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

**Authors** Scott Wentland and Gary Cornwall, U.S. Bureau of Economic Analysis;
Jeremy Moulton, University of North Carolina Chapel-Hill[1]

**Contact** Scott.Wentland@bea.gov

**Date** June 2023

**Abstract** This paper develops a new method for valuing land, a key asset on a nation's balance sheet. The method first employs an unsupervised machine learning method, kmeans clustering, to discretize unobserved heterogeneity, which we then combine with a supervised learning algorithm, gradient boosted trees (GBT), to obtain property-level price predictions and estimates of the land component. Our initial results from a large national dataset show this approach routinely outperforms hedonic regression methods (as used by the U.K.'s Office for National Statistics, for example) in out-of-sample price predictions. To exploit the best of both methods, we further explore a composite approach using model stacking, finding it outperforms all methods in out-of-sample tests and a benchmark test against nearby vacant land sales. In an application, we value residential, commercial, industrial, and agricultural land for the entire contiguous U.S. from 2006-2015. The results offer new insights into valuation and demonstrate how a unified method can build national and subnational estimates of land value from detailed, parcel-level data. We discuss further applications to economic policy and the property valuation literature more generally.

**Keywords** Land valuation, machine learning, non-financial assets, environmental-economic accounting, hedonic valuation, Big Data, national accounts, economic policy

**JEL Code** E01, Q56, Q24, R14, C80, G12, C01, C38, C55

[1]The authorship order was determined to be reverse alphabetical order arbitrarily by a coin toss. Gary lost. Disclaimer: Data provided by Zillow through the Zillow Transaction and Assessment Dataset (ZTRAX). More information on accessing the data can be found at zillow.com/ztrax. The results and opinions are those of the author(s) and do not reflect the position of Zillow Group. We thank John Clapp, Erwin Diewert, Arman Khachiyan, Will Larson, Thies Lindenthal, Joe Nichols, Christoph Nolte, Tusan Nguyen, and Yun Zhou for helpful comments and suggestions on prior drafts. We would also like to thank participants at the following conferences/seminars for invaluable feedback: 2023 FHFA Seminar Series, 2023 American Real Estate Society Annual Meeting, 2022 USDA-ERS Brown Bag Seminar Series, 2022 University of Cambridge Department of Land Economy Seminar, 2022 ASSA/SGE Annual Meeting, 2021 Federal Committee on Statistical Methodology Annual Conference, 2021 North American Meeting of the Urban Economics Association, and the 2021 North American Meetings of the Regional Science Association.

*"... 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)

*"The heart of the SNA describes how labour, capital and natural resources including land are used to produce goods and services. These goods and services are used for the three economic activities recognized in the SNA, production, consumption and accumulation."*

— U.N. System of National Accounts 2008, §3.19

# 1. Introduction

Land is, quite literally, a foundational asset for any economy. Extending back to at least Smith (1776), economists have long understood that for households and firms land is both a key input to production and a substantial asset. Prior research has estimated that, in aggregate, not only is land a considerable asset in its own right (*e.g.*, Davis (2009), Larson (2015), Wentland et al. (2020)), 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. This literature has argued that the infamous *housing* boom and bust of the 2000s is often mischaracterized, instead suggesting that it would be more aptly called a *land* boom and bust (Davis et al., 2017, 2021), citing evidence that much of the fluctuation in the value of residential property can be attributed to the underlying price of land (see also Kuminoff and Pope (2013)). Given both its economic significance, policy relevance,[2] 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 or property valuation more generally?

[2]Beyond its macroeconomic significance as an asset, from an economic policy standpoint taxation of land has been put forth by economists since George (1884) as one of the more efficient forms of taxation. As a result, most countries around the world have either a land tax or property tax on real estate, although one of the frequently cited issues of land taxation is the inherent difficulty in disentangling land from structure value McMillen and Zabel (2022). In the U.S., property tax revenue (which land is a large component of this tax base) was $615 billion in 2020, which was larger than corporate income taxes ($276 billion), for example. Source: BEA NIPA Tables 1.12 & 3.5 - `https://apps.bea.gov/iTable/index_nipa.cfm`.

This paper shows ML methods can indeed do both. We cultivate a unique approach that pairs ML methods, kmeans clustering and gradient boosted trees (GBT), with a linear hedonic regression method to estimate land value at scale, generating granular parcel-level estimates of residential, commercial, industrial, and agricultural land that we then use to construct national and subnational values. We find a composite approach outperforms more conventional hedonic methods (as used by the U.K.'s Office for National Statistics, for example) when we benchmark predicted values from these models against observable market prices (*i.e.,* in out-of-sample tests predicting transaction prices of single-family properties and in comparing land values of developed land near vacant land sales). While the primary contribution of this paper is methodological, we employ the new method in an application to provide a proof-of-concept detailing how it can scale from microdata to macroeconomic statistics. Specifically, we construct an aggregate valuation of private land for the entire contiguous U.S. from 2006-2015, leveraging microdata from Zillow (ZTRAX) containing millions of property transactions and detailed information corresponding to each property.

Using this new approach, we find private land in the contiguous U.S. was worth an estimated $24 trillion in 2015, or approximately $19,050 per acre, with large variation by geography and land-use category. For example, we find that residential land in dense urban areas of the Pacific census division (as defined by the U.S. Census Bureau) was worth an average of $3,966,805 per acre in 2015, while agricultural land was worth an estimated $12,275 per acre on average. The U.S. national time-series dynamics we observe with residential land are consistent with procyclical land value fluctuations over the infamous boom-bust-recovery period of the last two decades. However, we find a great deal of variation in these dynamics, as some regions and land-use types experienced much milder cycles (*i.e.,* much flatter peak to trough) and the timing of this trough also varied by region over this period. In addition, we use the composite approach to estimate land leverage, or the percentage of the total (*i.e.,* land + structure) property value comprised by land value for each category and region. Although more stable than land prices, we find land leverage can still have sizable fluctuations over time and can vary substantially across regions. On average, land leverage was around 2/3 for single-family properties in New England during most years of our sample, for example, but is only about 1/4 to 1/3 in the South Atlantic census division. Finally, our pilot estimates show how this method could help produce a set of land accounts consistent with international statistical standards set out in the System of Environmental-Economic Accounting Central Framework (SEEA-CF) and incorporate land onto the balance sheet as a distinct "non-produced, non-financial asset" prescribed by the U.N. System of National Accounts (SNA). Because this method is constructed using property-level data, an enduring potential benefit of this project is that the method can be easily taken off the shelf by researchers, policy analysts, appraisers, local tax assessors, statistical agencies, central banks, and others with local microdata (using our code or their own adaptation of it), generating a through line from micro to macro data.

### 1.1. Hedonic vs. cost-based approaches to land valuation, trends in accounting, and fair market value — why now?

Conceptually, one novelty of our proposed approach is that it exemplifies how $21^{st}$ century methods and data may have caught up to the accounting standards' valuation principles. When valuing products, services, and assets, national economic accounts follow accounting standards set out by the SNA (2008) and SEEA-CF (2012), which prioritize valuation via observable transactions in the market to obtain their fair market value whenever possible. When measuring Gross Domestic Product, for example, not all products and services will have observable market values (*e.g.*, government expenditures) or sufficient transaction data, necessitating alternatives and even cost-based methods to proxy for market value. Indeed, a central issue with valuing a nonfinancial asset like land at market value on the balance sheet is that these assets are often heterogeneous (differing in size, location, quality, etc.), sold infrequently, and commonly bundled with another asset like a structure in the transaction. In these circumstances, when similar but not identical assets are sold at market prices, the SNA recommends methods using market transactions of similar products/assets, which should then make adjustments for quality and other quantifiable differences.[3] This broadly characterizes the hedonic approach to land valuation, which exploits variation in market prices across heterogeneous assets to estimate the marginal value of each property characteristic with the intent to separate the value of land and structure components.[4] Recently, the U.K.'s Office for National Statistics (ONS) has applied a hedonic method to valuing land in the U.K., adapting the hedonic model used for their House Price Index (HPI) for this purpose (Johannsson and Nguyen, 2022).

The hedonic approach has gained traction in the $21^{st}$ century primarily as detailed transaction and property characteristic data has made it more tractable to do at a national scale.[5] Pragmatically, cost-based approaches are still deployed in the national accounts as a response to lack of sufficiently high

[3]Regarding valuation using market prices: "When market prices for transactions are not observable, valuation according to market-price-equivalents provides an approximation to market prices. In such cases, market prices of the same or similar items when such prices exist will provide a good basis for applying the principle of market prices. Generally, market prices should be taken from the markets where the same or similar items are traded currently in sufficient numbers and in similar circumstances. If there is no appropriate market in which a particular good or service is currently traded, the valuation of a transaction involving that good or service may be derived from the market prices of similar goods and services by making adjustments for quality and other differences." §3.123, SNA 2008.

[4]Regarding hedonic valuation methods: "A more general and powerful method of dealing with changes in quality is to make use of estimates from hedonic regression equations. Hedonic regression equations relate the observed market prices of different models to certain measurable price-determining characteristics. Provided sufficiently many differentiated models are on sale at the same time, the estimated regression equation can be used to determine by how much prices vary in relation to each of the characteristics or to predict the prices of models with different mixes of characteristics that are not actually on sale in the period in question... The technique has also been used for housing by regressing house prices (or rents) on characteristics such as area of floor space, number of rooms or location..." §15.83-84, SNA 2008.

[5]In their description of their HPI model, ONS's notes that it relies on transaction and property data going back to 1995 in England and Wales, but as recent as 2004 and 2005 for Scotland and Northern Ireland, respectively. The U.S. data shares a similar constraint, as we have a great deal of data going back to the mid-1990s, but it is most complete in the early to mid-2000s for all regions. We return to this discussion in the data section below.

quality market price data for certain assets and products. Hence, a common approach to land valuation by both national statistical offices and the broader academic literature has been some variation of the residual method, which usually rely on construction cost data (which is often high-quality and readily available). For properties containing a structure, this method first estimates the value of the structure based on construction cost data, then subtracts this value from the total property value, either transacted or estimated value, where the value left over (i.e., the residual) is the value of the land component.[6] This method should be most accurate for new properties, when the structure was just built, and the market land value should track the residual derived from this replacement cost relatively well (McMillen and Zabel, 2022). As the property ages, residual approaches also account for depreciation of the structure over time to approximate the land value of a new property for the non-new housing stock. This is why the residual method is also called the "depreciated cost" approach.[7] More recently, Clapp and Lindenthal (2022) and others have developed hybrid variations of this approach, which allow for a more nuanced decoupling of structure and land value, and the evolution of these values separately over time.[8]

There are both practical and conceptual challenges with the residual method and other methods that relies on cost-derived estimates for the national accounts. From a practical standpoint, Statistics Denmark, for example, had employed a variation of the residual approach where they used construction cost data in combination with a depreciation schedule of the structure to estimate land value. The 2009 OECD manual, *Measuring Capital*, relays the following anecdote based on the Danish experience and how market dynamics can produce strange results using this method:

> *"During the recession in the late 1980s, real estate prices declined whereas acquisition prices for new buildings increased as shown in the figure below. In the PIM estimations of the net stock of buildings, it was assumed that the prices of existing buildings (for a given age) followed the prices of new buildings which increased steadily. With decreasing prices for real estate and increasing prices for buildings, the residual – the value of land – declined. However, the decrease was so large that the value of land becomes negative for some years during the recession. A negative value for land is not an economically meaningful result."*
> – OECD 2009, p.162

[6] The OECD Manual Measuring Capital (2009) notes: "Information on the price and quantity of structures and buildings without land is often more readily available when data on the stock of dwellings uses the perpetual inventory method with investment series for structures and buildings from the national accounts. Investment surveys on construction permit relatively easy collection of information on the value of structures excluding land." (p. 155)

[7] This includes the pioneering work by Davis and Heathcote (2007), Davis and Palumbo (2008), and more recent variations using finer, more detailed data like Davis et al. (2021), among numerous others using a variation of the residual/depreciation cost method. For a more comprehensive review of the residual method and related literature, see also Clapp et al. (2021).

[8] McMillen and Zabel (2022) describe the Clapp and Lindenthal (2022) approach as a hybrid between a bundled goods approach and the residual method, which is related conceptually to the "land share" approach of Bourassa and Hoesli (2022). For more detail, see McMillen and Zabel's (2022) summary of the of the *Journal of Housing Economics'* special issue on land valuation where these and other methods are showcased.

Indeed, McMillen and Zabel (2022) noted that, "this somewhat embarrassing outcome is not uncommon" (p. 4), as researchers have made various adjustments to either the residual method itself, or have imposed an arbitrary floor on its value.[9] From a conceptual standpoint, an additional drawback of residual methods is that accounting principles in the SNA recommend valuation of market-price-equivalents, generally favoring fair market value over historic cost-based accounting methods (at least for non-new properties). While cost-based methods are widely used in the national economic accounts, as noted above, these methods are primarily used if "no appropriate market" exists or market price data is not available for a particular good, service, or asset, where cost can be used as a less-than-ideal substitute for market price.[10]

The rise of "Big Data" and machine learning in the $21^{st}$ century has changed this pragmatic dynamic, not only in the national accounts, but in private sector accounting as well. There has been a longstanding debate on fair value accounting versus historic cost accounting, highlighting trade-offs with each approach and circumstances where one may be preferred over the other (Jaijairam et al., 2013). U.S. GAAP standards, for example, had historically recommended cost accounting methods for non-financial assets like property, plants, and equipment (PPE), while IFRS standards follow fair market value methods. However, Warren Jr. et al. (2015) observe that Big Data is facilitating the convergence of the two standards towards fair market value, "with Big Data ... helping to construct a global accounting regime with fair value accounting as a key cornerstone ... [as] it will enhance measurement process through new forms of evidence to support management's accounting for transactions" (p. 404). Recent literature has also shown how machine learning has been adapted to improve firm-level accounting estimates and why it is important for accounting going forward Ding et al. (2020); Bertomeu (2020). Broadly, national accounts and official statistics are on board with this trend and recognize now as an inflection point.[11] In their introduction to the NBER volume title *Big Data for Twenty-First-Century Economic Statistics*, Abraham et al. (2019) conveyed precisely this sentiment:

> *"The message of the papers in this volume is that Big Data are ripe for incorporation into the production of official statistics. In contrast to the situation two decades ago, modern data science methods for using Big Data have advanced sufficiently to make the more systematic incorporation of these data into official statistics feasible."*
> – Abraham et al. (2019), p. 3

[9] In response to the Denmark scenario, the OECD (2009) manual has a nod to a the hedonic approach and methods relying primarily on market prices in the second-hand market: "a way forward would be to use asset prices from the second-hand market, combined with quality characteristics of transacted real estate... [and] this is a very difficult task, but might be necessary if reliable and consistent estimates for the value of buildings, land and real properties should be produced." p. 163.

[10] Regarding cost-based alternatives: "If there is no appropriate market from which the value of a particular non-monetary flow or stock item can be taken by analogy, its valuation may be derived from prices that are established in less closely related markets. Ultimately, some goods and services can only be valued by the amount that it would cost to produce them currently..." §3.135, SNA 2008. We will return to a related cost conceptual issue in section 2.

[11] See also Moyer and Dunn (2020) for a discussion of Big Data and data science applications in the national economic accounts.

## 1.2. Contributions to the valuation literature and national accounts

This paper makes several contributions to the academic literature. Methodologically, this is the first paper to apply a two-step machine learning approach, gradient boosting trees paired with kmeans clustering, to land valuation on a national scale in a way that conceptually tracks a hedonic method. As noted above, one reason the hedonic method is used by ONS in the U.K. is that the richness of their data is now reasonably well-suited to explain much of the variation in property prices, as the characteristics in their regression model explain about 80% of the variation in home prices (Johannsson and Nguyen, 2022). We find that our ML approach delivers a number of advantages over land valuation using hedonic models like those used by ONS or in the academic literature like Kuminoff and Pope (2013), Diewert et al. (2015), Wentland et al. (2020), among numerous others. Using the Zillow ZTRAX data, we compare the performance of multiple models (a variation of the ONS model, the Wentland et al. (2020) model, and our GBT model) when predicting prices of single-family residential properties, the land-use category where the data is richest and our results are most comparable to the literature. On average, the ML approach models the sales price outcome substantially better than linear (OLS) hedonic models, as evidenced by a significant reduction in root-mean-square-error and mean-absolute-error in the out-of-sample test set. Since the structure and land value are essentially decomposed from the coefficients that predict sale price, this is an important benchmark given that any error in the model's sales price prediction may be reflected in the error of its components, and thus the land value estimate.

Second, our adaption of an unsupervised machine learning approach, kmeans clustering, as an alternative to geographic-specific fixed effects provides a novel path forward for hedonic modeling and property valuation more generally. As we describe in more detail later in the paper, it transforms the modeling process into one that more closely mirrors the approach of professional appraisers. In the U.S., for instance, mortgage underwriters generally require a professional appraisal of the property. These appraisals typically assess the market value of the property based on nearby, comparable properties (or "comps"). Location is obviously an important determinant of market value, as any real estate agent will repeat three times; however, it is not the only determinant of value or source of time-invariant heterogeneity assessed in the appraisal. Appraisers will often draw on comps located further away (outside a ZIP code or census tract or some other small unit researchers use for spatial fixed effects) if the other characteristics of the property are a better match in terms of predicting price. The kmeans clustering algorithm mimics this approach more systematically by allowing the data to generate groups of comparable properties that balance this location (latitude, longitude) trade-off and minimize variation in certain property characteristics (*e.g.,* bedrooms, bathrooms, etc.).

In the context of property valuation, we show the benefit of the kmeans clustering approach is two-fold. First, it provides a tractable alternative to the well-known "thin cell problem" in urban economics, where granular spatial fixed effects, such as census tract, often contain too few observations in a given time period. Recent work by Davis et al. (2021), for example, employ ZIP code fixed effects to account for geographic-specific heterogeneity, where the initial data includes $18,322$ ZIP codes nationally. Other studies use census tracts or block groups with even finer spatial granularity. However, there is a well-known trade-off 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 or ZIP code, resulting in many of the usual overfitting problems and sensitivity to within-group outlier sales.[12] We show that a kmeans clustering approach can generate larger, yet more relevant, groups for predicting price, as it minimizes variation in characteristics of the property and location much like an appraiser (*e.g.,* a cluster of predominantly 4-bedroom, 3-bathroom homes in a location that crosses several census tracts or ZIP codes will have far less variation in bedrooms and bathrooms than a given single census tract or ZIP code). This allows our ML model to incorporate 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/MAE statistics) by grouping more homogeneous homes across greater dimensions than geography alone. Second, this approach systematically discretizes the unobserved heterogeneity akin to the approach described by Bonhomme et al. (2022), mimicking the practice of professional appraisers by grouping more homogenous homes across observables. As Bonhomme et al. (2022) pioneered in an analogous application, the unobserved heterogeneity is highly related to these characteristics over which we are generating clusters, where clustering allows us to "discret[ize] heterogeneity as a dimension reduction device rather than as a substantive assumption about population unobservables" (p. 2).[13] In an era where Big Data and large numbers of fixed effects (and interactions) are the norm in applied microeconomics more generally, a key contribution of this paper is to demonstrate an early empirical application of the Bonhomme et al. (2022) concept that likely has broader applications outside of land valuation.

Third, we show how a simple model stacking method adapted from the forecasting literature can further improve on the ML approach by creating a weighted composite measure. In our comparison of methods, we identify circumstances where the linear hedonic method proposed by Wentland et al. (2020) performed relatively well at predicting property prices (*i.e.*, the ML method is not strictly better in every single circumstance). Thus, we find the combined linear hedonic (HD) and GBT composite measure outperforms all methods in out-of-sample tests predicting property prices of single-family homes. As an out-of-sample test, we compare all methods' predictions against sales of nearby vacant land. The GBT-alone and composite measures of land value track the value of vacant land sales closest when the vacant land market is most robust, at the national peak of new housing starts during our sample period, which is when we would expect vacant land values to be most representative of nearby developed properties.[14]

[12] Davis et al. (2021)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 final results.

[13] For example, a cluster of predominantly 4-bedroom, 3-bathroom homes in a given location is more likely to have similar upgraded kitchen features (*i.e.*, unobservable to researchers with this kind of data or "drive-by appraisers" with similar information) than homes within a given census tract composed of a hodgepodge of 2-bedroom, 1.5 bathroom homes and 5-bedroom, 3-bathroom homes. We return to this point in our description of the kmeans clustering methodology below.

[14] Our goal was to develop a method steered by transaction prices of developed land to avoid selection bias issues of using vacant land alone. However, our predictions should still should be reasonably in line with vacant land prices when we compare apples-with-apples and vacant land markets are most robust. Both developed and vacant properties are relevant

Given its performance in out-of-sample tests, we then deploy the composite method more broadly to estimate land value for the contiguous U.S. (and 9 census divisions) for residential, commercial, industrial, and agricultural land for each year over a decade (2006-2015).

Finally, in addition to the contributions to the academic literature outlined above, this research is highly relevant to economic measurement and policy, as accounting for land on the national balance sheet is a notable gap in the national economic accounts for most countries. Although land is clearly a significant asset, there is virtually no available information directly quantifying the aggregate value of land itself in the official accounts (either in the U.S. or the vast majority of countries around the world).[15] 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. While Wentland et al. (2020) and ONS in the U.K. provide examples of how this might be accomplished with a linear hedonic valuation approaches, we build on this by putting forth a unifying, data-driven, composite ML method that substantially improves on prior approaches and would be replicable by any country around the world with similar property data available.

More generally, the incorporation of detailed land accounts into the national accounts is part of a broader international trend in the 21$^{st}$ century in expanding the scope of the national economic accounts to include more non-produced capital or "natural capital" that quantify the value of our natural resources (Boyd et al. (2018)), 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 2021. Yet, the U.S. does not currently produce any formal environmental-economic accounts. Given that land is an asset at the intersection of the traditional (SNA) national accounts and environmental accounts, as outlined in the System of Environmental-Economic Accounting Central Framework (SEEA-CF), valuing land presents a logical starting point for expanding the scope of what the national accounts measure in the United States. A more systematic, transparent approach to modeling can provide more confidence in the results by the public; and, if similar methods are used across countries for national accounts, then it would facilitate comparability of the resulting statistics. In the Discussion section of this paper, we return to this point, discussing potential next steps for this work in the context of the national accounts. In addition to potentially building a national

transaction prices. When vacant land transaction prices should exhibit the less selection bias, in periods when there are a lot of vacant land sales and a lot of development, the difference in our predictions and vacant land price should also be less. We find exactly this result. We return to this point in our discussion of the results comparing vacant land to predictions.

[15] Land is categorized as a "non-produced, non-financial asset" on a country's balance sheet in the SNA standard. In the U.S., the national asset balance sheet is part of the Integrated Macroeconomic Accounts, which is jointly produced by the Bureau of Economic Analysis and the Federal Reserve Board. Land accounts are also a distinct set of satellite accounts in the UN SEEA Central Framework. Prior work by Davis (2009), Larson (2015), and Wentland et al. (2020) has cultivated a number of different methods to remedy this gap. As of the writing of this draft, BEA has not adopted a particular method for land valuation, nor does BEA officially endorse any particular method at this time.

account based on these estimates for macro applications, like Davis et al. (2021), once published, we intend to make our land value estimates available at a variety of subnational levels to all who would find them useful in their research, policy-making, or other decision-making.[16]

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

## 2.1. 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 directly measures 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. Even 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.[17] Thus, these core problems (*i.e.*, the fact that not all land sells in a given period, properties are heterogeneous, and that the land that does sell is typically bundled with a structure) has spawned a deep literature that utilizes additional information to get at the underlying value of land in a more sophisticated way.

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. A recent symposium of papers by the *Journal of Housing Economics* has also included studies that estimate land value from teardowns,[18] variations of vacant

[16] 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.

[17] 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).

[18] See McMillen and Zabel (2022), which builds on a number of papers using a similar approach, including: Gedal and Ellen (2018); Munneke and Womack (2015); McMillen and O'Sullivan (2013); Dye and McMillen (2007); Munneke (1996);

land interpolations,[19] and a number of other innovative methods.[20]

These indirect approaches reasonably assume that the value of the property is the value of the bundled components of 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; Clapp and Lindenthal, 2022). For example, land may appreciate in value over time while the associated structure depreciates through wear-and-tear or consumption of fixed capital (with some exceptions and limitations, *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 $p_s S$, and the value of the plot of land, $p_l L$, can be written as:

$$V = p_S S + p_L L \tag{1}$$

where $S$ is the size of the structure, $L$ is the land area, and $p_S$ and $p_L$ are prices of a unit of $S$ and $L$ respectively. The challenge then is to best 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). Indirect methods differ primarily on how land value is decoupled from the property's total value. As we noted in the introduction above, the residual approach – or some variation thereof – is often used by both governments and academics, as they generally rely on construction or builder's costs as replacement costs (*e.g.,* Davis and Heathcote (2007); Davis and Palumbo (2008)). Other variations of this incorporate 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 by 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. However, even under circumstances where researchers have ideal cost data to pin down the cost of the structure most accurately, the key question before us is: is this the right conception of land value for the national accounts?

Rosenthal and Helsley (1994).

[19]See Albouy and Shin (2022) and Larson and Shui (2022) for original adaptations of interpolating land prices from vacant land sales using Bayesian and Kriging methods, respectively. For other innovative approaches using vacant land, see also Nolte (2020), Albouy et al. (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. We return to this point later when we discuss comparisons to vacant land value.

[20]Zabel (2022),Bourassa and Hoesli (2022), and Longhofer and Redfearn (2022) present novel variations of hedonic and land-share methods.

## 2.2. How should land be valued? A national accounts perspective

Though our review above is not exhaustive, we should acknowledge here an important takeaway from the literature: there are numerous, reasonable approaches to land valuation that exploit different types of data to get at this fundamentally difficult question. 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). However, there are two important aspects from the SNA's valuation principles that make the hedonic approach compelling over the cost-based approaches. The first, which we discussed at some length in the introduction, is the SNA's emphasis on using observable market values to the extent possible, which itself is contingent on available data. The second important aspect of SNA valuation that is relevant here, which we had not touched on in the introduction, is the idea that the SNA measures value in the market in whatever context goods, services, or assets are exchanged. The standard emphasizes that, “a market price should not necessarily be construed as equivalent to a free market price” (SNA 2008, §3.119). The context may be competitive, monopolistic, or somewhere in between – market value is what prevails in the (imperfect) markets we observe.[21]

Key assumptions underlying many cost-based approaches, however, are neoclassical assumptions about competitive markets and rational consumers, which clarify the link between construction cost, structure value, and market prices. In a competitive market, competition among builders and contractors should imply that the long-run average (economic) cost of a new structure should approximate its market value. By extension, this would suggest cost data are good proxies for market value and are broadly representative. Further, homes built on vacant lots of land would, rationally, be built to their highest and best use (HBU), as the residual land value then reflects this scenario. Clapp and Lindenthal (2022) summarize these assumptions and the residual approach being derived from the work of Alonso (1964), Muth (1969), and Mills (1972) – which is referred to as AMM theory.[22] The assumptions underlying AMM theory may very well characterize a sizable portion of the market in the U.S., but prior empirical work on construction costs in the U.S. cast doubt on the strongest form of these assumptions. Somerville (1999) and Gyourko and Saiz (2006), for example, document large differences in construction costs, and substantial heterogeneity in competitive environments, across U.S. regional markets. If these markets do not closely approximate perfectly competitive markets, then conceptually the SNA standard would not favor a valuation method that assumes a residual value derived from free market structure price across the board.

[21] Specifically, the SNA goes on to states: “that is, a market transaction should not be interpreted as occurring exclusively in a purely competitive market situation. In fact, a market transaction could take place in a monopolistic, monopsonistic, or any other market structure. Indeed, the market may be so narrow that it consists of the sole transaction of its kind between independent parties.” SNA 2008, §3.119

[22] Clapp and Lindenthal (2022) note that, “In AMM theory, land values are dependent on a structure that is built to maximize the present value of the location, *i.e.*, HBU structure. Land value at the time of new construction is a residual equal to the HBU property value less the construction costs. . . [where] construction cost equals structure value” (p. 1-2).

This is not necessarily an indictment of AMM theory or methods derived from it that focus on HBU value; on the contrary, AMM theory and these methods that try to pin down HBU value are incredibly useful for a variety of purposes. One of the explicit purposes for Clapp and Lindenthal (2022), among numerous other studies deriving residual-based land value, is to advance land valuation for the purposes of improving assessments related to taxation. A key argument for land taxes among Georgism proponents is that a tax on land incentivizes development (George, 1884), which is, in its strongest form, implicitly a HBU value concept relying on how that land might be used in an efficient market. HBU value is also highly useful for developers for similar reasons. Nevertheless, because residual/cost-based approaches rely on a concept of value that assumes land and structure to be in a more ideal state of HBU, from an SNA perspective, most common residual approaches are not ideal for the national accounts.

The hedonic approach, on the other hand, takes sale prices of properties as they are in the market, however competitive or monopolistic that market may have been that produced those prices. This approach regresses actual sale prices of properties we observe in the market on a variety of detailed characteristics of the land and structure, which yields an estimate of the market value of the structure using variation in the data from comparable structures and properties. One recent study by Rambaldi and Tan (2019) described 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 on available data. This allows for a nuanced, location-specific estimate based on observed market prices as opposed to costs.[23]

Consider a simple example. Suppose we observe three developed property sales adjacent to one another. To keep the numbers simple, one sells for \$300,000, the second sells for \$400,000, and the third property sells for \$500,000 in the same period. The first two properties sit on identical plots of land (say, 1 acre), but the square footage of the second's structure is twice as large (say, 1,000 vs. 2,000 sqft.). The third property has an identical structure as the first one (also 1,000 sqft.), but now sits instead on 2 acres. In this scenario, the hedonic model lines up with intuition. Comparing the first and third properties with identical structures, the extra acre yielded a \$200,000 higher sale price. Comparing the first and second properties, an extra 1,000 square feet of living area yielded a \$100,000 increase in sale price. Thus, a regression that explains 100% of the variation in sale prices here would simply yield these values as coefficients on square footage (in 000s of sqft) and acreage if we regressed these exact data points in a linear hedonic model. While this stylized example abstracts away from numerous complicating factors when working with real data (like location differences, time period differences, other property characteristics and market dynamics),[24] the basic intuition is that we are using variation in observed

[23]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).

[24]Note, even if there were unobservables here, like the fact that these properties may have different numbers of bathrooms

market prices and deducing the marginal value from variation in property characteristics. This method is agnostic about whether these structures were built for their "highest and best use" and simply infers what its fair market value is based on what the marginal characteristics are selling for on the market, as the property currently exists, and based on the revealed preferences of the market as we find it.

### 2.3. The hedonic approach – a baseline method suited to Big Data

Our data, which we will discuss in more detail in section 3, contains detailed information about transactions and property characteristics. Generally, this type of data is well-suited to a hedonic approach to estimate land value, as we alluded above in the discussion of its use by the U.K. ONS,[25], albeit with some well-known drawbacks. We adapt (and tweak) the hedonic approaches used in Johannsson and Nguyen (2022) and Wentland et al. (2020) to establish a baseline approach for comparison to our ML approach described later in Section 4. The hedonic model typically relies on a standard ordinary least squares regression model and is generally less intricate than more advanced techniques used by Zillow's proprietary automated valuation model, for example, or our ML variant. For residential properties we first estimate the following for each time period (3 year overlapping window) and state separately:

$$log(P) = \alpha + X\beta + D\gamma + D_x\zeta + Q\lambda + \epsilon \tag{2}$$

where $P$ is an $N \times 1$ vector of observed market prices, $X$ is an $N \times K$ matrix of characteristics which are pertinent to the development of $P$ (*e.g.,* number of bedrooms, bathrooms, garages, square footage of the living area, acreage, whether the structure has a basement, porch, etc.), $D$ is an $N \times J$ indicator matrix where $D = 1$ if $i \in j$ and $0$ otherwise where $j$ indexes the location (*e.g.,* census tract), $D_x$ is a set of interaction terms where both square footage and acreage of the parcel have been interacted with the location indicator, and finally $Q$ is an $N \times T$ indicator matrix where $Q = 1$ if $i \in t$ and $0$ otherwise.[26]

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 similar estimates using finer-level geographic fixed effects like census block groups.[27] Although this approach is intensive for processing, it allows the valuation of structure square footage and acreage to vary by a finer geography

or quality of flooring, the observable variation is sufficient to explain 100% of the variation in prices. This can occur if the unobservables are identical and highly/perfectly correlated with observables, or if the unobservables' marginal values are insignificant. We return to this idea of correlated unobserved heterogeneity in our discussion of kmeans clustering below.

[25] In addition to Wentland et al. (2020) and Kuminoff and Pope (2013) mentioned above, there are a number of other instructive hedonic studies, including but not limited to: Gong et al. (2018), Burnett-Isaacs et al. (2020), Rambaldi et al. (2015), and Diewert et al. (2015).

[26] 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 in individual states, 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. When data becomes more universally complete across states and regions, we see no reason not to expand the model to include it. However, we leave extensions to this model that exploit more variables to future work.

[27] Smaller geographic units, like block groups and blocks, have fewer sales, so the advantages of finer location controls

than typically available. This is key, as the valuation of these attributes can vary widely across areas within a state (either for demand-side reasons OR supply-side reasons due to regional variation in construction markets as described by Somerville (1999) and Gyourko and Saiz (2006), among others. For example, an additional tenth of an acre for a property in San Francisco, will be valued much differently than the same amount of space in Sacramento, which this model with interactions allows for the acreage coefficient to differ by location.[28]

For the ONS model, we simplify the hedonic model to constrain it to a narrower set of covariates found in the U.K. HPI model, which include: total rooms, total bedrooms, a binary measure of age (new/old), local socioeconomic indicators,[29] and fixed effects covering ZIP code, land use code, and year. While the data in the ZTRAX dataset does not exactly align with the U.K. model, we view this as a close approximation of how their model would perform in the U.S. if the data aligned more precisely. One might also think of it as a coarser hedonic model than in Wentland et al. (2020), but one that is still well-aligned with prior literature employing hedonic methods for this purpose.

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:

$$\tilde{lv} = \exp^{\alpha + D\hat{\gamma} + D_{x,acreage}\hat{\zeta} + Q\hat{\lambda}} \times \exp^{.5\nu^2} \tag{3}$$

where $\tilde{lv}$ is a parcel level land value prediction, and $\nu$ is the root-mean-square-error of the in-sample fit for the regression outlined in equation 2. Because we used relatively fine (spatially small) location fixed effects, all time-invariant local amenities and environmental benefits within each tract (and within the period of estimation) 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 (to the extent they are capitalized here).

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

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. We have also explored a variety of other specifications to improve model fit and predictions, including a linear dependent variable, where sale price is not logged.

[28]This interactive fixed effect approach is commonly used in the hedonic valuation literature for housing and land (*e.g.*, Kuminoff and Pope (2013) and Wentland et al. (2020)). 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.

[29]The socioeconomic indicators used in the U.K. model are somewhat U.K.-specific, so we used available local characteristics in the U.S. as a close proxy: we have replaced it with measures of affluence from the Socioeconomic Status and Demographic Characteristics of ZIP Code Tabulation Areas. See https://www.openicpsr.org/openicpsr/project/120462/version/V1/view for more information.

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 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 over 30 sales in the three year window for each model. If this condition was not met within a given tract and period, we estimated models for the remaining census tracts using higher-level county (FIPS) level geographic fixed effects.[30] 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 noisy predictions for areas with sales marginally above these thresholds, we cull any outliers above the $1^{st}$ percentile or above the $99^{th}$ percentile. These adjustments ensured that model coefficients were not driven by erroneous or mis-measured data, small samples, or outliers.[31] 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 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.[32]

## 2.4. Extending the hedonic model beyond single-family residential

Due to the relatively smaller number of sales for non-SFR properties, we take a few deviations from the approach described above when we extend the hedonic model to other residential, commercial, industrial, and agricultural properties. We thus estimated the models separately by census division (*i.e.*, a group of states) rather than a single state and used a five-year rather than three-year window. This allows the coefficient estimates of the property characteristics to be derived from more data in order to reduce the influence of outliers. We also specified the regression to just use census tract location fixed effects (or county if tract is missing) rather than the two separate models (census tract or county) as we used for SFR properties. The non-SFR residential properties use the same hedonic controls as the SFR, while the commercial and industrial regressions are limited to only age, square footage (interacted with location), and logged acreage (interacted with location) due to the limited number of relevant characteristics

[30] We lump all remaining counties together under one location fixed effect that do not have enough sales (after removing all census tracts that met the sales threshold) within the time period.

[31] One potential issue with the hedonic approach, or any prediction-based multivariate method, is multicollinearity. 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 may 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. We examined the correlations between acreage and square footage of the structure in our data in untabulated analysis. Somewhat surprisingly, we found the correlation was not particularly high in the U.S. (usually falling within 0.2-0.4).

[32] For out-of-sample tests comparing predicted prices to actual prices, we use only 80% of the sample, by census tract, and hold out a random selection of 20%. We return to this point below in our discussion of the out-of-sample tests.

available in our data.[33] The agricultural land models are estimated using county fixed effects and include square footage, logged acreage, and an indicator for a structure, which is also estimated using the entire Census Division. We return to a discussion of the data limitations for these land use types at the end of the paper.

# 3. Data description

This section describes the property-specific microdata we use to generate national estimates from millions of data points spanning much of the U.S., along with a number of choices made to clean or restrict the data for producing higher quality estimates. Specifically, we use the Zillow Transaction and Assessment Dataset (ZTRAX) that was made available to researchers in academia and government for a limited period of time (through September 2023).[34] It contains market transaction data as well as a large set of individual property characteristics for sales recorded in local tax assessor's data.[35] Coverage is generally representative of the United States' national market, initially comprising 374 million detailed transaction records across more 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.[36] 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.

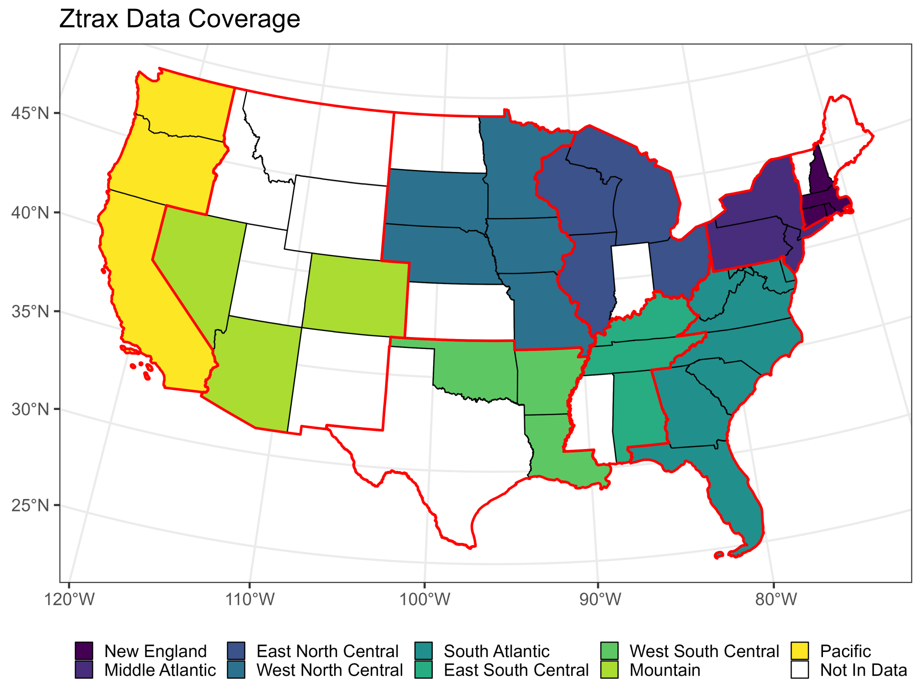
[33]There is a small, but growing literature on valuation of commercial land and developing price indices for non-SFR properties like condominiums/apartments, which draw from data sources with different (and in some cases a richer set of) property characteristics for these land-use types or take an alternative empirical approach. For recent examples, see Nichols et al. (2013); Diewert and Shimizu (2015); Diewert et al. (2015); Diewert and Shimizu (2017a,b) and Burnett-Isaacs et al. (2020).

[34]As we discuss further in Section 7 below, there are a number of limitation to this dataset, and some of them are straightforward to remedy. Our employer has purchased data from another data provider, Black Knight, that will allow us to extend this analysis beyond ZTRAX's current discontinuation date in 2023. Long term availability of national microdata is important for replicating this method in the future.

[35]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. Non-proprietary code used to generate the results for this paper is available upon request to the authors.

[36]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 price data but are missing substantial data in the ZTRAX vintage we use for this study. We omit these states as well. However, our employer has recently purchased supplemental data from Black Knight that contains sale prices and other relevant information for property transactions in all of these states, which we may use for filling these data gaps.

**Figure 1. ZTRAX Sale Price Coverage in the Continental U.S.**




**Note: Some states do not require public disclosure of sale prices, resulting in missing price data.**

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. 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.[37] A key aspect of this dataset 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 dataset from Zillow comes in a somewhat raw form. We therefore 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 to identify these transactions), 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

[37] 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).

might signal a misclassification of a property, where a building with 40 bathrooms may actually be a commercial office building. Hence, we dropped extreme values for price and home characteristics for our estimates, which is a common practice for recent research using this particular data.[38]

We also culled the regression samples to limit the influence of outliers on the coefficients. We retain properties with acreage above zero and below 5,000 acres. We use land use codes and acreage to classify properties into the land types based on detailed land-use codes as described in Wentland et al. (2020): dense urban, urban, single-family, rural, commercial, industrial, and agricultural. We initially removed properties that had extreme values in absolute terms, like a structure smaller than 50 square feet (agricultural land does not use this constraint) and a price lower than $1,000 or above $30 million. We then culled by price at the 2.5$^{th}$ and 97.5$^{th}$ percentile by year, land group, and county. We culled homes with square footage (a home's living area) below 2.5$^{th}$ or above the 97.5$^{th}$ 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 2.5$^{th}$ 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 properties (aside from agricultural) 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. 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. One novelty of this time period is that it offers great variation in time-series dynamics, as it includes intense periods of boom, bust, and recovery in the U.S. real estate market.

Finally, given that we will be comparing methods in out-of-sample tests, we split our data into an 80% training sample and a 20% test sample. This split is stratified by census tract to ensure that no census tract is left out of either sample by chance. Overall, our training set includes 26,415,128 observations over the sample period while the test set contains 6,608,198 observations. It is important to note that the training sample is used by all model structures (*e.g., our hedonic model, the ONS model, etc.*) to estimate the necessary parameters and performance is judged based on price predictions for the test set via an appropriate loss function (*e.g.,* root-mean-square error, mean-absolute-error, etc.). Summary statistics for the training, test, and assessment set are outlined in Table 1.

[38]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), Wentland et al. (2020), Moulton et al. (2018)). 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 this paper makes. See also Gindelsky et al. (2023) for a comparison and usage of common variables in ZTRAX versus the American Community Survey (ACS).

**Table 1. ZTRAX Summary Statistics – Single-Family Residential**

| | Variables | 1st Qu. | Median | Mean | 3rd Qu. | St. Dev. |
|---|---|---|---|---|---|---|
| | Acreage | 0.2 | 0.2 | 0.4 | 0.5 | 0.40 |
| | Square Footage | 1256 | 1680 | 1923.8 | 2345 | 925.80 |
| | Total Rooms | 0 | 6 | 4.5 | 6.7 | 3.20 |
| | Total Bedrooms | 3 | 3 | 2.8 | 3.2 | 1.20 |
| **Assessment Set** | Total Baths | 1 | 2 | 1.9 | 2.4 | 0.80 |
| *45,516,219 Observations* | Number of Stories | 1 | 1 | 1.3 | 2 | 0.50 |
| | Porch | 0 | 0 | 0.3 | 1 | 0.50 |
| | Basement | 0 | 0 | 0.2 | 0 | 0.40 |
| | Year Built | 1952 | 1971 | 1968.7 | 1993 | 29.30 |
| | Price | 114,000 | 190,000 | 246,101.5 | 313,491.5 | 391,543.60 |
| | Acreage | 0.1 | 0.2 | 0.3 | 0.3 | 0.40 |
| | Square Footage | 1306 | 1740 | 1961.3 | 2398 | 896.30 |
| | Total Rooms | 0 | 6 | 4.5 | 6.6 | 3.20 |
| **Sales Training Set** | Total Bedrooms | 3 | 3 | 2.8 | 3.4 | 1.20 |
| *26,415,128 Observations* | Total Baths | 1.5 | 2 | 2 | 2.5 | 0.80 |
| | Number of Stories | 1 | 1 | 1.3 | 2 | 0.50 |
| | Porch | 0 | 0 | 0.3 | 1 | 0.50 |
| | Basement | 0 | 0 | 0.2 | 0 | 0.40 |
| | Year Built | 1956 | 1981 | 1975.6 | 2002 | 29.30 |
| | Price | 114,000 | 190,000 | 245,962.3 | 313,500 | 356,691.30 |
| | Acreage | 0.1 | 0.2 | 0.3 | 0.3 | 0.40 |
| | Square Footage | 1307 | 1741 | 1962.1 | 2400 | 896.60 |
| | Total Rooms | 0 | 6 | 4.5 | 6.6 | 3.20 |
| **Sales Test Set** | Total Bedrooms | 3 | 3 | 2.8 | 3.4 | 1.20 |
| *6,608,198 Observations* | Total Baths | 1.5 | 2 | 2 | 2.5 | 0.80 |
| | Number of Stories | 1 | 1 | 1.3 | 2 | 0.50 |
| | Porch | 0 | 0 | 0.3 | 1 | 0.50 |
| | Basement | 0 | 0 | 0.2 | 0 | 0.40 |
| | Year Built | 1956 | 1981 | 1975.6 | 2002 | 29.30 |
| **Vacant Land Transactions** | Price | 15,000 | 42,500 | 120,590 | 134,900 | 205,777.30 |
| *1,035,517 Observations* | Acreage | 0.23 | 0.33 | 0.59 | 0.91 | 0.54 |

**Note: The data available for this project initially covers 36 of 48 states in the continental United States. The transactions we use occur between 2002 and 2016 and account for more than $8,000,000,000,000 in market value. The average number of transactions per year is just over 2.1 million. The Assessment data is a snapshot of all single family houses between 2014 and 2016. The sales test set was created by sampling randomly without replacement 20%, by census tract, of the overall sales transactions.**

# 4. Methodology – adapting machine learning for hedonic valuation

## 4.1. Unobserved heterogeneity, kmeans clustering, and the appraiser's problem

The hedonic valuation of land begins with predicting the overall price of the property from its components (land + structure). By breaking down the price of a property into its individual components, we can evaluate the impact of marginal changes to the property (*e.g.*, adding a bathroom) and ultimately back out the price of the property without its structure components based on variation in market prices and property characteristics, as discussed above. Hence, given that the hedonic method begins with a prediction model, our initial motivation for exploring an ML method is to evaluate whether we can gain new insights into valuing land for each property by making more accurate predictions with a method more tailored for prediction accuracy (i.e, measuring accuracy using a loss function such as root-mean-squared-error or root-mean-absolute error). Moreover, in order to avoid what is often called the "black box critique" of ML methods, our goal across the proceeding three subsections is to describe our methodology and the underlying mechanisms in sufficient depth so as to allow for replicability (along with our final code to be made public upon publication) and facilitate feedback for further improvement/refinement of the approach.

An important concern about this approach, however, is that the data describing the property may not be exhaustive and there could be relevant unobserved differences in properties that buyers/sellers can observe but we as modelers/appraisers cannot. That is, we have a rich data set of observable characteristics for each property (square footage, number of stories, acreage, etc.), but there is a great deal that we do not observe about the property. Does the home in question have a high-quality roof, or one in need of repair? Does the home have updated appliances? A more modern floor plan or architectural style? High-quality flooring or windows? The answers to these are often not available to modelers (although, the data continue to improve over time) and many professional appraisers. This is a fundamental problem for appraisers, particularly for "drive-by appraisals" or summary appraisal where the appraiser does not enter the home and evaluates comparable properties using a Sales Comparison Approach.[39] The information available to them is often far more limited than the information available to the actual buyers and sellers setting market prices, given that a host of unobservables likely play a role in negotiations and price-setting in property markets.

To address this unobserved heterogeneity problem, we employ a similar two-step group fixed-effects process to Bonhomme et al. (2022), which addresses an analogous issue of unobservables associated with individuals in the labor market. Their paper uses a kmeans algorithm as a classification step to group individuals in the labor market whose latent types, which need not be discrete, are most similar.

[39] Regulations on the time period of recent sales and what constitutes a "comparable sale" vary by state, lender, and over time. Typically, the time period contains a one-year look back and mandatory justifications if comparable characteristics are not available. Fannie Mae offers a detailed introduction to the Sales Comparison Approach appraisals in their Selling Guide (p. 596-601).

Similarly, our assumption (and implicitly the assumption of appraisers using comps as the primary basis for valuation) is that homes with similar observables have a stable distribution of unobservables within a given group – mirroring the logic of Bonhomme et al. (2022). For example, four-bedroom homes with more than two bathrooms are likely to have a similar distribution of unobservables, which differ from two-bedroom, one-bathroom homes. Just as Bonhomme et al. (2022) used group classifications as fixed effects in subsequent models to account for unobserved heterogeneity across individuals, we group properties by relevant observables in the assessment set, using these in lieu of geographic-based fixed effects.

In a way, the two-step group fixed-effects method also mirrors the way in which appraisers evaluate nearby sales of comparable properties by grouping along relevant observables. For example, if the subject house is a four-bedroom, three-bathroom house with a two-car garage sitting on a quarter of an acre of land, an appraiser will identify similar nearby properties that have sold in the recent past and correct for differences among both observables and unobservables (to our data set). While an appraiser will attempt to stay within the same neighborhood and school district, often because this is a large driver of home prices, there is no guarantee that sufficient nearby sales exist. Additionally, there is no guarantee that the appraisal process will respect other geographic boundaries used my modelers to proxy for location such as census tracts or block-groups, either.[40]

To provide context, we have included an [abbreviated] appraisal report from a local brokerage in Maryland in Figure A the appendix. Note that, along observable dimensions (to our data) such as square footage, number of bedrooms, and number of bathrooms, the three comparable properties are largely similar to the subject home. However, there is heterogeneity in other characteristics such as the quality of construction, the presence of a fence around the the property line, and level of finish in below grade (*e.g.*, basement) floors. Moreover, in this example, while two of the properties are nearby, less than $0.35$ miles, the third comparable is over two miles away; and, two out of the three comparables are in a different location with respect to the geopolitical boundary. By using the kmeans algorithm to group structures based on the observable characteristics we are essentially allowing these fixed effects to act as an appraiser grouping comps on observables, albeit with a substantially larger set of comparables and a more systematic approach. Like Bonhomme et al. (2022), the appraiser is assuming this process discretizes the remaining unobservable heterogeneity by creating relatively homogeneous set of houses along observable dimensions.

Thus, we cluster the assessment data in ZTRAX, which includes the near universe of properties, over a multi-dimensional space that includes the following characteristics: location (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. Each cluster represents the universe of houses within a state that an appraiser would consider "comparable" to

[40]For example, Fannie Mae's (2022) Selling Guide describes this comps process in some detail on p. 598-601 of the Guide.

a subject home. We then apply these time-invariant clusters from the housing stock to those houses that transacted on the market. Returning to the unobserved heterogeneity, this process also then assumes that the distribution of unobserved characteristics is relatively stable within the cluster and thus the influence of that heterogeneity on our predictions will be minimized.

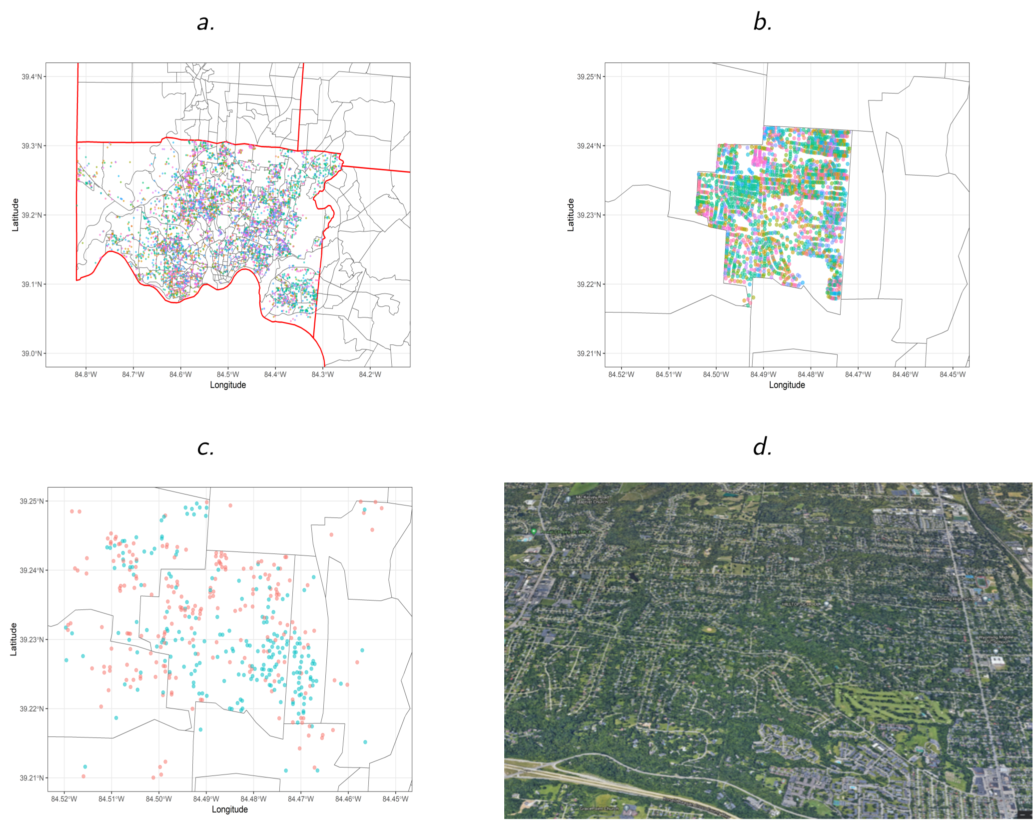
**Table 2. Distribution of Within Cluster versus Within Tract Standard Deviations: Ohio Example**

| | | Minimum | First Quartile | Median | Mean | Third Quartile | Max |
|---|---|---|---|---|---|---|---|
| Generated Clusters | Sales Price (Sales) | 29,053 | 42,518 | 52,993 | 61,832 | 75,187 | 229,704 |
| | Square Footage | 160.10 | 326.90 | 407.60 | 424.20 | 511.90 | 769.40 |
| | Acreage | 0.129 | 0.383 | 0.464 | 0.454 | 0.529 | 0.722 |
| | Bedrooms | 0.000 | 0.139 | 0.224 | 0.258 | 0.393 | 0.762 |
| | Bathrooms | 0.000 | 0.202 | 0.285 | 0.297 | 0.401 | 0.808 |
| Census Tract | Sales Price (Sales) | 4,734 | 32,825 | 43,512 | 50,556 | 58,743 | 582,193 |
| | Square Footage | 40.31 | 378.33 | 479.64 | 479.85 | 569.97 | 1058.54 |
| | Acreage | 0.002 | 0.100 | 0.248 | 0.298 | 0.492 | 0.918 |
| | Bedrooms | 0.000 | 0.596 | 0.658 | 0.655 | 0.719 | 1.528 |
| | Bathrooms | 0.000 | 0.435 | 0.549 | 0.542 | 0..638 | 1.226 |

**Note: The values above represent a single state (Ohio) to illustrate the reduction in within cluster variation over observable hedonic elements. In Ohio there are 370 clusters with an average of 7,142 (5,821) homes per cluster. There are 2,947 census tracts in Ohio, each of which as an average of 913 (852) homes. Each cluster can be thought of as a set of comparables that could be used by an appraiser to establish market value.**

Much like Bonhomme et al. (2022), which used this process as a dimensionality reduction device, this process reduces the dimensionality of our price prediction problem. Within a single cluster, this process generates a more homogeneous set of homes along the clustering variables. Note that we do not include square footage or acreage in the clustering algorithm and, as a result, the majority of within cluster price variation is loaded on to these other continuous variables of interest. To be plainer, if all houses within a cluster have the same observable characteristics and a stable, mean zero set of unobservable characteristics, then price variation within cluster comes from the size of the plot (acreage) and the size of the structure (square footage) as well as any location effects. In Ohio, for example, the average standard deviation on the number of bedrooms within our kmeans constructed clusters is less than 40% of that of the within cluster standard deviation of census tracts (0.393 versus 0.655), which we show in Table 2. The maximum standard deviation in our constructed clusters is less than half that of census tracts. Meanwhile, the size of the clusters is significantly larger than a tract with an average of 7,142 (5,821) homes per cluster as compared to an average tract size of 913 (852) homes. This reduction varies across states with some states such as California having as little as 10% of the within cluster variation as compared to that of the census tracts.

**Figure 2. Clustering and Boundaries: An Example**




**Note: To illustrate our clustering approach, we have plotted a sample of single family residences in Hamilton County Ohio (Figure 2a). In Figure 2b we pick a single tract within that county which represents the suburb of Wyoming and plot all single family residences in the tract color coded by cluster assignment. Figure 2c isolates two clusters in that area showing how clusters can cross geo-political boundaries such as census tracts. Finally, Figure 2d is a satellite image of that census tract showing the borders are created by artificial landmarks (roads) which may or may not make sense as a delineation in a fixed effect hedonic type model.**

Finally, Figure 2 illustrates how these clusters can cross the common geospatial boundaries used in the hedonic real estate and urban economics literature. In Figure 2a we show a sample of the properties in Hamilton County Ohio (the location of Cincinnati, Ohio) color coded by cluster. In Figure 2b we have isolated a single census tract in the suburb Wyoming. Every property is accounted for, and each is assigned a cluster which is similarly color coded. In Figure 2c we have that same census tract but isolating down to two individual clusters to show how they can cross the tract boundaries. We call your attention to Figure 2d which shows that the eastern boundary of the census tract from Figure 2b is a road, and our clustering algorithm allows for houses on one side of that road to be compared to the other; something an appraiser would almost certainly do, but could be obscured by the use of tract fixed effects.

### 4.2. Gradient boosted trees (GBT) paired with kmeans clustering – a new approach

Following Bonhomme et al. (2022), we use these data-driven fixed effects in a second stage estimation step, employing a gradient boosted trees (GBT) modeling framework to estimate the price as accurately as possible. 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), $\frac{1}{2}(y_i - f(x_i)^2)$, with gradient, $-\delta(y_i, f(x_i))/\delta(f(x_i) = y_i - f(x_i)$. 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. This differs from say a random forest which builds a number of independent trees and then averages the predictions. 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 Algorithm 1 we have outlined the generic framework of a gradient tree boosting algorithm (Friedman et al., 2000; Friedman, 2001, 2002).

**Algorithm 1** Gradient Boosting

Input:

Data, $\mathcal{D} = (X, Y)$, and a differentiable loss function, $L(y - i, F(x))$.

Initialize model with a $f_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_i^N L(y_i, \gamma)$

1. For $m = 1$ to $M$:
   (a) Compute $r_{im} = -\left[\frac{-\delta L(y_i, F(x_i))}{\delta F(x_i)}\right]_{f=f_{m-1}}$
   (b) Fit a regression tree to the target $r_{im}$ giving terminal regions $R_{jm}$ for $j = 1, \ldots, J_m$.
   (c) For $j = 1, \ldots, J_m$ compute $\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma)$
   (d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.
2. Output $\hat{f}(x) = f_M(x)$.

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

$$salesprice = f(latitutde, longitude, sqft, acreage, cluster, yearbuilt). \tag{4}$$

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 flip-side 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).

To account for properties that are not sold in a given period, we use the predictive model based on properties that are sold to project onto the near universe of properties (which is called "assessment data" in ZTRAX, as the underlying data comes from local assessors' offices) for both the linear hedonic and GBT models. This set of properties includes a large number of houses that are not typically observed on the market. This means we predict the price of homes that are sold in a given period as well as homes that may never be on the market based on these observable characteristics.

There are a couple of things we would like to highlight about this framework. First, as the tree splits along the clustering variable, any subsequent splits produce within cluster terminal nodes. For example, suppose the first split is along the cluster dimension, then any subsequent splits will be of houses that are homogeneous along the observable characteristics and the terminal node variation on the structure price will come from the square footage. Second, as the tree branches along latitude and longitude, post cluster split, it is dividing this comparable set of structures into fine grids of geography, in some cases much finer than census tracts or even blocks, in others (such as sparsely populated suburban areas) the geography may be larger than census tracts or even counties. The terminal nodes produced are relatively homogeneous structures that vary in size within a small geographic region, albeit a region which is ultimately rectangular.

Finally, similar to the hedonic approach, we treat the plot *"as if vacant"* by reducing the square footage of the structure to zero for the purposes of valuing the market value of the land. To do this, our trained model predicts the new price if the structure characteristic (sqft) is zeroed out. Since tree based algorithms do not differentiate between $sqft = 0$ and $sqft \leq 500$ we make a small correction to the structure price by predicting the change in price from increasing every property's square footage by the smallest in its cluster. The difference between this new, larger structure prediction and our original price prediction forms our within cluster correction term. The end result is our prediction of the land value and it can be written as:

$$\begin{aligned} lv_{i,t} &= \tilde{P}_{i,t|sqft=0} - \gamma_{c,t} \\ \gamma_{c,t} &= \tilde{P}_{i,t|sqft=sqft_i+min(sqft_{c,t})} - \tilde{P}_{i,t|sqft=sqft_i}, \end{aligned} \tag{5}$$

where $lv_{i,t}$ is the land value for property $i$ in time $t$, $\tilde{P}_{i,t|sqft=0}$ is the predicted price of property $i$ in time $t$ conditional upon the structure's square footage being reduced to zero, and $\gamma_{c,t}$ is the correction term applied to each $i \in c$.

To calculate the price-per-acre at a property-level, we divide the estimated land value, $lv_{i,t}$ by the observed acreage for the property. For a property with land value of $10,000$ that sits on $0.25$ acres of land this would imply a price-per-acre of $10,000\$/0.25acres = 40,000$ dollars per acre. We do this at an individual level so that we can then aggregate to any geographic level, $j$, by calculating $ppa_j = \sum_{i\in j}^{n} lv_i / \sum_{i\in j}^{n} acreage_i$. [41]

[41] Note that this measure of price-per-acre is one possible value measure and is different than say the average price-per acre, which would be calculated as $\bar{ppa}_j = n_{i\in j}^{-1} \sum_{i\in j}^{n} ppa_i$. The first is the price-per-acre of properties in the $j^{th}$ region

### 4.3. Improvement in price prediction and the case for model stacking

Recall that our raison d'être for the approaches described above is developing an "as if vacant" market value estimate for land underneath privately owned structures, as granular microdata and new methods should produce better valuations. It then begs a number of (answerable empirical) questions. First, how much better are the price predictions using this microdata? Second, how much better are the price predictions when we deviate from a traditional hedonic analysis and move to the two-staged machine learning structure outlined in the previous section? And third, if there are circumstances where one is better than the other, can we cultivate a composite approach via model stacking that predicts prices even better? We answer all three of these questions in this subsection, motivating the final method used to derive our bottomline results in the next section.

As we mentioned in the introduction, in 2022, the Office of National Statistics (ONS), the national statistical office of the United Kingdom, released new estimates of the land underlying buildings and structures.[42] For brevity, we will not cover the full model here, but simply note that the key differences are twofold:

1. The data being used is less detailed than that available in the Ztrax data set. For example the structure characteristics are limited to number of rooms, type of dwelling, a binary indicator of dwelling age (old/new), and an indicator if the buyer is a first time buyer or former owner occupier. Like our hedonic model the model used to produce these land estimates includes location (at the county or London borough level) and property use (*e.g.* fixed effects).
2. To supplement this, the ONS model includes socioeconomic indicators (known as ACORN) which are likely correlated to unobserved structure characteristics such as number of bedrooms, bathrooms, etc. Moreover, they interact this indicator with the dwelling type and first time buyer indicators.

In an effort to contextualize the improvement in price predictions, both from the richer microdata available through Ztrax and a progressively more adaptable modeling structure, we approximate the ONS model on our data and compare the out-of-sample price predictions across each model (ONS, our linear hedonic model, and the ML supported model).[43]

weighted by their relative importance (*e.g.,* smaller more expensive plots of land are more valuable than larger, cheaper land) whereas the second is the price-per-acre of the average plot in the $j^{th}$ region.

[42] See 'Improving estimates of land underlying other buildings and structures in the national balance sheet, UK: 2022' for a full accounting of the ONS methodology and release information.

[43] While the ACORN measure of neighborhood status has no analogue in the U.S. statistics we have replaced it instead with measures of affluence from the Socioeconomic Status and Demographic Characteristics of ZIP Code Tabulation Areas. See https://www.openicpsr.org/openicpsr/project/120462/version/V1/view for more information.

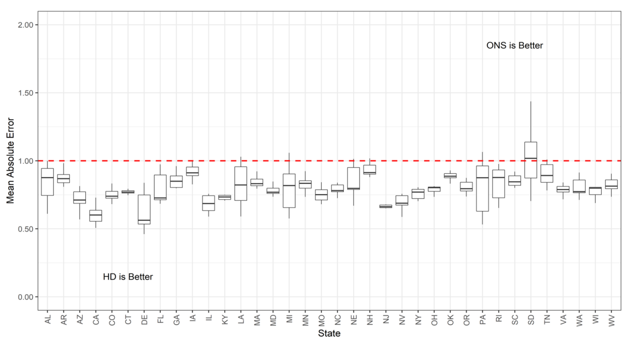
In Figure 3 we show the distribution, by state, of mean-absolute-error ratios for each model. Specifically, we calculate these ratios as:

$$\frac{\text{MAE}_a}{\text{MAE}_b} = \frac{\sum_{i,a}^{n \in test} |\tilde{p}_{i,a} - p_i|}{\sum_{i,b}^{n \in test} |\tilde{p}_{i,b} - p_i|}, \tag{6}$$

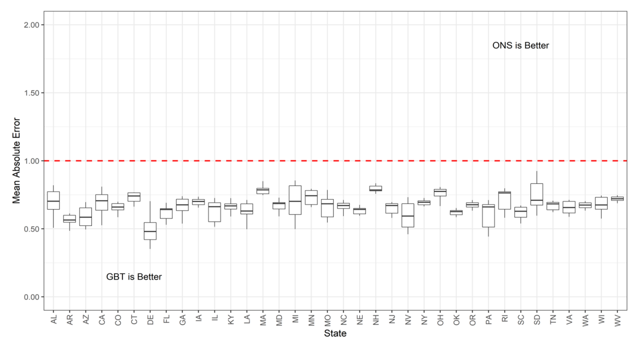
where $\tilde{p}_{i,a}$ is the predicted price for the $i^{th}$ observation for model $a$ for each state in the ZTRAX dataset for which we have sale price data. A value of less one here indicates that model $a$ has lower MAE, out-of-sample, than that of model b. For example, in Figure 3a model $a$ is the hedonic model we specified earlier in Section 2.3, and model $b$ is the approximation of the specification used by ONS.

**Figure 3. Mean Absolute Error Comparison: Distribution by State from 2004-2015**

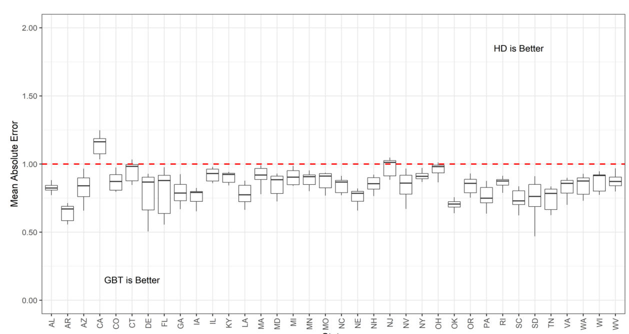
*a. Linear Hedonic Model versus ONS Model*




*b. Gradient Boosted Trees versus ONS Model*




*c. Gradient Boosted Trees versus Linear Hedonic Model*




**Note: All comparisons are made using the out-of-sample transaction set which is 20% of the sales sample by census tract. To construct these plots we take the ratio of Mean Absolute Errors for each model and plot the resulting distribution across years for each state. For example, in Panel 3a, we have divided the out-of-sample MAE of the proposed linear hedonic model by the model put forth by the U.K. Office of National Statistics. A value less than one indicates that the MAE of the linear hedonic model contained herein is lower than that of the model proposed by ONS.**

The results in Panel (a) of Figure 3 show the mean-absolute-error distributions by state over the 2004-2016 range are lower for the hedonic model we propose than the simpler ONS model, indicating that a richer data set and a more granular level of fixed-effects with appropriate interactions is likely a better predictor of overall price. There are exceptions within each state as some years may favor the ONS version of the model over that we have proposed and in a state such as South Dakota, where sales are thin and the data is very limited a coarser model (ONS) can perform better. Moving over to Panel (b), we see further improvement with all state-year distributions favoring the combination of data driven clustering and gradient boosted trees over the (approximate) ONS model across the board. However, the results from Panel (c) comparing our hedonic with GBT shows that for some states (like California) and state-year combinations (New Jersey, Ohio, and Connecticut for example) the more granular hedonic model can out-perform our machine learning approach in out-of-sample price prediction accuracy in some circumstances. The potential reasons for this are manifold. For example, in some states neighborhoods and census tracts are more homogeneous than others, limiting the value-added of kmeans clustering. Recall that we noted above California census tracts are more homogeneous across observable characteristics, as clusters there only showed a modest reduction in variance over observables. Further, being a relatively large volume market, the richness of the California sales data likely contributes to the performance of the linear hedonic method. While we leave further investigation of these differences to future research, the main takeaway from this comparison is that more granular data opens the door for improved performance of linear hedonic methods and the GBT method provides enhanced predictive accuracy in most (but not all) state-year combinations.

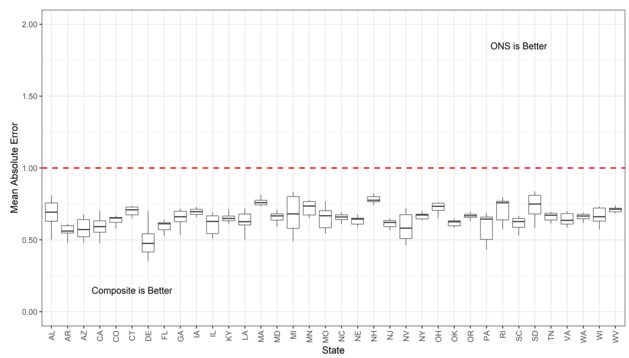
Although GBT outperforms all other models in price predictions in the vast majority of circumstances, there is still some ambiguity in which model we should prefer for a unified method estimating land value for the entire country. We thus draw on a rich literature surrounding forecast averaging (see Granger and Ramanathan (1984), Elliott and Timmermann (2004), Timmermann (2006), and Hansen (2008) among many others for examples), which argues that we do not need to choose a single model. In fact, since some state-year prices have lower out-of-sample error when using the hedonic model, and (the majority of) others are better predicted using our machine learning approach, the forecast averaging literature provides a straightforward solution: combine the predictions in order to generate composite price and land value predictions weighted in favor of the better model in each context. While there are many ways to combine forecasts – arithmetic average, eigenvector weighting, and complete subset regression with information theoretic weighting, to name a few – we have chosen one of the more straightforward ways to combine our forecasts, a simple regression. We implement our forecast combination using the following equation:

$$p_{i,j,t} = \alpha_{j,t} + \beta_1 \tilde{p}^{HD}_{i,j,t} + \beta_2 \tilde{p}^{GBT}_{i,j,t} \tag{7}$$

where $p_{i,j,t}$ is the observed price for the $i^th$ observation in the $j^th$ state in the $t^{th}$ period from the 20% test set, $\tilde{p}^{HD}_{i,j,t}$ is the price prediction for that same property by the hedonic model we outlined in Section 2, $\tilde{p}^{GBT}_{i,j,t}$ is the price prediction for that same property by the machine learning model we outlined in Section 4, and finally $\alpha_{j,t}$ is a state-year specific bias correction term. It is important to note that these weights need not sum to one, nor must they both be positive. Yet, the intuition is straightforward, as the composite value gives greater weight to a given model prediction if that model predicts the sale price of the property in a given state more accurately.

**Figure 4. Mean Absolute Error Comparison: Distribution by State from 2004-2015**

*a. Composite Forecast versus ONS Model*




**Note: All comparisons are made using the out-of-sample transaction set. To construct these plots we take the ratio of Mean Absolute Errors for each model and plot the resulting distribution across years for each state.**

We have included in an online appendix (Figure B) a figure which outlines the distribution of weights and the bias correction term by year for all states. The predictions put forth by the our GBT model are nearly uniformly preferred by weight over those produced by the hedonic model though both distributions are clearly different from zero. The bias terms tend to be negative overall which indicates that the predictions we do tend to over-predict relative to the true value. These weights are also informative as the overall mean-absolute-error, relative to the ONS model, is significantly less for the composite prediction compared to either of our original models. Figure 4 we see that the ratio of mean-absolute-error between the composite predictions and the ONS model is completely in favor of the composite for all state-year combinations.

We then apply these weights to the land values directly by the following equation,

$$\tilde{lv}_{i,j,t}^{comp} = \hat{\alpha}_{j,t} + \hat{\beta}_1 \tilde{lv}_{i,j,t}^{HD} + \hat{\beta}_2 \tilde{lv}_{i,j,t}^{GBT}, \tag{8}$$

where the land values are calculated from equation 3 and 5 respectively. Our explicit assumption here is that forecast error from the price is equally weighted between structural error and land error and thus the weights are not different. This composite method is the default method for the results reported in the next section.

# 5. Results

One challenge with granular, property-level land value predictions is that we generate millions of results over a decade-long sample, which can then be reported in countless ways. Thus, in this section, we proceed by reporting a handful of tables and figures that are useful for illustrating national and regional trends for 2006-2015, but only scratches the surface of how this data can be reported. First, we begin by reporting land value and leverage in Table 3 for single-family residential land underlying structures (labeled Suburban Residential in subsequent tables), using the composite method combining GBT and hedonic methods described in the previous section. This category is both the most valuable land in aggregate and, for comparison with other studies, it is one of the most common types of land valued in the academic literature. Second, in Tables 4 and 5, we provide price-per-acre estimates for the remaining residential categories (dense urban, urban, and rural) as well as agricultural, commercial, and industrial land. Third, in Table 6, we provide aggregates estimates of land value for the contiguous U.S. for all land groups, broken down by census division. In the online appendix, we further break out the results by state-year combinations and the final results, once published, will include further geographic disaggregations.[44]

## 5.1. Single-family residential (suburban) land value and land leverage results

Table 3 shows the price-per-acre and land leverage for single-family residential land from 2006 through 2015 across nine census divisions in the United States. These results highlight tremendous variation both across and within regions over time, conforming to the already well-established time-series dynamics that land value experienced a bust following the 2006-07 highs in the real estate markets, bottoming out over the next few years, and subsequently rebounding over the latter half of the sample period. While volatile over this period, the Pacific region, for example, maintained the highest value for single-family residential land, averaging nearly $850,000 per acre over this decade. The regions with the least expensive land value for single-family residential property were in the South. Based on the predicted prices and land values leverage (*i.e.*, the ratio of land value to price), was anywhere between a low of 18% to a high of 71% during the decade with again, the lowest leverage values appearing in the West South Central division while the highest leverage appeared in the New England division.

[44] We use census divisions and regions defined by the Census Bureau in subsequent tables and figures for a variety of reasons. Aesthetically, these aggregations can fit on a page in a single, relatively easy to read, table or figure. Given that some states are missing sale price data, another benefit to using divisions and regions is that we can aggregate to the national level if we assume that the missing states are reasonably represented in the division by the states we do have in the ZTRAX data. We return to this limitation in the Discussion section below.

## Table 3. Division Single Family Residences: Composite Values

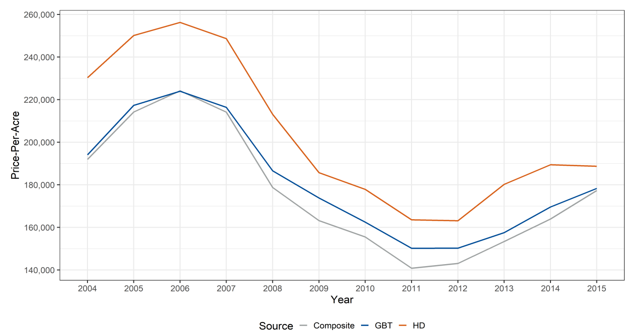
| | Division | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | New England | 434,588 | 419,234 | 369,838 | 337,840 | 326,284 | 314,062 | 303,167 | 307,869 | 310,125 | 320,281 |
| | Middle Atlantic | 364,889 | 358,524 | 333,303 | 306,809 | 292,496 | 263,358 | 253,027 | 252,038 | 256,660 | 265,617 |
| | East North Central | 248,280 | 235,753 | 195,028 | 180,380 | 159,612 | 141,387 | 138,216 | 150,598 | 164,825 | 173,337 |
| | West North Central | 232,791 | 229,238 | 195,179 | 184,856 | 173,290 | 171,461 | 180,573 | 183,412 | 198,405 | * |
| Price-Per-Acre | South Atlantic | 205,955 | 186,170 | 153,845 | 114,920 | 99,239 | 92,102 | 100,516 | 108,945 | 122,842 | 139,027 |
| | East South Central | 60,523 | 58,054 | 51,751 | 52,870 | 78,011 | 94,258 | 90,055 | 93,358 | 98,028 | * |
| | West South Central | 43,882 | 46,819 | 46,459 | 41,903 | 40,330 | 40,029 | 30,493 | 31,501 | 31,145 | 36,275 |
| | Mountain | 447,296 | 419,146 | 348,187 | 268,319 | 280,345 | 269,931 | 289,090 | 323,258 | 365,238 | 420,114 |
| | Pacific | 1,156,633 | 1,107,416 | 777,423 | 646,459 | 665,014 | 621,963 | 638,000 | 808,057 | 936,628 | 1,084,222 |
| | New England | 0.71 | 0.71 | 0.68 | 0.68 | 0.65 | 0.65 | 0.66 | 0.66 | 0.68 | 0.70 |
| | Middle Atlantic | 0.49 | 0.49 | 0.49 | 0.48 | 0.47 | 0.44 | 0.43 | 0.41 | 0.40 | 0.40 |
| | East North Central | 0.49 | 0.48 | 0.46 | 0.47 | 0.42 | 0.39 | 0.36 | 0.36 | 0.36 | 0.38 |
| | West North Central | 0.49 | 0.50 | 0.42 | 0.43 | 0.40 | 0.43 | 0.44 | 0.43 | 0.43 | * |
| Leverage | South Atlantic | 0.37 | 0.34 | 0.34 | 0.30 | 0.25 | 0.23 | 0.24 | 0.24 | 0.26 | 0.29 |
| | East South Central | 0.30 | 0.28 | 0.24 | 0.26 | 0.44 | 0.58 | 0.54 | 0.54 | 0.54 | * |
| | West South Central | 0.23 | 0.23 | 0.22 | 0.20 | 0.19 | 0.19 | 0.14 | 0.13 | 0.13 | 0.15 |
| | Mountain | 0.42 | 0.41 | 0.41 | 0.36 | 0.39 | 0.39 | 0.39 | 0.38 | 0.39 | 0.42 |
| | Pacific | 0.56 | 0.55 | 0.49 | 0.45 | 0.43 | 0.42 | 0.42 | 0.45 | 0.48 | 0.51 |

**Note: Recall that price-per-acre is calculated as the sum of all land values in an area divided by the sum of all acreage in that area in the assessment set. This is fundamentally a different centrality measure than the price-per-acre of the average plot, though both are reasonable. Values for West North Central and East South Central in 2015 have been suppressed due to data issues. Leverage is calculated by dividing the predicted land value by the predicted price and averaging over the region. In this sense leverage is that of the average plot of land in the region. All dollars are nominal.**

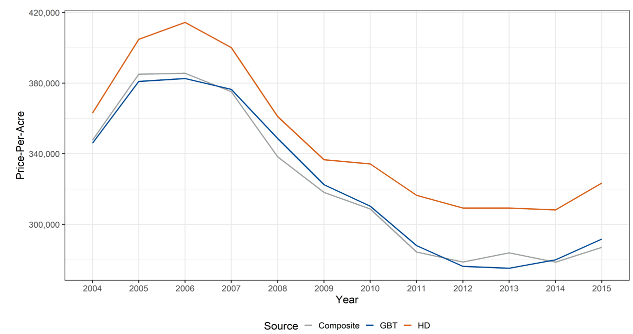
In Figure 5 we collapse single family residences down to the four regions of the continental U.S. (as designated by the U.S. Census Bureau) to illustrate the time-series dynamics across regions and three different models. All models in all regions show procyclical movement in land prices, consistent with the notion that land prices fluctuated directly with the demand shocks to the real estate markets over this period. Overall the price-per-acre of the two-step kmeans-GBT method outlined in Section 4 shows lower overall land values for all four divisions than the hedonic model. The composite value tends to be closer to the former rather than the latter, though they are not equivalent. This hides some of the variation that would be seen between states, as there are certainly states, such as California, where the composite value is almost perfectly in between or even favoring the hedonic values overall.

## Figure 5. Region Price-Per-Acre

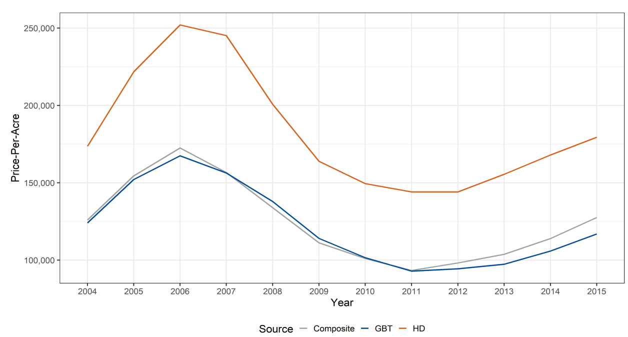
*a. Midwest Region*




*b. North East Region*




*c. South Region*




*d. West Region*


**Note: In each plot we have grouped the states according to their regional designation from the U.S. Census Bureau. These are weighted by the number of homes in the assessment set so that larger states will have more influence in the plot. Composite figures are computed by using by state-by-year weights from the observed sales price and predicted prices from each model. *Please note, y-axis scale is not common across each of the sub-figures.***

One takeaway we gleaned from these regional comparisons (and state-by-state comparisons in the online appendix) is that the ML method contained herein does "more with less." When a market has highly detailed data and a swift flow of transactions, the hedonic model tends to do quite well predicting the price and thus land value. In markets, either by state or land type, where the market is thinner, the ML model tends to have fewer issues with extreme values and better processes heterogeneity among individual parcels.[45]

[45] We recognize that it is nearly impossible to provide results of parcel level land values aggregated to every geography a reader might want to evaluate. Nonetheless, as part of this research, we are currently developing a "shiny app" that will allow users to, on demand, aggregate the land values to numerous subnational levels of geography they prefer. Moreover, this tool will allow for the user to examine a subset or even an individual state by county or census tract, for example, and download the appropriate data. Note we do not plan to provide individual property values at this time due to potential legal restrictions with the data; and, data limitations for some states, counties, tracts, etc. would also prevent estimates for some subnational geographies. To be clear, the data repository we describe above would be an extension of this

In Table 4, we show our composite estimates for all residential land types by census division, as well as agricultural land estimates by division. Recall that the primary difference in urban and dense urban (as defined by the NLUD data we use) is that dense urban areas have smaller plots (<.1 acre), which dominate sales of residential properties in dense cities.

**Table 4. Division Residential and Agricultural Price-Per-Acre: Composite Values**

| | Division | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | New England | 4,390,728 | 4,064,087 | 3,782,702 | 3,568,738 | 3,532,994 | 3,739,399 | 4,214,188 | 4,830,544 | 5,264,961 | 5,583,918 |
| | Middle Atlantic | 3,514,665 | 3,347,908 | 3,183,585 | 2,937,404 | 2,957,134 | 2,887,010 | 2,715,331 | 2,814,853 | 2,871,901 | 2,829,367 |
| | East North Central | 1,901,898 | 1,765,491 | 1,579,959 | 1,363,429 | 1,268,454 | 1,078,910 | 951,014 | 996,262 | 1,153,150 | 1,197,307 |
| | West North Central | 1,342,927 | 1,300,216 | 1,306,655 | 1,059,317 | 1,013,786 | 788,484 | 765,985 | 800,619 | 981,398 | 949,944 |
| Dense Urban | South Atlantic | 2,698,156 | 2,629,256 | 2,044,937 | 1,519,384 | 1,377,169 | 1,120,682 | 1,113,927 | 1,234,078 | 1,288,969 | 1,461,937 |
| | East South Central | 1,204,170 | 1,191,990 | 1,036,609 | 1,026,781 | 992,264 | 939,277 | 867,551 | 900,278 | 964,688 | 969,555 |
| | West South Central | 619,508 | 728,811 | 604,254 | 707,326 | 675,782 | 604,476 | 636,644 | 567,717 | 522,754 | 557,946 |
| | Mountain | 1,709,801 | 1,317,103 | 1,059,386 | 541,712 | 523,359 | 482,635 | 704,046 | 1,069,966 | 1,215,946 | 1,618,904 |
| | Pacific | 4,878,001 | 4,430,268 | 3,255,403 | 2,534,171 | 2,551,315 | 2,267,257 | 2,142,890 | 2,947,532 | 3,654,961 | 3,966,805 |
| | New England | 895,796 | 822,416 | 691,209 | 588,353 | 560,019 | 523,496 | 550,734 | 629,823 | 670,927 | 690,115 |
| | Middle Atlantic | 740,863 | 700,978 | 658,297 | 609,182 | 629,507 | 599,269 | 575,881 | 595,190 | 635,986 | 649,422 |
| | East North Central | 299,072 | 242,486 | 215,176 | 203,054 | 196,448 | 178,757 | 153,682 | 157,982 | 187,219 | 210,601 |
| | West North Central | 490,972 | 455,156 | 390,404 | 328,442 | 336,871 | 268,539 | 271,378 | 309,871 | 384,791 | 382,061 |
| Urban | South Atlantic | 637,606 | 587,715 | 429,748 | 239,477 | 209,601 | 198,877 | 232,715 | 257,590 | 292,301 | 353,634 |
| | East South Central | 226,738 | 235,269 | 209,574 | 195,602 | 194,201 | 191,165 | 188,252 | 196,666 | 211,082 | 236,268 |
| | West South Central | 252,046 | 234,162 | 202,337 | 214,566 | 215,406 | 165,504 | 194,971 | 228,901 | 282,243 | 262,876 |
| | Mountain | 472,937 | 454,159 | 438,306 | 264,313 | 291,133 | 366,627 | 437,811 | 469,693 | 525,665 | 609,871 |
| | Pacific | 967,605 | 868,165 | 640,102 | 490,939 | 471,143 | 446,159 | 431,577 | 539,841 | 722,674 | 732,325 |
| | New England | 39,289 | 36,233 | 32,715 | 28,245 | 27,917 | 24,185 | 23,325 | 22,100 | 23,067 | 23,979 |
| | Middle Atlantic | 9,488 | 9,242 | 8,545 | 7,875 | 7,945 | 7,406 | 6,681 | 8,282 | 9,049 | 8,114 |
| | East North Central | 8,173 | 6,363 | 5,932 | 4,965 | 4,526 | 4,605 | 5,482 | 6,563 | 7,536 | 7,870 |
| | West North Central | 6,906 | 6,955 | 6,959 | 5,579 | 5,557 | 5,438 | 5,196 | 4,980 | 6,673 | 6,180 |
| Rural | South Atlantic | 16,062 | 14,718 | 14,108 | 12,312 | 11,499 | 10,529 | 10,647 | 9,879 | 10,272 | 11,352 |
| | East South Central | 2,054 | 2,056 | 2,037 | 2,041 | 2,052 | 2,000 | 2,018 | 2,018 | 2,148 | 2,230 |
| | West South Central | 3,258 | 3,200 | 3,177 | 3,192 | 3,175 | 3,134 | 3,256 | 3,371 | 3,640 | 3,745 |
| | Mountain | 22,490 | 22,654 | 21,362 | 17,875 | 17,011 | 15,323 | 15,428 | 16,343 | 18,128 | 19,517 |
| | Pacific | 28,777 | 30,414 | 25,703 | 21,761 | 19,235 | 18,719 | 17,035 | 19,144 | 21,834 | 24,224 |
| | New England | 13,782 | 11,749 | 12,285 | 13,592 | 10,694 | 10,483 | 11,663 | 3,679 | 15,367 | 25,770 |
| | Middle Atlantic | 4,594 | 4,938 | 5,673 | 4,930 | 4,625 | 4,576 | 4,481 | 4,455 | 4,627 | 4,480 |
| | East North Central | 5,093 | 5,031 | 4,976 | 4,787 | 4,921 | 5,023 | 5,479 | 5,796 | 5,996 | 6,272 |
| | West North Central | 4,360 | 4,387 | 4,764 | 4,504 | 4,426 | 4,402 | 5,565 | 5,220 | 5,608 | 5,320 |
| Agricultural | South Atlantic | 8,778 | 8,218 | 6,413 | 5,169 | 4,894 | 4,351 | 4,009 | 3,971 | 4,400 | 5,152 |
| | East South Central | 2,899 | 2,759 | 2,649 | 2,469 | 2,328 | 2,276 | 2,417 | 2,380 | 2,452 | 2,832 |
| | West South Central | 1,887 | 2,047 | 1,805 | 1,352 | 1,553 | 1,909 | 2,365 | 2,654 | 3,188 | 3,087 |
| | Mountain | 5,404 | 7,009 | 5,723 | 4,248 | 3,220 | 3,137 | 3,438 | 3,446 | 3,291 | 3,986 |
| | Pacific | 8,456 | 8,587 | 8,289 | 6,551 | 6,880 | 6,880 | 7,180 | 8,664 | 9,816 | 12,275 |

**Note: For this table we have again summed the land values in a division and divided it by the sum of the acreages in that division for a measure of price-per-acre. Single family residence value can be found in Table 3. All dollar values are nominal USD.**

research paper as experimental/developmental, and not an official statistical product of BEA. We intend to make this available upon final publication of this paper.

Not surprisingly, dense urban land is by far the most valuable land in terms of price-per-acre, which can be as high as $4-5 million per acre in some divisions (New England and Pacific), but only a million or less in other areas in the U.S. (like in the South). Urban land is substantially cheaper, as it is predominantly sold on larger plots just outside the CBD of most cities (*i.e.*, most often the areas between "the suburbs" and "the city"), but its value generally falls between urban land and its suburban SFR alternative. While their focus was only single-family residential property, a broader takeaway from these results is that they conform to the general dynamic reported in Davis et al. (2021) and numerous other studies that show a steep price gradient away from density. Moreover, rural residential property, which largely consists of large parcels (>2.5 acres) and other rural land-use types (*e.g.*, mobile/manufactured homes), is the cheapest residential land type, conforming to this broader density story. But, it should be noted that rural land is far closer in value to agricultural land than suburban SFR land, which is intuitive given the location of rural and agricultural land more generally. In census divisions like New England, for example, rural land is relatively expensive, given the density of the states is also relatively high; however, Table 4 also shows agricultural land is similarly high due to its high opportunity cost of being converted into rural land. We do not see quite the same degree of this dynamic in the Pacific division, however, likely do to different density and how much further away rural areas are from densely populated areas in the American West as compared to the East Coast.

**Table 5. Regional Non-Residential Price-Per-Acre: Composite Values**

| | Division | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Industrial | New England | 264,834 | 240,673 | 278,049 | 231,625 | 208,919 | 172,049 | 212,051 | 151,694 | 152,244 | 144,839 |
| | Middle Atlantic | 287,901 | 313,559 | 333,352 | 284,977 | 283,289 | 285,061 | 284,502 | 264,872 | 314,253 | 387,342 |
| | East North Central | 185,345 | 198,609 | 176,237 | 156,989 | 134,830 | 142,393 | 112,939 | 160,233 | 142,674 | 175,560 |
| | West North Central | 194,129 | 210,856 | 184,506 | 151,394 | 157,154 | 136,883 | 128,426 | 158,893 | 181,290 | 150,434 |
| | South Atlantic | 205,621 | 197,092 | 188,199 | 147,447 | 141,103 | 121,589 | 106,123 | 136,190 | 154,305 | 170,128 |
| | East South Central | 49,286 | 67,664 | 68,699 | 67,002 | 67,097 | 77,079 | 76,464 | 44,352 | 54,124 | 64,464 |
| | West South Central | 56,133 | 57,388 | 71,639 | 68,062 | 62,352 | 81,359 | 89,423 | 92,742 | 94,908 | 88,059 |
| | Mountain | 253,935 | 280,651 | 308,962 | 221,715 | 231,193 | 253,301 | 270,693 | 278,862 | 346,988 | 372,977 |
| | Pacific | 391,233 | 469,532 | 457,849 | 400,352 | 395,858 | 437,453 | 374,092 | 360,895 | 496,777 | 574,450 |
| Commercial | New England | 448,719 | 473,367 | 490,004 | 404,556 | 408,964 | 414,921 | 423,793 | 366,330 | 413,789 | 363,923 |
| | Middle Atlantic | 628,399 | 786,175 | 698,084 | 614,971 | 571,250 | 617,417 | 597,235 | 589,558 | 652,674 | 721,703 |
| | East North Central | 287,068 | 295,920 | 266,207 | 225,368 | 207,810 | 210,658 | 171,966 | 237,716 | 233,440 | 251,440 |
| | West North Central | 374,195 | 392,757 | 297,098 | 236,333 | 245,726 | 220,230 | 212,644 | 219,386 | 266,732 | 270,761 |
| | South Atlantic | 269,131 | 295,491 | 281,760 | 239,798 | 236,666 | 230,968 | 230,112 | 233,310 | 258,187 | 286,738 |
| | East South Central | 92,578 | 96,044 | 96,954 | 82,869 | 84,121 | 88,687 | 105,229 | 109,732 | 117,065 | 152,510 |
| | West South Central | 178,718 | 223,737 | 178,479 | 167,578 | 139,569 | 147,469 | 152,428 | 159,266 | 211,474 | 236,729 |
| | Mountain | 774,877 | 838,801 | 798,861 | 659,530 | 559,095 | 530,936 | 512,820 | 588,706 | 665,155 | 764,276 |
| | Pacific | 663,015 | 672,698 | 610,105 | 557,577 | 547,155 | 568,834 | 566,147 | 606,654 | 896,146 | 1,012,348 |

**Note: For this table we have again summed the land values in a division and divided it by the sum of the acreages in that division for a measure of price-per-acre. All dollar values are nominal USD.**

In Table 5 we report composite estimates for industrial and commercial land. While still showing some pro-cyclical dynamics, compared to residential land values, our estimates of industrial and commercial land values over this period are somewhat flatter over this decade. Commercial land is generally more valuable than industrial land, which is likely due to a number of well-documented factors like differences in location. For example, if commercial land is more likely to be located in more densely populated areas near residential land, then we would expect land to reflect both this amenity value and opportunity cost. However, we should again express some caution with our estimates of non-residential land, which are derived from a coarser set of data. We return to this point in the Discussion section below.

## 5.2. National and census division results for all land types

The price estimates in the prior subsection provide important information about property markets over this period; however, as a more general point, we should emphasize that prices tell only part of a larger story. The national economic accounts aggregate economic activity by measuring national income and expenditures in GDP, for example, which is the sum total of relevant prices *and* quantities. For the national balance sheet (in the Integrated Macroeconomic Accounts), BEA and the Federal Reserve value assets in these terms as well. Hence, we follow a similar approach by (Wentland et al., 2020) that uses detailed land-use data to provide corresponding quantities of land for the contiguous United States to construct a pilot accounting of private land as an asset. The National Land Use Database (NLUD) provides a nearly exhaustive accounting of land use in the contiguous U.S., which leverages detailed data from numerous sources to depict how land is used across the categories relevant for this study.[46] One drawback of this source is that the NLUD was initially developed for only a single year, 2010, which we use here. Because land-use does not change particularly rapidly (*e.g.*, once a property is built residential, it generally stays that way for decades, given the relatively long lifespan of most structures), a snapshot of land-use is sufficient for a proof-of-concept account; however, a regularly produced NLUD or equivalent would be essential for production of an official account. We return to this point in the next section, as we discuss how this data would need to be augmented or even replaced by official sources if BEA would transition this proof-of-concept work into an official account.

Table 6 accounts for the total asset value of land by census division, land type, and year. Overall, private land in the contiguous U.S. was worth a staggering 27.27 trillion nominal dollars in 2006. By 2011 this had dropped by nearly 36% to 17.8 trillion dollars but largely had recovered by 2015 (24.1 trillion).[47] Nearly 20% of the 2006 value was in single family housing alone; and, all residential (dense urban, urban, single family/suburban, and rural) accounting for nearly 73% of the total land value in 2006. The relative ordering of the asset values by region is as expected. The most valuable region, by aggregate

[46]The NLUD was derived from "two-dozen publicly-available, national spatial datasets – predominately based on census housing, employment, and infrastructure, as well as land cover from satellite imagery... result[ing] in 79 land use classes" Theobald (2014). In the online appendix (See Tables A1 through A4), we show how we collapsed the land-use categories from the NLUD to the corresponding categories in ZTRAX's land-use designations, directly following the classification scheme in (Wentland et al., 2020).

[47]To calculate this we use our price-per-acre measure (the price) times the acreage (quantity) in that particular land group as denoted by the 2010 figures from NLUD (see (Wentland et al., 2020) for a similar exercise).

private land value, is the Pacific region, which includes California, Oregon, and Washington. This is not surprising considering the well-documented evidence of high property prices in those states. The least valuable region is the East South Central, which includes states such as Kentucky, Tennessee, and Alabama. Though overall values are ordered as expected, there is significant heterogeneity between the values of individual land types, some of which is driven by the relative size of that land type in the area.

**Table 6. Land Value Totals by Division–part 1**

| | | NLUD 2010 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Dense Urban | 237 | 1,156 | 1,050 | 772 | 601 | 605 | 537 | 508 | 699 | 866 | 940 |
| | Urban | 2,415 | 2,337 | 2,097 | 1,546 | 1,186 | 1,138 | 1,077 | 1,042 | 1,304 | 1,745 | 1,769 |
| | Suburban | 1,629 | 1,481 | 1,804 | 1,266 | 1,053 | 1,083 | 1,013 | 1,039 | 1,316 | 1,526 | 1,766 |
| Pacific | Rural | 9,893 | 285 | 301 | 254 | 215 | 190 | 185 | 169 | 189 | 216 | 240 |
| | Commercial | 611 | 405 | 411 | 373 | 341 | 334 | 348 | 346 | 371 | 548 | 619 |
| | Industrial | 261 | 102 | 123 | 119 | 104 | 103 | 114 | 98 | 94 | 130 | 150 |
| | Agricultural | 78,480 | 664 | 674 | 651 | 514 | 540 | 540 | 563 | 680 | 770 | 963 |
| | Dense Urban | 81 | 138 | 107 | 86 | 44 | 42 | 39 | 57 | 87 | 98 | 131 |
| | Urban | 1,383 | 654 | 628 | 606 | 366 | 403 | 507 | 605 | 650 | 727 | 843 |
| | Suburban | 1,263 | 448 | 529 | 440 | 339 | 354 | 341 | 365 | 408 | 461 | 531 |
| Mountain | Rural | 7,587 | 171 | 172 | 162 | 136 | 129 | 116 | 117 | 124 | 138 | 148 |
| | Commercial | 521 | 404 | 437 | 416 | 344 | 291 | 277 | 267 | 307 | 347 | 398 |
| | Industrial | 212 | 54 | 59 | 65 | 47 | 49 | 54 | 57 | 59 | 74 | 79 |
| | Agricultural | 218,751 | 1,182 | 1,533 | 1,252 | 929 | 704 | 686 | 752 | 754 | 720 | 872 |
| | Dense Urban | 49 | 66 | 64 | 64 | 52 | 50 | 39 | 38 | 39 | 48 | 47 |
| | Urban | 1,377 | 676 | 627 | 538 | 452 | 464 | 370 | 374 | 427 | 530 | 526 |
| | Suburban | 1,246 | 290 | 286 | 243 | 230 | 216 | 214 | 225 | 229 | 247 | * |
| West North Central | Rural | 11,073 | 76 | 77 | 77 | 62 | 62 | 60 | 58 | 55 | 74 | 68 |
| | Commercial | 510 | 191 | 200 | 152 | 121 | 125 | 112 | 108 | 112 | 136 | 138 |
| | Industrial | 268 | 52 | 57 | 49 | 41 | 42 | 37 | 34 | 43 | 49 | 40 |
| | Agricultural | 269,990 | 1,177 | 1,184 | 1,286 | 1,216 | 1,195 | 1,188 | 1,502 | 1,409 | 1,514 | 1,436 |
| | Dense Urban | 148 | 281 | 261 | 234 | 202 | 188 | 160 | 141 | 147 | 171 | 177 |
| | Urban | 2,872 | 859 | 696 | 618 | 583 | 564 | 513 | 441 | 454 | 538 | 605 |
| | Suburban | 2,640 | 655 | 622 | 515 | 476 | 421 | 373 | 365 | 398 | 435 | 458 |
| East North Central | Rural | 24,793 | 203 | 158 | 147 | 123 | 112 | 114 | 136 | 163 | 187 | 195 |
| | Commercial | 715 | 205 | 212 | 190 | 161 | 149 | 151 | 123 | 170 | 167 | 180 |
| | Industrial | 441 | 82 | 88 | 78 | 69 | 59 | 63 | 50 | 71 | 63 | 77 |
| | Agricultural | 95,720 | 488 | 482 | 476 | 458 | 471 | 481 | 524 | 555 | 574 | 600 |
| | Dense Urban | 98 | 61 | 71 | 59 | 69 | 66 | 59 | 62 | 56 | 51 | 55 |
| | Urban | 2,066 | 521 | 484 | 418 | 443 | 445 | 342 | 403 | 473 | 583 | 543 |
| | Suburban | 2,318 | 102 | 109 | 108 | 97 | 93 | 93 | 71 | 73 | 72 | 84 |
| West South Central | Rural | 22,875 | 75 | 73 | 73 | 73 | 73 | 72 | 74 | 77 | 83 | 86 |
| | Commercial | 809 | 145 | 181 | 144 | 136 | 113 | 119 | 123 | 129 | 171 | 192 |
| | Industrial | 388 | 22 | 22 | 28 | 26 | 24 | 32 | 35 | 36 | 37 | 34 |
| | Agricultural | 207,344 | 391 | 424 | 374 | 280 | 322 | 396 | 490 | 550 | 661 | 640 |

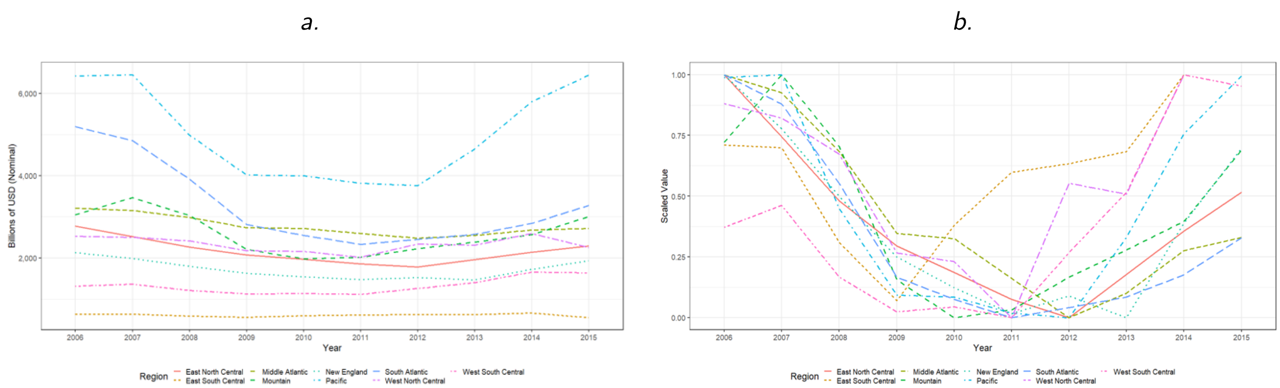
**Note: Acres are in thousands of acres. All dollar figures are in billions of nominal dollars.**

| | | NLUD 2010 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Dense Urban | 32 | 39 | 38 | 33 | 33 | 32 | 30 | 28 | 29 | 31 | 31 |
| | Urban | 797 | 181 | 188 | 167 | 156 | 155 | 152 | 150 | 157 | 168 | 188 |
| | Suburban | 1,810 | 110 | 105 | 94 | 96 | 141 | 171 | 163 | 169 | 177 | * |
| East South Central | Rural | 29,328 | 60 | 60 | 60 | 60 | 60 | 59 | 59 | 59 | 63 | 65 |
| | Commercial | 412 | 38 | 40 | 40 | 34 | 35 | 37 | 43 | 45 | 48 | 63 |
| | Industrial | 240 | 12 | 16 | 16 | 16 | 16 | 18 | 18 | 11 | 13 | 15 |
| | Agricultural | 64,973 | 188 | 179 | 172 | 160 | 151 | 148 | 157 | 155 | 159 | 184 |
| | Dense Urban | 210 | 567 | 552 | 429 | 319 | 289 | 235 | 234 | 259 | 271 | 307 |
| | Urban | 3,049 | 1,944 | 1,792 | 1,310 | 730 | 639 | 606 | 710 | 785 | 891 | 1,078 |
| | Suburban | 5,116 | 1,054 | 952 | 787 | 588 | 508 | 471 | 514 | 557 | 628 | 711 |
| South Atlantic | Rural | 44,969 | 722 | 662 | 634 | 554 | 517 | 473 | 479 | 444 | 462 | 510 |
| | Commercial | 886 | 238 | 262 | 250 | 212 | 210 | 205 | 204 | 207 | 229 | 254 |
| | Industrial | 375 | 77 | 74 | 71 | 55 | 53 | 46 | 40 | 51 | 58 | 64 |
| | Agricultural | 67,551 | 593 | 555 | 433 | 349 | 331 | 294 | 271 | 268 | 297 | 348 |
| | Dense Urban | 232 | 815 | 777 | 739 | 681 | 686 | 670 | 630 | 653 | 666 | 656 |
| | Urban | 1,462 | 1,083 | 1,025 | 962 | 891 | 920 | 876 | 842 | 870 | 930 | 949 |
| | Suburban | 2,171 | 792 | 778 | 724 | 666 | 635 | 572 | 549 | 547 | 557 | 577 |
| Middle Atlantic | Rural | 19,415 | 184 | 179 | 166 | 153 | 154 | 144 | 130 | 161 | 176 | 158 |
| | Commercial | 311 | 195 | 245 | 217 | 191 | 178 | 192 | 186 | 183 | 203 | 224 |
| | Industrial | 151 | 43 | 47 | 50 | 43 | 43 | 43 | 43 | 40 | 47 | 58 |
| | Agricultural | 21,632 | 99 | 107 | 123 | 107 | 100 | 99 | 97 | 96 | 100 | 97 |
| | Dense Urban | 61 | 268 | 248 | 231 | 218 | 216 | 228 | 257 | 295 | 321 | 341 |
| | Urban | 669 | 599 | 550 | 462 | 394 | 375 | 350 | 368 | 421 | 449 | 462 |
| | Suburban | 1,176 | 511 | 493 | 435 | 397 | 384 | 369 | 357 | 362 | 365 | 377 |
| New England | Rural | 10,836 | 426 | 393 | 354 | 306 | 303 | 262 | 253 | 239 | 250 | 260 |
| | Commercial | 196 | 88 | 93 | 96 | 79 | 80 | 81 | 83 | 72 | 81 | 71 |
| | Industrial | 90 | 24 | 22 | 25 | 21 | 19 | 15 | 19 | 14 | 14 | 13 |
| | Agricultural | 15,761 | 217 | 185 | 194 | 214 | 169 | 165 | 184 | 58 | 242 | 406 |
| U.S. National Totals | | 1,264,975 | 27,265 | 26,919 | 23,154 | 19,313 | 18,617 | 17,834 | 18,431 | 19,913 | 22,653 | 24,099 |

**Note: Acres are in thousands of acres. All dollar figures are in billions of nominal dollars.**

A careful examination of the table will also reveal that, while it is officially the case that the [Great] Recession ended in June of 2009 (as dated by the NBER), many regions did not experience the trough until 2011-2013. To make the time-series dynamics clearer, we graph the total land asset value by each division in Figure 6a, and in Figure 6b we provide a min-max transformation that better illustrates peak-trough dynamics. A new insight from this account, which unlike (Wentland et al., 2020) provides a yearly accounting of land value, is that the value of private land in the U.S. bottomed out over five years, which varied regionally. All nine census divisions peaked in 2006 or 2007; yet, some experienced the bottom of the trough in 2009, 2010, 2011, 2012, and even 2013. This is not immediately apparent viewing pricing data alone, and one of the many more nuanced insights a national account can offer by aggregating total value by the product of prices and quantities. Given that the U.S. had numerous policies related to the bust in asset prices over this time period, the timing, absolute values, and regional variation are all potentially highly relevant data points that could inform future policymakers if this type of data were available going forward.

**Figure 6. U.S. Private Land Asset Value by Division**




**Note: Here we have plotted the aggregate land value by division in Figure 6a. While the scale hides some of the variation, especially in less expensive divisions, you can clearly see a procyclical pattern emerging. In Figure 6b we transformed the value using a min-max transformation. This allows us to see the peak for each division (where the max=1), as well as when each division experienced the trough (where min value=0).**

# 6. Robustness to further out-of-sample testing: vacant land comparison

Though they are not used for deriving the land value estimates reported above, the Ztrax data set contains data on more than one million vacant residential land transactions over our sample period. As a robustness check, we compare our land estimates for single-family residential property to nearby sales of vacant land. However, we should express caution in such a comparison, given the previously discussed selection bias issues with vacant land.[48] With these limitations in mind we use the vacant land transactions as an out-of-sample validation of our land estimates because, while imperfect, they are actual transactions that take place in the marketplace.

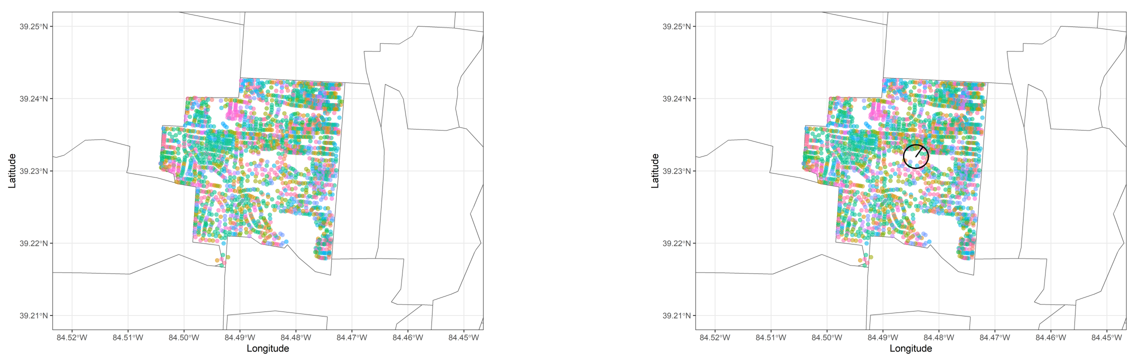
So, while vacant land transactions may not be representative of all developed land, the aforementioned selection issues should be mitigated in at least a couple circumstances. First, they should be more representative of fair-market land value of property in the immediate, adjacent area than areas further away. For instance, much of vacant land may be sold at the outskirts of developed areas; so, while vacant land on the far reaches of the suburbs may not extrapolate well to developed land near the central business district (CBD), it may proxy reasonably well for other nearby land at the outskirts. Second, when there are periods of higher development and thus higher volume of vacant land sales in residential developments, we should expect this higher volume to be more representative than periods when sales

[48]For example, a vacant plot may be smaller, more oddly shaped, or geographically/environmentally undesirable (*e.g.*, it could be the lowest elevation in the area and thus the most likely to flood). Second, despite their being over one million transactions, these take place over the full time span of the data and across the entire geographical region examined. In practice, for each state in each time period there is significantly less data from which we can compare.

are scant. When these conditions are satisfied, we should observe more comparable developed land values and vacant land values.

Thus, to generate a more apples-to-apples comparison of vacant land and nearby developed land, we use the geolocation of a sold parcel we draw a circle with radius of one-tenth of a mile around the plot.[49] For example, in Figure 7 we have plotted the assessment set first shown in Figure 2b. If this tract were to have a vacant plot sold as indicated in Figure 7, then our donut would be as indicated. This polygon would serve as our comparable area with respect to local geography. We exclude the first $1/100^{th}$ of a mile so as to avoid same year plot sales post development. This leaves us with a small donut shaped polygon with total area of approximately $0.03$ square-miles, or $19$ square-acres. Using this polygon, we identify properties nearby in our assessment set for which we have land estimates. If there were no nearby properties we simply dropped that plot for the purposes of this exercise, keeping a small, yet more comparable, subsample of the data for comparison.

**Figure 7. Matching Vacant Land Plots to Nearby Developed Plots**




**Note: To match nearby developed plots with vacant land transactions we create a buffer of 0.10 miles with an exclusion range of 0.01 miles. In some cases, such as those illustrated above, there may be many comparable developed properties while in others there may be far fewer. Vacant land transactions are limited to those zoned for single family residential use.**

The end result is over half a million vacant land transactions across the 36 states examined. In Table 7 we have provided the number of vacant plots by year along with the median (observed) price and acreage for those plots. Note that, while half a million vacant land transactions seems like a large sample at first glance, the within-year number of transactions varies between $28,000$ and $67,000$. This could further be broken down across the 36 states and should be apparent that, over time and geography this sample is not very large. Using the donut buffer described above, each transacted vacant plot is matched to, on average, forty nearby [developed] plots from the assessment set. [50]

[49]We have also done this with larger radius circles but as the radius grows the likelihood of being near comparable plots of land decreases since these are not walking distance radii but rather as the crow flies.

[50]This is not to say that every plot has many nearby matches; in fact, some vacant plots are only matched to a single nearby developed plot within the geography we have outlined, and some have as many as $200$ nearby comparisons.

**Table 7. Vacant Land Comparison by Year**

| Year | Vacant Plots | Median Price | Median Acreage | Nearby Price | Nearby Acreage | Nearby Plots |
|---|---|---|---|---|---|---|
| 2004 | 48,750 | 50,000 | 0.38 | 52,569 | 0.36 | 40.21 |
| 2005 | 67,772 | 66,900 | 0.42 | 66,082 | 0.40 | 37.69 |
| 2006 | 54,022 | 65,000 | 0.46 | 69,181 | 0.43 | 40.40 |
| 2007 | 43,855 | 57,500 | 0.45 | 61,863 | 0.44 | 41.55 |
| 2008 | 33,212 | 44,803 | 0.42 | 49,515 | 0.41 | 42.42 |
| 2009 | 28,705 | 35,000 | 0.41 | 37,201 | 0.40 | 43.31 |
| 2010 | 29,968 | 30,000 | 0.43 | 35,654 | 0.41 | 42.55 |
| 2011 | 28,075 | 30,000 | 0.43 | 34,602 | 0.42 | 41.14 |
| 2012 | 32,838 | 30,000 | 0.43 | 36,357 | 0.41 | 41.15 |
| 2013 | 41,736 | 39,500 | 0.43 | 43,892 | 0.41 | 41.69 |
| 2014 | 50,145 | 47,500 | 0.42 | 51,857 | 0.40 | 42.09 |
| 2015 | 65,566 | 79,696 | 0.35 | 63,391 | 0.37 | 41.08 |

**Note: The Ztrax dataset contains just over one million vacant land transactions for the 36 states examined. Here we are comparing the median price of these vacant plots within each year to nearby [developed] plots from the assessment set. The land prices used are the composite prices which are the weighted average of the hedonic regression and machine learning predictions. Each vacant plot is matched to a developed plot within an interval of $(0.01, 0.10)$ miles from the vacant plot. The Nearby Plots column provides the average number of matches per vacant plot. Note that vacant land transactions are not used in our estimation of the parcel level land values and thus this is an out-of-sample comparison.**

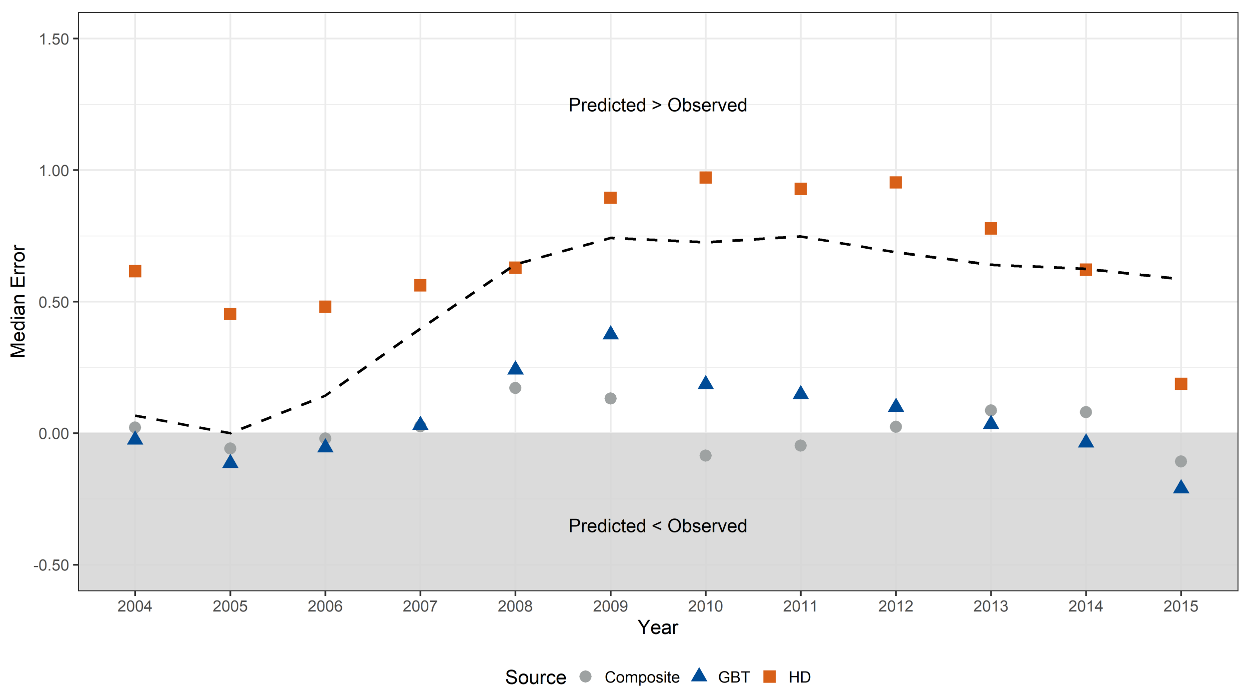
The summary statistics for this comparison show that developed plots tend to be slightly smaller in terms of median acreage and – as expected if one were to think of possible selection bias in vacant land – the median price for nearby plots is slightly higher. Note that we have restricted the conversation to our composite estimate of the parcel-level land value, that is the weighted combination of land values from the hedonic model and machine learning model as outlined in Equation 8. With a higher price and a lower acreage, the overall price-per-acre for the developed plots is higher than that of the vacant land, consistent with a selection bias argument. However, overall, the vacant land prices and the land prices from the developed properties are in the same ballpark, giving us more confidence that our results are not radically different than what we might observe in market transactions of land-alone.

Figure 8 shows that this comparison between our estimates and nearby vacant land sales is consistent with the economic intuition outlined above. Vacant land transactions showed the least bias/error when compared to our nearby valuations of developed land for all models during 2004-2006 (peaking in 2005). This corresponded to the peak of the housing construction boom in the U.S. Indeed, over the entire time-series, as we find a strong negative correlation between vacant land bias and new housing starts, a key national indicator for residential construction/investment.[51] Given the negative correlation, we graph

[51] This series is produced by Census and HUD, which can be found at: `https://fred.stlouisfed.org/series/HOUST1F`.

the inverse of New Housing Starts nationally on Figure 8, which shows that the series peaked in 2005, collapsed through 2009, and stagnated through 2015. Thus, if our models are estimating the true market value of this subset of land, this mirrored bias over the business cycle provides evidence consistent with what we would expect. Taken together, the out-of-sample price validation for vacant land and property sale prices of developed land provide robust evidence that our models, and the composite method in particular, robustly track market values in line with the spirit SNA valuation principles for the national accounts.[52]

**Figure 8. Prediction Error for Vacant Land Mirrors New Housing Starts**




**Note: Here we have plotted the mean error for each year produced by comparing the observed vacant land sales and the nearby land estimates using developed properties. The dashed line represents new housing starts (NHS) through the transformation $1 - \text{NHS}/\max(\text{NHS})$ with peak new housing starts occurring in 2005. Orange squares represent the error produced when comparing estimates from the linear hedonic model to vacant land prices while blue triangles represent the error from gradient boosted trees. The grey circles represent the error produced by an OLS weighted composite forecast between the two methods. The error of all three methods – linear hedonic, gradient boosted trees, and composite – are all correlated with changes in new housing starts.**

[52] Clapp and Lindenthal (2022) provide an additional test to compare models, which uses the land value as a determinant of price. In untabulated tests we have found that our composite measure also performs well using their novel validation approach; however, for brevity we omit the results and they are available upon request.

# 7. Discussion

Though the main contribution of this paper is methodological, the estimates produced by the application of this method to ZTRAX data demonstrate one way both ML methods and Big Data could come together to produce a pilot national account for land value for the United States. However, we should reiterate here that these estimates are not yet official statistics produced by BEA. Instead, these estimates provide a proof-of-concept that are both illustrative from an academic standpoint and a practical standpoint. For this to be an official account, a number of data limitations would need to be addressed, which we discuss below.

The primary data limitation that would need to be addressed would be filling gaps in both the price and quantity data. We have already mentioned ZTRAX's chief limitation, which is that it does not include sale price data for states whose local municipalities do not disclose final sale price data. And, in some states where some municipalities do disclose this information, there is sufficient missing data that we omit the states when we derive our estimates (*e.g.*, Louisiana, Maine, and Vermont). An assumption we make in our national and census division estimates above is that the omitted and missing states are reasonably proxied by their neighboring states for the aggregate estimates; but, for some divisions this assumption may strain credulity (*e.g.*, the West South Central division missing Texas). Data outside the ZTRAX dataset is available for purchase by various Big Data vendors, which include sale price data from non-disclosure states that could potentially fill this gap.[53]

Although price data is available for Hawaii and Alaska, the NLUD does not yet include these states, nor does it include the U.S. territories. BEA produces national economic accounts for all U.S. states and territories, drawing on data sources that are representative of all localities. Future work would need to update the NLUD to include these states/territories for a true national account. The United States Geological Survey (USGS) is currently doing pioneering work on expanding the scope of land use and land cover data in the U.S., including products by the Land Change Monitoring, Assessment, and Projection (LCMAP) that account for land at fine levels of detail. Though they have not yet developed a comparable land-use product, a regularly produced land-use product like the NLUD used in this study would provide a tractable path forward for a comprehensive official national (and subnational) land account.[54]

As an alternative approach to using a (not yet available) land-use data source, another potential path forward would be to use the land leverage estimates produced in this study to apply to existing figures in the national accounts for the asset value of real estate underlying structures. The current balance

[53] BEA has very recently purchased detailed data from another vendor, which has sale price data for the non-disclosure states from non-municipal sources (like Multiple Listing Services), which we are currently exploring as an avenue to fill this important gap and potentially update the estimates through 2022. Thus, the remedy for this limitation is on the horizon, so-to-speak.

[54] See Wentland et al. (2020) for a discussion of a number of other challenges with using this data for land valuation, including issues with dense urban and non-SFR properties in particular and the lack of detailed commercial/industrial structure characteristics.

sheets produced by BEA and the Federal Reserve in the Integrated Macroeconomic Accounts provide some aggregate figures for real estate, but the current configuration is not broken down precisely as we have done in this paper. With some modifications, land leverage estimates derived from this study could provide a breakout of land and structure value, limiting real estate value to the scope in which it is currently measured. But, given the regional variation in land leverage we observe in the data, we are reluctant to apply these leverages to national estimates as currently constituted on the balance sheet. A key takeaway from our regional figures is that regional variation matters; and, leverage changes both over time and across regions. A more accurate version of the land leverage approach would require use of the subnational data from which the real estate values were constructed, which is data not available to us currently. One advantage of the approach proposed in this paper is that it uses data sources that the general public can also access, facilitating transparency in the national accounts. On the other side of that coin, an alternative approach using internal data would not offer this benefit.

Finally, a key omitted land-use category is public lands, which, in the United States, are quite substantial in terms of acreage. Depending on how one interprets who these assets belong to in an accounting sense, it may not be necessary from the standpoint of the SNA to value these lands as an asset on the balance sheet. However, much of these lands are used for purposes with private value (*e.g.*, National Parks that generate revenues, leased grazing lands, etc.) and play a role in our economy as measured by GDP. From an environmental-economic accounting standpoint, these lands are in-scope of the accounts as defined by SEEA-CF. Because our method relies on market prices to guide valuation, our approach would likely need to be augmented (*e.g.*, constraining the sample to lands adjacent to or around public lands). But, this presents a number of conceptual and practical challenges that, in the interest of brevity, we leave for future work to explore further. Nevertheless, the SNA prescribes cost-based methodologies for accounting for government services in GDP, so having an alternative methodology for public sector valuation is also a potential path forward for public lands.

# 8. Conclusion

In the $21^{st}$ century, the increasing promulgation of large datasets (so-called Big Data) in combination with more advanced methods (like ML) present an opportunity for national statistics offices to exploit new ways to create more accurate, timely, and detailed estimates of products, services, and assets (Abraham et al., 2019). To answer the call of this new era, we cultivate a new approach to land valuation that leverages both Big Data and ML methods to provide new pilot estimates of private land value in the contiguous U.S. for a decade (2006-2015). Our results underscore the potential importance of private land as a quantitatively significant asset on our national balance sheet, as private land in the U.S. was worth an estimated $24 trillion in 2015. Considering that U.S. net wealth in 2015 (Q4) was about $81 trillion, this represents nearly 30% of net wealth assets as measured by the Financial Accounts of the U.S..[55] Another takeaway from our results in this paper is that the time-series dynamics of land

[55]As we alluded to in the prior section, comparisons with the balance sheet should be taken with a grain of salt. As it is currently constructed, it is not necessarily apples-to-apples and has a variety of differences between what is currently

value, while generally procyclical over this period, did not align precisely across the U.S.. In fact, while land value reached its peak in 2006 or 2007 for each of the nine census divisions, there was substantial regional variation in the timing of the trough. Regional bottoming out occurred as early as 2009 for some census divisions and as late as 2013 for others. This highlights an important point for economic policymakers and future researchers, that regional variation in land value differs substantially in both severity *and* timing of peak-trough dynamics.

The approach we introduce in this paper opens the door to a host of new extensions. Methodologically, our two-step ML approach, pairing kmeans clustering with gradient boosted trees (GBT), provides a substantial increase in price prediction accuracy over traditional hedonic approaches in the vast majority of circumstances. A key conclusion of this paper is that our model stacking approach, which produces a composite land value based on a weighted combination of ML and hedonic models, outperforms all models individually in terms of price prediction and other out-of-sample tests. Future work can further augment these models, use additional data, and/or incorporate additional models in the stacking procedure to improve the accuracy of these estimates even further.

In some ways, the potential uses and applications of this approach may also present a new frontier for future work. Not only can future work build on these methods, but as we make our code available to everyone, national statistical offices, academic researchers, professional appraisers, and others can take this approach off the shelf to create micro or macro estimates. Whether users want to generate land values, property values, and even borrow our clustering methods (to augment quasi-experimental research designs in urban economics), the transparent methods we provide here may offer countless avenues of new inquiry. In a new era where data is becoming increasingly plentiful, and accounting standards have an increasing emphasis on accurate market values, having comparable and reliable methods may facilitate a host of new applications. A goal of our research is to advance these ends, bridging the data and methods of micro-research and macroeconomic accounts.

measured in the Financial Accounts and how we measure land in this paper. We use this net wealth figure purely for a reference point and not to imply that this is part of an official estimate. The full time series for U.S. net wealth can be found here: https://fred.stlouisfed.org/series/BOGZ1FL892090005Q.

# References

Abraham, K. G., R. S. Jarmin, B. Moyer, and M. D. Shapiro (2019). Big data for 21st century economic statistics.

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

Albouy, D. and M. Shin (2022). A statistical learning approach to land valuation: Optimizing the use of external information. *Journal of Housing Economics*, 101873.

Alonso, W. (1964). The historic and the structural theories of urban form: their implications for urban renewal. *Land Economics 40*(2), 227–231.

Barr, J., F. H. Smith, and S. J. Kulkarni (2018). What's manhattan worth? a land values index from 1950 to 2014. *Regional science and urban economics 70*, 1–19.

Bertomeu, J. (2020). Machine learning improves accounting: discussion, implementation and research opportunities. *Review of Accounting Studies 25*(3), 1135–1155.

Bonhomme, S., T. Lamadon, and E. Manresa (2022). Discretizing unobserved heterogeneity. *Econometrica 90*(2), 625–643.

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

Bourassa, S. C. and M. Hoesli (2022). Hedonic, residual, and matching methods for residential land valuation. *Journal of Housing Economics*, 101870.

Boyd, J. W., K. J. Bagstad, J. C. Ingram, C. D. Shapiro, J. E. Adkins, C. F. Casey, C. S. Duke, P. D. Glynn, E. Goldman, M. Grasso, et al. (2018). The natural capital accounting opportunity: let's really do the numbers. *BioScience 68*(12), 940–943.

Burnett-Isaacs, K., N. Huang, and W. E. Diewert (2020). Developing land and structure price indices for ottawa condominium apartments. *Journal of Official Statistics 36*(4), 763–802.

Clapp, J. M., J. P. Cohen, and T. Lindenthal (2021). Are estimates of rapid growth in urban land values an artifact of the land residual model? *The Journal of Real Estate Finance and Economics*, 1–49.

Clapp, J. M. and T. Lindenthal (2022). Urban land valuation with bundled good and land residual assumptions. *Available at SSRN 4080868*.

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

Coomes, O. T., G. K. MacDonald, and Y. I. P. de Waroux (2018). Geospatial land price data: A public good for global change science and policy. *BioScience 68*(7), 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), 350–359.

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

Davis, M. A., W. D. Larson, S. D. Oliner, and J. Shui (2021). The price of residential land for counties, zip codes, and census tracts in the united states. *Journal of Monetary Economics 118*, 413–431.

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

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

Diewert, E. and C. Shimizu (2017a). Alternative approaches to commercial property price indexes for tokyo. *Review of Income and Wealth 63*(3), 492–519.

Diewert, E. and C. Shimizu (2017b). Alternative land price indexes for commercial. *Journal of Productivity Analysis 3*, 211–248.

Diewert, W. and C. Shimizu (2015). A conceptual framework for commercial property price indexes. *Journal of Statistical Science and Application 3*(9-10), 131–152.

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

Ding, K., B. Lev, X. Peng, T. Sun, and M. A. Vasarhelyi (2020). Machine learning improves accounting estimates: Evidence from insurance payments. *Review of Accounting Studies 25*(3), 1098–1134.

Dye, R. F. and D. P. McMillen (2007). Teardowns and land values in the chicago metropolitan area. *Journal of urban economics 61*(1), 45–63.

Elliott, G. and A. Timmermann (2004). Optimal forecast combinations under general loss functions and forecast error distributions. *Journal of Econometrics 122*(1), 47–79.

Friedman, J., T. Hastie, and R. Tibshirani (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics 28*(2), 337–407.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis 38*(4), 367–378.

Gedal, M. and I. G. Ellen (2018). Valuing urban land: Comparing the use of teardown and vacant land sales. *Regional Science and Urban Economics 70*, 190–203.

George, H. (1884). *Progress and poverty: An inquiry into the cause of industrial depressions, and of increase of want with increase of wealth, the remedy*. W. Reeves.

Gindelsky, M., J. Moulton, K. Wentland, and S. Wentland (2023). When do property taxes matter? tax salience and heterogeneous policy effects. *Tax Salience and Heterogeneous Policy Effects (April 4, 2023)*.

Gindelsky, M., J. Moulton, and S. A. Wentland (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., J. de Haan, et al. (2018). Accounting for spatial variation of land prices in hedonic imputation house price indices: a semi-parametric approach. *Journal of official statistics 34*(3), 695–720.

Granger, C. W. and R. Ramanathan (1984). Improved methods of combining forecasts. *Journal of forecasting 3*(2), 197–204.

Gyourko, J. and A. Saiz (2006). Construction costs and the supply of housing structure. *Journal of regional Science 46*(4), 661–680.

Hansen, B. E. (2008). Least-squares forecast averaging. *Journal of Econometrics 146*(2), 342–350.

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

Jaijairam, P. et al. (2013). Fair value accounting vs. historical cost accounting. *Review of Business Information Systems (RBIS) 17*(1), 1–6.

Johannsson, K. and T. Nguyen (2022). Improving the estimates of land underlying other buildings and structures in the national balance sheet, uk: 2022.

Kuminoff, N. V. and J. C. Pope (2013). The value of residential land and structures during the great housing boom and bust. *Land Economics 89*(1), 1–29.

Larson, W. (2015). New estimates of value of land of the united states.

Larson, W. D. and J. Shui (2022). Land valuation using public records and kriging: Implications for land versus property taxation in cities. *Journal of Housing Economics*, 101871.

Longhofer, S. D. and C. L. Redfearn (2022). Estimating land values using residential sales data. *Journal of Housing Economics*, 101869.

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

McMillen, D. and J. Zabel (2022). Special issue on land valuation: Introduction. *Journal of Housing Economics 58*(PB).

Mills, E. S. (1972). Studies in the structure of the urban economy.

Moulton, J. G., N. J. Sanders, and S. A. Wentland (2018). Toxic assets: How the housing market responds to environmental information shocks.

Moyer, B. C. and A. Dunn (2020). Measuring the gross domestic product (gdp): The ultimate data science project.

Munneke, H. J. (1996). Redevelopment decisions for commercial and industrial properties. *Journal of Urban Economics 39*(2), 229–253.

Munneke, H. J. and K. S. Womack (2015). Neighborhood renewal: the decision to renovate or tear down. *Regional Science and Urban Economics 54*, 99–115.

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

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

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), 29577–29583.

Nolte, C., K. J. Boyle, A. M. Chaudhry, C. M. Clapp, D. Guignet, H. Hennighausen, I. Kushner, Y. Liao, S. Mamun, A. Pollack, et al. (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., R. R. McAllister, C. S. Fletcher, et al. (2015). Decoupling land values in residential property prices: Smoothing methods for hedonic imputed price indices. Technical report, 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. In *International Conference on Real Estate Statistics. EuroStat. Luxembourg. https://ec. europa. eu/eurostat/about/opportunities/conferences*.

Rosenthal, S. S. and R. W. Helsley (1994). Redevelopment and the urban land price gradient. *Journal of Urban Economics 35*(2), 182–200.

Smith, A. (1776). The wealth of nations (e. cannan, ed.). *New York: The Modern Library*.

Somerville, C. T. (1999). The industrial organization of housing supply: Market activity, land supply and the size of homebuilder firms. *Real Estate Economics 27*(4), 669–694.

Theobald, D. M. (2014). Development and applications of a comprehensive land use classification and map for the us. *PloS one 9*(4), e94628.

Timmermann, A. (2006). Forecast combinations. *Handbook of economic forecasting 1*, 135–196.

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

Warren Jr., J. D., K. C. Moffitt, and P. Byrnes (2015). How big data will change accounting. *Accounting horizons 29*(2), 397–407.

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

Zabel, J. (2022). A matching method for land valuation. *Journal of Housing Economics 58*, 101878.

# Appendices

## A Appendix Figures

Figure A1: An Appraisal Report

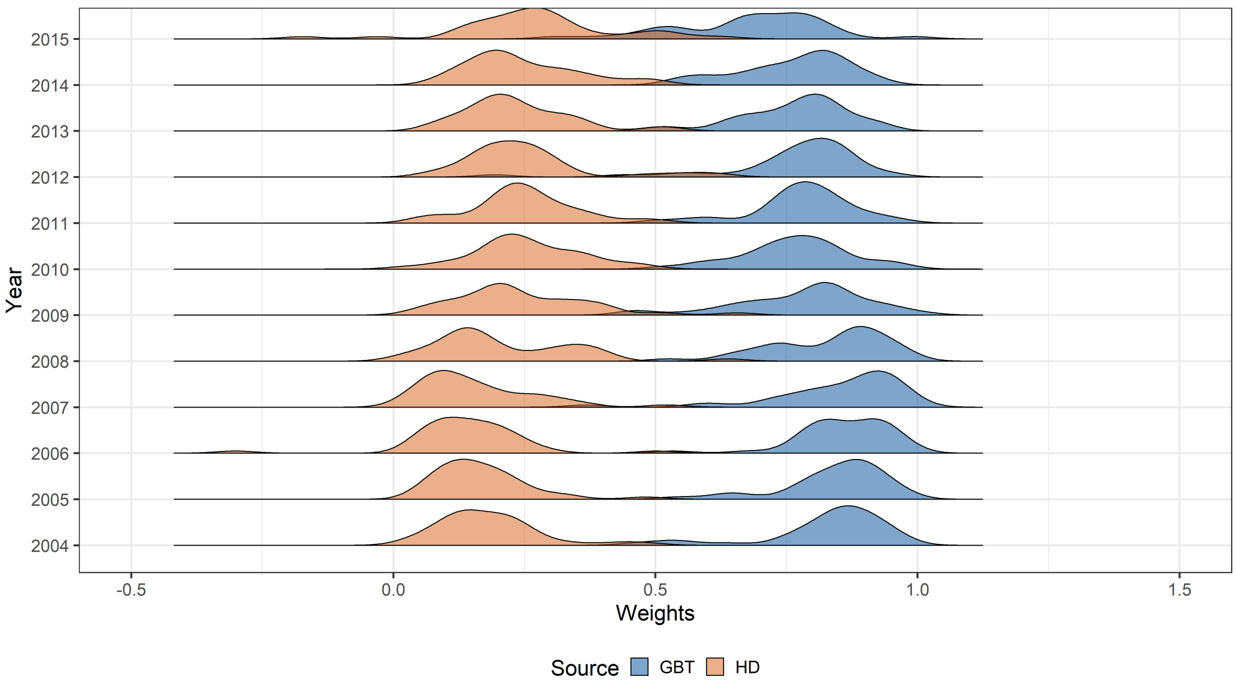
SALES COMPARISON APPROACH

| FEATURE | SUBJECT | COMPARABLE SALE NO. 1 | | COMPARABLE SALE NO. 2 | | COMPARABLE SALE NO. 3 | |
|---|---|---|---|---|---|---|---|
| [redacted] | | [redacted] | | [redacted] | | [redacted] | |
| Proximity to Subject | | 0.20 miles NNE | | 0.35 miles WSW | | 2.06 miles W | |
| Sale Price | $ 565,000 | | $ 550,000 | | $ 492,500 | | $ 635,000 |
| Sale Price/Gross Liv. Area | $ 167.76 sq. ft. | $ 220.00 sq. ft. | | $ 188.77 sq. ft. | | $ 204.84 sq. ft. | |
| Data Source(s) | Insp: 10/2011 | MRIS # [redacted] DOM: 34 | | MRIS # [redacted] DOM: 65 | | MRIS # [redacted] DOM: 5 | |
| Verification Source(s) | Tax/Contract | Tax Record / R. E. Broker | | Tax Record / R. E. Broker | | Tax Record / R. E. Broker | |
| VALUE ADJUSTMENTS | DESCRIPTION | DESCRIPTION | +(-) $ Adjustment | DESCRIPTION | +(-) $ Adjustment | DESCRIPTION | +(-) $ Adjustment |
| Sale or Financing Concessions | 10000 | Conv Mtg $2,000 CH - Typ | | Conv Mtg $10,000 CH - Typ | | VA Mtg $10,000 CH - Typ | |
| Date of Sale/Time | 10/01/11 | 09/30/11 C: 07/11 | 0 | 08/22/11 C: 08/11 | 0 | 08/24/11 C: 06/11 | 0 |
| Location | Valley Mede | St Johns Green | 0 | Valley Mede | | Turf Valley Overlk | -25,000 |
| Leasehold/Fee Simple | Fee Simple | Fee Simple | | Fee Simple | | Fee Simple | |
| Site | 17,660 Sq. Ft. | 16,068 Sq. Ft. | 0 | 21,867 Sq. Ft. | 0 | 14,026 Sq. Ft. | 0 |
| View | Residential | Residential | | Residential | | Parkland | -25,000 |
| Design (Style) | Colonial | Colonial | | Colonial | | Colonial | |
| Quality of Construction | 25% Brk,Vinyl | 25% Brk,Vinyl | | 5% Brk,Vinyl | 10,000 | Alum Siding | 10,000 |
| Actual Age | 23 Yrs | 26 Yrs | 0 | 40 Yrs | 10,000 | 18 Yrs | 0 |
| Condition | Average | Avg+, Upd Kit | -15,000 | Average | | Avg+, Upd Kit | -15,000 |
| Above Grade Room Count | Total 12, Bdrms. 4, Baths 2.5 | Total 10, Bdrms. 4, Baths 2.5 | | Total 10, Bdrms. 4, Baths 2.5 | | Total 12, Bdrms. 4, Baths 2.5 | |
| Gross Living Area 35 | 3,368 sq. ft. | 2,500 sq. ft. | 30,400 | 2,609 sq. ft. | 26,600 | 3,100 sq. ft. | 9,400 |
| Basement & Finished | 1380 Sq Ft | 1320 Sq Ft | 0 | 1232 Sq Ft | 0 | 1400 Sq Ft | 0 |
| Rooms Below Grade | RR,FR,Den,Bath | RR, Bath | 14,000 | RR | 20,000 | Unfinished | 30,000 |
| Functional Utility | Average | Average | | Average | | Average | |
| Heating/Cooling | FA/CAC | FA/CAC | | FA/CAC | | FA/CAC | |
| Energy Efficient Items | Insul W/D | Insul W/D | | Insul W/D | | Insul W/D | |
| Garage/Carport | 2/Gar Att | 2/Gar Att | | 2/Gar Att | | 2/Gar Att | |
| Porch/Patio/Deck | Deck | Patio | 2,000 | Deck | | Deck | |
| OTHER | 1 F/P | 1 F/P | | 1 F/P | | 1 F/P | 1,500 |
| OTHER | Fence | Fence | | Covd Porch | -2,000 | None | 3,000 |
| OTHER | None | Ingr Pool | -10,000 | None | | None | |
| Net Adjustment (Total) | | [X] + [ ] - | $ 21,400 | [X] + [ ] - | $ 64,600 | [ ] + [X] - | $ 11,100 |
| Adjusted Sale Price of Comparables | | Net Adj. 3.9 % Gross Adj. 13.0 % | $ 571,400 | Net Adj. 13.1 % Gross Adj. 13.9 % | $ 557,100 | Net Adj. -1.7 % Gross Adj. 18.7 % | $ 623,900 |

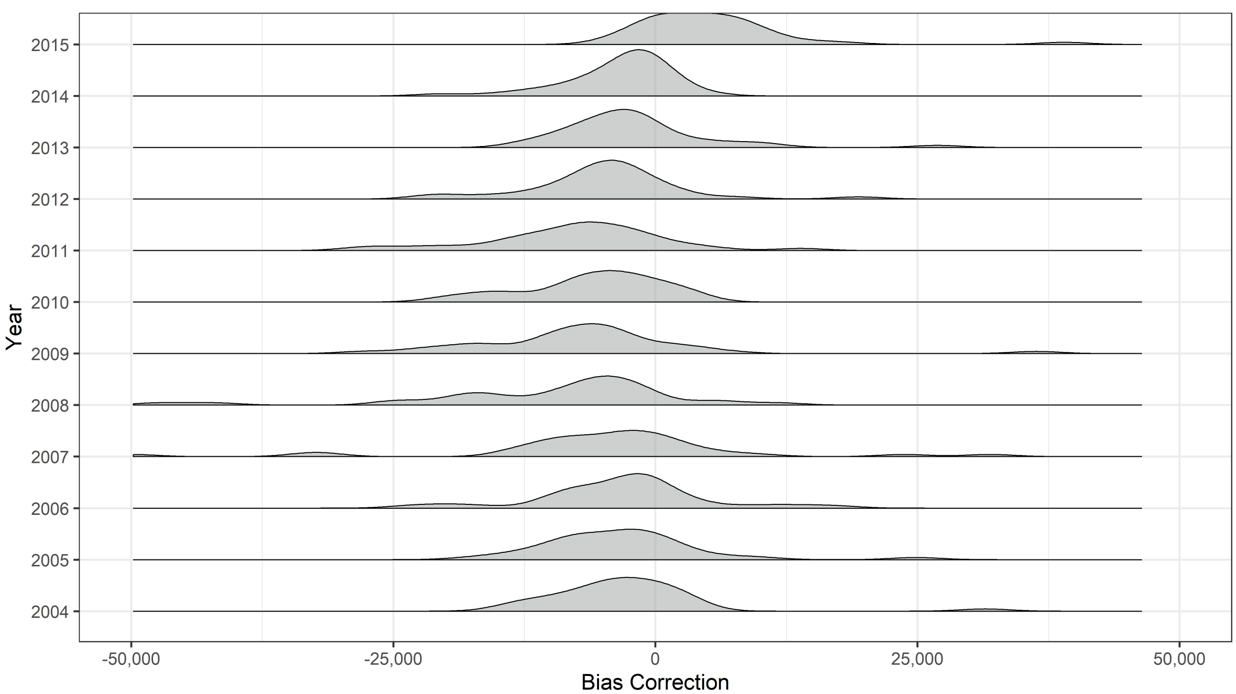
**Note: This figure is one example <https://www.thewendyslaughterteam.com/blog/appraisals by a local Realtor in Columbia, MD describing the appraisal process to clients, which also includes a useful visual of an appraisal report. Citation of this does not constitute BEA's official endorsement of a particular realtor or any information on their website for the general public.**

Figure A2: Forecast Combination Weights

(a)




(b)




**Note:** The composite forecasts are constructed using a linear regression (**?**). The densities in Figure A2a represent the distribution of weights across the thirty-six states for each year and each estimate. Note that these weights do not necessarily sum to one and as a result one can only examine the relative positions of the distributions. Figure A2b outlines the distribution of intercepts (also known as the Bias Correction term) from the combination regressions.

## B Appendix Tables

Table A1: Land Use/Type Cross-Walk from Zillow Data to National Land Use Database: Residential

| Land Type | NLUD | Zillow |
|---|---|---|
| Dense Urban Residential | Households - Dense Urban | Dense Urban < 0.1 acres |
| Urban Residential | Households - Urban | Urban 0.1 to 1 acres |
| | | RI101 Duplex (2 Units, Any Combination) |
| | | RI102 Triplex (3 Units, Any Combination) |
| | | RI107 High-Rise Apartment |
| | | RI108 Boarding House Rooming House Apt Hotel Transient Lod |
| | | RI112 Apartment (Generic) |
| | | RR104 Townhouse |
| | | RR105 Cluster Home |
| | | RR106 Condominium |
| | | RR107 Cooperative |
| | | RR108 Row House |
| | | RR114 Zero Lot Line |
| | | RR116 Patio Home |
| | | RR119 Garden Home |
| | | RR120 Landominium |
| Suburban Residential | Households - Suburban | *These categories below 2.5 acres* |
| | | RR000 Residential General |
| | | RR101 Single Family Residential |
| | | RR113 Bungalow |
| | | RR999 Inferred Single Family Residential |
| Rural Residential | Households - Exurban/Rural | *These categories above 2.5 acres* |
| | | RR000 Residential General |
| | | RR101 Single Family Residential |
| | | RR113 Bungalow |
| | | RR999 Inferred Single Family Residential |
| | | |
| | | *All in these Categories* |
| | | RI109 Mobile Home Park, Trailer Park |
| | | RR102 Rural Residence |
| | | RR103 Mobile Home |
| | | RR115 Manufactured, Modular, Prefabricated Homes |

Table A2: Land Use/Type Cross-Walk from Zillow Data to National Land Use Database: Commercial

| Land Type | NLUD | Zillow |
|---|---|---|
| Commercial | Offices - NAICS 51-56<br>Retail - NAICS 44-45 | CM000 Communication<br>CM100 Cable Tv Station<br>CO101 Commercial/Office/Residential Mixed Used<br>CO102 Commercial/Industrial Mixed Use<br>CO103 Professional Building<br>CO104 Professional Building Multi-Story<br>CO105 Office Building<br>CO106 Office Building Multi-Story<br>CO107 Dental Building<br>CO108 Medical Building<br>CO109 Financial Building<br>CO110 Condominium Offices<br>CO111 Skyscraper, Highrise<br>CO112 Common Area - Commercial Office<br>CO113 Mobile Commercial Units<br>CR000 Commercial - General<br>CR101 Retail Store - General<br>CR102 Multi-Story Store<br>CR103 Store/Office (Mixed Use)<br>CR104 Department Store<br>CR105 Department Store Multi-Story<br>CR106 Mall, Shopping Center<br>CR107 Shopping Plaza, Mini-Mall<br>CR108 Neighborhood Shopping Center, Strip Mall, Enterprise Zone<br>CR109 Grocery, Supermarket<br>CR110 Veterinary, Animal Hospital<br>CR111 Restaurant<br>CR112 Fast Food, Drive Thru Restaurant<br>CR113 Take Out Restaurant (Fast Food)<br>CR114 Bakery<br>CR115 Bar, Tavern<br>CR116 Liquor Store<br>CR117 Convenience Store<br>CR118 Gas Station<br>CR119 Service Station - Full Service<br>CR120 Service Station With Convenience Store<br>CR121 Truck Stop<br>CR122 Vehicle Rentals And Vehicle Sales<br>CR123 Auto Repair, Garage<br>CR124 Car Wash<br>CR125 Dry Cleaner, Laundry<br>CR126 Service Shop<br>CR127 Florist, Nursery, Greenhouse<br>CR128 Wholesale Outlet, Discount Store<br>CR129 Printer - Retail<br>CR130 Mini-Warehouse, Storage<br>CR131 Day Care, Preschool<br>CR132 Hotel<br>CR133 Motel<br>CR134 Hotel/Motel<br>CR135 Hotel Resort<br>CR136 Casino<br>CR137 Parking Garage, Parking Structure<br>CR138 Parking Lot<br>CR139 Funeral Home, Mortuary<br>CR140 Stores & Apartments<br>CR141 Commercial Building, Mail Order, Show Room<br>CR142 Appliance Store<br>CR143 Kennel<br>CR144 Laundromat<br>CR145 Nightclub, Cocktail Lounge<br>CR146 Farm Supply & Equipment<br>CR147 Garden Center, Home Improvement<br>CR148 Commercial Condominium<br>CR149 Drug Store Pharmacy<br>CR150 Bed & Breakfast<br>CR151 Shopping Center Common Area |

Table A3: Land Use/Type Cross-Walk from Zillow Data to National Land Use Database: Industrial

| Land Type | NLUD | Zillow |
|---|---|---|
| Industrial | Manufacturing NAICS 31-33 | IH000 Industrial Heavy - General |
| | | IH101 Distribution Warehouse |
| | | IH102 Mining |
| | | IH103 Storage Yard |
| | | IH104 Distillery, Brewery, Bottling |
| | | IH105 Refinery, Petroleum Products |
| | | IH106 Mill |
| | | IH107 Factory |
| | | IH108 Processing Plant |
| | | IH109 Lumberyard, Building Materials |
| | | IH110 Shipyard, Storage |
| | | IH111 Slaughter House, Stockyard |
| | | IH112 Waste Disposal, Sewage |
| | | IH113 Quarries |
| | | IH114 Heavy Manufacturing |
| | | IH115 Labor Camp |
| | | IH116 Winery |
| | | IH117 Chemical Plant |
| | | IH118 Foundry, Industrial Plant |
| | | IH119 Cannery |
| | | IH120 Bulk Storage, Tanks |
| | | IH121 Grain Elevator |
| | | IH122 Dump Site |
| | | IH123 Cold Storage |
| | | IH124 Transportation - Industrial |
| | | IN000 Industrial - General |
| | | IN101 Manufacturing (Light) |
| | | IN102 Light Industrial |
| | | IN103 Warehouse |
| | | IN104 Storage Yard, Open Storage |
| | | IN105 Food Packing, Packing Plant |
| | | IN106 Assembly Plant |
| | | IN107 Food Processing |
| | | IN108 Recycling |
| | | IN109 Condominium (Industrial) |
| | | IN110 Laboratory, Research Facility, R&D Facility |
| | | IN111 Industrial Park |
| | | IN112 Multi-Tenant Industrial Building |
| | | IN113 Marine Facility, Boat Repairs |
| | | IN114 Lumber & Wood Products Mfg |
| | | IN115 Paper Product Mfg & Related Products |
| | | IN116 Printing & Publishing |
| | | IN117 Loft Building |
| | | IN118 Construction/Contracting Services |
| | | IN119 Common Area (Industrial) |

Table A4: Land Use/Type Cross-Walk from Zillow Data to National Land Use Database: Agricultural

| Land Type | NLUD | Zillow |
|---|---|---|
| Agricultural | Farms NAICS 111<br>Livestock NAICS 112 | AG000 Agricultural General<br>AG101 Farm (Irrigated Or Dry)<br>AG103 Poultry Farm<br>AG106 Crop Land, Field Crops, Row Crops<br>AG107 Orchard (Fruit, Nut)<br>AG108 Vineyard<br>AG113 Grove<br>AG116 Horticulture, Growing Houses, Ornamental<br>AG118 Truck Crops<br>AG121 Rural Improved, Nonresidential<br>VL108 Agricultural, Unimproved Vacant Land<br>AG102 Dairy Farm<br>AG104 Ranch<br>AG105 Range Land, Grazing Land<br>AG112 Pasture, Meadow<br>AG114 Feedlot<br>AG115 Livestock |

Table A5: SFR Hedonic PPA

| State | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ALABAMA | * | * | 81,145 | 70,003 | 66,155 | 64,506 | 68,231 | 70,732 | 66,343 | 68,727 | 68,265 | 67,693 |
| ARIZONA | 408,691 | 559,906 | 675,371 | 663,568 | 499,420 | 388,045 | 377,311 | 335,183 | 368,299 | 419,166 | 454,010 | 498,084 |
| ARKANSAS | 53,711 | 70,840 | 76,584 | 75,217 | 66,811 | 65,443 | 66,632 | 63,935 | 56,448 | 69,025 | 74,751 | 71,825 |
| CALIFORNIA | 1,427,330 | 1,693,454 | 1,750,386 | 1,582,709 | 1,152,677 | 1,012,939 | 1,078,160 | 1,014,426 | 1,046,745 | 1,263,332 | 1,424,628 | 1,566,155 |
| COLORADO | 467,510 | 499,986 | 524,712 | 520,410 | 465,900 | 419,160 | 405,335 | 382,996 | 410,481 | 454,694 | 513,269 | 594,851 |
| CONNECTICUT | 255,785 | 286,898 | 292,934 | 284,365 | 251,873 | 222,039 | 216,676 | 205,178 | 192,116 | 194,523 | 198,823 | 202,700 |
| DELAWARE | 302,464 | 358,669 | 406,319 | 579,020 | 655,571 | 487,704 | 353,364 | 293,044 | 278,752 | 311,813 | 331,246 | 355,215 |
| FLORIDA | 382,135 | 534,004 | 635,053 | 556,288 | 371,188 | 258,090 | 237,424 | 227,595 | 247,753 | 272,775 | 307,741 | 349,902 |
| GEORGIA | 117,902 | 129,248 | 147,531 | 160,833 | 136,371 | 109,756 | 96,426 | 82,725 | 82,970 | 96,991 | 117,200 | 124,427 |
| ILLINOIS | 517,304 | 575,268 | 618,059 | 595,587 | 489,142 | 375,687 | 343,480 | 296,543 | 280,049 | 306,452 | 336,061 | 359,358 |
| IOWA | 165,712 | 192,764 | 201,185 | 201,002 | 187,031 | 174,809 | 169,910 | 169,205 | 181,568 | 192,770 | 203,480 | 202,852 |
| KENTUCKY | 68,749 | 88,969 | 114,144 | 121,237 | 114,080 | 107,830 | 99,292 | 100,211 | 99,363 | 106,637 | 105,009 | 113,217 |
| LOUISIANA | * | 157,871 | 94,553 | 210,330 | 227,466 | 194,504 | 165,925 | 238,384 | 85,451 | 59,157 | 54,575 | 51,486 |
| MARYLAND | 412,550 | 493,920 | 523,668 | 522,374 | 456,052 | 399,418 | 364,066 | 349,156 | 347,074 | 367,872 | 373,933 | 378,153 |
| MASSACHUSETTS | 396,001 | 428,886 | 410,096 | 382,781 | 333,137 | 309,563 | 309,303 | 290,254 | 278,710 | 285,533 | 273,918 | 280,705 |
| MICHIGAN | 272,053 | 285,355 | 260,628 | 214,752 | 153,194 | 136,196 | 147,562 | 129,011 | 121,947 | 183,595 | 186,331 | 162,515 |
| MINNESOTA | 321,622 | 343,015 | 337,334 | 323,642 | 274,668 | 250,277 | 248,031 | 224,346 | 228,445 | 239,830 | 252,739 | 268,372 |
| MISSOURI | 166,865 | 156,795 | 163,716 | 151,356 | 137,941 | 130,789 | 134,158 | 128,007 | 132,952 | 126,155 | 127,884 | 94,930 |
| NEBRASKA | 223,528 | 182,395 | 108,675 | 134,651 | 139,022 | 171,585 | 178,657 | 184,654 | 201,598 | 213,492 | 217,493 | 220,120 |
| NEVADA | 864,128 | 1,108,222 | 1,026,636 | 916,743 | 656,646 | 501,520 | 489,744 | 777,959 | 729,349 | 605,525 | 656,970 | 671,144 |
| NEW HAMPSHIRE | 154,273 | 170,762 | 171,357 | 163,730 | 140,538 | 125,291 | 124,920 | 119,570 | 120,521 | 122,379 | 125,338 | 137,679 |
| NEW JERSEY | 678,413 | 779,433 | 798,584 | 757,769 | 660,444 | 592,715 | 579,159 | 549,434 | 519,634 | 522,866 | 541,676 | 563,105 |
| NEW YORK | 369,964 | 404,805 | 431,584 | 420,692 | 397,469 | 360,902 | 357,155 | 332,303 | 337,231 | 334,741 | 335,624 | 354,874 |
| NORTH CAROLINA | 98,499 | 127,546 | 143,645 | 153,393 | 142,063 | 126,095 | 113,383 | 108,189 | 109,014 | 117,025 | 122,715 | 128,935 |
| OHIO | 163,515 | 171,628 | 166,342 | 151,352 | 126,940 | 120,268 | 116,754 | 105,945 | 103,483 | 107,060 | 115,188 | 117,437 |
| OKLAHOMA | 86,562 | 124,742 | 137,983 | 148,005 | 148,288 | 143,102 | 129,518 | 126,230 | 135,857 | 138,533 | 154,334 | 151,597 |
| OREGON | 352,630 | 430,514 | 509,560 | 532,297 | 491,316 | 421,623 | 396,635 | 370,879 | 378,430 | 417,246 | 445,272 | 491,243 |
| PENNSYLVANIA | 232,214 | 260,847 | 270,447 | 274,860 | 258,981 | 274,839 | 278,605 | 266,717 | 256,257 | 253,086 | 246,373 | 265,201 |
| RHODE ISLAND | 401,359 | 452,339 | 453,091 | 419,841 | 352,713 | 304,309 | 294,718 | 313,928 | 398,144 | 373,997 | 324,647 | 338,271 |
| SOUTH CAROLINA | 148,637 | 145,798 | 137,861 | 138,914 | 123,016 | 104,229 | 85,976 | 73,193 | 83,663 | 100,716 | 106,243 | 107,300 |
| SOUTH DAKOTA | * | 34,078 | 143,929 | 619,647 | 265,440 | 338,285 | 163,644 | 247,906 | 209,452 | 254,624 | 239,685 | * |
| TENNESSEE | 90,830 | 103,637 | 110,983 | 112,424 | 101,351 | 89,948 | 80,869 | 74,131 | 72,788 | 78,812 | 90,831 | 98,966 |
| VIRGINIA | 208,785 | 226,441 | 277,421 | 250,216 | 219,136 | 194,006 | 194,037 | 182,780 | 183,984 | 191,064 | 190,268 | 197,818 |
| WASHINGTON | 390,682 | 468,595 | 548,277 | 563,684 | 510,312 | 440,485 | 420,277 | 379,160 | 380,311 | 401,298 | 434,652 | 472,820 |
| WEST VIRGINIA | 51,910 | 78,149 | 88,976 | 94,833 | 86,707 | 86,795 | 85,919 | 86,397 | 83,923 | 85,416 | 81,917 | 93,033 |
| WISCONSIN | 454,765 | 413,963 | 420,630 | 416,844 | 376,477 | 337,544 | 322,717 | 296,097 | 296,807 | 303,724 | 312,005 | 310,286 |

Table A6: SFR Hedonic Leverage

| State | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ALABAMA | * | * | 0.53 | 0.41 | 0.38 | 0.40 | 0.40 | 0.45 | 0.40 | 0.41 | 0.40 | 0.39 |
| ARIZONA | 0.53 | 0.55 | 0.56 | 0.58 | 0.58 | 0.58 | 0.57 | 0.56 | 0.54 | 0.53 | 0.54 | 0.56 |
| ARKANSAS | 0.45 | 0.57 | 0.54 | 0.50 | 0.47 | 0.51 | 0.49 | 0.49 | 0.48 | 0.51 | 0.51 | 0.44 |
| CALIFORNIA | 0.65 | 0.65 | 0.65 | 0.65 | 0.63 | 0.62 | 0.62 | 0.60 | 0.59 | 0.59 | 0.60 | 0.61 |
| COLORADO | 0.59 | 0.58 | 0.57 | 0.56 | 0.54 | 0.51 | 0.49 | 0.49 | 0.50 | 0.50 | 0.51 | 0.53 |
| CONNECTICUT | 0.52 | 0.52 | 0.51 | 0.50 | 0.49 | 0.48 | 0.47 | 0.46 | 0.44 | 0.44 | 0.44 | 0.45 |
| DELAWARE | 0.54 | 0.58 | 0.55 | 0.60 | 0.61 | 0.69 | 0.59 | 0.55 | 0.53 | 0.55 | 0.56 | 0.58 |
| FLORIDA | 0.53 | 0.56 | 0.59 | 0.58 | 0.53 | 0.49 | 0.49 | 0.48 | 0.47 | 0.48 | 0.49 | 0.51 |
| GEORGIA | 0.49 | 0.49 | 0.50 | 0.52 | 0.50 | 0.46 | 0.43 | 0.42 | 0.41 | 0.42 | 0.45 | 0.45 |
| ILLINOIS | 0.56 | 0.57 | 0.58 | 0.56 | 0.53 | 0.49 | 0.44 | 0.41 | 0.40 | 0.40 | 0.41 | 0.42 |
| IOWA | 0.55 | 0.56 | 0.56 | 0.55 | 0.52 | 0.50 | 0.48 | 0.48 | 0.49 | 0.50 | 0.49 | 0.46 |
| KENTUCKY | 0.31 | 0.36 | 0.40 | 0.43 | 0.43 | 0.42 | 0.39 | 0.39 | 0.38 | 0.39 | 0.37 | 0.37 |
| LOUISIANA | * | 1.07 | 0.63 | 1.35 | 1.26 | 1.17 | 1.08 | 1.58 | 0.57 | 0.38 | 0.34 | 0.31 |
| MARYLAND | 0.61 | 0.61 | 0.60 | 0.60 | 0.59 | 0.58 | 0.57 | 0.56 | 0.55 | 0.54 | 0.54 | 0.54 |
| MASSACHUSETTS | 0.55 | 0.56 | 0.55 | 0.53 | 0.52 | 0.51 | 0.51 | 0.50 | 0.49 | 0.47 | 0.54 | 0.57 |
| MICHIGAN | 0.66 | 0.66 | 0.62 | 0.58 | 0.54 | 0.51 | 0.48 | 0.45 | 0.47 | 0.48 | 0.45 | 0.43 |
| MINNESOTA | 0.54 | 0.54 | 0.52 | 0.51 | 0.51 | 0.51 | 0.49 | 0.48 | 0.46 | 0.44 | 0.44 | 0.45 |
| MISSOURI | 0.51 | 0.45 | 0.48 | 0.45 | 0.44 | 0.43 | 0.43 | 0.43 | 0.42 | 0.41 | 0.41 | 0.35 |
| NEBRASKA | 0.58 | 0.45 | 0.33 | 0.39 | 0.40 | 0.42 | 0.45 | 0.47 | 0.49 | 0.49 | 0.47 | 0.46 |
| NEVADA | 0.65 | 0.69 | 0.61 | 0.61 | 0.60 | 0.61 | 0.60 | 1.05 | 0.97 | 0.65 | 0.62 | 0.60 |
| NEW HAMPSHIRE | 0.51 | 0.53 | 0.54 | 0.53 | 0.51 | 0.51 | 0.51 | 0.51 | 0.49 | 0.48 | 0.47 | 0.48 |
| NEW JERSEY | 0.66 | 0.66 | 0.65 | 0.64 | 0.62 | 0.62 | 0.62 | 0.62 | 0.59 | 0.57 | 0.58 | 0.58 |
| NEW YORK | 0.61 | 0.60 | 0.60 | 0.59 | 0.59 | 0.56 | 0.58 | 0.55 | 0.55 | 0.53 | 0.52 | 0.53 |
| NORTH CAROLINA | 0.40 | 0.46 | 0.48 | 0.48 | 0.46 | 0.44 | 0.41 | 0.41 | 0.41 | 0.42 | 0.42 | 0.42 |
| OHIO | 0.58 | 0.59 | 0.57 | 0.56 | 0.53 | 0.51 | 0.48 | 0.47 | 0.45 | 0.43 | 0.43 | 0.41 |
| OKLAHOMA | 0.38 | 0.48 | 0.50 | 0.52 | 0.52 | 0.50 | 0.47 | 0.46 | 0.49 | 0.48 | 0.48 | 0.46 |
| OREGON | 0.57 | 0.59 | 0.59 | 0.60 | 0.60 | 0.59 | 0.59 | 0.59 | 0.58 | 0.57 | 0.56 | 0.57 |
| PENNSYLVANIA | 0.58 | 0.60 | 0.59 | 0.59 | 0.58 | 0.59 | 0.59 | 0.58 | 0.55 | 0.54 | 0.53 | 0.52 |
| RHODE ISLAND | 0.57 | 0.58 | 0.59 | 0.58 | 0.56 | 0.55 | 0.54 | 0.57 | 0.55 | 0.52 | 0.50 | 0.52 |
| SOUTH CAROLINA | 0.64 | 0.55 | 0.49 | 0.47 | 0.45 | 0.41 | 0.36 | 0.32 | 0.35 | 0.40 | 0.39 | 0.35 |
| SOUTH DAKOTA | * | 0.05 | 0.25 | 1.61 | 0.36 | 0.61 | 0.46 | 0.68 | 0.52 | 0.55 | 0.53 | * |
| TENNESSEE | 0.52 | 0.53 | 0.52 | 0.51 | 0.49 | 0.46 | 0.44 | 0.42 | 0.41 | 0.42 | 0.45 | 0.44 |
| VIRGINIA | 0.50 | 0.47 | 0.48 | 0.45 | 0.47 | 0.43 | 0.46 | 0.44 | 0.42 | 0.42 | 0.41 | 0.41 |
| WASHINGTON | 0.58 | 0.59 | 0.60 | 0.57 | 0.56 | 0.55 | 0.56 | 0.54 | 0.53 | 0.52 | 0.52 | 0.51 |
| WEST VIRGINIA | 0.31 | 0.44 | 0.47 | 0.47 | 0.44 | 0.43 | 0.43 | 0.43 | 0.40 | 0.40 | 0.38 | 0.39 |
| WISCONSIN | 0.80 | 0.68 | 0.67 | 0.66 | 0.64 | 0.62 | 0.58 | 0.57 | 0.57 | 0.55 | 0.54 | 0.52 |

Table A7: SFR Gradient Boosted Trees PPA

| State | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ALABAMA | * | * | 56,484 | 56,280 | 34,582 | 48,006 | 23,951 | 22,560 | 30,909 | 31,998 | 31,793 | 33,240 |
| ARIZONA | 219,785 | 317,626 | 396,083 | 375,746 | 314,277 | 210,375 | 190,772 | 153,196 | 164,563 | 202,408 | 223,137 | 269,972 |
| ARKANSAS | 26,069 | 32,261 | 37,239 | 39,716 | 36,305 | 33,261 | 30,144 | 31,085 | 25,162 | 23,374 | 23,410 | 28,498 |
| CALIFORNIA | 1,006,465 | 1,272,812 | 1,347,152 | 1,354,420 | 884,524 | 670,368 | 646,596 | 606,630 | 629,205 | 839,324 | 1,026,151 | 1,269,754 |
| COLORADO | 418,321 | 434,077 | 429,257 | 426,251 | 375,913 | 343,969 | 339,595 | 331,934 | 370,004 | 441,379 | 500,840 | 589,530 |
| CONNECTICUT | 370,744 | 417,398 | 429,574 | 423,713 | 408,184 | 374,868 | 370,592 | 356,642 | 344,567 | 346,096 | 336,944 | 337,261 |
| DELAWARE | 334,867 | 354,329 | 408,923 | 363,920 | 416,190 | 367,869 | 373,594 | 305,147 | 304,778 | 283,751 | 286,875 | 328,798 |
| FLORIDA | 231,004 | 309,882 | 359,730 | 313,368 | 257,718 | 154,426 | 109,158 | 90,807 | 109,632 | 118,862 | 145,098 | 168,592 |
| GEORGIA | 97,957 | 101,032 | 103,112 | 93,136 | 89,211 | 72,940 | 71,749 | 50,429 | 45,589 | 47,339 | 57,288 | 66,562 |
| ILLINOIS | 514,255 | 571,467 | 582,597 | 574,614 | 445,590 | 397,467 | 347,371 | 295,977 | 309,456 | 328,636 | 365,859 | 407,688 |
| IOWA | 152,599 | 161,041 | 154,946 | 162,042 | 157,120 | 148,236 | 162,605 | 158,207 | 170,314 | 170,293 | 175,920 | 194,096 |
| KENTUCKY | 103,796 | 104,609 | 107,244 | 101,566 | 100,141 | 102,067 | 92,360 | 82,522 | 75,961 | 72,672 | 80,621 | 98,026 |
| LOUISIANA | 16,546 | 22,291 | 25,927 | 30,097 | 30,540 | 32,672 | 28,761 | 25,207 | 19,503 | 17,185 | 14,054 | 25,309 |
| MARYLAND | 314,788 | 388,217 | 441,279 | 447,411 | 392,989 | 350,792 | 319,299 | 303,231 | 300,150 | 318,320 | 335,728 | 350,803 |
| MASSACHUSETTS | 531,135 | 568,347 | 559,391 | 537,515 | 478,558 | 460,130 | 451,219 | 437,531 | 404,840 | 396,041 | 408,056 | 433,881 |
| MICHIGAN | 177,761 | 185,583 | 189,738 | 175,339 | 149,494 | 122,459 | 121,105 | 110,502 | 95,429 | 112,777 | 122,473 | 124,122 |
| MINNESOTA | 422,550 | 440,473 | 458,125 | 427,406 | 349,637 | 317,973 | 300,432 | 306,983 | 314,840 | 327,213 | 378,603 | 400,675 |
| MISSOURI | 130,178 | 129,298 | 141,194 | 136,207 | 117,499 | 132,249 | 131,036 | 110,022 | 115,382 | 105,080 | 105,148 | 94,469 |
| NEBRASKA | 155,241 | 166,413 | 173,392 | 188,272 | 189,172 | 177,991 | 154,161 | 151,663 | 150,178 | 167,658 | 169,702 | 196,067 |
| NEVADA | 524,129 | 527,216 | 550,896 | 519,001 | 417,672 | 251,211 | 246,915 | 193,089 | 175,380 | 191,574 | 256,022 | 349,502 |
| NEW HAMPSHIRE | 204,700 | 221,766 | 225,495 | 218,730 | 187,897 | 170,220 | 167,235 | 162,815 | 167,848 | 175,529 | 179,010 | 196,971 |
| NEW JERSEY | 441,342 | 512,948 | 517,143 | 510,772 | 454,010 | 417,341 | 377,489 | 321,505 | 305,530 | 290,884 | 310,328 | 318,734 |
| NEW YORK | 314,593 | 363,681 | 369,319 | 363,071 | 336,830 | 301,293 | 289,743 | 274,879 | 276,631 | 286,308 | 290,178 | 301,600 |
| NORTH CAROLINA | 55,285 | 78,744 | 81,257 | 78,353 | 79,213 | 78,239 | 73,081 | 70,682 | 67,771 | 71,043 | 68,106 | 67,754 |
| OHIO | 128,903 | 125,246 | 120,475 | 113,042 | 100,643 | 114,955 | 89,591 | 87,495 | 83,048 | 86,006 | 89,319 | 91,591 |
| OKLAHOMA | 40,169 | 49,310 | 55,152 | 54,229 | 57,348 | 50,774 | 51,689 | 50,301 | 39,336 | 44,700 | 41,873 | 40,094 |
| OREGON | 388,630 | 448,893 | 508,916 | 540,270 | 507,531 | 469,350 | 401,046 | 363,800 | 378,466 | 417,933 | 470,789 | 545,782 |
| PENNSYLVANIA | 199,891 | 196,697 | 196,168 | 199,321 | 205,547 | 196,071 | 189,161 | 168,516 | 154,947 | 150,758 | 149,854 | 157,569 |
| RHODE ISLAND | 472,945 | 500,675 | 504,565 | 495,447 | 417,483 | 340,277 | 338,997 | 295,928 | 296,783 | 319,705 | 332,708 | 368,920 |
| SOUTH CAROLINA | 71,328 | 73,778 | 73,615 | 81,560 | 74,843 | 75,471 | 62,971 | 57,657 | 60,018 | 66,848 | 77,328 | 89,931 |
| SOUTH DAKOTA | * | 211,632 | 195,697 | 238,123 | 346,192 | 382,554 | 330,457 | 255,626 | 220,869 | 200,598 | 212,099 | 204,793 |
| TENNESSEE | 41,179 | 41,968 | 45,501 | 43,262 | 43,985 | 42,390 | 40,268 | 41,303 | 38,175 | 32,419 | 34,200 | 42,576 |
| VIRGINIA | 233,034 | 291,155 | 309,182 | 288,419 | 217,415 | 210,094 | 209,314 | 206,535 | 226,709 | 225,156 | 241,661 | 258,251 |
| WASHINGTON | 369,932 | 401,446 | 474,363 | 496,362 | 466,983 | 428,962 | 417,651 | 354,171 | 357,112 | 408,279 | 423,960 | 476,628 |
| WEST VIRGINIA | 90,045 | 94,622 | 90,134 | 86,067 | 84,934 | 94,294 | 85,666 | 81,125 | 69,867 | 77,298 | 72,864 | 94,817 |
| WISCONSIN | 297,552 | 348,698 | 365,351 | 355,691 | 341,537 | 325,803 | 305,931 | 285,182 | 268,984 | 280,443 | 317,212 | 325,718 |

Table A8: SFR Gradient Boosted Trees Leverage

| State | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ALABAMA | * | * | 0.36 | 0.34 | 0.21 | 0.31 | 0.15 | 0.15 | 0.20 | 0.21 | 0.20 | 0.20 |
| ARIZONA | 0.29 | 0.32 | 0.35 | 0.35 | 0.37 | 0.31 | 0.29 | 0.25 | 0.25 | 0.26 | 0.27 | 0.31 |
| ARKANSAS | 0.23 | 0.27 | 0.29 | 0.29 | 0.27 | 0.27 | 0.24 | 0.25 | 0.20 | 0.18 | 0.18 | 0.20 |
| CALIFORNIA | 0.47 | 0.50 | 0.51 | 0.54 | 0.46 | 0.40 | 0.37 | 0.36 | 0.36 | 0.40 | 0.43 | 0.49 |
| COLORADO | 0.53 | 0.52 | 0.49 | 0.48 | 0.45 | 0.43 | 0.42 | 0.43 | 0.45 | 0.49 | 0.51 | 0.54 |
| CONNECTICUT | 0.78 | 0.79 | 0.79 | 0.78 | 0.79 | 0.80 | 0.79 | 0.79 | 0.78 | 0.78 | 0.76 | 0.76 |
| DELAWARE | 0.58 | 0.63 | 0.69 | 0.61 | 0.66 | 0.64 | 0.63 | 0.58 | 0.59 | 0.52 | 0.51 | 0.57 |
| FLORIDA | 0.37 | 0.38 | 0.39 | 0.36 | 0.37 | 0.29 | 0.22 | 0.19 | 0.22 | 0.21 | 0.24 | 0.25 |
| GEORGIA | 0.41 | 0.39 | 0.37 | 0.33 | 0.34 | 0.31 | 0.33 | 0.25 | 0.23 | 0.22 | 0.24 | 0.26 |
| ILLINOIS | 0.56 | 0.57 | 0.55 | 0.54 | 0.47 | 0.50 | 0.45 | 0.42 | 0.45 | 0.44 | 0.46 | 0.50 |
| IOWA | 0.49 | 0.49 | 0.45 | 0.46 | 0.44 | 0.41 | 0.45 | 0.44 | 0.46 | 0.44 | 0.44 | 0.46 |
| KENTUCKY | 0.42 | 0.42 | 0.40 | 0.37 | 0.37 | 0.39 | 0.35 | 0.32 | 0.29 | 0.27 | 0.29 | 0.34 |
| LOUISIANA | 0.13 | 0.17 | 0.20 | 0.21 | 0.20 | 0.22 | 0.20 | 0.17 | 0.14 | 0.12 | 0.09 | 0.16 |
| MARYLAND | 0.50 | 0.50 | 0.53 | 0.53 | 0.52 | 0.52 | 0.50 | 0.48 | 0.48 | 0.48 | 0.50 | 0.52 |
| MASSACHUSETTS | 0.77 | 0.77 | 0.76 | 0.75 | 0.73 | 0.75 | 0.73 | 0.73 | 0.76 | 0.75 | 0.78 | 0.84 |
| MICHIGAN | 0.44 | 0.44 | 0.46 | 0.47 | 0.48 | 0.46 | 0.45 | 0.42 | 0.37 | 0.37 | 0.37 | 0.35 |
| MINNESOTA | 0.72 | 0.71 | 0.72 | 0.69 | 0.64 | 0.66 | 0.62 | 0.67 | 0.65 | 0.63 | 0.69 | 0.69 |
| MISSOURI | 0.37 | 0.36 | 0.39 | 0.38 | 0.35 | 0.41 | 0.41 | 0.37 | 0.38 | 0.34 | 0.34 | 0.37 |
| NEBRASKA | 0.41 | 0.42 | 0.47 | 0.50 | 0.50 | 0.44 | 0.38 | 0.38 | 0.36 | 0.39 | 0.37 | 0.40 |
| NEVADA | 0.38 | 0.35 | 0.34 | 0.35 | 0.39 | 0.30 | 0.30 | 0.26 | 0.23 | 0.21 | 0.25 | 0.32 |
| NEW HAMPSHIRE | 0.69 | 0.69 | 0.70 | 0.70 | 0.66 | 0.66 | 0.65 | 0.66 | 0.66 | 0.67 | 0.65 | 0.67 |
| NEW JERSEY | 0.45 | 0.46 | 0.44 | 0.44 | 0.42 | 0.43 | 0.40 | 0.36 | 0.35 | 0.32 | 0.33 | 0.33 |
| NEW YORK | 0.55 | 0.57 | 0.55 | 0.54 | 0.52 | 0.50 | 0.49 | 0.47 | 0.48 | 0.48 | 0.47 | 0.47 |
| NORTH CAROLINA | 0.22 | 0.29 | 0.28 | 0.25 | 0.26 | 0.27 | 0.26 | 0.26 | 0.25 | 0.25 | 0.24 | 0.22 |
| OHIO | 0.46 | 0.43 | 0.41 | 0.41 | 0.41 | 0.49 | 0.38 | 0.39 | 0.36 | 0.36 | 0.34 | 0.34 |
| OKLAHOMA | 0.17 | 0.20 | 0.21 | 0.20 | 0.21 | 0.18 | 0.19 | 0.18 | 0.14 | 0.15 | 0.14 | 0.13 |
| OREGON | 0.63 | 0.62 | 0.60 | 0.61 | 0.61 | 0.65 | 0.58 | 0.57 | 0.58 | 0.57 | 0.60 | 0.64 |
| PENNSYLVANIA | 0.53 | 0.48 | 0.45 | 0.46 | 0.49 | 0.48 | 0.48 | 0.44 | 0.41 | 0.39 | 0.39 | 0.39 |
| RHODE ISLAND | 0.68 | 0.65 | 0.66 | 0.67 | 0.65 | 0.60 | 0.60 | 0.56 | 0.55 | 0.57 | 0.57 | 0.62 |
| SOUTH CAROLINA | 0.33 | 0.31 | 0.28 | 0.28 | 0.27 | 0.30 | 0.27 | 0.25 | 0.26 | 0.27 | 0.30 | 0.32 |
| SOUTH DAKOTA | * | 0.76 | 0.56 | 0.72 | 0.56 | 0.81 | 0.85 | 0.64 | 0.54 | 0.46 | 0.47 | 0.48 |
| TENNESSEE | 0.26 | 0.24 | 0.24 | 0.21 | 0.23 | 0.23 | 0.22 | 0.23 | 0.22 | 0.18 | 0.17 | 0.20 |
| VIRGINIA | 0.50 | 0.51 | 0.50 | 0.49 | 0.43 | 0.44 | 0.44 | 0.43 | 0.47 | 0.44 | 0.45 | 0.47 |
| WASHINGTON | 0.53 | 0.51 | 0.52 | 0.51 | 0.52 | 0.54 | 0.55 | 0.50 | 0.49 | 0.52 | 0.50 | 0.51 |
| WEST VIRGINIA | 0.54 | 0.51 | 0.45 | 0.42 | 0.42 | 0.48 | 0.43 | 0.41 | 0.34 | 0.37 | 0.34 | 0.40 |
| WISCONSIN | 0.52 | 0.58 | 0.59 | 0.57 | 0.58 | 0.59 | 0.55 | 0.55 | 0.52 | 0.52 | 0.57 | 0.56 |

Table A9: SFR Composite PPA

| State | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ALABAMA | * | * | 60,929 | 55,720 | 36,273 | 53,298 | 75,873 | 77,938 | 84,861 | 36,362 | 47,219 | 38,414 |
| ARIZONA | 228,217 | 331,032 | 411,516 | 368,358 | 316,585 | 226,639 | 227,422 | 186,999 | 197,397 | 233,022 | 263,210 | 315,110 |
| ARKANSAS | 26,346 | 26,885 | 34,407 | 35,471 | 36,821 | 29,689 | 27,108 | 31,042 | 26,729 | 21,031 | 30,254 | 36,023 |
| CALIFORNIA | 1,184,669 | 1,438,573 | 1,507,175 | 1,386,489 | 985,337 | 849,578 | 807,863 | 759,658 | 803,926 | 1,032,300 | 1,178,791 | 1,387,070 |
| COLORADO | 408,754 | 420,514 | 406,977 | 384,555 | 371,261 | 352,200 | 355,826 | 311,667 | 373,668 | 430,873 | 491,569 | 566,871 |
| CONNECTICUT | 333,619 | 395,633 | 397,627 | 396,876 | 338,856 | 313,040 | 308,960 | 295,983 | 286,843 | 285,047 | 269,245 | 271,107 |
| DELAWARE | 331,627 | 361,912 | 418,082 | 388,550 | 422,770 | 374,845 | 358,901 | 295,502 | 293,799 | 293,934 | 294,091 | 329,541 |
| FLORIDA | 241,452 | 328,764 | 378,543 | 305,380 | 240,764 | 157,704 | 116,321 | 107,381 | 124,645 | 143,719 | 171,918 | 208,132 |
| GEORGIA | 98,216 | 98,961 | 104,406 | 91,977 | 83,521 | 66,419 | 70,317 | 49,267 | 45,863 | 56,943 | 67,364 | 73,801 |
| ILLINOIS | 501,630 | 559,019 | 575,152 | 561,458 | 429,694 | 362,438 | 323,338 | 264,989 | 281,725 | 306,886 | 337,145 | 392,263 |
| IOWA | 157,556 | 166,659 | 160,059 | 162,720 | 156,353 | 142,264 | 151,172 | 149,287 | 168,237 | 167,220 | 178,977 | 189,588 |
| KENTUCKY | 87,401 | 98,102 | 106,698 | 105,963 | 100,522 | 93,602 | 90,042 | 77,144 | 75,571 | 77,941 | 87,778 | 95,576 |
| LOUISIANA | 40,733 | 51,029 | 44,648 | 48,899 | 64,177 | 58,996 | 33,700 | 46,521 | 34,521 | 29,926 | 18,702 | 35,784 |
| MARYLAND | 324,800 | 399,372 | 449,247 | 453,676 | 381,886 | 338,240 | 312,136 | 296,677 | 304,240 | 323,168 | 339,880 | 359,340 |
| MASSACHUSETTS | 493,461 | 530,764 | 514,477 | 480,125 | 407,303 | 391,575 | 389,028 | 375,857 | 359,561 | 376,599 | 353,205 | 332,201 |
| MICHIGAN | 192,004 | 202,484 | 202,104 | 182,278 | 147,803 | 121,226 | 124,158 | 108,672 | 98,819 | 117,499 | 127,528 | 137,790 |
| MINNESOTA | 384,061 | 416,438 | 441,364 | 404,393 | 308,025 | 283,911 | 278,525 | 283,336 | 293,370 | 300,406 | 330,845 | 349,604 |
| MISSOURI | 131,421 | 131,613 | 143,270 | 133,478 | 111,274 | 125,441 | 124,977 | 106,219 | 113,631 | 108,460 | 106,567 | 94,753 |
| NEBRASKA | 161,575 | 167,933 | 185,819 | 197,032 | 180,384 | 175,386 | 149,104 | 143,419 | 151,437 | 179,690 | 176,638 | 201,009 |
| NEVADA | 560,222 | 560,165 | 561,788 | 516,079 | 431,006 | 282,536 | 295,921 | 311,878 | 262,091 | 298,337 | 293,738 | 418,029 |
| NEW HAMPSHIRE | 197,824 | 216,644 | 216,243 | 209,022 | 175,109 | 160,246 | 157,630 | 154,139 | 164,009 | 163,673 | 171,267 | 190,057 |
| NEW JERSEY | 494,618 | 568,862 | 570,781 | 544,335 | 493,339 | 474,865 | 442,251 | 374,403 | 371,968 | 385,923 | 393,278 | 422,408 |
| NEW YORK | 331,898 | 373,031 | 384,722 | 378,042 | 348,119 | 314,085 | 310,656 | 290,361 | 288,825 | 296,897 | 291,753 | 305,767 |
| NORTH CAROLINA | 53,121 | 72,056 | 80,588 | 79,762 | 78,677 | 75,103 | 71,943 | 68,234 | 63,230 | 70,373 | 71,758 | 78,799 |
| OHIO | 126,818 | 126,519 | 121,122 | 110,316 | 99,685 | 110,040 | 89,367 | 84,230 | 81,392 | 85,838 | 94,103 | 98,921 |
| OKLAHOMA | 42,123 | 53,111 | 58,969 | 55,334 | 54,878 | 48,432 | 56,169 | 45,711 | 38,625 | 49,553 | 42,879 | 49,872 |
| OREGON | 379,587 | 442,266 | 498,927 | 530,785 | 498,898 | 441,735 | 382,248 | 344,770 | 367,993 | 418,511 | 461,563 | 527,609 |
| PENNSYLVANIA | 203,225 | 207,614 | 209,254 | 213,232 | 214,275 | 209,644 | 197,310 | 178,461 | 170,309 | 162,025 | 163,425 | 173,156 |
| RHODE ISLAND | 453,404 | 485,528 | 477,605 | 475,969 | 391,739 | 326,538 | 324,552 | 291,998 | 302,587 | 330,743 | 327,780 | 357,432 |
| SOUTH CAROLINA | 73,400 | 73,263 | 74,051 | 84,609 | 73,873 | 74,460 | 67,809 | 50,294 | 56,110 | 69,231 | 81,139 | 93,909 |
| SOUTH DAKOTA | * | 205,135 | 286,967 | 393,343 | 404,694 | 353,815 | 246,685 | 274,303 | 196,267 | 227,255 | 215,336 | * |
| TENNESSEE | 41,408 | 41,294 | 47,432 | 43,586 | 42,191 | 40,786 | 37,456 | 38,182 | 37,500 | 34,736 | 39,103 | 47,093 |
| VIRGINIA | 232,030 | 284,958 | 310,241 | 289,286 | 210,871 | 193,826 | 195,578 | 186,641 | 217,645 | 216,774 | 229,766 | 238,697 |
| WASHINGTON | 368,814 | 406,387 | 484,769 | 499,625 | 461,810 | 418,313 | 407,968 | 345,276 | 340,981 | 400,985 | 425,391 | 474,704 |
| WEST VIRGINIA | 76,549 | 86,536 | 93,909 | 89,838 | 93,018 | 90,892 | 85,102 | 76,582 | 68,373 | 81,934 | 81,956 | 113,045 |
| WISCONSIN | 319,428 | 349,045 | 367,688 | 350,328 | 330,541 | 319,558 | 302,978 | 270,193 | 268,171 | 281,185 | 313,683 | 319,662 |

Table A10: SFR Composite Leverage

| State | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ALABAMA | * | * | 0.39 | 0.34 | 0.22 | 0.36 | 0.49 | 0.55 | 0.54 | 0.24 | 0.29 | 0.23 |
| ARIZONA | 0.30 | 0.33 | 0.36 | 0.34 | 0.37 | 0.33 | 0.34 | 0.31 | 0.29 | 0.30 | 0.32 | 0.36 |
| ARKANSAS | 0.24 | 0.24 | 0.27 | 0.26 | 0.28 | 0.24 | 0.22 | 0.25 | 0.22 | 0.17 | 0.23 | 0.25 |
| CALIFORNIA | 0.54 | 0.56 | 0.56 | 0.55 | 0.51 | 0.49 | 0.45 | 0.45 | 0.45 | 0.48 | 0.49 | 0.53 |
| COLORADO | 0.51 | 0.50 | 0.46 | 0.43 | 0.44 | 0.44 | 0.45 | 0.40 | 0.46 | 0.48 | 0.50 | 0.52 |
| CONNECTICUT | 0.70 | 0.74 | 0.72 | 0.72 | 0.66 | 0.67 | 0.66 | 0.66 | 0.66 | 0.65 | 0.61 | 0.60 |
| DELAWARE | 0.58 | 0.64 | 0.68 | 0.65 | 0.68 | 0.66 | 0.60 | 0.57 | 0.56 | 0.53 | 0.51 | 0.57 |
| FLORIDA | 0.38 | 0.40 | 0.41 | 0.35 | 0.35 | 0.30 | 0.23 | 0.23 | 0.24 | 0.25 | 0.28 | 0.31 |
| GEORGIA | 0.41 | 0.38 | 0.37 | 0.32 | 0.32 | 0.29 | 0.32 | 0.25 | 0.23 | 0.26 | 0.28 | 0.29 |
| ILLINOIS | 0.55 | 0.56 | 0.55 | 0.53 | 0.46 | 0.46 | 0.43 | 0.38 | 0.41 | 0.41 | 0.42 | 0.48 |
| IOWA | 0.50 | 0.50 | 0.46 | 0.46 | 0.44 | 0.41 | 0.42 | 0.42 | 0.46 | 0.43 | 0.45 | 0.44 |
| KENTUCKY | 0.36 | 0.40 | 0.39 | 0.39 | 0.38 | 0.37 | 0.35 | 0.30 | 0.29 | 0.29 | 0.31 | 0.33 |
| LOUISIANA | 0.01 | 0.39 | 0.33 | 0.34 | 0.42 | 0.39 | 0.23 | 0.32 | 0.24 | 0.20 | 0.12 | 0.21 |
| MARYLAND | 0.51 | 0.52 | 0.53 | 0.53 | 0.51 | 0.51 | 0.48 | 0.48 | 0.49 | 0.49 | 0.51 | 0.53 |
| MASSACHUSETTS | 0.71 | 0.71 | 0.70 | 0.68 | 0.62 | 0.64 | 0.64 | 0.63 | 0.68 | 0.74 | 0.69 | 0.67 |
| MICHIGAN | 0.47 | 0.48 | 0.49 | 0.49 | 0.48 | 0.46 | 0.46 | 0.42 | 0.38 | 0.38 | 0.38 | 0.39 |
| MINNESOTA | 0.66 | 0.66 | 0.70 | 0.66 | 0.57 | 0.58 | 0.57 | 0.62 | 0.61 | 0.57 | 0.60 | 0.60 |
| MISSOURI | 0.37 | 0.37 | 0.40 | 0.38 | 0.34 | 0.40 | 0.40 | 0.36 | 0.37 | 0.36 | 0.34 | 0.37 |
| NEBRASKA | 0.42 | 0.42 | 0.50 | 0.52 | 0.48 | 0.43 | 0.37 | 0.36 | 0.36 | 0.41 | 0.38 | 0.42 |
| NEVADA | 0.41 | 0.37 | 0.35 | 0.35 | 0.40 | 0.34 | 0.36 | 0.42 | 0.35 | 0.33 | 0.29 | 0.38 |
| NEW HAMPSHIRE | 0.67 | 0.67 | 0.67 | 0.67 | 0.62 | 0.62 | 0.62 | 0.63 | 0.65 | 0.62 | 0.62 | 0.65 |
| NEW JERSEY | 0.50 | 0.50 | 0.49 | 0.47 | 0.46 | 0.49 | 0.47 | 0.42 | 0.43 | 0.43 | 0.42 | 0.44 |
| NEW YORK | 0.58 | 0.58 | 0.57 | 0.56 | 0.55 | 0.53 | 0.53 | 0.50 | 0.50 | 0.50 | 0.47 | 0.48 |
| NORTH CAROLINA | 0.21 | 0.27 | 0.27 | 0.25 | 0.26 | 0.26 | 0.26 | 0.26 | 0.24 | 0.25 | 0.25 | 0.26 |
| OHIO | 0.45 | 0.43 | 0.42 | 0.40 | 0.40 | 0.47 | 0.38 | 0.37 | 0.35 | 0.36 | 0.36 | 0.36 |
| OKLAHOMA | 0.18 | 0.22 | 0.23 | 0.20 | 0.20 | 0.17 | 0.20 | 0.17 | 0.14 | 0.17 | 0.14 | 0.16 |
| OREGON | 0.61 | 0.60 | 0.58 | 0.59 | 0.60 | 0.61 | 0.55 | 0.54 | 0.56 | 0.57 | 0.58 | 0.61 |
| PENNSYLVANIA | 0.54 | 0.50 | 0.48 | 0.49 | 0.51 | 0.52 | 0.51 | 0.47 | 0.45 | 0.42 | 0.42 | 0.43 |
| RHODE ISLAND | 0.65 | 0.63 | 0.63 | 0.65 | 0.60 | 0.58 | 0.58 | 0.57 | 0.56 | 0.59 | 0.55 | 0.58 |
| SOUTH CAROLINA | 0.34 | 0.31 | 0.28 | 0.29 | 0.27 | 0.30 | 0.29 | 0.22 | 0.25 | 0.28 | 0.31 | 0.33 |
| SOUTH DAKOTA | * | 0.68 | 0.95 | 1.13 | 0.20 | 0.81 | 0.69 | 0.68 | 0.49 | 0.52 | 0.48 | * |
| TENNESSEE | 0.26 | 0.23 | 0.25 | 0.21 | 0.22 | 0.22 | 0.21 | 0.22 | 0.21 | 0.19 | 0.20 | 0.22 |
| VIRGINIA | 0.50 | 0.51 | 0.51 | 0.50 | 0.42 | 0.41 | 0.42 | 0.41 | 0.46 | 0.43 | 0.44 | 0.45 |
| WASHINGTON | 0.53 | 0.51 | 0.52 | 0.51 | 0.52 | 0.53 | 0.54 | 0.49 | 0.47 | 0.51 | 0.50 | 0.50 |
| WEST VIRGINIA | 0.46 | 0.49 | 0.47 | 0.43 | 0.45 | 0.46 | 0.42 | 0.39 | 0.33 | 0.39 | 0.38 | 0.47 |
| WISCONSIN | 0.56 | 0.58 | 0.59 | 0.56 | 0.56 | 0.58 | 0.55 | 0.52 | 0.52 | 0.52 | 0.56 | 0.55 |