata Utilities for Geocoding and Generating Travel Time and Travel Distance Information," *Stata Journal* 11 (2011), 106.

Ricardo, David, *On the Principles of Political Economy, and Taxation* (London: John Murray, 1821).

Roback, Jennifer, "Wages, Rents, and the Quality of Life," *Journal of Political Economy* 90 (1982), 1257–1258.

Zeger, S. L., and M. R. Karim, "Generalized Linear Models with Random Effects: A Gibbs Sampling Approach," *Journal of the American Statistical Association* 86 (1991), 79–86.

APPENDIX A

Additional Data Notes

When a CMSA contains multiple PMSAs, we treat each PMSA as its own MSA for purposes of estimation and reporting. For instance, we treat the Washington, DC-MD-VA-WV, and Baltimore, MD, PMSAs as separate MSAs, although they are both parts of the Washington-Baltimore DC-MD-VA-WV CMSA. For New York City, we use the Empire State Building as the city center rather than city hall, following Haughwout, Orr, and Bedoll (2008). We treat each named city in an MSA with a hyphenated city name as having its own city center. For instance, we treat Minneapolis–St. Paul, MN-WI, as containing two distinct cities, Minneapolis, MN, and St. Paul, MN. However, we treat such cities as belonging to one MSA for purposes of aggregating and reporting.

In the CoStar data, we consider twelve of the most common proposed uses, which are neither mutually exclusive nor collectively exhaustive. We consider an observation to feature a structure when the transaction record includes the fields for "Bldg Type," "Year Built," "Age," or the phrase "Business Value Included" in the field "Sale Conditions." We geocoded the lot sales using the Stata "geocode" module of Ozimek and Miles (2011). In addition to the exclusions discussed in the main text, we also exclude outlier observations with a listed price of less than \$100 per acre or a lot size over 5,000 acres, or farther than 60 miles away from the city center. We also exclude lots we could not geocode successfully.

Median lot size is 3.5 acres versus a mean of 26 acres. Land sales occur more frequently in the beginning of our sample period, with 21.7% of our sample from 2005 and 11.4% from 2010. Residential uses are common but by no means predominant in the sample—17.6% of properties have a proposed use of single-family, multifamily, or apartments—23.4% is being held for development or investment, and 16% of the sample had no listed proposed use.

APPENDIX B

Computation

1. Estimation of Land Value Gradients: α_{jt} *and* $\boldsymbol{\delta}_j$

For notational convenience we rewrite the model in equation (1) as

$$\ln r_{ijt} = Z'_{ijt}\gamma_j + X_{ijt}\beta + e_{ijt}, \quad e_{ijt} \sim N(0, \sigma_e^2), \tag{B1}$$

where $Z'_{ijt} = \left[1, D_{ij}, D_{ij}^2, D_{ij}^3, D_{ij}^4, 1_{ijt}^{2006}, 1_{ijt}^{2007}, 1_{ijt}^{2008}, 1_{ijt}^{2009}, 1_{ijt}^{2010}\right]$, with $D_{ij} = D(\mathbf{z_{ij}}, \mathbf{z_j^c})$, and where 1_{ijt}^s is an indicator variable that takes value 1 if $s = t$ and 0 otherwise. The parameter vector γ_j collects city- and time-specific parameters with a multivariate normal prior distribution,

$$\begin{aligned}\gamma_j &= \left[\alpha_j, \delta_{j1}, \delta_{j2}, \delta_{j3}, \delta_{j4}, \alpha_{j,2006}, \alpha_{j,2007}, \alpha_{j,2008}, \alpha_{j,2009}, \alpha_{j,2010}\right]' \\ &\sim N(m_{\gamma,j}, V_{\gamma,0}),\end{aligned} \tag{B2}$$

where

$$m_{\gamma,j} = \begin{pmatrix} a_0 + a_1 \ln A_j \\ \mathbf{b_0} + \mathbf{b_1} \ln A_j \\ \boldsymbol{\tau} \end{pmatrix} \quad \text{and} \quad V_{\gamma,0} = \begin{pmatrix} \Sigma_{\alpha\alpha} & \boldsymbol{\Sigma_{\alpha\delta}} & \mathbf{0_{(1\times5)}} \\ \boldsymbol{\Sigma_{\delta\alpha}} & \boldsymbol{\Sigma_{\delta\delta}} & \mathbf{0_{(4\times5)}} \\ \mathbf{0_{(5\times1)}} & \mathbf{0_{(5\times4)}} & \boldsymbol{\Sigma_{\tau\tau}} \end{pmatrix} \tag{B3}$$

with $\boldsymbol{\tau} = [\tau_{2006}, \tau_{2007}, \tau_{2008}, \tau_{2009}, \tau_{2010}]'$ and $\boldsymbol{\Sigma_{\tau\tau}} = diag([\sigma_{2005}^2, \sigma_{2006}^2, \sigma_{2007}^2, \sigma_{2008}^2, \sigma_{2009}^2, \sigma_{2010}^2]')$. Conditional on fixed and variance parameters ($\theta = [\beta, a_0, a_1, \mathbf{b_0}, \mathbf{b_1}, \boldsymbol{\tau}, \sigma_e^2, \Sigma_{\alpha\alpha}, \boldsymbol{\Sigma_{\delta\alpha}}, \boldsymbol{\Sigma_{\delta\delta}}, \boldsymbol{\Sigma_{\tau\tau}}]$) and observed data for city j, the posterior distribution of γ_j follows the multivariate normal distribution

$$\gamma_j|\theta, Data \sim N\left(\widetilde{m}_{\gamma,j}(\theta),\ \widetilde{V}_{\gamma,j}(\theta)\right) \tag{B4}$$

with a posterior mean as the weighted average between the prior mean ($m_{\gamma,0}$) and the fixed-effect estimate $\widehat{\gamma}_j = (Z_j'Z_j)^{-1}\left[Z_j'\left(\ln r_j - X_j\beta\right)\right]$:

$$\begin{aligned}\widetilde{m}_{\gamma,j}(\theta) &= W_j(\theta)m_{\gamma,j} + \left[I - W_j(\theta)\right]\widehat{\gamma}_j(\theta) \\ &\text{where } W_j(\theta) = \left[V_{\gamma,0}^{-1} + \sigma_e^{-2}(Z_j'Z_j)\right]^{-1} V_{\gamma,0}^{-1}.\end{aligned} \tag{B5}$$

Here we write Z_j, $\ln r_j$, and X_j as matrices that stack elements only relevant for the city j. The weighting matrix depends on the number of observations in the city j (n_j), the relative size of the prior variance ($V_{\gamma,\theta}$), and the idiosyncratic error variance (σ_e^2). The posterior variance is

$$\widetilde{V}_{\gamma,j}(\theta) = \left[V_{\gamma,0}^{-1} + \sigma_e^{-2}(Z_j'Z_j)\right]^{-1}. \tag{B6}$$

It is well known that the posterior mean $\widetilde{m}_{\gamma,j}(\theta)$ is the best linear unbiased predictor for γ_j given θ and the observed data. In our application, we do not know θ. Instead of taking a full Bayesian approach and putting a prior on θ, we take the empirical Bayesian approach in which θ is calibrated by maximizing the following marginal likelihood (Laird & Louis, 1989):

$$\widehat{\theta} \in argmax_\theta\ L(data|\theta) = \int p(\ln r|Z, X, \theta, \gamma)d\gamma, \tag{B7}$$

where the γ is integrated out from the likelihood function using an improper prior, $p(\gamma) \propto 1$ (see Harville, 1977). Then we treat $\widehat{\theta}$ as a known and fixed quantity and use the following posterior distribution for the computation of land values and the prediction,

$$\gamma_j|Data \sim N\left(\widetilde{m}_{\gamma,j}(\widehat{\theta}),\ \widetilde{V}_{\gamma,j}(\widehat{\theta})\right). \tag{B8}$$

One of the potential shortcomings of this approach is that it neglects uncertainty coming from the estimation of θ, and the resulting posterior distribution for γ_j underestimates uncertainty. Fortunately, we have a relatively large amount of data about θ (about 67,000 observations in total). Second, the practicality of our shrinkage estimator is evaluated by the out-of-sample forecasting evaluation. However, we note that a full Bayesian approach is possible (Zeger & Karim, 1991) at the cost of even longer computation time. We choose to take the empirical Bayes approach because of the out-of-sample evaluation of our shrinkage procedure.

2. Point Predictions for Land Values

Once we obtain the posterior distribution of γ_j, we can generate land value predictions. For the cross-validation exercise, we generate and evaluate point predictions for the log-price of the land parcels in the city j at time t with characteristic X_{ijt}^* and Z_{ijt}^* as the mean of the posterior predictive distribution. In the standard case when we observe at least some data in city j, the point prediction for the value of a land parcel is

$$\widehat{\ln r_{ijt}} = \int \ln r_{ijt} p(\ln r_{ijt}|data, X^*_{ijt}, Z^*_{ijt}) d\ln r_{ijt}$$
$$= \int\int \ln r_{ijt} p(\ln r_{ijt}|data, X^*_{ijt}, Z^*_{ijt}, \gamma_j) p(\gamma_j|data) d\gamma_j d\ln r_{ijt}$$
$$= Z^{*'}_{ijt} \widetilde{m}_{\gamma,j}(\widehat{\theta}) + X^{*'}_{ijt}\widehat{\beta}, \tag{B9}$$

where $p(\ln r_{ijt}|data, X^*_{ijt}, Z^*_{ijt}, \gamma_j)$ is defined by equation (B1) at $\widehat{\theta}$, and $p(\gamma_j|data)$ is the multivariate probability density function with mean $\widetilde{m}_{\gamma,j}\widehat{\theta}$ and variance-covariance matrix $\widetilde{V}_{\gamma,j}(\widehat{\theta})$.

We can also generate predictions for the land in cities where we do not have observed transaction prices. This is based on our "metacity" for a city with area A_j, using the prior with estimated hyperparameters, $\widehat{\theta}$. In this case, our prediction is just

$$\widehat{\ln r_{ijt}} = Z^{*'}_{ijt} m_{\gamma,j}(\widehat{\theta}) + X^{*'}_{ijt}\widehat{\beta}. \tag{B10}$$

3. Computation of Land Values

For each census tract l in city j in year t, we calculate the predicted land value r_{ljt} at the tract centroid and assign that average value to the entire tract.

$$R_{jt} = \sum_{l=1}^{L} \widehat{r}_{ljt} A_l, \tag{B11}$$

where A_l is the tract area we use the mean of the predictive distribution for r_{ljt} as the predicted land value. That is,

$$\widehat{r}_{ljt} = \int \exp(r_{ljt}) p(r_{ljt}|data, X^{**}_{ljt}, Z^{**}_{ljt}) dr_{ljt}$$
$$= \int\int \exp(r_{ljt}) p(r_{ljt}|data, X^{**}_{ljt}, Z^{**}_{ljt}, \gamma_j) p(\gamma_j|data) dr_{ljt} d\gamma_j$$
$$= \exp\left(Z^{**'}_{ljt} m_{\gamma,j}(\widehat{\theta}) + X^{**'}_{ljt}\widehat{\beta} + \widehat{\sigma}^2_e/2 + Z^{**'}_{ljt} V_{\gamma,j}(\widehat{\theta}) Z^{**'}_{ljt}/2\right), \tag{B12}$$

where the last two terms are due to the log-normal correction. We can also estimate values for cities with no observed land sales using only the prior.

Since our land data are incomplete, some land characteristics such as lot sizes and planned uses (a subvector of X^{**}_{ljt}) are unknown at the tract centroid. Therefore, we predict these characteristics based on what we do know of the land, namely, its location. To do this, we decompose the predicted land value in the following manner:

$$\widehat{r}_{ljt} = \int \exp(r_{ljt}) p(r_{ljt}|data, X^{**}_{ljt}, Z^{**}_{ljt}) dr_{ljt}$$
$$= \int\int \exp(r_{ljt}) p(r_{ljt}|data, X^{**}_{ljt}, Z^{**}_{ljt}, \gamma_j) \times p(\gamma_j, X^{**}_{ljt}|data, Z^{**}_{ljt}) dr_{ljt} d\gamma_j dX^{**}_{ljt}$$
$$= \int\int \exp(r_{ljt}) p(r_{ljt}|data, X^{**}_{ljt}, Z^{**}_{ljt}, \gamma_j) p(\gamma_j|data) \times p(X^{**}_{ljt}|data, Z^{**}_{ljt}) dr_{ljt} d\gamma_j dX^{**}_{ljt}$$
$$= \exp\left(Z^{**'}_{ljt} m_{\gamma,j}(\widehat{\theta}) + \widehat{\sigma}^2_e/2 + Z^{**'}_{ljt} V_{\gamma,j}(\widehat{\theta}) Z^{**'}_{ljt}/2\right) \times \int \exp\left(X^{**'}_{ljt}\widehat{\beta}\right) p(X^{**}_{ljt}|data, Z^{**}_{ljt}) dX^{**}_{ljt}, \tag{B13}$$

where the uncertainty about the unobserved land characteristic at the tract centroid is captured by the predictive distribution function of X^{**}_{ljt} in the last integral. We construct a model for each unobserved element in X^{**}_{ljt} using observed characteristics of the tract l in city j. More specifically, the sth element in X^{**}_{ljt} is modeled as

$$X^{**}_{s,ljt} = \alpha^x_{s,j} + \delta^x_{s,j} D_{lj} + \gamma_s C_{lj} + e_{s,ljt}, \quad e_{s,ljt} \sim i.i.d.N(0, \sigma^2_s), \tag{B14}$$

where D_{lj} is the distance metric based on the distance between the tract centroid and the city center, and C_{lj} is log distance to coast from the tract centroid. Then we replace unobserved elements in X^{**}_{ljt} in equation B13 with their predicted values.

This technique is based on a similar but simpler version of the hierarchical model used for land prices. The intercept and coefficient on the distance to the city center are allowed to vary across MSAs, but using only an affine function as opposed to a quartic polynomial. The coefficient on the distance to the coast is fixed.

Because these coefficients are not known, we estimate them using the observed transaction data with the similar prior specification and assumption employed for the estimation of model for the land price. More specifically, the prior distribution for city-specific parameters $\alpha^x_{s,j}$ and $\delta^x_{s,j}$ follows a multivariate normal distribution. The mean vector is an affine function of each city's log urban area, and the variance-covariance matrix is allowed to have nonzero off-diagonal elements. We impose similar exogeneity assumptions for $\alpha^x_{s,j}$, $\delta^x_{s,j}$, and $e_{s,ljt}$. Finally, we assume that each element in X^{**}_{ljt} is correlated only through observed tract characteristics D_{lj} and C_{lj} (equation B14). Because estimation and prediction for the land price and land characteristics are performed conditional on distance variables, we do not assume any specific distributional form for observed distance variables D_{ij} and C_{ij}. However, we assume that the marginal density of D_{ij} puts nonzero positive value on the entire MSA area. This last assumption implies that if we do not have a transaction observation at a specific census tract, this missingness is completely random, and we would eventually collect observations from this tract as the sample size goes to infinity.

4. Cross-Validation

Cross-validation techniques help to determine the most appropriate econometric specification and evaluate the effectiveness of the shrinkage model. We design a pseudo out-of-sample prediction exercise that quantifies the potential gains or losses from different models. For this exercise, we take cities that have at least fifty observations per year for at least two years. This leads to 58 cities with 55,155 total observations. Then, for each city j,

1. Randomly choose n_{hold} observations out of n_{jt} observations for each time $t = 2005, 2006, \ldots, 2010$ in city j. We keep those $6 \times n_{hold}$ observations as well as the remaining sample of data from other cities.
2. Estimate each of models using the method described in section 1.
3. Generate predictions for sample held out in step 1 for city j based on the method in section 2.
4. Compute and store the prediction error for this holdout sample. $\{e_{j,r,1}, e_{j,r,2}, \ldots, e_{j,r,(n_{jt}-n_{hold})}\}$ (forecast errors are defined as predicted minus actual).
5. Repeat steps 1 to 4 for $r = 1, 2, \ldots, R$.
6. Repeat steps 1 to 5 for each city $j = 1, \ldots, J$.
7. Compute aggregated out-of-sample prediction evaluation statistics. For example, the MSE for the city j is computed as

$$MSE(j) = \frac{1}{R \times (n_{j,t} - n_{hold})} \sum_{r=1}^{R} \sum_{i=1}^{(n_{jt}-n_{hold})} e^2_{r,j,i}, \tag{B15}$$

where we set $R = 30$. We perform for $n_{hold} = 3$ (small sample size) and $n_{hold} = 30$ (moderate sample size) for each city. About 35% of MSAs in our sample have observations less than or equal to eighteen observations (approximately three per year in our data set), and about 81% of MSAs in our sample have observations less than equal to 180 (which is approximately thirty per year in our data set). We report average $MSE(j)$ over $j = 1, 2, \ldots 58$.

The unshrunken estimates are based on the fixed-effect estimation. The estimator is defined as $\widehat{\gamma}_j = (Z'_j Z_j)^{-1}(Z'_j(\ln r_j - X_j\beta)$ and used in equation B5.

TABLE APPENDIX

TABLE A1.—ESTIMATED COEFFICIENTS ON COVARIATES IN PREFERRED SPECIFICATION

Covariate	Estimated Coefficient	Standard Error	t-Statistic	p-Value
Log Lot Size	−0.543	0.0037	−146.134	0.000
(Log Lot Size Squared)/100	−3.053	0.1592	−19.176	0.000
(Log Lot Size Cubed)/1000	3.601	0.2498	14.415	0.000
Log Distance to Coast	−0.052	0.0043	−12.196	0.000
Planned use				
None Listed	−0.182	0.0112	−16.193	0.000
Commercial	−0.380	0.0599	−6.354	0.000
Industrial	−0.346	0.0141	−24.578	0.000
Retail	0.255	0.0134	18.963	0.000
Single Family	0.003	0.0133	0.202	0.840
Multifamily	−0.139	0.0198	−7.055	0.000
Office	0.046	0.0148	3.129	0.002
Apartment	0.288	0.0196	14.713	0.000
Hold for Development	−0.073	0.0118	−6.171	0.000
Hold for Investment	−0.283	0.0195	−14.523	0.000
Mixed Use	0.250	0.0265	9.438	0.000
Medical	0.171	0.0355	4.810	0.000
Parking	0.076	0.0373	2.044	0.041

This table reports the coefficients on the covariates from the preferred specification in table 1 from the main body of the text, which applies shrinkage to a model with a quartic polynomial in log distance to the city center plus 1 mile.