Revisiting Automated Valuation Model Disparities in Majority-Black Neighborhoods

New Evidence Using Property Condition and Artificial Intelligence

Linna Zhu, Michael Neal, and Caitlin Young

May 2022

Automated valuation models (AVMs) represent the promise of greater efficiency and lower costs for the mortgage industry. But research has suggested that AVMs can produce racially disparate outcomes—namely, higher error as a percentage of value in majority-Black neighborhoods—that highlight the importance of technological equity. Potential inequities produced by AVMs may reflect data omission. But they may also result from racial disparities in model inputs or from the modeling techniques AVMs use.

In this brief, we build on our previous study by testing each of these possibilities. We find that gathering additional data on property condition and employing more sophisticated artificial intelligence techniques can help us more accurately assess the percentage magnitude of AVM error and its underlying contributors. But even with data improvement and artificial intelligence, we still find evidence that the percentage magnitude of AVM error is greater in majority-Black neighborhoods. This indicates that we cannot reject the role historic discrimination has played in the evaluation of home values. But we also suggest more research exploring the dimensions of data and modeling to ensure the homebuying process benefits everyone seeking to achieve or maintain the American dream.

Background

The value of one's home is a key part of a household's assets and net worth. Residential appraisals are the primary method through which properties are valued for home purchases or mortgage refinances. And the home value combined with any new mortgage captures the amount of housing equity at origination.

But recent analysis has suggested that appraised estimates of homes in Black neighborhoods may systematically underestimate the property's value (Howell and Korver-Glenn 2018; Narragon et al.

2021). Undervaluation may improve borrowers' ability to renegotiate to a lower contract price, thus improving homebuying affordability (Fout and Yao 2016). At the same time, lower appraised property values may contribute to the broader racial gap in the financial benefits associated with homeownership (Neal et al. 2021).

The good news is that AVMs may reduce racial bias in home values appraisers have estimated (Williamson and Palim 2022). AVMs facilitate mortgage transactions by reducing the human input in residential property valuations. Hypothetically, reducing human input should reduce racial disparities in property valuations. But a previous paper we wrote suggested that AVMs can produce racially disparate home value outcomes (Neal et al. 2020).

In that report, which compared majority-Black and majority-white census tracts in the Atlanta, Memphis, and Washington, DC, core-based statistical areas (CBSAs), we did not find systematic evidence that AVMs undervalued sales prices in majority-Black neighborhoods or majority-white neighborhoods. And the absolute AVM error, measured as the absolute-value distance between the AVM estimate and sales price, was greater, on average, in majority-white neighborhoods than in majority-Black ones.

But the percentage magnitude of AVM error, which measures the absolute difference as a share of the sales price, was greater in majority-Black neighborhoods. This indicates that the degree of absolute AVM error in majority-Black neighborhoods is magnified by the significantly lower home prices in majority-Black neighborhoods. Yet, even after controlling for property differences, neighborhood conditions, and turnover, a neighborhood's majority race was still a significant determinant of the percentage magnitude of AVM error.

These findings illustrate that AVMs both undervalue and overvalue sales prices, both of which can be harmful. Undervaluation can limit wealth gains for homeowners seeking to refinance or sell their home. But overvaluation may result in credit risk holders underestimating risk and may speed up irrational inflation of property values, potentially resulting in a future home price correction (PAVE 2022). Finally, lower home values in majority-Black neighborhoods, partly reflecting historic discrimination, increase the risk of AVM error. Although we do not find systematic undervaluation bias in AVMs, we do observe that our AVM produced a racially disparate outcome in the form of a greater percentage magnitude of AVM error in majority-Black neighborhoods than in majority-white neighborhoods.

Analyzing the percentage magnitude of AVM error suggests that AVMs' racial disparities partly reflect the key *inputs* that have contributed to systematically lower home values in majority-Black neighborhoods. The magnitude of error also suggests that understanding AVMs' racial *outcomes* requires analysis of both tails of the AVM estimate distribution, not just the bottom one.

History and research have illustrated the role historic discrimination has played in determining home values in Black neighborhoods (Neal, Choi, and Walsh 2020). Another potential contributor to AVM error is a lack of data.[1] For example, AVMs rely on a large amount of historical sales data and may include home price forecasts to make them more responsive to current housing market conditions. But many AVMs do not have a strong sense of a property's condition.[2] The absence of these data could weaken AVM accuracy and contribute to a greater percentage magnitude of AVM error.[3]

Property condition is a key contributor to a home's value. But property condition may vary by race. For example, Black homeowners are more likely than white homeowners to live in inadequate housing (Neal, Choi, and Walsh 2020). This difference may contribute to greater absolute AVM error by potentially overvaluing homes in majority-Black neighborhoods. At the same time, a deluge of distressed home sales—which are greater in the majority-Black neighborhoods we assessed but can also have poorer property conditions—may result in the undervaluation of other homes in the neighborhood if that distressed home is immediately used as a comparable sale in the mortgage transaction of a nondistressed home (Conklin, Coulson, and Diop 2022).

In our previous study, we did not control for property condition. To strengthen our AVM analysis, this brief includes a measure of property condition in the analysis of percentage magnitude AVM error. Incorporating a measure of property condition into our analysis could help us understand the role data omission plays in producing AVM error.

The rest of this brief proceeds as follows. First, we describe our measure of property condition and illustrate why it is a reasonable indicator of what an appraiser may assess. We then update our regression analysis with this new measure and report its impact. Next, we identify modeling weaknesses of a standard econometric ordinary least squares (OLS) model and offer a substitute algorithm based on artificial intelligence. We then report how changing modeling techniques and adopting machine learning tools, in addition to adding the property condition variable, can produce a more accurate assessment of the percentage magnitude of AVM error. We end with an interpretation of our results and offer key policy implications and concluding thoughts on the direction of future research.

Data on Property Condition

To capture property condition, we use a measure called the exterior condition rating (ECR). The property intelligence firm CAPE Analytics provided us property-level ECRs.

CAPE Analytics creates and applies computer vision algorithms to high-resolution images captured from airplanes to create structured data that include the ECR. The ECR covers all a parcel's visible external features, including roofs, yards, driveways, and debris. The rating is measured on a five-point scale from severe to excellent (severe, poor, fair, good, and excellent). Table 1 provides the five-point scale definitions.

TABLE 1

CAPE Analytics Exterior Condition Rating Scale Definitions

Rating	Definition
Excellent	Parcel condition falls within the best 5% of parcels
Good	Parcel condition falls within the best 20% but not the best 5% of parcels
Fair	Parcel condition is average (50% of parcels)
Poor	Parcel condition falls within the worst 23% but not the worst 2% of parcels
Severe	Parcel condition falls within the worst 2% of parcels
Unknown	Parcel could be assigned a property condition

Source: CAPE Analytics.

In the previous report, we analyzed Atlanta, Georgia; Memphis, Tennessee; and Washington, DC. Each city had a significant Black population share and produced solid property-level pairings between AVM estimates and sales prices to analyze. In each city, instead of using the entire CBSA, we used the counties with strong historical deeds data that we could match with the AVM data. These counties are a small proportion of the total number of counties in each CBSA but account for the majority of the CBSA population. The Atlanta, Memphis, and Washington, DC, counties account for 17 percent, 22 percent, and 33 percent of the total counties in their CBSAs, respectively, and 63 percent, 74 percent, and 56 percent of their respective populations.

In this analysis, we match our property records data for these three metropolitan areas with the ECRs from CAPE Analytics based on property latitudes and longitudes, parcel lot assessor parcel numbers, and transaction dates. The match rates are 98 percent for Atlanta, 90 percent for Memphis, and 44 percent for Washington, DC. For Atlanta and Memphis, the small share of unmatched properties was proportionately distributed between majority-Black and majority-white neighborhoods and thus do not skew the overall distribution. The match rate is so low for Washington, DC, because a sizeable portion of observations in the property records data do not have valid coordinates coded to the rooftop level. By using only assessor parcel numbers and transaction dates, we cannot match those observations with valid property records in the CAPE Analytics database. Therefore, we exclude Washington, DC, and use only Atlanta and Memphis in this analysis. Despite excluding observations from Washington, DC, we still replicate the results from our previous report, giving us confidence to incorporate the ECR measure.

For our analysis, we collapse the five-point ECR scale from CAPE Analytics into three categories: poor (includes poor and severe), fair, and good (includes good and excellent). Table 2 presents the ECR distributions based on the grouped categories for the matched sample within the Atlanta and Memphis CBSAs.

In the Atlanta and Memphis CBSAs, single-family properties in majority-Black neighborhoods are more likely to have a poor rating and are less likely to have a fair or good rating than those in majority-white neighborhoods (table 2). In Atlanta, 46 percent of single-family properties in majority-Black neighborhoods had a poor rating in 2018, compared with 34 percent in majority-white neighborhoods. In Memphis, 44 percent of single-family properties in majority-Black neighborhoods had a poor rating, compared with 34 percent in majority-white neighborhoods.

TABLE 2

ECR Distribution in the Atlanta and Memphis CBSAs

CBSA	ECR	Majority-Black neighborhoods	Majority-white neighborhoods
Atlanta-Sandy Springs-Roswell, GA	Good	9%	13%
Atlanta-Sandy Springs-Roswell, GA	Fair	45%	52%
Atlanta-Sandy Springs-Roswell, GA	Poor	46%	34%
Memphis, TN-MS-AR	Good	10%	14%
Memphis, TN-MS-AR	Fair	46%	52%
Memphis, TN-MS-AR	Poor	44%	34%

Source: Urban Institute calculations using data from the American Community Survey, CAPE Analytics, and a major property records provider.

Note: CBSA = core-based statistical area; ECR = exterior condition rating.

Intuitively, an assessment of property condition reflects both external and internal adequacy (Neal, Choi, and Walsh 2020). Before examining the impact of our ECR measure on the percentage magnitude of AVM error, we first establish that external property condition is a reasonable proxy for the property condition overall, both inside and out. To do so, we calculate the polychoric correlation[4]—the correlation between two categorical variables—between exterior property conditions and interior structural conditions, using American Housing Survey (AHS) data.

The AHS is a recognized source of information on property condition, albeit with a limited suite of variables and geographic granularity. We use the survey's information on roofs and outside walls across owner-occupied homes nationwide to assess exterior conditions, and we use its information on fundamental or structural problems, such as floors, windows, foundations, and peeling paint, to assess interior conditions. We find a polychoric correlation of 0.67 between exterior and interior conditions.

This polychoric correlation should be regarded as a lower-bound estimate of the true strength of the correlation because of the AHS's limited variables to capture a property's exterior condition. Compared with AHS variables that cover only roofs and outside walls, the ECRs in our analysis cover all a parcel's visible external features, including roofs, yards, driveways, and debris. Because the ECR variable in our analysis is a more comprehensive measure of exterior condition, its correlation with interior condition should be greater than 0.67, suggesting that it should be a reasonable proxy for the property condition overall.

How Much AVM Error Can Be Explained by Property Conditions?

To determine how the ECR contributes to the percentage magnitude of AVM appraisal inaccuracy in the Atlanta and Memphis CBSAs, we first conduct the OLS regressions, with 2018 as our analysis period, focused only on single-family home purchases. In addition to the year of data, we follow the model specification in our previous report and control for key neighborhood characteristics affecting the percentage magnitude of AVM inaccuracy. These neighborhood characteristics are grouped along four dimensions: home values, differences in properties within a neighborhood, neighborhood conditions, and turnover rates. Table 3 presents summary statistics of those variables.

TABLE 3

Summary Statistics

	Black Neighborhood		White Neighborhood	
Variable	**Mean**	**SD**	**Mean**	**SD**
Home value	127,756	80,969	329,443	204,256
Property age	46.4	24.3	37.5	22.1
Standard deviation of neighborhood property ages	14.0	7.1	12.5	6.5
Percentage deviation of neighborhood property values	43.2%	14.8%	34.4%	11.3%
Gentrified neighborhood	7.5%	26.3%	2.1%	14.3%
Share of neighborhood distressed home sales	15.7%	20.8%	5.0%	13.7%
Neighborhood median household income	46,198	16,657	92,312	30,955
Neighborhood number of households	2,320	1,328	2,630	1,214
Turnover rate at neighborhood level	8.8%	4.1%	7.5%	3.3%
ECR				
Good	9.0%	29.0%	13.4%	34.0%
Fair	45.2%	50.0%	52.0%	50.0%
Poor	45.8%	50.0%	34.0%	47.0%

Source: Urban Institute calculations using data from the American Community Survey, CAPE Analytics, and a major property records provider.

Note: ECR = exterior condition rating; SD = standard deviation.

Using the variables in table 3, we conduct a regression analysis using OLS to examine the ECR's impact on the percentage magnitude of inaccuracy. Table 4 presents the results of these regressions. In all the regressions, we include county fixed effects to control for local factors. The dependent variable is the percentage magnitude of AVM inaccuracy. A positive sign in the coefficient means the independent variable is associated with a higher percentage magnitude of inaccuracy. For example, the coefficient of the percentage deviation of neighborhood property values (0.422***) shows that a 1 percentage-point increase in the percentage deviation of neighborhood property values leads to a 42 basis-point increase in the percentage magnitude of inaccuracy. In this example, the three asterisks indicate that the coefficient is statistically significant at the 99 percent confidence level.

The results in table 4 indicate that an ECR rating worse than good would raise the percentage magnitude of AVM error. Relative to an otherwise similar property with a good rating, a property with a fair rating would increase the AVM's percentage magnitude of error by 2.72 percentage points. Similarly, relative to a property with a good rating, a property with a poor rating would further increase AVM inaccuracy, increasing the percentage magnitude of error by 4.35 percentage points. In this case, the magnitude of the coefficient means that for a home with an average sales price of $250,000, having a poor rating is associated with a $10,875 greater percentage AVM error than a property with a good rating, holding all other attributes constant.

As we hypothesized, adding property condition to our regression analysis reduces the impact of the neighborhood's majority race on the percentage magnitude of error, but only slightly. After controlling for the ECR, the magnitude of this Black neighborhood coefficient is slightly reduced from 3.593 percentage points in column 4 to 3.499 percentage points in column 5. This indicates that even when controlling for property condition, location in a majority-Black neighborhood rather than a majority-white one still raises the percentage magnitude of error by 3.499 percentage points. The difference is a $4,549 greater percentage AVM error for a home with an average sales price of $130,000 in a majority-Black neighborhood, compared with a property with the same attributes and sales price in a majority-white neighborhood. This result is significant at the 99 percent confidence level.

TABLE 4

Regression Results

	Dependent Variable: Percentage Magnitude of AVM Inaccuracy				
	(1)	(2)	(3)	(4)	(5)
Black neighborhood	21.024*** (0.393)	4.816*** (0.504)	4.040*** (0.499)	3.593*** (0.542)	3.499*** (0.542)
Log (Home value)		-15.785*** (0.316)	-12.535*** (0.328)	-10.358*** (0.402)	-10.075*** (0.403)
Standard deviation of neighborhood property ages			0.155*** (0.028)	0.058** (0.029)	0.059** (0.028)
Percentage deviation of neighborhood property values (%)			0.453*** (0.014)	0.422*** (0.014)	0.422*** (0.014)
Share of neighborhood distressed home sales (%)				-0.005 (0.010)	-0.006 (0.010)
Gentrified neighborhood				2.155*** (0.817)	2.174*** (0.817)
Log (Neighborhood median household income)				-4.116*** (0.652)	-4.153*** (0.652)
Log (Number of households in neighborhood)				-5.107*** (0.377)	-5.016*** (0.377
Neighborhood-level turnover rate (%)				-0.246*** (0.049)	-0.230*** (0.049)
ECR: Fair					2.718*** (0.530)
ECR: Poor					4.350*** (0.546)
Constant	13.860*** (0.752)	213.240*** (4.063)	156.708*** (4.336)	219.975*** (6.708)	213.170*** (6.765)
County fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	62,609	62,609	62,606	62,606	62,606
R^2	0.086	0.121	0.138	0.142	0.143
Adjusted R^2	0.086	0.121	0.138	0.142	0.143
Residual standard error	41.587 (df=62602)	40.784 (df=62601)	40.383 (df=62596)	40.297 (df=62591)	40.276 (df=62589)

	Dependent Variable: Percentage Magnitude of AVM Inaccuracy				
	(1)	**(2)**	**(3)**	**(4)**	**(5)**
F-statistics	981.818*** (df=6; 62606)	1,230.664*** (df=7; 62601)	1,115.744*** (df=9; 62596)	739.866*** (df=14; 62591)	652.256*** (df=16; 62589)

Source: Urban Institute calculations using data from the American Community Survey, CAPE Analytics, and a major property records provider.

Note: AVM = automated valuation model; df = degrees of freedom; ECR = exterior condition rating.

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Adding the ECR to our regressions demonstrates that property condition is correlated with percentage AVM error, but the ECR variable does not significantly increase the OLS model's goodness of fit, as represented by the *R*-squared. Adding the ECR increased our model's *R*-squared only from 0.142 to 0.143—that is, 14.3 percent of the observed variation in the percentage magnitude of inaccuracy can be explained by our model's inputs.

It is not always an issue for a variable to have a limited impact on the *R*-squared when it also has a statistically significant coefficient, as is the case here, because they represent different measures. The coefficient's statistical significance indicates the strength of the relationship between the independent variable (ECR) and the dependent variable (AVM percentage magnitude of inaccuracy), while the *R*-squared represents the model's goodness of fit. Still, we perform a few tests to narrow the explanation for our *R*-squared value.

One potential source of this seemingly divergent result is multicollinearity (Shrestha 2020). Multicollinearity occurs when the multiple linear regression analysis includes several variables that are significantly correlated not only with the dependent variable but with each other. We investigate whether any of our independent variables are "collinear" with each other by performing a variance inflation factor (VIF) test (table 5).[5] The general rule is that VIFs exceeding 5 warrant further investigation, while VIFs exceeding 10 are signs of severe multicollinearity requiring correction. Based on the VIF results, ECR is not highly correlated with other independent variables, ruling out multicollinearity as an explanation for the low *R*-squared value.

TABLE 5

Variance Inflation Factor Test Results

	GVIF	Df	GVIF^(1/(2*Df))
Black neighborhood	2.769905	1	1.664303
Log (Home value)	3.475924	1	1.864383
Standard deviation of neighborhood property ages	1.450068	1	1.204188
Percentage deviation of neighborhood property values (%)	1.487771	1	1.219742
Neighborhood distressed home sales share (%)	1.128677	1	1.062392
Gentrified neighborhood	1.075338	1	1.036985
Log (Neighborhood median household income)	4.045038	1	2.011228
Log (Number of households in neighborhood)	1.353252	1	1.163294
Neighborhood-level turnover rate (%)	1.259577	1	1.122309
ECR	1.033595	2	1.008295
County fixed effects	2.642358	5	1.102045

Source: Urban Institute calculations using data from the American Community Survey, CAPE Analytics, and a major property records provider.

Note: Df = degrees of freedom; ECR = exterior condition rating; GVIF = generalized variance inflation factor.

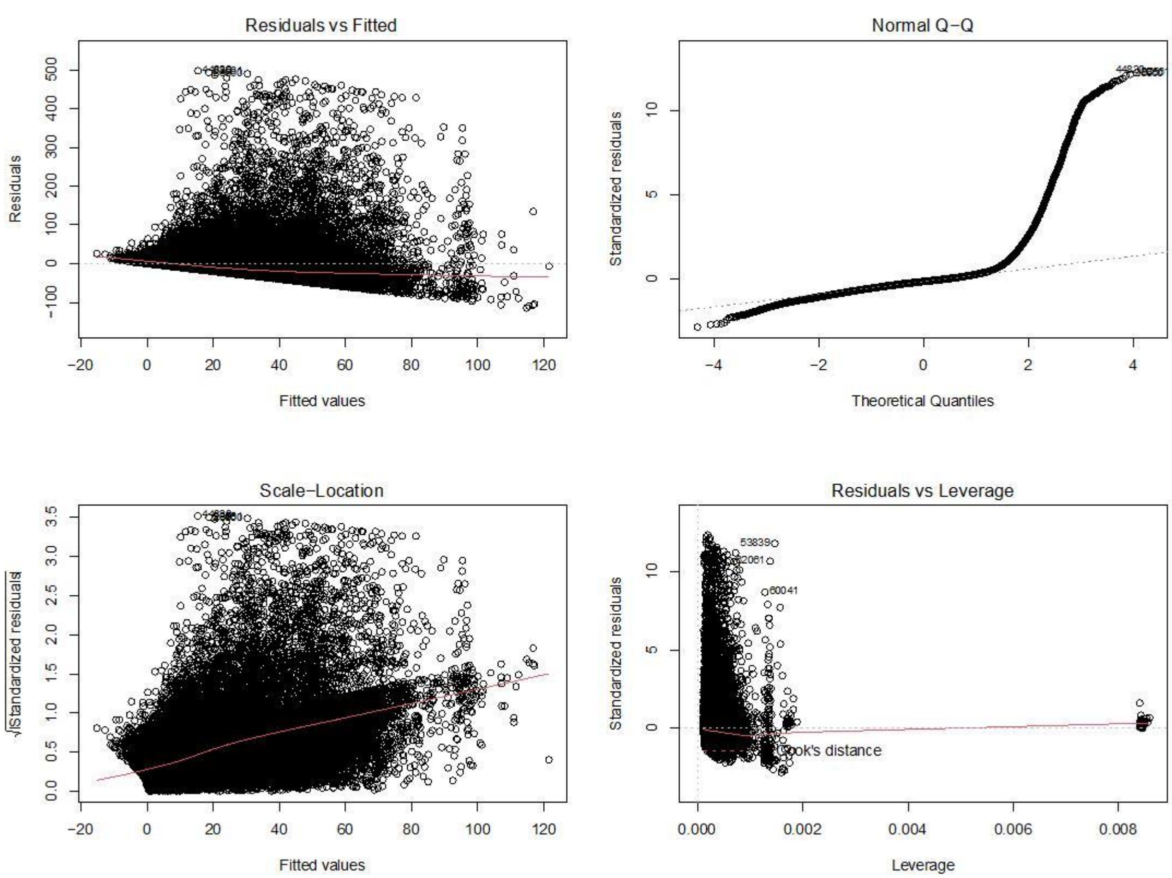
Another potential explanation of weak model fit as measured by the *R*-squared is the data's underlying structure. Our results suggest that a linear regression may not be the type of regression best suited to the data's spread. To confirm whether an OLS regression is the best approach for examining property condition's impact on AVM accuracy, we run several diagnostic tests.

Linear regression usually makes several assumptions about the data: (1) a linear relationship between the dependent variable and the independent variables; (2) normality of the residuals—that is, the residual errors are assumed to be normally distributed; (3) homoscedasticity—that is, the residuals are assumed to have a constant variance; and (4) independence of residuals error terms. The four diagnostic test results shown in figure 1 suggest the data structure in this analysis does not meet those linear assumptions.

The residuals-versus-fitted plot indicates that randomness of the error term was not met. The Normal Q–Q plot shows that the residuals from our OLS regressions (column 5) are not normally distributed. In addition, the scale-location plot shows severe heteroscedasticity problem. All these suggest that OLS regression may not be the best approach.

FIGURE 1

Diagnostic Tests for Linear Regression Accuracy



URBAN INSTITUTE

Source: Urban Institute calculations using data from the American Community Survey, CAPE Analytics, and a major property records provider.

Nonparametric Supervised Machine Learning Approach: LightGBM

Nonparametric supervised machine learning (machine learning) is a highly innovative and effective vein in predictive data analysis[6] and has several advantages over traditional linear parametric methods such as OLS. First, machine learning methods fully use the available historical data. By repeatedly validating the model through training and prediction sets derived from existing data, the methods can map new data entries into specific dependent variables, based on relevant independent variables used to train the model. Second, they possess great capacities and effectiveness in handling interrelated variables (e.g., collinearity) (Aggarwal 2015), thus boosting the prediction accuracy from traditional regression methods. Third, machine learning methods do not assume linearity and can handle complex datasets that do not fulfill the requirements of traditional regression models.

LightGBM is among the most recent and most efficient machine learning prediction algorithms (Ke et al. 2017). It provides more regularized model formalization and better overfitting control (Ashari, Paryudi, and Tjoa 2013). It is also an algorithm that assumes no linearity, providing more appropriate handling to our complex dataset. We thus choose LightGBM as a nonparametric, tree-based machine learning counterpart to our OLS model. And this helps us explore the broader question of whether and how sophisticated artificial intelligence tools improve analysis of automated systems.

Methodology

We first partition the entire dataset to a training set (70 percent) and a testing set (30 percent). We set up cross-validation through a stratified k-fold (k = 5) process. We enter all relevant independent variables into the LightGBM model as predictors and enter the outcome variable, the percentage magnitude of AVM inaccuracy, as the prediction target. We then employ a Bayesian optimization procedure to obtain the model parameters that support the most accurate predictions of the target variable. We describe the methodology below.

DATA PARTITIONING AND MODEL VALIDATION

In this study, we divide the processed dataset for Memphis and Atlanta into two portions—the training set and the testing set—to regulate the efficiency of the machine learning procedures. The LightGBM model is trained using only the training set and tested using only the testing set. This split is vital to demonstrate and tune the model's response to new data being processed for the first time. For the robustness of the division, we put 70 percent of the data into the training portion and the remaining 30 percent into the testing portion.

To enhance the model's validity, accuracy, and robustness, we also employ a 5-fold cross-validation procedure on the training set. We adopt the k-fold (k = 5) cross-validation because of its efficiency and smoothness during the validation. Each dataset is randomly separated into k numbers of folds, where k-1 folds are used for training purposes, and the remaining fold is simultaneously used for testing. The results over the k testing folds are averaged at the end.

MODEL PARAMETERS

To tune the hyperparameters of the LightGBM model, in conjunction with the k-fold cross-validation procedure, we employ a Bayesian optimization procedure to obtain the model parameters that best predict the regression outcome. The parameter optimization boundaries are listed below:

- Learning rate: 0–1
- Number of leaves: 5–40
- Minimum gain to split: 0–10
- Minimum sum of hessian in leaf: 0–20

The final optimized LightGBM model has the following parameters:

- Number of threads: 6
- Number of leaves: 25
- Learning rate: 0.468
- Minimum gain to split: 1.823
- Minimum sum of hessian in leaf: 9.517

With those parameters, we now obtain our optimized LightGBM prediction model based on the 70 percent training set.

EVALUATION OF MODEL ACCURACY

Root mean square error (RMSE) is the standard deviation of the residuals (predicted errors) and is used to measure the accuracy of model prediction. We take the advantage of its strong interpretability, as it has the same unit as our regression target variable.

We test the RMSE for the LightGBM prediction model and compare it against the RMSE for the OLS model to test whether our LightGBM model makes more accurate predictions than the OLS model.

IDENTIFICATION OF AVM RACIAL DISPARITY: FEATURE IMPORTANCE

Shapley Additive Explanations (SHAP)[7] is a novel way of computing feature contribution toward the prediction while preserving the sum of contributions being equal to the final outcome. It is especially well suited for tree-based models. SHAP values calculate a feature's importance by comparing what a model predicts with and without the feature. Given that the order in which a model sees a feature can affect its predictions, SHAP values account for all possible orders to make sure all features are fairly compared.

To determine our predictors' relative importance and impact on the model outcome, we calculate the SHAP values for each predictor. Their SHAP values would allow us to delve deeper into the predictive model's complexity and partially unveil the machine learning black box. This would help us evaluate the impact of neighborhood race and ECR on predicted AVM error.

QUANTIFICATION OF FEATURE IMPORTANCE: SYNTHETIC CONTROL METHOD

Though the SHAP value could provide evidence on a specific feature's importance, it does not quantify the magnitude of the impact. Thus, to quantify the impact, we employed a synthetic control method to examine our identified racial disparity in AVM valuations, the ECR's impact , and the impact of the intersection of neighborhood majority race and the ECR. The results would shed light on whether systemic racism is a key factor behind the AVM error. Below, we discuss how we construct the synthetic data groups.

Results

LIGHTGBM HAS A GREATER PREDICTIVE POWER THAN THE OLS REGRESSIONS

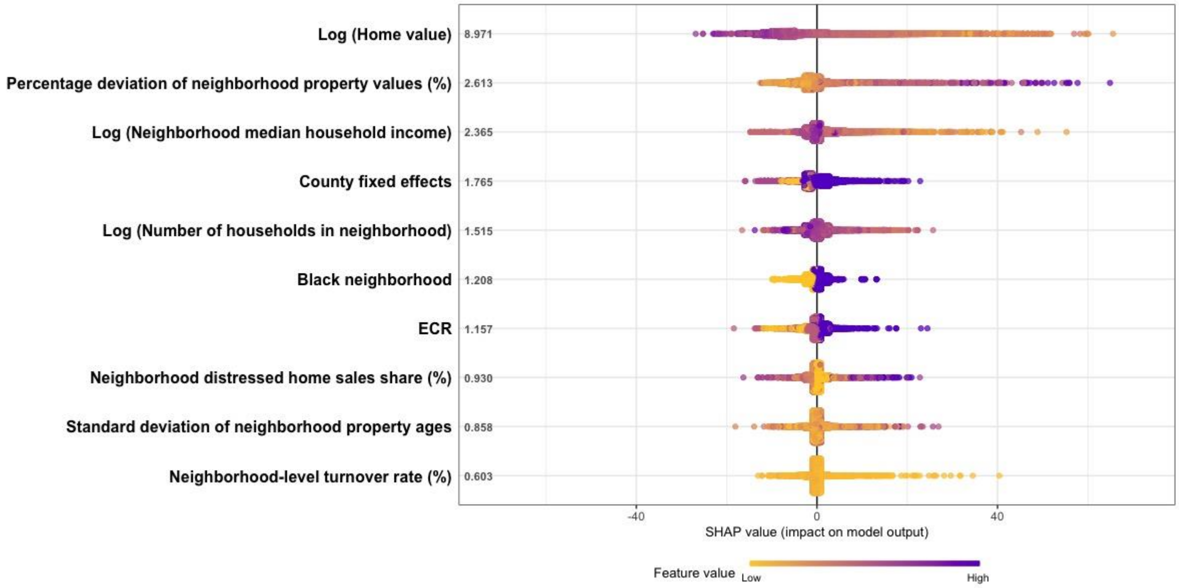
After completing data partitioning, model validation, and parameter tuning, our optimized LightGBM model produced an RMSE of 40.4. We apply the same data partitioning procedure to the OLS regression and got an RMSE of 46.2. This suggests that LightGBM produces a 5.8 percentage-point improvement in prediction accuracy. The magnitude of RMSE improvement does not differ between majority-Black and majority-white neighborhoods (only around a 0.05 percentage-point difference). Our results validate our selection of LightGBM over OLS regressions with respect to evaluating our identified AVM racial disparity. By relaxing the linear assumptions, this nonparametric, tree-based machine learning approach provides more appropriate handling of our complex dataset.

MAJORITY-BLACK NEIGHBORHOODS ARE ASSOCIATED WITH GREATER PREDICTED AVM INACCURACY

Figure 2 illustrates the SHAP values for each feature. The *y* axis displays the feature name, in order of importance from top to bottom. The value next to the variable name is the mean SHAP value. The *x* axis is the SHAP value. Each point represents a row from the training dataset. The gradient color represents that variable's original value. Continuous numerical variables, such as the log of home values, can contain the whole color spectrum. For dummy variables such as majority-Black neighborhood, it can take only two colors.

FIGURE 2

Shapley Additive Explanations Values



URBAN INSTITUTE

Source: American Community Survey, CAPE Analytics, and a major property records provider.
Note: ECR = exterior condition rating; SHAP = Shapley Additive Explanations.

For example, the percentage deviation of neighborhood property values is a key feature contributing to the AVM percentage error prediction, as it ranks as the second feature, after the home values in log form. The percentage deviation of neighborhood property values is associated with high and positive values on the target (figure 2). High value is shown by the color distribution in the color bar at the bottom of the figure, with purple indicating high percentage deviation of neighborhood property values. Positive value is indicated by the preponderance of purple observations on the right side of the zero vertical line, signaling the SHAP value is above zero. Given that it is a continuous variable, its color scheme contains the full color spectrum of the feature value from low (yellow) to high (purple). This suggests that greater heterogeneity with respect to property values in a neighborhood contributes to a greater predicted percentage magnitude of AVM error.

Figure 2 illustrates that the majority-Black neighborhood variable is associated with high and positive values on the target. The high value is because of the figure value color bar and is positive from the *x*-axis SHAP value. And because this is a dummy variable, its color scheme contains only two colors for low (yellow) and high (purple). This high and positive relationship suggests that compared with majority-white neighborhoods, AVM error in majority-Black neighborhoods is greater.

Similarly, the ECR is associated with high and positive values on the target. The ratings are coded as 1 (good), 2 (fair), and 3 (poor), so high ECR values means poor property conditions. This indicates that properties in poor condition are associated with greater AVM inaccuracy.

In addition, if we look at the order of importance from top to bottom, the majority-Black neighborhood variable ranks higher than the ECR, distressed sales share, and turnover rate variables. The SHAP values for neighborhood majority race combined with its ranking align with what we found in the OLS regressions, suggesting that even though an AVM algorithm does not have disparate input, such as race, it still can produce racial disparities.

HISTORIC RACISM MAY BE A KEY FACTOR BEHIND THE AVM ERROR

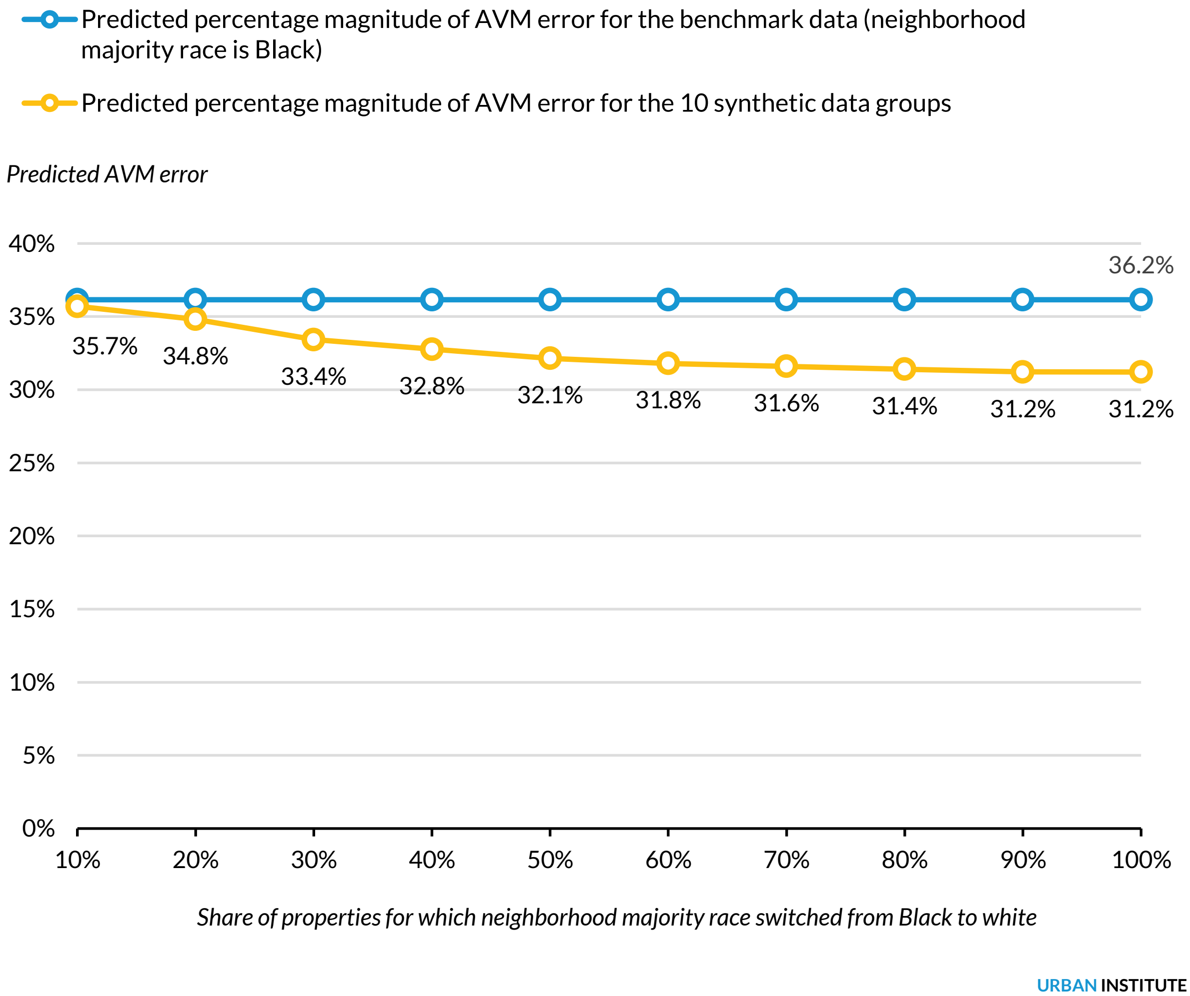
Though the SHAP value could show the importance of neighborhood majority race, it does not quantify the impact's magnitude. Thus, to quantify this racial disparity, we employ a synthetic control method.

We extract all properties in majority-Black neighborhoods from the test set and treat this selected group as our benchmark dataset. First, we predict the AVM error for the benchmark dataset using our constructed LightGBM model. Second, to create the corresponding synthetic datasets, we change the neighborhood majority race variable from Black to white, holding everything else constant. We create 10 synthetic datasets, with the share of properties of which the neighborhood race being altered starting from 10 percent to 100 percent. For the 10th synthetic dataset, all properties' neighborhood majority race has been switched from Black to white. Third, we predict the AVM error for each synthetic dataset using our constructed LightGBM model. We then obtain 10 corresponding predicted mean values of AVM error. Finally, we compare the predicted target variable from the synthetic datasets against the benchmark dataset. This difference between the two predicted values measures the racial disparity. The properties in the synthetic groups are the same as those in the benchmark group except that the neighborhood majority race is different.

The blue line in figure 3 illustrates the predicted percentage magnitude of AVM error for the benchmark data. The yellow line represents the predicted percentage magnitude of AVM error for the 10 synthetic data groups.[8] The *x* axis represents the share of properties for which neighborhood majority race switched from Black to white. As the share increases, the gap between the two lines widens. This indicates that, for example, if 60 percent of properties currently in majority-Black neighborhoods "move" to majority-white neighborhoods while keeping all other attributes constant, their associated predicted percentage magnitude of AVM error could decline from 36.2 percent to 31.8 percent, a 4.4 percentage-point difference. And if all properties currently in majority-Black neighborhoods "move" to majority-white neighborhoods, the predicted AVM error could decline by 5.0 percentage points, which is an upper-bound estimate of the racial disparity in AVMs. And it suggests that historic racism could be a key factor behind greater AVM error in majority-Black neighborhoods.

FIGURE 3

Racial Disparity: Neighborhood Majority Race



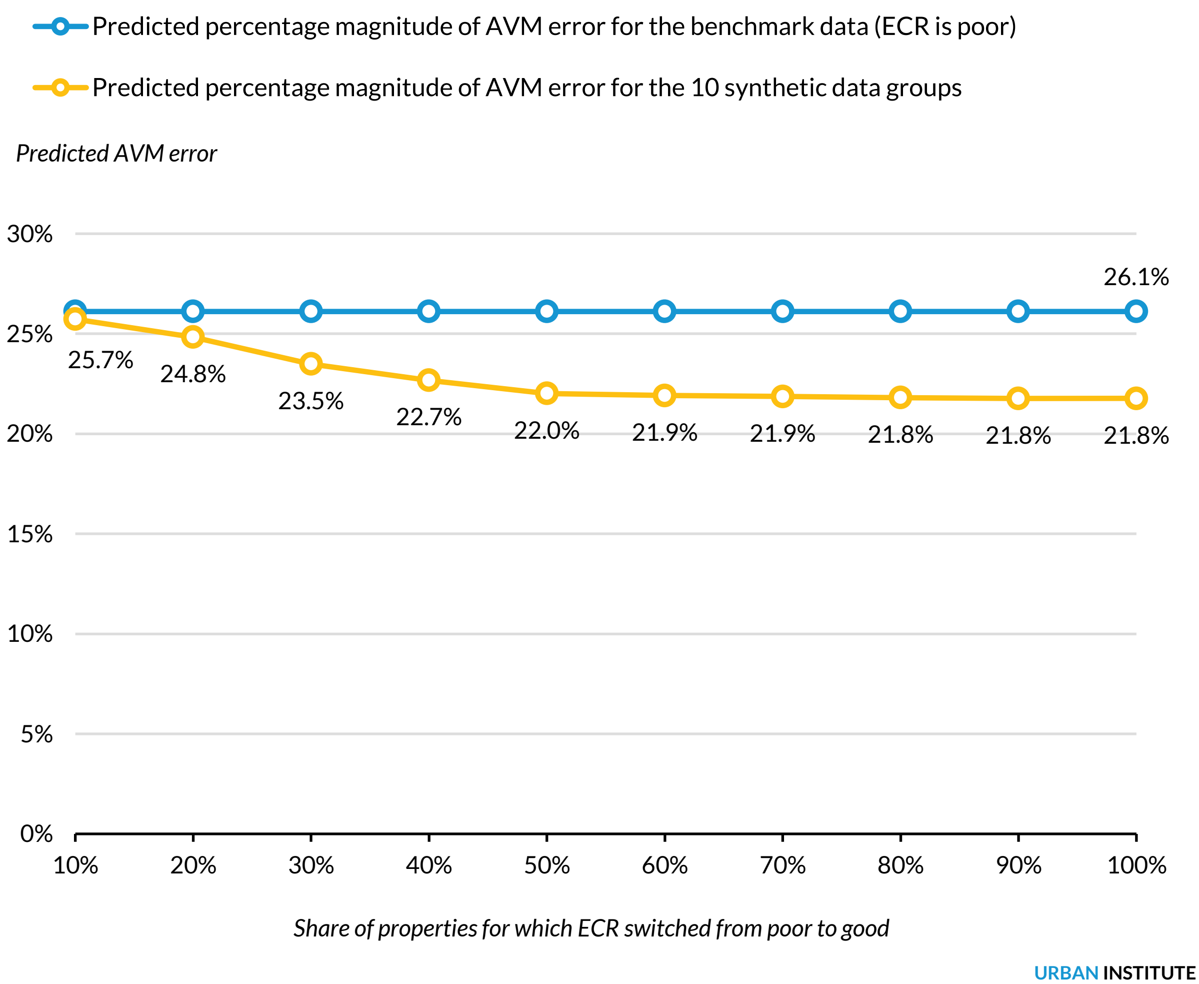
Source: Authors' calculations.
Note: AVM = automated valuation model.

To examine the ECR's impact, we apply a similar synthetic control approach. We use all properties with a poor ECR from the prediction set as our benchmark dataset (figure 4). We then create 10 corresponding synthetic datasets by changing the ECR from poor to good, holding everything else constant. Again, the share that switches from poor ECR to good ECR increases from 10 percent to 100 percent. For the 10th synthetic dataset, all properties' ECRs have changed from poor to good. We calculate the predicted AVM error for both the benchmark group (the blue line) and the synthetic data groups (the yellow line).

Figure 4 demonstrates that if all properties currently in poor condition upgraded to good condition, all else constant, their associated AVM error could decline from 26.1 percent to 21.8 percent, a 4.3 percentage-point difference. This provides further evidence that polices to improve housing adequacy could reduce the adverse impact of the percentage magnitude of AVM error.

FIGURE 4

Impact of ECR



Source: Authors' calculations.
Note: AVM = automated valuation model; ECR = exterior condition rating.

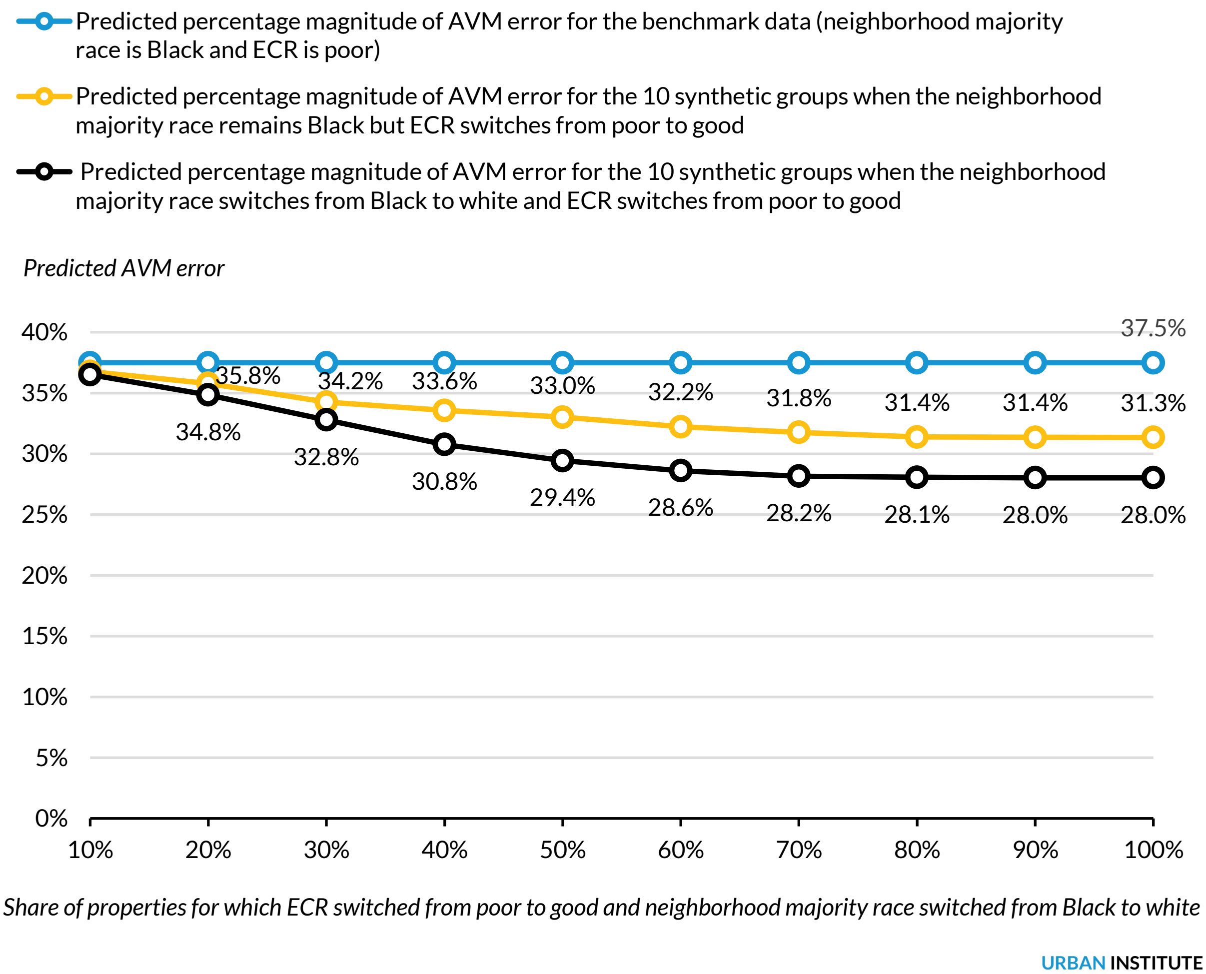
THE IMPACT OF THE INTERSECTION OF NEIGHBORHOOD MAJORITY RACE AND ECR

Lastly, we examine the intersection of neighborhood majority race and ECR. The benchmark data consist of properties in majority-Black neighborhoods rated as poor (the blue line in figure 5). Then, we change the benchmark data by first altering the ECR from poor to good, while keeping neighborhood majority race and all other attributes constant. As before, for each synthetic data group, we increase the share of change by 10 percent, ranging from 10 percent to 100 percent. Next, we change the neighborhood majority race from Black to white.

If we change all the properties in the benchmark group from a poor ECR to a good ECR, the predicted percentage magnitude of AVM error falls from 37.5 percent to 31.3 percent (figure 5). The AVM error will further decline to 28 percent once we flip the neighborhood majority race from Black to white. The gap between the yellow and black lines represents the impact of the intersection of neighborhood majority race and ECR. In other words, for two identical properties that both have improved their ECR from poor to good, the home located in a majority-Black neighborhood still experiences a 3.4 percentage-point greater percentage magnitude of AVM error, further suggesting that racial differences at the neighborhood level, which can reflect the impact of systemic discrimination, can play a role in producing percentage magnitude of AVM error.

FIGURE 5

Impact of the Intersection of Neighborhood Majority Race and ECR



Source: Authors' calculations.
Note: AVM = automated valuation model; ECR = exterior condition rating.

Implications

By incorporating a measure of property condition into our analysis, we can better understand the underlying contributors to AVM error. Notably, our analysis demonstrates that poor property conditions are associated with a greater percentage magnitude of AVM error and that including property conditions in AVM calculations is important for AVM developers looking to improve accuracy. As our analysis is based on AVM error in the Atlanta and Memphis CBSAs, we note these results may be specific to the AVM data we used and the geographic areas we analyzed. Further research will be needed to ascertain the universality of the percentage magnitude of AVM error determinants and which factors influence AVM accuracy in different metropolitan areas.

Implications for the Greater Use of Technology

If mortgage and housing providers are interested in determining the underlying contributors to the percentage magnitude of AVM error, OLS regression is helpful, based on our findings. But if the goal is to more accurately assess the shortcomings of AVM models and their underlying contributors using large and complex data, the housing industry should consider exploring algorithms the artificial intelligence community has developed. Artificial intelligence tools could enhance our understanding of the complexity of predictive algorithms and partially unveil the AVM black box.

Our nonlinear regression results demonstrate that using a LightGBM model that includes ECR data could reduce the predicted percentage magnitude of AVM error by 5.8 percentage points. This suggests that this nonparametric machine learning model more accurately copes with the complexity in variables of multiple dimensions. In addition, the SHAP values and the synthetic control approach shed light on the shortcomings of AVM algorithms that often get hidden in the black box.

With this machine learning tool, we could address the key weakness of our original OLS model. But other semiparametric or nonparametric machine learning tools may provide additional benefits, and we welcome further analysis to uncover the best solution. But even if the industry increases its use of machine learning tools, it is important to continually test for its impact on historically marginalized classes (Axelrod et al. 2022).

Implications for Appraisal Error

In our previous research, we showed that AVMs can produce racially disparate outcomes. These results may have reflected a lack of important data, the impact of systemic racism on home values, or a combination of the two. By updating this research with a measure of property condition—a likely driver of AVM error and a factor correlating with neighborhood racial makeup—our new analysis tested whether adding relevant property data could help explain the greater percentage magnitude of error AVMs produce in majority-Black neighborhoods. We find that including information on property condition does slightly reduce the impact of neighborhood racial composition on the percentage magnitude of AVM error. But this impact remains statistically significant, suggesting that even though adding key data may attenuate the full impact of neighborhood majority race on AVM error, it does not eliminate it.

These results contribute to the question of whether incorporating additional data into AVMs would better assess AVM error and improve AVM accuracy. Our results suggest that additional data can help explain the percentage magnitude of AVM error. And there is potential that adopting artificial intelligence tools could further reduce the percentage magnitude of AVM error. But our efforts do not eliminate the statistical significance of the neighborhood's majority race.

The continued significance of a neighborhood's majority race suggests that additional data may still be needed. In addition, the promising impact of machine learning models in more accurately measuring the percentage magnitude of AVM error and its key contributors suggests that exploring other artificial intelligence models could build on these findings. But even with data and modeling improvements, we still find evidence that the percentage magnitude of AVM error is greater in majority-Black neighborhoods. As a result, we cannot yet reject the role of historic racism, which has persistently penetrated through home values, property conditions, and neighborhood conditions. And inequities in each of these dimensions can produce lower home values, less adequate housing, and lower household incomes across majority-Black neighborhoods. By including, in the model, variables that demonstrate clear racial disparities, the AVM algorithm is more likely to produce relatively greater disparate outcomes in the form of percentage magnitude of AVM error.

Policy Recommendations

Problems with appraisals have come under renewed scrutiny (Narragon et al. 2021).[9] And this has captured federal policymakers' attention (PAVE 2022). By reducing the potential for human bias, increasing efficiency, and lowering costs, AVM use has the potential to reach a large scale. But ensuring AVM use produces accurate outcomes in all communities is critical. By not addressing the potential for error, particularly in neighborhoods with a history of low home values, increased use of AVMs may perpetuate past shortcomings. Our previous and updated findings have several implications for policies that could address racial disparities in AVM error.

Improve transparency in artificial intelligence and machine learning models. Using racially disparate data in artificial intelligence and machine learning models can contribute to racially disparate outcomes as well. Consequently, giving researchers and the public greater insight into the models' inputs would allow us to understand when and why racially disparate outcomes occur. Analyzing these models with a well-established auditing system could systemize AVM assessments (Akinwumi, Rice, and Sharma 2022). In addition, policy guardrails are necessary to ensure accountability and to protect consumers from discrimination or inequitable outcomes.

Expand and improve the data AVM algorithms use. Our analysis demonstrates that including data on property conditions can improve AVM accuracy. Including additional relevant inputs and improving the precision of existing variables can reduce AVM error.

Expand and improve access to renovation financing. Our analysis suggests that fair or poor property condition increases the percentage magnitude of error relative to properties in good condition. And research suggests that Black homeowners are more likely to live in inadequate housing relative to white homeowners. Although many municipalities offer renovation assistance, the amount of total financial assistance provided is often not enough to cover all affected homeowners. At the federal level, the Federal Housing Administration and the government-sponsored enterprises Fannie Mae and Freddie Mac have programs that support home improvements. But programmatic obstacles have limited their effectiveness.[10]

Strictly enforce the Fair Housing Act in the AVM space. The Fair Housing Act prohibits both disparate treatment and disparate impact based on race. As big data and automation continue to penetrate government and industrial organizations, ensuring these users understand what racially disparate estimations look like and how to address it will be important (Akinwumi et al. 2021). Even neutral algorithms need to be evaluated under the lens of disparate impact, because they can still disproportionately disadvantage certain groups.

Encourage appropriate regulatory oversight of AVMs and machine learning models. Our analysis illustrates that racial inequalities bake in home values, property conditions, and neighborhood conditions. By including, into the model, variables embedded with systemic racism, the AVM algorithm is more likely to produce a greater percentage magnitude of inaccuracy by replicating past discrimination. The Interagency Task Force on Property Appraisal and Valuation Equity developed a transformative set of actions that include a nondiscrimination quality control standard on AVMs. Encouraging regulatory oversight of AVMs and machine learning models will help ensure AVMs do not rely upon biased data that could reinforce past discrimination.

We welcome additional research into automation in the housing sector, anticipating that it can benefit from both traditional research and artificial intelligence and machine learning methods. As data become more ubiquitous and modeling techniques become more sophisticated, the ability to implement these tools will increase. Potential cost savings, especially when combined with labor shortages, will increase the incentive to adopt these tools. And shock events such as the COVID-19 pandemic, which made person-to-person contact risky, will only quicken adoption. But in addition to the efficiency gains, understanding the equity implications of automated decisionmaking tools, including AVMs, will be important going forward.

Conclusion

AVMs gained popularity before the pandemic by eliminating the need for an appraisal and hence reducing costs. The use of these models increased dramatically during the crisis, as AVMs allowed for the facilitation of mortgage transactions when appraisals proved difficult. AVMs will likely remain a key part of the mortgage industry toolkit going forward. It is important to realize that, even if human input is reduced, these models can still produce racial disparities. And this could have adverse effects for households of color and the communities in which they live, potentially contributing to racial disparities in wealth accumulation.

A portion of these shortcomings with AVMs can be overcome with additional data. And additional inequities can be eliminated with more sophisticated modeling techniques. An expansion in data and modeling may further reduce racial disparities in AVMs. But we cannot reject the role historic racism has played in the evolution of home values. To guard against the possibility of a lack of transparency or interpretability when incorporating sophisticated machine learning techniques, we suggest that AVMs be continuously audited for their potential impact on historically marginalized groups. At the same time, exploration of new data and modeling improvements to AVMs should continue, particularly through the use of machine learning, and with the appropriate regulatory oversight of AVMs in place as their use evolves. Understanding and eliminating the percentage magnitude of AVM error will ensure all communities experience equitable benefits from homeownership.

Notes

1 "Executive Order on Advancing Racial Equity and Support for Underserved Communities through the Federal Government," White House, January 20, 2021, https://www.whitehouse.gov/briefing-room/presidential-actions/2021/01/20/executive-order-advancing-racial-equity-and-support-for-underserved-communities-through-the-federal-government/.

2 James Dawkins, "What AVMs Can and Can't Do in the Appraisal Process," HousingWire, February 18, 2022, https://www.housingwire.com/articles/what-avms-can-and-cant-do-in-the-appraisal-process/.

3 Buchak and coauthors (2020) suggest that the lack of property condition may result in buyers' adverse selection. They also note that adverse selection is comparatively worse in less liquid markets. This finding would contribute to worse absolute AVM error in majority-white neighborhoods where turnover is lower. But it is offset by racial disparities in other key inputs, including the property condition.

4 See the entry for the polychoric correlation coefficient in Salkind (2010).

5 "10.7–Detecting Multicollinearity Using Variance Inflation Factors," Pennsylvania State University, Eberly College of Science, accessed April 25, 2022, https://online.stat.psu.edu/stat462/node/180/.

6 "Supervised Learning," IBM, accessed April 25, 2022, https://www.ibm.com/cloud/learn/supervised-learning.

7 See https://github.com/slundberg/shap.

8 We ran the model 10 times against the same dataset, which is the 30 percent test set. Because we randomly select shares of properties in the test set using a with-replacement approach, we would get slightly different results every time we created these 10 synthetic datasets. Our 70 percent–30 percent train-test split and the fivefold cross-validation ensures that the overall monotonically decreasing trend in figure 3 is statistically robust.

In addition, because we are doing a with-replacement selection here, we would get slightly different results every time we create these 10 synthetic datasets.

9 National Fair Housing Alliance, "Groundbreaking Report Identifies Bias and Systemic Barriers in Real Estate Appraisals," press release, January 19, 2022, https://nationalfairhousing.org/groundbreaking-report-identifies-bias-and-systemic-barriers-in-real-estate-appraisals/; and Michael Neal and Peter J. Mattingly, "Increasing Diversity in the Appraisal Profession Combined with Short-Term Solutions Can Help Address Valuation Bias for Homeowners of Color," *Urban Wire* (blog), Urban Institute, July 1, 2021, https://www.urban.org/urban-wire/increasing-diversity-appraisal-profession-combined-short-term-solutions-can-help-address-valuation-bias-homeowners-color.

10 Laurie Goodman and Edward Golding, "Institutional Investors Have a Comparative Advantage in Purchasing Homes That Need Repair," *Urban Wire* (blog), Urban Institute, October 20, 2021, https://www.urban.org/urban-wire/institutional-investors-have-comparative-advantage-purchasing-homes-need-repair.

References

Aggarwal, Charu C. 2015. *Data Mining: The Textbook*, vol. 1. New York: Springer.

Akinwumi, Michael, Lisa Rice, and Snigdha Sharma. 2022. "Purpose, Process, and Monitoring: A New Framework for Auditing Algorithmic Bias in Housing and Lending." Washington, DC: National Fair Housing Alliance.

Akinwumi, Michael, John Merrill, Lisa Rice, Kareem Saleh, and Maureen Yap. 2021. "An AI Lending Policy Agenda for the Federal Financial Regulators." Washington, DC: Brookings Institution.

Ashari, Ahmad, Iman Paryudi, and A. Min Tjoa. 2013. "Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool." *International Journal of Advanced Computer Science and Applications* 4 (11).

Axelrod, Judah, Alena Stern, Michael Neal, and Linna Zhu. 2022. "Comment Letter on NIST Initial Draft of Artificial Intelligence Risk Management Framework." Washington, DC: Urban Institute.

Buchak, Greg, Gregor Matvos, Tomasz Piskorski, and Amit Seru. 2020. *iBuyers: Liquidity in Real Estate Markets?* National University of Singapore, Institute of Real Estate and Urban Studies.

Conklin, James, N. Edward Coulson, and Moussa Diop. 2022. *Distressed Comps*. New York: SSRN.

Fout, Hamilton, and Vincent Yao. 2016. *Housing Market Effects of Appraising below Contract*. Washington, DC: Fannie Mae; Atlanta: Georgia State University.

Howell, Junia, and Elizabeth Korver-Glenn. 2018. "Neighborhoods, Race, and the Twenty-First Century Housing Appraisal Industry." *Sociology of Race and Ethnicity* 4 (4): 473–90. https://doi.org/10.1177/2332649218755178.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." In *Advances in Neural Information Processing Systems 30*, edited by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Viswanathan, and Roman Garnett, 3149–57. Long Beach, CA: Neural Information Processing Systems.

Narragon, Melissa, Danny Wiley, Doug McManus, Vivian Li, Kangli Li, Xue Wu, and Kadiri Karamon. 2021. "Racial and Ethnic Valuation Gaps in Home Purchase Appraisals." McLean, VA: Freddie Mac.

Neal, Michael, Jung Hyun Choi, Kathryn Reynolds, Joe Schilling, Gideon Berger, Elizabeth Champion, and Caitlin Young. 2021. *Why Do Households of Color Own Only a Quarter of the Nation's Housing Wealth When They Compose a Third of the Nation's Households?* Washington, DC: Urban Institute.

Neal, Michael, Jung Hyun Choi, and John Walsh. 2020. *Before the Pandemic, Homeowners of Color Faced Structural Barriers to the Benefits of Homeownership*. Washington, DC: Urban Institute.

Neal, Michael, Sarah Strochak, Linna Zhu, and Caitlin Young. 2020. "How Automated Valuation Models Can Disproportionately Affect Majority-Black Neighborhoods." Washington, DC: Urban Institute.

PAVE (Interagency Task Force on Property Appraisal and Valuation Equity). 2022. *Action Plan to Advance Property Appraisal and Valuation Equity: Closing the Racial Wealth Gap by Addressing Mis-valuations for Families and Communities of Color*. Washington, DC: PAVE.

Salkind, Neil J., ed. 2010. *Encyclopedia of Research Design.* New York: Sage Publications.

Shrestha, Noora. 2020. "Detecting Multicollinearity in Regression Analysis." *American Journal of Applied Mathematics and Statistics* 8 (2): 39–42.

Williamson, Jake, and Mark Palim. 2022. "Appraising the Appraisal: A Closer Look at Divergent Appraisal Values for Black and White Borrowers Refinancing Their Home." Washington, DC: Fannie Mae.

About the Authors

Linna Zhu is a research associate with the Housing Finance Policy Center at the Urban Institute. Her research centers on housing economics, aging, and real estate finance. Zhu holds a BA in economics from Renmin University of China, an MS in finance from Johns Hopkins University, and a PhD in public policy and management from the University of Southern California**.**

Michael Neal is a principal research associate in the Housing Finance Policy Center. Previously, he worked at Fannie Mae, where he was a director of economics in the Economic and Strategic Research division. Before his service at Fannie Mae, Neal was the assistant vice president at the National Association of Home Builder's Economic and Housing Policy department. As a housing economist, Neal has an in-depth knowledge of housing market trends and has provided expert analysis and commentary on housing to media outlets around the country. Previously, Neal worked for Congress's Joint Economic Committee, the Federal Reserve System, the Congressional Budget Office, and Goldman Sachs. Neal has a bachelor's degree in economics from Morehouse College and a master's degree in public administration from the University of Pennsylvania.

Caitlin Young is a policy analyst in the Housing Finance Policy Center. She graduated from the University of North Carolina at Chapel Hill with double majors in political science, for which she received highest honors, and in economics. She authored her senior thesis on the impact of state social welfare policy on recidivism outcomes. Before joining Urban, she interned with the Center for Community Self-Help, a community development financial institution.

Acknowledgments

This brief was funded by the Urban Institute.

The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples.

We thank CAPE Analytics for providing property-level data on property condition.

URBAN INSTITUTE

500 L'Enfant Plaza SW
Washington, DC 20024

www.urban.org

ABOUT THE URBAN INSTITUTE

The nonprofit Urban Institute is a leading research organization dedicated to developing evidence-based insights that improve people's lives and strengthen communities. For 50 years, Urban has been the trusted source for rigorous analysis of complex social and economic issues; strategic advice to policymakers, philanthropists, and practitioners; and new, promising ideas that expand opportunities for all. Our work inspires effective decisions that advance fairness and enhance the well-being of people and places.