J Real Estate Finan Econ (2020) 60:40–52
https://doi.org/10.1007/s11146-019-09712-0

Appraisal Accuracy and Automated Valuation Models in Rural Areas

Alexander N. Bogin[1] · Jessica Shui[1]

Published online: 8 August 2019

Abstract
Accurate and unbiased property value estimates are essential to credit risk management. Along with loan amount, they determine a mortgage's loan-to-value ratio, which captures the degree of homeowner equity and is a key determinant of borrower credit risk. For home purchases, lenders generally require an independent appraisal, which, in addition to a home's sales price, is used to calculate a value for the underlying collateral. A number of empirical studies have shown that property appraisals tend to be biased upwards, and over 90 percent of the time, either confirm or exceed the associated contract price. Our data suggest that appraisal bias is particularly pervasive in rural areas where over 25 percent of rural properties are appraised at more than five percent above contract price. Given this significant upward bias, we examine a host of alternate valuation techniques to more accurately estimate rural property values.

Keywords Automated valuation models · Appraisal · Property value · Rural

JEL Classification G21 · L85 · R3

Introduction

Accurate property value estimates are an essential component of the mortgage underwriting process. Along with the loan amount, they determine a mortgage's loan-to-value (LTV) ratio, which captures the degree of homeowner equity and the credit risk of a loan. For home purchases, lenders generally require an independent appraisal, which,

✉ Jessica Shui
jessica.shui@fhfa.gov

Alexander N. Bogin
alexander.bogin@fhfa.gov

[1] Federal Housing Finance Agency, Office of Policy Analysis & Research, 400 7th Street SW, Washington, DC 20219, USA

in addition to a home's sales price, is used to determine a value for the underlying collateral. A number of empirical studies have shown that property appraisals tend to be biased upwards, and over 90% of the time, either confirm or exceed the associated contract price.[1] This upward appraisal bias is often particularly pronounced in rural areas where there are fewer comparable sales and more heterogeneity across homes. In fact, our data suggest that more than 25% of rural appraisals exceed the associated contract price by more than 5%. Given the extent and ubiquity of appraisal bias in rural areas, we create a series of alternate automated property value estimates, using a number of machine learning algorithms, to more accurately value the collateral underlying rural purchase-money mortgages.

Appraisals are performed by experts with specialized knowledge about local housing markets, but they can face pressures—either apparent or perceived—to arrive at a value estimate at or above the contract price to ensure that a sale goes through. Since the Great Recession, the majority of single-family conforming loans have been sold to or securitized by the government sponsored enterprises (GSEs). For purchase money loans, the GSEs require that a borrower's LTV ratio be calculated as the loan amount over the lesser of the contract price and an independent appraisal estimate. If the appraisal estimate is less than the contract price, a borrower may need to increase his or her down payment to stay at their desired LTV range and not incur a higher interest rate on the loan. Alternatively, the contract price could be renegotiated. In either case, additional difficulties may arise for the borrower or seller and thus appraised values that are "too low" may increase the chance that the sale could fall through.[2]

Property appraisals are most often estimated based upon the recent sales prices of three to five comparable properties.[3] This leads to a certain degree of subjectivity. Appraisers have a number of different options in terms of what comparable sales they select as reference transactions. Because properties can resemble each other along many different dimensions, the appraiser has some latitude in terms of what comparable properties he or she chooses. Further, when a comparable property has characteristics that differ from the subject property (e.g., condition, square footage, number of bedrooms), appraisers will manually make adjustments to a comparable sales price to reflect these differences. Finally, appraisers can assign different weights to each comparable based upon the degree of applicability to the subject property. In each of these stages, there is opportunity to push up appraised values so that they either confirm or exceed the contract price. Eriksen et al. (2016) find evidence of significant upward bias at each of these three stages of the appraisal process.

Lang and Nakamura (1993), Blackburn and Vermilyea (2007), and Ding (2014) find that appraisal bias is amplified in rural areas, which are often characterized by fewer comparable sales and more heterogeneity across homes. We find a similar result using appraisal data for Fannie Mae and Freddie Mac acquisitions from 2012 through 2016.

[1] Cho and Megbolugbe (1996), Horne and Rosenblatt (1996), and Calem et al. (2017) find that between 90 to 95% of appraisals come in at or above contract price. See Yiu et al. (2006) for a detailed review of the literature. Consistent with existing literature, we define appraisal bias as the percentage deviation of the appraised value from the contract price. For an individual appraisal, it is possible that the appraised value exceeding the contract price is in fact an accurate estimate of the real house value. However, it is highly unlikely to observe such a systematically skewed relationship without at least some level of bias.

[2] Fout and Yao (2016) find that about 32% of negative appraisals result in the transaction falling through.

[3] Dotzour (1990) finds that the sales comparison approach tends to provide a more accurate measure of value than the cost approach.

As illustrated in Fig. 1a and 1b, both urban and rural areas are subject to appraisal bias, but such bias is exacerbated in rural areas. Specifically, we find that approximately 25% of properties in rural areas compared to 12.7% in urban areas are appraised at more than 5% above contract price.

Given observed appraisal bias, several researchers have considered including automated property value estimates as an alternate, and potentially unbiased, measure of the underlying value of the collateral when calculating credit risk (LaCour-Little and Malpezzi 2003; Kelly 2007; Agarwal et al. 2015; and Calem et al. 2017). We explore a similar question, but concentrate our attention on rural areas.[4] Specifically, we estimate a series of AVMs[5] and examine their computational burden and out-of-sample predictive accuracy in rural areas. In an effort to explore a wide array of specifications, we include two tree-based approaches to help capture non-linear relationships.[6]

The paper is structured as follows. In section two, we discuss the data used to estimate a series of AVMs. In section three, we evaluate several different AVM models and select a preferred specification based upon both computation time and out-of-sample predictive ability. We conclude in section four.

Data

Our analysis draws on the Uniform Appraisal Dataset (UAD) from Q4 2012 to Q1 2016. It contains every active appraisal record associated with loan applications submitted to Fannie Mae and Freddie Mac during the study period. The UAD provides us with information on structural characteristics, neighborhood attributes, and historical sales prices for rural homes.

We define rural in the same manner as the Appraisal Institute.[7] Specifically, it is defined as "Pertaining to the country as opposed to urban or suburban; land under an agricultural use; areas that exhibit relatively slow growth with less than 25% development" (The Dictionary of Real Estate Appraisal – 4th Edition). Appraisers use this definition to determine the neighborhood characteristics of a subject property when filling out a Uniform Residential Appraisal Report.

The Enterprises agreed to adopt the Home Valuation Code of Conduct (HVCC), which became effective in May 2009. This strengthened requirements for submitting each appraisal associated with a loan application to the Enterprises. In the Uniform Residential Appraisal Report, appraisers are required to document property attributes for subject and comparable properties as well as appraisal approaches adopted. Starting in 2012, the data captured by lenders' appraisal reports conform to the UAD, and are digitized, compiled, and submitted to the Enterprises to support loans lenders wish to

[4] Compared to AVM estimates in urban areas, AVM estimates in rural areas may face additional scrutiny due to lack of data.

[5] We do not seek to use contemporaneous data and make real-time AVM predictions in this research. Instead, our work is focused on highlighting the pros and cons of machine learning algorithms and comparing their performance with more traditional methodologies.

[6] Many other researchers, for example Pace and Hayunga (2018) and Villupuram and Johnson (2018), have also explored machine learning algorithms in property valuations.

[7] This definition is consistent with the Enterprises' guidance.

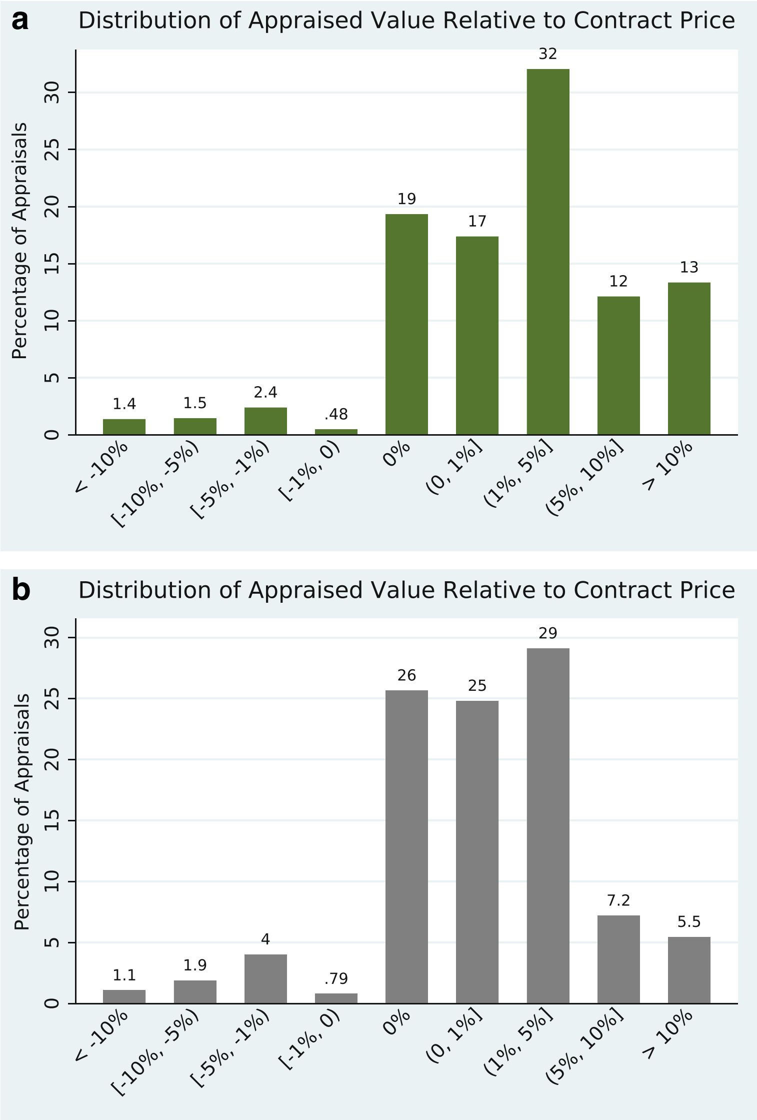

Fig. 1 **a** Distribution of Appraised Value Relative to Contract Price for Rural Areas. **b** Distribution of Appraised Value Relative to Contract Price for Urban Areas

sell to the Enterprises. The data consists of about 18 million unique appraisal records for subject properties and 2.1 million for purchase-money mortgages.[8] Of the 2.1 million appraisals for purchase-money mortgages, 420,370[9] are associated with rural properties, and hence form the full sample in our analysis.

[8] The same appraisal is often submitted to both Fannie Mae and Freddie Mac.

[9] We apply a series of standard data filters, which include censoring observations associated with extreme/implausible values for several structural attributes (i.e., number bedrooms, number of bathrooms, square footage, and age). To further minimize the influence of outliers, we remove observations with sales prices in the top and bottom 1% of the price distribution.

For purposes of model development, We randomly split our full sample into two subsamples – a training dataset and a test dataset. The training dataset contains 80% of our observations and is used to estimate or parameterize each of our models. The test dataset contains the remaining 20% of observations and is used to measure each model's out-of-sample performance.

Table 1 provides summary statistics for the full sample, and the training and test samples. As detailed in Panel A, the average sales price for rural buyers in the full sample is approximately $220,000.[10] The average rural property has approximately 3 bedrooms, 2 bathrooms, and 1,870 square feet of living area. Although rural properties may have different values for their land and amenities, their basic structural attributes are qualitatively similar to those of the full purchase-money mortgage sample in the UAD. The average rural property is approximately 32 years old[11] with a 3.57 (out of 5) appraiser-rated overall condition and a 2.99 (out of 5) quality of construction. Among the purchase-money rural mortgages, properties in the following states have the largest representation: MI (7.35%), TX (6.94%), OH (4.93%), GA (4.87%), and PA (4.77%). As shown in Panels B and C, the average properties in the training and the test samples are very similar to the representative property in the full sample.

AVM Techniques and their out-of-Sample Performance

Using the UAD, we begin our analysis by estimating and exploring the out-of-sample performance of a series of AVMs. There are a number of different approaches to estimating home values. We focus our attention on a subset of techniques that allow us to more fully capture household heterogeneity by incorporating information on both structural (e.g., number of bedrooms, number of bathrooms, square footage) and neighborhood attributes (e.g., location, proximity to amenities, view). We begin our model[12] selection process with a hedonic regression estimated using ordinary least squares (OLS). This is one of the most commonly used techniques for price estimation and will serve as a baseline as we evaluate five alternate and more involved estimation techniques.

As mentioned in Section 2, we use the training dataset to train each of our models and estimate the parameters and use the test dataset to measure each model's out-of-sample performance. For each model, we focus our attention on two performance metrics – the R^2 and the root mean squared error (RMSE). The R^2 statistic is a relative measure of fit for calculating the proportion of variance in the dependent variable, which is captured through the model. The RMSE is an absolute measure of fit, which calculates the sample standard deviation of the residuals and provides an overall measure of model accuracy.

Table 2 provides model fit statistics for each AVM technique. As detailed, a standard hedonic results in an out-of-sample R^2 value of 0.6803 and an RMSE of 0.3188. The R^2 value indicates that a standard hedonic is able to explain approximately 68.03% of the variation in log sales price for single-family homes in rural areas. These metrics serve as a baseline as we explore a series of alternate estimators. While an overall RMSE and

[10] This is about $55,000 lower than the average sales price across all purchase-money mortgages in the UAD.

[11] This is about twice as old as the average property in the full purchase-money mortgage sample in the UAD.

[12] In this paper, we use model and algorithm interchangeably to refer to each specific AVM technique.

Table 1 Summary Statistics for Rural Purchase-Money Mortgages

Variables	Mean	SD	P25	P75
Panel A: Full Sample				
Sales Price ($)	219,555	129,047	130,000	275,000
Number of Bathrooms	1.9027	0.6532	1.1	2.1
Number of Bedrooms	3.0330	0.7721	3	3
Square Footage	1,869.55	712.1829	1,360	2,239
Age of the House (Years)	32.2496	30.2857	11	44
Overall Condition (1-5)	3.5662	0.6140	3	4
Quality of Construction (1-5)	2.9921	0.8461	3	4
Number of Observations	420,370			
Panel B: Training Sample (80%)				
Sales Price ($)	219,428	128,909	130,000	275,000
Number of Bathrooms	1.9018	0.6537	1.1	2.1
Number of Bedrooms	3.0326	0.7724	3	3
Square Footage	1,869.165	712.2785	1,360	2,239
Age of the House (Years)	32.2350	30.2350	11	44
Overall Condition (1-5)	3.5665	0.6143	3	4
Quality of Construction (1-5)	2.9933	0.8450	3	4
Number of Observations	336,216			
Panel C: Test Sample (20%)				
Sales Price ($)	220,066	129,593	130,900	276,000
Number of Bathrooms	1.9062	0.6513	1.1	2.1
Number of Bedrooms	3.0346	0.7706	3	3
Square Footage	1,871.09	711.8031	1,363	2,240
Age of the House (Years)	32.3078	30.4878	11	44
Overall Condition (1-5)	3.5649	0.6126	3	4
Quality of Construction (1-5)	2.9873	0.8505	3	4
Number of Observations	84,154			

This table reports the summary statistics of the attributes and conditions of the house and the transaction price. Each observation corresponds to a single appraisal record in the sample. The top five states in our sample are Michigan (7.35%), Texas (6.94%), Ohio (4.93%), Georgia (4.87%), and Pennsylvania (4.77%).

R^2 are useful in describing average model fit, they fail to provide sufficient information on the success (or lack thereof) of model fit in the tails of the price distribution. For instance, a particular model may perform well when estimating the value of an average priced home, but fail to explain sufficient variation for lower or higher priced units. To explore how our model fits the tails of the price distribution, we calculate separate R^2 values for the top and bottom quartiles of the sales price distribution.[13]

[13] As detailed, R^2 are actually higher in the tails of the price distribution. While this result may seem counterintuitive, it is simply a reflection of the proportional nature of the statistic. Both the residual sum of squares and total sum of squares increase as we move away from the middle of the price distribution, but the total sum of squares increases at a faster rate. In other words, absolute fit (as captured by RMSE) is deteriorating, but proportional fit (or the percentage of explained variation) is actually increasing.

Table 2 AVM Out-of-Sample Performance Metrics: Model Fit Statistics

Specifications		R^2			RMSE		
		Baseline	P25	P75	Baseline	P25	P75
Tree-Based Estimators	Random Forest	0.7224	0.7700	0.8193	0.2971	0.3749	0.3215
	Boosting	0.6755	0.7273	0.7772	0.3212	0.4082	0.3569
Standard Hedonic	OLS	0.6803	0.7452	0.7729	0.3188	0.3946	0.3603
Shrinkage Estimators *lambda.1se*	Elastic Net	0.6788	0.7404	0.7699	0.3196	0.3983	0.3627
	Lasso	0.6795	0.7422	0.7713	0.3192	0.3969	0.3616
	Ridge	0.6761	0.7335	0.7654	0.3209	0.4036	0.3662
Shrinkage Estimators *lambda.min*	Elastic Net	0.6803	0.7447	0.7727	0.3188	0.3950	0.3605
	Lasso	0.6803	0.7448	0.7727	0.3188	0.3949	0.3605
	Ridge	0.6765	0.7344	0.7660	0.3207	0.4029	0.3658

This table reports the out-of-sample performance (measured by R^2 and RMSE) for different model specifications employing all records (All) in the test dataset and the subsamples associated with the sales price in the top (P75) and the bottom quartile (P25) of the sales price distribution. In terms of shrinkage estimators, for model simplicity, we choose by default *lambda.1se*, the largest value of lambda such that the cross-validated error is within one standard deviation of the minimum mean cross-validated error. Alternatively, if *lambda.min*—the lambda that minimize the mean cross-validated error—is chosen, R^2 and RMSE for shrinkage estimators improve marginally (for example, from 0.6688 to 0.6700 for Elastic Net) and the model fit for Lasso and Elastic Net become equally good as for OLS and Boosting.

Due to our relatively small sample size, we also explore a series of shrinkage estimators, which introduce a degree of bias in exchange for a lower variance. Specifically, shrinkage estimators incorporate a penalty function in the estimation process which pushes (or shrinks) coefficients values towards zero. This bias-variance tradeoff has been shown to lead to smaller mean squared errors when applied to out-of-sample data.

Two of the more popular shrinkage estimators are ridge regression and lasso (least absolute shrinkage and selection operator) regression. Ridge regression includes a penalty function, which biases the value of model coefficients towards zero. Lasso regression goes one step further and aides with variable selection by eliminating certain covariates altogether (this is achieved by shrinking the associated coefficients all the way to zero). Consistent with other literature and for model simplicity, we choose *lambda.1se* as the benchmark for all shrinkage estimation models. L*ambda.1se* is the largest value of lambda such that the cross-validated error is within one standard deviation of the minimum mean cross-validated error, whereas *lambda.min* is the lambda that minimizes the mean cross-validated error. With the latter, R^2 and RMSE for shrinkage estimators improve marginally at the cost of making models more complicated. As detailed in the bottom row in Table 2, in our benchmark specification for shrinkage estimators, the ridge regression performs marginally worse than our baseline estimator with an out-of-sample R^2 value of 0.6761 and an RMSE of 0.3209. The additional flexibility garnered through variable selection improves model fit statistics for the lasso regression, but increases in accuracy are relative minor. The lasso regression is associated with an out-of-sample R^2 value of 0.6795 and an RMSE of 0.3192 (the fifth bottom row in Table 2).[14] While these fit statistics are marginally better than our baseline hedonic model, the added accuracy may not be worth the increase in model complexity.

Next, we test elastic net, which is a hybrid estimator that combines features of both the ridge and lasso regressions. Specifically, elastic net combines the ridge and lasso penalties into a single function and has been shown to outperform both precursor models on data with highly correlated predictors. However, as detailed in Table 2, we find the performance of elastic net lies between those of lasso and ridge, evidenced by both R^2 and RMSE.

While our linear estimators have yet to significantly improve upon a standard hedonic, machine learning tree-based models may provide more "value-added" by mapping non-linear relationships. The first tree-based model we test is random forest. Random forest involves building a number of decision trees on a series of bootstrapped training samples. In an effort to decorrelate the individual trees, whenever a split is considered, potential candidate variables are drawn from a random sample of m predictors where $m<p$ and p is the full set of predictors.[15] This ultimately results in a composite estimate characterized by low variance because it is based upon the average of many uncorrelated decision trees. As detailed in Table 2, random forest significantly improves upon our baseline estimator with an out-of-sample R^2 value of 0.7224 and an RMSE of 0.2971.

Though the random forest estimator increases explanatory power by approximately 6.1% relative to our baseline model, this improved model fit comes at the cost of significant computational time. Required CPU time is over 48 hours versus the minutes it takes to re-estimate a standard hedonic regression. The processing time is calculated including both the time for training and for out-of-sample estimation with the following specifics: 1) as shown in Table 1, our training dataset contains 336,216 observations and our test dataset contains 84,154 observations; 2) each model contains 115 explanatory variables except that for random forest we further break down the data by state (to increase efficiency) and estimate for each state the same specification netting out the 49 state FEs; 3) processing time for all algorithms are based on the same dual-core CPU @ 2.6GHz with 8GB memory and 1600 Max RAM speed. For most algorithms, training takes up over 90% of the time. To the extent that one has a more powerful CPU, the processing time will reduce accordingly.

Another drawback of random forest is that it may suffer from significant overfitting. We show the baseline out-of-sample and in-sample fitting in Table 3. In-sample R^2 is calculated using the training dataset. Without surprise, random forest suffers greatly from overfitting while other algorithms do not.

We next explore a second tree-based model called boosting. Boosting involves sequentially growing a set of decision trees. The first tree is fitted to the outcome variable, which results in a set of residuals. The second tree is fitted to these first-stage residuals and then added back into the initially fitted function, $\hat{f}$. This new function produces an updated set of residuals and the process repeats. With each iteration, the

[14] If using *lambda.min*, then the out-of-sample R^2 value is 0.6803 and the RMSE is 0.3188.

[15] Oftentimes, m is set equal to the square root of p.

Table 3 AVM In-sample vs Out-of-sample performance metrics

Specifications		Baseline			
		R^2		RMSE	
		Out-of-sample	In-sample	Out-of-sample	In-sample
Tree-Based Estimators	Random Forest	0.7224	0.9436	0.2971	0.1337
	Boosting	0.6755	0.6749	0.3212	0.3211
Standard Hedonic	OLS	0.6803	0.6801	0.3188	0.3185
Shrinkage Estimators *lambda.1se* (Benchmark)	Elastic Net	0.6788	0.6786	0.3196	0.3192
	Lasso	0.6795	0.6793	0.3192	0.3189
	Ridge	0.6761	0.6759	0.3209	0.3206
Shrinkage Estimators *lambda.min*	Elastic Net	0.6803	0.6801	0.3188	0.3185
	Lasso	0.6803	0.6801	0.3188	0.3185
	Ridge	0.6765	0.6763	0.3207	0.3204

This table reports the in-sample performance (measured by R^2 and RMSE) in comparison to the out-of-sample performance. In-sample and out-of-sample statistics are calculated employing the training and the test dataset respectively. Random Forest, though outperforming other, suffers from overfitting.

boosting approach slowly improves $\hat{f}$ in high variance areas. A shrinkage parameter, λ, controls the rate at which the boosting approach learns, while the degree of model complexity is controlled by d, which determines the number of splits in each tree. Interestingly, the boosting approach does not improve upon the random forest results, evidenced by a lower R^2. The boosting model results in an out-of-sample R^2 value of 0.6755 and an RMSE of 0.3212 (see Table 2). These fit statistics are virtually the same as our standard hedonic, which suggests that boosting may not be worth its added computational burden (approximately 24 hours of CPU time), at least when it comes to home price valuation in rural areas.

After obtaining a baseline OLS fit and examining the initial performance of several alternative algorithms, we explore additional ways of improving the accuracy of our estimates. One of the most straightforward ways to improve accuracy is to increase the sample size of the training dataset. However, adding more rural data does not seem to be a viable option here due to data scarcity. Therefore, as an alternative, we add a sample of urban data to the existing rural training sample with a one-to-one or one-to-two ratio of number of observations. Table 4 shows the results. In addition to the baseline model specification, the models employing two mixed samples always include an additional rural dummy, which has a significant negative impact on the sales price.

For random forest, we expect the additional data to help overcome our overfitting problem since a larger sample will lead to smaller distinctions between individual trees within the forest. Not surprisingly, results show that random forest outperforms other algorithms with an even bigger gap. In other words, though the urban data introduces some bias, random forest does not seem to be sensitive to such bias and its out-of-sample fit improves significantly, as detailed in Table 4 comparing the baseline R^2 to the one in the adjacent column. At some point

Table 4 AVM Out-of-Sample Performance Metrics: Baseline vs Mixed samples

Specifications		R^2			RMSE		
		Baseline	1:0.5	1:1	Baseline	1:0.5	1:1
Tree-Based Estimators	Random Forest	0.7224	0.8992	0.8971	0.2971	0.1790	0.1808
	Boosting	0.6755	0.6519	0.6436	0.3212	0.3327	0.3366
Standard Hedonic	OLS	0.6803	0.6677	0.6539	0.3188	0.3250	0.3317
Standard Hedonic	OLS w/ interactions	0.6803	0.6762	0.6717	0.3188	0.3208	0.3231
Shrinkage Estimators *lambda.1se* (Benchmark)	Elastic Net	0.6788	0.6667	0.6531	0.3196	0.3255	0.3321
	Lasso	0.6795	0.6661	0.6525	0.3192	0.3258	0.3324
	Ridge	0.6761	0.6613	0.6481	0.3209	0.3281	0.3345
Shrinkage Estimators *lambda.min*	Elastic Net	0.6803	0.6676	0.6539	0.3188	0.3251	0.3317
	Lasso	0.6803	0.6676	0.6539	0.3188	0.3251	0.3317
	Ridge	0.6765	0.6624	0.6486	0.3207	0.3276	0.3342

This table reports the out-of-sample performance (measured by R^2 and RMSE) for different model specifications employing all records (All) in the test dataset. The ratios (1:0.5 and 1:1) illustrate how the sample is constructed. For example, ratio 1:0.5 means that for two rural records in the training dataset we mix one urban record in the sample. In addition to baseline model specification, the models employing two mixed samples always include an additional rural dummy, which has a significant negative impact on the sales price. Both R^2 and RMSE are calculated only for the records in the baseline test dataset so that all columns are easily comparable.

however, the trade-off between bias and variance reaches a point where the cost of additional bias from including more urban sales outweighs the benefit of lower variance due to the increased sample size. Hence, the out-of-sample performance for mixing with a one-to-one ratio is not significantly better than with a two-to-one ratio.

While this effort helps with random forest, it does not help with any other algorithms we explore in this paper. One possible explanation is that our estimations from other algorithms are already quite precise, so the marginal cost of additional bias outweighs the marginal benefit of increased sample size when we mix rural and urban samples with a two-to-one ratio. However, some may argue that it is not a fair comparison since random forest automatically considers the interactions between the urban variable and other existing variables while others algorithms do not once the urban data is added. Therefore, we add two-way interaction terms of the urban dummy with every other existing variable in the OLS regression. Though this is not a full approximation of random forest since it does not consider multi-way interactions, it still gives us an idea of to what degree this effort improves the OLS performance. Comparing the fourth and the third rows in Table 4, we conclude that though it helps a little to include the additional interaction terms, this effort does not get us anywhere close to the performance of random forest.

A caveat that is worth mentioning lies in the uniqueness and the richness of our appraisal data. Consider the public records data where many property attributes are not available, results derived employing the appraisal data, where such attributes are nicely populated, may not be easily generalized. To proxy the public

Table 5 AVM Out-of-Sample Performance Metrics: Generalization

Specifications		R^2			RMSE		
		Baseline	Urban	Fewer Variables	Baseline	Urban	Fewer Variables
Tree-Based Estimators	Random Forest	0.7224	0.8921	0.4865	0.2971	0.1887	0.4041
	Boosting	0.6755	0.6373	0.4699	0.3212	0.3459	0.4106
Standard Hedonic	OLS	0.6803	0.6381	0.4714	0.3188	0.3456	0.4100
Shrinkage Estimators *lambda.1se* (Benchmark)	Elastic Net	0.6788	0.6366	0.4696	0.3196	0.3463	0.4106
	Lasso	0.6795	0.6374	0.4695	0.3192	0.3459	0.4107
	Ridge	0.6761	0.6311	0.4686	0.3209	0.3489	0.4111
Shrinkage Estimators *lambda.min*	Elastic Net	0.6803	0.6380	0.4714	0.3188	0.3456	0.4100
	Lasso	0.6803	0.6380	0.4714	0.3188	0.3456	0.4100
	Ridge	0.6765	0.6326	0.4703	0.3207	0.3482	0.4104

This table reports the out-of-sample performance (measured by R^2 and RMSE) separately employing the rural and the urban sample. Columns "Fewer Variables" present results estimated from models with basic house characteristics only—number of bedrooms and bathrooms, age of the house, number of stories, state and year of the most recent sale. Columns "Urban" contain results estimated using the urban sample with the baseline specification. All R^2 and RMSE are calculated only for the records in the baseline test dataset and the urban test dataset with the same number of observations so that all columns are easily comparable.

records data, we limit our variables to basic house characteristics only—number of bedrooms and bathrooms, age of the house, number of stories, state, and year of the most recent sale. Results are shown in the rightmost columns in the R^2 and the RMSE sections in Table 5. When limiting the number of the explanatory variables, while the performance rank maintains, the gap almost closes between random forest and other algorithms. In general, the richness of the data has the largest beneficial impact on the random forest algorithm. In addition, we employ a different sample—the urban sample—with the same specification as the baseline to test whether our results are robust to the uniqueness of the rural sample (Table 5 Columns "Urban"). We find that the performance rank maintains while the gap between random forest and others prevails, which suggests that our results hold in general. However, given that random forest may perform even better with other samples, it is still encouraged to consider the trade-off between algorithms based on the specific sample employed.

As a summary, Fig. 2 illustrates our baseline result comparing actual versus predicted values for each of the aforementioned models. As illustrated, all but random forest result in similar out-of-sample fits. It is important to note that these results may not extrapolate to urban samples where neighborhoods are much more compact and there are more similarities among nearby properties.

While random forest has the best out-of-sample predictive accuracy, it is computationally burdensome and may not be replicable without access to a powerful server. In addition, it can potentially suffer from significant overfitting. Therefore, we would encourage researchers to consider a standard hedonic estimator in credit modeling amid those concerns.

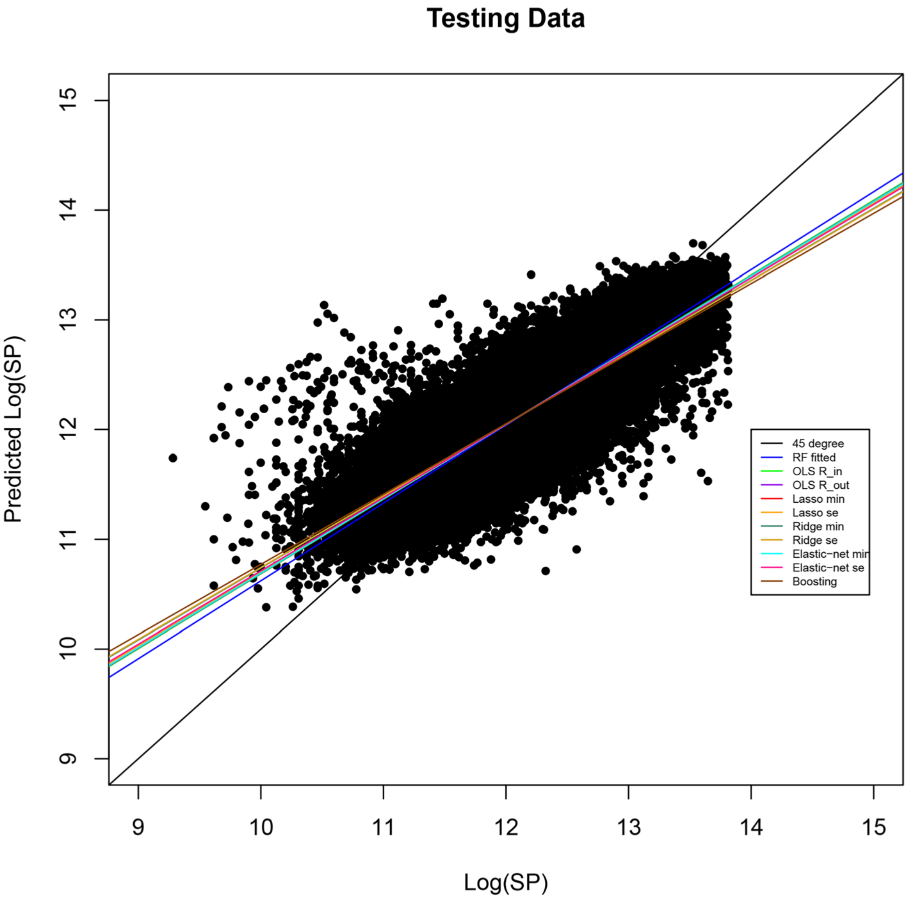

Fig. 2 AVM Out-of-Sample Performance Metrics: Predicted vs. Actual

Conclusion

A number of empirical studies have shown that property appraisals tend to be biased upwards, and may overstate the true value of the underlying collateral.[16] This upward bias is often exacerbated in rural areas where there are fewer comparable sales and more heterogeneity across homes. Based on our data, approximately 25% of rural appraisals exceed the associated contract price by 5% or more. Given the extent of upward bias in rural appraisals, we explore a wide array of AVM techniques in search of an estimator, potentially unbiased, to more accurately value the collateral underlying rural purchase-money mortgages. Our tree-based random forest estimator performs the best in terms of out-of-sample fit, but is also the most computationally burdensome. In the face of computing constraints or lack of a powerful server, we believe that a standard hedonic offers an excellent alternative.

[16] Valuation bias is not unique to the real estate industry. Michaely and Womack (1999), White (2010), and Bolton et al. (2007) have examined similar market pressures and their varied impact on other financial sectors.

Acknowledgements We thank Andy Leventis, Michela Barba, and Will Doerner for their support and to the participants at the American Real Estate Society conference and DC Real Estate Valuation Symposium for their comments that greatly improved this research. We thank Bob Witt and Ming-Yuen Meyer-Fong for sharing their expertise and Sam Frumkin for providing helpful suggestions.

Disclaimer The analysis and conclusions are those of the authors alone and should not be represented or interpreted as conveying an official FHFA analysis, opinion, or endorsement.

References

Agarwal, S., Ben-David, I., & Yao, V. (2015). Collateral valuation and borrower financial constraints: Evidence from the residential real estate market. *Management Science, 61*(9), 2220–2240.

Blackburn, M., & Vermilyea, T. (2007). The role of information externalities and scale economies in home mortgage lending decisions. *Journal of Urban Economics, 61*(1), 71–85.

Bolton, P., Freixas, X., & Shapiro, J. (2007). Conflicts of interest, information provision, and competition in the financial services industry. *Journal of Financial Economics, 85*(2), 297–330.

Calem, P. S., Lambie-Hanson, L., & Nakamura, L. I. (2017). Appraising home purchase appraisals.

Cho, M., & Megbolugbe, I. F. (1996). An empirical analysis of property appraisal and mortgage redlining. *The Journal of Real Estate Finance and Economics, 13*(1), 45–55.

Ding, L. (2014). *The pattern of appraisal Bias in the Third District during the housing crisis*. Philadelphia Federal Reserve: Working Paper.

Dotzour, M. (1990). An empirical analysis of the reliability and precision of the cost approach in residential appraisal. *Journal of Real Estate Research, 5*(1), 67–74.

Eriksen, M. D., Fout, H. B., Palim, M., & Rosenblatt, E. (2016). Contract price confirmation bias: Evidence from repeat appraisals. Working paper.

Fout, H., & Yao, V. (2016). Housing market effects of appraising below contract. Working paper.

Horne, D., & Rosenblatt, E. (1996). Property appraisals and moral hazard. Working paper.

Kelly, A. (2007). *Appraisals, automated valuation models, and mortgage default*. Federal Housing Finance Agency: Working Paper.

LaCour-Little, M., & Malpezzi, S. (2003). Appraisal quality and residential mortgage default: Evidence from Alaska. *The Journal of Real Estate Finance and Economics, 27*(2), 211–233.

Lang, W. W., & Nakamura, L. I. (1993). A model of redlining. *Journal of Urban Economics, 33*(2), 223–234.

Michaely, R., & Womack, K. L. (1999). Conflict of interest and the credibility of underwriter analyst recommendations. *The Review of Financial Studies, 12*(4), 653–686.

Pace, K., & Hayunga D. (2018). Combining random forests with spatiotemporal modeling to improve prediction of real estate prices. Working paper.

Villupuram, S., & Johnson, E. (2018). The value of curb appeal: A machine learning approach. Working paper.

White, L. J. (2010). Credit-rating agencies and the financial crisis: Less regulation of CRAs is a better response. *Journal of international banking law, 25*(4), 170.

Yiu, C. Y., Tang, B. S., Chiang, Y. H., & Choy, L. H. T. (2006). Alternative theories of appraisal Bias. *Journal of Real Estate Literature, 14*(3), 321–344.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.