For What It's Worth: Measuring Land Value in the Era of Big Data and Machine Learning

Scott Wentland[a+]
U.S. Bureau of Economic Analysis

Jeremy Moulton
University of North Carolina – Chapel Hill

Gary Cornwall[+]
U.S. Bureau of Economic Analysis

December 19, 2021

Abstract

We adapt a machine learning method to provide new estimates for land valuation in the United States, pairing this approach with "big data" from Zillow. Because this data includes detailed information from hundreds of millions of property transactions covering much of the US, the heterogeneous nature of this data serves as fertile ground for highlighting some of the practical limitations of linear hedonic regression techniques for land valuation, a common method for mass appraisal of land. We first construct traditional hedonic estimates of land value at the parcel-level for most of the US as a baseline, focusing on single-family residential properties in our initial analysis. We then modify the hedonic approach by using a machine learning method, gradient boosting trees, for comparison. The results demonstrate how a machine learning approach can more effectively address issues of sparse data with spatial controls or thin cells at fine levels of geography (like census tracts or block groups). Our initial estimates also show that the machine learning method offers a substantial improvement in prediction of single-family sale prices (i.e., a 75% reduction in RMSE, on average) and great potential for further applications in constructing aggregate measures of land value beyond the cases we pilot here.

Keywords: land valuation, national accounts, non-financial assets, environmental-economic accounting, hedonic valuation, Big Data, machine learning, gradient boosted trees
JEL Classifications: E01, Q56, Q24, R14, C80, G12

[a] *Contact Author*. Office of the Chief Economist, 4600 Silver Hill Rd, Suitland, MD 20746; scott.wentland@bea.gov. The authorship order was determined to be reverse alphabetical order arbitrarily by a coin toss. Gary lost.

[+]*Disclaimer*: Any views expressed here are those of the authors and not necessarily those of the Bureau of Economic Analysis or the U.S. Department of Commerce. Data provided by Zillow through the Zillow Transaction and Assessment Dataset (ZTRAX). More information on accessing the data can be found at http://www.zillow.com/ztrax. The results and opinions are those of the author(s) and do not reflect the position of Zillow Group.

"…the commodities which compose the whole annual produce of the labour of every country, taken complexly, must resolve itself into the same three parts, and be parceled out among different inhabitants of the country, either as the wages of their labour, the profits of their stock, or the rent of their land…Wages, profit, and rent, are the three original sources of all revenue as well as of all exchangeable value." –Adam Smith (1776, The Wealth of Nations – Book 1, Chapter VI)

"*Buy land – they're not making it anymore*." -Mark Twain[*]

1. For what it's worth: Introduction

Land is, quite literally, a foundational asset for any economy. Economists, extending back to at least Adam Smith (1776), have long understood that for households and firms the value of the land underlying their structure is often a substantial portion of the overall value of their property; and, in some cases (especially with agricultural land) it constitutes nearly the entire asset value or flow of rents. Recent research has estimated that, in aggregate, not only is land a substantial asset in its own right,[1] but the fluctuations in its value can play a pivotal role in the business cycle, as illustrated by the real estate boom and bust that coincided with the Great Recession in 2007-2009. Some literature has pointed out that the infamous housing boom and bust of the 2000s if often mischaracterized, instead suggesting that it would more aptly called a *land* boom and bust (Davis et al 2017; Davis et al 2021), citing evidence that much of the fluctuation in value of residential property can be attributed to the underlying price of land (see also Kuminoff and Pope 2013). Given both its economic significance and the diversity of approaches used in the literature to investigate the value of this asset, the purpose of this paper is to revisit a timeless question using new methods and new data: how much is land worth? More specifically, can modern machine learning (ML) methods using "Big Data" from across the United States deliver significant advantages over prior approaches and provide new insights into this question?

We show that ML methods indeed can; and, by employing this approach using microdata from a national dataset like Zillow's ZTRAX data, we provide a tractable way to construct national and subnational estimates of land value that solves several critical issues that arise with commonly used methods. In addition to several contributions to the academic literature, which we discuss in more detail below, this research marks a step toward filling a critical gap in the national economic

[*] Like many sayings attributed to Mark Twain, this one is likely apocryphal, but variations of this statement have been attributed to him for decades. Variations of this quote have also been attributed to Will Rogers. The original source for this particular version of the quote is unknown, but attributing it to Mark Twain sounds better.

[1] Wentland et al (2020) estimated that the value of private land in the contiguous 48 US states was approximately $25 trillion in 2015. Using a different approach, Larson (2015) estimated land in the US to be worth $23 trillion.

accounts. Despite the significance of land as an asset, there is virtually no available information directly quantifying the aggregate value of land itself in the national accounts (either in the US or the vast majority of countries around the world).[2] This fact might be surprising to classical economists like Adam Smith, who mention land explicitly in his early writings on national output, as well as modern day economists and decision-makers who use aggregate data from the national income and product accounts (NIPA) like gross domestic product (GDP) to understand a wide variety of national economic phenomena. Because land is described as a "non-produced, non-financial asset" (a category Mark Twain might appreciate) on a country's balance sheet in the System of National Accounts (also known as the SNA – the statistical standard governing national income accounting), countries rarely provide direct aggregate estimates of this particular line item on a national scale.[3] Instead, statistics developed in the 20th century have traditionally focused their accounts on goods and services and the produced capital assets related to their production.

In the 21st century, however, there has been broad international interest in expanding the scope of the national accounts to include more non-produced capital or "natural capital" with environmental economic accounts that quantify the value of our natural resources (Boyd et al 2018), including land, along with a greater interest in information on land prices in particular (Coomes et al. 2018). In fact, the UN has recently reported that over 90 countries produce at least one SEEA-based environmental economic account as of 2020.[4] Yet, the US does not currently produce any formal environmental economic accounts. Given that land is an asset at the intersection of the traditional national accounts as described by the SNA and environmental accounts proposed in the System of Environmental-Economic Accounting (SEEA), valuing land presents a logical starting point for expanding the scope of what the national accounts explicitly

[2] This is not to suggest that the value of land is omitted from our national accounts entirely. There are, however, indirect ways land enters into the accounts. For example, in GDP, the value of land (while not separately estimated) is in rent paid for some types of land or as part of an intermediate input/component of final housing services or business spending on fixed assets. The Integrated Macro Accounts, which house the national balance sheet of assets, includes real estate among household nonfinancial assets, but does not separately break out structure versus land value specifically. In the US, these balance sheets generally exclude land from corporate and government assets.

[3] There are some exceptions to this, as a handful of countries provide some information about the value of land in their accounts, including Australia, South Korea, U.K., and Canada (Wentland et al. 2020). But these accounts are either limited in scope or, as of the writing of this paper, they have coarser estimates that do not use "big data" methodological approaches like the one outlined in this paper.

[4] For more information on environmental economic accounts and the corresponding statistical standards countries use to implement them (System of Environmental-Economic Accounting (SEEA) Central Framework and Ecosystem Services Accounts), see https://seea.un.org/content/frequently-asked-questions.

measure in the US.[5] Thus, one of the primary objectives of this paper is to cultivate a new method to decouple land value from structure value, using single family residential (SFR) properties as a pilot case, which could then be readily applied to a national data set that could be used to construct aggregate estimates of land value in a formal account.

To measure the aggregate value of land, we begin small. Conceptually, we are interested in unpacking the structure and land values from an individual property's value, as the most common type of real estate transaction we observe is one that couples the structure with the underlying land. Given that individual property's total value or price is what we most commonly observe,[6] our method here draws on fine-grained data from Zillow's ZTRAX dataset, which contains detailed information on hundreds of millions of property transactions over the last couple decades across the US. It also contains detailed property characteristic information for a large percentage of the universe of properties in the US, including information about the land itself (like its acreage and land use type) along with structure characteristics (like number of bedrooms, bathrooms, rooms, square footage of the living area, and much more). Methodologically, we use the variation in this rich data to estimate the value of land at the property-level using an ML method and a common hedonic approach, which both deconstruct the total transaction value of each property based on the estimated value of its components. The core idea of our approach is to use detailed market transaction data to estimate what the value of the land is by subtracting the marginal value of the components associated with the structure (as if the structure was "zeroed out"). While we discuss the specifics of each method in the proceeding sections, we should emphasize at the outset that because we estimate land value at such a micro level, this approach is flexible enough to aggregate to any geographic level, including, of course, to state and national levels. Given that large national datasets are becoming more commonly used in the most recent literature (e.g., Davis et al 2021, Nolte 2020, Wentland et al 2020), we interpret this micro-to-macro approach to be the new standard in the land valuation literature. Moreover, this approach is consistent with a broader movement toward exploring the utility of "big data" for constructing official national statistics, which is a recent trend adopted by statistical agencies (see, for example,

[5] See Wentland et al (2020) for further discussion of the SNA, SEEA, and the nuances of how land is defined and measured in the statistical manuals.

[6] In the next section, we will discuss the market for vacant land and how studies have used vacant land sales to generalize land value to improved land with structures. For various reason that we discuss later, our focus in this version of the working paper is on improved land with structures.

Moyer and Dunn (2020) and Abraham et al. (2019) for summaries of recent applications throughout the US government).

While our goal is to eventually build out a full accounting of land value across residential, commercial, industrial, and agricultural land in the US, our initial focus in this paper is on single family homes for two reasons. First, in aggregate, the land underlying single family residential (SFR) property is by far the most valuable land use category in the US (Wentland et al 2020). Methodological improvements to accurately measure this type of land thus contribute the most to accurately measuring the value of all privately owned land collectively. This is also one reason why SFR land is the focus of much of the land valuation literature to which we are drawing comparisons with our new ML approach (e.g., Davis et al. 2021, Davis et al. 2017, Kuminoff and Pope 2013). Thus, the second reason why our initial focus is on SFR land is that most of our analysis is in drawing comparisons to methods in this literature (and to the hedonic approach in particular, as used by Kuminoff and Pope (2013) and Wentland et al. (2020), for example). If we can demonstrate significant advantages for our approach on this particular type of property, it motivates broader use of this method where the literature is thinner and comparisons are less direct. We return to this point in the last section of the paper as we discuss further applications and how this approach could be used to construct land accounts broader in scope.

This paper makes a number of contributions to the literature. Methodologically, this is the first paper to apply on a national scale a particular machine learning approach, gradient boosting trees, to land valuation in a way that conceptually mirrors a hedonic method. We find that this ML approach delivers a number of specific advantages over land valuation using hedonic models like those used by Kuminoff and Pope (2013), Diewert et al. (2015), Wentland et al. (2020), and numerous others. On average, the ML approach models the sale price outcome far better than a linear (OLS) hedonic model, as evidenced by a substantial reduction (about 75%) in RMSE in the price prediction. Since the structure and land value are essentially decomposed from the coefficients that predict sale price, any error in the model's sale price prediction is likely to be reflected in the error of its components, and thus the land value estimate. That is, a substantial improvement in the models predictive accuracy builds confidence in the decomposed land value by reducing the scope for large error on the outcome (on average).

A second contribution is that we provide a tractable alternative to geographic/location fixed effects (like census tracts/block groups/blocks, zip codes, etc.) that could be useful for numerous applications of large datasets like we use. From a modeling standpoint, a benefit of spatial fixed effects like census tracts or blocks is that they can account for fine location-specific heterogeneity, like neighborhood-specific amenities, which are critical determinants of property values. Recent work by Davis et al. (2020) employ zip code fixed effects, for example, where the initial data includes 18,322 zip codes nationally, and other studies use census tracts or block groups with even finer spatial granularity. However, there is a well-known tradeoff here, econometrically. As the level of fixed effects becomes more fine-grained (i.e., the number goes up), there are fewer observations per group in the sample. At some point there may be very few sales in a given census tract, for example, resulting in many of the usual overfitting problems. Davis et al. (2020) and Wentland et al. (2020) sidestep this issue by establishing some arbitrary cutoff that eliminates geographies that include fewer than 50 sales, for example. However, deciding what this cutoff should be is inevitably *ad hoc* and can have a substantial impact on the results. We propose a more systematic approach here. Instead of relying on preset geographic boundaries and arbitrary thresholds for sale counts, we employ a *k*-means clustering approach that generates clusters within our sample that minimize not only latitude/longitude distance but also characteristics of the home (like bedrooms, bathrooms, number of rooms) that are important for determining the outcome (i.e., sale price). This allows our ML model to use fewer fixed effects (or clusters) in order to avoid the small N problems among small geographic areas, while preserving high performance for model fit (as shown by our RMSE statistics) by grouping more homogenous homes across greater dimensions than geography alone. In an era where big datasets and large numbers of fixed effects (and interactions) are the norm, a key contribution we would like to highlight is how this approach can employ fixed effects more effectively.

Finally, although the initial focus of this working paper is to explore the advantages of a new ML approach in comparison to a hedonic approach to land valuation, a broader goal of this project is to generate aggregate land value estimates using novel data that could be used to generate new values for the national economic accounts and/or complementary satellite accounts based on SEEA accounting standards. While we build on Wentland et al. (2020) and others by employing an ML method that has tangible advantages, like reducing the *ad hoc* modeling decisions involved in fixed effects, we should also emphasize a broader point is that reducing the *ad hoc* decisions in

the process of producing official statistics is a meaningful goal. A more systematic, transparent approach to modeling can provide more confidence in the results by the public; and, if the same method is used across countries for national accounts, for example, it would facilitate comparability of the resulting statistics (minimizing arbitrary decisions that go into modeling). In the final section of this paper, we return to this point, discussing potential next steps for this work. In addition to potentially building a national account based on these estimates for macro applications, like Davis et al. (2021), once published, we intend to make all of our land value estimates available at a variety of subnational levels (tract, zip, county, state levels) like FHFA to all who would find them useful in their research or decision-making.

2. Measuring land value: Conceptual background, literature, and the hedonic approach

2.A. How is land valued? Some background and discussion of recent literature

Broadly speaking, there are two ways to value land using market data. One approach might be to directly measure land value by observing what land (without a structure) sells for on the open market and use the market prices and quantities we observe to total an aggregate value of land, much like one would tabulate the aggregate value of any commodity, good, or service. But, for a number of reasons, using price and quantity data alone will not suffice in a vast majority of circumstances. In the case of agricultural land, where this approach might seem most reasonable given that many of the transactions will not include a structure of any kind, price and quantity alone might not be enough information to generate a reasonable estimate because of the problem that not all land sells in a given period, and thus the market sample may not be representative of the land off the market. Further, there is still significant heterogeneity even among agricultural land in terms of soil quality, geographic proximity to markets and infrastructure, and numerous other factors that require more than simply prices and quantities.[7] Thus, these core problems (i.e., the fact that not all land sells in a given period and that the land that does sell is typically bundled with a structure) has spawned a deep literature in using additional information to get at the underlying value of land in a more sophisticated way.

[7] The *2015 Eurostat-OECD Compilation Guide on Land Estimation* includes a variety of caveats when discussing this method, even in nearly ideal conditions. It states: "the direct method is normally preferred by countries for the valuation of agricultural land on which no buildings or structures are situated…[however] since the value of land is highly dependent on several factors e.g., location, land use and the presence of nearby facilities, such information should be incorporated in the land price data" (p. 60).

The more common approach to land valuation can be described then as an indirect method, which refers to a set of approaches that use additional information to estimate the value of land from some other value (like a total property value containing both the structure and land) or an extrapolation from vacant land sales (to similar properties with structures, for example). According to the 2015 Eurostat-OECD Compilation Guide on Land Estimation, these include the residual, land-to-structure ratio (also called land leverage), and hedonic approaches. These indirect approaches reasonably assume that the value of the property is the value of the bundled components of the land and associated structure(s). Conceptually, land and the structure(s) are assumed to be separable assets, and the values of these bundled components do not necessarily move together (Bostic et al. 2007). That is, land is often found to be an appreciating asset over long periods of time while the associated structure is found to be a depreciating asset over time (with some exceptions, e.g., historic structures). In its simplest form, we might think of this as a linear and additive model where the selling price of a property V, the value of the structure p_sS, and the value of the plot of land p_LL can be written as:

$$V = p_sS + p_LL$$

where S is the size of the structure, L is the land area, and p_s and p_L are the prices of a unit of S and L, respectively. The challenge then is how best to determine either p_s or p_L, given that we have information on V, L, and S in real estate sales data, or we might be able to infer structure value in other ways (e.g., construction cost data).

The different indirect methods differ primarily on how land value is decoupled from the property's total value. The residual approach (or some variation thereof) is often used by both governments and academics. Residual methods typically rely on construction or builder's costs as replacement costs (e.g., Davis and Heathcoat 2007, Davis and Palumbo 2008) or demolition costs factored into "teardowns," which are near substitutes for vacant land (e.g., Rosenthal and Helsley 1994, Dye and McMillen 2007). Davis et al. (2021) employ a novel cost-based residual approach using very detailed appraisal records. Their dataset constitutes a very large portion of single-family homes in the U.S. and they provide land value results for various geographies, which we use later in the paper for comparison purposes.

There are a number of other novel approaches to land valuations that may use vacant land transactions to extrapolate to those with structures (or the approach may not neatly fit into these

categories, as they use a hybrid approach). Several examples that use vacant land transactions in various ways include Nolte (2020), Albouy, Ehrlich, and Shin (2018), Barr et al (2018), Turner et al. (2014), Nichols et al. (2013), Combes et al. (2019), and Haughwout et al. (2008). While these studies take a number of sophisticated approaches to try to address various drawbacks to using vacant land, a fundamental issue with using vacant land transactions is that vacant land may suffer from important systematic selection issues and unobservable differences. Vacant land is often land that was previously selected over for development for various difficult-to-observe reasons; or, geographically it is more likely to be on the outskirts of developed core areas of metropolitan areas; or, it may have unobservable differences like whether water or sewer lines are in place. A recent working paper by Larson and Shui (2021) takes a novel approach adapted from Davis et al. (2020) by using Kriging, a spatial interpolation approach using data from Maricopa County, Arizona. Though our review here may not be exhaustive, we should acknowledge here that a key takeaway from the literature is that there are numerous reasonable approaches to land valuation that exploit different types of data to get at this fundamentally difficult question to pin down with precision: what is land worth? In fact, the Eurostat-OECD manual on best practices for land valuation (*2015 Compilation Guide on Land Estimation*), acknowledges that no method is perfect, and states that, "there is no 'best' method; which of these approaches should be used, heavily depends on the available data sources" (p. 66).

2.B The hedonic approach – a relatively simple baseline method suited to "Big Data"

As we discuss in more detail in the next section, our data contains very detailed information about transactions and property characteristics. Generally, this type of data is well-suited to a hedonic approach to estimate land value.[8] This approach uses detailed information on by regressing sales prices of properties on a variety of characteristics of the land and structure, which yields an estimate of the market value of the structure (not just its cost) using variation in the data from comparable structures and properties. One recent study by Rambaldi and Tan (2019) observed that a key advantage of the hedonic method is that "it is a revealed preference method that estimates the contribution of each characteristic to the overall price" (Rambaldi and Tan 2019, p. 5) as the coefficients each represent an incremental or marginal contribution to the price based

[8] In addition to Wentland et al. (2020) and Kuminoff and Pope (2013) mentioned above, there are a number of other instructive hedonic studies, including for example Gong and Haan (2018), Burnett-Issacs et al. (2016), Rambaldi et al. (2015), and Diewert et al. (2015).

on available data. This allows for a nuanced, location-specific estimate based on observed market prices as opposed to costs.[9]

We adapt and tweak the hedonic approach used in Wentland et al. (2020) to establish a baseline approach for comparison to our ML approach that is both common in the land valuation literature and well suited to the data. The hedonic approach generally relies on a standard ordinarily least squares (OLS) regression model and is generally less intricate than more advanced machine learning techniques used by Zillow's proprietary automated valuation model or our machine learning variant. For residential properties, we first estimated the following hedonic model for each time period (3 year overlapping window) and state separately:

$$log(Residential\ Property\ Sale\ Price_{ijt}) = \alpha + \sum \beta\, X_i + \gamma LOCATION_j + \sum \delta\, sq.ft._i * LOCATION_j + \sum \varphi \log(acreage_i) * LOCATION_j + \rho YEAR_t + \varepsilon$$

where X is a set of physical characteristics (number of total rooms; bedrooms; bathrooms; floors; the structure's year built in relation to the median; living area measured by square feet; natural log of the lot size measured by acreage; and separate indicators equal to 1 if the home had a pool, had a basement, or had a porch or had a missing value for each of these variables); LOCATION represents location (census tract, county, or state depending on number of sales at location level) fixed effects; and YEAR includes year-by-quarter time fixed effects for single family homes (or simply year fixed effects for the other property types due to fewer sales observations) to account for time-varying heterogeneity.[10]

We interact the location fixed effects with structure square footage and logged acreage, respectively. For practical reasons, we initially use census tract fixed effects, although we obtained

[9] The hedonic valuation fits with the idea of land value put forth in the *2015 Guide* stating that: "on the balance sheet land should be valued at its current market price (SNA 2008 paragraph 13.16, ESA 2010 paragraph 7.33)…When market prices for transactions are not observable, valuation according to market-price-equivalents provides an approximation to market prices. For example, if the market price of a certain piece of land is not available, prices of land with a comparable use and location could be used" (p. 25).

[10] The Zillow ZTRAX dataset contains quite a bit more information about individual properties that would help with valuation, but we chose the variables with extensive coverage across all states in the dataset. When compared to a fuller model that includes many more home characteristics than we end up using, the marginal gain in precision was small compared to the potential loss in observations due to missing data in states/localities that do not regularly report certain variables. In some cases, where a key variable like the structure's square footage is not reported widely in a particular state or municipality, we ran the regression without this variable separately.

similar estimates using finer-level geographic fixed effects like census block groups.[11] Although this approach is intensive for processing, it allows the valuation of structure square footage and acreage to vary by a finer geography than typically available. This is key, as the valuation of these attributes can vary widely across areas within a state. For example, an additional tenth of an acre for a property in New York City, will be valued much differently than the same amount of space in Albany, which this model with interactions allows for this coefficient to be different by location.[12]

Within each state and period, we then used these coefficients to compute a land price prediction for each property in each year, using each three-year overlapping window. Our model generates a total price prediction for each individual property based on its characteristics. We used the value of the property's location and acreage to obtain the underlying nominal land value of each property, based on the following calculation:[13]

$$Residential\ Land\ Value_{ijt} = e^{\alpha+\gamma LOCATION_j+\sum \varphi \log(acreage_i)*LOCATION_j+\rho TIME_t} \times e^{\frac{1}{2}e^{RMSE^2}}$$

Because we used relatively fine (spatially small) location fixed effects, all time-invariant local amenities within each tract (and within the period of estimation), including land-cover types (and by extension corresponding ecosystem services associated with each land-cover type) will be incorporated into the tract coefficients valuing location. Thus, each land value we estimated for each property will account for net market value of location-specific amenities

Due to the nature of the data, several issues arise with the hedonic model that prompt *ad hoc* decisions to rectify. One issue in the hedonic estimation of land value is that the tails of the distribution can often produce extreme values, particularly when there are thin cells (i.e., states and years with land-use categories having few sales and some extreme sales), from which the

[11] Smaller geographic units like block groups and blocks have fewer sales, so the advantages of finer location controls need to be balanced with thinness of sales within these areas (which can create some noisiness in the estimates). The interactions also become problematic for estimation of too many fixed effects in most statistical software packages, however. We have also explored a variety of other specifications to improve model fit and predictions, including a semi-log specification, where sale price is logged, but these models produced similar results overall.

[12] This approach is used commonly in the hedonic valuation literature for housing and land (e.g., Kuminoff and Pope 2013). As we discuss in more detail below, we require a minimum number of transactions to occur within a location (e.g., tract) over a given period, pooling observations that do not meet this threshold at a higher geographic level (e.g., county) in a separate regression.

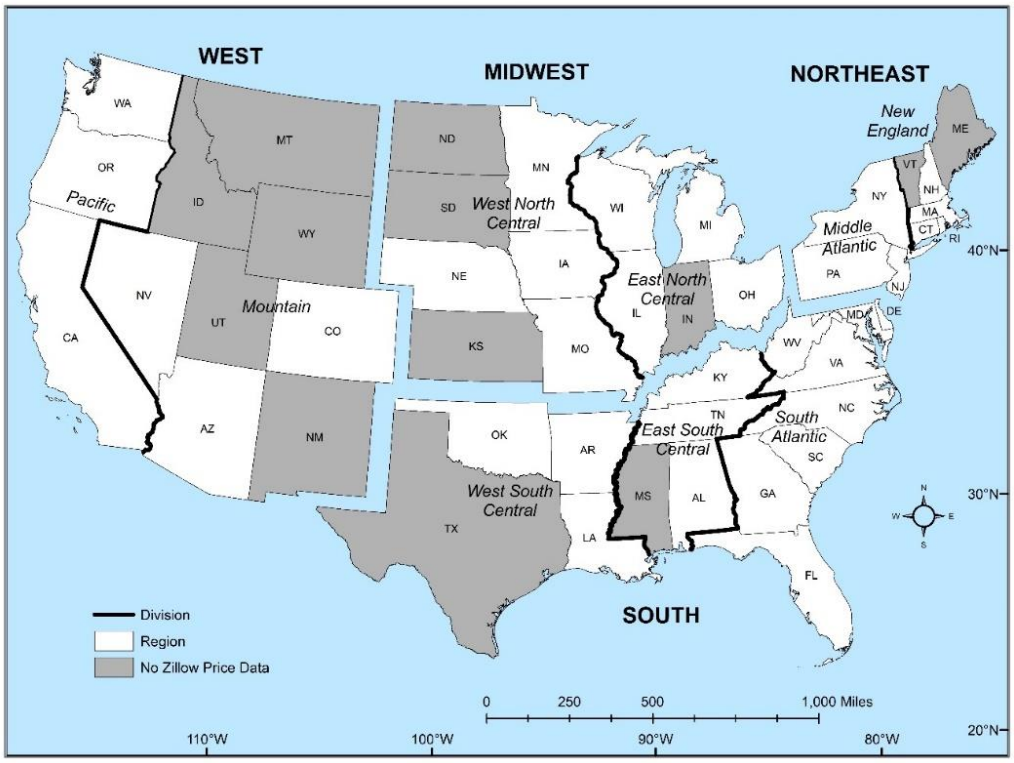
[13] Note that since the outcome variable is logged, we have smeared the prediction following Duan (1983).

model generates a (semi-log) linear prediction. To avoid making predictions for thin cells, like Davis et al. (2021), we establish a threshold under which we do not allow observations to be modeled using that fine-grained of a fixed effect. Specifically, we specified that a given tract have at least 30 sales in the three year window for each model. If this condition was not met within a given tract a period, we iteratively estimated models with higher-level geographic fixed effects (i.e., the same model only using county (FIPS)-level fixed effects and a separate model using the entire state). Moreover, one reason why we use a three-year running window is that a single year of data will often yield noisier prediction results for hedonic models using fine fixed-effects, making this threshold of N a more binding constraint for more of the dataset.

Because there may be leftover noisy predictions for areas with sales marginally above these thresholds, we made adjustments to outliers in the following ways. First, we create our land value measure using the predicted land value from the model using census tract fixed effects. We then replace any missing or outlier predictions by first replacing those in the extreme ends of the distribution (less than the 1st percentile or above the 99th percentile – a prior version of our model followed Davis et al. 2021 using a hard-coded threshold of $200 price per acre) with the predictions from a county fixed effects model. Any further missing predictions are taken from the state-level model. This tiered fixed effect approach ensures we are not systematically throwing away data from rural areas, for example, where the number of sales over a particular time window may fall underneath our threshold. Finally, we then cull any outliers above the 1st percentile or above the 99th percentile. These adjustments ensured that model coefficients were not driven by erroneous or mis-measured data, small samples, or outliers.[14] Nonetheless, a key takeaway from how we deal with these problems, the thin cell problem and outlier problems, should be that we (and many others), if we are to be transparent about our method and design choices, must communicate a lengthy description of the nuances and arbitrary thresholds have to establish to run these models and get reliable, reasonable results. We return to this point as a potential problem that data-driven methods like machine learning can help solve in less arbitrary, more systematic ways.

[14] One potential issue with the hedonic approach, or any prediction-based multivariate method, is multicollinearity, where the acreage of a property could be highly correlated with the size of the structure (square footage), particularly for land in dense urban areas. This could produce bias or imprecise estimates of land value if there is a mechanical relation between these two variables such that value is not meaningfully separable. In untabulated analysis, we examined the correlations between acreage and square footage of the structure in our data, finding the correlation was not high in the U.S. (usually falling within 0.2-0.4).

Figure 1. Census Regions and Divisions of the United States (Source: U.S. Census), showing states with and without Zillow data.



3. …in the era of big data…: Data description

As we alluded to in the introduction and prior section, one of the novelties of this study and only a few very recent studies like Davis et al. (2021), Nolte (2020), and Wentland et al. (2020) is that we leverage very fine, property-specific microdata to generate national estimates from millions of data points that span much of the US. Specifically, our study uses the Zillow Transaction and Assessment Dataset (ZTRAX) dataset that has been recently made available to researchers in academia and government for a limited period of time (through September 2023). It contains market transaction data as well as a large set of individual property characteristics for sales recorded in local tax assessor's data.[15] Coverage is generally representative of the United States' national market, initially comprising 374 million detailed transaction records across more

[15] Data are provided by Zillow through the Zillow Transaction and Assessment Dataset (ZTRAX). More information on accessing the data can be found at http://www.zillow.com/ztrax. The results and opinions do not reflect the position of Zillow Group. Nonproprietary code used to generate the results for this paper is available upon request to the authors.

than 2,750 counties (i.e., 91.5% of U.S. counties). Not all U.S. states require disclosure of sale prices, so while our data cover a large portion of the country, the price data in particular have some limitations in coverage, specifically for 13 (mostly rural) states.[16] The data include detailed information on each individual home's sale price, sale date, mortgage information, foreclosure status, and other information commonly disclosed by a local tax assessor's office for each real estate transaction.

We join each transaction to each property's characteristics into a single dataset to be used for our analysis, so that each transaction has the corresponding property characteristic data from the assessment dataset. Specifically, the assessment data include a number of characteristics found on Zillow's website or a local tax assessor's office describing a property: the size of the structure on the property (in square feet), lot size (in acres), number of rooms, bedrooms and bathrooms, year built, and various other characteristics.[17] A key aspect of these data is that it contains detailed information about each property's location (address and latitude-longitude) such that this fine-level spatial data can be linked to any level of geography commonly used in hedonic property analysis.

The data from Zillow is originally packaged in a somewhat raw form. As a result, we scrutinized missing data and extreme values as part of our initial culling of outliers and general cleaning. Some outliers may arise because they are either foreclosures or non-arm's length transactions (which we omit using variables such as the document type), but others are typos in the source data (e.g., where a municipality records the number of bathrooms as 40), or the property itself is unusual enough that it would not be an appropriate fit for a model (e.g., if the home did, in fact, have 40 bathrooms, it is unlikely that each bathroom is valued in the same way as other, more typical properties). Or, this might signal a misclassification of a property, where a building with 40 bathrooms may actually be a commercial office building. Hence, we dropped extreme

[16] Because some states do not require mandatory disclosure of the sale price, we currently do not have price data for the following states: Idaho, Indiana, Kansas, Mississippi, Montana, New Mexico, North Dakota, South Dakota, Texas, Utah, and Wyoming. In addition, some states like Louisiana, Maine, and Vermont have substantial missing data in our current sample, and we omit these states as well.

[17] Zillow's ZTRAX data contain separate transaction files by state, where all transactions need to be linked to corresponding assessment records. With guidance from Zillow, we were able to merge the bulk of the data, but not without some data loss (which figures into the size of our final sample).

values for price and home characteristics for our analysis, which is a common practice for recent research using this particular data.[18]

For this analysis we retain single family residences with non-zero acreage, then cull those with acreage less than the 2.5th percentile of acreage and above 2.5 acres.[19] We removed properties that had a structure smaller than 50 square feet and a price lower than $1,000. We then culled by price at the 2.5th and 97.5th percentile by year and county. We culled homes with square footage (a home's living area) above the 97.5th percentile and year built (we use year built – median year built so that the intercept is for a home built in the median year) below the 2.5th percentile. Homes were also winsorized using total rooms at 11, bedrooms at five, bathrooms at four, and number of floors at three, thus confining the influence of outliers in our hedonic model. We remove from our model any indicators for the presence of a porch, basement, and garage if less than 5 percent or more than 95 percent of properties in the land-use type and period had the amenity (we use 1 and 99 percent for presence of a pool). We remove variables if more than 75 percent of properties in the land-use type and period were missing and recode to the average if less than 5 percent were missing. Lastly, we remove from our sample any residential properties that do not provide some form of structure size (either square footage, bedrooms and bathrooms, or total number of rooms). While the Zillow dataset contains a vast number of property characteristics, we primarily relied on the variables above, which have the most coverage nationally to limit how much data we discarded in our initial analysis.[20] We limited the sample years to 2002 through 2015, as data for those years are most complete for the vast majority of the states in our sample and, for our baselines hedonic model we use a rolling three year window of sales to estimate land value for each year. One advantage of this time period is that it provides a lot of variation in the data, as this period includes the intense periods of boom, bust, and recovery in the U.S. real estate market.

[18] See Nolte et al. (2021) for a broad discussion of best practices using the Zillow ZTRAX data, which cites some of our prior work using this data (e.g., Gindelsky et al. (2019)). This is a very useful guide to using the Zillow data; and, while some of the precise thresholds and cutoffs may differ, we follow many of the general suggestions of this paper makes.

[19] Homes with more than 2.5 acres are classified as rural residential and will be considered in a future version of this paper.

[20] In untabulated regressions, we conducted a sensitivity analysis for subsets of the sample that employed more property characteristics to determine whether the results are sensitive to omitted variables for which we can control. Our results were generally robust to omitting variables that have more limited coverage.

4. …and machine learning: Methodology

4.A. Gradient boosting trees and k-means clustering – a machine learning approach to land valuation

At its core, the hedonic valuation of land necessitates some form of prediction for the overall price of the property (land + structure) and its components. Recall that once we decompose the price of a property into its constituent parts, the idea is that we can evaluate the marginal effect of changes to the property (e.g., an extra bathroom), or rather extrapolate what the price would be in the absence of a structure altogether. We do this as a baseline in our traditional hedonic model described above, as our objective is to isolate the value of the land from the remaining elements. Because not all properties sell in a given period, we project a predictive model from single family homes that sell onto the assessment data which contains the near universe of houses, most of which are not observed on the market in the periods observed (or may never be on the market). Thus, we turn to Gradient Boosted Machines to help predict the price of both homes on the market and homes which have not, and may not ever, see the market (Friedman, et al., 2000; Friedman, 2001; Friedman, 2002). The motivation for doing this is that because the hedonic approach relies on a prediction model to generate land value estimates, if we can predict the overall price more accurately, as measured by a loss function such as root-mean-squared-error or root-mean-absolute error, we might have better insight into the relative value of land for each property. This is precisely a strength of machine learning approaches like gradient boosting.

Gradient boosting is a learning algorithm which combines individual weak learners [decision trees] through iterative construction such that each subsequent tree attempts to correct the mistakes of its predecessor. The gradient being evaluated depends on the loss function chosen given the context of the modeling. In this case we have chosen the L2 loss function (least squares), $\frac{1}{2}\big(y_i - f(x_i)\big)^2$, with gradient $-\delta L\big(y_i, f(x_i)\big)/\delta f(x_i) = y_i - f(x_i)$. In each iteration, a tree is built on a random sub-sample of the data and this tree is of fixed depth. In our case we have chosen an interaction depth of four to limit the possibility of overfitting for each individual tree. Note that for each iteration the target is not the sales price of each individual home but rather the residuals of the previous iteration. The learning rate, or how big of a step along the gradient, is limited for each tree to the default parameter of $\gamma = .1$. In the Algorithm 1 table below, we have outlined the generic framework of a gradient tree boosting algorithm.

Algorithm 1: Gradient Tree Boosting Algorithm	**(Hastie, et al., 2017 pp. 361)**
	1. Initialize $f_0(x) = argmin_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$
	2. For $m = 1 \; to \; M$:
	(a) For $i = 1, \dots, N$ compute $$r_{im} = -\left[\frac{-\delta L\left(y_i, f(x_i)\right)}{\delta f(x_i)}\right]_{f=f_{m-1}}$$
	(b) Fit a regression tree to the targets r_{im} giving terminal regions $R_{jm}, j = 1, \dots, J_m$.
	(c) For $j = 1, \dots, J_m$ compute $$\gamma_{jm} = argmin_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$$
	(d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} \, I\left(x \in R_{jm}\right)$
	3. Output $\hat{f}(x) = f_M(x)$.

Recall that we are not particularly interested in drawing causal inferences on the marginal effect of adding an additional bathroom, rather we are first interested in minimizing the error on our predicted price. We would like to be able to easily strip the predicted price of the overall contribution made by the structure itself conditional upon the observable characteristics. The choice of features then becomes very important because we would like to avoid splitting on elements such as number of bathrooms in the housing unit. Additionally, it is well known that the first second and third most important elements of home price are: location, location, and location.

In a generic linear hedonic model this amounts to having location level fixed effects (e.g., census tract/block) interacted with observable features of the property including plot size, square footage, etc. This leads to a rapid expansion in the number of variables in the information set. For example, the hedonic model representing single family housing in California has approximately 10,000 right hand side variables between observed elements of the structure, location fixed effects, time fixed effects, and included interaction terms. This introduces a fundamental problem with many fixed effects in linear models, which we discussed in some detail above. While tree algorithms are good at dealing with many features, conceptually, recall that our goal is to remove the structure all together. To do this we first cluster the data using a nearest-neighbors algorithm over a multi-dimensional space including latitude, longitude, number of bedrooms, number of

bathrooms, total rooms, the presence of a porch and/or basement, the presence of a garage, the number of stories in the structure, and the year the structure was built. This means that, within a given cluster, we are minimizing the variance of the properties over these dimensions.

Within a single cluster this process would generate a cluster that, for example, might have predominately three-bedroom, two-bathroom houses with a porch, basement, and garage all within a similar geographic location and built roughly in the same period. The implicit assumption using this method is that, within a confined geographic area, once we have a set of houses with nearly identical observable elements (e.g., number of bedrooms) then price is going to vary over two general dimensions: first, the size (measured in square footage of the house) and second, the land and location effects. Note that we did not include square footage in the clustering elements and thus, within the cluster, the model loads the price variation within that cluster for the structure specifically on the square footage of the house. Also note that we did not include acreage in the clustering elements and thus the types of plots these houses sit on is allowed to vary.

Conceptually, when we exclude land or the home square footage in the clustering algorithm, the within cluster variation is going come from the land itself (acreage) and the size of the structure, not its features (as we are already comparing homes with similar features within the cluster by design). So, while there is substantial heterogeneity in structure features in any given neighborhood, even in a development built by the same builder, a k-means clustering approach essential assembles geographic clusters of relatively homogenous homes that primarily vary on these other dimensions (size and acreage). For example, in California, the average standard deviation on the number of bedrooms within our k-means constructed clusters is less than ten percent that of the within cluster standard deviation of census blocks (0.065 versus 0.709), which we show in Table 1 below. For bathrooms, the generated clusters exhibit slightly more than ten percent of the average variation (0.065 versus 0.575). Thus, while characteristics like the numbers of bedrooms and bathrooms are common determinants of price differences, by reducing the variation along these dimensions and location, this approach allows for a more apples-to-apples comparison among properties within clusters than small geographic fixed effects like census tracts or block groups. One might think of this approach as similar to how a professional appraiser would select nearby “comparables” or “comps,” which may be outside one’s census tract, block, or even zipcode, but is still geographically close and contains very similar features.

For each state-year we apply the gradient boosting algorithm above to the sales data with the estimating equation:

$$salesprice = f(latitude, longitude, sqft, acreage, cluster, yearbuilt).$$

Our location effects in this case are latitude, longitude, cluster, and year built; where year built is both an imperfect proxy for structure quality (depreciation) and potentially for the unobserved land amenities of the property (i.e., the flipside to the vacant land selection bias – land developed earlier, within a certain geographic location/cluster, likely has more positive unobservable amenities and infrastructure than properties built more recently in that area). While this is not strictly true as houses are rebuilt over time, thus resetting the development year, we think, on average, it is likely a reasonable proxy for these qualities.

To back out the structure value we then predict a new sales price based on a five percent increase in the square footage of the home and take the difference in prices to obtain the marginal value of the structure (conditional one being within a cluster of similarly featured homes). This price difference is then linearly extrapolated to the entire square footage of the property and the resulting structure value is subtracted from the sales price (predicted in the case of the assessment data). That is,

$$sv_i = 20(\tilde{P}_{i|sqft+} - \tilde{P}_{i|sqft})$$

where sv_i is the structure value for the i^{th} property, $\tilde{P}_{i|sqft+}$ represents our predicted price with the marginal increase in square footage (holding all other features at their original values), and $\tilde{P}_{i|sqft}$ is the original predicted price based on observed square footage. This means that the land value is the difference between the original predicted price and the structure value, $lv_i = \tilde{P}_{i|sqft} - sv_i$. To prevent negative land values, we-top coded the value of the structure at the full predicted price of the property.

To calculate the price-per-acre we divide the estimated land value, lv_i, by the observed acreage for the property. For a property with land value 10,000 dollars that sits on 0.25 acres of land this would imply a price-per acre of $\$10{,}000/.25\ acres = 40{,}000$ dollars per acre. We do this at an individual level so that we can then aggregate to any geographic level, j, by averaging over all properties in that unit, $n_{i\in j}^{-1}\sum_{i\in j} ppa_i$. Since we formed our geographic fixed effects

through the k-means (data driven) process our clusters do not correspond to any fixed geographic unit (e.g., census tracts) and as a result within cluster measure of price-per-acre would not be comparable across standard geographies, this is the reason we have predicted price-per-acre at the individual property level.

4.B How do k-means clusters stack up against other spatial fixed effects?

While census tracts are designed with some geographic and population homogeneity in mind, the housing stock within these boundaries varies widely. Variation in data is not inherently a bad thing for price predictions; it is, however, problematic when there is a small number of sales. Specifically, in a given census tract or block group that has, for example, 50 sales in a given year – how many of those sales have precisely 4 bedrooms and 2.5 bathrooms (and have a pool, basement, etc.)? Thus, a census tract or block group, despite having "enough sales" may have only a couple good "comps" for appraisal/valuation purposes. A k-means cluster contains far more similar homes such that even if there are 50 sales in a given year, a much higher proportion of those homes will have very similar features. We see this in the data by comparing the standard deviation of a few important characteristics within our constructed clusters versus census block groups.

Table 1 shows that the average within cluster standard deviation of square footage for our constructed clusters is significantly less than that of the census tracts (two sample t-stat of -19.013) but is still substantial (i.e., we would expect there to be some correlation between the number of bedrooms or bathrooms and the overall square footage of the property, hence the decrease in variation). This means that, while we have minimized the variation within cluster across elements such as number of bedrooms and bathrooms, we still have a significant amount of variation in the square footage of the structure to exploit to estimate the structure value in the gradient boosting model (within clusters). Additionally, Table 1 shows that the variation in acreage is larger in the generated clusters. This means that we are grouping houses with similar characteristics (number of bedrooms, bathrooms, etc.) across a wider range of plot and house sizes than one would find using traditional geographic fixed effects. Conditional upon the price predictions being better from the boosting approach, we might have more confidence the underlying decomposition of the components.

Table 1: Within Cluster Standard Deviation Summary Statistics (California Assessment Data)						
	Minimum	First Quartile	Median	Mean	Third Quartile	Max
Generated Clusters: Square Feet	206.9	328.8	451.1	440.7	526.4	812.2
Generated Clusters: Acreage	0.052	0.178	0.242	0.256	0.340	0.548
Generated Clusters: Bedrooms	0.000	0.035	0.245	0.264	0.425	0.857
Generated Clusters: Bathrooms	0.000	0.104	0.212	0.237	0.370	0.840
Census Block Group: Square Feet	0.000	346.6	441.1	474.4	578.5	1908.5
Census Block Group: Acreage	0.000	0.030	0.057	0.132	0.156	1.311
Census Block Group: Bedrooms	0.000	0.605	0.707	0.709	0.807	2.121
Census Block Group: Bathrooms	0.000	0.458	0.571	0.575	0.683	2.121

Overall, this approach includes the individual characteristics and locational coordinates of each property in our feature set this clustering, which allows us to collapse the dimensionality and include only the cluster identifier. The assumption here is that within cluster variation from the omitted variables is very low and that the variation in price because of those characteristics is primarily captured in the between variation of the clusters. The number of clusters for each state is a function of the number of observations available in both the assessment and sales set and are time invariant. We explored several different numbers of clusters but settled on $N^{1/3}$ number of clusters which gives us geographical units smaller than counties but larger than census tracts on average.[21] The average number of clusters per state is 99 with the smallest number of clusters appearing in South Dakota (45) with the most appearing in California (184). The total number of clusters across all 36 states in the dataset is 3,574.[22] The average number of properties in the average cluster (assessment set only) is 10,891. The centroids for each cluster are generated using a K-means algorithm (Kuhn & Johnson, 2013; Hastie, et al., 2017; and Chen et al., 2021 for a discussion).

[21] For example, in California we have 184 clusters compared to 55 counties and 8057 census tracts. The top five clusters by sample size account for 706,869 observations out of a possible 6,402,740 in the assessment set. The top five largest clusters by geography account for a combined 165,441 acres while the five smallest account for 2028 combined acres. The median acreage per cluster is 7,132.98 and the median number of properties within the cluster is 29,071.

[22] For context, according to the U.S. Census Bureau there are 3,143 county and county equivalents in the 50 states and the District of Columbia (not including territories). Our data consists of 34 states and represents 1,855 of the possible 3,143 counties. There are 61,615 census tracts in the 34 states based on the 2010 census.

To illustrate how our clusters compare to common geographic units we have plotted a county level map of Orange County, California in Figure 2. Orange County is as an example of a large county with a large variation in single family homes. The borders you see are those of the census tracts within Orange County, of which there are 582 (1,822 census block-groups). Note that in dense urban areas the census tracts are considerably smaller in surface area relative to suburban and rural tracts within the county. The points on the map represent the 530,533 unique single-family residences in the county assessment data. These properties account for 92,035 acres of land out of the nearly 600,000 acres in the county.[23] The points are colored according to their cluster membership, of which we have 47 clusters in the county. It is important to note that clusters are not confined to any individual tract and can be non-contiguous over the geography. Remember, we are minimizing distance over several attributes in the clustering matrix including latitude, longitude, total rooms, number of bedrooms, number of bathrooms, etc. and there is no requirement that the clusters remain contiguous in such a structure. In fact, the more alike the houses are within the cluster the larger the geographic range is allowed by the algorithm – which is not unlike how professional appraisers evaluate "comps." Moreover, of the 47 clusters that appear in Orange County, a total of 46 of them have some observations in the cluster that appear outside Orange County. Despite their difference in location as measured by latitude and longitude the houses outside Orange County are highly like those inside.

[23] Note that we are focused exclusively on single family residences at this time. While we have data that covers other land types (e.g., commercial, dense urban dwellings, agricultural, etc.) those fall outside the purview of this study.

Figure 2: Clusters are not the same as tracts – an example from Orange County, CA

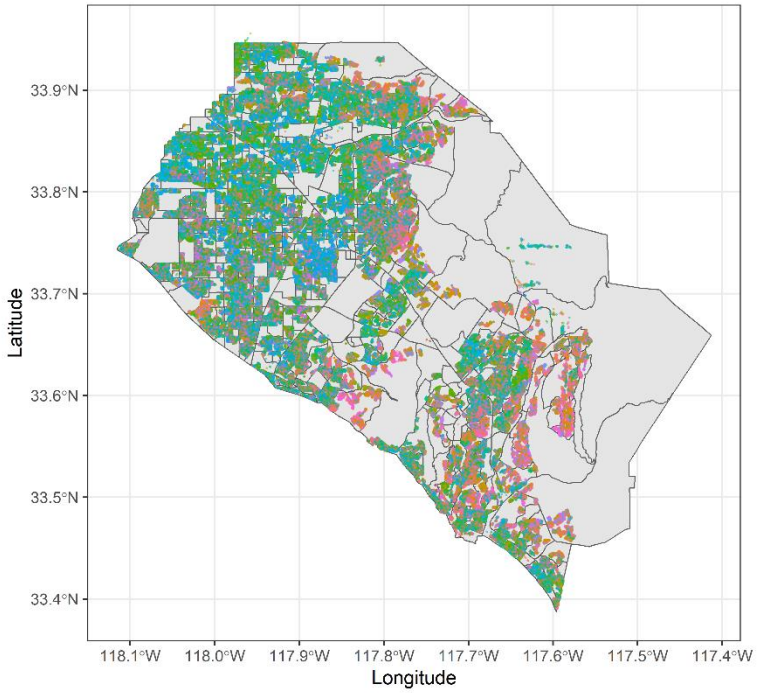


Figure 2 Note: Here we have plotted a map of Orange County, CA with the corresponding census tracts. Each point on the map represents a property found in the assessment data set and it is color coded by assigned cluster. From this map it is easy to see that the data-generated clusters do not follow tract borders and often are overlapping, disjoint, and/or include multiple tracts. There are 47 clusters present in Orange County, CA as compared to 582 Census Tracts.

5. Measuring Land Value in the Era of Big Data and Machine Learning: Some Results

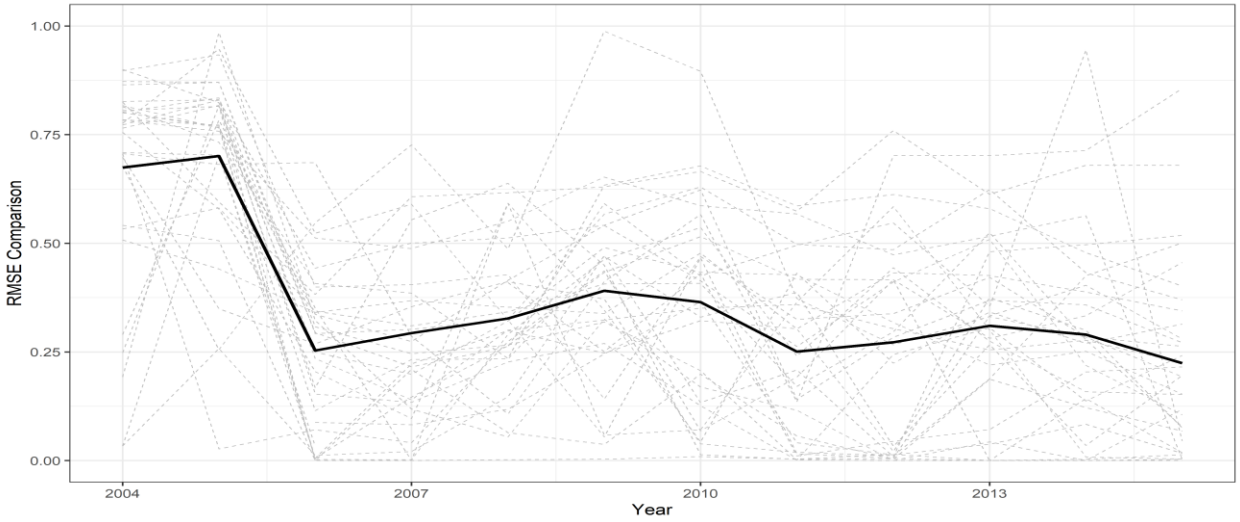
5.A. Comparing Gradient Boosted Trees to a Linear Hedonic Baseline - Evidence from RSME

Before we examine the resulting land values of these models, we first compare the accuracy of our prediction models by comparing the root-mean-square errors (RMSE) of our ML approach with our baseline hedonic approach adapted from Wentland et al (2020).[24] In Figure 3 below, we have plotted the ratio of root-mean-square errors, $e_t = \text{RMSE}_{\text{GBT,t}}/\text{RMSE}_{\text{LH,t}}$, produced by each model (Gradient Boosted Trees and Linear Hedonics respectively) for each state over a decade of data, from 2004 through 2015. This period covers much of the boom and bust in the US housing market, which provides tremendous variation in the data, which we view as a nice setting for this

[24] There are a number of ways to compare models, but a common approach in this literature (e.g., Bencure et al (2019)) and the property appraisal literature more generally is to compare the RMSE.

kind of test. If $e_t = 1$ then both models produce the same loss when predicting the sales price (though their predictions need not be the same at the individual). An $e_t > 1$ implies that for the given state the linear hedonic model has lower loss in prediction than the gradient boosting alternative, and $e_t < 1$ the opposite. The figure clearly shows that for all states in all time periods the predictions of the gradient boosted tree are better than that of the alternative (where, for an average state, the RMSE of the gradient boosted model was about a quarter of the hedonic model in most years). We view the clear and substantial improvement in RMSE for the ML model as a key result of the paper, documenting a particular advantage of the gradient boosting tree method for this key first step of land valuation.

Figure 3. Gradient Boosted Trees Has Lower RMSE than Linear Hedonic Models



Note: Here we have plotted the ratio $RMSE_{GBT}/RMSE_{LH}$ by state (dashed lines). The average ratio of RMSE across all states for each year is plotted by the solid black line. A value of one indicates that the two modeling methods produce the same loss on prediction. A value greater than one means that the Linear Hedonic model has a lower RMSE than that of the Gradient Boosted Trees while a value less than one indicates the opposite.

5.B. Comparing state-level results across methods

While we can aggregate to any level of geography using this approach, our initial analysis begins with states because, more practically, states are easy to compare in a single figure or table and they provide a quick sense of whether our models are on the right track because of the well-known variation in property values (e.g., coastal states are generally higher valued, on average).

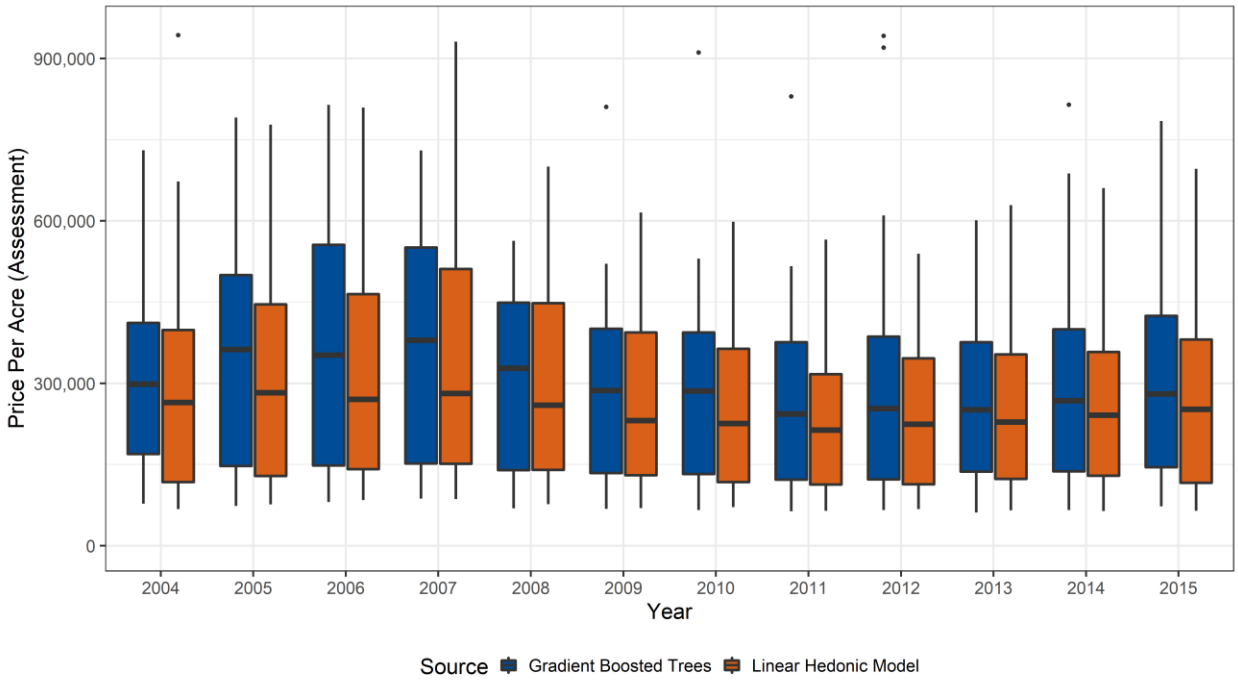
In Figure 4 we have plotted the resulting box plots showing the cross-state variation in each year for the price-per-acre estimate for the states in our dataset. Generally, for both methods, we see a familiar time-series pattern in the data over this time period, as the distribution of land values drifted upward in the boom of the mid-2000s and declined after 2007, coinciding with the real estate bust and subsequent recover into the 2010s.

We show the state-by-state values in the Appendix. Tables A and B in the Appendix include the land value for each state and year, measured as the price per acre, separately for the hedonic and machine learning models. Both models report significant variation in the estimated price per acre across the country. The estimates from the hedonic model range from low values in the $60,000s for Alabama to high values over $2 million for California. The machine learning estimates range from slightly smaller values in the $50,000s for Arkansas to almost $2 million in California. The estimates in both models reflect that state level trends in price per acre followed the boom and bust associated with the Great Recession. There is heterogeneity in how the price per acre adjusted along the business cycle, while states like Arizona, California, Florida, and Nevada experienced explosive growth up to 2007, followed by significant depreciation, other states like Kentucky, Missouri, and Pennsylvania were relatively flat in comparison.

Tables C and D contain the land value to home price ratios for each state and year from the hedonic and machine learning models. The value of land using the hedonic model is roughly half of the total price of the home. The ratio using the estimates from the machine learning model is a bit higher at 69 percent. Both of our estimates are considerably larger than the 32 percent average land value to price ratio in Davis et al (2021). Consistent with Davis et al (2021), it appears that land value is relatively more expensive during expansionary periods. For example, Florida, a posterchild for the housing bubble, experienced an almost 10 percentage point drop in the land to price ratio from 54 percent in 2006 to 44 percent in 2012 using the hedonic estimates and 15 percentage point drop using the machine learning estimates (52 percent to 37 percent). Of the thirty-four states in the final sample, the state with the lowest average price-per-acre (as calculated using the GBT method) over the time periods available was Tennessee which has an average price-per-acre of $79,517 ($6,443) while California provided the highest average price-per-acre at $1,189,564 ($287,360). In the linear hedonic model, the state with the lowest average price-per-acre is Alabama at $71,487 ($8,731) though Tennessee is third at $82,247 ($8,710). California

again takes the top spot with an average price-price-acre of $1,728,107 ($340,856). The median state as measured by average price-per-acre over all the sample periods is Minnesota (using the GBT method) with a value of $342,981 ($51,629) and Nebraska (using the linear hedonic model) $224,961 ($32,244) respectively.

Figure 1*:* Price-Per-Acre by Prediction Method at the State Level – Box Plot Distributions



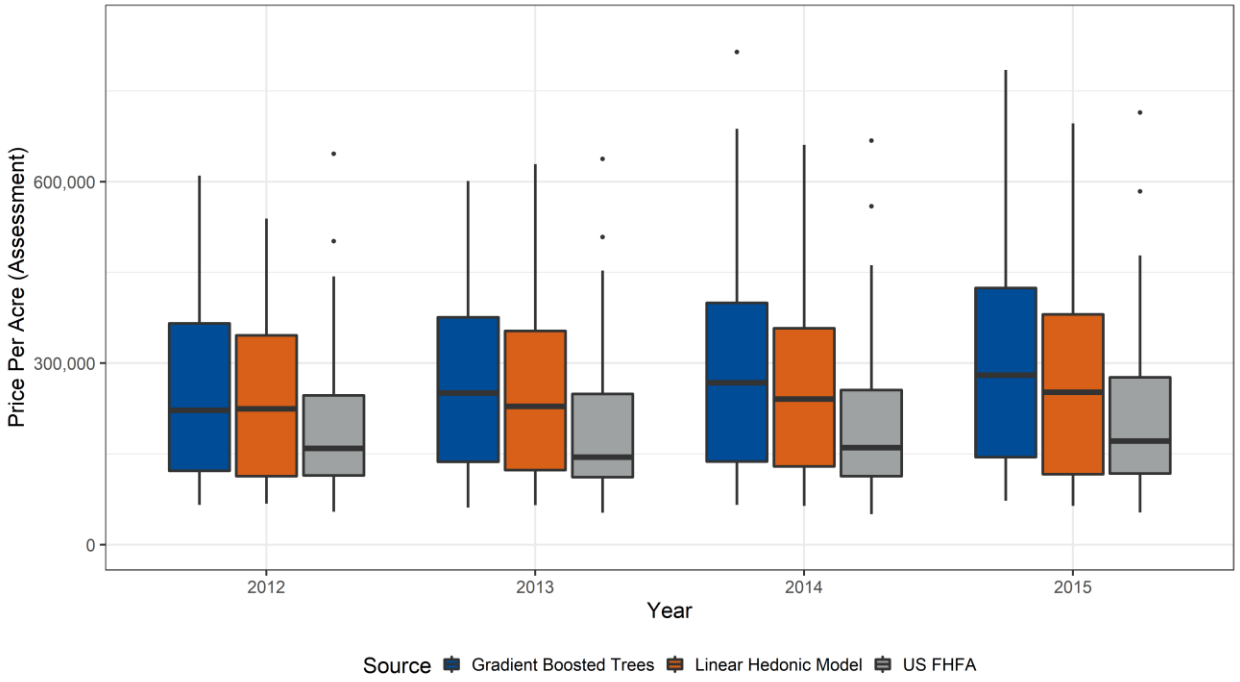
Note: This figure shows the distribution of the average price-per-acre at the state level over our sample period, comparing the land value results from machine learning method (blue) with the results from the baseline linear hedonic method (orange).

When we compare the distributions of the results across methods, in most cases the interquartile ranges appear to be relatively similar in width however the location of the predictions from the GBM are consistently higher than that of the linear hedonic model. This is bolstered by the medians, the GBT median is generally higher than that of the linear hedonic model. The entire distribution produced by the GBT method is locationally to the right of that produced by the linear hedonic model though it is less skewed as the outliers (e.g., California) have a substantially lower average price-price-per acre.[25]

[25] Our plan is to have an online appendix where individual states can be evaluated based on price-per-acre, land-value-ratio, and other metrics of interest by any interested parties. This will be in the form of tables, some of which you will find in Appendix (indicator), and interactive dashboards via Shiny in R.

In Figure 5 we have plotted the intersection of both methods with estimates (or "experimental dataset") produced by the U.S. FHFA, which can be found on their website (Davis et al 2021).[26] Overall, the interquartile range of Davis et al. (2021) is narrower than the alternatives we have presented here, and the median is on the low end of the hedonic model interquartile range, and in some cases outside the interquartile range produced by the GBT method. Additionally, this comes with a change in the outliers as well. For example, estimates provided by the FHFA researchers in Davis et al. (2021) show that New York is the most expensive state as measured by price-per-acre with an average in 2013 of $1,634,500 dollars. In contrast, the price-per-acre estimates are significantly lower for the GBT model ($389,850 for sales set) and linear hedonic model ($356,288 for sales set).[27]

Figure 2: Price-Per-Acre at the State Level – Box Plot Distributions Comparing Three Methods



Note: This figure shows the distribution of the average price-per-acre at the state level over the 4 year sample period that overlaps with Davis et al. (2021), comparing the land value results from machine learning method (blue) with the results from the baseline linear hedonic method (orange) and the Davis et al. (2021) method (gray).

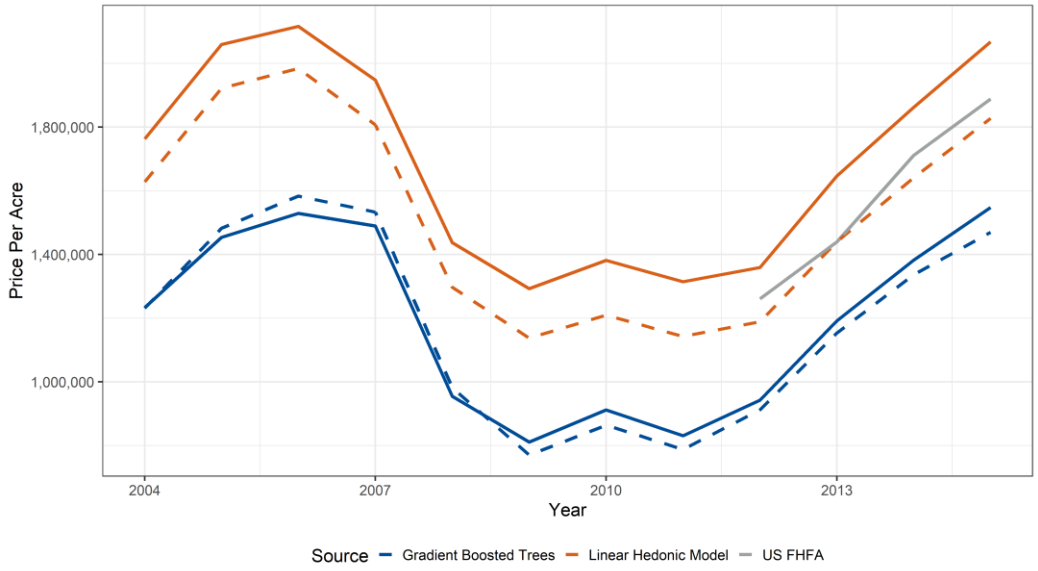
[26] Their dataset can be found here: https://www.fhfa.gov/PolicyProgramsResearch/Research/PaperDocuments/Land-Prices_DLOS_2019_9Oct.xlsx

[27] For the assessment set the estimates provided by both models are $458,071 and $353,270 respectively. This could imply that there is a sampling difference between the data we use here and that used by Davis et al. 2021.

Finally, we turn to a single state example to illustrate some important differences across methods and samples, which are not readily apparent in comparing the distribution of the results in the box plots above. Within a single state, we can see more intuitively the differences and similarities in the results. Figure 6 shows the time series for the average price per acre in California over our sample period, comparing all three methods. At first glance, we can that both the ML and hedonic approach show the dramatic boom and bust in land prices, which was particularly pronounced in California (known as one of the "sand states," with Arizona, Florida, and Nevada, that were hit particularly hard during the real estate boom and bust). Both models show a peak in land prices in California in 2006 and a steep drop through 2009 and recovery thereafter (with the gradient boosted model shifted lower than the hedonic in this example state). The estimates from FHFA researchers and co-authors in Davis et al. (2021) corresponds with the slope of the recovery in land prices in both of our models. As shown in our Appendix tables, we should note that it is not always the case that our gradient boosted model is shifted lower than the hedonic estimates. The ML estimates are, in fact, often higher. However, while the price levels are shifted, we should note that in Figure 7, where we graph the land leverage (land-to-price ratio) for each method in California, we find more similar and somewhat less volatile results, indicating that land leverage in California over this period was somewhat more stable than land prices.

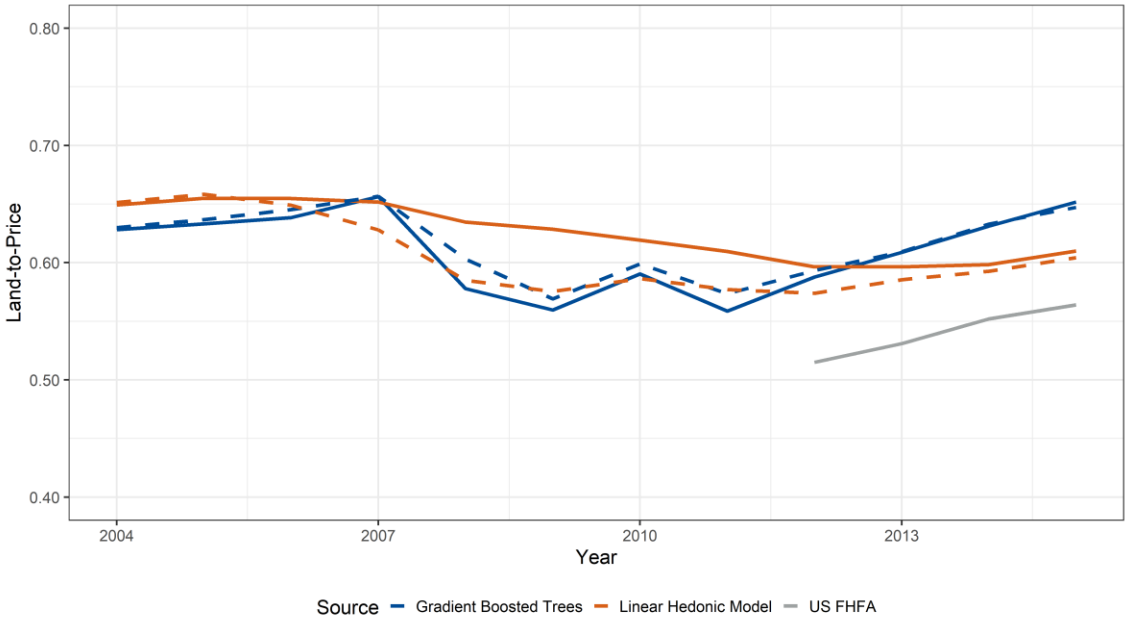
The comparison in Figure 6 also highlights another issue and an important feature of this data. What we usually find when we compare results within states, which is illustrated in this example, is that the sample matters. Specifically, if we calculate a price per acre using only the properties in the sales or transactions dataset in a given year, then we find a significant difference as compared to the same calculation applied to the universe of properties in the assessment set. Recall that the assessment set consists of virtually the universe properties, whether or not they sold in that particular year. The results from California show a consistent pattern when we apply this comparison to other states, which is that we often see a disproportionate amount of high quality homes (in terms of having better property characteristics) not on the market in any given year. This results in the sales data set often being shifted lower than the assessment set, as shown in this example. And, the size of this shift is not uniform over the time-series, with this issue become more pronounced at various points in the boom-bust cycle. Thus, a key feature of this kind of data is that we can weight the results to fit the more appropriate full sample of data.

Figure 3: Price Per Acre across Methods and Samples: California



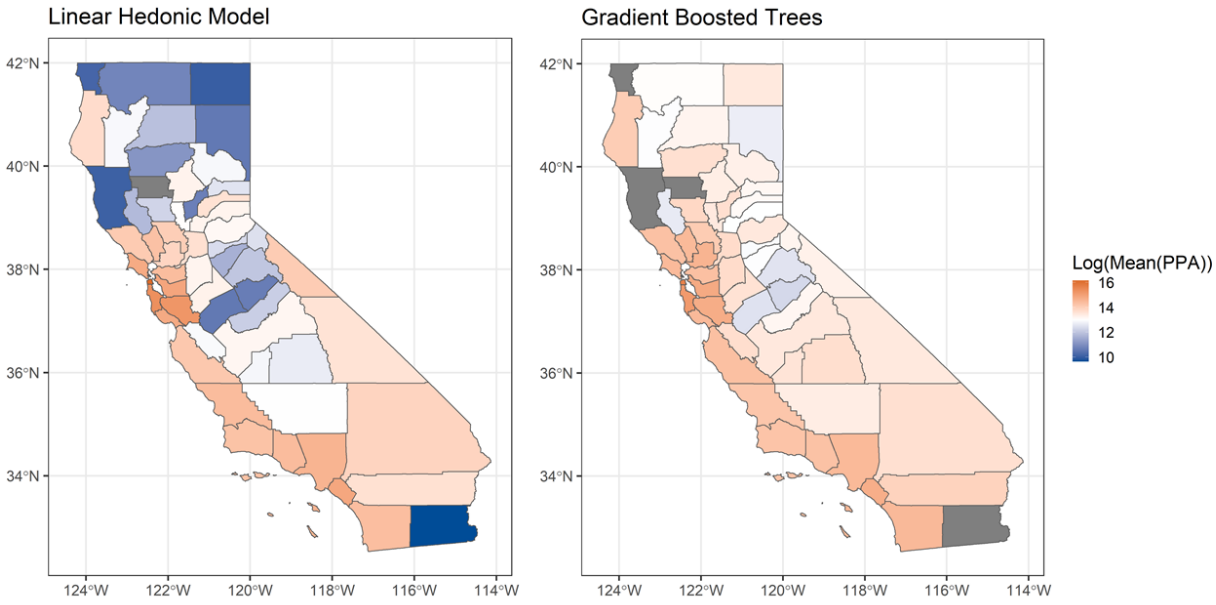
Note: Here we have plotted the time series of price per acre for the state of California. The dashed line indicates price per acre derived completely from the sales set. The solid line represents the price per acre derived from the assessment set. The differences in both the GBT and LH models across the different sets is due to a compositional change between houses that were on the market versus the universe of houses in California. The grey solid line for 2012-2015 represents estimates provided by the FHFA research in Davis et al. (2021).

Figure 7: Land to Price Ratio: California



Note: Here we have plotted the time series of the land-to-price ratio. This is derived by dividing the estimated land value for each property by its predicted price and then averaging overall properties within the year. Again, the dashed line indicates a ratio derived from the sales set and the solid lines indicate ratios which originate with the assessment set. Estimates provided by the FHFA in Davis et al. (2021) for 2012-2015 are represented by the grey solid line.

Figure 8: County Level Price-per-Acre Variation: California



Note: Here we have plotted the average price per acre (log) by county in California to highlight the geographic variation produced by the two methods. On the left is the linear hedonic model which has a much wider distribution of predicted sales prices for the assessment set. This leads to a larger difference in price-per-acre by county as compared to the Gradient-Boosted-Trees method with data driven clustering.

6. Discussion

The results from this paper show that a machine learning (ML) approach to land valuation, using gradient boosting trees and *k*-means clustering techniques, provides tangible advantages over a more traditional linear hedonic method, like the one used by Wentland et al. (2020) or Kuminoff and Pope (2013), for example. Specifically, we find that the ML method substantially reduces the RMSE of the model prediction by as much as 75% on average, resulting in a markedly more accurate model fit. Because the hedonic approach depends on decoupling land value and structure value from this model, the ML land value is less encumbered by overall model-specific error, often yielding very different results. Moreover, because we employ a *k*-means clustering approach to geographic-based fixed effects, we document a new avenue for future researchers for avoiding the pitfalls of overfitting and arbitrary decisions to exclude certain observations from the sample because of the small-N fixed effects problem. We show that the *k*-means approach creates clusters

of comparable properties or "comparables" in ways that substantially reduce the within-cluster variation of key characteristics (which market participants use as shorthand to compare homes like bedrooms and bathrooms) to a far greater extent than traditional geographic-based fixed effects like census tracts or block groups.

An additional benefit of this research is that we show a proof-of-concept demonstration that an ML approach like the one we employ here can be used to produce land value results at any geographic level, including state or national estimates of land value, provided that the data covered a this domain. Wentland et al. (2020) and Nolte et al (2021) document a number of drawbacks to the Zillow ZTRAX dataset for the purposes of land valuation at a national level, with the main drawback being the limited availability of sale price data in some states. However, given that this is the "Era of Big Data," we are hopeful that proprietary sources will adapt to be able to fill these gaps. If this were the case, we can easily extend the ML approach here, which we applied to single family residential land as a pilot, more broadly to all types of land (as in Wentland et al. (2020)), building a national estimate of land value from national data. Alternatively, we could use this data as a national sample in order to establish more precise land leverage (land-to-price ratio) estimates, which we could then apply to existing values in the accounts for total real estate values that include both structure and land value. The possibilities that this kind of data brings to the fore, with land valuation and other applications, are virtually endless; but, as we show in this paper, ML approaches can provide helpful tools in shaping this data into meaningful statistics.

References

Abraham, K.G., Jarmin, R.S., Moyer, B. and Shapiro, M.D., 2019. Introduction to" Big Data for 21st Century Economic Statistics". In *Big Data for 21st Century Economic Statistics*. University of Chicago Press.

Albouy, D., Ehrlich, G. and Shin, M., 2018. Metropolitan land values. *Review of Economics and Statistics*, *100*(3), pp.454-466.

Barr, J., Smith, F.H. and Kulkarni, S.J., 2018. What's Manhattan worth? A land values index from 1950 to 2014. *Regional Science and Urban Economics*, 70, pp.1-19.

Bencure, J.C., Tripathi, N.K., Miyazaki, H., Ninsawat, S. and Kim, S.M., 2019. Development of an innovative land valuation model (iLVM) for mass appraisal application in sub-urban areas using AHP: an integration of theoretical and practical approaches. *Sustainability*, *11*(13), p.3731.

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

Boyd, J., K. Bagstad, J. C. Ingram, C. Shapiro, J. Adkins, C. F. Casey, C. Duke, P. Glynn, E. Goldman, M. Grasso, J. Hass, J. Johnson, G. Lange, J. Matuszak, A. Miller, K. Oleson, S. Posner, C. Rhodes, F. Soulard, M. Vardon, F. Villa, B. Voigt, S. Wentland, "The Natural Capital Accounting Opportunity: Let's Really Do the Numbers," *BioScience*, Volume 68, Issue 12, December 2018, Pages 940–943.

Burnett-Issacs, K., Huang, N., and Diewert, W. E. (2016). *Developing Land and Structure Price Indexes for Ottawa Condominium Apartments*. Discussion Paper 16-09, Vancouver School of Economics, University of British Columbia, Vancouver, BC, Canada.

Chen, Jeffrey C., Edward A. Rubin, and Gary J. Cornwall. 2021. Data Science for Public Policy. New York: Springer.

Combes, P.P., Duranton, G. and Gobillon, L., 2019. The costs of agglomeration: House and land prices in French cities. *Review of Economic Studies*, 86(4), pp.1556-1589.

Coomes, O.T., MacDonald, G.K. and de Waroux, Y.L.P., 2018. Geospatial land price data: a public good for global change science and policy. *BioScience*, *68*(7), pp.481-484.

Davis, M.A., 2009. The price and quantity of land by legal form of organization in the United States. *Regional Science and Urban Economics*, *39*(3), pp.350-359.

Davis, M. and J. Heathcote. 2007. The price and quantity of residential land in the United States, *Journal of Monetary Economics*, 54, issue 8, p. 2595-2620, https://EconPapers.repec.org/RePEc:eee:moneco:v:54:y:2007:i:8:p:2595-2620.

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

Davis, M. and M. Palumbo. 2008, The price of residential land in large US cities, *Journal of Urban Economics*, 63, issue 1, p. 352-384, https://EconPapers.repec.org/RePEc:eee:juecon:v:63:y:2008:i:1:p:352-384.

Davis, M.A., Larson, W.D., Oliner, S.D. and Shui, J., 2021. The price of residential land for counties, ZIP codes, and census tracts in the United States. *Journal of Monetary Economics*, *118*, pp.413-431.

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

Duan, N., 1983. Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association*, 78(383), pp.605-610.

Dye, R. F. and D.P. McMillen. 2007. Teardowns and Land Values in the Chicago Metropolitan Area. Journal of Urban Economics, 61, 45-63. http://dx.doi.org/10.1016/j.jue.2006.06.003

European Union / OECD (2015): Eurostat-OECD compilation guide on land estimation. Luxembourg: Publications Office of the European Union. https://ec.europa.eu/eurostat/documents/3859598/6893405/KS-GQ-14-012-EN-N.pdf

Federal Reserve (2019), Financial Accounts of the United States - Table Z.1 Current release March 7, 2019 (2018 Q4 release), Balance Sheet Tables B.101, B.103, B.104. https://www.federalreserve.gov/releases/z1/current/default.htm

Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." *Annals of Statistics* (2001): 1189-1232.

Friedman, Jerome H. "Stochastic gradient boosting." *Computational Statistics and Data Analysis.* (2002): 367-378.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)." *The Annals of Statistics* 28, no. 2 (2000): 337-407.

Gindelsky, M., Moulton, J. and Wentland, S.A., 2019. Valuing housing services in the era of big data: A user cost approach leveraging Zillow microdata. In *Big Data for 21st Century Economic Statistics*. University of Chicago Press.

Gong, Y. and de Haan, J., 2018. Accounting for Spatial Variation of Land Prices in Hedonic Imputation House Price Indices: a Semi-Parametric Approach. *Journal of Official Statistics*, *34*(3), pp.695-720.

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

Hastie, Trevor, Trevor Hastie, Robert Tibshirani, and J. H. Friedman. 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer.

Kuhn, Max, and Kjell Johnson. 2013. *Applied predictive modeling*. Vol. 26. New York: Springer.

Kuminoff, N.V. and J.C. Pope. 2013. The Value of Residential Land and Structures during the Great Housing Boom and Bust. Land Economics, Feb 2013, vol. 89, issue 1, p. 1-29.

Larson, W. 2015. "New Estimates of Value of Land of the United States," BEA Working Paper. https://www.bea.gov/research/papers/2015/new-estimates-value-land-united-states

Larson, W. and Shui, J., 2020. Land Valuation using Public Records and Kriging: Implications for Land versus Property Taxation in Cities (No. 20-01). Federal Housing Finance Agency.

Moyer, B.C. and Dunn, A., 2020. Measuring the Gross Domestic Product (GDP): The Ultimate Data Science Project. *Harvard Data Science Review*, 2(1).

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

Nolte, C., 2020. High-resolution land value maps reveal underestimation of conservation costs in the United States. *Proceedings of the National Academy of Sciences*, 117(47), pp.29577-29583.

Nolte, C., Boyle, K.J., Chaudhry, A.M., Clapp, C.M., Guignet, D., Hennighausen, H., Kushner, I., Liao, Y., Mamun, S., Pollack, A. and Richardson, J., 2021. Studying the Impacts of Environmental Amenities and Hazards with Nationwide Property Data: Best Data Practices for Interpretable and Reproducible Analyses. Available at SSRN.

Rambaldi, A.N., McAllister, R.R. and Fletcher, C.S., 2015. *Decoupling land values in residential property prices: smoothing methods for hedonic imputed price indices* (No. 549). University of Queensland, School of Economics.

Rambaldi, A.N. and M.S. Tan, 2019. Land Value Indices and The Land Leverage Hypothesis in Residential Housing. International Conference on Real Estate Statistics. 20-22 February 2019, Luxembourg. https://www.real-estate-statistics.eu/

Rosenthal, S. and R. Helsley, (1994), Redevelopment and the Urban Land Price Gradient, *Journal of Urban Economics*, 35, issue 2, p. 182-200, https://EconPapers.repec.org/RePEc:eee:juecon:v:35:y:1994:i:2:p:182-200.

Sirmans, C.F. and Slade, B.A., 2012. National transaction-based land price indices. *Journal of Real Estate Finance and Economics*, *45*(4), pp.829-845.

Smith, Adam. 1776. *The Wealth of Nations*. Edited by Edwin Cannan, 1904. Reprint edition 1937. New York, Modern Library.

Turner, M.A., Haughwout, A. and Van Der Klaauw, W., 2014. Land use regulation and welfare. *Econometrica*, 82(4), pp.1341-1403.

Wentland, S.A., Ancona, Z.H., Bagstad, K.J., Boyd, J., Hass, J.L., Gindelsky, M. and Moulton, J.G., 2020. Accounting for land in the United States: Integrating physical land cover, land use, and monetary valuation. *Ecosystem Services*, 46, p.101178.

Appendix

Table A: Price Per Acre by State – Hedonic Model

State	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Alabama	67,661	78,513	86,874	88,607	76,781	71,774	71,092	64,551	67,540	65,213	64,163	64,165
Arizona	398,642	544,188	642,711	619,836	500,136	395,098	392,107	344,928	378,479	432,774	471,420	517,849
Arkansas	90,970	97,960	103,444	93,959	81,429	69,211	72,112	74,583	77,373	80,708	75,537	74,973
California	1,762,692	2,059,673	2,116,231	1,947,605	1,436,938	1,292,281	1,381,109	1,313,976	1,359,117	1,646,481	1,862,106	2,067,508
Colorado	492,702	502,330	511,508	496,006	448,326	427,116	414,150	391,712	420,608	466,331	519,570	594,002
Connecticut	257,919	286,823	297,033	288,134	259,639	231,017	226,030	213,646	199,691	200,130	202,641	205,407
Delaware	271,490	329,041	382,684	613,781	700,412	477,382	337,815	316,748	303,100	331,729	349,729	387,283
Florida	519,032	679,603	731,325	629,178	421,916	305,563	278,548	261,095	278,669	332,573	370,733	418,236
Georgia	115,291	125,182	132,463	133,211	116,804	100,510	96,199	83,090	82,112	91,964	100,754	110,681
Illinois	536,571	596,855	635,814	616,801	506,599	394,952	363,983	315,946	301,244	324,055	353,991	376,594
Iowa	191,428	196,960	196,316	194,469	185,582	177,544	173,480	173,370	180,891	189,175	197,752	201,115
Kentucky	115,900	104,630	107,965	111,682	109,237	107,550	104,313	103,887	104,093	108,200	107,360	113,845
Maryland	395,660	481,084	517,256	511,203	448,615	394,129	366,544	349,831	350,905	367,872	379,356	380,870
Massachusetts	398,578	428,722	413,147	384,358	340,322	318,401	319,083	296,774	286,129	301,805	275,564	278,672
Michigan	286,350	285,110	263,584	217,658	159,483	137,701	150,729	129,357	116,554	175,932	188,313	169,004
Minnesota	331,458	350,504	340,743	322,543	272,508	251,239	255,711	226,246	228,707	244,454	262,161	273,991
Missouri	173,347	173,418	173,113	165,284	147,895	141,474	144,134	139,625	140,634	134,505	132,496	94,282
Nebraska	264,487	279,979	186,202	185,650	177,147	212,641	206,701	210,979	224,583	228,676	240,974	251,973
Nevada	943,303	1,156,121	1,044,245	930,945	680,932	525,338	502,971	431,045	437,633	628,813	660,709	696,212
New Hampshire	174,058	186,307	184,634	176,263	154,271	137,567	135,385	129,026	134,599	135,543	139,380	151,855
New Jersey	672,518	777,788	809,282	775,769	685,257	615,577	598,401	565,547	539,139	542,656	556,822	574,668
New York	386,557	421,050	448,867	444,223	424,686	384,568	368,058	347,981	346,194	353,270	357,956	378,649
North Carolina	109,939	125,363	138,999	148,649	140,097	130,240	117,448	113,275	113,399	123,537	129,361	137,632
Ohio	169,961	172,666	167,761	152,688	129,094	123,177	117,232	104,956	103,673	107,089	113,919	116,451
Oklahoma	100,037	123,140	142,610	151,300	149,652	144,395	131,493	130,004	142,291	146,970	160,168	161,673
Oregon	372,910	443,761	520,317	542,400	502,488	437,272	416,797	386,070	383,821	420,665	447,893	492,129
Pennsylvania	242,815	264,246	275,702	281,454	263,843	281,536	291,700	281,052	271,790	262,057	257,530	278,184
Rhode Island	401,351	452,521	449,080	417,715	356,544	306,963	297,839	303,865	392,583	381,315	319,768	327,818
South Carolina	117,790	130,176	129,535	118,855	108,091	97,610	89,280	77,748	89,361	99,819	105,366	103,367
Tennessee	80,613	82,731	88,102	86,087	84,269	78,009	76,208	71,232	69,491	76,206	82,084	92,676
Virginia	223,624	255,320	265,429	255,846	220,918	205,046	204,573	191,783	198,547	205,669	212,670	215,294
Washington	401,048	470,249	545,507	563,033	514,253	447,977	426,711	383,135	387,486	409,254	442,843	472,462
West Virginia	68,819	76,289	84,422	89,413	84,461	82,602	81,006	79,074	77,910	80,470	81,089	95,027
Wisconsin	352,054	411,437	431,035	427,441	382,661	347,877	335,229	310,976	304,808	316,626	323,281	322,151

Source: Zillow ZTRAX
Notes: Estimated price per acre are calculated as the total land value from the hedonic model divided by the total SFR acreage in a given state/year.

Appendix

Table B: Price Per Acre by State – Machine Learning Model

State	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Alabama	94,265	102,652	104,488	97,990	*	83,809	*	*	*	*	*	82,245
Arizona	353,318	499,882	592,405	568,575	436,841	351,782	314,388	283,735	316,855	346,377	379,111	422,865
Arkansas	269,177	73,593	80,486	*	68,802	68,102	65,684	63,423	65,813	61,419	65,688	72,657
California	1,232,957	1,453,698	1,529,099	1,489,088	954,586	810,719	911,471	830,229	941,995	1,191,358	1,381,994	1,547,576
Colorado	565,486	550,610	599,142	600,011	528,381	520,717	530,411	516,321	559,906	601,071	687,517	784,546
Connecticut	410,203	457,122	469,992	471,484	440,271	401,069	384,486	382,568	365,322	375,834	379,580	375,653
Delaware	298,429	362,680	352,300	379,788	380,914	347,825	317,348	318,742	319,311	327,386	328,871	*
Florida	357,946	447,000	481,620	459,532	348,339	233,498	221,837	212,457	198,239	250,881	267,669	322,637
Georgia	135,455	144,426	142,886	138,554	122,100	111,929	106,552	91,659	92,204	115,762	124,891	128,513
Illinois	607,710	675,773	697,098	683,757	563,145	462,828	462,286	403,300	407,266	441,702	470,530	503,335
Iowa	198,662	207,107	204,456	200,681	216,137	204,740	215,219	201,497	218,723	218,237	212,443	216,896
Kentucky	169,505	147,233	138,651	144,928	136,747	133,067	121,428	117,285	119,968	136,950	137,344	130,172
Maryland	406,561	525,680	555,829	550,915	489,141	428,631	396,524	366,312	392,529	410,887	420,427	433,635
Massachusetts	539,641	571,560	557,598	523,262	486,990	464,625	468,657	429,509	367,506	373,239	399,859	393,210
Michigan	249,116	244,522	230,351	205,537	158,276	157,121	148,120	125,951	122,151	148,415	158,589	169,893
Minnesota	386,191	424,104	402,261	393,631	337,420	295,134	292,710	266,717	284,641	323,919	347,568	361,478
Missouri	209,870	228,538	234,794	215,469	216,339	218,162	193,171	184,848	197,657	185,177	194,183	161,381
Nebraska	293,646	247,626	220,211	224,023	222,268	225,605	227,306	202,115	225,760	238,153	228,365	238,218
Nevada	730,447	790,921	814,112	729,788	537,406	399,103	360,973	324,971	320,454	387,599	453,916	522,347
New Hampshire	212,893	227,609	229,242	207,513	185,927	169,749	174,852	157,583	168,083	172,323	189,562	205,041
New Jersey	534,488	608,002	581,551	573,602	520,569	498,640	451,397	437,928	423,983	400,542	*	444,126
New York	384,033	432,691	468,417	465,997	449,116	395,086	417,583	379,165	406,437	396,020	426,327	429,378
North Carolina	117,506	142,710	144,416	142,179	133,018	129,243	116,109	116,404	122,294	132,242	130,346	134,495
Ohio	168,021	175,262	158,721	151,761	137,580	133,907	120,843	118,344	108,176	111,783	117,575	124,793
Oklahoma	109,160	120,339	126,426	121,945	113,308	111,614	109,842	120,965	109,847	105,871	118,078	107,029
Oregon	415,979	490,205	573,937	585,527	526,985	472,079	435,657	416,207	422,523	458,817	506,153	565,826
Pennsylvania	239,448	262,641	294,350	299,038	317,931	278,538	278,874	220,409	208,998	213,633	212,325	226,530
Rhode Island	460,908	513,766	526,490	519,857	431,134	362,398	387,249	347,527	361,055	364,555	372,273	392,897
South Carolina	132,221	144,548	148,209	148,552	128,650	132,533	127,558	119,406	122,562	127,218	130,433	148,404
Tennessee	77,276	82,921	91,886	87,147	78,886	74,772	78,788	70,746	73,435	73,527	77,994	86,823
Virginia	318,225	377,755	371,783	354,321	327,633	328,487	294,722	300,704	*	316,203	328,599	346,144
Washington	418,180	561,043	746,451	580,712	518,697	439,725	427,594	390,423	*	446,029	465,635	548,575
West Virginia	146,952	139,262	144,533	148,925	139,760	134,649	118,294	119,936	119,331	123,197	127,453	165,410
Wisconsin	411,739	422,002	402,166	410,443	371,308	358,068	347,862	287,756	281,325	324,789	324,466	343,195

Source: Zillow ZTRAX

Notes: Estimated price per acre are calculated as the total land value from the machine learning model divided by the total SFR acreage in a given state/year. * represents omitted, outlier estimates that are based on small cells or erroneous data (as the ML approach was not culled with the same filter as the linear hedonic estimates).

Appendix

Table C: Land Value to Price Ratio by State – Hedonic Model

State	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Alabama	51%	43%	42%	40%	37%	36%	36%	36%	35%	30%	32%	33%
Arizona	51%	53%	54%	56%	58%	58%	57%	55%	54%	52%	54%	57%
Arkansas	58%	64%	58%	51%	46%	45%	58%	58%	63%	55%	58%	40%
California	65%	65%	65%	65%	63%	63%	62%	61%	60%	60%	60%	61%
Colorado	57%	57%	55%	54%	52%	50%	49%	48%	49%	49%	50%	51%
Connecticut	50%	50%	50%	49%	47%	45%	44%	43%	41%	40%	41%	41%
Delaware	35%	62%	60%	61%	62%	65%	60%	61%	58%	58%	59%	61%
Florida	53%	54%	54%	53%	50%	47%	45%	45%	44%	46%	45%	46%
Georgia	40%	48%	46%	45%	44%	43%	42%	41%	39%	39%	40%	40%
Illinois	55%	58%	58%	56%	53%	50%	47%	44%	43%	42%	42%	43%
Iowa	42%	61%	56%	54%	52%	50%	49%	49%	48%	48%	48%	47%
Kentucky	48%	43%	46%	48%	44%	43%	43%	42%	40%	40%	38%	39%
Maryland	59%	60%	60%	66%	60%	59%	58%	56%	55%	54%	54%	54%
Massachusetts	55%	55%	54%	53%	51%	50%	50%	48%	47%	41%	56%	59%
Michigan	66%	63%	60%	56%	53%	49%	46%	43%	43%	44%	43%	41%
Minnesota	51%	52%	50%	51%	50%	51%	50%	47%	44%	44%	45%	45%
Missouri	43%	46%	52%	44%	46%	45%	46%	50%	49%	46%	46%	44%
Nebraska	51%	51%	49%	46%	43%	48%	47%	47%	47%	48%	48%	47%
Nevada	73%	71%	61%	60%	60%	62%	59%	56%	56%	64%	61%	60%
New Hampshire	57%	57%	56%	56%	54%	53%	52%	52%	51%	49%	49%	50%
New Jersey	64%	65%	64%	64%	62%	62%	59%	58%	60%	58%	51%	61%
New York	64%	64%	64%	63%	64%	62%	61%	60%	59%	58%	58%	58%
North Carolina	40%	43%	44%	45%	44%	43%	40%	39%	39%	40%	41%	41%
Ohio	58%	58%	58%	56%	54%	52%	49%	46%	44%	43%	42%	41%
Oklahoma	34%	46%	51%	52%	50%	48%	45%	46%	49%	48%	49%	45%
Oregon	56%	59%	59%	59%	59%	58%	58%	58%	56%	54%	53%	54%
Pennsylvania	62%	57%	58%	58%	57%	58%	59%	59%	56%	54%	53%	53%
Rhode Island	56%	57%	57%	56%	55%	54%	52%	54%	53%	51%	49%	49%
South Carolina	48%	51%	47%	42%	39%	38%	35%	34%	36%	38%	36%	35%
Tennessee	46%	43%	43%	41%	42%	41%	41%	39%	37%	39%	39%	41%
Virginia	25%	27%	27%	28%	50%	43%	43%	42%	43%	41%	38%	39%
Washington	59%	59%	59%	57%	56%	55%	56%	54%	53%	51%	51%	49%
West Virginia	44%	48%	51%	55%	53%	53%	40%	39%	36%	39%	37%	38%
Wisconsin	85%	68%	67%	67%	64%	62%	60%	59%	57%	56%	54%	52%

Source: Zillow ZTRAX

Notes: Estimated land value to price ratio are calculated as the total land value from the hedonic model divided by the total predicted price in a given state/year.

Appendix

Table D: Land Value to Price Ratio by State – Machine Learning Model

State	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Alabama	62%	63%	59%	56%	*	54%	*	*	*	*	*	46%
Arizona	58%	68%	63%	62%	62%	61%	58%	60%	61%	59%	58%	*
Arkansas	57%	55%	52%	51%	49%	42%	43%	42%	37%	42%	41%	46%
California	51%	53%	49%	47%	45%	46%	45%	43%	44%	50%	49%	47%
Colorado	67%	70%	68%	66%	60%	58%	60%	56%	58%	58%	56%	59%
Connecticut	65%	64%	59%	57%	60%	58%	62%	58%	60%	57%	53%	51%
Delaware	66%	56%	50%	53%	50%	50%	45%	44%	44%	48%	47%	43%
Florida	64%	67%	65%	66%	65%	64%	62%	59%	62%	62%	62%	63%
Georgia	78%	77%	76%	74%	75%	76%	75%	72%	77%	82%	82%	80%
Illinois	64%	61%	58%	56%	52%	60%	55%	49%	49%	49%	48%	48%
Iowa	65%	69%	65%	65%	64%	63%	62%	59%	59%	62%	63%	60%
Kentucky	53%	60%	61%	56%	60%	63%	57%	57%	60%	56%	58%	68%
Maryland	74%	68%	56%	58%	58%	53%	54%	49%	52%	52%	47%	46%
Massachusetts	57%	52%	51%	50%	50%	48%	45%	45%	43%	43%	45%	48%
Michigan	71%	71%	71%	66%	65%	66%	68%	63%	65%	64%	67%	69%
Minnesota	55%	55%	50%	50%	48%	52%	47%	48%	47%	43%	65%	45%
Missouri	64%	65%	67%	66%	67%	63%	67%	61%	66%	62%	63%	62%
Nebraska	44%	49%	47%	44%	43%	44%	41%	42%	44%	46%	44%	42%
Nevada	60%	62%	56%	57%	57%	58%	52%	53%	47%	46%	45%	45%
New Hampshire	48%	51%	53%	53%	52%	53%	49%	48%	48%	45%	46%	48%
New Jersey	46%	48%	48%	44%	40%	39%	39%	43%	38%	36%	37%	34%
New York	67%	68%	69%	67%	65%	65%	63%	65%	64%	62%	64%	66%
North Carolina	60%	62%	66%	66%	71%	67%	68%	57%	54%	54%	53%	55%
Ohio	66%	67%	69%	70%	67%	65%	69%	67%	67%	65%	64%	65%
Oklahoma	59%	58%	54%	52%	47%	53%	52%	51%	52%	50%	48%	50%
Oregon	46%	45%	47%	43%	40%	40%	43%	39%	40%	38%	38%	39%
Pennsylvania	*	57%	56%	*	48%	51%	48%	48%	47%	43%	46%	45%
Rhode Island	61%	62%	58%	58%	61%	65%	58%	61%	*	58%	58%	60%
South Carolina	63%	73%	84%	62%	60%	57%	57%	56%	*	57%	55%	59%
Tennessee	72%	69%	67%	68%	65%	65%	58%	59%	56%	57%	56%	62%
Virginia	70%	72%	66%	66%	63%	64%	63%	55%	53%	58%	56%	56%
Washington	63%	63%	64%	66%	58%	56%	59%	56%	59%	61%	63%	65%
West Virginia	64%	63%	65%	64%	60%	62%	62%	64%	64%	64%	66%	69%
Wisconsin	81%	81%	81%	80%	80%	81%	77%	79%	77%	79%	79%	79%

Source: Zillow ZTRAX

Notes: Estimated land value to price ratio are calculated as the total land value from the machine learning model divided by the total predicted price in a given state/year. * represents omitted, outlier estimates that are based on small cells or erroneous data (as the ML approach was not culled with the same filter as the linear hedonic estimates).