Université du Québec en Outaouais

Département des sciences administratives

Working Paper No. 3

Hedonic Housing Price Models for the United States: A Multi-Method Comparison of Parametric, Quantile, and Machine Learning Approaches

Simon-Pierre Boucher

Département des sciences administratives
Université du Québec en Outaouais

simon-pierre.boucher@uqo.ca

Gatineau – Pavillon Alexandre-Taché
283, boulevard Alexandre-Taché
Gatineau, Québec, Canada J9A 1L8

October 2026

Version 1.2

Abstract

This paper compares econometric and machine-learning approaches to hedonic housing valuation on 788,842 active Zillow listings covering all 50 U.S. states and the District of Columbia, and asks how much of the machine-learning advantage survives when models must price homes in markets they have not seen. A semi-log OLS model with 62 regressors ($R^2 = 0.634$) yields listing-price gradients consistent with the hedonic literature; ZIP3 fixed effects raise out-of-sample R^2 to 0.725 and residual Moran's $I = 0.27$ shows strong spatial dependence. Allowing for that dependence matters: cluster-robust standard errors are, on median, 8 (ZIP3) to 12 (state) times the HC3 errors, and only 33 of 63 coefficients remain significant under state clustering. State-by-state estimation shows that the pooled coefficients average markedly different local price functions (living-area elasticities from 0.06 to 1.15) and that a fully interacted model gains 6.1 points of out-of-sample R^2 over state fixed effects. Quantile regression reveals distributional heterogeneity, with paired cluster-bootstrap Wald tests rejecting equality of the 10th- and 90th-percentile coefficients for 10 of 13 key attributes. XGBoost reaches $R^2 = 0.833$ under random validation but 0.425 under a ten-state holdout and a median of 0.653 across 51 leave-one-state-out models; once 0.6% of corrupt records are removed, OLS is as accurate out of state (0.539) as the best tree ensemble. The gap is traced to the prediction of market price levels rather than within-market structure (within-state $R^2 = 0.62$), to region dummies and to "neighbourhood" variables that are state-median imputations for 93% of listings, and to boosting capacity: held-out-state accuracy peaks after 75 rounds and declines thereafter while within-training accuracy keeps rising. SHAP rankings are stable across ensembles ($\rho > 0.89$), and the third-ranked feature is a state identifier. A coding error that set ln(lot) to zero for 16.5% of listings is corrected (lot elasticity 0.095, not 0.018). Estimates are interpreted throughout as conditional associations in asking prices, not willingness-to-pay parameters.

Keywords: hedonic price function, listing prices, quantile regression, gradient boosting, SHAP, spatial cross-validation, cluster-robust inference, automated valuation model

JEL Classification: R31, C21, C45, C52

1. Introduction

Housing is the dominant asset in the typical American household's portfolio, the collateral underpinning the largest class of household debt, and a central object of study in urban economics, household finance and public policy. How housing attributes map into prices matters well beyond academia: property-tax assessment, mortgage underwriting and the automated valuation models (AVMs) that increasingly mediate transactions all rest on some version of that mapping. The hedonic framework—formalised by Rosen (1974) on the consumer-theoretic foundations of Lancaster (1966), with empirical roots in Waugh (1928) and Court (1939)—provides the canonical approach: decompose observed prices into the implicit valuations of constituent characteristics. After five decades of applied hedonic research, three methodological tensions remain unresolved, and the arrival of machine-learning valuation has sharpened rather than settled them.

First, the standard OLS hedonic model imposes a log-linear relationship between attributes and prices, a choice made for robustness under attribute omission (Cropper et al., 1988) but one that may approximate the non-linear, interactive structure of housing markets poorly. Second, by estimating conditional mean effects, OLS constrains implicit prices to be constant across the price distribution, a restriction the quantile-regression literature has shown to be untenable (Zietz et al., 2008; McMillen, 2008; Liao and Wang, 2012). Third, while gradient-boosted models deliver large predictive gains in property valuation (Bourassa et al., 2010; Kok et al., 2017; Mayer et al., 2019), their black-box character limits adoption where economic interpretation matters (Mullainathan and Spiess, 2017; Athey and Imbens, 2019)—and, as this paper documents, their headline accuracy can be a partial illusion created by the way they are evaluated and by the geography hidden in their inputs.

This paper confronts the three tensions on a single dataset of 788,842 active Zillow listings covering all 50 U.S. states and the District of Columbia, with 62 regressors spanning structural attributes, lot characteristics, amenities, neighbourhood quality, market status and six interaction terms. We estimate three families of models: semi-log OLS with progressively finer geographic fixed effects and, in this revision, cluster-robust inference and state-by-state estimation; quantile regression at five quantiles with inter-quantile Wald tests whose covariance is obtained by a paired cluster bootstrap; and gradient-boosted ensembles (XGBoost, LightGBM) with SHAP-based interpretation (Lundberg and Lee, 2017), evaluated under a random split, a ten-state holdout, grouped five-fold cross-validation by state and leave-one-state-out validation.

The core methodological contribution is a systematic examination of *spatial leakage* in the evaluation of hedonic models. Housing data are spatially dependent: nearby properties share unobserved local price determinants (Dubin, 1998; Basu and Thibodeau, 1998). Random train–test splits therefore let information flow from training to test observations through spatial proximity and

inflate out-of-sample performance, a phenomenon well documented in ecology and geostatistics (Roberts et al., 2017; Ploton et al., 2020; Meyer and Pebesma, 2021, 2022) but rarely confronted in housing economics. We quantify its consequences and, going beyond the previous version of this paper, identify its mechanics. Under random validation XGBoost reaches $R^2 = 0.833$ against 0.630 for OLS. Under the ten-state holdout the fully grown model reaches 0.425, the median of 51 leave-one-state-out models is 0.653, and—a correction of the previous version, which reported a catastrophic OLS failure caused by corrupt records—the semi-log regression is as accurate out of state (0.539) as the best tree ensemble. The gap is not a matter of training-set size (a size-matched random split loses 0.4 points) and is not located in within-market structure (within-state R^2 in held-out states is 0.62 for every specification). It is located in the prediction of market price *levels*, and three features of the data and the method produce it: region dummies, which partition training states' levels; "neighbourhood" variables that are state-median imputations for 93% of listings and therefore encode the state; and boosting capacity, since held-out-state accuracy peaks after 75 rounds and declines thereafter while within-training accuracy keeps rising, so that standard early stopping selects an over-fitted model.

Our contribution is methodological and empirical, not causal. Following the identification literature (Bartik, 1987; Epple, 1987; Ekeland et al., 2004; Kuminoff et al., 2013; Bishop et al., 2020), we interpret all estimates as conditional associations in *listing* prices rather than structural willingness-to-pay parameters, and we engage the listing-price literature (Horowitz, 1992; Genesove and Mayer, 2001; Han and Strange, 2016) when drawing that distinction. Within that discipline the paper makes five contributions.

First, geographic granularity and inference in OLS. Progressively finer spatial fixed effects raise out-of-sample R^2 from 0.630 to 0.725, in line with the specification advice of Kuminoff et al. (2010); residual Moran's I of 0.27 shows that regional controls leave most of the spatial structure unabsorbed; and cluster-robust standard errors by ZIP3 and by state are, on median, 8 and 12 times the HC3 errors previously reported, reducing the number of significant coefficients from 57 to 33 of 63. Second, one market or fifty-one. Estimating the hedonic equation separately by state shows living-area elasticities from 0.06 to 1.15 and sign reversals in about twenty states for four attributes; a fully interacted model gains 6.1 points of out-of-sample R^2 over state fixed effects, giving empirical content to Rosen's single-market condition and the submarket literature (Bourassa et al., 2003; Goodman and Thibodeau, 2003). Third, formal distributional heterogeneity. Inter-quantile tests with a paired ZIP3-cluster bootstrap reject equality of the 10th- and 90th-percentile coefficients for 10 of 13 key attributes, against 11 under the independence-based test of the previous version; the pattern—garage and living-area gradients concentrated at the bottom, pool and lot-size gradients at the top—is stable across ten subsamples. Fourth, the anatomy of the machine-learning advantage, summarised above and developed in Sections 5.6–5.8. Fifth, cross-model stability of SHAP rankings ($\rho = 0.89$–0.99

across XGBoost, LightGBM and random forests), together with a concrete demonstration of why stability is not structure: the third most important feature is, for 93% of listings, a state identifier.

A sixth result is a data finding with substantive consequences. Documenting the share of imputed values and the rule actually applied for every regressor revealed that the published lot-size gradient (0.018, "tripling" when imputed observations were dropped) was produced by a coding error that set $\ln(\text{lot})$ to zero for 16.5% of listings; the corrected elasticity is 0.095. We report this openly, together with the other corrections, in the belief that an audit trail is part of the contribution.

For practitioners the message is direct: in AVM deployment contexts that require generalisation to unfamiliar markets, validation design is the difference between a model that appears excellent and one that transfers; geographic and imputed features that help in-sample can be counterproductive out of market; and model capacity should be tuned on spatially blocked evaluation sets. For researchers, the results argue that random-split accuracy comparisons between machine-learning and hedonic models overstate the machine-learning advantage whenever the deployment question involves new locations, and that conventional robust standard errors on national scraped data overstate the precision of hedonic gradients.

The remainder of the paper is organised as follows. Section 2 situates the study in the hedonic, quantile, machine-learning and spatial-validation literatures. Section 3 describes the data, the listing-price caveat, and the imputation and coding rules that condition the results. Section 4 presents the econometric and machine-learning methods, the inference, and the validation designs. Section 5 reports the results across the three model families. Section 6 presents imputation, coding, winsorisation and subsample-stability checks. Section 7 interprets the findings against the literature, Section 8 states the limitations, and Section 9 concludes.

2. Literature Review

This paper sits at the intersection of four literatures: the hedonic pricing tradition in housing economics, the econometrics of distributional heterogeneity, the machine-learning strand of property valuation, and the methodological literature on validating predictive models under spatial dependence. We review each in turn, emphasising how it bears on the design choices of this study and where the gap addressed by the paper lies.

2.1. Hedonic Pricing: Origins, Theory and the Interpretation of Implicit Prices

The practice of regressing prices on product characteristics predates its theoretical justification by several decades. Waugh (1928) priced quality attributes of vegetables, Court (1939) constructed hedonic price indexes for automobiles, and Griliches (1961) revived the method for quality-adjusted price measurement. In housing, Ridker and Henning (1967) provided the first regression-based

estimates of how a local disamenity, air pollution, is capitalized into residential property values, inaugurating the property-value approach to valuing local public goods. The theoretical foundation arrived with Lancaster (1966), who recast goods as bundles of characteristics, and Rosen (1974), who showed that in a competitive market the equilibrium price schedule $P(\mathbf{z})$ traces out a double envelope of bid and offer functions, so that the gradient $\partial P/\partial z_k$ equals the marginal implicit price of attribute z_k. Sirmans et al. (2005) synthesise 125 empirical applications, documenting robust positive gradients for living area, bathrooms and garages and negative gradients for property age, the qualitative pattern our conditional-mean estimates reproduce at national scale; Malpezzi (2003) and Palmquist (2005) survey the method's uses and pitfalls in housing and environmental valuation.

A central and often under-appreciated distinction is that first-stage hedonic gradients are equilibrium price phenomena, not preference parameters. Recovering willingness-to-pay functions requires a second stage whose identification demands exclusion restrictions or multiple markets (Bartik, 1987; Epple, 1987; Ekeland et al., 2004). Modern syntheses formalise when hedonic estimates can be trusted: Kuminoff et al. (2013) survey the equilibrium-sorting framework that now underpins structural interpretation of housing-market regressions, and Bishop et al. (2020) distil best practices for credible hedonic estimation, emphasising fine spatial controls, robustness to specification and transparent treatment of data limitations. Particularly relevant to our fixed-effects progression, Kuminoff et al. (2010) show in large-scale simulations that specifications with rich spatial fixed effects dramatically outperform sparse cross-sectional specifications in recovering marginal willingness to pay when omitted spatial variables correlate with amenities. Our region → state → ZIP3 exercise provides complementary, purely predictive evidence on the same point.

Rosen's framework also has a spatial scope condition that a national dataset strains: a single price schedule exists only within a market in which all participants face the same schedule. The housing literature has long recognised that metropolitan areas, let alone countries, consist of segmented submarkets with distinct hedonic functions (Goodman and Thibodeau, 1998). Bourassa et al. (2003) and Goodman and Thibodeau (2003) show that estimating separate hedonic equations by submarket improves prediction accuracy over a pooled equation with location dummies, and Bourassa et al. (2007) find that submarket-specific equations and spatial methods are complements in house-price prediction. A pooled national regression such as ours therefore estimates a listing-weighted average of local price functions. Rather than treating that average as the object of interest, we measure how much the pooled coefficients hide by estimating the hedonic equation separately for each of the 51 jurisdictions and by comparing the predictive accuracy of pooled, fixed-effects and fully interacted specifications (Section 5.3). Throughout, we follow the identification literature's counsel and interpret coefficients as conditional associations—listing-price capitalization gradients—rather than structural demand parameters.

2.2. Functional Form, Dummy Variables and Inference

The semi-log specification became the workhorse of applied hedonics after Cropper et al. (1988) showed in Monte Carlo experiments that simple forms outperform flexible ones when some attributes are omitted or measured with error, precisely the situation of scraped listing data; Hill (2013) surveys the subsequent evolution of hedonic specification practice. Earlier work had already cautioned against atheoretic flexibility: Halvorsen and Pollakowski (1981) demonstrated the sensitivity of Box–Cox hedonic estimates, and Anglin and Gençay (1996) documented gains from semiparametric estimation of the price function. Two technical points from this literature are applied directly below. The proportional effect of a dummy variable in a semi-log equation is $\exp(\beta) - 1$, not β (Halvorsen and Palmquist, 1980), with the small-sample correction of Kennedy (1981); and inference on regressors that vary at an aggregate level, or on spatially correlated data, requires standard errors that allow for within-group dependence (Moulton, 1990; Cameron and Miller, 2015; Conley, 1999), a point whose quantitative importance for national hedonic regressions we document in Section 5.1.5. The functional-form literature frames one of our central questions: how much of the predictive advantage of gradient-boosted trees reflects genuine non-linearity that the semi-log form misses, and how much reflects something else—in our case, the memorisation of local price levels.

2.3. Spatial Dependence in Housing Markets

Housing data are spatial data. Can (1992) showed that ignoring spatial autocorrelation biases hedonic coefficient estimates; Dubin (1998) explains why residential prices are spatially correlated (shared unobserved neighbourhood attributes, market-mediated spillovers); and Basu and Thibodeau (1998) document strong spatial autocorrelation in Dallas transaction prices even after conditioning on rich attribute sets. The formal apparatus of spatial econometrics—spatial lag, spatial error and spatial Durbin models—is developed in Anselin (1988) and LeSage and Pace (2009), with Anselin (2010) providing a retrospective. We deliberately stop short of estimating spatial econometric models: our contribution is diagnostic. Using Moran's I (Moran, 1950) on residuals, we quantify the spatial dependence that remains after 62 attributes and regional controls, and we compare it across estimators. Pace and Hayunga (2020) pursue a complementary strategy, using trees and forests to examine what information hedonic and spatial model residuals still contain; their finding that machine-learning methods extract signal from residual spatial structure anticipates our result that tree ensembles absorb most of the residual spatial autocorrelation of OLS without any coordinate information.

2.4. Quantile Regression and Distributional Heterogeneity

Quantile regression (Koenker and Bassett, 1978; Koenker and Hallock, 2001; Koenker, 2005) replaces the conditional mean with a family of conditional quantiles, allowing implicit prices to vary across the price distribution. Zietz et al. (2008) introduced the approach to housing, finding that square footage and other attributes are priced differently at different quantiles; Mak et al. (2010) document quantile-varying gradients in Hong Kong; and Liao and Wang (2012) combine quantile regression with spatial methods. McMillen (2008) shows that changes in the *distribution* of house prices over time are driven largely by changes in coefficients rather than in characteristics, evidence that quantile-specific pricing is economically meaningful rather than a statistical curiosity. Waltl (2019) provides a comprehensive quantile analysis of the Sydney market, documenting systematic variation across both price segments and locations. Our contribution to this strand is scale and formality: we estimate five quantiles on a national sample, verify coefficient stability across ten independent subsamples, and test tail differences with inter-quantile Wald statistics whose covariance is obtained by a paired cluster bootstrap, a step that earlier housing applications, which typically compared quantile paths graphically, did not take.

2.5. Listing Prices, Search and Seller Behaviour

Because our dependent variable is an asking price, the microstructure of listing behaviour matters for interpretation. Horowitz (1992) models the list price as a commitment device in seller search; Knight (2002) shows that overpricing lengthens time on market and reduces eventual sale prices; Genesove and Mayer (2001) demonstrate that loss aversion leads sellers facing nominal losses to set systematically higher asking prices; and Han and Strange (2016) show that the asking price directs buyer search, so that its information content varies across market segments. Two implications follow. First, listing-price gradients need not equal transaction-price gradients, and the wedge is likely correlated with attributes. Second, the wedge affects all three modelling frameworks identically, so comparisons *across* frameworks, our primary object of interest, are unaffected even where levels must be interpreted cautiously. Section 3.2 uses Zillow's own valuation to bound the size of the wedge in our sample.

2.6. Capitalization of Local Public Goods and Amenities

Several regressors proxy local public goods, for which a mature capitalization literature exists. Oates (1969) initiated the study of property-tax and public-spending capitalization; Black (1999) used school-attendance boundaries to isolate the value of school quality, finding that parents pay 2.5% more for a 5% increase in test scores, about half the naive hedonic estimate; and Bayer et al. (2007) embed boundary discontinuities in an equilibrium sorting model, showing that naive cross-sectional

estimates confound school quality with neighbour characteristics. This literature disciplines our reading of the school-rating and tax-rate coefficients, which in our data are, in addition, imputed at the state level for most listings. Pivo and Fisher (2011) document a walkability premium using Walk Score, the measure we employ, and Duncan et al. (2011) show that the correspondence between Walk Score and underlying walkability constructs varies across metropolitan areas. The climate capitalization literature (Baldauf et al., 2020; Bernstein et al., 2019; Murfin and Spiegel, 2020) identifies a class of price-relevant risk variables that are entirely missing from our data.

2.7. Machine Learning in Property Valuation

The econometrics profession has converged on a division of labour in which machine learning excels at prediction ($\hat{y}$) problems while classical methods target parameter ($\hat{\beta}$) problems (Mullainathan and Spiess, 2017; Varian, 2014; Athey and Imbens, 2019). In real estate this maps onto automated valuation models. Bourassa et al. (2010) compare methods for exploiting spatial dependence in prediction; Antipov and Pokryshevskaya (2012) introduced random forests to mass appraisal; Park and Bae (2015) document early machine-learning gains in county-level data; Kok et al. (2017) describe the shift from manual appraisal to big-data valuation; and a growing body of applied comparisons finds tree ensembles ahead of linear hedonic models on random holdouts (Pérez-Rave et al., 2019; Hu et al., 2019; Ho et al., 2021). Mayer et al. (2019) show that the advantage of gradient boosting over hedonic regression persists when models are re-estimated over time, and Bogin and Shui (2020) document that AVM accuracy deteriorates in rural areas where comparable data are sparse, an early indication that accuracy depends on how well the deployment population is represented in training. Steurer et al. (2021) catalogue the metrics appropriate for evaluating AVM performance, emphasising that headline R^2 figures conceal economically relevant tail behaviour. The tree-ensemble methods we deploy—random forests (Breiman, 2001), gradient boosting (Friedman, 2001) and their implementations XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017)—dominate tabular prediction tasks of this kind. Against this backdrop, our contribution is to ask not *whether* boosting beats OLS under random validation (it does, by 20 percentage points of R^2) but *what that gap is made of*, and whether it survives when the model must price homes in markets it has not seen.

2.8. Explainability: SHAP and Its Limits

SHAP values (Lundberg and Lee, 2017), rooted in the cooperative-game solution concept of Shapley (1953) and computable efficiently for trees via TreeSHAP (Lundberg et al., 2020), have become the standard tool for interpreting ensemble predictions, and Lorenz et al. (2023) demonstrate their use for real estate market analysis. The explainability literature itself urges caution: local

surrogate explanations can be unstable (Ribeiro et al., 2016), and Rudin (2019) argues that post-hoc explanations of black-box models should not be conflated with intrinsically interpretable modelling in high-stakes settings. In the hedonic context the danger is specific: SHAP values are prediction decompositions, sensitive to feature correlation, and do not satisfy the equilibrium conditions under which Rosen's gradient equals a marginal implicit price (Rosen, 1974; Bishop et al., 2020). We use SHAP to describe the predictive structure of the fitted model, probe the robustness of that description across three tree ensembles, and show with the school-rating variable how a feature that is in fact a state identifier can rank third in importance.

2.9. Validation Under Spatial Dependence

A methodological literature developed largely in ecology and geostatistics warns that random cross-validation overstates predictive skill whenever observations are spatially dependent, because information leaks from training to test folds through spatial proximity. Roberts et al. (2017) systematise blocking strategies for structured data; Valavi et al. (2019) provide the standard software implementation of spatial blocking; Schratz et al. (2019) show that both performance assessment and hyperparameter tuning are distorted when the resampling ignores spatial structure; Ploton et al. (2020) demonstrate that large-scale ecological mapping models with excellent random-CV scores lose most of their skill under spatial validation; and Meyer and Pebesma (2021, 2022) formalise the "area of applicability" of spatial prediction models and show that global machine-learning maps are routinely assessed on data that do not represent the prediction space. The lesson is not uncontested: Wadoux et al. (2021) show that when the goal is accuracy over a sampled region—an interpolation problem—spatial cross-validation can be pessimistically biased and design-based random validation is appropriate. This debate sharpens our design: predicting prices in unobserved states is an extrapolation task, for which held-out-region validation is the relevant benchmark, while the random split answers the interpolation question. Reporting both, and showing that models and feature sets rank differently under each, is what the debate prescribes. We are not aware of an earlier application of leave-region-out validation to a national hedonic housing model, nor of one that separates the level and within-market components of out-of-region error.

2.10. Research Gap

Three gaps emerge. First, the hedonic and AVM literatures rarely meet: machine-learning papers report accuracy without engaging identification and interpretation, while econometric papers report coefficients without assessing predictive generalisation, and few of either kind report inference that allows for spatial dependence. Second, the spatial-validation insights of ecology have barely penetrated housing economics, despite housing being a canonically spatial asset; the interaction

between geographic *features*, boosting *capacity* and validation *design* is unexplored at national scale. Third, the literature offers little guidance on how explanation tools such as SHAP behave across model families in housing applications, or on how they respond to the imputation choices that scraped data force on the analyst. This paper addresses all three, at the scale of 788,842 listings spanning every U.S. state, while maintaining the interpretive discipline the identification literature demands.

3. Data and Variable Construction

3.1. Data Source and Coverage

The data are an extract of active for-sale listings on Zillow, the largest online real estate marketplace in the United States, collected in 2025–2026. The raw extract contains 839,313 residential properties described by 116 variables: the asking price, structural characteristics, geocoordinates, neighbourhood accessibility scores, school information, tax records, and listing-status tags. Every observation carries the status `FOR_SALE`; 85.4% are single-family homes and 14.6% condominiums; 89.2% are listed by an agent, 6.5% are new-construction listings or plans, and 1.7% are for-sale-by-owner listings. The median listing had been on Zillow for 72 days at the time of extraction (interquartile range 35–134 days).

The sample is a snapshot of the supply side of the market, not of the housing stock. Four sources of non-representativeness deserve emphasis. First, Zillow coverage varies across regions and depends on the completeness of local multiple-listing-service feeds. Second, active listings over-represent segments with long marketing times and under-represent segments that clear quickly. Third, off-market properties, unlisted for-sale-by-owner homes and recently transacted properties are absent. Fourth, Southern states account for 56.6% of the sample against roughly 38% of U.S. housing units, reflecting the high listing volumes of Sun Belt markets. We therefore do not claim national representativeness and interpret the estimates as parameters of the population of active listings in the extraction window.

3.2. Listing Prices versus Transaction Prices

The dependent variable is the asking price, not a transaction price. Asking prices are strategic objects: they serve as commitment devices in seller search (Horowitz, 1992), direct buyer attention (Han and Strange, 2016), respond to loss aversion among sellers facing nominal losses (Genesove and Mayer, 2001), and systematic overpricing lengthens time on market and lowers the eventual sale price (Knight, 2002). The wedge between asking and transaction prices therefore varies with market conditions, property type and price tier, and it may be correlated with the attributes whose gradients

we estimate. Throughout the paper the estimated coefficients are read as *listing-price capitalization gradients*: conditional associations between attributes and asking prices, not transaction-price implicit prices in the sense of Rosen (1974).

The extract does not contain the eventual sale price of the listed homes, so the wedge cannot be measured directly. It does contain Zillow's own automated valuation, the Zestimate, for 55.0% of the sample. Table 1 reports the distribution of the ratio of the asking price to the Zestimate. The median listing is priced 2.0% above Zillow's valuation, 91.4% of listings are priced above it, and the interquartile range of the ratio is narrow (1.008–1.037). The premium is largest in the South (median 1.023) and smallest in the West and Northeast (1.015–1.016). Because the Zestimate is itself a model fitted partly to listing information, the ratio is a lower bound on the information content of the wedge rather than a measure of it; it does, however, indicate that asking prices in this sample sit systematically but modestly above an independent valuation benchmark, and that the dispersion of the premium is small relative to the dispersion of prices.

Table 1. Ratio of Asking Price to Zestimate, by Census Region

Quantile	All	Northeast	Midwest	South	West
p_{10}	1.001	0.988	0.999	1.003	0.999
p_{25}	1.008	1.002	1.007	1.011	1.005
p_{50}	1.020	1.016	1.018	1.023	1.015
p_{75}	1.037	1.034	1.033	1.041	1.030
p_{90}	1.057	1.055	1.051	1.061	1.047
Share above 1	91.4%	78.2%	88.1%	95.1%	88.0%
N	433,738	32,969	58,622	251,274	90,873

Notes: Listings with a non-missing Zestimate (55.0% of the analytical sample; ratios outside $[0.2, 5]$ excluded). The Zestimate is Zillow's automated valuation at the time of extraction. Generated by `scripts/06_revision_inference.py`.

3.3. Sample Construction

Table 2 documents the sample construction. Five sequential filters remove 6.0% of the raw records, leaving 788,842 listings in 50 states and the District of Columbia, 26,477 five-digit ZIP codes and 885 three-digit ZIP prefixes (ZIP3). The price window of \$10,000–\$10,000,000 excludes land-only and placeholder listings at the bottom and a small number of trophy properties at the top.

Table 2. Sample Construction

Step	N remaining	Removed	Criterion
Raw Zillow records	839,313	—	—
Valid price	817,473	21,840	$\$10{,}000 < P < \$10{,}000{,}000$
Valid living area	797,380	20,093	$200 <$ sqft $< 20{,}000$
Valid bedrooms/bathrooms	790,799	6,581	$1 \leq$ bed ≤ 10, $1 \leq$ bath ≤ 10
Valid geocoordinates	789,199	1,600	Non-missing latitude/longitude
U.S. states and DC only	788,842	357	Exclude PR, VI
Final analytical sample	**788,842**	**50,471**	**(6.0% removed)**

Notes: Filters applied sequentially. The filters act on price, living area, room counts and location only; secondary attributes are not screened at this stage (see Section 3.6).

3.4. Variable Description

The specification includes 62 regressors organised in six blocks.

Structural attributes (8). The natural logarithm of living area (ln sqft), bedrooms, bathrooms, property age (2026 − year built) and its square, stories, the bathroom-to-bedroom ratio and living area per bedroom. Living area enters in logarithms to accommodate the concavity of the size–price relationship.

Lot characteristics (1). The natural logarithm of lot size (ln lot); its coding is discussed in Section 3.5.

Amenity and quality indicators (11). Binary indicators for swimming pool (39.8% prevalence), spa (6.1%), basement (24.3%), fireplace (39.6%), garage (67.2%), waterfront location (8.9%), central air conditioning (68.1%), forced-air heating (25.8%) and hardwood flooring (26.7%); the number of parking spaces; and a composite luxury score (0–7) counting premium features mentioned in the listing.

Neighbourhood and location (7). Walk Score (0–100), Bike Score, Transit Score, the average GreatSchools rating of nearby schools (1–10), the number of nearby schools, the distance to the nearest school, and the local effective property-tax rate. Walk Score is the accessibility measure for which a capitalization premium has been documented (Pivo and Fisher, 2011); school ratings and tax rates proxy the local public goods whose capitalization is the subject of a large identification literature (Oates, 1969; Black, 1999; Bayer et al., 2007). As Section 3.5 shows, the three school variables are observed for only 7% of listings, which fundamentally changes how they must be read.

Market status (5). Condominium indicator (14.6%), homeowners'-association membership (37.1%), log annualised HOA fees ($\ln(1 + \text{HOA}_{\text{annual}})$), new construction (5.9%) and foreclosure

status (0.5%).

Interaction terms (6) and categorical controls (24). Six interactions (ln sqft × age, pool × South, waterfront × ln sqft, basement × North, condominium × Walk Score, age × luxury score) and dummy sets for roof type (6), construction material (8), foundation type (7) and Census region (3, Midwest reference). Each categorical set includes an "Unknown" category that absorbs listings for which the attribute is not reported.

3.5. Missing Data, Imputation and Coding

Scraped listing data are incomplete in ways that are rarely random. Table 3 reports, for every regressor input, the share of the analytical sample for which the raw field is missing and the rule that the sample-construction stage applied. Three facts emerge that earlier versions of this paper did not report and that condition the interpretation of several results.

Table 3. Missing Raw Inputs and Imputation Rules (Analytical Sample, N = 788,842)

Regressor input (raw field)	Missing (%)	Rule applied
Average school rating	93.0	state median
School count	93.0	state median (= 3 in every state)
Nearest-school distance	93.0	state median
Transit Score	70.2	set to 0
HOA fee (ln HOA)	62.7	set to 0
Garage indicator	32.8	set to False
Roof, construction, foundation type	31.2	'Unknown' category
Flooring (hardwood dummy)	29.8	dummy set to False
Stories	25.6	set to 1
Parking spaces	24.3	set to 0
Year built (age, age^2, interactions)	18.9	state median year built
Lot size (ln lot)	15.7	state median in lot_size_sqft; ln(lot) coded 0
Cooling type (central-air dummy)	11.4	dummy set to False
Heating type (forced-air dummy)	8.3	dummy set to False
Bike Score	6.0	state median
Walk Score	3.1	state median

Notes: Missingness measured in the raw extract for the 788,842 retained listings. "State median" denotes imputation by the median of observed values within the listing's state. Climate-risk factors (flood, fire, heat, wind, air) are missing for every record and excluded. Generated by `scripts/06_revision_inference.py`.

First, the three school variables are missing for 93.0% of listings and were imputed with state medians. For 93% of the sample, "average school rating" therefore takes one of 51 values and

identifies the state rather than the quality of nearby schools; the imputed school count is three in every state. Any coefficient or importance measure attached to these variables must be read as a state-level location effect, a point we return to when interpreting the negative school-rating coefficient (Section 5.1.3) and the prominence of school rating in the SHAP rankings (Section 5.9). Second, year built is missing for 18.9% of listings and replaced by the state median, so that the age distribution contains a spike of imputed values at the state-specific median age. Third, lot size is missing for 15.7% of listings and recorded as zero for a further 0.8%; although the lot-size field itself was imputed with state medians, the regressor actually entering the models, $\ln(\text{lot})$, was computed before imputation and takes the value 0 for all 130,271 such listings (16.5% of the sample). The published lot-size results therefore reflect a de facto missing-value code located eight log units below the mass of the data; Section 6.2 re-estimates the model with a correct coding and shows that the baseline lot-size gradient was understated by a factor of five.

The remaining rules are conventional but not innocuous. Transit Score is set to zero when missing (70.2%), on the assumption that properties without a score lie outside transit-served areas; the garage indicator is set to *false* when the field is missing (32.8%), so that it partly measures the completeness of the listing; and the central-air, forced-air and hardwood indicators are derived from free-text cooling, heating and flooring fields that are missing for 8–30% of listings. Setting missing indicators to zero is a special case of the missing-indicator method whose bias properties are analysed by Jones (1996): the coefficient on such a variable mixes the attribute effect with the effect of reporting the attribute. We treat these coefficients as descriptive and use missingness indicators as a robustness device in the machine-learning section (Section 5.8).

3.6. Data Quality

The sample filters screen price, size, room counts and location, but not secondary attributes. Table 4 counts the records whose secondary attributes are implausible. The number of parking spaces exceeds 20 for 1,133 listings and 1,000 for 67 (the maximum is 1.1×10^{11}); stories exceed 10 for 1,607 listings; 422 lots exceed one square mile; 121 listings report a year built after 2026 (age set to zero) and 1,838 a year built before 1826 (age capped at 200). These records are few—4,970 listings, 0.6% of the sample, fail at least one of the four screens used below—but they matter for the linear models: in the unstandardised OLS the parking-space coefficient is 1.9×10^{-12} with a standardised standard error of 0.099, i.e. the variable is uninformative because its variance is dominated by a handful of corrupt entries. For the geographic-holdout comparison of Section 5.6 these records are decisive: the catastrophic out-of-state failure of OLS reported in the previous version of this paper ($R^2 \approx -2 \times 10^6$) is entirely attributable to linear extrapolation on corrupt parking and story counts in the held-out states, not to the regional dummies. Tree ensembles are unaffected because they split on ranks. Bike Score exceeds its nominal maximum of 100 for 38 listings (maximum 248) and is

retained; results are robust to capping it at 100 (Section 6.3).

Table 4. Implausible Attribute Values and Coding Artefacts in the Analytical Sample

Check	Listings	Share (%)
parking_spaces > 20	1,133	0.14
parking_spaces > 1,000	67	0.01
stories > 10	1,607	0.20
lot_size > 1 sq. mile (640 acres)	422	0.05
lot_size > 100 acres	2,568	0.33
lot_size recorded as 0 (raw)	6,115	0.78
ln(lot) coded 0 (missing or zero lot)	130,271	16.51
year_built > 2026 (age set to 0)	121	0.02
year_built < 1826 (age capped at 200)	1,838	0.23
age == 0	18,845	2.39
bike_score > 100	38	0.00
school rating imputed (state median)	733,929	93.04

Notes: Counts in the analytical sample (N = 788,842). The four screens used to define the trimmed samples of Sections 5.6 and 5.3 are parking spaces > 20, stories > 10, lot size > 640 acres and age = 200 (4,970 listings in total). Generated by `scripts/06_revision_inference.py`.

3.7. Descriptive Statistics

Table 5 presents summary statistics. The median listing price is $399,900 and the mean $621,737, a right skew that the logarithmic transformation largely removes (Figure 1). The median home has 1,820 square feet, three bedrooms and two bathrooms, and is 36 years old. Walk and Transit Scores are low on average (medians of 16 and 0), consistent with a listing stock dominated by suburban and exurban single-family homes.

Table 5. Descriptive Statistics (N = 788,842)

Variable	Mean	Std. Dev.	Min	Median	Max
Continuous variables					
Listing price ($)	621,737	817,965	10,300	399,900	9,999,999
Living area (sqft)	2,093	1,181	208	1,820	19,600
Bedrooms	3.25	1.09	1	3	10
Bathrooms	2.52	1.16	1	2	10
Property age (years)	42.2	31.5	0	36	200
Walk Score (0–100)	25.6	26.5	0	16	100
Bike Score	34.9	18.1	1	31	248
Transit Score (0–100)	8.9	18.3	0	0	100
Avg. school rating (1–10)	6.02	0.90	1	6	10
Property tax rate (%)	1.10	0.53	0	1.0	4.0
Luxury score (0–7)	1.36	1.29	0	1	7
Binary variables (prevalence)					
Swimming pool	39.8%				
Garage	67.2%				
Fireplace	39.6%				
Basement	24.3%				
Waterfront	8.9%				
HOA	37.1%				
Central air	68.1%				
Hardwood floors	26.7%				
Condominium	14.6%				
New construction	5.9%				
Foreclosure	0.5%				
Regional distribution					
South	56.6%				
West	18.9%				
Midwest	16.1%				
Northeast	8.4%				

Notes: Listing prices, not transaction prices. Property age computed as 2026 – year built after imputation; imputed values included. The school rating is a state median for 93% of listings (Table 3).

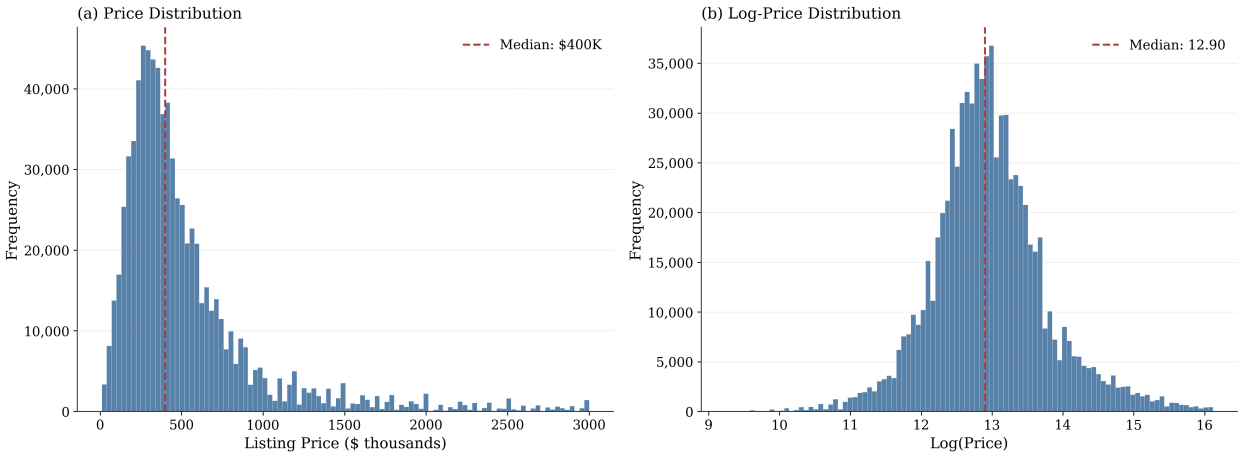


Figure 1. Distribution of listing prices: (a) levels; (b) natural logarithm. The logarithmic transformation reduces right skewness and approximates normality.

4. Econometric and Machine-Learning Methodology

4.1. Identification, Prediction and Interpretation

None of the models estimated in this paper identifies causal willingness-to-pay parameters. In Rosen's framework the hedonic price function $P(\mathbf{z})$ is an equilibrium object whose gradient equals the marginal implicit price of each attribute only within a single market in which buyers and sellers face the same price schedule (Rosen, 1974); recovering preferences from it requires a second stage whose identification depends on exclusion restrictions or on variation across markets (Bartik, 1987; Epple, 1987; Ekeland et al., 2004). Our data pool listings from fifty-one jurisdictions and thousands of local markets. A pooled hedonic regression on such data estimates an average of market-specific price functions, weighted by each market's share of listings, and its coefficients are conditional associations in asking prices: the average difference in log listing price between properties that differ by one unit in attribute x_k, holding the other included attributes constant. Quantile coefficients are the corresponding quantile-specific associations and SHAP values are model-specific prediction decompositions. Table 6 summarises these distinctions, and Section 5.3 measures directly how much the pooled coefficients hide by estimating the price function state by state.

Table 6. Interpretation of Estimates Across Modelling Frameworks

	OLS coefficient	Quantile coefficient	SHAP value
Object	Conditional-mean association	Conditional-quantile association	Prediction contribution
Interpretation	Semi-elasticity or elasticity (unstandardised); SD effect (standardised)	Quantile-specific semi-elasticity	Model-specific contribution to the prediction
Causal?	No, unless identified	No, unless identified	No
Structural WTP?	Only under single-market and exclusion assumptions	Same	No
Primary use	Inference on average gradients	Distributional heterogeneity	Interpretation of predictions

4.2. Semi-Log Hedonic Specification

The baseline model is the semi-log specification

$$\ln P_i = \alpha + \sum_{k=1}^{K} \beta_k x_{ik} + \sum_{j=1}^{J} \gamma_j d_{ij} + \sum_{m=1}^{M} \delta_m \, x_{im_1} x_{im_2} + \mu_{g(i)} + \varepsilon_i, \tag{1}$$

where P_i is the asking price of listing i, x_{ik} are continuous and binary attributes, d_{ij} are categorical dummies, $x_{im_1}x_{im_2}$ are the six interaction terms and $\mu_{g(i)}$ is a location effect at the level g of Census region (the baseline, three dummies), state (50 effects) or ZIP3 (885 effects). The semi-log form is the workhorse of applied hedonics because it is robust to omitted and mismeasured attributes (Cropper et al., 1988; Halvorsen and Pollakowski, 1981), precisely the situation of scraped listing data.

Coefficient interpretation. We report two versions of Eq. (1). In the standardised version every continuous regressor is scaled to zero mean and unit variance, so that β_k measures the association between a one-standard-deviation increase in x_k and log price and coefficients are comparable across variables with different units. In the unstandardised version the coefficient on a logged regressor is an elasticity and the coefficient on a continuous regressor in natural units is a semi-elasticity, $100\,\beta_k$ per cent per unit. Binary regressors are never standardised. For a dummy variable the proportional effect is not β_k but

$$g_k = \exp(\beta_k) - 1, \qquad \hat{g}_k^{\mathrm{K}} = \exp\!\big(\hat{\beta}_k - \tfrac{1}{2}\widehat{\mathrm{Var}}(\hat{\beta}_k)\big) - 1, \tag{2}$$

where the first expression is the exact transformation of Halvorsen and Palmquist (1980) and the second is the approximately unbiased estimator of Kennedy (1981). With standard errors below 0.04 for every dummy in our sample the Kennedy correction is below 0.1 percentage point, and we report $100(\exp(\hat{\beta}_k) - 1)$ throughout.

Inference. The previous version of this paper reported heteroskedasticity-robust HC3 standard errors (White, 1980; MacKinnon and White, 1985). HC3 treats listings as independent. Residuals of a hedonic regression on spatial data are not independent: listings in the same neighbourhood share unobserved local price determinants, and several regressors (tax rate, the imputed school and lot variables, region) vary only at the state or ZIP level. In that setting the Moulton problem applies (Moulton, 1990): robust standard errors on group-level or spatially correlated regressors can understate sampling variability by an order of magnitude. We therefore also report cluster-robust standard errors (Cameron and Miller, 2015),

$$\widehat{\mathrm{Var}}_{\mathrm{CR}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\Big(\sum_{c=1}^{C}\mathbf{X}_c'\hat{\mathbf{e}}_c\hat{\mathbf{e}}_c'\mathbf{X}_c\Big)(\mathbf{X}'\mathbf{X})^{-1}\cdot\frac{C}{C-1}\,\frac{N-1}{N-K}, \tag{3}$$

where c indexes clusters, $\mathbf{X}_c$ and $\hat{\mathbf{e}}_c$ stack the regressors and OLS residuals of cluster c, and the clusters are either the 885 ZIP3 areas or the 51 states. Eq. (3) is consistent under arbitrary within-cluster correlation and heteroskedasticity as $C \to \infty$; with 51 state clusters the asymptotics are less comfortable, so the state-clustered results are read as a conservative bound. A spatial HAC estimator (Conley, 1999) would allow dependence to decay smoothly with distance across cluster boundaries; we do not pursue it because the cluster estimates already change the inference qualitatively (Section 5.1.5).

4.3. Quantile Regression

Quantile regression (Koenker and Bassett, 1978; Koenker and Hallock, 2001; Koenker, 2005) models the τ-th conditional quantile of log price,

$$Q_\tau(\ln P_i \mid \mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}(\tau), \qquad \tau \in (0,1), \tag{4}$$

with $\hat{\boldsymbol{\beta}}(\tau) = \arg\min_{\mathbf{b}} \sum_i \rho_\tau(\ln P_i - \mathbf{x}_i'\mathbf{b})$ and check function $\rho_\tau(u) = u\,(\tau - \mathbb{1}\{u < 0\})$. We estimate Eq. (4) at $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$ with the full set of 62 standardised regressors. Because each fit of the interior-point algorithm on the full sample is costly, estimation uses a random subsample of 150,000 listings (seed 42) and we verify stability by repeating the median regression on ten independent subsamples (Section 6.4). Goodness of fit is measured by the pseudo-R^2 of Koenker and Machado (1999), $R^1(\tau) = 1 - \hat{V}(\tau)/\tilde{V}(\tau)$, the proportional reduction in the check-function loss relative to a constant-only model.

Inter-quantile tests. To test whether a gradient differs between the lower and upper tails we consider $H_0 : \beta_k(0.10) = \beta_k(0.90)$ with statistic

$$z_k = \frac{\hat{\beta}_k(0.10) - \hat{\beta}_k(0.90)}{\sqrt{\widehat{\mathrm{Var}}[\hat{\beta}_k(0.10)] + \widehat{\mathrm{Var}}[\hat{\beta}_k(0.90)] - 2\,\widehat{\mathrm{Cov}}[\hat{\beta}_k(0.10), \hat{\beta}_k(0.90)]}}. \tag{5}$$

The two quantile estimates are computed on the same sample and are positively correlated; the asymptotic covariance is proportional to $\min(\tau_1, \tau_2) - \tau_1\tau_2 = 0.01$ against variances proportional to $\tau(1-\tau) = 0.09$ (Koenker, 2005, ch. 3). The previous version of the paper set the covariance term to zero, which makes the test conservative under independence of observations but ignores spatial dependence altogether. We now obtain the denominator of Eq. (5) from a paired bootstrap: in each of 200 replications both quantile regressions are re-estimated on a resample of the same 150,000 listings, drawn either observation by observation or by ZIP3 cluster (resampling whole ZIP3 areas with replacement), and the standard deviation of $\hat{\beta}_k(0.10) - \hat{\beta}_k(0.90)$ across replications is the bootstrap standard error of the difference. The cluster version allows for within-ZIP3 dependence of the kind that the clustered OLS standard errors address.

4.4. Gradient-Boosted Ensembles and Benchmarks

XGBoost (Chen and Guestrin, 2016) implements gradient boosting (Friedman, 2001): an additive ensemble of regression trees $f_1, \dots, f_T$ fitted stagewise by minimising

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i^{(t-1)} - f_t(\mathbf{x}_i)\right)^2 + \Omega(f_t), \qquad \Omega(f) = \gamma T_f + \tfrac{1}{2}\lambda\|\mathbf{w}\|^2 + \alpha\|\mathbf{w}\|_1, \tag{6}$$

where T_f is the number of leaves and $\mathbf{w}$ the vector of leaf values. We use up to 1,000 rounds, maximum depth 8, learning rate 0.05, row and column subsampling of 0.8, minimum child weight 5, $\alpha = 0.1$, $\lambda = 1$, and early stopping after 50 rounds without improvement on a 10% evaluation set drawn at random from the training data. LightGBM (Ke et al., 2017) is fitted with matching hyperparameters. We also report random forests (500 trees, depth 20; Breiman, 2001), Ridge ($\alpha = 1$), Lasso ($\alpha = 0.001$) and Elastic Net as linear benchmarks, and OLS with state and ZIP3 fixed effects. Hyperparameters were set at values that are standard for tabular data of this size and were not tuned by cross-validation; the consequences of this choice for the geographic-holdout results are examined in Section 5.6.2.

Evaluation metrics. On the log scale we report R^2, the root mean squared error and the mean absolute error; on the dollar scale, following the recommendations of Steurer et al. (2021) for

automated valuation models, the mean and median absolute percentage error,

$$\text{MdAPE} = \text{median}_i \left| \frac{\hat{P}_i - P_i}{P_i} \right| \times 100, \qquad \hat{P}_i = \exp(\widehat{\ln P_i}), \tag{7}$$

which is robust to the heavy right tail of dollar errors. Differences in test-set R^2 between models are accompanied by bootstrap confidence intervals obtained by resampling test observations, or whole ZIP3 clusters of test observations, 1,000 times, and by a paired test on squared-error differences in the spirit of Diebold and Mariano (1995).

4.5. Validation Designs

Because listing prices are spatially dependent, the validation protocol is a substantive choice rather than a technical one. A random split assesses interpolation within the markets represented in the training data; a spatially blocked split assesses extrapolation to markets that the model has not seen (Roberts et al., 2017; Valavi et al., 2019; Meyer and Pebesma, 2021, 2022). Neither is universally preferable: for interpolation objectives, blocked validation can be pessimistically biased (Wadoux et al., 2021). We therefore report both and interpret each against its own deployment question, using four designs.

Random split. An 80/20 split with seed 42 (631,073 training and 157,769 test listings). Because 99.4% of the five-digit ZIP codes in the test set also appear in the training set, nearly every test listing has training neighbours in its own ZIP code.

Ten-state holdout. California, New York, Texas, Florida, Ohio, Colorado, North Carolina, Washington, Illinois and Georgia—the five largest states by sample size and five further states chosen for geographic diversity—are withheld entirely; models are trained on the remaining forty states and the District of Columbia (350,527 listings) and evaluated on the 438,315 held-out listings. This single split was the only geographic design in the previous version. It confounds extrapolation with a 44% smaller training set and rests on one arbitrary partition, so we add two designs.

Grouped five-fold cross-validation by state. The 51 jurisdictions are assigned to five folds of roughly equal size (`GroupKFold`); each fold is predicted by a model trained on the other four, so that every listing is predicted once by a model that has seen no listing from its state, and the pooled predictions are scored.

Leave-one-state-out. For each of the 51 jurisdictions a model is trained on the other fifty and evaluated on the omitted one. This design yields a distribution of out-of-state performance over 51 markets, each predicted from the largest possible training set, and isolates the market-level bias of the predictions.

Within-state accuracy. Pooled out-of-state R^2 mixes two errors: the error in predicting the price *level* of a market the model has never seen, and the error in predicting *within-market* price

differences. We separate them by also reporting the within-state R^2,

$$R^2_{\text{within}} = 1 - \frac{\sum_i \left[(y_i - \bar{y}_{s(i)}) - (\hat{y}_i - \bar{\hat{y}}_{s(i)})\right]^2}{\sum_i (y_i - \bar{y}_{s(i)})^2}, \tag{8}$$

where $\bar{y}_{s(i)}$ and $\bar{\hat{y}}_{s(i)}$ are the test-set means of actual and predicted log prices in listing i's state. R^2_{within} is the accuracy an analyst would obtain who knew each held-out market's average price level but nothing else about it.

Ablation and feature variants. XGBoost is estimated on six nested feature sets (structural; + lot; + amenities; + neighbourhood; + market status; full model with interactions, categorical controls and region dummies) and on four variants of the full model that remove the region dummies, the three school variables, both, or add six missingness indicators. Every variant is evaluated under both the random and the ten-state design.

4.6. SHAP-Based Interpretation and Cross-Model Stability

SHAP values (Lundberg and Lee, 2017), rooted in the solution concept of Shapley (1953) and computed exactly for tree ensembles by TreeSHAP (Lundberg et al., 2020), decompose each prediction additively,

$$f(\mathbf{x}) = \mathbb{E}[f(\mathbf{X})] + \sum_{k=1}^{K} \phi_k(\mathbf{x}), \tag{9}$$

where $\phi_k(\mathbf{x})$ is the contribution of feature k to the deviation of the prediction from its mean. We compute TreeSHAP values on 10,000 random test listings and summarise global importance by the mean absolute contribution $\overline{|\phi_k|}$. To assess whether the resulting ranking is an artefact of one algorithm we compute $\overline{|\phi_k|}$ for XGBoost, LightGBM and the random forest and report pairwise Spearman rank correlations. SHAP values are prediction-level contributions: they are model-dependent, share contributions among correlated features, decompose predictions rather than the data-generating process, and are not Rosen's gradient $\partial P/\partial z_k$ (Rudin, 2019; Lorenz et al., 2023). We use them to describe which features the fitted ensembles rely on, not to estimate marginal willingness to pay.

4.7. Spatial Autocorrelation Diagnostics

Moran's I (Moran, 1950) on model residuals e_i is

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j w_{ij}(e_i - \bar{e})(e_j - \bar{e})}{\sum_i (e_i - \bar{e})^2}, \tag{10}$$

with a row-standardised k-nearest-neighbour weight matrix ($k = 8$) built from listing coordinates. Under the null of no spatial autocorrelation $\mathbb{E}[I] = -1/(N-1)$; we assess significance by 999 random permutations. Because I on 788,842 listings is computationally prohibitive, it is computed on three independent random subsamples of 5,000 listings, for the residuals of the baseline OLS (full sample) and, in the revision, for the test-set residuals of OLS, LightGBM and XGBoost under the random split, which shows how much residual spatial structure the tree ensembles absorb.

5. Estimation Results

5.1. OLS Baseline and Geographic Fixed Effects

5.1.1. Standardised Baseline Specification

Table 7 presents the OLS results with standardised continuous regressors. The model achieves $R^2 = 0.634$ with a root mean squared error of 0.489 in log-price units. The F-statistic of 18,712.5 rejects the null of jointly zero slopes; with 788,842 observations this is uninformative, and the more relevant question, how precisely individual gradients are estimated once spatial dependence is allowed for, is taken up in Section 5.1.5. The standard errors in Tables 7 and 8 are HC3 standard errors and are reported for comparability with the previous version; they should be read together with Table 10.

Table 7. OLS Hedonic Regression Results (Dependent Variable: ln Listing Price)

Variable	Coefficient	HC3 SE	
Panel A: Structural attributes (standardised)			
ln(Living Area)†	0.3119	0.0035	***
Bedrooms†	−0.0675	0.0023	***
Bathrooms†	0.2359	0.0026	***
Age†	−0.0418	0.0139	**
Age²†	0.0507	0.0019	***
Stories†	0.0004	0.0003	
Bath/Bed Ratio†	−0.0208	0.0021	***
Panel B: Lot and amenities			
ln(Lot Size)†‡	0.0697	0.0010	***
Pool	0.0286	0.0026	***
Spa	0.0327	0.0027	***
Basement	−0.0424	0.0023	***
Garage	0.1094	0.0015	***
Waterfront	−0.1349	0.0325	***
Central Air	0.0693	0.0014	***
Hardwood Floors	0.1246	0.0014	***
Luxury Score†	0.0443	0.0018	***
Panel C: Neighbourhood (standardised)			
Walk Score†	−0.0109	0.0012	***
Bike Score†	0.0717	0.0010	***
Transit Score†	0.0524	0.0008	***
Avg. School Rating†¶	−0.0491	0.0007	***
Property Tax Rate†	−0.0292	0.0008	***

Notes: Table continues on the next page; estimation notes follow the continuation.

Table 7. OLS Hedonic Regression Results (continued)

Variable	Coefficient	HC3 SE	
Panel D: Market status			
Condominium	−0.0370	0.0037	***
HOA Member	−0.1077	0.0024	***
ln(HOA Fees + 1)†	0.0439	0.0014	***
New Construction	0.1035	0.0029	***
Foreclosure	−0.4064	0.0098	***
Panel E: Interactions (standardised)			
ln(sqft) × Age†	−0.1100	0.0140	***
Pool × South	0.0081	0.0012	***
Waterfront × ln(sqft)†	0.1371	0.0092	***
Basement × North	−0.0255	0.0012	***
Condo × Walk Score†	0.0816	0.0013	***
Age × Luxury†	0.0440	0.0015	***
Panel F: Regional dummies (reference: Midwest)			
Northeast	0.4869	0.0030	***
South	0.0220	0.0028	***
West	0.4008	0.0031	***
R^2	0.6343		
Adj. R^2	0.6342		
RMSE	0.4887		
F-statistic	18,712.5		
N	788,842		

Notes: HC3 heteroskedasticity-robust standard errors; cluster-robust standard errors are 2–36 times larger (Table 10) and the stars should be read accordingly. †Standardised (zero mean, unit variance) before estimation; coefficients are effects of a one-standard-deviation increase. ‡Published coding with ln(lot) = 0 for 16.5% of listings (Section 6.2). ¶State median for 93% of listings. Binary variables are not standardised; their percentage effects are $100(e^{\beta} - 1)$. Categorical controls (roof 6, construction 8, foundation 7 dummies) included but not shown (Appendix B). *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ (HC3).

5.1.2. Unstandardised Specification for Economic Interpretation

Table 8 reports key coefficients from the unstandardised specification. The R^2 is identical (0.634); only the scale of the coefficients differs. The elasticity of the asking price with respect to living area is 0.63: a 10% larger home is listed at a 6.3% higher price, holding the other attributes constant. Each additional bathroom is associated with a $100(e^{0.2035} - 1) = 22.6\%$ higher asking price, while each additional bedroom, conditional on total area, is associated with a 6.0% *lower* price, the familiar room-count versus room-size trade-off (Sirmans et al., 2005). The foreclosure discount is $100(e^{-0.4064} - 1) = -33.4\%$; the Kennedy correction of Eq. (2) changes this figure by less than 0.01 percentage point.

Table 8. Unstandardised OLS: Key Coefficients in Natural Units

Variable	Coefficient	HC3 SE	Interpretation	
Constant	7.7316	0.0495		***
ln(Living Area)	0.6281	0.0070	Elasticity 0.63	***
Bedrooms	−0.0620	0.0022	−6.0% per additional bedroom	***
Bathrooms	0.2035	0.0023	+22.6% per additional bathroom	***
Age (years)	−0.0013	0.0004	−0.13% per year	**
Age^2 ($\times 10^{-5}$)	1.2	0.05	Positive quadratic	***
ln(Lot Size)[‡]	0.0182	0.0003	Elasticity 0.02; 0.095 recoded	***
Walk Score (per point)	−0.0004	0.0000	−0.04% per point	***
Bike Score (per point)	0.0040	0.0001	+0.40% per point	***
Transit Score (per point)	0.0029	0.0000	+0.29% per point	***
Avg. School Rating (per point)[¶]	−0.0544	0.0008	−5.3% per rating point	***
Property Tax Rate (per ppt)	−0.0553	0.0015	−5.4% per ppt	***
Luxury Score (per unit)	0.0344	0.0014	+3.5% per unit	***

Notes: Same sample and specification as Table 7 with continuous variables in natural units; $N = 788{,}842$; HC3 standard errors. [‡]Published coding of lot size; the corrected elasticity is reported in Table 29. [¶]State median for 93% of listings. Percentage effects for dummies follow Eq. (2).

5.1.3. Counter-Intuitive Coefficient Signs

Several coefficients have signs that a reader expecting implicit prices would not anticipate. They are instructive precisely because the data documentation of Section 3.5 explains most of them.

Negative school rating (−5.3% per rating point). The school rating is a state median for 93% of listings. The coefficient therefore compares states with higher median school ratings to states with lower ones, conditional on the region dummies and the (also largely state-level) tax rate, and it says nothing about the valuation of nearby schools. The boundary-discontinuity literature, which isolates

school quality from neighbourhood composition, consistently finds positive valuations (Black, 1999; Bayer et al., 2007); our coefficient is not evidence against those findings but an illustration of what a state-level proxy entering a national cross-section can produce. Under state-clustered standard errors the coefficient is not significantly different from zero ($t = -1.9$, Table 10).

Negative waterfront main effect (−0.135, standardised). The waterfront indicator is interacted with standardised ln(sqft), and the interaction is strongly positive (+0.137). The main effect is the waterfront premium at the point where the standardised interaction equals zero, which is not an economically meaningful property. The net waterfront association is positive for larger properties. The coefficient is also imprecise once spatial dependence is allowed for (state-clustered standard error 0.46) and unstable across subsamples (Section 6.1), as expected for an attribute held by 8.9% of listings concentrated in a few coastal and lake markets.

Negative Walk Score (−0.04% per point). Walk Score is highly correlated with Bike Score ($r \approx 0.56$) and Transit Score ($r \approx 0.62$). With three collinear accessibility measures in the regression the premium is split across coefficients and cannot be attributed variable by variable; a walkability premium is well documented when accessibility enters alone (Pivo and Fisher, 2011). The Walk Score coefficient is insignificant under clustered inference ($t = -0.5$).

Negative basement (−0.042) and HOA membership (−0.108). Basements are concentrated in older homes in colder regions and, conditional on age, region and size, the indicator proxies for construction vintage and layout. HOA membership enters alongside the log of HOA fees (+0.044), so the dummy captures the property-type effect of planned communities and condominiums while the fee variable captures service quality.

These signs are common in hedonic regressions with many correlated controls. They do not indicate misspecification, but they do show that attribute-by-attribute economic interpretation of a national pooled regression is limited, a point made more forcefully by the state-by-state estimates of Section 5.3.

5.1.4. Progressive Geographic Fixed Effects

Table 9 documents how explanatory power increases with geographic granularity. Replacing the three region dummies by 50 state fixed effects raises the in-sample R^2 from 0.634 to 0.681 and the out-of-sample R^2 on the random 20% holdout from 0.630 to 0.678; 885 ZIP3 fixed effects raise them further to 0.729 and 0.725. The 9.5-percentage-point out-of-sample gain from geographic granularity alone, with no additional property-level information, is the first indication that unobserved local price levels are the dominant omitted variable in Eq. (1).

Table 9. OLS R^2 Under Progressively Finer Geographic Controls

Specification	Geographic FE	N_{FE}	In-sample R^2	Out-of-sample R^2
Baseline (4 regions)	Census region	3	0.634	0.630
State FE	State	50	0.681	0.678
ZIP3 FE	3-digit ZIP	885	0.729	0.725

Notes: Same 59 non-geographic regressors in all specifications; geographic fixed effects replace the three regional dummies. Out-of-sample R^2 on the random 20% holdout (seed 42). The 885 ZIP3 prefixes include a code for listings with a missing ZIP.

The gap between the ZIP3 model ($R^2 = 0.725$) and XGBoost under random validation ($R^2 = 0.833$) is often read as the contribution of non-linearity. That reading is not warranted here: XGBoost does not observe ZIP codes, and Section 5.8 shows that it recovers location from state-level proxies, so the two models exploit different information and the 10.8-point difference cannot be attributed to functional form alone.

5.1.5. Inference Under Spatial Dependence

Table 10 re-estimates the standardised baseline with standard errors clustered by ZIP3 (885 clusters) and by state (51 clusters), Eq. (3). The consequences are large and systematic. Across the 63 coefficients, ZIP3-clustered standard errors are on median 7.6 times the HC3 values and state-clustered standard errors 11.6 times. The inflation is smallest for attributes that vary within neighbourhoods (foreclosure, 2.3–3.8; living area, 4.6–8.9) and largest for variables that are themselves spatial aggregates or state-level imputations: school rating (21–35), tax rate (20–36), Bike Score (12–27) and the region dummies (13–43). This is the Moulton problem in its textbook form (Moulton, 1990). The number of coefficients significant at the 5% level falls from 57 under HC3 to 35 under ZIP3 clustering and 33 under state clustering. The structural gradients that carry the paper's economic interpretation survive—living area, bathrooms, garage, hardwood floors, central air, luxury score, foreclosure and the Northeast and West differentials all have state-clustered t-statistics above 2.6—but the school-rating, tax-rate, Walk Score, pool, waterfront, condominium and age coefficients do not. The quantile and SHAP results below should be read with the same caution: with a Moran's I of 0.27 in the residuals, the effective sample size for spatially aggregated regressors is in the hundreds, not the hundreds of thousands.

Table 10. OLS Standard Errors Under Independence, ZIP3 Clustering and State Clustering

		Standard error			Ratio to HC3		
Variable	Coef.	HC3	ZIP3	State	ZIP3	State	t (state)
ln(Living Area)	0.3119	0.0035	0.0159	0.0308	4.6	8.9	10.13
Bedrooms	−0.0675	0.0023	0.0131	0.0287	5.6	12.2	−2.35
Bathrooms	0.2359	0.0026	0.0161	0.0251	6.1	9.6	9.40
Property Age	−0.0418	0.0139	0.0648	0.1196	4.7	8.6	−0.35
Property Age^2	0.0507	0.0019	0.0077	0.0188	4.0	9.8	2.70
ln(Lot Size)	0.0697	0.0010	0.0098	0.0144	10.3	15.0	4.84
Pool	0.0286	0.0026	0.0304	0.0544	11.7	21.0	0.52
Garage	0.1094	0.0015	0.0105	0.0160	7.0	10.6	6.83
Waterfront	−0.1349	0.0325	0.2479	0.4567	7.6	14.1	−0.30
Central Air	0.0693	0.0014	0.0147	0.0229	10.6	16.5	3.03
Hardwood Floors	0.1246	0.0014	0.0093	0.0177	6.6	12.6	7.05
Luxury Score	0.0443	0.0018	0.0100	0.0114	5.7	6.5	3.89
Walk Score	−0.0109	0.0012	0.0130	0.0212	11.2	18.3	−0.51
Bike Score	0.0717	0.0010	0.0123	0.0269	12.2	26.8	2.66
Transit Score	0.0524	0.0008	0.0191	0.0127	23.3	15.5	4.13
Avg. School Rating	−0.0491	0.0007	0.0150	0.0254	20.7	35.2	−1.93
Property Tax Rate	−0.0292	0.0008	0.0155	0.0285	19.7	36.3	−1.02
Condo	−0.0370	0.0037	0.0416	0.0608	11.2	16.4	−0.61
HOA	−0.1077	0.0024	0.0227	0.0288	9.5	12.1	−3.74
New Construction	0.1035	0.0029	0.0234	0.0380	8.1	13.1	2.72
Foreclosure	−0.4064	0.0098	0.0227	0.0372	2.3	3.8	−10.91
Region: Northeast	0.4869	0.0030	0.0555	0.1294	18.5	43.1	3.76
Region: South	0.0220	0.0028	0.0361	0.0653	13.1	23.8	0.34
Region: West	0.4008	0.0031	0.0494	0.1061	15.9	34.2	3.78
Coefficients significant at 5% (of 63)		57	35	33			
Median SE ratio across all coefficients					7.6	11.6	

Notes: Standardised baseline specification of Table 7, N = 788,842. Cluster-robust standard errors follow Eq. (3) with 885 ZIP3 clusters or 51 state clusters. Generated by `scripts/06_revision_inference.py`.

5.1.6. OLS Diagnostics

Figure 2 presents diagnostic plots. The residual distribution has excess kurtosis (5.55 against 3.0 under normality) and mild positive skewness (0.18); the Jarque–Bera statistic is 217,078 ($p < 0.001$). The heteroskedastic pattern in the residuals-versus-fitted plot motivates robust standard errors; the clustered standard errors above address the dependence that the plot cannot show.

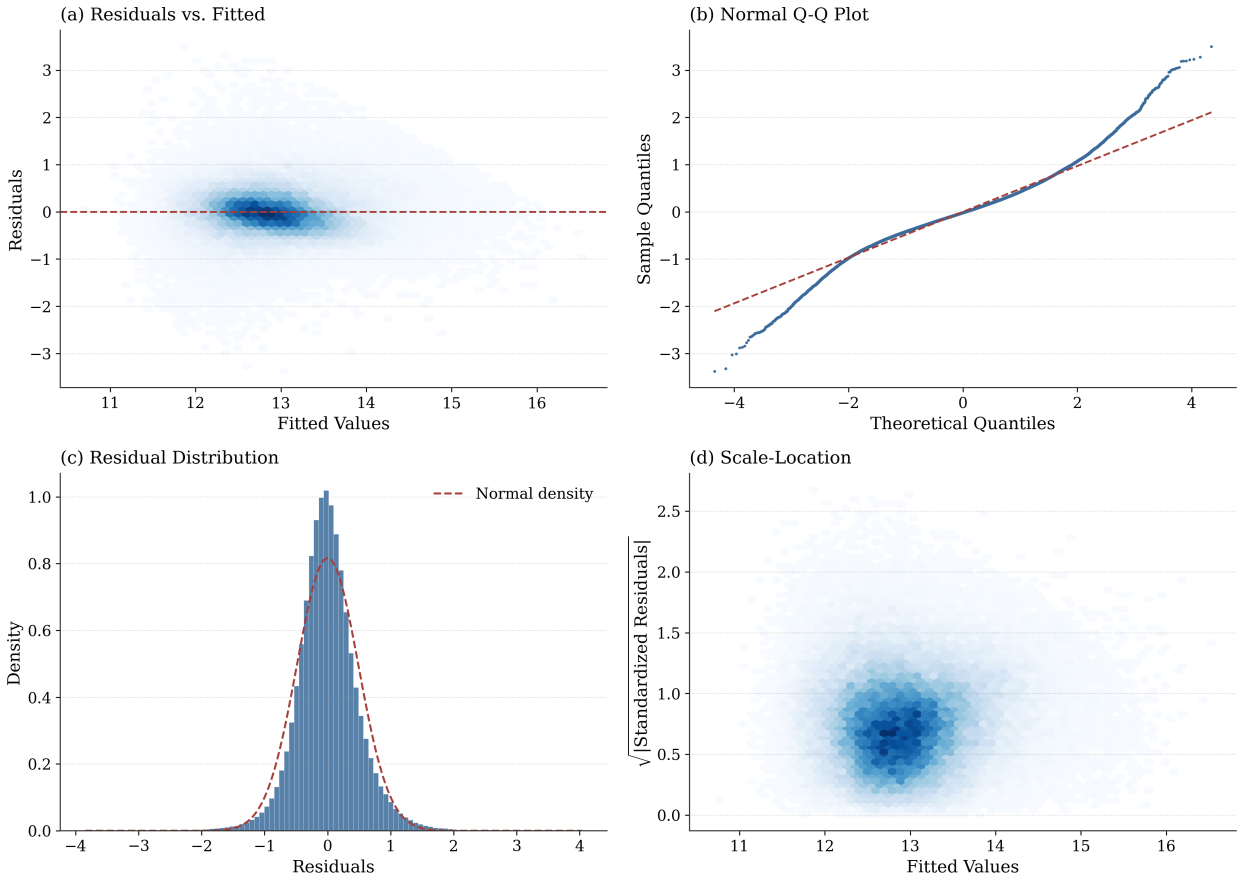


Figure 2. OLS diagnostic plots: (a) residuals versus fitted values; (b) normal Q–Q plot; (c) residual distribution; (d) scale–location.

5.2. Spatial Autocorrelation

Table 11 reports Moran's I on the residuals of the baseline OLS, computed on three independent random subsamples of 5,000 listings with an eight-nearest-neighbour weight matrix. The mean I is 0.2745 (standard deviation 0.0076 across subsamples), and every permutation test rejects the null of no spatial autocorrelation. The stability across subsamples (coefficient of variation 2.8%) indicates a structural feature of the data rather than a sampling artefact: four Census regions and 59 attributes leave most of the local price structure in the error term.

Table 11. Moran's I on Baseline OLS Residuals: Subsample Stability

Subsample	I	z-statistic	p-value	
Subsample 1	0.2742	41.8	< 0.001	***
Subsample 2	0.2839	43.5	< 0.001	***
Subsample 3	0.2652	40.5	< 0.001	***
Mean	0.2745			
Std. Dev.	0.0076			

Notes: Row-standardised k-nearest-neighbour weights, $k = 8$; 5,000 listings per subsample; 999 permutations. Residuals from the baseline OLS fitted on the random training set.

Table 12 extends the diagnostic to the test-set residuals of the three main models under the random split, on the same subsamples. Moran's I falls from 0.263 for OLS to 0.110 for LightGBM and 0.079 for XGBoost. The tree ensembles, which observe no coordinates and no ZIP codes, absorb roughly 70% of the residual spatial autocorrelation of OLS. They can only do so by recovering location from the features available to them, and Section 5.8 identifies which features serve that purpose. The finding mirrors Pace and Hayunga (2020), who show that trees and forests extract exactly the spatial information that hedonic residuals retain.

Table 12. Moran's I on Test-Set Residuals, Random Split: OLS versus Tree Ensembles

Model	Subsample 1	Subsample 2	Subsample 3	Mean	Std. Dev.	z range
OLS	0.2571	0.2764	0.2554	0.2629	0.0095	39.8–42.8
LightGBM	0.1078	0.1147	0.1070	0.1098	0.0034	16.5–17.4
XGBoost	0.0835	0.0787	0.0752	0.0791	0.0034	11.5–12.4

Notes: Same weights and subsample sizes as Table 11, drawn from the 157,769 test listings of the random split; all $p < 0.001$. Generated by `scripts/06_revision_inference.py`.

5.3. One Market or Fifty-One? State-by-State Estimates

Rosen's price function is defined within a market. To measure how much a single national equation hides, we estimate the unstandardised specification separately for each of the 51 jurisdictions (region dummies dropped, lot size recoded as in Section 6.2, 4,970 implausible records removed) and compare the predictive accuracy of pooled, state-fixed-effects and fully interacted specifications on the random holdout. Table 13 summarises the distribution of the state coefficients and Figure 3

displays two of them.

Table 13. State-by-State OLS Coefficients: Distribution Across 51 Jurisdictions (Natural Units)

	Pooled,	Distribution of state coefficients					States with
Variable	state FE	p_{10}	Median	p_{90}	Min	Max	opposite sign
ln(Living Area)	0.625	0.316	0.506	0.720	0.062	1.149	0
Bedrooms	−0.066	−0.073	−0.020	0.035	−0.130	0.057	19
Bathrooms	0.204	0.052	0.149	0.199	0.019	0.285	0
Property Age	−0.002	−0.015	−0.002	0.008	−0.021	0.016	16
ln(Lot Size)	0.087	0.034	0.091	0.141	−0.011	0.165	2
Pool	0.056	−0.067	0.008	0.110	−0.140	0.313	22
Garage	0.095	0.029	0.071	0.139	−0.005	0.232	2
Central Air	0.066	−0.042	0.090	0.225	−0.259	0.309	10
Hardwood Floors	0.100	0.016	0.075	0.163	−0.054	0.280	2
Luxury Score	0.008	−0.052	0.009	0.055	−0.097	0.115	22
Foreclosure	−0.378	−0.539	−0.344	−0.131	−0.660	0.000	3
Condo	−0.120	−0.406	−0.114	0.054	−0.932	0.589	10
HOA	−0.006	−0.151	−0.020	0.074	−0.482	0.253	23

Notes: Unstandardised specification of Table 8 with lot size recoded (Table 29) and implausible records removed (N = 783,872). "Pooled, state FE" adds 50 state fixed effects to the pooled regression. State regressions by ordinary least squares within each jurisdiction; the last column counts states whose coefficient has the opposite sign from the pooled estimate. Generated by `scripts/06_revision_inference.py`.

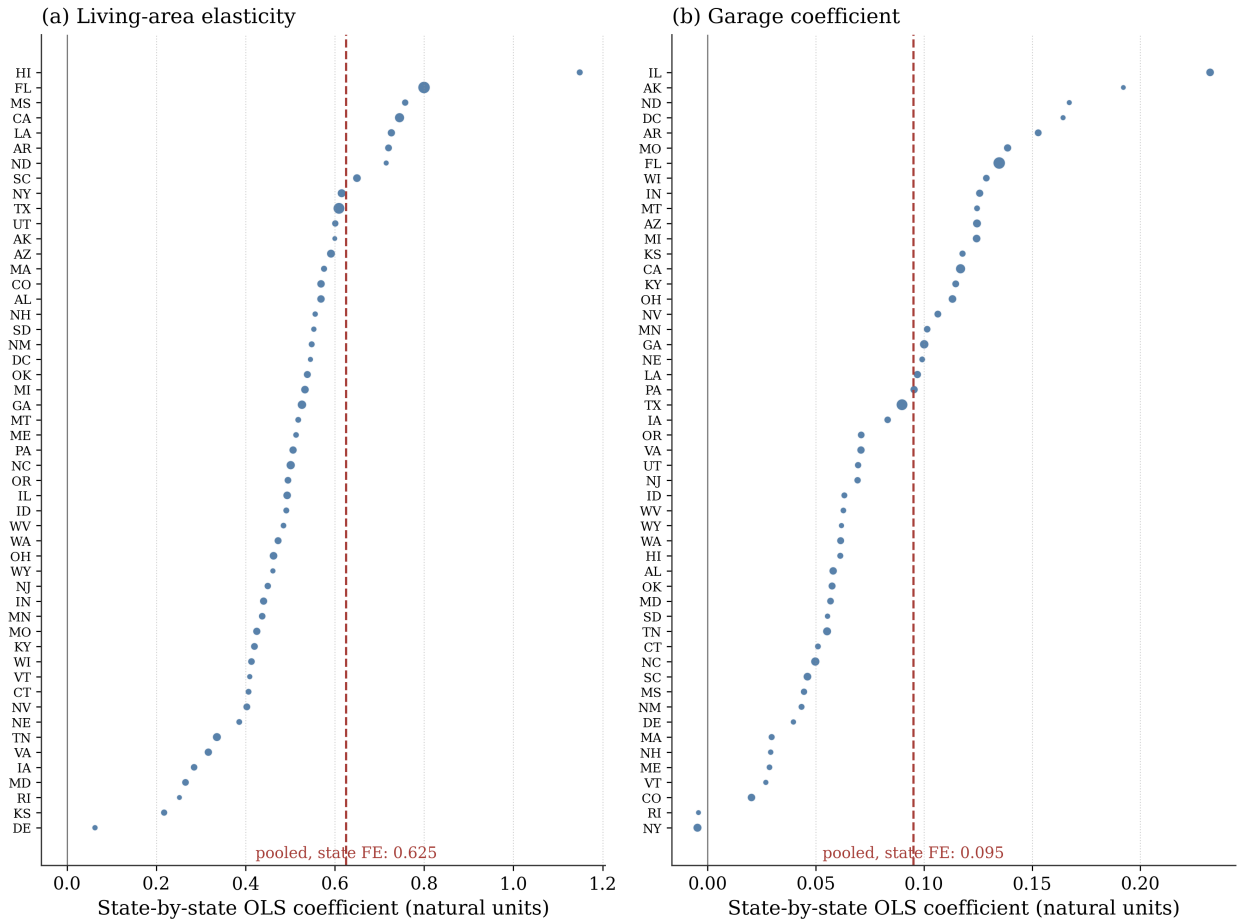


Figure 3. State-by-state OLS coefficients, natural units: (a) living-area elasticity; (b) garage coefficient. Marker size proportional to the square root of the state's sample size; the dashed line is the pooled coefficient with state fixed effects.

The pooled living-area elasticity of 0.625 is an average of state elasticities ranging from 0.06 to 1.15, with an inter-decile range of 0.32–0.72; Florida (0.80), California (0.75) and New York (0.62) price space far more steeply than Tennessee (0.34) or Ohio (0.46). The bathroom coefficient ranges from 0.02 to 0.29 across states and the garage coefficient from −0.005 to 0.23. For several attributes the pooled coefficient is not even representative of the sign: the bedroom coefficient is positive in 19 states, the pool coefficient negative in 22, the luxury-score coefficient negative in 22 and the HOA coefficient positive in 23. Only living area and bathrooms have the same sign in every state.

Table 14 translates this heterogeneity into predictive terms. On the random holdout the pooled regression without geographic controls reaches $R^2 = 0.619$; adding state fixed effects raises it to 0.690; allowing every slope to differ by state raises it to 0.751. Slope heterogeneity across states is therefore worth 6.1 percentage points of out-of-sample R^2, almost as much as the state price levels themselves (7.1 points). The fully interacted linear model, with no non-linearity beyond the semi-log form, recovers close to half of the gap between OLS with state fixed effects and XGBoost.

The implication for the interpretation of the pooled coefficients is the one drawn by the submarket literature (Bourassa et al., 2003; Goodman and Thibodeau, 2003): the national hedonic equation is a listing-weighted average of distinct local price functions, and its coefficients describe no particular market.

Table 14. Random-Holdout R^2 of Pooled, Fixed-Effects and State-by-State OLS

Specification	Out-of-sample R^2
Pooled OLS, no geographic controls	0.618
Pooled OLS + 50 state fixed effects	0.690
State-by-state OLS (fully interacted)	0.751

Notes: Random 80/20 split (seed 42) restricted to the trimmed sample (627,102 training, 156,770 test listings); natural units, lot size recoded. The fully interacted model estimates a separate regression in each state on the training listings of that state.

5.4. Quantile Regression and Inter-Quantile Tests

5.4.1. Coefficient Estimates Across Quantiles

Table 15 presents quantile-regression coefficients at the five estimated quantiles; the pseudo-R^2 of Koenker and Machado (1999) is 0.583, 0.598, 0.601, 0.580 and 0.544 at $\tau = 0.10, 0.25, 0.50, 0.75, 0.90$, so the specification fits the centre of the conditional distribution better than its upper tail.

Table 15. Quantile Regression Coefficients at Selected Quantiles

Variable	$\tau = 0.10$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.90$	OLS
ln(Living Area)	0.358***	0.313***	0.275***	0.237***	0.251***	0.312***
Bedrooms	−0.085***	−0.076***	−0.064***	−0.037***	−0.034***	−0.068***
Bathrooms	0.157***	0.195***	0.234***	0.259***	0.259***	0.236***
Age	−0.430***	−0.201***	−0.066**	−0.007	−0.026	−0.042**
ln(Lot Size)‡	0.045***	0.051***	0.061***	0.070***	0.086***	0.070***
Pool	0.008	0.008	0.019***	0.061***	0.094***	0.029***
Garage	0.205***	0.146***	0.102***	0.061***	0.020***	0.109***
Luxury Score	0.018***	0.038***	0.054***	0.068***	0.072***	0.044***
Foreclosure	−0.494***	−0.473***	−0.379***	−0.327***	−0.338***	−0.406***
Northeast	0.339***	0.395***	0.471***	0.553***	0.603***	0.487***
West	0.365***	0.372***	0.376***	0.399***	0.450***	0.401***
Pseudo-R^2	0.583	0.598	0.601	0.580	0.544	

Notes: Random subsample of 150,000 listings; continuous variables standardised; all 62 regressors included, selected coefficients shown. ‡Published coding of lot size. Stars from the default kernel-based standard errors, which assume independent observations: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

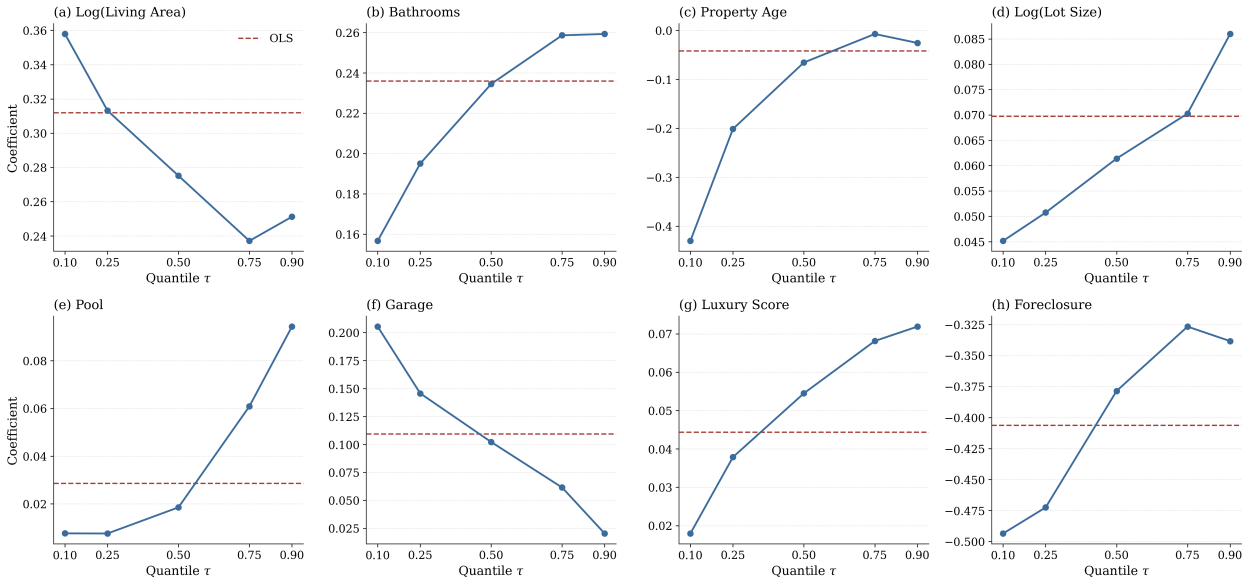


Figure 4. Quantile-regression coefficients across the conditional price distribution. Dashed lines are OLS estimates; coefficients are on standardised variables.

Four patterns stand out. The living-area gradient declines from 0.358 at $\tau = 0.10$ to 0.237 at

$\tau = 0.75$: additional space is associated with proportionally larger price differences at the lower end of the conditional distribution. The bathroom gradient rises monotonically from 0.157 to 0.259. The garage gradient declines from 0.205 to 0.020, a tenfold difference, consistent with a garage differentiating lower-priced homes and being nearly universal, hence unpriced, higher up. The pool gradient is insignificant below the median and reaches 0.094 at $\tau = 0.90$, a $100(e^{0.094} - 1) = 9.9\%$ difference. The age main effect moves from -0.430 at $\tau = 0.10$ to an insignificant -0.026 at $\tau = 0.90$; because age enters with its square and two interactions, this coefficient is the marginal association at the sample mean of the interacted variables and is the least precisely estimated of the set (Section 6.4).

5.4.2. Inter-Quantile Tests ($\tau = 0.10$ versus $\tau = 0.90$)

Table 16 tests the equality of the $\tau = 0.10$ and $\tau = 0.90$ coefficients for the thirteen key variables. Three statistics are reported for each difference: the z-statistic of the previous version, which treats the two quantile estimates as independent and the observations as independent; a paired observation bootstrap, which supplies the covariance term of Eq. (5); and a paired ZIP3-cluster bootstrap, which also allows for spatial dependence.

Table 16. Inter-Quantile Wald Tests, $\hat{\beta}(0.10) - \hat{\beta}(0.90)$, with Bootstrap Covariance

		z-statistic				
Variable	Difference	Indep.	Boot. (obs.)	Boot. (ZIP3)	95% CI (ZIP3)	
Garage	+0.185	28.92	24.43	11.78	[0.149, 0.211]	***
Bathrooms	−0.103	−10.48	−7.99	−5.06	[−0.140, −0.059]	***
Region: Northeast	−0.264	−21.43	−16.32	−4.88	[−0.361, −0.136]	***
ln(Living Area)	+0.107	9.52	6.76	4.15	[0.063, 0.168]	***
ln(Lot Size)	−0.041	−9.49	−9.16	−4.09	[−0.062, −0.024]	***
Property Age	−0.405	−7.89	−5.83	−3.85	[−0.607, −0.183]	***
Luxury Score	−0.054	−6.82	−6.93	−3.73	[−0.079, −0.026]	***
Bedrooms	−0.051	−5.95	−4.58	−3.44	[−0.083, −0.029]	***
Pool	−0.087	−7.80	−6.48	−2.09	[−0.151, 0.002]	*
Foreclosure	−0.155	−4.18	−2.31	−2.06	[−0.325, −0.040]	*
Region: West	−0.086	−6.26	−5.91	−1.65	[−0.193, −0.001]	
Waterfront	+0.222	1.86	1.36	0.71	[−0.514, 0.660]	
Property Age^2	+0.010	1.40	0.92	0.61	[−0.018, 0.046]	

Notes: Both quantile regressions estimated on the same 150,000-listing subsample (seed 42). "Indep." uses the kernel-based standard errors of each quantile regression and sets the covariance to zero (previous version). Bootstrap columns use 200 paired replications in which both quantile regressions are re-estimated on a resample of the subsample, drawn by observation or by ZIP3 cluster (resampling the subsample's ZIP3 areas with replacement). Stars refer to the ZIP3-cluster bootstrap: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Variables sorted by $|z|$ (ZIP3). Generated by `scripts/07_revision_qr_bootstrap.py`.

Supplying the covariance term and allowing for spatial dependence changes the inference materially. The ZIP3-cluster bootstrap standard errors of the differences are 2.3 times the independence-based values on median (range 1.7–4.4), and the observation bootstrap alone, which supplies the covariance but keeps independence, leaves 11 rejections of 13. Under the ZIP3-cluster bootstrap the null of equal coefficients at $\tau = 0.10$ and $\tau = 0.90$ is rejected at the 5% level for 10 of the 13 key variables: living area, bedrooms, bathrooms, age, lot size, pool, garage, luxury score, foreclosure, Northeast. The differences that do not survive are age squared ($z = 0.61$), waterfront ($z = 0.71$), West ($z = -1.65$). The pool and foreclosure differences are marginal ($z \approx 2.1$; their percentile intervals touch zero). The largest and most robust differences remain those of garage ($z = 11.8$), living area ($z = 4.2$) and bathrooms ($z = -5.1$): the conditional distribution of asking prices is differentially stretched and compressed by the structural attributes, while the tail differences of the rarer amenities and of the region dummies are estimated with far less precision than the

independence-based z-statistics suggested. The pattern in Figure 4 therefore stands for the attributes that matter most, with uncertainty bands several times wider than those reported in the previous version.

5.5. Machine Learning Under Random Validation

Table 17 summarises out-of-sample performance under the random 80/20 split. XGBoost achieves $R^2 = 0.833$, a 20.3-percentage-point improvement over OLS; its median absolute percentage error on the dollar scale is 16.5%, against 27.2% for OLS. Regularised linear models (Ridge, Lasso, Elastic Net) offer no improvement over OLS, confirming that the parametric specification is not overfit. The ZIP3 fixed-effects model reaches 0.725.

Table 17. Out-of-Sample Predictive Performance: Random Validation

	Log-price scale			Dollar scale		
Model	R^2	RMSE	MAE	MAPE (%)	MdAPE (%)	MAE ($)
OLS	0.6301	0.4911	0.3619	39.8	27.2	240,941
Ridge ($\alpha = 1$)	0.6301	0.4911	0.3619	—	—	—
Lasso ($\alpha = 0.001$)	0.6280	0.4925	0.3627	—	—	—
Elastic Net	0.6145	0.5014	0.3681	—	—	—
OLS + state FE	0.6775	0.4586	—	—	—	—
OLS + ZIP3 FE	0.7253	0.4232	—	—	—	—
Random forest	0.7842	0.3751	0.2636	28.4	18.5	180,116
LightGBM	0.8088	0.3531	0.2519	26.9	18.2	169,495
XGBoost	**0.8330**	**0.3300**	**0.2327**	**24.7**	**16.5**	**157,822**

Notes: Training 631,073 (80%), test 157,769 (20%), seed 42. MAPE = mean absolute percentage error; MdAPE = median absolute percentage error, Eq. (7). Dollar metrics obtained by exponentiating predicted log prices. Dollar-scale metrics for the regularised and fixed-effects linear models were not computed.

Table 18 attaches sampling uncertainty to these comparisons. Under an observation bootstrap of the test set the 95% interval for the XGBoost–OLS difference is [0.200, 0.206]; resampling whole ZIP3 areas of the test set widens it to [0.193, 0.213], and the paired test on squared-error differences rejects equal accuracy by a wide margin. The ranking XGBoost > LightGBM > OLS is therefore not a sampling accident, even after spatial dependence in the test set is allowed for. The intervals also show that the cluster bootstrap inflates the uncertainty of every R^2 by a factor of four to five relative to the observation bootstrap, the same lesson as in Section 5.1.5.

Table 18. Bootstrap Confidence Intervals for Test-Set R^2 and Paired Differences

Quantity	Point estimate	95% interval	Bootstrap SE
Observation bootstrap			
R^2 OLS	0.6301	[0.6266, 0.6335]	0.0017
R^2 LightGBM	0.8088	[0.8067, 0.8108]	0.0011
R^2 XGBoost	0.8330	[0.8311, 0.8349]	0.0010
R^2 XGBoost - R^2 OLS	0.2029	[0.2004, 0.2055]	0.0013
R^2 XGBoost - R^2 LightGBM	0.0242	[0.0236, 0.0249]	0.0003
R^2 LightGBM - R^2 OLS	0.1787	[0.1764, 0.1812]	0.0012
ZIP3-cluster bootstrap			
R^2 OLS	0.6301	[0.6162, 0.6419]	0.0067
R^2 LightGBM	0.8088	[0.7972, 0.8181]	0.0055
R^2 XGBoost	0.8330	[0.8224, 0.8415]	0.0050
R^2 XGBoost - R^2 OLS	0.2029	[0.1934, 0.2130]	0.0051
R^2 XGBoost - R^2 LightGBM	0.0242	[0.0226, 0.0258]	0.0008
R^2 LightGBM - R^2 OLS	0.1787	[0.1695, 0.1880]	0.0050

Notes: 1,000 bootstrap replications of the 157,769 random-split test listings, resampling observations or the 885 ZIP3 clusters of the test set. The paired test on squared-error differences (not shown) gives a mean loss difference of 0.132 (SE 0.0009) for OLS minus XGBoost and 0.016 (SE 0.0002) for LightGBM minus XGBoost. Generated by `scripts/06_revision_inference.py`.

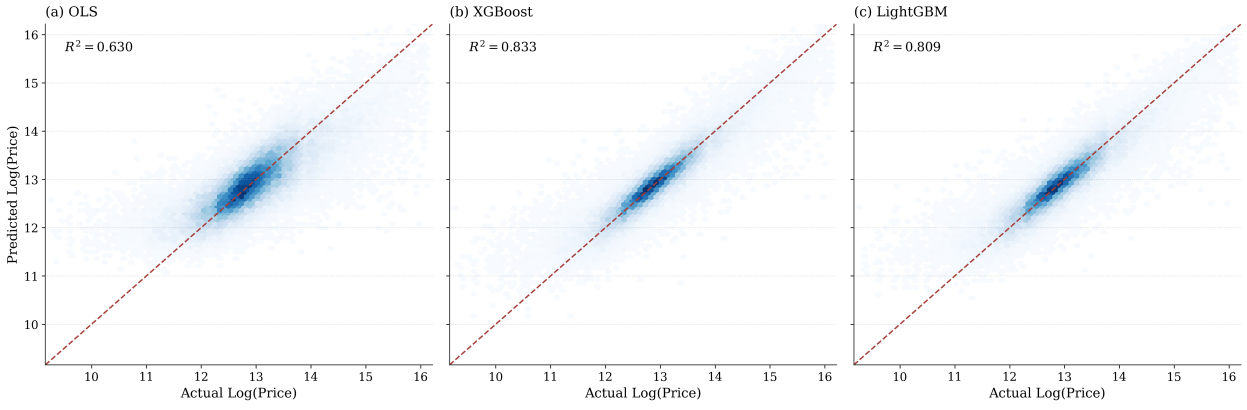


Figure 5. Predicted versus actual log prices under random validation: (a) OLS; (b) XGBoost; (c) LightGBM.

5.6. Machine Learning Under Geographic Validation

5.6.1. The Ten-State Holdout, Revisited

Table 19 reports the full XGBoost model under three designs: the random split, a random split with the training-set size of the geographic design, and the ten-state holdout. Shrinking the random training set from 631,073 to 350,527 listings costs 0.4 points of R^2 (0.833 to 0.829). Withholding ten states costs 40.8 points (0.425). Training-set size therefore explains almost none of the gap; the rest is extrapolation. The within-state R^2 of Eq. (8) is 0.618: an analyst who knew each held-out state's average price level would recover 62% of the within-state variation from the model's predictions. The between-state component accounts for 12.4% of the variance of log prices in the held-out set, so the collapse from 0.618 to 0.425 is due less to the size of the between-state variation than to the model mispredicting state price levels in a way that adds error rather than removing it. The median absolute percentage error rises from 16.5% to 29.8%.

Table 19. Full XGBoost Model Under Random, Size-Matched Random and Ten-State Designs

Design	N_{train}	N_{test}	R^2	R^2_{within}	RMSE	MdAPE (%)
Random 80/20 split	631,073	157,769	0.833	—	0.330	16.5
Random split, geographic training size	350,527	438,315	0.829	—	0.334	16.7
Ten-state geographic holdout	350,527	438,315	0.425	0.618	0.612	29.8

Notes: 62 features, 1,000 boosting rounds. "Random, geographic training size" draws 350,527 training listings at random. R^2_{within} removes state means of actual and predicted log prices in the test set, Eq. (8). Generated by `scripts/05_revision_geo_validation.py`.

The previous version of this paper reported that OLS "fails catastrophically" under the same holdout ($R^2 < 0$) and attributed the failure to the regional dummies. Table 20 corrects this. The published figure, $R^2 = -1.96 \times 10^6$, is produced by 204 held-out listings with corrupt parking or story counts (Table 4) on which the linear model extrapolates to absurd values; the median absolute percentage error of the same predictions is an unremarkable 27.9%. Once the 4,970 implausible records are removed from training and test sets, OLS with the same 62 regressors reaches $R^2 = 0.539$ in the held-out states, with a within-state R^2 of 0.560 and an MdAPE of 27.9%. On the same trimmed test set the 1,000-round XGBoost model reaches 0.426 and the 100-round model 0.544. Under extrapolation to unseen states, the semi-log hedonic regression is therefore as accurate as the best tree ensemble and more accurate than the one the paper reported. The regional dummies cost OLS only 1.8 points (0.539 versus 0.521 without them).

Table 20. Ten-State Holdout: OLS With and Without Implausible Records, and XGBoost on the Same Test Set

Model	N_{test}	R^2	R^2_{within}	RMSE	MdAPE (%)
OLS, all observations	438,315	-2.0×10^6	—	1129	27.9
OLS, implausible records removed	435,553	0.539	0.560	0.547	27.9
OLS, no region dummies, implausible removed	435,553	0.521	0.546	0.558	30.9
XGBoost, 100 rounds, same trimmed test set	435,553	0.544	—	0.544	27.1
XGBoost, 1,000 rounds, same trimmed test set	435,553	0.426	—	0.611	29.7

Notes: "Implausible values removed" drops the 4,970 listings failing the four screens of Table 4 from both training and test sets (348,319 training listings). XGBoost rows: full model trained on all 350,527 training listings, scored on the trimmed test set. Generated by `scripts/05_revision_geo_validation.py` and `06_revision_inference.py`.

5.6.2. Boosting Rounds and the Two Published Numbers

The previous version reported two geographic-holdout values for the full XGBoost model, 0.547 and 0.425, and attributed the difference to "training-set composition". Figure 6 and Table 21 identify the actual cause. Held-out-state R^2 rises to a peak of 0.550 after 75 boosting rounds and then declines monotonically to 0.425 at 1,000 rounds, while R^2 on the early-stopping set, drawn from the training states, keeps rising from 0.77 to 0.83, as does random-split test R^2. Beyond roughly 100 rounds every additional tree improves the fit within the training states and degrades it outside them. The 0.547 figure corresponds to a model stopped near the peak; the 0.425 figure to the fully grown model. Early stopping on a random subset of the training data, the standard practice and the one used here, cannot detect this because the evaluation set shares the training states' price levels. The result is the boosting analogue of the spatial-tuning problem documented by Schratz et al. (2019): hyperparameters selected under random resampling are not the hyperparameters that maximise performance under spatial extrapolation.

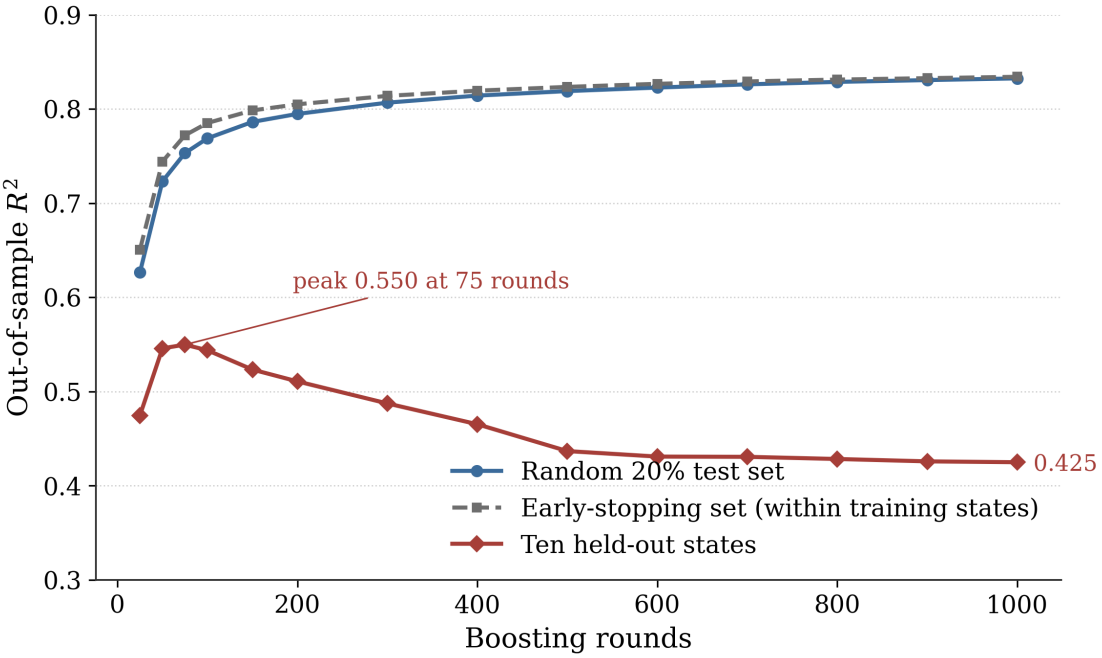


Figure 6. Out-of-sample R^2 of the full XGBoost model as a function of boosting rounds, on the random 20% test set, on the early-stopping set drawn from the training states, and on the ten held-out states.

Table 21. Out-of-Sample R^2 by Number of Boosting Rounds

Rounds	Random test set	Early-stopping set	Ten held-out states
25	0.627	0.651	0.475
50	0.723	0.744	0.546
75	0.753	0.772	0.550
100	0.769	0.785	0.544
200	0.795	0.805	0.511
300	0.807	0.814	0.487
500	0.819	0.824	0.437
700	0.826	0.829	0.431
1000	0.833	0.834	0.425

Notes: Full model, 62 features. The random-test and early-stopping columns come from the model trained on the random 80% split; the held-out-state column from the model trained on the forty remaining states. Generated by `scripts/05_revision_geo_validation.py`.

5.6.3. Grouped Cross-Validation and Leave-One-State-Out

A single ten-state partition is one draw from a large set of possible partitions, and it withholds 56% of the data. Table 22 reports grouped five-fold cross-validation in which each fold holds out a different set of states. Pooled over the five folds, every listing being predicted by a model that saw no listing from its state, the full model reaches $R^2 = 0.620$ and a within-state R^2 of 0.657; fold-level R^2 ranges from 0.504 to 0.717. The ten-state holdout, at 0.425, is thus a pessimistic draw: it withholds the five largest states at once, so that the training set loses most of the Sun Belt and West Coast price variation.

Table 22. Grouped Five-Fold Cross-Validation by State, XGBoost

Fold	States	N_{test}	R^2	R^2_{within}	RMSE	MdAPE (%)
Full model (62 features)						
Fold 1	6	157,887	0.504	0.642	0.577	29.1
Fold 2	9	157,911	0.631	0.671	0.448	25.1
Fold 3	12	157,982	0.583	0.629	0.544	27.6
Fold 4	12	157,366	0.717	0.704	0.408	22.0
Fold 5	12	157,696	0.656	0.647	0.495	25.9
Pooled over the five folds	51	788,842	0.620	0.657	0.498	25.8
Without school variables (59 features)						
Fold 1	6	157,887	0.457	0.628	0.603	30.9
Fold 2	9	157,911	0.649	0.660	0.437	24.4
Fold 3	12	157,982	0.571	0.628	0.552	27.9
Fold 4	12	157,366	0.704	0.689	0.417	22.3
Fold 5	12	157,696	0.605	0.630	0.530	27.1
Pooled over the five folds	51	788,842	0.597	0.645	0.513	26.4

Notes: Folds formed by `GroupKFold` on state; each model is trained on the other four folds with 1,000 rounds. "Pooled" scores the concatenated out-of-fold predictions. Generated by `scripts/05_revision_geo_validation.py`.

Leave-one-state-out validation (Table 23, Figure 7; state-level detail in Appendix C) gives the most complete picture. The median held-out-state R^2 of the full model is 0.653, with an inter-decile range of 0.42–0.74; forty-four of the 51 jurisdictions exceed 0.5, and only one, California, has a negative R^2 (−0.02). The listing-weighted mean is 0.564, pulled down by the largest states: Florida is predicted with $R^2 = 0.44$, Texas 0.68, California −0.02, Georgia 0.73 and North Carolina 0.56. The median across the ten states of the original holdout, each held out alone, is 0.60, against 0.68 for the other 41 jurisdictions, so the ten-state figure of 0.425 reflects the simultaneous removal of

the largest and most expensive markets rather than a property of any one of them. Level bias is the dominant error. The median absolute mean error across states is 0.11 log points; California is under-priced by 0.54 log points on average (about 42%), Hawaii by 0.50, and Louisiana and West Virginia are over-priced by 0.42 and 0.48. The states with the worst R^2 are precisely those whose price level lies farthest from what the rest of the country implies for homes with their attributes. The median absolute percentage error across states is 22.5%, against 16.5% under the random split. Stopping after 100 rounds improves held-out R^2 in 17 of 51 states, mostly the large ones (median across states 0.634), so the capacity effect of Section 5.6.2 is concentrated where the training set loses the most information. Removing the school variables lowers the median to 0.622. Taken together, the three geographic designs bracket the out-of-market accuracy of the fully grown model between 0.43 (ten large states withheld together) and 0.65 (one state at a time), with within-state accuracy near 0.62 throughout: the model transfers within-market structure, not market levels.

Table 23. Leave-One-State-Out Validation: Distribution of Accuracy Across 51 Held-Out Jurisdictions

	R^2 across states						$R^2 < 0$	Median	Median
Model	Median	p_{10}	p_{90}	Min	Max	Weighted	(states)	|bias|	MdAPE
Full model	0.653	0.417	0.738	−0.016	0.770	0.564	1	0.111	22.5
No school variables	0.622	0.421	0.745	−0.199	0.808	0.541	1	0.119	23.6

Notes: 51 models per variant, each trained on all listings outside the held-out jurisdiction (1,000 rounds). "Weighted" is the listing-weighted mean of state R^2. |bias| is the absolute mean error of predicted minus actual log price in the held-out state. Generated by `scripts/05_revision_geo_validation.py`.

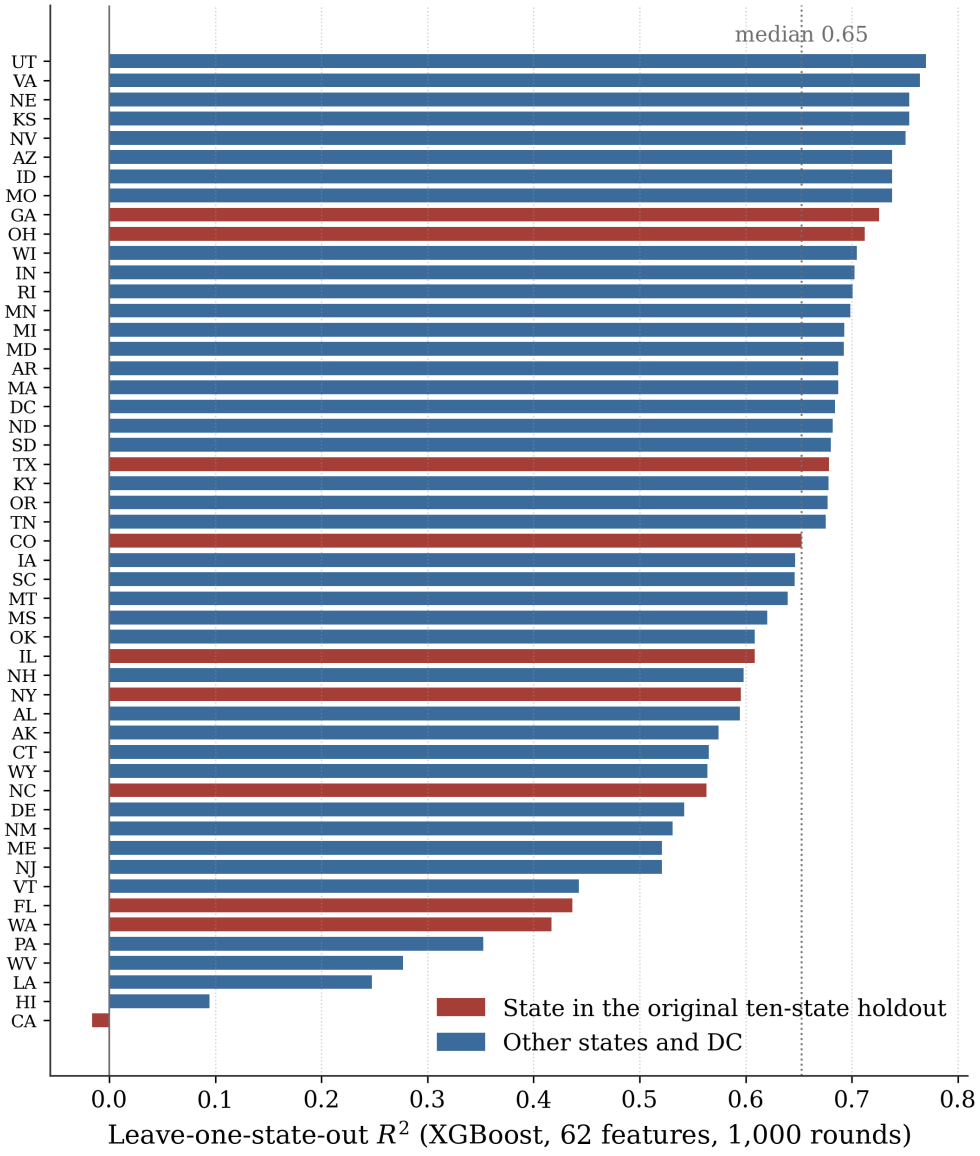


Figure 7. Leave-one-state-out R^2 of the full XGBoost model by held-out state. States of the original ten-state holdout are shaded differently.

5.7. Ablation: What Drives the Machine-Learning Gain?

Table 24 reports the feature-ablation analysis for XGBoost under both designs.

Table 24. XGBoost Ablation Analysis: Marginal Contribution of Feature Groups

Stage	Feature set	N_{features}	Random R^2	Ten-state R^2	Δ ten-state
1	Structural only	8	0.498	0.348	—
2	+ Lot	9	0.546	0.376	+0.028
3	+ Amenities	20	0.638	0.449	+0.072
4	+ Neighbourhood	27	0.814	0.483	+0.035
5	+ Market status	32	0.824	0.501	+0.018
6	Full (+ interactions, categorical, region)	62	0.833	0.425	−0.076

Notes: Each row adds a feature group to the previous stage; 1,000 rounds. Random R^2 on the 20% random holdout, ten-state R^2 on the geographic holdout.

Three findings emerge. First, the neighbourhood block (Walk, Bike and Transit Scores, school rating, school count, school distance, tax rate) contributes by far the largest random-validation gain, +17.6 points, but only +3.5 points under the ten-state design. Given Section 3.5, this is not a statement about local amenities: three of the seven variables are state medians for 93% of listings and the tax rate takes 243 distinct values across the country, so the block supplies the trees with a near-perfect encoding of the listing's state. Under random validation that encoding is worth 17.6 points because the test listings' states are in the training set; under the ten-state design the imputed values of the held-out states have never been seen and the encoding is nearly worthless. Second, the full model loses 7.6 points of ten-state R^2 relative to stage 5 while gaining 0.9 points under random validation. Section 5.8 shows that the region dummies are responsible. Third, eight structural variables alone give $R^2 = 0.348$ in the held-out states, so basic size and room counts carry predictive content that transfers across markets.

5.8. Geographic Features, School Variables and Missingness Indicators

Table 25 evaluates variants of the full model under both designs. Removing the three region dummies costs 0.3 points of random R^2 and gains 9.4 points of ten-state R^2 (0.425 to 0.519); the within-state R^2 is unchanged (0.618 to 0.621), so the dummies damage the prediction of state price *levels*, not of within-state structure. Trees use the region dummies to partition the training states' price levels, and in a held-out state the resulting partition is wrong by construction. Adding raw coordinates has the opposite effect in the previous version's experiment (random 0.870, ten-state 0.464): the most flexible location encoding yields the largest interpolation gain and a large extrapolation loss.

Table 25. Variants of the Full XGBoost Model Under Random and Ten-State Validation

		Random	Ten-state holdout			
Variant	Feat.	R^2	R^2	R^2 (100 r.)	R^2_{within}	MdAPE
Full model	62	0.833	0.425	0.544	0.618	29.8
No school variables	59	0.817	0.377	0.541	0.596	31.5
No region dummies	59	0.830	0.519	0.520	0.621	28.4
No school variables, no region dummies	56	0.797	0.494	0.518	0.599	27.9
Full model + 6 missingness indicators	68	0.835	0.446	0.568	0.623	29.0

Notes: 1,000 rounds unless stated. "School variables" are average school rating, school count and nearest-school distance (state medians for 93% of listings). "Missingness indicators" flag listings whose raw year built, lot size, school rating, Walk Score, garage or Transit Score was missing. Generated by `scripts/05_revision_geo_validation.py`.

The school variables behave differently from the region dummies. Removing them costs 1.6 points of random R^2 (0.833 to 0.817) and 4.8 points of ten-state R^2 at 1,000 rounds (0.425 to 0.377), although nothing changes at 100 rounds (0.544 against 0.541). The state-median school rating is thus a useful feature even out of state: a held-out state's imputed rating is a new value, but the trees route it to the leaf of the training state with the nearest rating, and because state median ratings are correlated with state price levels this produces a crude but informative guess of the unseen state's level. The mechanism is the same memorisation of state price levels, operating through a variable labelled "school quality". Removing both the school variables and the region dummies yields the most transferable specification in terms of dollar error (ten-state MdAPE 27.9%) at a cost of 3.6 points of random R^2. Adding six missingness indicators improves every metric slightly (random 0.835, ten-state 0.446, within-state 0.623), consistent with the view that the pattern of missing fields is itself informative about the listing.

5.9. SHAP Analysis and Cross-Model Stability

5.9.1. Feature Importance Rankings

Table 26 reports the twenty features with the largest mean absolute SHAP values in the XGBoost model. Living area (0.174) and bathrooms (0.161) dominate. School rating (0.088), the West and Northeast region dummies (0.075 and 0.043), lot size (0.075), the luxury score (0.071) and the tax rate (0.069) form a second tier.

Table 26. Top 20 Features by Mean Absolute SHAP Value (XGBoost)

Rank	Feature	Mean $\lvert\phi\rvert$
1	ln(Living Area)	0.1740
2	Bathrooms	0.1610
3	Avg. School Rating[¶]	0.0881
4	Region: West	0.0751
5	ln(Lot Size)[‡]	0.0747
6	Luxury Score	0.0713
7	Property Tax Rate	0.0694
8	Nearest School Distance[¶]	0.0472
9	Region: Northeast	0.0425
10	Bike Score	0.0400
11	ln(HOA Fees + 1)	0.0398
12	Garage	0.0378
13	Age	0.0369
14	Hardwood Floors	0.0311
15	Waterfront $\times$ ln(sqft)	0.0292
16	Sqft per Bedroom	0.0289
17	Walk Score	0.0276
18	Central Air	0.0273
19	Region: South	0.0250
20	Transit Score	0.0213

Notes: TreeSHAP values on 10,000 random test listings. Mean $\lvert\phi\rvert$ is the average absolute contribution to the deviation of the predicted log price from its mean. [¶]State median for 93% of listings. [‡]Coded 0 for 16.5% of listings. These are predictive contributions, not implicit prices.

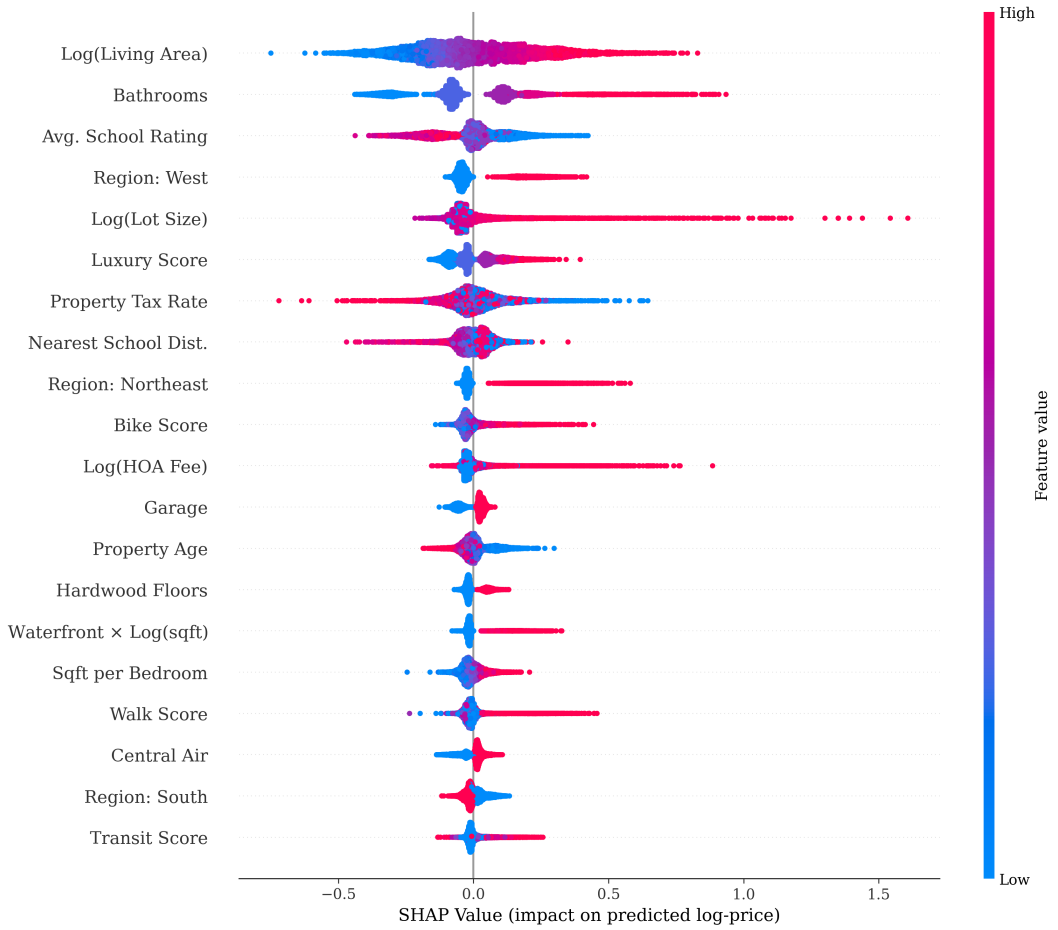


Figure 8. SHAP summary plot, XGBoost. Each dot is one listing; horizontal position is the feature's contribution to the predicted log price; colour indicates the feature value (red high, blue low). Predictive contributions, not causal effects.

The ranking must be read with Table 3 in hand. Four of the top nine features—school rating, the two region dummies and nearest-school distance—are location identifiers rather than property attributes, and together with the tax rate they are the channel through which the model memorises state price levels. The third-ranked "school rating" ranks third because it is, for 93% of listings, the state. This is a concrete instance of the general warning that SHAP decomposes the fitted prediction function, not the data-generating process (Lundberg and Lee, 2017; Rudin, 2019): a feature can be important because of what it proxies.

Figure 9 shows how the SHAP contribution of eight features varies with the feature value. The strongly concave living-area contribution, the saturation of the bathroom contribution and the threshold-like behaviour of the tax rate are genuine non-linearities the semi-log form cannot capture; the step pattern of the school-rating panel reflects its discrete, state-level support.

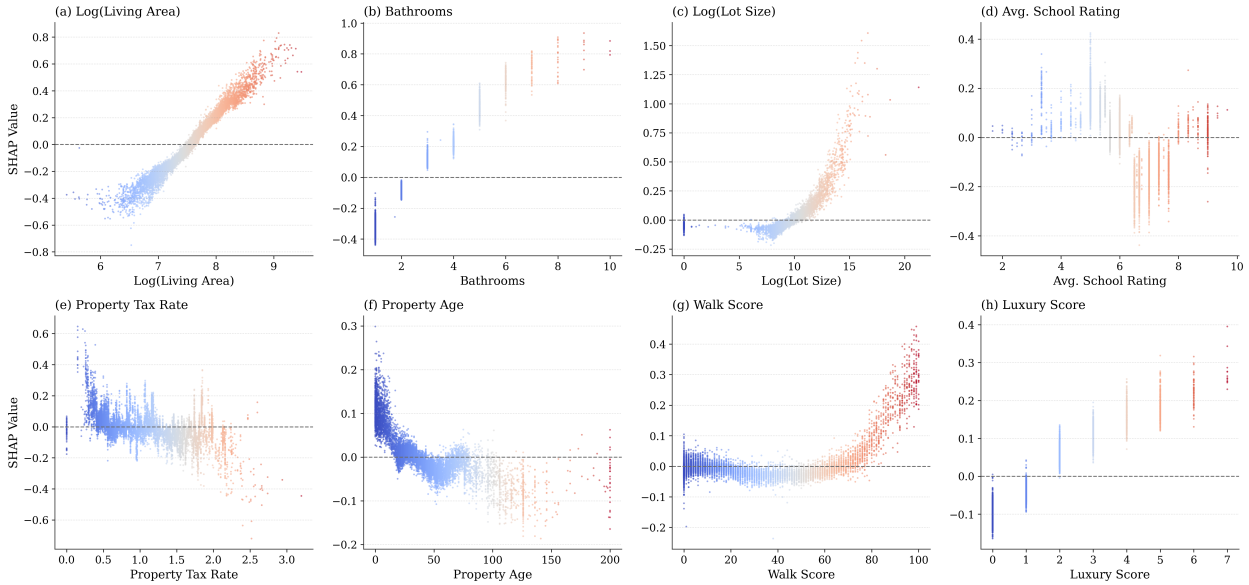


Figure 9. SHAP dependence plots for eight features. Each panel shows how the SHAP contribution varies with the feature value.

5.9.2. Cross-Model SHAP Stability

Table 27 reports pairwise Spearman rank correlations of mean absolute SHAP values across the three tree ensembles. XGBoost and LightGBM produce nearly identical rankings ($\rho = 0.990$); the random forest correlates at $\rho = 0.906$ and 0.892 with the two boosters. The same six features—living area, bathrooms, school rating, West, lot size and luxury score—rank in the top six for all three models. Stability across algorithms establishes that the ranking describes the predictive structure of the data as encoded in these 62 features; it does not convert any entry of the ranking into an implicit price, and in the case of the school variables it reflects the stability of a state identifier.

Table 27. Cross-Model SHAP Importance Stability: Spearman Rank Correlations

	XGBoost	LightGBM	Random forest
XGBoost	1.000	0.9898	0.9060
LightGBM	0.9898	1.000	0.8917
Random forest	0.9060	0.8917	1.000

Notes: Spearman ρ over the 62 mean $|\phi|$ values of each model, computed on the same 10,000 test listings.

5.10. Geographic Distribution of Listing Prices

Figure 10 maps the listing prices of a 50,000-listing subsample and Figure 11 shows their distribution by Census region. The coastal price gradient is visible, with California, the Northeast corridor, Hawaii and Colorado at the top. The map also illustrates the coverage issue: listing density is highly uneven, and sparse areas reflect few Zillow listings rather than few homes.

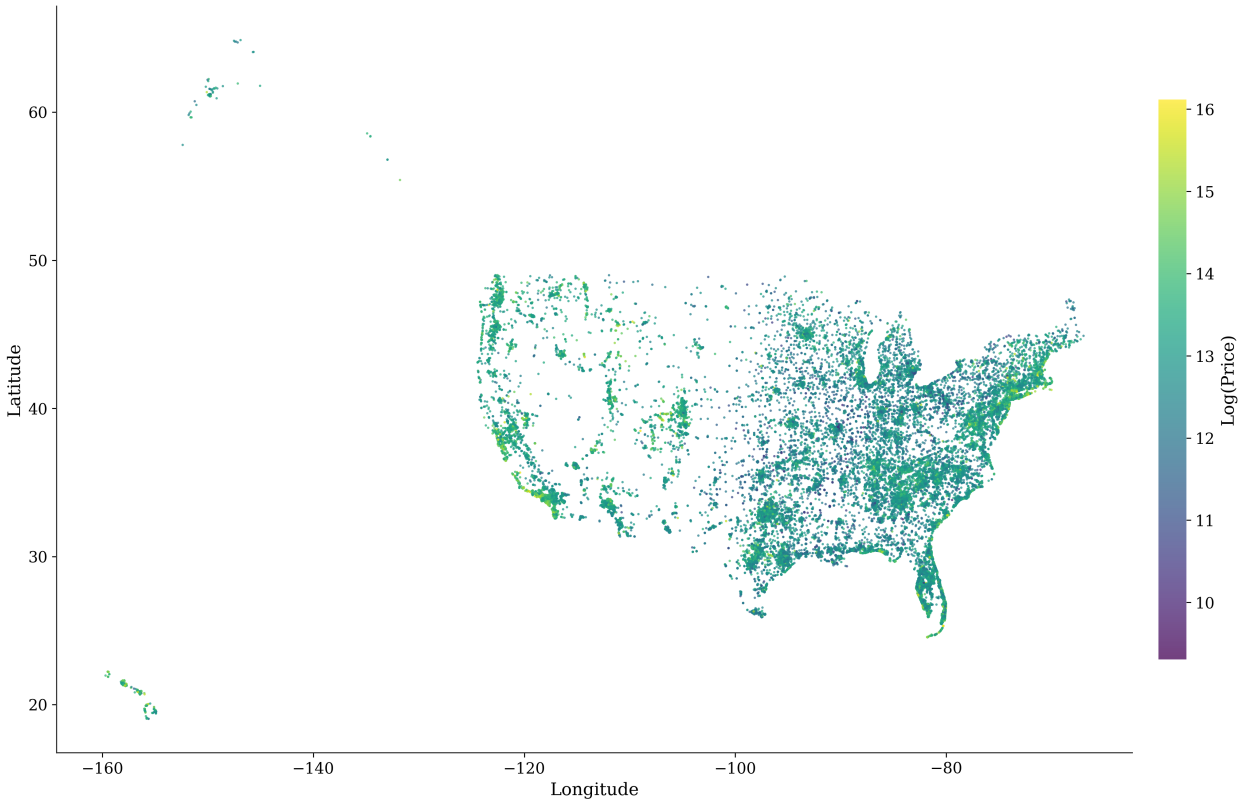


Figure 10. Geographic distribution of log listing prices, 50,000-listing random subsample. Colour indicates log price; listing density is not uniform.

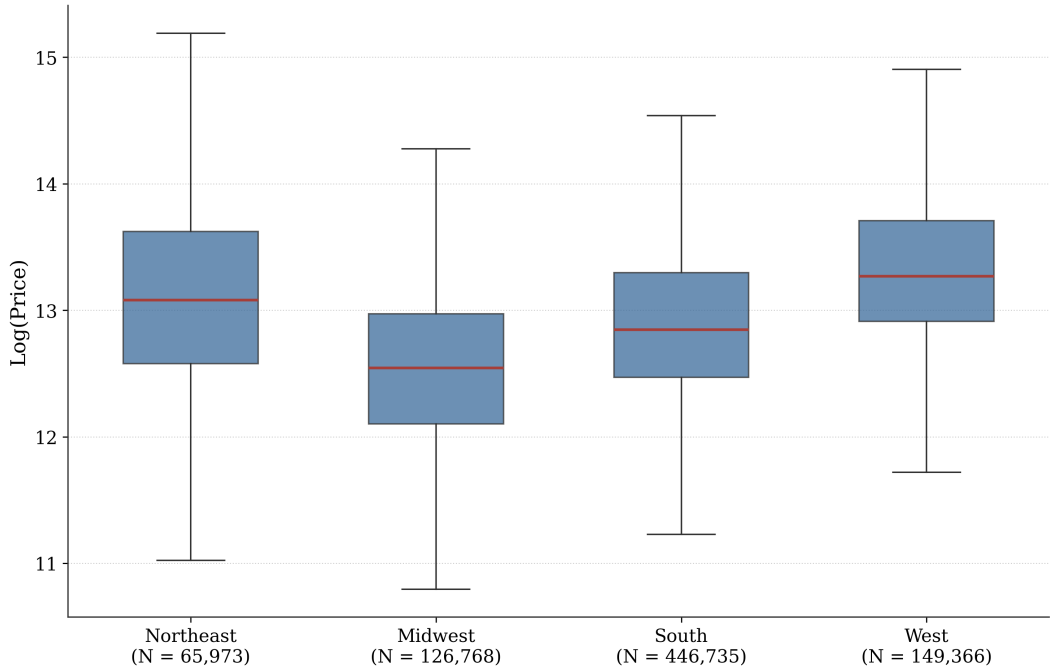


Figure 11. Distribution of log listing prices by Census region.

6. Robustness Checks

6.1. Imputation Sensitivity

Two variables with substantial missingness, year built (18.9%) and lot size (15.7%), were imputed with state medians at the sample-construction stage. Table 28 re-estimates the unstandardised baseline on subsamples that exclude the imputed listings. Dropping listings with imputed year built changes the lot-size coefficient from 0.018 to 0.019 and raises R^2 by 0.4 points; dropping listings with imputed lot size raises the lot-size coefficient to 0.059 and R^2 by 1.6 points; dropping both gives 0.059 and 0.654.

Table 28. Imputation Sensitivity: Unstandardised OLS on Complete-Case Subsamples

Specification	N	R^2	$\hat{\beta}_{\ln(\text{lot})}$	Notes
Full sample (baseline)	788,842	0.634	0.018	Published coding
Drop imputed year built	640,055	0.638	0.019	19% fewer listings
Drop imputed lot size	664,900	0.651	0.059	16% fewer listings
Drop both	534,803	0.654	0.059	32% fewer listings

Notes: Same 62-regressor unstandardised specification throughout; $\hat{\beta}_{\ln(\text{lot})}$ in natural units. The complete-case subsamples still contain the 6,329 listings with a reported lot size of zero, for which $\ln(\text{lot}) = 0$. Other coefficients are broadly stable across subsamples except the waterfront main effect, which ranges from 0.003 to 0.533.

The previous version read the threefold increase in the lot-size coefficient as evidence that "state-median imputation attenuates the lot-size gradient". That reading was wrong, because the regressor was not state-median imputed: as documented in Section 3.5, $\ln(\text{lot})$ equals zero for every listing whose lot size was missing or recorded as zero. The next subsection re-estimates the model with the coding the paper had described.

6.2. Lot-Size Coding

Table 29 compares five codings of lot size in the unstandardised baseline. Replacing the zeros by the log of the state-median lot size raises the elasticity of the asking price with respect to lot size from 0.018 to 0.095 (HC3 standard error 0.0007) and R^2 from 0.634 to 0.648; adding a missing-lot indicator leaves the elasticity at 0.095 and attaches a -2.2% discount to listings without a reported lot; restricting the sample to listings with a reported, positive lot gives 0.095 as well. The living-area elasticity falls from 0.628 to 0.596 under the corrected coding, because part of the lot-size gradient had been loading on living area. Winsorising the recoded lot at its 99.5th percentile changes the elasticity by 0.001.

Table 29. Lot-Size Coding and the Lot-Size Elasticity (Unstandardised OLS)

Coding	N	R^2	$\hat{\beta}_{\ln \text{lot}}$	(SE)	$\hat{\beta}_{\ln \text{sqft}}$	Indicator
Published: ln(lot) = 0 if missing or zero	788,842	0.634	0.018	(0.0003)	0.628	—
Recoded: log of state-median lot if missing or zero	788,842	0.648	0.095	(0.0007)	0.596	—
Recoded + missing-lot indicator	788,842	0.648	0.095	(0.0007)	0.597	−0.022
Complete cases (lot reported, positive)	658,785	0.660	0.095	(0.0007)	0.546	—
Recoded, lot winsorised at p99.5	788,842	0.647	0.096	(0.0007)	0.595	—

Notes: 62-regressor unstandardised specification; "missing or zero" covers 130,271 listings (16.5%). State medians computed from reported positive lots. Generated by `scripts/06_revision_inference.py`.

Three conclusions follow. The published lot-size elasticity of 0.02 is a coding artefact and the correct baseline figure is 0.095, five times larger and in line with the lot gradients reported in the meta-analysis of Sirmans et al. (2005). The "imputation attenuation" story of the previous version is replaced by a mechanical one: 130,271 observations placed eight log units below the mass of the data acted as a missing-value dummy and pulled the slope toward zero. And the episode illustrates why Table 3 documents the coding actually applied to each regressor, not the rule the analyst intended.

6.3. Winsorisation Sensitivity

To assess the influence of extreme values the previous version winsorised lot size at its 99.5th percentile (3,944 listings) and capped Bike Score at 100 (38 listings). Table 30 reproduces the result. R^2 rises from 0.634 to 0.640 and the lot-size coefficient from 0.018 to 0.058; the other coefficients move little. The large change in the lot coefficient is, again, not a winsorisation effect: the winsorised variable was recomputed from the state-median-imputed lot-size field, so this specification is effectively the recoded specification of Table 29 with a cap, and the residual difference between 0.058 and 0.095 reflects the 6,329 zero lots that remained at ln(lot) = 0. With the recoding in place, winsorisation at the 99.5th percentile is immaterial (0.095 versus 0.096).

Table 30. Winsorisation Sensitivity: Baseline versus Winsorised OLS

Metric / coefficient	Baseline	Winsorised
R^2	0.634	0.640
$\hat{\beta}_{\ln(\text{lot})}$	0.018	0.058
$\hat{\beta}_{\ln(\text{sqft})}$	0.628	0.610
$\hat{\beta}_{\text{bathrooms}}$	0.204	0.211
$\hat{\beta}_{\text{garage}}$	0.109	0.111

Notes: Lot size winsorised at its 99.5th percentile (≈ 3.04 million sqft; 3,944 listings) and recomputed from the imputed lot-size field; Bike Score capped at 100 (38 listings). Unstandardised OLS, N = 788,842.

6.4. Quantile-Regression Subsample Stability

Table 31 repeats the median regression on ten random subsamples of 150,000 listings and reports the coefficient of variation and sign stability of the key coefficients. Ten of the thirteen have coefficients of variation below 8% and the same sign in every subsample. Three do not: pool (CV 33.4%), waterfront (42.2%) and age (78.6%, sign reversed in two of ten subsamples).

Table 31. Quantile-Regression Stability at $\tau = 0.50$ Across Ten Random Subsamples

Variable	CV (%)	Sign stability (%)	Assessment
West	1.0	100	Highly stable
Northeast	1.6	100	Highly stable
ln(Lot Size)	2.7	100	Highly stable
ln(Living Area)	2.9	100	Highly stable
Bathrooms	2.9	100	Highly stable
Garage	4.1	100	Highly stable
Luxury Score	4.8	100	Highly stable
Age^2	5.3	100	Highly stable
Foreclosure	7.1	100	Stable
Bedrooms	7.3	100	Stable
Less stable variables			
Pool	33.4	100	Imprecise
Waterfront	42.2	100	Imprecise
Age	78.6	80	Imprecise

Notes: CV = standard deviation divided by the absolute mean of the coefficient across ten random subsamples of 150,000 listings (seeds in `src/wp3/config.py`). Sign stability = share of subsamples in which the coefficient has the sign of its mean.

The previous version attributed the instability of the age coefficient to "geographic heterogeneity across subsamples". The subsamples are random draws from the national sample and have the same geographic composition, so that explanation cannot be right. The instability is a precision problem that the paper's own OLS output already displays: the standardised age coefficient has an HC3 standard error of 0.014, ten times that of the other structural coefficients, because age enters together with its square, with $\ln(\text{sqft}) \times \text{age}$ and with age $\times$ luxury, and the three terms are nearly collinear once 19% of ages are set to a state median. A coefficient of -0.04 with a standard error of this order will change sign in a fifth of random subsamples. The pool and waterfront coefficients are imprecise for the reason Section 5.1.5 makes explicit: they are concentrated in a few markets, and their state-clustered t-statistics are 0.5 and -0.3. The stability table thus agrees with the clustered inference on which coefficients the data pin down and which they do not.

6.5. Robustness Summary

Table 32 collects the robustness evidence, distinguishing the checks carried over from the previous version from those added in the revision.

Table 32. Robustness Summary

Check	Baseline	Alternative	Finding
OLS specification and inference			
Region FE → state FE	$R^2 = 0.630$	$R^2 = 0.678$	+4.8 pp out of sample
Region FE → ZIP3 FE	$R^2 = 0.630$	$R^2 = 0.725$	+9.5 pp out of sample
State FE → state-by-state	$R^2 = 0.690$	$R^2 = 0.751$	Slopes differ across markets (+6.1 pp)
HC3 → state-clustered SE	57 of 63 significant	33 of 63	Median SE ratio 11.6
Lot-size coding	$\hat{\beta}_{\ln \text{lot}} = 0.018$	0.095	Zero-coding artefact; elasticity ×5
Winsorisation (lot p99.5, Bike ≤ 100)	$R^2 = 0.634$	$R^2 = 0.640$	Immaterial once lot recoded
Spatial diagnostics			
Moran's I, OLS (3 subsamples)	$I = 0.2745$	SD 0.0076	Strong, stable autocorrelation
Moran's I, OLS vs XGBoost residuals	0.263	0.079	Trees absorb 70% of spatial structure
Quantile regression			
QR stability (10 draws)	10 of 13 CV < 8%	Age CV 78.6%	Imprecision, not heterogeneity
Inter-quantile Wald, covariance	11 of 13 rejected (indep.)	10 of 13 (ZIP3 bootstrap)	Rejections fall from 11 to 10; SE ratio 2.3
Machine learning			
Random vs ten-state holdout	$R^2 = 0.833$	$R^2 = 0.425$	−40.8 pp; within-state $R^2 = 0.618$
Size-matched random split	$R^2 = 0.833$	$R^2 = 0.829$	Training size explains 0.4 pp
Boosting rounds, ten-state	0.550 at 75	0.425 at 1,000	Overfitting to training states
OLS, ten-state, implausible removed	$R^2 < 0$ (published)	$R^2 = 0.539$	OLS on par with XGBoost out of state
Grouped 5-fold by state	—	$R^2 = 0.620$	Folds 0.50–0.72
Leave-one-state-out (51)	—	median $R^2 = 0.653$	Range -0.02–0.77; level bias dominates
Remove region dummies	ten-state 0.425	0.519	+9.4 pp; within-state unchanged
Remove school variables	ten-state 0.425	0.377	State-median rating encodes state level
Add missingness indicators	0.833 / 0.425	0.835 / 0.446	Missingness pattern informative
Add latitude/longitude	0.833 / 0.425	0.870 / 0.464	Coordinates memorise location
Bootstrap CI, XGBoost − OLS R^2	0.203	[0.193, 0.213]	Robust to ZIP3 resampling
SHAP stability			
XGBoost vs LightGBM	—	$\rho = 0.990$	Near-identical rankings
XGBoost vs random forest	—	$\rho = 0.906$	Same top six features
LightGBM vs random forest	—	$\rho = 0.892$	Same top six features

Notes: "pp" = percentage points. Unless stated, $N = 788{,}842$ and machine-learning results use 1,000 boosting rounds.

7. Discussion

The results support four messages, one per modelling framework, and a fifth that cuts across them. For each we interpret the estimates against the literature, note where they agree with published findings and where they diverge, and state what the revision changed.

7.1. Message 1: The Pooled Hedonic Regression Is Informative About Averages, Not Markets

The semi-log OLS model explains 63.4% of the variation in log asking prices with 62 regressors, within the range typical of large cross-sectional hedonic studies (Sirmans et al., 2005; Malpezzi, 2003), and its headline gradients align with the meta-analytic record: a living-area elasticity of 0.63, a bathroom premium of 22.6%, a negative conditional bedroom gradient (the standard result once total area is held constant), and a 33.4% foreclosure discount. The last figure exceeds the 27% forced-sale discount estimated from Massachusetts transactions by Campbell et al. (2011), as expected if distressed listings are positioned below market before negotiation. Once lot size is coded as the paper had described, the lot elasticity is 0.095 rather than 0.02, also within the published range. Regularised linear benchmarks achieve the same accuracy as OLS, so the parametric specification is not overfit.

Two results of the revision qualify this picture. First, inference. With spatially correlated residuals and regressors that vary at the state or ZIP level, HC3 standard errors understate uncertainty by an order of magnitude; clustering by state leaves 33 of 63 coefficients significant where HC3 left 57. The gradients that survive are the structural ones—living area, bathrooms, garage, hardwood, central air, luxury score, foreclosure, the Northeast and West differentials—and the ones that do not are exactly the spatially aggregated proxies (school rating, tax rate, Walk Score) and the rare, geographically concentrated attributes (pool, waterfront). This is Moulton's pitfall in a hedonic setting, and it implies that many of the "highly significant" coefficients reported in large scraped-data hedonic studies that use conventional robust standard errors would not survive clustering.

Second, aggregation. The state-by-state estimates show that the pooled coefficients are averages of markedly different local price functions: the living-area elasticity ranges from 0.06 to 1.15 across jurisdictions, and for bedrooms, pools, the luxury score and HOA membership the pooled sign is wrong for roughly twenty states. Allowing the slopes to vary by state is worth 6.1 points of out-of-sample R^2, nearly as much as the state price levels themselves. This is the empirical content of Rosen's single-market condition and of the submarket literature (Goodman and Thibodeau, 1998; Bourassa et al., 2003; Goodman and Thibodeau, 2003): a national hedonic equation is a useful summary of average capitalization gradients in asking prices, but it is not the price function of any market and should not be used as one.

The fixed-effects progression—from four regions ($R^2 = 0.634$) to states (0.681) to 885 ZIP3

areas (0.729)—shows that a large share of the "unexplained" variation is spatial, echoing Kuminoff et al. (2010) and Bishop et al. (2020) on the primacy of spatial controls. The residual Moran's I of 0.27 after regional controls is comparable to the residual autocorrelation documented in metropolitan samples by Basu and Thibodeau (1998) and confirms that the spatial error structure modelled by the spatial econometrics literature (Anselin, 1988; LeSage and Pace, 2009; Dubin, 1998) is present at national scale.

7.2. Message 2: Quantile Regression Adds Distributional Insight, With Wider Uncertainty Than Reported

Quantile regression reveals heterogeneity that the mean masks, and its shape matches the housing literature. The declining living-area gradient (0.358 at $\tau = 0.10$ against 0.251 at $\tau = 0.90$) mirrors Zietz et al. (2008); the rising pool and lot-size gradients are consistent with amenity-driven luxury pricing documented at upper quantiles in Hong Kong (Mak et al., 2010) and Sydney (Waltl, 2019); the garage gradient, ten times larger at the bottom decile than the top, fits the view that basic functional amenities differentiate lower-priced homes and are unpriced where they are universal. Such quantile-dependent pricing is one face of market segmentation (Goodman and Thibodeau, 1998) and the cross-sectional analogue of McMillen (2008), who finds that house-price distributions move through coefficients rather than characteristics.

The revision replaces the independence-based inter-quantile test with a paired cluster-bootstrap test. With a paired ZIP3-cluster bootstrap for the covariance of the two quantile estimates, 10 of the 13 inter-quantile differences remain significant at the 5% level (11 under the independence-based test), and the bootstrap standard errors are on median 2.3 times the independence-based values. The structural gradients (garage, living area, bathrooms, lot size) carry the distributional story; the tail differences of age squared, waterfront, West do not survive. The age coefficient's instability across subsamples, previously attributed to geographic heterogeneity, is a precision problem created by near-collinearity among age, its square and two interactions; it is a reason not to interpret the national age gradient literally, not evidence about markets.

7.3. Message 3: The Machine-Learning Advantage Is Real for Interpolation and Absent for Extrapolation

XGBoost improves on OLS by 20.3 points of R^2 under random validation, with a ZIP3-cluster bootstrap interval of [0.193, 0.213], squarely within the gains reported in the ML-valuation literature (Bourassa et al., 2010; Park and Bae, 2015; Mayer et al., 2019; Pérez-Rave et al., 2019; Ho et al., 2021). The geographic designs overturn the reading that this reflects transferable attribute–price structure. Under the ten-state holdout the fully grown model reaches $R^2 = 0.425$; across the 51

leave-one-state-out models the median R^2 is 0.653; and, most tellingly, the semi-log OLS regression, once 0.6% of corrupt records are removed, reaches 0.539 in the held-out states—as much as the best tree ensemble and more than the fully grown one. The "catastrophic failure" of OLS reported in the previous version was an artefact of linear extrapolation on corrupt parking and story counts.

The revision also identifies the mechanics of the collapse, which the previous version could only describe. Three channels operate. (i) *Level memorisation.* The within-state R^2 in held-out states is 0.62 for every variant of the model, so the trees transfer within-market structure reasonably well; what they get wrong is the price level of a market they have not seen, and the region dummies make this worse because they partition training states' levels. (ii) *Imputation as geography.* Three of the seven "neighbourhood" features are state medians for 93% of listings. They supply the trees with a near-perfect state encoding, which is why the neighbourhood block is worth 17.6 points under random validation and 3.5 under the ten-state design, and why the school-rating variable ranks third in SHAP importance. The conjecture of the previous version—that Walk Score and school ratings "carry market-specific meaning" in the sense of Duncan et al. (2011)—is not needed to explain the result; the imputation explains it. (iii) *Capacity.* Held-out-state accuracy peaks after about 75 boosting rounds and falls by 12 points by round 1,000 while within-training-state accuracy keeps improving. Early stopping on a random evaluation set cannot see this, so the standard tuning protocol selects a model that is over-fitted to the training markets, the boosting analogue of the spatial-tuning problem of Schratz et al. (2019). The two geographic-holdout figures of the previous version, 0.547 and 0.425, were the two ends of this curve.

The scope of the conclusion follows the debate opened by Wadoux et al. (2021). For interpolation, valuing a home in a market represented in the training data, random validation is informative, the geographic features earn their keep, and the 20-point advantage of boosting is real; the Moran's I of tree residuals (0.08 against 0.26 for OLS) shows how much local structure the ensembles capture. For extrapolation to unrepresented markets, blocked validation is the relevant benchmark (Roberts et al., 2017; Valavi et al., 2019; Meyer and Pebesma, 2021, 2022), and under it the ranking of models changes, the ranking of feature sets changes, and the optimal number of boosting rounds changes. The practical rules for AVM deployment are therefore: match the validation protocol to the deployment question; report within-market and level accuracy separately, because an AVM that knows a market's level from a handful of local sales recovers most of the within-state accuracy; prefer the leaner, geography-free specification when new markets are in scope; and tune capacity on a spatially blocked evaluation set. Bogin and Shui (2020) reach a related conclusion for rural areas where AVMs lack comparable data: accuracy depends on how well the deployment population is represented in training.

7.4. Message 4: SHAP Interprets Prediction, Not Economics

SHAP values identify living area, bathrooms, school rating, region, lot size and the luxury score as the dominant predictive contributors, and this ranking is stable across three tree ensembles (ρ = 0.89–0.99), which addresses the concern that importance rankings are artefacts of one fitted model (Ribeiro et al., 2016). But stability is not structure, and the revision supplies a sharper illustration than the previous version could: the third most important feature is, for 93% of listings, the state. The ranking is stable across algorithms because every algorithm finds the same state encoding useful. Three cautions from the literature apply directly. SHAP decomposes predictions, not the data-generating process, and a feature earns importance by proxying unobservables (Lundberg and Lee, 2017; Lundberg et al., 2020); correlated features share contributions in ways that defeat attribute-level economic reading (Lorenz et al., 2023); and post-hoc explanation of a black box is not a substitute for an interpretable model when the stakes are high (Rudin, 2019). In Rosen's terms, SHAP values are not implicit prices, and the stability we document should raise confidence in SHAP as a description of *this prediction technology*, not as a measurement of *market valuation*.

7.5. Message 5: Data Documentation Is Part of the Method

Three of the revision's most consequential findings—the school variables as state identifiers, the zero-coded lot size, and the corrupt parking and story counts behind the published OLS failure—are data-documentation findings, not modelling findings. None of them was visible in the published tables; each changed a substantive conclusion. Scraped listing data are incomplete in structured ways, and the rules used to fill them (state medians, zeros, "unknown" categories) create variables whose statistical behaviour differs from the attribute they are named after. The missing-indicator analysis of Jones (1996) and the measurement cautions of Bishop et al. (2020) apply with full force. A table like Table 3, reporting for every regressor the share of imputed values and the rule actually applied in the code, is a necessary part of any hedonic study on such data, and the interpretation of coefficients, quantile paths and SHAP rankings should start from it.

7.6. Synthesis

Across the five messages a single theme recurs: each framework is reliable within the question it was built to answer, and only there. OLS with rich spatial controls and clustered inference yields stable, literature-consistent average gradients in asking prices, and the state-by-state estimates mark the limit of their generality; quantile regression reveals distributional structure that survives a conservative bootstrap for most but not all attributes; boosting delivers interpolation gains that are real and extrapolation gains that are not, with the gap traceable to level memorisation through imputed and regional features and to unchecked capacity; and SHAP describes the predictive machine without

licensing economic claims. The methodological corollary—that validation design, feature coding and model capacity interact, and that random-split comparisons overstate the machine-learning advantage for out-of-market questions—is the paper's most transferable lesson for the hedonic and AVM literatures alike.

8. Limitations

The limitations fall into four groups: the nature of the dependent variable, the construction of the regressors, the scope of the inference, and the design of the validation exercises.

The dependent variable is an asking price. Asking prices embed seller strategy (Horowitz, 1992; Genesove and Mayer, 2001; Han and Strange, 2016) and differ from transaction prices by a wedge that is plausibly correlated with attributes. The ratio of asking price to Zillow's valuation has a median of 1.02 and a narrow interquartile range, which bounds the average wedge in this sample but not its covariance with attributes; the Zestimate is itself partly a function of listing information. Listing-price gradients are therefore not transaction-price implicit prices, and the sample—active listings in a single 2025–2026 window, with the South over-represented—describes the supply side of the market at one moment, with no time dimension and no repeat-sales control. Climate-risk variables, which the capitalization literature shows to be priced (Baldauf et al., 2020; Bernstein et al., 2019; Murfin and Spiegel, 2020), are absent from the extract and plausibly contribute to the residual spatial autocorrelation.

The regressors inherit the incompleteness of scraped data. The three school variables are state medians for 93% of listings, the tax rate varies at the county level, and the year-built, garage, Transit Score and parking variables are imputed for 19–70% of listings by rules that make them partly indicators of reporting completeness. The lot-size regressor was coded to zero for 16.5% of listings in the published estimates; the revision corrects the coding for the OLS results but the quantile, ablation and SHAP results retain it, which is innocuous for the tree models (they treat the zero as a category) and understates the lot gradient in the quantile paths. A further 0.6% of listings carry corrupt secondary attributes that the linear models cannot tolerate. The upstream sample-construction code was lost before the 2026 restructuring, so the analytical sample is reproducible only from the stored artefacts; the revision documents the codings that it finds in those artefacts rather than the rules the original scripts were meant to apply.

No estimate in the paper identifies a causal effect. Cross-sectional listing data cannot separate preferences, supply constraints and sorting (Ekeland et al., 2004; Kuminoff et al., 2013), and the state-by-state estimates show that the pooled coefficients are averages of local price functions that differ in magnitude and sometimes in sign. The inference is now cluster-robust by ZIP3 and by state, but the state-clustered estimates rest on 51 clusters of very unequal size, and a spatial HAC

estimator (Conley, 1999) or a wild cluster bootstrap would be more defensible; spatial lag, error and geographically weighted models (Anselin, 1988; LeSage and Pace, 2009) are not estimated. The quantile regressions are estimated on 150,000 listings for computational reasons, and their inter-quantile tests on a 200-draw paired bootstrap; the standardised age coefficient is too imprecise to be interpreted.

The validation exercises have known blind spots. The ten-state holdout withholds 56% of the data and is a pessimistic draw among state partitions, as the grouped and leave-one-state-out results show (median leave-one-state-out R^2 of 0.653 against 0.425). Leave-one-state-out itself is an extrapolation benchmark whose relevance depends on the deployment question; for interpolation within represented markets, blocked validation is pessimistically biased (Wadoux et al., 2021) and the random-split results remain the appropriate benchmark. Finer blocking at the county or metropolitan level, which would trace the continuum between the two extremes, is not possible with the extract's county identifiers (missing for 97% of listings). Hyperparameters were not tuned, and the boosting-round analysis shows that tuning under random resampling would not have helped the extrapolation results; tuning under spatially blocked resampling (Schratz et al., 2019) is left for future work. Finally, SHAP values describe the fitted ensembles' use of the 62 features, including the imputed ones, and their cross-model stability supports the ranking but not an economic interpretation.

9. Conclusion

This paper has compared three approaches to hedonic housing-price analysis—semi-log OLS with progressively finer spatial controls, quantile regression, and gradient-boosted ensembles with SHAP interpretation—on a national sample of 788,842 U.S. Zillow listings, and has subjected the comparison to the validation and inference standards that spatial data require. Each framework answers a different question, and the revision documents the boundaries of each answer more precisely than the previous version could.

The semi-log regression yields average listing-price gradients that align with the meta-analytic record—a living-area elasticity of 0.63, a lot-size elasticity of 0.095 once a coding error is corrected, a 22.6% bathroom premium and a 33.4% foreclosure discount—and the progression from region to ZIP3 fixed effects raises out-of-sample R^2 from 0.630 to 0.725. Two qualifications are new. Cluster-robust standard errors are an order of magnitude larger than the HC3 errors previously reported, and only 33 of 63 coefficients remain significant under state clustering; the casualties are the spatially aggregated proxies and the rare, concentrated amenities. And the pooled coefficients are averages of state price functions whose slopes differ enough that a fully interacted model gains 6.1 points of out-of-sample R^2 over state fixed effects and that four attributes change sign in about

twenty states.

Quantile regression shows that the mean masks distributional heterogeneity: the garage gradient is ten times larger at the 10th than at the 90th conditional percentile, the pool gradient emerges only above the median, and the living-area gradient declines across the distribution. Paired cluster-bootstrap Wald tests confirm these differences for 10 of 13 key attributes, with standard errors several times larger than the independence-based ones reported previously.

Gradient boosting improves prediction from $R^2 = 0.630$ to 0.833 under random validation, a gain whose cluster-bootstrap interval excludes anything below 0.19. That gain does not transfer to markets the model has not seen. Under the ten-state holdout the fully grown model reaches 0.425, the median of 51 leave-one-state-out models is 0.653, and the semi-log regression, once 0.6% of corrupt records are removed, is as accurate out of state (0.539) as the best tree ensemble. The revision traces the collapse to three mechanisms: memorisation of state price levels through region dummies, through tax rates and, above all, through "neighbourhood" variables that are in fact state-median imputations for 93% of listings; a within-state accuracy of 0.62 that is stable across specifications, so that the error is in levels rather than within-market structure; and a capacity effect by which held-out-state accuracy peaks after 75 boosting rounds and declines thereafter while within-training accuracy keeps rising, so that standard early stopping selects an over-fitted model. SHAP rankings are stable across three ensembles, and the third-ranked feature is a state identifier.

The central contribution is methodological. Model evaluation in spatially dependent housing data hinges on validation design, and the random-versus-blocked distinction developed in ecology (Roberts et al., 2017; Ploton et al., 2020; Meyer and Pebesma, 2021; Wadoux et al., 2021) transfers to hedonic economics with three additions that this paper makes concrete: out-of-market error should be decomposed into level and within-market components; imputed and aggregated features should be audited for the geography they encode; and model capacity should be tuned on spatially blocked, not random, evaluation sets. Studies comparing machine learning to hedonic regression on random splits alone will overstate the practical machine-learning advantage for any application involving new markets, and studies reporting conventional robust standard errors on national scraped data will overstate the precision of their gradients.

Several extensions follow. Spatial econometric estimation and spatial HAC or wild-cluster inference would model the dependence we only diagnose; replication on transaction prices would quantify the listing-price wedge; climate-risk variables are a first-order omission; county or metropolitan blocking would map the continuum between interpolation and extrapolation if location identifiers were complete; spatially blocked hyperparameter tuning would establish how much of the extrapolation gap is recoverable; and an observed, rather than imputed, school-quality measure would allow the capitalization question to be asked at all. We regard leave-region-out validation, reported alongside random validation and decomposed into level and within-region accuracy, as a

reasonable default for future comparisons of machine-learning and hedonic models.

The broader lesson is not that machine learning replaces hedonic econometrics, nor the reverse. Prediction within observed markets, prediction in new markets, distributional heterogeneity and economic interpretation are different questions; each of the tools examined here is strong on at most one of them, and housing-market analysis will continue to require all of them—each validated against the question it is meant to answer, and each read with the data's imputation rules in hand.

References

Anglin, Paul M. and Ramazan Gençay, "Semiparametric estimation of a hedonic price function," *Journal of Applied Econometrics*, 1996, *11* (6), 633–648.

Anselin, Luc, *Spatial Econometrics: Methods and Models*, Dordrecht: Kluwer Academic Publishers, 1988.

_ , "Thirty years of spatial econometrics," *Papers in Regional Science*, 2010, *89* (1), 3–25.

Antipov, Evgeny A. and Elena B. Pokryshevskaya, "Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics," *Expert Systems with Applications*, 2012, *39* (2), 1772–1778.

Athey, Susan and Guido W. Imbens, "Machine learning methods that economists should know about," *Annual Review of Economics*, 2019, *11*, 685–725.

Baldauf, Markus, Lorenzo Garlappi, and Constantine Yannelis, "Does climate change affect real estate prices? Only if you believe in it," *Review of Financial Studies*, 2020, *33* (3), 1256–1295.

Bartik, Timothy J., "The estimation of demand parameters in hedonic price models," *Journal of Political Economy*, 1987, *95* (1), 81–88.

Basu, Sabyasachi and Thomas G. Thibodeau, "Analysis of spatial autocorrelation in house prices," *Journal of Real Estate Finance and Economics*, 1998, *17* (1), 61–85.

Bayer, Patrick, Fernando Ferreira, and Robert McMillan, "A unified framework for measuring preferences for schools and neighborhoods," *Journal of Political Economy*, 2007, *115* (4), 588–638.

Bernstein, Asaf, Matthew T. Gustafson, and Ryan Lewis, "Disaster on the horizon: The price effect of sea level rise," *Journal of Financial Economics*, 2019, *134* (2), 253–272.

Bishop, Kelly C., Nicolai V. Kuminoff, H. Spencer Banzhaf, Kevin J. Boyle, Kathrine von Gravenitz, Jaren C. Pope, V. Kerry Smith, and Christopher D. Timmins, "Best practices for using hedonic property value models to measure willingness to pay for environmental quality," *Review of Environmental Economics and Policy*, 2020, *14* (2), 260–281.

Black, Sandra E., "Do better schools matter? Parental valuation of elementary education," *Quarterly Journal of Economics*, 1999, *114* (2), 577–599.

Bogin, Alexander N. and Jessica Shui, "Appraisal accuracy and automated valuation models in rural areas," *Journal of Real Estate Finance and Economics*, 2020, *60* (1–2), 40–52.

Bourassa, Steven C., Eva Cantoni, and Martin Hoesli, "Spatial dependence, housing submarkets, and house price prediction," *Journal of Real Estate Finance and Economics*, 2007, *35* (2), 143–160.

_ , _ , and _ , "Predicting house prices with spatial dependence: A comparison of alternative methods," *Journal of Real Estate Research*, 2010, *32* (2), 139–160.

_ , Martin Hoesli, and Vincent S. Peng, "Do housing submarkets really matter?," *Journal of Housing Economics*, 2003, *12* (1), 12–28.

Breiman, Leo, "Random forests," *Machine Learning*, 2001, *45* (1), 5–32.

Cameron, A. Colin and Douglas L. Miller, "A practitioner's guide to cluster-robust inference," *Journal of Human Resources*, 2015, *50* (2), 317–372.

Campbell, John Y., Stefano Giglio, and Parag Pathak, "Forced sales and house prices," *American Economic Review*, 2011, *101* (5), 2108–2131.

Can, Ayse, "Specification and estimation of hedonic housing price models," *Regional Science and Urban Economics*, 1992, *22* (3), 453–474.

Chen, Tianqi and Carlos Guestrin, "XGBoost: A scalable tree boosting system," in "Proceedings of the 22nd ACM SIGKDD" 2016, pp. 785–794.

Conley, Timothy G., "GMM estimation with cross sectional dependence," *Journal of Econometrics*, 1999, *92* (1), 1–45.

Court, Andrew T., "Hedonic price indexes with automotive examples," in "The Dynamics of Automobile Demand," General Motors Corporation, 1939, pp. 99–117.

Cropper, Maureen L., Leland B. Deck, and Kenneth E. McConnell, "On the choice of functional form for hedonic price functions," *Review of Economics and Statistics*, 1988, *70* (4), 668–675.

Diebold, Francis X. and Roberto S. Mariano, "Comparing predictive accuracy," *Journal of Business & Economic Statistics*, 1995, *13* (3), 253–263.

Dubin, Robin A., "Spatial autocorrelation: A primer," *Journal of Housing Economics*, 1998, *7* (4), 304–327.

Duncan, Dustin T., Jared Aldstadt, John Whalen, Steven J. Melly, and Steven L. Gortmaker, "Validation of Walk Score for estimating neighborhood walkability: An analysis of four US metropolitan areas," *International Journal of Environmental Research and Public Health*, 2011, *8* (11), 4160–4179.

Ekeland, Ivar, James J. Heckman, and Lars Nesheim, "Identification and estimation of hedonic models," *Journal of Political Economy*, 2004, *112* (S1), S60–S109.

Epple, Dennis, "Hedonic prices and implicit markets: Estimating demand and supply functions for differentiated products," *Journal of Political Economy*, 1987, *95* (1), 59–80.

Friedman, Jerome H., "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, 2001, *29* (5), 1189–1232.

Genesove, David and Christopher Mayer, "Loss aversion and seller behavior: Evidence from the housing market," *Quarterly Journal of Economics*, 2001, *116* (4), 1233–1260.

Goodman, Allen C. and Thomas G. Thibodeau, "Housing market segmentation," *Journal of Housing Economics*, 1998, *7* (2), 121–143.

_ and _ , "Housing market segmentation and hedonic prediction accuracy," *Journal of Housing Economics*, 2003, *12* (3), 181–201.

Griliches, Zvi, "Hedonic price indexes for automobiles: An econometric analysis of quality change," in "The Price Statistics of the Federal Government," NBER, 1961, pp. 173–196.

Halvorsen, Robert and Henry O. Pollakowski, "Choice of functional form for hedonic price equations," *Journal of Urban Economics*, 1981, *10* (1), 37–49.

_ and Raymond Palmquist, "The interpretation of dummy variables in semilogarithmic equations," *American Economic Review*, 1980, *70* (3), 474–475.

Han, Lu and William C. Strange, "What is the role of the asking price for a house?," *Journal of Urban Economics*, 2016, *93*, 115–130.

Hill, Robert J., "Hedonic price indexes for residential housing: A survey, evaluation and taxonomy," *Journal of Economic Surveys*, 2013, *27* (5), 879–914.

Ho, Winky K. O., Bo-Sin Tang, and Siu Wai Wong, "Predicting property prices with machine learning algorithms," *Journal of Property Research*, 2021, *38* (1), 48–70.

Horowitz, Joel L., "The role of the list price in housing markets: Theory and an econometric model," *Journal of Applied Econometrics*, 1992, *7* (2), 115–129.

Hu, Lirong, Shenjing He, Zixuan Han, He Xiao, Shiliang Su, Min Weng, and Zhongliang Cai, "Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies," *Land Use Policy*, 2019, *82*, 657–673.

Jones, Michael P., "Indicator and stratification methods for missing explanatory variables in multiple linear regression," *Journal of the American Statistical Association*, 1996, *91* (433), 222–230.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu, "LightGBM: A highly efficient gradient boosting decision tree," in "Advances in Neural Information Processing Systems" 2017, pp. 3146–3154.

Kennedy, Peter E., "Estimation with correctly interpreted dummy variables in semilogarithmic equations," *American Economic Review*, 1981, *71* (4), 801.

Knight, John R., "Listing price, time on market, and ultimate selling price," *Real Estate Economics*, 2002, *30* (2), 213–237.

Koenker, Roger, *Quantile Regression* Econometric Society Monographs, Cambridge: Cambridge University Press, 2005.

_ **and Gilbert Bassett**, "Regression quantiles," *Econometrica*, 1978, *46* (1), 33–50.

_ **and José A. F. Machado**, "Goodness of fit and related inference processes for quantile regression," *Journal of the American Statistical Association*, 1999, *94* (448), 1296–1310.

_ **and Kevin F. Hallock**, "Quantile regression," *Journal of Economic Perspectives*, 2001, *15* (4), 143–156.

Kok, Nils, Eija-Leena Koponen, and Carmen A. Martínez-Barbosa, "Big data in real estate? From manual appraisal to automated valuation," *Journal of Portfolio Management*, 2017, *43* (6), 202–211.

Kuminoff, Nicolai V., Christopher F. Parmeter, and Jaren C. Pope, "Which hedonic models can we trust to recover the marginal willingness to pay for environmental amenities?," *Journal of Environmental Economics and Management*, 2010, *60* (3), 145–160.

_ **, V. Kerry Smith, and Christopher Timmins**, "The new economics of equilibrium sorting and policy evaluation using housing markets," *Journal of Economic Literature*, 2013, *51* (4), 1007–1062.

Lancaster, Kelvin J., "A new approach to consumer theory," *Journal of Political Economy*, 1966, *74* (2), 132–157.

LeSage, James and R. Kelley Pace, *Introduction to Spatial Econometrics*, Chapman and Hall/CRC, 2009.

Liao, Wen-Chi and Xizhu Wang, "Hedonic house prices and spatial quantile regression," *Journal of Housing Economics*, 2012, *21* (1), 16–27.

Lorenz, Felix, Jonas Willwersch, Marcelo Cajias, and Franz Fuerst, "Interpretable machine learning for real estate market analysis," *Real Estate Economics*, 2023, *51* (5), 1178–1208.

Lundberg, Scott M. and Su-In Lee, "A unified approach to interpreting model predictions," in "Advances in Neural Information Processing Systems" 2017, pp. 4765–4774.

_ , Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, 2020, *2* (1), 56–67.

MacKinnon, James G. and Halbert White, "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties," *Journal of Econometrics*, 1985, *29* (3), 305–325.

Mak, Stephen, Lennon H. T. Choy, and Winky K. O. Ho, "Quantile regression estimates of Hong Kong real estate prices," *Urban Studies*, 2010, *47* (11), 2461–2472.

Malpezzi, Stephen, "Hedonic pricing models: A selective and applied review," in Tony O'Sullivan and Kenneth Gibb, eds., *Housing Economics and Public Policy*, Blackwell, 2003, pp. 67–89.

Mayer, Michael, Steven C. Bourassa, Martin Hoesli, and Donato Scognamiglio, "Estimation and updating methods for hedonic valuation," *Journal of European Real Estate Research*, 2019, *12* (1), 134–150.

McMillen, Daniel P., "Changes in the distribution of house prices over time: Structural characteristics, neighborhood, or coefficients?," *Journal of Urban Economics*, 2008, *64* (3), 573–589.

Meyer, Hanna and Edzer Pebesma, "Predicting into unknown space? Estimating the area of applicability of spatial prediction models," *Methods in Ecology and Evolution*, 2021, *12* (9), 1620–1633.

_ and _ , "Machine learning-based global maps of ecological variables and the challenge of assessing them," *Nature Communications*, 2022, *13*, 2208.

Moran, P. A. P., "Notes on continuous stochastic phenomena," *Biometrika*, 1950, *37* (1–2), 17–23.

Moulton, Brent R., "An illustration of a pitfall in estimating the effects of aggregate variables on micro units," *Review of Economics and Statistics*, 1990, *72* (2), 334–338.

Mullainathan, Sendhil and Jann Spiess, "Machine learning: An applied econometric approach," *Journal of Economic Perspectives*, 2017, *31* (2), 87–106.

Murfin, Justin and Matthew Spiegel, "Is the risk of sea level rise capitalized in residential real estate?," *Review of Financial Studies*, 2020, *33* (3), 1217–1255.

Oates, Wallace E., "The effects of property taxes and local public spending on property values," *Journal of Political Economy*, 1969, *77* (6), 957–971.

Pace, R. Kelley and Darren Hayunga, "Examining the information content of residuals from hedonic and spatial models using trees and forests," *Journal of Real Estate Finance and Economics*, 2020, *60* (1–2), 170–180.

Palmquist, Raymond B., "Property value models," in Karl-Göran Mäler and Jeffrey R. Vincent, eds., *Handbook of Environmental Economics*, Vol. 2, Amsterdam: Elsevier, 2005, pp. 763–819.

Park, Byeonghwa and Jae Kwon Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data," *Expert Systems with Applications*, 2015, *42* (6), 2928–2934.

Pérez-Rave, Jorge Iván, Juan Carlos Correa-Morales, and Favián González-Echavarría, "A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes," *Journal of Property Research*, 2019, *36* (1), 59–96.

Pivo, Gary and Jeffrey D. Fisher, "The walkability premium in commercial real estate investments," *Real Estate Economics*, 2011, *39* (2), 185–219.

Ploton, Pierre, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, Nicolas Picard, Vivien Rossi, Carsten F. Dormann, Guillaume Cornu, Gaëlle Viennois, Nicolas Bayol, Alexei Lyapustin, Sylvie Gourlet-Fleury, and Raphaël Pélissier, "Spatial validation reveals poor predictive performance of large-scale ecological mapping models," *Nature Communications*, 2020, *11*, 4540.

Ribeiro, Marco Túlio, Sameer Singh, and Carlos Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," in "Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining" 2016, pp. 1135–1144.

Ridker, Ronald G. and John A. Henning, "The determinants of residential property values with special reference to air pollution," *Review of Economics and Statistics*, 1967, *49* (2), 246–257.

Roberts, David R., Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann, "Cross-validation strategies for data with temporal and spatial dependence," *Ecography*, 2017, *40* (8), 913–929.

Rosen, Sherwin, "Hedonic prices and implicit markets: Product differentiation in pure competition," *Journal of Political Economy*, 1974, *82* (1), 34–55.

Rudin, Cynthia, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, 2019, *1* (5), 206–215.

Schratz, Patrick, Jannes Muenchow, Eugenia Iturritxa, Jakob Richter, and Alexander Brenning, "Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data," *Ecological Modelling*, 2019, *406*, 109–120.

Shapley, Lloyd S., "A value for n-person games," in Harold W. Kuhn and Albert W. Tucker, eds., *Contributions to the Theory of Games II*, Princeton University Press, 1953, pp. 307–317.

Sirmans, Stacy, David Macpherson, and Emily Zietz, "The composition of hedonic pricing models," *Journal of Real Estate Literature*, 2005, *13* (1), 1–44.

Steurer, Miriam, Robert J. Hill, and Norbert Pfeifer, "Metrics for evaluating the performance of machine learning based automated valuation models," *Journal of Property Research*, 2021, *38* (2), 99–129.

Valavi, Roozbeh, Jane Elith, José J. Lahoz-Monfort, and Gurutzeta Guillera-Arroita, "blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models," *Methods in Ecology and Evolution*, 2019, *10* (2), 225–232.

Varian, Hal R., "Big data: New tricks for econometrics," *Journal of Economic Perspectives*, 2014, *28* (2), 3–28.

Wadoux, Alexandre M. J.-C., Gerard B. M. Heuvelink, Sytze de Bruin, and Dick J. Brus, "Spatial cross-validation is not the right way to evaluate map accuracy," *Ecological Modelling*, 2021, *457*, 109692.

Waltl, Sofie R., "Variation across price segments and locations: A comprehensive quantile regression analysis of the Sydney housing market," *Real Estate Economics*, 2019, *47* (3), 723–756.

Waugh, Frederick V., "Quality factors influencing vegetable prices," *Journal of Farm Economics*, 1928, *10* (2), 185–196.

White, Halbert, "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 1980, *48* (4), 817–838.

Zietz, Joachim, Emily N. Zietz, and G. Stacy Sirmans, "Determinants of house prices: A quantile regression approach," *Journal of Real Estate Finance and Economics*, 2008, *37* (4), 317–333.

A. Additional Figures

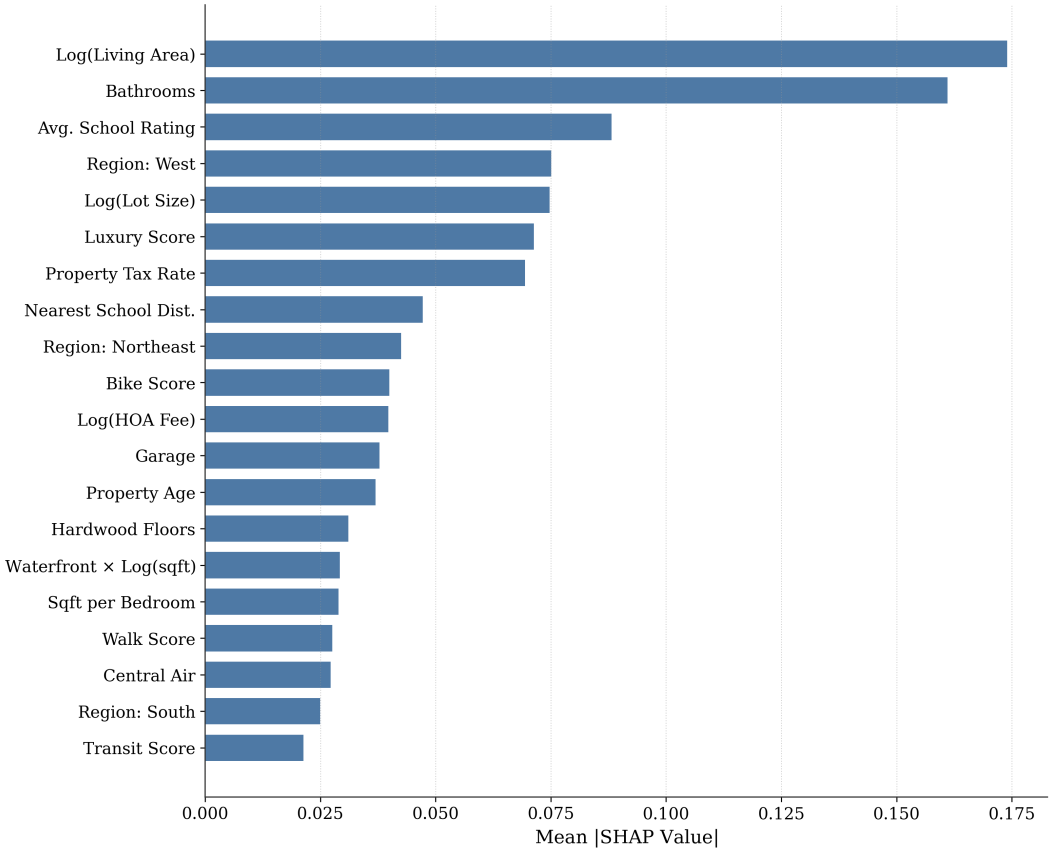


Figure 12. Mean absolute SHAP values, XGBoost, top 20 features. These values are predictive contributions, not causal effects; the school-rating variable is a state median for 93% of listings (Section 3.5).

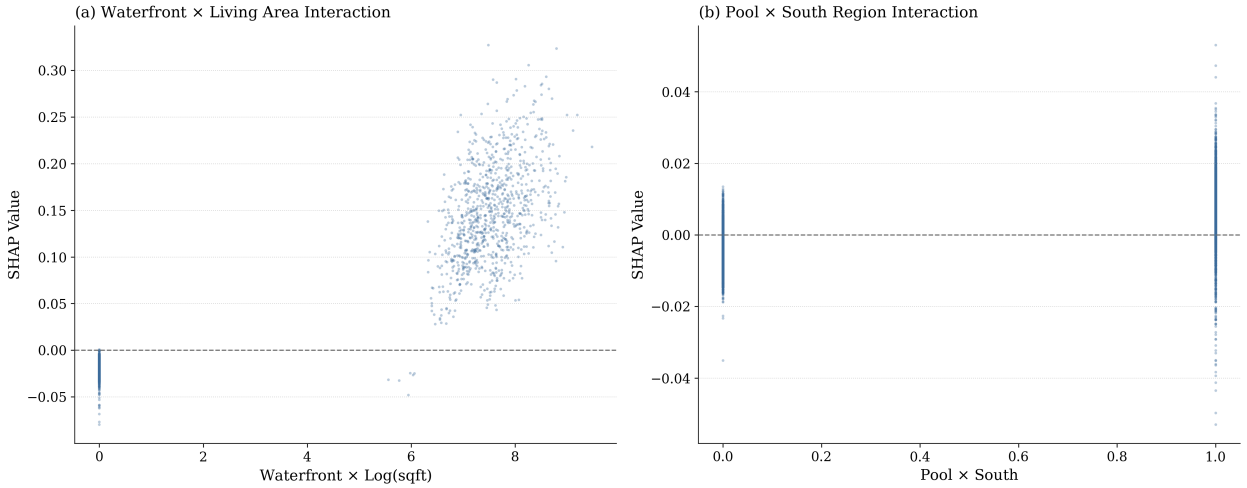


Figure 13. SHAP values for interaction terms: (a) waterfront × ln(living area); (b) pool × South region.

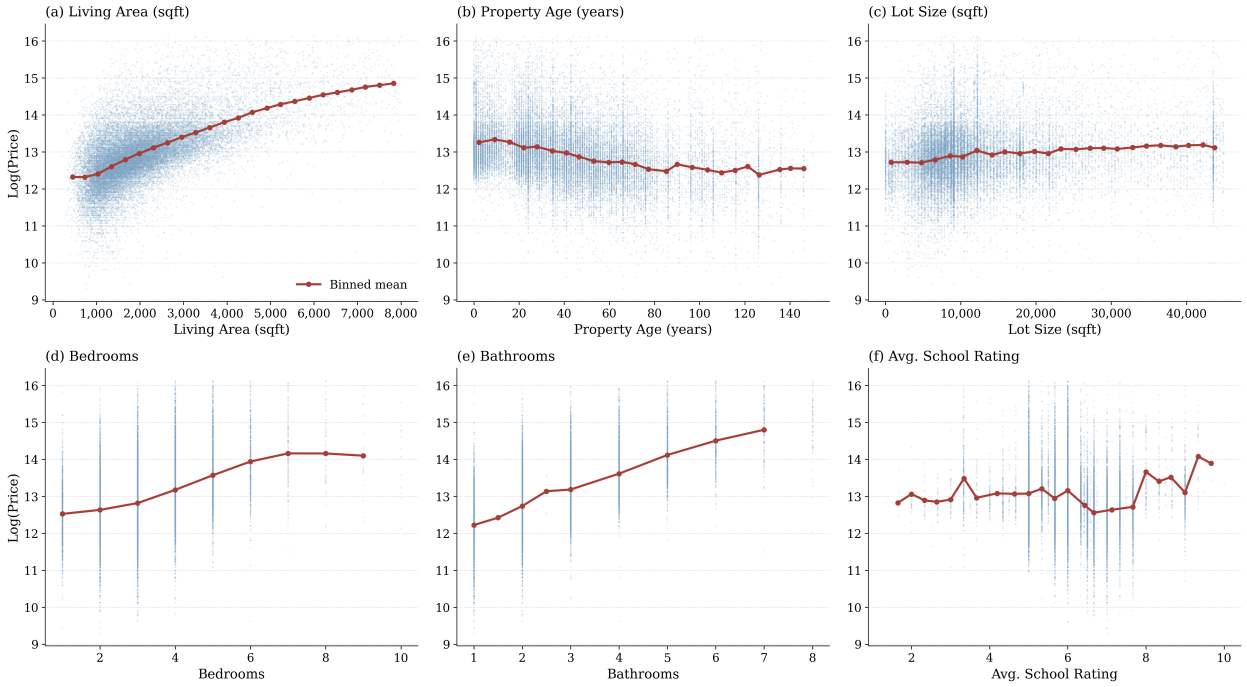


Figure 14. Unconditional bivariate relationships between key attributes and log listing price (raw scatter plots and binned means). They do not control for other attributes and should not be read as conditional effects.

B. Full OLS Coefficient Table

Table 33 reports the complete set of standardised OLS coefficients, including all categorical controls, with HC3 standard errors. Cluster-robust standard errors for the main coefficients are in Table 10; across all 63 coefficients, ZIP3-clustered standard errors are on median 7.6 times the HC3 values and state-clustered standard errors 11.6 times.

Table 33. Full OLS Regression Results Including All Categorical Controls (Standardised)

Variable	Coeff.	SE	Sig.	Variable	Coeff.	SE	Sig.
Constant	12.510	0.010	***	Roof: Metal	0.070	0.008	***
ln(Living Area)†	0.312	0.003	***	Roof: Shingle/Comp	−0.074	0.008	***
Bedrooms†	−0.068	0.002	***	Roof: Tile	−0.110	0.008	***
Bathrooms†	0.236	0.003	***	Roof: Slate	0.008	0.017	
Age†	−0.042	0.014	**	Roof: Other	0.033	0.008	***
Age²†	0.051	0.002	***	Roof: Unknown	0.004	0.008	
Stories†	0.000	0.000		Constr: Concrete/Block	0.269	0.003	***
Bath/Bed Ratio†	−0.021	0.002	***	Constr: Stucco	0.149	0.002	***
Sqft/Bed†	−0.002	0.002		Constr: Stone	0.145	0.005	***
ln(Lot Size)†‡	0.070	0.001	***	Constr: Wood/Frame	0.103	0.002	***
Pool	0.029	0.003	***	Constr: Vinyl	0.012	0.002	***
Spa	0.033	0.003	***	Constr: Steel	0.027	0.014	*
Basement	−0.042	0.002	***	Constr: Other	0.075	0.002	***
Fireplace	−0.002	0.002		Constr: Unknown	0.177	0.002	***
Garage	0.109	0.002	***	Found: Concrete	0.217	0.005	***
Waterfront	−0.135	0.032	***	Found: Crawlspace	0.181	0.006	***
Central Air	0.069	0.001	***	Found: Slab	0.122	0.005	***
Forced Air	−0.023	0.002	***	Found: Pier/Raised	0.161	0.006	***
Hardwood	0.125	0.001	***	Found: Stone	0.032	0.010	**
Parking†§	0.000	0.099		Found: Other	0.144	0.006	***
Luxury†	0.044	0.002	***	Found: Unknown	0.174	0.005	***
Walk†	−0.011	0.001	***	Northeast	0.487	0.003	***
Bike†	0.072	0.001	***	South	0.022	0.003	***
Transit†	0.052	0.001	***	West	0.401	0.003	***
School Rating†¶	−0.049	0.001	***	sqft×Age†	−0.110	0.014	***
School Count†¶	−0.002	0.001	***	Pool×South	0.008	0.001	***
School Dist.†¶	−0.016	0.001	***	WF×sqft†	0.137	0.009	***
Tax Rate†	−0.029	0.001	***	Base×North	−0.025	0.001	***
Condo	−0.037	0.004	***	Condo×Walk†	0.082	0.001	***
HOA	−0.108	0.002	***	Age×Lux†	0.044	0.001	***
ln(HOA+1)†	0.044	0.001	***				
New Constr.	0.103	0.003	***				
Foreclosure	−0.406	0.010	***				

R^2 = 0.634, Adj. R^2 = 0.634, N = 788,842, F = 18,712.5, RMSE = 0.489

Notes: HC3 robust standard errors; significance stars refer to HC3 and overstate precision (Section 5.1.5). †Standardised continuous variables. ‡Published coding, ln(lot) = 0 for 16.5% of listings; see Table 29. §Variance dominated by corrupt entries (Table 4). ¶State median for 93% of listings. Reference categories: Roof = Flat; Construction = Brick; Foundation = Basement; Region = Midwest. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

C. Leave-One-State-Out Results by State

Table 34 reports, for each of the 51 jurisdictions, the accuracy of the full XGBoost model trained on the other fifty (Section 5.6.3). The mean error is the average of predicted minus actual log price in the held-out state, i.e. the level bias of the out-of-state prediction; a value of +0.10 means that the model over-prices the state by about 10%.

Table 34. Leave-One-State-Out Accuracy of the Full XGBoost Model, by Held-Out State

State	N	R^2 (1,000 rounds)	R^2 (100 rounds)	Mean error	MdAPE (%)
UT	6,656	0.770	0.730	+0.003	16.8
VA	15,831	0.764	0.724	−0.095	21.4
NE	3,969	0.754	0.741	+0.029	16.2
KS	6,009	0.754	0.721	+0.044	21.4
NV	10,786	0.751	0.774	−0.119	18.0
AZ	23,980	0.738	0.718	+0.113	22.2
ID	4,537	0.738	0.685	+0.005	18.0
MO	14,833	0.738	0.726	+0.066	19.2
GA	32,035	0.726	0.688	+0.014	18.8
OH	19,486	0.712	0.706	−0.063	20.2
WI	9,148	0.705	0.639	+0.002	19.1
IN	14,367	0.702	0.685	+0.048	17.7
RI	1,056	0.701	0.665	−0.156	19.2
MN	8,723	0.698	0.676	−0.036	21.5
MI	19,258	0.693	0.666	+0.032	23.5
MD	9,706	0.692	0.741	+0.106	24.8
AR	10,866	0.687	0.670	+0.123	20.3
MA	5,438	0.687	0.623	−0.154	25.2
DC	1,414	0.684	0.724	−0.111	23.2
ND	1,344	0.682	0.658	+0.029	20.8
SD	2,070	0.680	0.646	+0.019	19.0
TX	101,804	0.679	0.613	+0.090	22.5
KY	10,766	0.678	0.671	−0.076	18.6
OR	9,524	0.677	0.664	−0.063	19.3
TN	24,658	0.675	0.662	−0.140	21.8
CO	17,566	0.653	0.634	−0.021	22.5
IA	8,287	0.647	0.656	−0.109	22.6
SC	20,396	0.646	0.611	+0.012	21.4
MT	3,974	0.640	0.603	−0.028	25.2
MS	6,799	0.620	0.545	+0.145	22.0
OK	12,319	0.609	0.645	+0.253	28.4
IL	19,274	0.608	0.623	−0.044	30.2
NH	2,390	0.598	0.609	−0.117	21.6
NY	25,095	0.595	0.519	−0.177	40.6
AL	16,966	0.594	0.587	+0.182	28.2
AK	814	0.575	0.592	−0.004	25.3
CT	4,232	0.565	0.586	−0.292	27.1
WY	1,660	0.564	0.524	+0.207	39.9
NC	29,315	0.563	0.558	−0.232	22.9
DE	2,011	0.542	0.447	−0.116	19.7
NM	4,726	0.531	0.498	+0.294	35.8
ME	3,345	0.521	0.496	+0.033	29.8
NJ	7,374	0.521	0.626	−0.250	26.2
VT	1,754	0.443	0.350	−0.311	31.9
FL	133,620	0.437	0.512	−0.356	31.4
WA	11,968	0.417	0.494	−0.282	25.6
PA	15,289	0.353	0.496	+0.456	53.4
WV	3,229	0.277	0.295	+0.481	46.6
LA	15,000	0.248	0.357	+0.421	42.7
HI	5,023	0.095	0.212	−0.503	43.8
CA	48,152	−0.016	0.240	−0.537	40.6

Notes: Each row is a separate model trained on all listings outside the state. Sorted by R^2 at 1,000 rounds. Mean error in log points. Generated by `scripts/05_revision_geo_validation.py`.

D. Robustness Agenda

The following checks and extensions have been completed (✓) or remain open.

- ✓ ZIP3-level fixed effects (Section 5.1.4).
- ✓ Cluster-robust inference by ZIP3 and by state (Section 5.1.5).
- ✓ State-by-state estimation and fully interacted prediction (Section 5.3).

1. County-level and MSA-level fixed effects (county identifiers are missing for 97% of listings in the extract).
2. Spatial lag, spatial error and spatial HAC estimation.

- ✓ Inter-quantile Wald tests with a paired cluster-bootstrap covariance (Section 5.4).
- ✓ Re-estimation excluding imputed observations and with corrected lot-size coding (Sections 6.1–6.2).
- ✓ Missingness indicators in the machine-learning model (Section 5.8).

3. CatBoost benchmark with native categorical handling.
4. Nested cross-validation for hyperparameter tuning, with spatially blocked inner folds (Schratz et al., 2019).

- ✓ SHAP stability across XGBoost, LightGBM and random forest (Section 5.9).

5. Integration of climate-risk data (First Street Foundation or FEMA flood maps).
6. Replication on transaction prices.
7. GAM or spline models to decompose the machine-learning gain into non-linearity versus spatial partitioning.
8. Comparison of the sample distribution with ACS and Census benchmarks by state and price tier.
9. Calibration plots by price decile for all models.

- ✓ Winsorisation sensitivity (Section 6.3).
- ✓ Feature ablation for XGBoost (Section 5.7).
- ✓ XGBoost with and without geographic features and school variables (Section 5.8).
- ✓ Boosting-round sensitivity under geographic holdout (Section 5.6.2).
- ✓ Size-matched random split (Section 5.6).
- ✓ Grouped five-fold and leave-one-state-out cross-validation (Section 5.6.3).
- ✓ Moran's I subsample stability and comparison across estimators (Section 5.2).
- ✓ Quantile-regression stability across ten subsamples (Section 6.4).
- ✓ Bootstrap confidence intervals for test-set R^2 differences (Section 5.5).

E. Reproducibility Checklist

Table 35. Reproducibility Checklist

Item	Status	Notes
Data		
Data source identified	Yes	Zillow active listings, 2025–2026
Sample construction documented	Yes	Table 2
Missing data rates reported	Yes	Table 3, all regressor inputs
Imputation rules documented	Yes	Table 3
Imputation sensitivity tested	Yes	Tables 28, 29, 25
Implausible values documented	Yes	Table 4
Methodology		
Model specifications documented	Yes	Eq. (1)–(10)
Hyperparameters reported	Yes	Section 4.4
Random seeds reported	Yes	Seed 42; QR stability seeds in `src/wp3/config.py`
Validation designs documented	Yes	Section 4.5
Results		
Point estimates reported	Yes	Tables 7–27
Standard errors	Yes	HC3 and cluster-robust (Table 10)
Bootstrap inference	Yes	Tables 16, 18
Effect sizes in natural units	Yes	Table 8
Robustness		
Sensitivity to data quality	Yes	Sections 6.1–6.2
Sensitivity to model choice	Yes	Table 17
Sensitivity to validation design	Yes	Tables 19–22
Cross-model stability	Yes	Table 27
Code and data availability		
Code available	Yes	`src/`, `scripts/01–08`, pinned `requirements.txt`
Upstream sample-construction code	No	Lost before the 2026-08-05 restructuring; stored artefacts are canonical
Data available	On request	Subject to Zillow terms of service