Université du Québec en Outaouais

Département des sciences administratives

Working Paper No. 3

Hedonic Housing Price Models for the United States: Validation Design, Market Price Levels and the 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

November 2026

Version 2.0

Replication code: repository `wp3_uqo` (https://git.spboucher.ai; mirror https://github.com/spboucher-ai/wp3-hedonic-housing-us), branch `review-20261103`: this PDF and the code and results that produced it are committed together as the child of commit `61dd598` (the hash is captured at build time, before that commit exists). Data availability: Appendix E.

Abstract

Whether gradient-boosted trees price homes better than a hedonic regression depends on how the comparison is validated. On 783,872 Zillow asking prices covering every U.S. state, semi-log OLS, quantile regression and boosted ensembles are estimated on one corrected design matrix and compared under seven validation designs with ZIP3-cluster bootstrap intervals. Under random and ZIP-blocked splits XGBoost leads OLS by 13.2–18.7 points of out-of-sample R^2; under state blocks and leave-one-state-out the lead shrinks to 4.8–6.3 points; under a single ten-state holdout the two are indistinguishable. Out-of-state error is dominated by market price levels: twenty local listings raise leave-one-state-out R^2 from 0.635 to 0.699. State-level feature encodings inflate random-split accuracy and transfer only partly. Wild-cluster inference by state leaves 23 of 63 OLS slopes significant, 8 of 14 inter-quantile differences survive multiplicity control, and SHAP rankings are stable across ensembles. Estimates are listing-price associations, not willingness to pay.

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

JEL Classification: R31, C21, C45, C52

1. Introduction

Housing is the largest asset in most American households' portfolios and the collateral behind the largest class of household debt. How attributes map into prices therefore matters for property-tax assessment, mortgage underwriting and the automated valuation models (AVMs) that increasingly mediate transactions. The hedonic framework of Rosen (1974), built on Lancaster (1966) and with empirical roots in Waugh (1928) and Court (1939), remains the canonical way to decompose prices into the implicit valuations of their characteristics. Three questions about its implementation remain open. The semi-log regression imposes a log-linear, additive structure chosen for its robustness to omitted attributes (Cropper et al., 1988) but not for its fidelity to the interactions of real markets. It also restricts implicit prices to be constant across the price distribution, a restriction the quantile-regression literature rejects (Zietz et al., 2008; McMillen, 2008). And gradient-boosted trees, which deliver large accuracy gains in property valuation (Bourassa et al., 2010; Mayer et al., 2019), are opaque (Mullainathan and Spiess, 2017; Athey and Imbens, 2019) and, as this paper shows, owe a design-dependent share of their advantage to the way they are evaluated.

This paper asks a specific version of the third question: how much of the machine-learning advantage over a hedonic regression survives when the model must price homes in markets it has not seen, and what produces the part that does not? The setting is a single extract of 788,842 active Zillow listings covering the fifty states and the District of Columbia, reduced to 783,872 listings once implausible records are screened, with 62 regressors spanning structural attributes, lot size, amenities, neighbourhood measures, market status and six interactions. Three families of models are estimated on one corrected design matrix: semi-log OLS with progressively finer geographic fixed effects, cluster-robust and wild-cluster-bootstrap inference and state-by-state estimation; quantile regression at five quantiles with paired cluster-bootstrap inter-quantile tests under multiplicity control; and gradient-boosted ensembles (XGBoost, LightGBM) with SHAP interpretation (Lundberg and Lee, 2017). The OLS-versus-ensemble comparison is run under seven validation designs—a random split, a size-matched random split, grouped five-fold cross-validation blocked by five-digit ZIP, by three-digit ZIP and by state, leave-one-state-out over the 51 jurisdictions, and a single ten-state holdout—with the number of boosting rounds selected inside the training groups, five seeds per fit and ZIP3-cluster bootstrap intervals on every difference.

The answer is that the comparison depends on the validation design in a way that is monotone in the spatial scale of the blocks. Under the random split XGBoost exceeds OLS by 18.7 points of out-of-sample R^2 (0.833 against 0.6456; 95% ZIP3-cluster interval 17.6 to 19.9). Blocking by five-digit ZIP leaves 16.8 points, blocking by three-digit ZIP 13.2, blocking by state 4.8 (interval 3.5 to 6.2), and leave-one-state-out 6.3 (interval 4.7 to 7.6), with XGBoost ahead in 44 of the 51 held-out states. Only under the single ten-state partition, which withholds 55.6% of the listings

including the five largest states, are the two indistinguishable: the OLS–XGBoost difference is +2.6 points with an interval of $[-0.1, 7.4]$, and LightGBM remains ahead of OLS by 2.4 points there. A tree ensemble's advantage is therefore real for interpolation within observed markets, smaller but present for extrapolation to new states, and absent only in one partition that the other designs identify as an unfavourable draw.

The second finding concerns what the out-of-state error is made of. Decomposing the held-out sum of squared errors into a within-state and a between-state (level) component shows that, under the ten-state holdout, a fully grown 1,000-round XGBoost model leaves 25.0% of the test variance of log prices as state-level error, against 8.5% for OLS; selecting the rounds inside the training states reduces the ensemble's level share to 17.6%, and under leave-one-state-out it is 7.7%. Because the error is in levels, it is correctable with little local information: estimating each held-out state's mean log price from twenty randomly drawn local listings raises leave-one-state-out R^2 for XGBoost from 0.635 to 0.699, 83% of the gain an oracle level correction would deliver (0.712). The ablation shows where the level information comes from. Adding the neighbourhood block to a structural-plus-amenity model raises random-split R^2 by 16.0 points but out-of-state R^2 by only 3.8; the tax rate alone adds 8.6 points in sample and subtracts 7.2 out of state, fifty state dummies add 10.2 in sample and -0.4 out of state, and raw coordinates add 20.0 in sample and subtract 5.1 out of state. The three school variables, which are state medians for 93.0% of listings, behave differently: removing them lowers out-of-state R^2 (0.4901 against 0.5148 under the ten-state holdout with nested rounds), and replacing them with a bare state identifier does not recover the loss. State-level encodings in the features thus inflate random-split accuracy, but they are partly transferable, and the boosting-capacity effect documented below—held-out-state accuracy peaks after about 69 rounds and declines by 12.0 points by round 1,000—runs through the region dummies: without them the curve is flat.

Three papers are closest to this one. Deppner and Cajias (2024) apply spatial cross-validation to random forests and XGBoost on Frankfurt asking rents, benchmark OLS and spatial-autoregressive models, and show that tuning under random resampling selects more boosting rounds than tuning under spatial resampling; the capacity observation of this paper replicates theirs on a national housing sample and locates it in the interaction between rounds and regional dummies. What they do not do, and this paper does, is to run the hedonic regression itself through every blocked design, to attach cluster-bootstrap intervals to the gap at each scale, to separate level from within-market error, and to show that the level error is cheaply correctable. Kolbe et al. (2021) compare hedonic regressions on German listings and transactions and find that asking prices differ from sale prices in composition and in implicit prices; their evidence is why this paper reads every coefficient as a listing-price association and why it does not claim that the listing–transaction wedge affects the three model families identically. Rico-Juan and Taltavull de La Paz (2021) run the same triad—hedonic OLS, quantile regression and a random forest with attribute ranking—on asking prices in one

Spanish city; relative to them this paper adds national scale, spatial-dependence-robust inference for the parametric and quantile models, geographic validation for the ensembles, and a cross-ensemble test of whether SHAP rankings are stable. The spatial-validation methodology itself descends from Brenning (2012), Pohjankukka et al. (2017), Roberts et al. (2017) and Milà et al. (2022); housing economics has used it rarely, and not, to the author's knowledge, with a national leave-one-state-out design that decomposes the error into level and within-market parts. Table 1 summarises the comparison.

Table 1. Position of This Paper Relative to the Three Closest Studies

	Deppner and Cajias (2024)	Kolbe et al. (2021)	Rico-Juan and Taltavull de La Paz (2021)	This paper
Sample	Frankfurt asking rents	German listings and transactions	Asking prices, one Spanish city	U.S. asking prices, every state and DC
Model families	Random forest, XGBoost; OLS and spatial-autoregressive benchmarks	Hedonic regressions	OLS, quantile regression, random forest with attribute ranking	OLS, quantile regression, XGBoost, LightGBM, random forest
Spatially blocked validation of the ensembles	Yes (spatial cross-validation; tuning under random vs spatial resampling)	—	No	Yes (seven designs, incl. leave-one-state-out)
Hedonic regression run through every blocked design, cluster-bootstrap interval on the gap at each scale	No	—	No	Yes
Level vs within-market error; feasible level correction	No	—	No	Yes
Spatial-dependence-robust inference for OLS and quantile gradients	—	—	No	Yes (wild cluster bootstrap; paired ZIP3-cluster bootstrap)
Cross-ensemble stability of SHAP rankings	—	—	No	Yes
Listing–transaction wedge	—	Documented (composition and implicit prices)	—	Read as the reason every estimate is a listing-price association

Notes: Restates the comparison of this section and Section 2.6; "—" marks a dimension outside that study's question as characterised there. No cell adds a claim about the three studies beyond what the text states.

The remaining results discipline the parametric and distributional models. Cluster-robust standard errors by state are on median 11.7 times the HC3 errors and a wild cluster bootstrap by state, which with 51 clusters of very unequal size is the appropriate reference (MacKinnon and

Webb, 2017), leaves 23 of 63 slopes significant at 5% and 9 after Holm's adjustment. State-by-state estimation with ZIP3-cluster standard errors rejects slope homogeneity for every attribute; the living-area elasticity has an inter-decile range of 0.32–0.72 across states, and a fully interacted model gains 6.1 points of out-of-sample R^2 over state fixed effects. Of 14 inter-quantile differences between the 10th and 90th conditional percentiles, 10 are significant by a paired ZIP3-cluster bootstrap and 8 after Romano–Wolf or Holm control. SHAP importance rankings have Spearman correlations of 0.905–0.987 across XGBoost, LightGBM and a random forest, and the third-ranked feature is, for 93.0% of listings, a state identifier. A coding step that set ln(lot) to zero for 16.5% of listings is corrected throughout; the lot elasticity is 0.10 rather than 0.02.

For AVM practice the implication is that the validation design must match the deployment question and that out-of-market error should be reported in its level and within-market parts, because a model that knows a market's level from a few local listings recovers most of its within-market accuracy. For research, random-split comparisons between machine-learning and hedonic models overstate the machine-learning advantage for any question involving new locations, and conventional robust standard errors on national scraped data overstate the precision of hedonic gradients. The paper proceeds as follows: Section 2 positions the study, Section 3 describes the data and the coding audit, Section 4 the estimators and validation designs, Section 5 the results, Section 6 the robustness checks, Section 7 the interpretation, Section 8 the limitations and Section 9 concludes.

2. Related Literature

The paper draws on four literatures: hedonic pricing and its identification, the econometrics of spatial dependence and of distributional heterogeneity, machine learning in property valuation, and the validation of predictive models on spatially dependent data. This section states what each contributes to the design and where the present study departs from the closest prior work.

2.1. Hedonic Pricing and the Interpretation of Gradients

Regressing prices on characteristics predates its theory: Waugh (1928) priced quality attributes of vegetables, Court (1939) built hedonic automobile indexes and Griliches (1961) revived the method for quality-adjusted price measurement, while Ridker and Henning (1967) provided the first regression estimates of how a local disamenity is capitalised into house values. Lancaster (1966) recast goods as bundles of characteristics and Rosen (1974) showed that in a competitive market the equilibrium price schedule $P(\mathbf{z})$ is a double envelope of bid and offer functions whose gradient $\partial P/\partial z_k$ is the marginal implicit price of attribute z_k. Sirmans et al. (2005) synthesise 125 applications and document robust positive gradients for living area, bathrooms and garages and negative gradients for age; Malpezzi (2003) and Palmquist (2005) survey the method's uses and

pitfalls.

Two scope conditions govern how the estimates below are read. First, first-stage gradients are equilibrium price phenomena, not preference parameters; recovering willingness to pay requires a second stage identified by exclusion restrictions or by variation across markets (Bartik, 1987; Epple, 1987; Ekeland et al., 2004). Kuminoff et al. (2013) survey the equilibrium-sorting framework that underpins structural interpretation, and Bishop et al. (2020) distil best practice for credible hedonic estimation: fine spatial controls, robustness to specification and transparent treatment of data limitations. Kuminoff et al. (2010) show in large-scale simulations that specifications with rich spatial fixed effects outperform sparse cross-sections when omitted spatial variables correlate with amenities; the region–state–ZIP3 progression of Section 5.1.3 is the predictive counterpart of that result. Second, Rosen's price function exists within a market in which all participants face the same schedule, and the submarket literature has long shown that metropolitan areas, let alone a country, contain segmented markets with distinct hedonic functions (Goodman and Thibodeau, 1998): estimating separate equations by submarket improves prediction over a pooled equation with location dummies (Bourassa et al., 2003; Goodman and Thibodeau, 2003), and submarket equations and spatial methods are complements (Bourassa et al., 2007). A pooled national regression therefore estimates a listing-weighted average of local price functions; Section 5.3 measures how much that average hides.

The semi-log form became the applied workhorse after Cropper et al. (1988) showed in Monte Carlo experiments that simple forms outperform flexible ones when attributes are omitted or mismeasured, the situation of scraped listing data; Halvorsen and Pollakowski (1981) documented the sensitivity of Box–Cox hedonic estimates, Anglin and Gençay (1996) the gains from semiparametric estimation and Hill (2013) the subsequent evolution of practice. Two technical points are applied directly: the proportional effect of a dummy in a semi-log equation is $\exp(\beta) - 1$ (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). With few and very unequal clusters, cluster-robust standard errors with normal critical values over-reject, and the wild cluster bootstrap is the appropriate reference (Cameron et al., 2008; MacKinnon and Webb, 2017; Webb, 2023); Carter et al. (2017) provide the effective number of clusters that quantifies the problem. Section 5.1.4 applies all three.

2.2. Spatial Dependence and Distributional Heterogeneity

Housing data are spatial data. Can (1992) showed that ignoring spatial autocorrelation biases hedonic estimates, Dubin (1998) explained why residential prices are spatially correlated and Basu and Thibodeau (1998) documented strong residual autocorrelation in Dallas transactions after rich

attribute controls. The formal apparatus of spatial lag, error and Durbin models is developed in Anselin (1988) and LeSage and Pace (2009), with a retrospective in Anselin (2010). This paper does not estimate spatial econometric models; its use of the literature is diagnostic. Moran's I (Moran, 1950) on residuals quantifies the dependence that remains after 62 attributes and regional controls, and the comparison across estimators follows Pace and Hayunga (2020), who use trees and forests to show that hedonic and spatial-model residuals still contain extractable spatial information. A spatial HAC estimator (Conley, 1999) is reported alongside the cluster estimators in Appendix B.

Quantile regression (Koenker and Bassett, 1978; Koenker and Hallock, 2001; Koenker, 2005) lets implicit prices vary across the conditional price distribution. Zietz et al. (2008) introduced it to housing, Mak et al. (2010) documented quantile-varying gradients in Hong Kong, Liao and Wang (2012) combined it with spatial methods and Waltl (2019) provided a comprehensive quantile analysis of Sydney; McMillen (2008) showed that changes in the distribution of house prices are driven largely by coefficients rather than characteristics, so that quantile-specific pricing is economically meaningful. The contribution here is scale and inference: five quantiles on a national sample, stability across ten subsamples, and inter-quantile tests whose covariance comes from a paired ZIP3-cluster bootstrap with step-down control of the family-wise error rate (Holm, 1979; Romano and Wolf, 2005) and of the false discovery rate (Benjamini and Hochberg, 1995).

2.3. Listing Prices

The dependent variable is an asking price. Horowitz (1992) models the list price as a commitment device in seller search, Knight (2002) shows that overpricing lengthens time on market and lowers the eventual sale price, Genesove and Mayer (2001) show that sellers facing nominal losses set higher asking prices and Han and Strange (2016) show that the asking price directs buyer search. Kolbe et al. (2021) test the consequences directly on German data: asking prices stochastically dominate sale prices, mainly because the composition of listed and sold properties differs, implicit prices estimated from listings differ from those estimated from transactions, and listing-based predictions of individual sale prices are biased upward with large error variance. Because the wedge is attribute-dependent, nothing guarantees that a semi-log regression and a tree ensemble absorb it identically, and the cross-framework comparisons below are comparisons of models of *asking* prices. Section 3.2 describes the Zillow valuation available for part of the sample and explains why it does not bound the wedge.

2.4. Capitalisation of Local Public Goods

Several regressors proxy local public goods for which a mature capitalisation literature exists. Oates (1969) initiated the study of property-tax and public-spending capitalisation; Black (1999) used

attendance boundaries to isolate 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) embedded boundary discontinuities in a sorting model, showing that cross-sectional estimates confound school quality with neighbour characteristics. These results discipline the reading of the school-rating and tax-rate coefficients, which in this extract are, in addition, state-level imputations for most listings. Walk Score enters as an accessibility measure; Pivo and Fisher (2011) document a walkability premium in *commercial* real-estate investments, and Duncan et al. (2011) show that the correspondence between Walk Score and underlying walkability constructs varies across metropolitan areas, so the residential premium is not assumed here. The climate-risk literature (Baldauf et al., 2020; Bernstein et al., 2019; Murfin and Spiegel, 2020) identifies price-relevant variables that are entirely absent from the extract.

2.5. Machine Learning in Property Valuation and Its Explanation

Economics has settled on a division of labour in which machine learning excels at prediction and classical methods at parameter estimation (Varian, 2014; Mullainathan and Spiess, 2017; Athey and Imbens, 2019). In real estate this maps onto AVMs: 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 on county data, Kok et al. (2017) describe the shift to big-data valuation, and applied comparisons on random holdouts find tree ensembles ahead of linear hedonic models for sale prices (Pérez-Rave et al., 2019; Ho et al., 2021) and for rents (Hu et al., 2019). Mayer et al. (2019) show that the advantage of boosting persists under re-estimation over time, Bogin and Shui (2020) that AVM accuracy deteriorates where comparable data are sparse, and Steurer et al. (2021) catalogue the metrics appropriate for AVM evaluation; Geerts et al. (2023) review 93 prediction papers and find spatial information under-used. The ensembles used here are random forests (Breiman, 2001), gradient boosting (Friedman, 2001) and its implementations XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017).

Post-hoc explanation of such models rests on SHAP values (Lundberg and Lee, 2017), rooted in Shapley (1953) and computed exactly for trees by TreeSHAP (Lundberg et al., 2020). Their use in housing is now established: Lorenz et al. (2023) derive economically sensible gradients from tree ensembles on European rents, Krämer et al. (2023) rank the determinants of German residential values with XGBoost and SHAP, Iban (2022) interpret tree-based mass-appraisal models and Potrawa and Tetereva (2022) price views with a SHAP-interpreted ensemble; Rico-Juan and Taltavull de La Paz (2021) compare a random forest with attribute ranking, spatial hedonic regression and quantile tools on Alicante asking prices and recommend combining them. The explainability literature itself cautions that local explanations can be unstable (Ribeiro et al., 2016) and that post-hoc

explanation is not a substitute for interpretable modelling in high-stakes settings (Rudin, 2019). In the hedonic context SHAP values are prediction decompositions that share contributions among correlated features and do not satisfy the equilibrium conditions under which Rosen's gradient is an implicit price (Bishop et al., 2020). What the housing applications have not reported is whether importance rankings agree across ensemble families, and what they measure when a feature is an imputed state-level constant; Section 5.8 addresses both.

2.6. Validation Under Spatial Dependence

A methodological literature developed in geostatistics, remote sensing and ecology shows that random cross-validation overstates predictive skill when observations are spatially dependent, because information leaks from training to test folds through proximity. Brenning (2012) introduced k-means spatial cross-validation, Pohjankukka et al. (2017) showed that ordinary cross-validation can be up to 40% more optimistic than spatial k-fold, Karasiak et al. (2022) documented the same pitfall in remote-sensing classification, Roberts et al. (2017) systematised blocking strategies for structured data and Valavi et al. (2019) implemented them; Schratz et al. (2019) showed that hyperparameter tuning is distorted when resampling ignores spatial structure, Ploton et al. (2020) that ecological mapping models with excellent random-validation scores lose most of their skill under spatial validation, and Meyer and Pebesma (2021, 2022) formalised the area of applicability of a spatial prediction model. The prescription is not that blocking is always right: Wadoux et al. (2021) show that when the target is accuracy over the sampled region, an interpolation problem, spatial cross-validation is pessimistically biased and design-based random validation is appropriate, and Milà et al. (2022) propose a leave-one-out scheme that matches the validation geometry to the prediction geometry. The design of this paper follows that prescription: predicting prices in unobserved states is an extrapolation task, blocks of increasing spatial scale trace the continuum between the two extremes, and each design is read against its own deployment question.

Housing economics has begun to use this toolkit. Deppner and Cajias (2024) apply spatial cross-validation to random forests and XGBoost on Frankfurt asking rents, nest it for tuning and assessment, benchmark OLS and spatial-autoregressive models and find that random resampling selects overly complex boosters; Ahmed and Hawkins (2026) reach similar conclusions across two cities. Relative to this work the present paper runs the hedonic regression itself through every blocked design at national scale, attaches cluster-bootstrap intervals to the gap at each spatial scale, decomposes the out-of-state error into level and within-market components and shows that the level component is cheaply correctable, and traces the capacity effect to its interaction with regional dummies. On the data side, Nolte et al. (2024) give guidance for nationwide property data on missing and mismeasured attributes; the coding audit of Section 3.5 applies that guidance to a scraped listing extract and shows that the imputation rules themselves become geographic features.

3. Data and Variable Construction

3.1. Source, Coverage and Availability

The data are a research extract of active for-sale listing pages on Zillow, the largest online real-estate marketplace in the United States, collected by the author into a DuckDB database. No extraction log survives; the collection window is inferred from the data, whose latest listing and price-history dates fall on 8 January 2026, so the snapshot was completed on that day. The raw table holds 839,313 listings described by 116 variables: the asking price, structural characteristics, geocoordinates, accessibility scores, school information, tax records and listing-status tags. Every record carries the status `FOR_SALE`; 87.4% are listed by an agent, 3.4% are new-construction plans, 2.9% new construction, 2.8% of unknown listing type, 1.7% for sale by owner and 0.7% new-construction specification homes. In the analytical sample 85.4% are single-family homes and 14.6% condominiums. The median listing had been on Zillow for 72 days at extraction (interquartile range 35–134).

Zillow's terms of use do not permit redistribution of the extract. The raw database and the derived analysis files are therefore not distributed; they are available from the author for verification, and the repository carries their sizes and SHA-256 checksums, the sample-construction code that rebuilds the analytical sample from the raw table, a frozen lookup for the free-text categorical fields, an entry point that regenerates every exhibit and the inline numbers of this paper, and tests that assert the rebuild against the stored artefacts (Appendix E).

The sample is a snapshot of the supply side of the market, not of the housing stock, and four sources of non-representativeness deserve emphasis. Zillow coverage varies across regions with the completeness of local multiple-listing feeds; active listings over-represent segments with long marketing times; off-market properties, unlisted for-sale-by-owner homes and recent transactions are absent; and the South accounts for 56.6% of listings, the West for 18.9%, the Midwest for 16.1% and the Northeast for 8.4%, with five states (FL, TX, CA, GA, NC) holding 43.7% of the sample. No comparison with Census housing counts was performed, so the paper does not claim national representativeness; the estimates describe the population of active listings in the extraction window.

3.2. Listing Prices versus Transaction Prices

The dependent variable is the asking 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). Kolbe et al. (2021) show on matched German data that the wedge between asking and

sale prices varies with attributes and that implicit prices estimated from listings differ from those estimated from transactions. Throughout the paper the coefficients are therefore read as *listing-price capitalisation gradients*, conditional associations between attributes and asking prices, and the comparisons across model families are comparisons of models of asking prices.

The extract does not contain the sale price of the listed homes, so the wedge cannot be measured. It contains Zillow's own valuation, the Zestimate, for 55.0% of listings. Because Zillow re-anchors the Zestimate to the list price once a home is listed, the ratio of asking price to Zestimate is not a bound on the wedge in either direction and is not used as one; its distribution and a comparison of the observables of listings with and without a Zestimate are reported in Appendix D for completeness. Listings with a Zestimate are newer to the market (standardised mean difference in days on Zillow of −0.21) and less often new construction (−0.25); the two groups are otherwise similar.

3.3. Sample Construction

Table 2 documents the sample. Five sequential filters on price, living area, room counts, coordinates and jurisdiction remove 6.0% of the raw records and leave 788,842 listings in the fifty states and the District of Columbia, 26,477 five-digit ZIP codes and 885 three-digit ZIP prefixes (ZIP3); 16 listings without a ZIP code form an additional ZIP3 group, so the clustering and fixed-effect exercises below use 886 ZIP3 groups. The price window of $10,000 to $10,000,000 excludes land-only and placeholder listings at the bottom and a small number of trophy properties at the top. Exact duplicates on ZIP, living area, bedrooms, bathrooms, year built and price—builder plan listings of the same model home and relisted properties—number 6,787 listings in 3,152 groups (0.86% of the sample); they are kept, and Section 6.3 shows that they do not affect the random-split comparison.

Table 2. Sample Construction from the Raw Zillow Extract

Step	N remaining	Removed	Criterion
Raw Zillow extract	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 and bathrooms	790,799	6,581	$1 \le$ bed ≤ 10, $1 \le$ bath ≤ 10
Valid coordinates	789,199	1,600	Non-missing latitude/longitude
50 states and DC only	788,842	357	Exclude PR, VI
Final analytical sample	788,842	50,471	6.0% of the raw extract removed

Notes: Filters applied sequentially to the 839,313 raw active-listing records (extraction window inferred from the data, see Section 3); the filters act on price, living area, room counts and location only; secondary attributes are screened later (Section 3.6). The sample is rebuilt from the raw DuckDB extract by `scripts/00_build_sample.py` and is bit-identical to the stored v1.1 sample. 6,787 listings (0.86%) belong to 3,152 groups of exact duplicates on ZIP, living area, rooms, year built and price; they are kept and their effect on the random split is reported in Table 31.

3.4. Variables

The specification includes 62 regressors in six blocks, plus an indicator for missing lot size introduced in Section 3.5. *Structural attributes* (eight): the natural logarithm of living area, bedrooms, bathrooms, property age (2026 − year built) and its square, stories, the bathroom-to-bedroom ratio and living area per bedroom. *Lot* (one): the natural logarithm of lot size. *Amenities and quality* (eleven): indicators for swimming pool, spa, basement, fireplace, garage, waterfront location, central air conditioning, forced-air heating and hardwood flooring, the number of parking spaces, and a luxury score (0–7) counting premium features mentioned in the listing. *Neighbourhood* (seven): 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, which takes 243 distinct values across the country. *Market status* (five): condominium, homeowners'-association membership, log annualised HOA fees, new construction and foreclosure. *Interactions and categorical controls* (six and twenty-four): ln sqft × age, pool × South, waterfront × ln sqft, basement × North, condominium × Walk Score and age × luxury score, and dummy sets for roof type (six), construction material (eight), foundation type (seven) and Census region (three, Midwest reference), each categorical set including an "Unknown" category. The free-text roof, construction and foundation fields are mapped to categories by a frozen lookup of the raw strings that is committed with the code. Table 3 reports summary statistics; the median listing is priced at $399,900 (mean $621,737), has 1,820 square feet, 3 bedrooms and 2 bathrooms and is 36 years old, and Walk and Transit Scores are low (medians of 16 and 0), consistent with a stock dominated by suburban and exurban single-family homes.

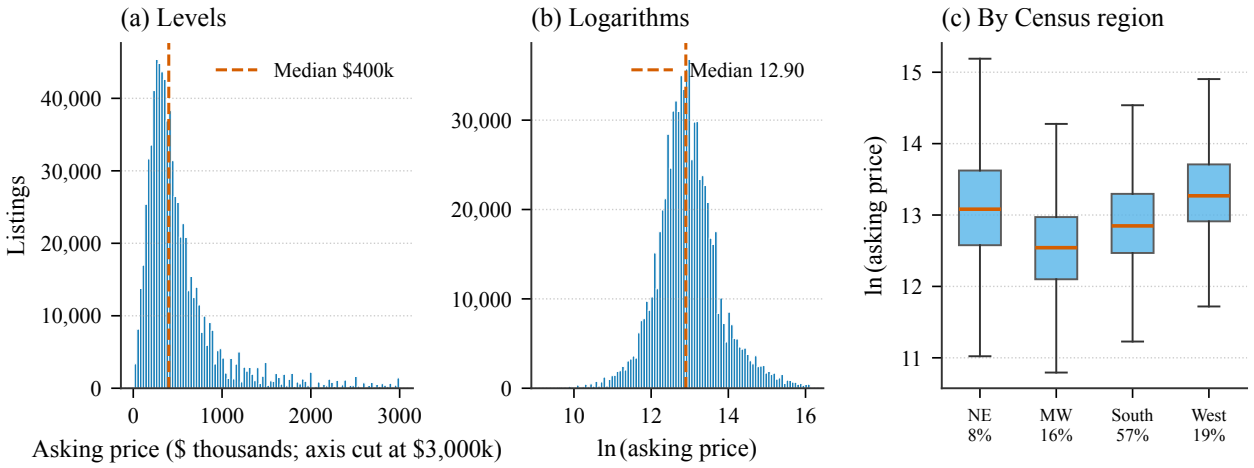


Figure 1. Distribution of asking prices in the analytical sample ($N = 788,842$ active listings; Table 3): (a) levels, histogram of 100 bins truncated at $3,000k for legibility (16,651 listings lie above the cut); (b) natural logarithm, the dependent variable of every model; (c) box plots of log prices by Census region (NE = Northeast, MW = Midwest; median, quartiles, whiskers at 1.5 interquartile ranges, outliers omitted; the share of listings is given under each region). Listing prices are asking prices, not transaction prices.

Table 3. Descriptive Statistics of the Analytical Sample ($N = 788,842$ Active Listings)

Variable	Mean	Std. dev.	Min	Median	Max
Panel A: Continuous variables					
Listing price ($)	621,737	817,965	10,300	399,900	9,999,999
Living area (sq ft)	2,093	1,181	208	1,820	19,600
Bedrooms	3.25	1.09	1.00	3.00	10.00
Bathrooms	2.52	1.16	1.00	2.00	10.00
Property age (years)	42.2	31.5	0.0	36.0	200.0
Lot size (sq ft, imputed field)	252,667	11,907,374	-130,680	10,249	2,011,339,440
Walk Score (0–100)	25.6	26.5	0.0	16.0	100.0
Bike Score (0–100 nominal)	34.9	18.1	1.0	31.0	248.0
Transit Score (0–100)	8.9	18.3	0.0	0.0	100.0
Avg. school rating (1–10)[¶]	6.02	0.90	1.00	6.00	10.00
Property tax rate (%)	1.10	0.53	0.00	1.08	3.73
Luxury score (0–7)	1.36	1.29	0.00	1.00	7.00
Panel B: Binary variables (prevalence, %)					
Swimming pool	39.8				
Garage	67.2				
Fireplace	39.6				
Basement	24.3				
Waterfront	8.9				
HOA member	37.1				
Central air	68.1				
Hardwood floors	26.7				
Condominium	14.6				
New construction	5.9				
Foreclosure	0.5				

Notes: Asking (listing) prices, not transaction prices. Property age is 2026 − year built after state-median imputation; imputed values are included in every row. The lot-size field is the state-median-imputed raw field, whose extreme values (negative, and above 2×10^9 sq ft) motivate the screens of Section 3.6. [¶]State median for the listings whose raw value is missing (Table 4), i.e. a state identifier for most of the sample. Regional shares: see Table 42.

3.5. Missing Data, Imputation and Coding

Scraped listing data are incomplete in ways that are rarely random, and the rules used to fill them determine what the regressors measure. Table 4 reports, for every regressor input, the share of the analytical sample for which the raw field is missing and the rule applied by the rebuilt sample-construction code. Three facts condition the interpretation of the results.

Table 4. Missing Raw Inputs and the Imputation Rule Applied to Each Regressor (Analytical Sample, $N = 788,842$)

Regressor input (raw field)	Missing	Share (%)	Rule applied
avg_school_rating	733,929	93.0	state median
school_count	733,891	93.0	state median
nearest_school_distance	733,891	93.0	state median
transit_score	554,073	70.2	set to 0
hoa_annual_cost (ln HOA)	494,744	62.7	set to 0
has_garage	258,523	32.8	indicator set to 0
roof_type / construction_material / foundation_type	246,005	31.2	"Unknown" category
flooring (hardwood dummy)	234,924	29.8	indicator set to 0
stories	202,087	25.6	state median[a]
parking_spaces	191,492	24.3	state median[a]
year_built (age, age^2, interactions)	148,787	18.9	state median year built (age and its terms)
lot_size_sqft (ln lot)	123,942	15.7	v1.1: ln(lot) = 0; v2: ln(state-median lot) + indicator[b]
cooling_type (central-air dummy)	89,598	11.4	indicator set to 0
heating_type (forced-air dummy)	65,345	8.3	indicator set to 0
bike_score	47,058	6.0	state median
walk_score	24,136	3.1	state median

Notes: Missingness measured in the raw extract for the 788,842 retained listings; rules are those actually applied by the rebuilt sample-construction code (`scripts/00_build_sample.py`), not the rules the v1.1 text described. "State median" is the median of observed values within the listing's state. [a]The v1.1 text said "set to 1" (stories) and "set to 0" (parking); the stored sample uses state medians. [b]In the v2 design ln(lot) is the log of the state-median lot for the 130,271 listings with a missing, zero, one-square-foot or negative lot (16.5%), with a lot-missing indicator (Table 5); parking spaces are winsorised at 20 and stories at 10. Climate-risk fields are missing for every record and excluded.

First, the three school variables are missing for 93.0% of listings and imputed with state medians. For those listings "average school rating" takes one of 51 values and identifies the state rather than the quality of nearby schools; the imputed school count is the same in every state. Any coefficient or importance measure attached to these variables is in large part a state-level location effect, a point that governs the reading of the negative school-rating coefficient (Section 5.1.2), of the ablation (Section 5.7.3) and of the SHAP ranking (Section 5.8). Second, year built is missing for 18.9% of listings and replaced by the state median, so the age distribution carries a spike at each state's median age. Third, lot size is missing for 15.7% of listings, recorded as zero for 6,115, as one square foot for 211 and as a negative number for 3 (Table 5). Although the lot-size field was state-median

imputed, the regressor that entered the earlier estimates, ln(lot), was computed before imputation and equalled zero for all 130,271 such listings (16.5% of the sample), a de facto missing-value code far below the mass of the data (the median imputed lot is 10,249 square feet). The design matrix used throughout this paper replaces that value by the log of the state-median lot and adds an indicator for the affected listings; Section 6.1 quantifies the consequences, which are a fivefold change in the lot elasticity.

Table 5. Reconciliation of the Listings Coded ln(lot) = 0 in the v1.1 Design

Class of raw lot-size value	Listings	Share of sample (%)
Raw lot size missing	123,942	15.71
Raw lot size recorded as 0	6,115	0.78
Raw lot size of 1 sq ft (ln = 0)	211	0.03
Raw lot size strictly between 0 and 1	0	0.00
Raw lot size negative	3	0.00
Total with ln(lot) = 0 in the v1.1 coding	130,271	16.51
Check: sum of the five classes	130,271	16.51

Notes: Analytical sample, $N = 788,842$. The v1.1 text reported 6,329 zero lots in one place and 6,115 in another: the raw field holds 6,115 zeros, 211 lots of exactly one square foot and 3 negative lots, all of which (together with the 123,942 missing values) received ln(lot) = 0 in the v1.1 design. The v2 design replaces these 130,271 values by the log of the state-median lot and adds a lot-missing indicator (Table 29).

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; the HOA fee is set to zero when missing (62.7%); and the central-air, forced-air and hardwood indicators are derived from free-text fields missing for 8.3–29.8% of listings. Setting a missing indicator to zero is a special case of the missing-indicator method whose bias properties Jones (1996) analyse: the coefficient mixes the attribute effect with the effect of reporting the attribute. These coefficients are treated as descriptive, and six missingness indicators (year built, lot size, school rating, Walk Score, garage, Transit Score) are added as an optional block in the machine-learning variants of Section 5.6.

3.6. Data Quality and the Trimmed Sample

The sample filters screen price, size, room counts and location but not secondary attributes. The number of parking spaces exceeds 20 for 1,133 listings and 1,000 for 67; stories exceed 10 for 1,607 listings; 422 lots exceed one square mile; 121 listings report a year built after 2026 and 1,838 a year built before 1826; Bike Score exceeds its nominal maximum of 100 for 38 listings. Table 6 lists these counts. Four screens—parking spaces above 20, stories above 10, lots above 640 acres and an

age at the 200-year cap—flag 4,970 listings (0.63% of the sample). Linear models cannot tolerate these records: in the original design the parking variable had a variance dominated by a handful of corrupt entries, and under a geographic holdout the linear extrapolation on them produced an out-of-state R^2 of −1,956,000 in which 1,856 held-out listings with absolute errors above two log points carried 100% of the squared error. Tree ensembles are unaffected because they split on ranks. The design matrix therefore winsorises parking spaces at 20 and stories at 10 for every model, and the *trimmed sample* of 783,872 listings that removes the 4,970 flagged records is the estimation sample for every model in the paper; the full sample of 788,842 listings is retained only where the text says so.

Table 6. Implausible Attribute Values and Coding Artefacts in the Analytical Sample

Check	Listings	Share (%)
Parking spaces > 20[s]	1,133	0.14
Parking spaces > 1,000	67	0.01
Stories > 10[s]	1,607	0.20
Lot size > 1 square mile (640 acres)[s]	422	0.05
Lot size > 100 acres	2,568	0.33
Lot size recorded as 0 (raw field)	6,115	0.78
ln(lot) coded 0 in v1.1 (missing, zero, 1 sq ft or negative lot)	130,271	16.51
Year built > 2026 (age set to 0)	121	0.02
Year built < 1826 (age capped at 200)[s]	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 ($N = 788,842$), computed by `scripts/06_revision_inference.py` from the stored sample. [s]The four screens that define the trimmed sample used by every v2 estimate (parking spaces > 20, stories > 10, lot > 640 acres, age = 200) remove 4,970 listings (0.63%), leaving $N = 783,872$. In the v2 design the remaining parking and story counts are winsorised at 20 and 10.

4. Econometric and Machine-Learning Methodology

4.1. Identification, Prediction and Interpretation

None of the models estimated here 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 schedule (Rosen, 1974); recovering preferences requires a second stage identified by exclusion restrictions or by variation across markets (Bartik, 1987; Epple, 1987; Ekeland et al., 2004). The data pool listings from 51 jurisdictions and thousands of local markets, so a pooled

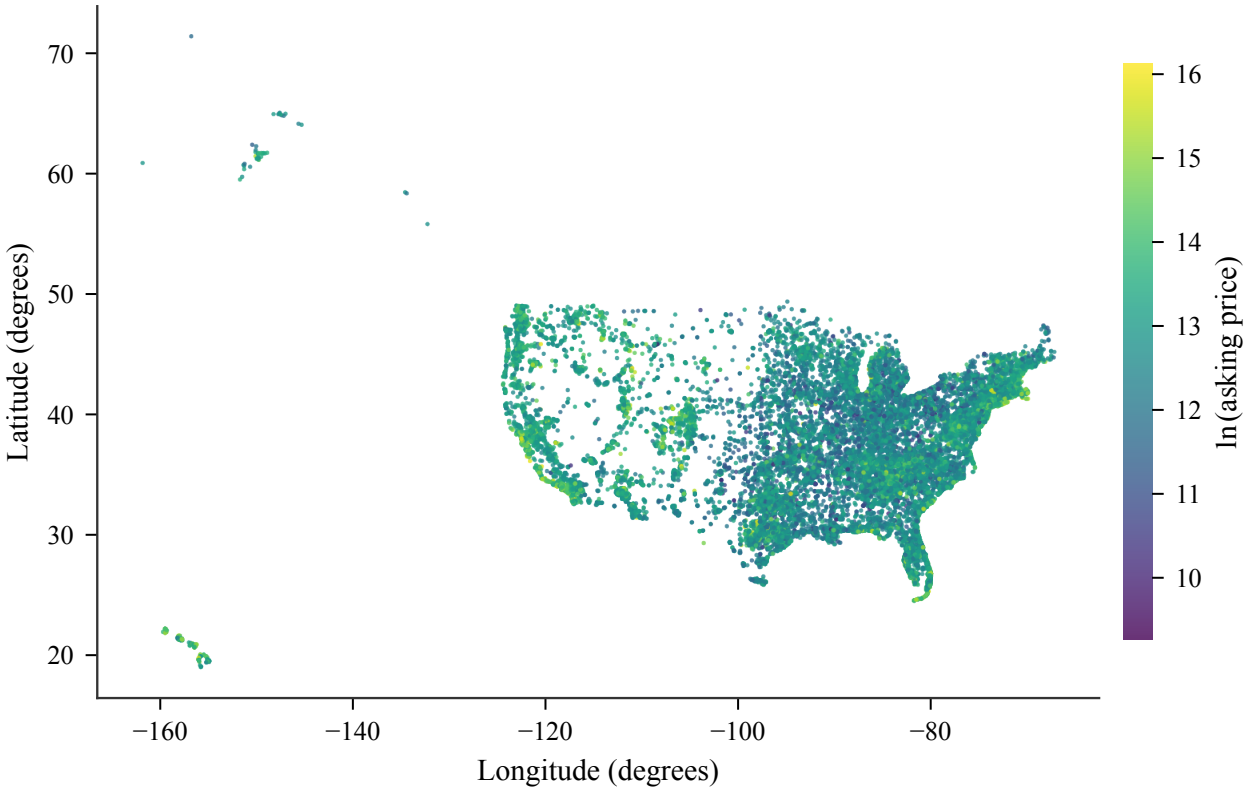


Figure 2. Log asking prices of a 50,000-listing random subsample of the analytical sample by location (viridis scale; plate carrée projection, aspect ratio 1.25). The coastal gradient and the uneven listing density are both visible: sparse areas reflect few Zillow listings in the extract, not few homes.

regression 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 7 summarises these distinctions; Section 5.3 measures how much the pooled coefficients hide by estimating the price function state by state.

Table 7. What Each Estimate Measures: OLS, Quantile Regression and SHAP

	OLS coefficient	Quantile coefficient	SHAP value
Object	Conditional-mean association	Conditional-quantile association	Contribution to one model's prediction
Reading	Semi-elasticity or elasticity (natural units); effect of one SD (standardised)	Quantile-specific semi-elasticity	Attribution of the fitted prediction, in log-price units
Causal?	No, unless separately identified	No, unless separately identified	No
Structural willingness to pay?	Only under single-market and exclusion assumptions	Same	No
Inference	Cluster-robust and wild-cluster bootstrap by state (Table 9)	Paired ZIP3-cluster bootstrap (Table 15)	Descriptive; stability across ensembles (Table 27)
Primary use	Average gradients and their precision	Distributional heterogeneity	Interpretation of the prediction technology

Notes: Conceptual summary of Section 4. All three objects are conditional associations in asking prices on the same design matrix (62 regressors plus a lot-missing indicator); none identifies willingness-to-pay parameters (Section 4.1).

4.2. Semi-Log Hedonic Specification

The baseline model is

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

where P_i is the asking price of listing i, x_{ik} are continuous and binary attributes (including the lot-missing indicator), d_{ij} are categorical dummies, $x_{im_1} x_{im_2}$ are the six interactions 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 applied workhorse because it is robust to omitted and mismeasured attributes (Cropper et al., 1988; Halvorsen and Pollakowski, 1981), the situation of scraped listing data. Two versions are reported: in the standardised version every continuous regressor is scaled to zero mean and unit variance, so that β_k is the association between a one-standard-deviation change in x_k and log price and coefficients are comparable across units; in the

natural-unit version the coefficient on a logged regressor is an elasticity and the coefficient on a regressor in natural units is a semi-elasticity. Binary regressors are never standardised. For a dummy the proportional effect is not β_k but

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

the exact transformation of Halvorsen and Palmquist (1980) and the approximately unbiased estimator of Kennedy (1981). The Kennedy correction is negligible under HC3 variances but not under the cluster-robust variances adopted below for imprecise dummies such as waterfront; Appendix B reports both.

Inference. Heteroskedasticity-robust HC3 standard errors (White, 1980; MacKinnon and White, 1985) treat listings as independent. Residuals of a hedonic regression on spatial data are not independent, and several regressors (tax rate, the imputed school and lot variables, region) vary only at the state or ZIP level, so the Moulton problem applies (Moulton, 1990). Cluster-robust standard errors (Cameron and Miller, 2015),

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

are reported with the 886 ZIP3 groups and the 51 states as clusters. Eq. (3) is consistent as $C \to \infty$; with 51 state clusters whose sizes range from 814 (Alaska) to 133,620 (Florida), the state-clustered t-statistics with normal critical values over-reject rather than under-reject (MacKinnon and Webb, 2017). Three devices address this. The effective number of clusters of Carter et al. (2017) is computed for every coefficient; a wild cluster bootstrap by state with the six-point weights of Webb (2023) and with Rademacher weights, 9,999 replications, imposing the null (Cameron et al., 2008; Roodman et al., 2019), gives the reference p-values; and significance counts over the 63 slopes are reported unadjusted and after the step-down adjustment of Holm (1979) and the false-discovery-rate adjustment of Benjamini and Hochberg (1995). A spatial HAC estimator (Conley, 1999) with Bartlett kernels of 25, 50 and 100 km, computed on a 100,000-listing subsample, is reported in Appendix B as a check that dependence decaying with distance gives standard errors between the HC and the state-cluster values.

4.3. Quantile Regression

Quantile regression (Koenker and Bassett, 1978; 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 $\rho_\tau(u) = u\,(\tau - \mathbb{1}\{u < 0\})$. Eq. (4) is estimated at $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$ with the full standardised design on a random subsample of 150,000 listings of the trimmed sample (seed 42), because each interior-point fit on the full sample is costly; convergence of every fit is reported, and stability is checked by repeating the median regression on ten independent subsamples (Section 6.2). Goodness of fit is the pseudo-R^2 of Koenker and Machado (1999), $R^1(\tau) = 1 - \hat{V}(\tau)/\tilde{V}(\tau)$, the proportional reduction in check-function loss relative to a constant-only model.

Inter-quantile tests. Whether a gradient differs between the tails is tested as $H_0\colon \beta_k(0.10) = \beta_k(0.90)$ with

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

The two estimates are computed on the same sample and are positively correlated, and the observations are spatially dependent, so the denominator is obtained from a paired bootstrap in which both quantile regressions are re-estimated on 999 resamples of the 150,000 listings drawn by ZIP3 cluster (whole ZIP3 areas with replacement); an observation-level paired bootstrap and the independence formula are reported for comparison. Inference is based on the percentile interval of the bootstrap distribution of the difference, and the family of 14 tests is controlled by Holm's procedure, by the Benjamini–Hochberg procedure and by the step-down method of Romano and Wolf (2005), which uses the joint bootstrap distribution. The same cluster bootstrap provides ZIP3-cluster standard errors for every coefficient at $\tau = 0.10$ and 0.90, which replace the independence-based kernel standard errors in the inference reported.

4.4. Gradient-Boosted Ensembles and Benchmarks

XGBoost (Chen and Guestrin, 2016) implements gradient boosting (Friedman, 2001): an additive ensemble of regression trees $f_1, \ldots, f_T$ 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. The hyperparameters are maximum depth 8, learning rate 0.05, row and column subsampling 0.8, minimum child weight 5, $\alpha = 0.1$ and $\lambda = 1$, values that are standard for tabular data of this size and were not tuned; the penalty terms of Eq. (6) are the only regularisation beyond the number of rounds. The number of rounds, the hyperparameter that governs capacity, is treated in three ways. Under the *nested* rule it is selected by an inner three-fold grouped cross-validation on the training groups of the design at hand (ZIP5, ZIP3

or state blocks for the blocked designs; the training states for the ten-state and leave-one-state-out designs; random folds for the random designs), so that no held-out observation informs the choice. Fixed rules of 100 and 1,000 rounds are reported alongside, the latter with early stopping after 50 rounds without improvement on a 10% random subset of the training data. Every boosted fit is repeated with five seeds (42–46) and reported as a mean with its seed standard deviation. LightGBM (Ke et al., 2017) is fitted with matching hyperparameters and the nested rule. A random forest (Breiman, 2001) with 500 trees, one third of the features per split and a minimum leaf of five, Ridge, Lasso and Elastic Net with penalties chosen by five-fold cross-validation, and OLS with state and ZIP3 fixed effects serve as benchmarks under the random split.

Evaluation. On the log scale the paper reports R^2, the root mean squared error and the mean absolute error; on the dollar scale, following Steurer et al. (2021), the mean and median absolute percentage error,

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

The difference in R^2 between OLS and each ensemble is the quantity of interest under every design. Its sampling uncertainty is obtained by resampling whole ZIP3 clusters of the evaluated listings 1,000 times and recomputing both R^2 values on each draw, which gives a 95% percentile interval that allows for spatial dependence in the test set; under the random split a paired test on squared-error differences with ZIP3-cluster standard errors, in the spirit of Diebold and Mariano (1995), is reported as well. Under leave-one-state-out the 51 paired state-level differences are summarised by a sign test and a Wilcoxon signed-rank test.

4.5. Validation Designs

Because listing prices are spatially dependent, the validation protocol is a substantive choice. A random split assesses interpolation within the markets represented in the training data; a spatially blocked split assesses extrapolation to markets the model has not seen (Roberts et al., 2017; Meyer and Pebesma, 2021); for interpolation objectives blocked validation is pessimistically biased (Wadoux et al., 2021). Seven designs are therefore reported, all on the trimmed sample, and each is read against its own deployment question.

Random split. An 80/20 split with seed 42 (627,102 training and 156,770 test listings). Nearly every test listing has training neighbours in its own ZIP code.

Size-matched random split. A random split with the training-set size of the ten-state design (348,298 training listings), which separates the effect of training-set size from the effect of extrapolation.

Grouped five-fold cross-validation by ZIP5, by ZIP3 and by state. The 26,477 five-digit ZIP

codes, the 886 ZIP3 groups and the 51 jurisdictions are each assigned to five folds (`GroupKFold`); each fold is predicted by a model trained on the other four, so that every listing is predicted once by a model that saw no listing from its own block, and the pooled out-of-fold predictions are scored. The three block sizes trace the continuum from neighbourhood-scale to state-scale extrapolation; fold membership is listed in Appendix C.

Leave-one-state-out. For each of the 51 jurisdictions a model is trained on the other fifty and evaluated on the omitted one, yielding a distribution of out-of-state accuracy over 51 markets, each predicted from the largest possible training set. Both the pooled R^2 over all out-of-state predictions and the distribution of state-level R^2 values are reported, with the median and the listing-weighted mean.

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 (348,319 listings) and evaluated on the 435,553 held-out listings (55.6% of the sample). This single partition confounds extrapolation with a much smaller training set and is one draw from a large set of possible partitions; it is retained because it is the harshest of the seven designs and because the other designs locate it within the range of outcomes.

Level and within-market error. Pooled out-of-state R^2 mixes the error in predicting the price *level* of a market the model has never seen with the error in predicting *within-market* price differences. The sum of squared errors decomposes exactly into a within-state part and a between-state part, $\mathrm{SSE} = \sum_i [(e_i - \bar{e}_{s(i)})]^2 + \sum_s n_s \bar{e}_s^2$, where $e_i = \hat{y}_i - y_i$ and $\bar{e}_s$ is the mean error in state s; both parts are reported as shares of the test variance of log prices. The *oracle 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, is the accuracy of an analyst who knew each held-out market's average price level exactly; it is an upper bound, not an achievable accuracy, and is labelled as such throughout. The achievable counterpart, the *feasible level correction*, estimates each held-out state's mean error from k listings drawn at random from that state ($k \in \{5, 20, 50, 100, 200\}$, 100 draws), shifts the remaining predictions by that estimate and scores them; it is reported with the interquartile range over draws.

Capacity, geography and ablation. The number of boosting rounds is varied over a grid from 25 to 1,000 under the ten-state design and scored on the held-out states and on the early-stopping subset of the training states, for each of the five seeds. A 2×2 factorial removes the region dummies and the three school variables from the full model, separately and jointly, under each rounds rule and each geographic design. A nested ablation adds feature blocks to XGBoost in the order structural,

lot, amenities, neighbourhood, market status and full model, and the neighbourhood block is split into its components (tax rate alone; latitude and longitude; fifty state dummies; the school block replaced by a state identifier; the block without Walk, Bike and Transit Scores), each under the random and the ten-state design with five seeds.

4.6. SHAP Interpretation and Spatial Diagnostics

SHAP values (Lundberg and Lee, 2017), rooted in 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. Eq. (9) is evaluated on 9,937 random test listings of the random split and global importance is the mean absolute contribution $\overline{|\phi_k|}$. For XGBoost and LightGBM the values are exact; for the random forest, on which exact TreeSHAP over 500 unpruned trees is intractable, the Saabas path-attribution approximation is used and the comparison is labelled accordingly. Pairwise Spearman rank correlations of $\overline{|\phi_k|}$ across the three ensembles measure whether the ranking is an artefact of one algorithm. SHAP values are model-dependent prediction decompositions that share contributions among correlated features and are not Rosen's gradient (Rudin, 2019; Lorenz et al., 2023); they describe which features the fitted ensembles rely on.

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

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

is computed with a row-standardised eight-nearest-neighbour weight matrix on the full random test set of 156,770 listings, for the residuals of OLS, LightGBM and XGBoost, with significance from 999 permutations. On the full test set the median distance to the eighth neighbour is 2.6 km, so the statistic measures neighbourhood-scale dependence; binary distance-band weights at 0–1, 1–5 and 5–25 km are reported to show how the dependence decays with distance.

5. Estimation Results

5.1. The Semi-Log Regression

5.1.1. Baseline Estimates

Table 8 reports the baseline of Eq. (1) on the trimmed sample of 783,872 listings for the main regressors in standardised and natural units with state-clustered inference and economic magnitudes; Table 9 places the HC3, ZIP3-cluster and state-cluster standard errors of the same coefficients side by side, and the complete coefficient table is Table 34 in Appendix A. The model has $R^2 = 0.649$ and a root mean squared error of 0.478 log points; the F-statistic of 18,943 rejects jointly zero slopes, which with this sample size is uninformative. The residuals have excess kurtosis (5.38) and mild positive skewness (0.12); the Jarque–Bera statistic is 187,380; Figure 3 shows the out-of-sample residual diagnostics on the random test set, whose fan shape motivates the robust and cluster-robust inference below.

In natural units the elasticity of the asking price with respect to living area is 0.59: a 10% larger home is listed at about 5.9% more, holding the other attributes constant. The lot elasticity is 0.097; under the original coding that set ln(lot) to zero for 16.5% of listings it was 0.018, an understatement by a factor of 5.5 analysed in Section 6.1. Each additional bathroom is associated with a 24.8% higher asking price and each additional bedroom, conditional on total area, with a 6.2% lower price, the familiar room-count versus room-size trade-off (Sirmans et al., 2005). The garage differential is 12.2% and the foreclosure discount 32.3%, by Eq. (2); the latter exceeds the 27% forced-sale discount that Campbell et al. (2011) estimate from Massachusetts transactions, as expected if distressed listings are positioned below market before negotiation. The Kennedy correction changes these percentages by less than 0.1 point under HC3 variances; under state-cluster variances it matters only for waterfront (−21.4% exact against −28.9% corrected; Table 36).

5.1.2. Coefficient Signs That Are Not Implicit Prices

Several coefficients have signs that a reader expecting implicit prices would not anticipate, and the coding audit of Section 3.5 explains most of them. The school rating is a state median for 93.0% of listings, so its coefficient (−5.2% per rating point) compares states with higher median ratings to states with lower ones, conditional on the region dummies and the largely state-level tax rate; it says nothing about the valuation of nearby schools, which boundary-discontinuity designs consistently find positive (Black, 1999; Bayer et al., 2007), and under state clustering its t-statistic is −2.01. The waterfront main effect is negative because the indicator is interacted with standardised ln(sqft) and the interaction is strongly positive, so the main effect is the premium at a point that is not an economically meaningful property; its state-cluster standard error is 0.45, as expected for an attribute

Table 8. OLS Hedonic Regression of Log Asking Price on the v2 Design: Standardised and Natural-Unit Coefficients with State-Clustered Inference

	Standardised regressors			Natural units		
Variable	Coef.	(SE)	p, wild	Coef.	(SE)	Economic magnitude
Structural attributes						
ln(living area)	0.291***	(0.033)	0.000	0.5868	(0.0663)	elasticity 0.587
Bedrooms	−0.070**	(0.028)	0.013	−0.0643	(0.0259)	−6.43% per unit
Bathrooms	0.256***	(0.023)	0.000	0.2213	(0.0203)	+22.13% per unit
Age (years)	0.012	(0.115)	0.931	0.0004	(0.0038)	+0.04% per unit
Age^2	0.044*	(0.021)	0.061	1.2×10^{-5}	(5.6×10^{-6})	quadratic term
Lot						
ln(lot size)	0.128***	(0.017)	0.000	0.0968	(0.0128)	elasticity 0.097
Lot missing (indicator)	−0.023	(0.020)	0.319	−0.0230	(0.0202)	−2.3% vs. reference
Amenities						
Pool	0.068	(0.052)	0.491	0.0685	(0.0521)	+7.1% vs. reference
Garage	0.115***	(0.016)	0.000	0.1148	(0.0160)	+12.2% vs. reference
Waterfront	−0.240	(0.450)	0.958	−0.2405	(0.4495)	−21.4% vs. reference
Central air	0.083***	(0.022)	0.003	0.0835	(0.0222)	+8.7% vs. reference
Hardwood floors	0.123***	(0.018)	0.000	0.1233	(0.0177)	+13.1% vs. reference
Luxury score (0–7)	0.004	(0.009)	0.650	0.0033	(0.0069)	+0.33% per unit
Neighbourhood						
Walk Score	0.037*	(0.019)	0.096	0.0014	(0.0007)	+0.14% per unit
Bike Score	0.079**	(0.027)	0.015	0.0044	(0.0015)	+0.44% per unit
Transit Score	0.054***	(0.012)	0.001	0.0030	(0.0007)	+0.30% per unit
Avg. school rating¶	−0.048	(0.024)	0.124	−0.0529	(0.0263)	−5.29% per unit
Property tax rate (%)	−0.019	(0.027)	0.582	−0.0359	(0.0511)	−3.59% per unit
Market status						
Condominium	−0.093	(0.054)	0.408	−0.0933	(0.0544)	−8.9% vs. reference
HOA member	−0.066	(0.029)	0.102	−0.0658	(0.0287)	−6.4% vs. reference
New construction	0.071**	(0.030)	0.030	0.0711	(0.0302)	+7.4% vs. reference
Foreclosure	−0.390***	(0.036)	0.000	−0.3895	(0.0363)	−32.3% vs. reference
Census region (reference: Midwest)						
Region: Northeast	0.477**	(0.124)	0.014	0.4765	(0.1238)	+61.0% vs. reference
Region: South	0.035	(0.062)	0.649	0.0351	(0.0616)	+3.6% vs. reference
Region: West	0.419***	(0.104)	0.008	0.4195	(0.1042)	+52.1% vs. reference
N	783,872					
R^2	0.649					
RMSE (log points)	0.478					
Clusters (states)	51					

Notes: Dependent variable: ln(asking price). v2 design matrix (lot recoded with a lot-missing indicator, parking winsorised at 20, stories at 10, 4,970 implausible listings removed; 62 regressors plus the indicator, of which 25 are shown; categorical controls for roof, construction and foundation and the remaining regressors are in Table 34). Standardised columns: continuous regressors scaled to zero mean and unit variance, so a coefficient is the association with a one-standard-deviation increase; binary regressors are not standardised. Standard errors in parentheses are cluster-robust by state (CR1, 51 clusters, Eq. (3)); "p, wild" is the wild cluster bootstrap p-value by state (Webb six-point weights, $B = 9,999$), the inference the paper relies on with 51 unbalanced clusters; stars follow the wild-bootstrap p-value: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Economic magnitude: elasticity for logged regressors, $100(e^{\beta} - 1)$ for indicators (Eq. (2); the Kennedy correction changes these by less than 0.1 pp, Table 36), 100β per unit otherwise. HC3 and ZIP3-clustered standard errors are compared in Table 9. ¶State median for the listings whose raw value is missing (Table 4), i.e. a state identifier for most of the sample. Estimates are conditional associations in asking prices, not willingness to pay.

Table 9. OLS Standard Errors Under Independence (HC3), ZIP3 Clustering and State Clustering, with Wild-Cluster-Bootstrap p-Values

		Standard error			State/	t	p, wild	G_{eff}
Variable	Coef.	HC3	ZIP3	State	HC3	(state)	(state)	(state)
Panel A: selected slopes								
ln(living area)	0.291	0.0032	0.0162	0.0329	10.2	8.85	0.000	21.1
Bedrooms	−0.070	0.0022	0.0127	0.0281	12.6	−2.49	0.013	17.2
Bathrooms	0.256	0.0027	0.0157	0.0235	8.9	10.91	0.000	14.2
Age (years)	0.012	0.0135	0.0650	0.1148	8.5	0.11	0.931	24.3
Age^2	0.044	0.0020	0.0090	0.0209	10.7	2.12	0.061	25.1
ln(lot size)	0.128	0.0009	0.0076	0.0170	19.4	7.54	0.000	17.4
Lot missing (indicator)	−0.023	0.0024	0.0269	0.0202	8.6	−1.14	0.319	12.3
Pool	0.068	0.0026	0.0279	0.0521	20.0	1.31	0.491	17.9
Garage	0.115	0.0015	0.0103	0.0160	10.9	7.15	0.000	17.8
Waterfront	−0.240	0.0331	0.2442	0.4495	13.6	−0.53	0.958	5.5
Central air	0.083	0.0014	0.0139	0.0222	16.3	3.76	0.003	17.5
Hardwood floors	0.123	0.0014	0.0094	0.0177	12.7	6.99	0.000	20.8
Luxury score (0–7)	0.004	0.0019	0.0085	0.0088	4.7	0.47	0.650	18.0
Walk Score	0.037	0.0012	0.0124	0.0192	16.3	1.91	0.096	18.9
Bike Score	0.079	0.0010	0.0122	0.0271	27.2	2.90	0.015	20.3
Transit Score	0.054	0.0008	0.0181	0.0124	15.4	4.35	0.001	14.2
Avg. school rating[¶]	−0.048	0.0007	0.0134	0.0237	33.7	−2.01	0.124	15.3
Property tax rate (%)	−0.019	0.0008	0.0146	0.0269	34.6	−0.70	0.582	14.8
Condominium	−0.093	0.0037	0.0402	0.0544	14.5	−1.72	0.408	7.8
HOA member	−0.066	0.0024	0.0225	0.0287	12.1	−2.29	0.102	13.4
New construction	0.071	0.0029	0.0228	0.0302	10.4	2.35	0.030	13.1
Foreclosure	−0.390	0.0097	0.0222	0.0363	3.7	−10.73	0.000	14.9
Region: Northeast	0.477	0.0030	0.0536	0.1238	41.2	3.85	0.014	12.5
Region: South	0.035	0.0027	0.0339	0.0616	23.1	0.57	0.649	19.4
Region: West	0.419	0.0030	0.0474	0.1042	34.4	4.02	0.008	21.1
Panel B: counts over all 63 slopes								
Slopes with $\lvert t\rvert > 1.96$		58	40	33				
Slopes with $\lvert t\rvert > t_{0.975}(G-1)$		58	40	33				
Slopes significant after Holm (5%)		54	23	17				
Slopes significant after Benjamini–Hochberg (5%)		58	36	26				
Slopes with wild-bootstrap $p < 0.05$ (Webb / Rademacher)				23 / 23				
after Holm / after Benjamini–Hochberg				9 / 20				
Median SE ratio to HC3			7.8	11.7				
State/HC3 ratio, p_{10}–p_{90} (min–max)				5.1–22.9 (3.7–41.2)				
Effective clusters (CSS), median; minimum			164; 22	14.2; 3.7				

Notes: Standardised v2 regression of Table 8, $N = 783,872$. Cluster-robust (CR1) standard errors follow Eq. (3) with 886 ZIP3 clusters (885 prefixes plus the code for a missing ZIP, 16 listings; reassigning them to the nearest ZIP3 changes no standard error by more than 0.03%) or 51 state clusters. "p, wild" is the wild cluster bootstrap p-value by state with Webb weights ($B = 9,999$; Rademacher weights give the same count at 5%). G_{eff} is the Carter–Schnepel–Steigerwald effective number of clusters for the coefficient. Panel B counts the 63 slopes (intercept excluded); $t_{0.975}(G-1)$ is 2.009 for states and 1.963 for ZIP3. Because the wild-bootstrap count is below the CR1 count, CR1 inference with 51 unbalanced clusters is not conservative here. Conley HAC standard errors are in Table 35.

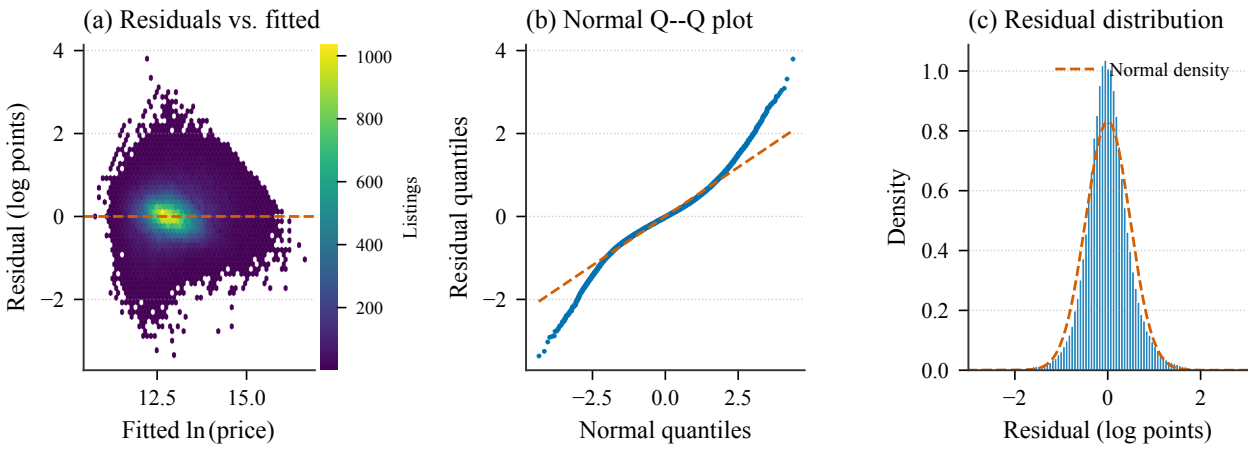


Figure 3. Diagnostics of the v2 OLS model on the 156,770 test listings of the random split (seed 42; out-of-sample residuals): (a) residuals against fitted values (hexagonal bins of a 100,000-listing random subset, viridis scale = number of listings); (b) normal quantile–quantile plot of the same subset; (c) residual density with a normal density of the same mean and variance. Test-set residual skewness 0.15 and kurtosis 5.43 (in-sample 0.12 and 5.38, Table 8); the fan-shaped spread motivates heteroskedasticity-robust errors and the clustered inference of Table 9 addresses the spatial dependence that this plot cannot show.

concentrated in a few coastal and lake markets. Walk Score enters alongside the correlated Bike and Transit Scores, so the accessibility premium is split across coefficients and cannot be attributed variable by variable. The negative basement and HOA-membership coefficients reflect, respectively, the correlation of basements with construction vintage and layout conditional on age and region, and the presence of the log HOA fee alongside the membership dummy. These signs do not indicate misspecification, but they show that attribute-by-attribute economic reading of a national pooled regression is limited, a point the state-by-state estimates of Section 5.3 make directly.

5.1.3. Geographic Fixed Effects

Table 10 documents how explanatory power rises with geographic granularity. On the random holdout, replacing the three region dummies by 50 state fixed effects raises out-of-sample R^2 from 0.646 to 0.690 (4.5 points), and 885 ZIP3 fixed effects raise it to 0.762 (11.6 points); the corresponding figures under the original coding are shown for reference. The 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). The remaining distance to XGBoost under random validation (7.1 points) is often read as the contribution of non-linearity; that reading is not warranted here, because XGBoost does not observe ZIP codes and Section 5.7.3 shows that it recovers location from state-level proxies, so the two models exploit different information.

Table 10. Out-of-Sample R^2 of OLS Under Progressively Finer Geographic Controls

Specification	Geographic control	N_{FE}	In-sample R^2	Out-of-sample R^2
Panel A: v2 design (trimmed sample, lot recoded), random 80/20 split				
Baseline (3 region dummies)	Census region	3	—	0.646
State fixed effects	State	50	—	0.690
ZIP3 fixed effects	3-digit ZIP	885	—	0.762
State-by-state (fully interacted)	State	51	—	0.752
Panel B: v1.1 design (full sample, $\ln(lot) = 0$ *coding), reproduced from the stored artefacts*				
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: Out-of-sample R^2 on the 20% random holdout (seed 42). Panel A: v2 design, training 627,102, test 156,770 listings; the fixed effects replace the three region dummies; the fully interacted row estimates a separate regression in each state on that state's training listings (Table 13; 51 separate regressions). Panel B: the v1.1 estimates reproduced by `scripts/00c_primary_models.py -legacy` (in-sample values from the stored fit; the state-FE in-sample value was not stored and is recomputed). The ZIP3 groups include the code for a missing ZIP.

5.1.4. Inference Under Spatial Dependence

The consequences of allowing for dependence are large and systematic (Table 9). Across the 63 slopes, ZIP3-cluster standard errors are on median 7.8 times the HC3 values and state-cluster standard errors 11.7 times, with an inter-decile range of 5.1–22.9 and a maximum of 41.2. The inflation is smallest for attributes that vary within neighbourhoods (foreclosure 3.7, living area 10.2) and largest for variables that are spatial aggregates or state-level imputations: school rating 33.7, tax rate 34.6, Bike Score 27.2 and the Northeast dummy 41.2, the pattern that Moulton (1990) describes for aggregate regressors. The number of slopes significant at 5% falls from 58 under HC3 to 40 under ZIP3 clustering and 33 under state clustering with normal critical values; after Holm's adjustment the state-cluster count is 17, and after Benjamini–Hochberg 26.

The state-cluster count with normal critical values is not a conservative bound. The effective number of state clusters is 14.2 at the median and 3.7 at the minimum, far below 51, and the wild cluster bootstrap by state, which is the appropriate reference with few and unequal clusters (MacKinnon and Webb, 2017), leaves 23 slopes significant at 5% with Webb weights (23 with Rademacher weights) and 9 after Holm. The gradients that carry the paper's economic interpretation survive—living area, bathrooms, garage, hardwood floors, central air, foreclosure and the Northeast and West differentials all have wild-bootstrap p-values below 0.05—but school rating, tax rate, Walk Score, pool, waterfront, condominium, HOA, luxury score and age do not. The Conley spatial-HAC standard errors of Table 35 lie between the HC and the state-cluster values (median ratio to HC0 of 2.4 at 25 km and 3.5 at 100 km against 4.2 for state clustering), confirming that the dependence decays with distance but extends well beyond the neighbourhood. The quantile and SHAP results

below are read with the same caution.

5.2. Spatial Autocorrelation of Residuals

Table 11 reports Moran's I (Eq. (10)) on the test-set residuals of the three main models under the random split, computed on all 156,770 test listings with eight-nearest-neighbour weights whose median eighth-neighbour distance is 2.6 km. For OLS, $I = 0.471$ ($z = 388.0$ under 999 permutations): four Census regions and 62 attributes leave most of the local price structure in the error term, and the dependence is strongest at the neighbourhood scale ($I = 0.639$ within 1 km, 0.207 at 5–25 km). LightGBM and XGBoost, which observe no coordinates and no ZIP codes, reduce the statistic to 0.288 and 0.236, absorbing about 50% of the residual spatial autocorrelation of OLS; they can only do so by recovering location from the features available to them, and Section 5.7.3 identifies which features serve that purpose. The finding mirrors Pace and Hayunga (2020), who show that trees and forests extract the spatial information that hedonic residuals retain. The earlier subsample-based statistics, computed on 5,000-listing national draws, measured regional rather than neighbourhood dependence and are superseded.

Table 11. Moran's I of Test-Set Residuals on the Full Random Test Set: OLS versus Tree Ensembles, by Neighbourhood Definition

		$k = 8$ nearest neighbours			Distance bands, I		
Model	N	I	z	d_8 (km)	0–1 km	1–5 km	5–25 km
Panel A: v2 predictions (trimmed random test set)							
OLS	156,770	0.471	388	2.6	0.639	0.442	0.207
LightGBM	156,770	0.288	234	2.6	0.250	0.136	0.041
XGBoost	156,770	0.236	195	2.6	0.188	0.105	0.027
Panel B: v1.2 stored predictions (untrimmed random test set)							
OLS	157,769	0.454	378	2.6	0.624	0.438	0.209
LightGBM	157,769	0.288	241	2.6	0.247	0.135	0.042
XGBoost	157,769	0.234	197	2.6	0.187	0.104	0.027

Notes: Residuals of the random-split test set (all listings, not subsamples). k-nearest-neighbour weights are row-standardised; z is the permutation z-statistic (999 permutations, all $p < 0.001$); d_8 is the median distance to the eighth neighbour, so the $k = 8$ statistic measures dependence at the neighbourhood scale. Distance-band weights are binary and unstandardised (I with its analytical z in the source file). The v1.1 subsample statistic ($I = 0.274$, three draws of 5,000 listings, $k = 8$) measured dependence at a regional scale because the eighth neighbour in a 5,000-listing subsample is tens of kilometres away; its three values are 0.2742, 0.2839, 0.2652 (population SD 0.0076, sample SD 0.0094; the z-statistics of the v1.2 table are not stored and are withdrawn).

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, the natural-unit specification is estimated separately in each of the 51 jurisdictions (region dummies dropped, constant columns dropped per state, ZIP3-cluster standard errors) and the predictive accuracy of pooled, state-fixed-effects and fully interacted specifications is compared on the random holdout. Table 12 summarises the distribution of the state coefficients and Figure 4 displays two of them with their confidence bands.

Table 12. State-by-State OLS Coefficients in Natural Units: Dispersion Across the 51 Jurisdictions, Sign Reversals with Inference, and Homogeneity Tests

	Pooled, state FE		State coefficients			States with				
Variable	Coef.	(SE)	p_{10}	Median	p_{90}	sign ≠	sign ≠, $\lvert t\rvert > 2$	CI excl. pooled	$\hat{\tau}$	p, hom.
ln(living area)	0.627	(0.030)	0.324	0.498	0.721	0	0	26	0.125	< 0.001
Bedrooms	−0.066	(0.011)	−0.073	−0.020	0.036	19	3	37	0.050	< 0.001
Bathrooms	0.203	(0.012)	0.052	0.150	0.199	0	0	31	0.043	< 0.001
Age (years)	−0.001	(0.002)	−0.016	−0.002	0.007	17	4	17	0.007	< 0.001
ln(lot size)	0.089	(0.005)	0.035	0.094	0.143	1	1	30	0.094	< 0.001
Pool	0.056	(0.023)	−0.078	0.008	0.110	23	6	23	0.105	< 0.001
Garage	0.096	(0.009)	0.029	0.073	0.138	2	0	17	0.084	< 0.001
Central air	0.065	(0.013)	−0.040	0.090	0.224	10	3	30	0.079	< 0.001
Hardwood floors	0.101	(0.008)	0.019	0.074	0.162	2	0	28	0.039	< 0.001
Luxury score (0–7)	0.006	(0.006)	−0.052	0.005	0.049	23	12	26	0.061	< 0.001
Foreclosure	−0.383	(0.017)	−0.544	−0.383	−0.222	1	0	14	0.096	< 0.001
Condominium	−0.105	(0.031)	−0.402	−0.073	0.134	14	5	18	0.387	< 0.001
HOA member	−0.006	(0.017)	−0.151	−0.022	0.085	22	5	16	0.052	< 0.001

Notes: Natural-unit v2 specification (lot recoded, trimmed sample, $N = 783,872$) estimated separately in each of the 51 jurisdictions by OLS with ZIP3-clustered standard errors; columns within each state drop constant regressors (`results/inference_v2/state_regressions_dropped_columns.csv`). "Pooled, state FE" adds 50 state fixed effects to the pooled regression (ZIP3-clustered SE in parentheses). "sign ≠" counts states whose point estimate has the opposite sign from the pooled coefficient; "sign ≠, $|t| > 2$" those where the opposite sign is also significant at 5%; "CI excl. pooled" those whose 95% interval excludes the pooled value. $\hat{\tau}$ is the DerSimonian–Laird random-coefficient standard deviation and "p, hom." the p-value of the cluster-robust Cochran Q test of equal coefficients across states. Small-state designs (e.g. DE, $n = 1,998$; HI, $n = 5,013$) remain near-singular (maximum condition number 1.2×10^{21}), so the extremes should be read with their intervals (Figure 4).

The pooled living-area elasticity of 0.63 averages state elasticities with an inter-decile range of 0.32–0.72; Florida (0.79), California (0.74) and New York (0.61) price space more steeply than Tennessee (0.34) or Ohio (0.47). The extremes, Delaware (0.06, $n = 1,998$) and Hawaii (1.13, interval 0.83–1.43), are small-state fits in which the living-area and bathroom coefficients are near mirror images, and the inter-decile range is the defensible summary. A cluster-robust test rejects equal coefficients across states for every variable in the table ($p < 0.001$), and the 95% interval of 26 states excludes the pooled living-area elasticity. Sign reversals are less common than point estimates suggest once inference is attached: the bedroom coefficient has the opposite sign from the pooled estimate in 19 states but is significantly so in only 3; the pool coefficient in 23 and 6; the luxury score in 23 and 12; HOA membership in 22 and 5. Living area, bathrooms, garage and

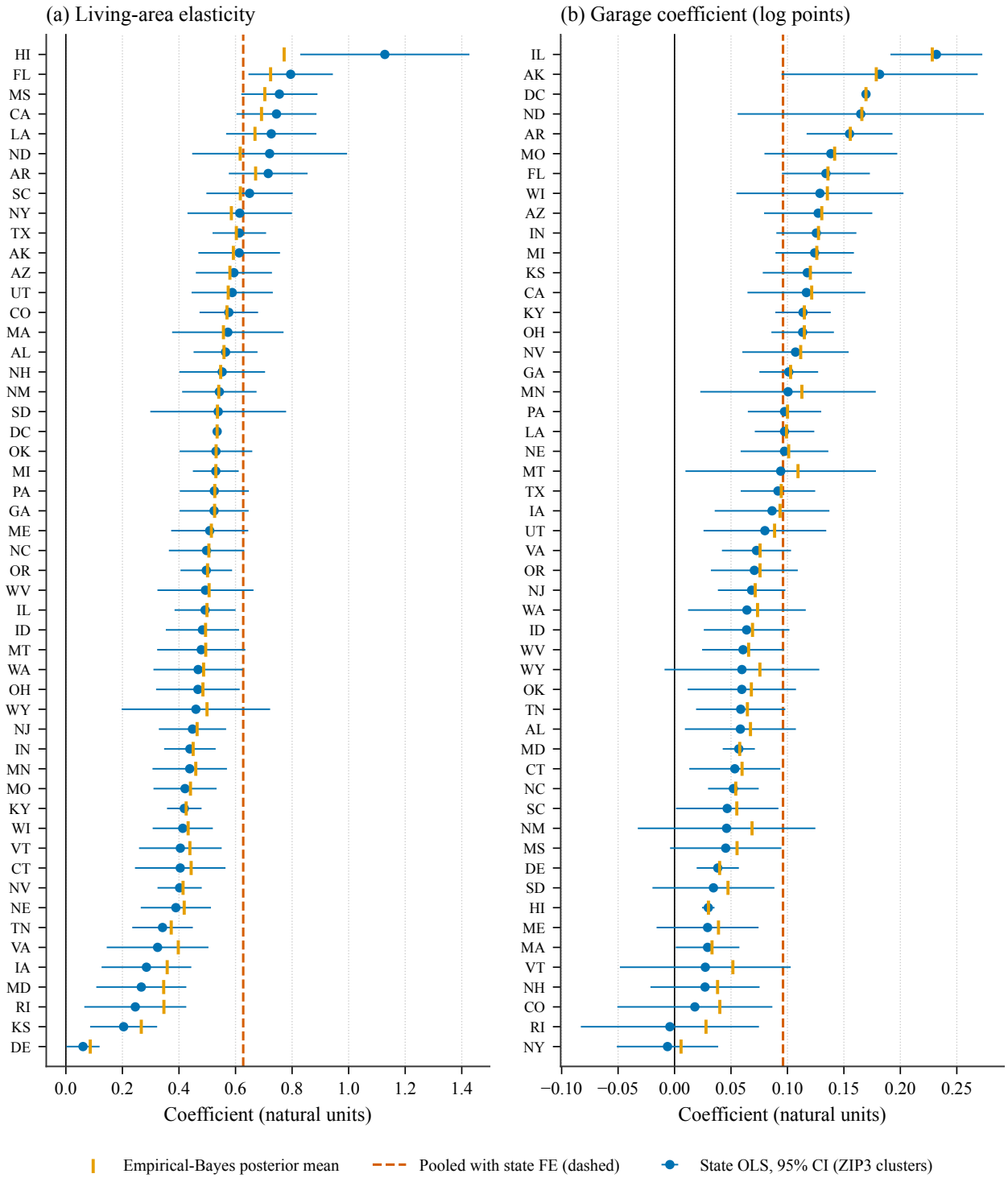


Figure 4. State-by-state OLS coefficients in natural units with inference (trimmed v2 sample, Table 12): (a) the elasticity of the asking price with respect to living area; (b) the garage coefficient. Circles are the state estimates with 95% confidence intervals from ZIP3-clustered standard errors; orange ticks are empirical-Bayes posterior means that shrink imprecise small-state estimates towards the precision-weighted mean; the dashed line is the pooled coefficient with state fixed effects. The extreme elasticities (DE at the bottom, HI at the top) carry the widest intervals.

hardwood floors have the same sign in essentially every state.

Table 13 translates the 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 0.752. Slope heterogeneity across states is therefore worth 6.1 points of out-of-sample R^2, almost as much as the state price levels themselves (7.2 points). The implication 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 13. Random-Holdout R^2 of Pooled, State-Fixed-Effects and Fully Interacted (State-by-State) OLS

Specification	N_{train}	N_{test}	Out-of-sample R^2
Pooled OLS, no geographic controls	627,102	156,770	0.619
Pooled OLS + 50 state fixed effects	627,102	156,770	0.690
State-by-state OLS (fully interacted)	627,102	156,770	0.752
Gain of state slopes over state fixed effects (pp)			6.1

Notes: Random 80/20 split (seed 42) of the trimmed v2 sample; natural units, lot recoded. The fully interacted model estimates a separate regression in each state on the training listings of that state and predicts its test listings. Sampling uncertainty of these R^2 values is of the order of the ZIP3-cluster bootstrap intervals of Table 38.

5.4. Quantile Regression and Inter-Quantile Tests

5.4.1. Coefficient Estimates Across Quantiles

Table 14 presents quantile-regression coefficients at the five quantiles on the corrected design; all five fits converged. The pseudo-R^2 of Koenker and Machado (1999) is 0.429, 0.415, 0.413, 0.435 and 0.472 at $\tau = 0.10, 0.25, 0.50, 0.75, 0.90$, so the specification fits the upper tail of the conditional distribution slightly better than its centre. The living-area gradient declines from 0.363 at $\tau = 0.10$ to 0.240 at $\tau = 0.90$: additional space is associated with proportionally larger price differences at the lower end of the conditional distribution. The bathroom gradient rises monotonically from 0.188 to 0.290. The garage gradient declines from 0.210 to 0.018, a ratio of 11.5, consistent with a garage differentiating lower-priced homes and being nearly universal, hence unpriced, higher up. The pool gradient rises from 0.034 to 0.107, a 11.2% difference at the top decile. The lot elasticity rises from 0.101 to 0.147 along the distribution, and the parking coefficient is at most 0.029 in absolute value at every quantile, where the uncorrected design had produced coefficients in the hundreds from a handful of corrupt entries. The age main effect moves from -0.252 at $\tau = 0.10$ to 0.089 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 6.2). Figure 5 traces the coefficient paths of eight attributes across the quantiles and shows how much wider the ZIP3-cluster-bootstrap intervals at the tails are than the independence-based bands.

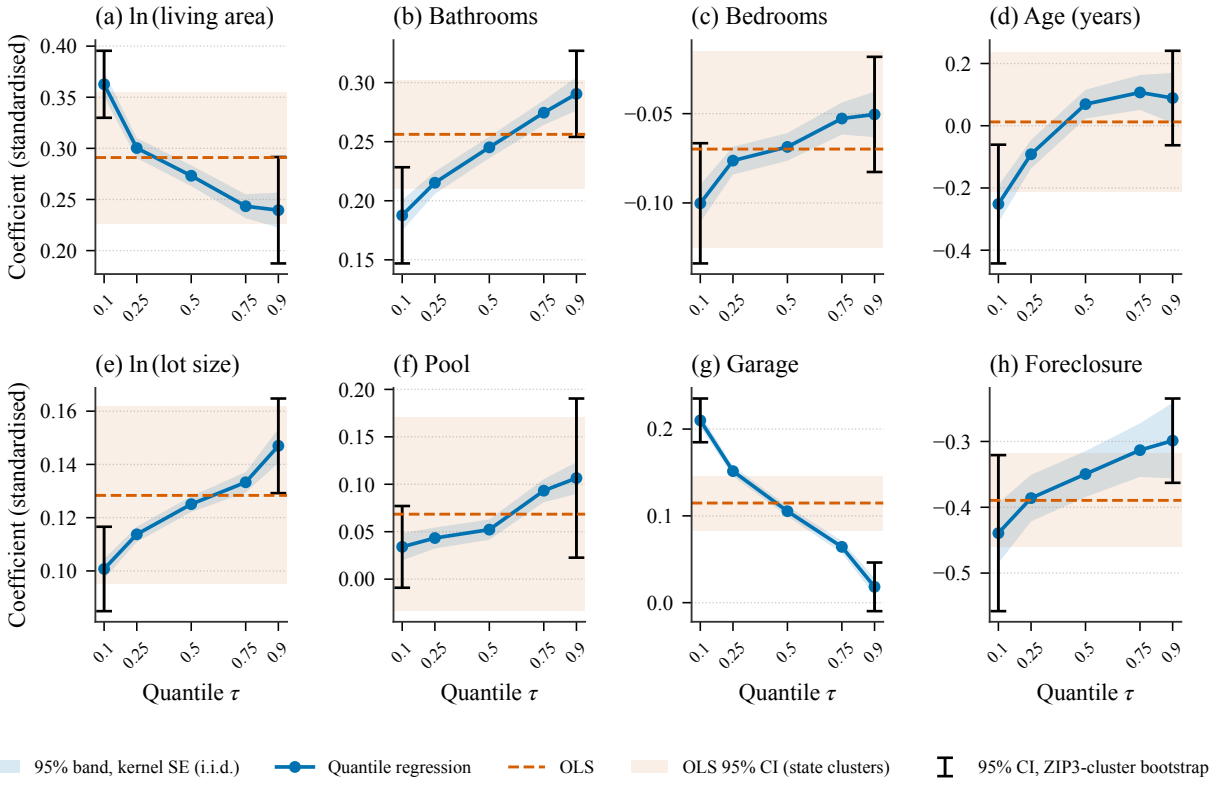


Figure 5. Quantile-regression coefficient paths on the v2 design for eight attributes (random subsample of 150,000 listings; continuous regressors standardised, Table 14). Shaded bands are 95% intervals from the kernel standard errors, which assume independent observations; the black bars at $\tau = 0.10$ and 0.90 are 95% intervals from the paired ZIP3-cluster bootstrap (999 draws, Table 37) and are several times wider. The dashed line is the OLS coefficient with its state-clustered 95% interval.

5.4.2. Inter-Quantile Tests

Table 15 tests the equality of the $\tau = 0.10$ and $\tau = 0.90$ coefficients, with the statistic of Eq. (5), for 14 variables (thirteen key attributes and the lot-missing indicator). The ZIP3-cluster bootstrap standard errors of the differences are on median 2.3 times the independence-based values (range 1.7–4.2). Under the independence formula 13 of 14 differences are significant at 5%; under the paired ZIP3-cluster bootstrap 10 are, by the percentile interval; after Holm's step-down and after the Romano–Wolf procedure 8 remain, and 10 under Benjamini–Hochberg. The largest and most robust differences are those of garage ($z = 12.63$), lot size, bathrooms, the Northeast dummy, living area and the luxury score; the age difference and the bedroom difference survive Holm; the age-squared and foreclosure differences are significant by the percentile interval but not after

Table 14. Quantile-Regression Coefficients on the v2 Design at Five Quantiles of the Conditional Log-Price Distribution, with the OLS Benchmark

Variable	$\tau = 0.10$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.90$	OLS
ln(living area)	0.363	0.300	0.273	0.243	0.240	0.291
	(0.006)	(0.005)	(0.005)	(0.006)	(0.009)	(0.003)
Bedrooms	−0.100	−0.076	−0.069	−0.053	−0.050	−0.070
	(0.005)	(0.004)	(0.004)	(0.005)	(0.006)	(0.002)
Bathrooms	0.188	0.215	0.245	0.275	0.290	0.256
	(0.007)	(0.005)	(0.005)	(0.005)	(0.007)	(0.003)
Age (years)	−0.252	−0.092	0.069	0.107	0.089	0.012
	(0.030)	(0.024)	(0.024)	(0.029)	(0.041)	(0.014)
Age^2	0.054	0.052	0.051	0.045	0.017	0.044
	(0.004)	(0.004)	(0.004)	(0.004)	(0.006)	(0.002)
ln(lot size)	0.101	0.114	0.125	0.133	0.147	0.128
	(0.002)	(0.001)	(0.002)	(0.002)	(0.003)	(0.001)
Lot missing (indicator)	−0.024	−0.028	−0.040	−0.032	−0.040	−0.023
	(0.007)	(0.005)	(0.005)	(0.006)	(0.008)	(0.002)
Parking spaces	0.029	0.026	0.023	0.017	0.011	0.021
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)
Pool	0.034	0.043	0.052	0.093	0.107	0.068
	(0.007)	(0.006)	(0.005)	(0.006)	(0.009)	(0.003)
Garage	0.210	0.152	0.105	0.064	0.018	0.115
	(0.004)	(0.003)	(0.003)	(0.004)	(0.005)	(0.001)
Luxury score (0–7)	−0.030	−0.008	0.011	0.025	0.035	0.004
	(0.005)	(0.004)	(0.004)	(0.005)	(0.007)	(0.002)
Foreclosure	−0.439	−0.386	−0.350	−0.313	−0.299	−0.390
	(0.023)	(0.018)	(0.018)	(0.021)	(0.029)	(0.010)
Waterfront	−0.068	−0.116	−0.245	−0.440	−0.559	−0.240
	(0.079)	(0.060)	(0.059)	(0.068)	(0.090)	(0.033)
Region: Northeast	0.342	0.379	0.450	0.548	0.611	0.477
	(0.008)	(0.006)	(0.006)	(0.007)	(0.010)	(0.003)
Region: West	0.412	0.384	0.381	0.412	0.481	0.419
	(0.008)	(0.007)	(0.007)	(0.008)	(0.011)	(0.003)
Pseudo-R^2, $R^1(\tau)$	0.429	0.415	0.413	0.435	0.472	—
Converged (max. 5,000 iterations)	yes	yes	yes	yes	yes	—
N			150,000			783,872

Notes: Quantile regressions on a random subsample of 150,000 listings (seed 42) of the trimmed v2 sample; all 62 regressors plus the lot-missing indicator included, 15 shown; continuous regressors standardised. Standard errors in parentheses are the default kernel (Koenker–Bassett) errors, which assume independent observations; ZIP3-cluster-bootstrap errors at $\tau = 0.10$ and 0.90 are in Table 37 (e.g. garage: kernel 0.0038 vs. cluster 0.0128 at $\tau = 0.10$) and inter-quantile tests with multiplicity control in Table 15; no stars are shown. The OLS column is the standardised regression of Table 9 with its HC3 error. Pseudo-R^2 is the Koenker–Machado $R^1(\tau)$ (the v1.1 table reported $1 - R^1(\tau)$, Appendix F). In the v1.1 design the corrupt parking column produced coefficients of several hundred at $\tau \geq 0.25$; on the winsorised v2 column its coefficients are at most 0.029 in absolute value.

multiplicity control; and the pool, waterfront, West and lot-missing differences are not significant under any cluster-based criterion. The conditional distribution of asking prices is thus differentially stretched and compressed by the structural attributes, while the tail differences of the rarer amenities and of the West dummy are estimated with far less precision than independence-based statistics suggest.

Table 15. Inter-Quantile Differences $\hat{\beta}(0.10) - \hat{\beta}(0.90)$ on the v2 Design: Paired Bootstrap Inference with Multiplicity Control

		z-statistic			95% CI	p-value (ZIP3 cluster bootstrap)			
Variable	Diff.	Indep.	Obs.	ZIP3	(percentile, ZIP3)	Normal	Holm	BH	RW
Garage	0.192	30.36	24.24	12.63	[+0.155, +0.214]*	0.000	0.000	0.000	0.000
ln(lot size)	−0.046	−12.16	−11.79	−5.35	[−0.062, −0.028]*	0.000	0.000	0.000	0.000
Bathrooms	−0.103	−10.50	−7.67	−5.32	[−0.142, −0.065]*	0.000	0.000	0.000	0.000
Region: Northeast	−0.269	−21.56	−16.74	−5.18	[−0.364, −0.162]*	0.000	0.000	0.000	0.000
ln(living area)	0.123	11.36	6.92	4.56	[+0.081, +0.186]*	0.000	0.000	0.000	0.000
Luxury score (0–7)	−0.065	−7.66	−6.61	−4.21	[−0.095, −0.034]*	0.000	0.000	0.000	0.000
Age (years)	−0.341	−6.67	−5.30	−3.29	[−0.519, −0.110]*	0.001	0.008	0.002	0.010
Bedrooms	−0.050	−5.90	−4.55	−2.73	[−0.088, −0.018]*	0.006	0.044	0.011	0.044
Age^2	0.037	4.77	3.89	2.60	[+0.004, +0.059]*	0.009	0.056	0.015	0.056
Foreclosure	−0.141	−3.75	−2.86	−2.19	[−0.274, −0.011]*	0.028	0.141	0.040	0.132
Pool	−0.072	−6.43	−4.95	−1.83	[−0.139, +0.012]	0.067	0.268	0.079	0.247
Waterfront	0.491	4.11	2.94	1.83	[−0.060, +0.979]	0.068	0.268	0.079	0.247
Region: West	−0.069	−5.06	−4.40	−1.37	[−0.169, +0.018]	0.170	0.341	0.184	0.335
Lot missing (indicator)	0.016	1.47	1.43	0.64	[−0.027, +0.068]	0.520	0.520	0.520	0.544
Rejections at 5% (of 14)		13	13	10	10	10	8	10	8

Notes: Both quantile regressions estimated on the same 150,000-listing subsample of the v2 design; 14 differences (13 key variables plus the lot-missing indicator), sorted by $|z|$ (ZIP3). "Indep." treats the two estimates and the observations as independent (kernel standard errors, covariance set to zero); "Obs." and "ZIP3" are paired bootstraps with 999 draws (999 converged in each scheme) resampling listings or whole ZIP3 areas (884 clusters) and re-estimating both quantile regressions on each draw. Columns 7–10 are the normal-approximation p-value of the ZIP3 scheme and its Holm, Benjamini–Hochberg and Romano–Wolf (step-down, bootstrap) adjustments across the 14 tests. A star marks a percentile interval that excludes zero. The v1.1 version reported "10 of 13" rejections from a 200-draw bootstrap without multiplicity control.

5.5. Machine Learning Under Random Validation

Table 16 summarises out-of-sample performance under the random 80/20 split of the trimmed sample. XGBoost achieves $R^2 = 0.833$ (seed standard deviation 0.0004), 18.7 points above OLS; its median absolute percentage error on the dollar scale (Eq. (7)) is 16.5% against 26.5% for OLS, and its mean absolute dollar error \$157,041 against \$236,762. LightGBM reaches 0.810 and the random forest 0.800. Ridge, Lasso and Elastic Net with cross-validated penalties reproduce OLS to the fourth decimal, because the chosen penalties are essentially zero at this sample size: the parametric specification is not overfit. The ZIP3 fixed-effects regression reaches 0.762. Resampling whole ZIP3 clusters of the test set gives a 95% interval of [17.6, 19.9] points for the XGBoost–OLS difference, and the paired test on squared-error differences gives a mean loss difference of 0.1216

with a ZIP3-cluster standard error of 0.0063 (t = 19.4; Table 17 reports every pair). Figure 6 shows that OLS compresses both tails of the predicted distribution while the ensembles track them. The ranking XGBoost > LightGBM > random forest > OLS under random validation is not a sampling accident, even after spatial dependence in the test set is allowed for. Exact duplicates across the split do not drive it (Section 6.3).

Table 16. Out-of-Sample Accuracy of Linear and Tree-Ensemble Hedonic Models Under Random Validation on the v2 Design

	Log-price scale				Dollar scale		
Model	R^2	(SD, seeds)	RMSE	MAE	MdAPE (%)	MAE ($)	Best iter.
OLS (62 regressors + lot indicator)	0.646	—	0.480	0.354	26.5	236,762	—
OLS + state fixed effects	0.690	—	0.448	0.329	24.5	221,398	—
OLS + ZIP3 fixed effects	0.762	—	0.393	0.283	20.7	194,629	—
Ridge (α by 5-fold CV)^r	0.646	—	0.480	0.354	26.5	236,761	—
Lasso (α by 5-fold CV)^r	0.646	—	0.480	0.354	26.5	236,760	—
Elastic net (α, ℓ_1 ratio by 5-fold CV)^r	0.646	—	0.480	0.354	26.5	236,760	—
XGBoost (1,000 rounds, early stopping)	0.833	(0.0004)	0.329	0.232	16.5	157,041	999
LightGBM (1,000 rounds, early stopping)	0.810	(0.0002)	0.351	0.251	18.1	168,073	1,000
Random forest (500 trees, 1/3 features, leaf 5)^r	0.800	—	0.360	0.250	17.2	172,687	—
Panel B: paired comparisons on the same test listings (ZIP3-cluster inference)							
R^2(OLS) − R^2(XGBoost), 95% CI	−0.187 [−0.199, −0.176]						
R^2(OLS) − R^2(LightGBM), 95% CI	−0.164 [−0.175, −0.153]						
Mean squared-error difference, OLS minus XGBoost	0.1216 (SE 0.0063, t = 19.4; i.i.d. SE 0.0009)						
Mean squared-error difference, LightGBM minus XGBoost	0.0148 (SE 0.0006, t = 25.7; i.i.d. SE 0.0002)						
Mean squared-error difference, OLS minus LightGBM	0.1067 (SE 0.0059, t = 18.1; i.i.d. SE 0.0008)						

Notes: Random 80/20 split (seed 42) of the trimmed v2 sample: training 627,102, test 156,770 listings; every model uses the same design matrix (62 regressors plus the lot-missing indicator). Boosters: 1,000 rounds with early stopping on a random 10% of the training set, seeds 42–46 (SD across seeds in parentheses); the best iteration is the mean over seeds. Dollar metrics exponentiate the predicted log price; MdAPE is the median absolute percentage error (Eq. (7); the mean APE is in the source file). ^{r}Re-specified in v2 with declared cross-validation grids because the v1.1 code and hyperparameters of these rows were lost (Appendix F); the penalised models select $\alpha \approx 0$ and add nothing to OLS at this sample size. Panel B: gap in R^2 with the ZIP3-cluster bootstrap interval of Table 38, and the mean difference in squared errors with its ZIP3-cluster standard error (885 clusters).

Table 17. Paired Squared-Error Differences Between Models on the Random Test Set with i.i.d. and ZIP3-Cluster Standard Errors

Predictions	Difference	Mean	SE, i.i.d.	SE, ZIP3	t, ZIP3
v2 (N2)	OLS minus XGBoost	0.1216	0.00086	0.00628	19.4
v2 (N2)	LightGBM minus XGBoost	0.0148	0.00021	0.00058	25.7
v2 (N2)	OLS minus LightGBM	0.1067	0.00079	0.00591	18.1

Notes: Mean of the listing-level difference in squared log-price errors on the 156,770 random-split test listings (v2 predictions, seed 42); the ZIP3-cluster standard error allows for dependence within the 885 ZIP3 areas of the test set and is about 7 times the i.i.d. value.

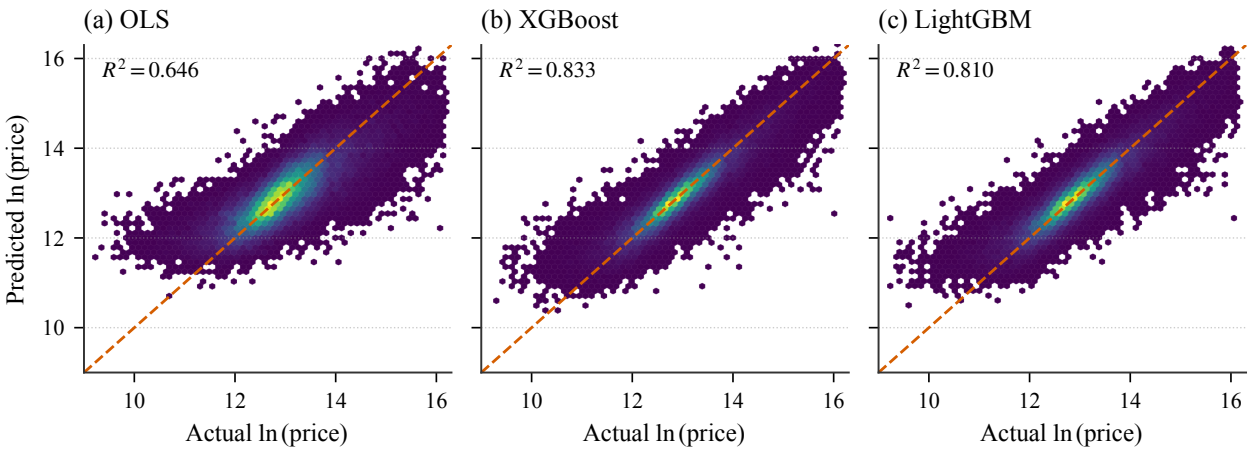


Figure 6. Predicted against actual log asking prices on the random-split test set (trimmed v2 sample, seed 42; 60,000-listing random subset per panel, hexagonal bins with a viridis count scale): (a) OLS, (b) XGBoost, (c) LightGBM; dashed line = 45-degree line. The R^2 printed in each panel is computed on all 156,770 test listings (Table 16). OLS compresses the tails; the ensembles track them.

5.6. Validation Design and the OLS–Ensemble Gap

5.6.1. Seven Designs

Table 18 summarises the comparison: OLS, XGBoost and LightGBM with nested rounds under each of the seven designs, with the gap and its interval. Table 19 reports OLS, XGBoost under the three rounds rules, XGBoost with the missingness indicators and LightGBM under every validation design, with the oracle within-state R^2, the dollar error and the ZIP3-cluster interval for the OLS–ensemble difference; Figure 7 plots the difference by design and Table 38 in Appendix C lists every interval. The ensemble advantage declines with the spatial scale of the blocks. It is 18.7 points under the random split, 18.1 under the size-matched random split, 16.8 under ZIP5 blocks, 13.2 under ZIP3 blocks, 4.8 under state blocks and 6.3 under leave-one-state-out, with intervals that exclude zero in every case; under the ten-state holdout OLS is ahead by 2.6 points with an interval of $[-0.1, 7.4]$, while LightGBM is ahead of OLS there by 2.4 points (interval $[0.4, 4.7]$). The results with and without the missingness indicators differ by less than two seed standard deviations under the geographic designs (Table 32).

Two features of the table deserve comment. First, training-set size explains almost none of the gap: shrinking the random training set to the size of the ten-state design costs XGBoost 0.4 points, a difference that exceeds the seed noise but is an order of magnitude smaller than any geographic effect. Second, the nested rule selects very different capacities across designs: 1,000 rounds under random and ZIP5 blocks, 820 under ZIP3 blocks, 155 under state blocks, 175 under the ten-state holdout and between 75 and 300 across the leave-one-state-out folds (mean 153). The rounds rule matters only where the blocks are states: under the ten-state holdout the fully grown 1,000-round model reaches

0.425 (seed standard deviation 0.015), the nested model 0.515 and the fixed 100-round model 0.538.

Table 18. The OLS-versus-Ensemble Accuracy Gap Depends on the Validation Design: Out-of-Sample R^2 and ZIP3-Cluster Bootstrap Intervals Under Seven Designs

		OLS	XGBoost		LightGBM	OLS − XGBoost		OLS − LightGBM	
Validation design	N_{test}	R^2	R^2	(SD)	R^2	Gap	95% CI	Gap	95% CI
Random	156,770	0.646	0.833	(0.0002)	0.809	−0.187	[−0.199, −0.176]	−0.164	[−0.175, −0.153]
Random, size-matched	435,574	0.649	0.829	(0.0003)	0.809	−0.181	[−0.191, −0.169]	−0.160	[−0.171, −0.149]
ZIP5 blocks	783,872	0.647	0.815	(0.0001)	0.799	−0.168	[−0.178, −0.158]	−0.152	[−0.163, −0.142]
ZIP3 blocks	783,872	0.635	0.767	(0.0002)	0.759	−0.132	[−0.145, −0.119]	−0.124	[−0.138, −0.111]
State blocks	783,872	0.580	0.628	(0.0013)	0.631	−0.048	[−0.062, −0.035]	−0.052	[−0.063, −0.041]
Leave-one-state-out	783,872	0.573	0.637	(0.0012)	0.643	−0.063	[−0.076, −0.047]	−0.070	[−0.085, −0.057]
Ten-state holdout	435,553	0.541	0.515	(0.0052)	0.565	0.026	[−0.001, +0.074]	−0.024	[−0.047, −0.004]

Notes: Trimmed v2 sample ($N = 783,872$), one design matrix for every model. R^2 is computed on the pooled out-of-fold predictions (for the random, size-matched and ten-state designs on the single test set). XGBoost and LightGBM use the number of boosting rounds selected by an inner GroupKFold(3) on the training groups ("nested"; never on the test states), mean over seeds 42–46 with the SD across seeds in parentheses; OLS has no tuning parameter. "Gap" is R^2(OLS) − R^2(ensemble) with its 95% percentile interval from 1,000 ZIP3-cluster bootstrap draws of the test listings (ZIP3 areas resampled with replacement, both models' predictions held fixed). The ten-state holdout withholds CA, CO, FL, GA, IL, NC, NY, OH, TX, WA; the size-matched random design uses a training set of the same size. Full results by comparator, including fixed 100- and 1,000-round XGBoost, are in Tables 19 and 38; LOSO state-level results in Table 21.

5.6.2. Blocked Cross-Validation at Three Scales

Table 20 reports the fold-level results of the grouped five-fold designs. Under ZIP5 blocks the five folds are nearly interchangeable (XGBoost 0.808–0.823), under ZIP3 blocks almost so, and under state blocks they are not: the fold holding Florida with six small states has $R^2 = 0.510$ for XGBoost and 0.456 for OLS, while the best fold reaches 0.717. As the block widens, pooled XGBoost R^2 falls from 0.815 to 0.767 to 0.628, and the oracle within-state R^2 from 0.781 to 0.727 to 0.651; OLS moves from 0.647 to 0.635 to 0.580. The ensemble loses more than OLS at every step, which is what one expects if part of its advantage is the memorisation of local price levels at scales that the semi-log regression with region dummies cannot represent. Finer blocking than ZIP5 is not attempted; county identifiers are missing for 97.0% of listings, but ZIP-based blocking covers the range from neighbourhood to state.

5.6.3. Leave-One-State-Out

Leave-one-state-out validation (Tables 21 and 22; Figure 8; state-level detail in Table 41) gives the most complete picture of out-of-state accuracy. The median held-out-state R^2 of XGBoost with nested rounds is 0.634, with an inter-decile range of 0.441–0.727, and 42 of the 51 jurisdictions exceed 0.5; for OLS the median is 0.551, the inter-decile range 0.376–0.623 and 33 states exceed 0.5. The listing-weighted means, 0.578 and 0.505, are lower than the medians because the largest states

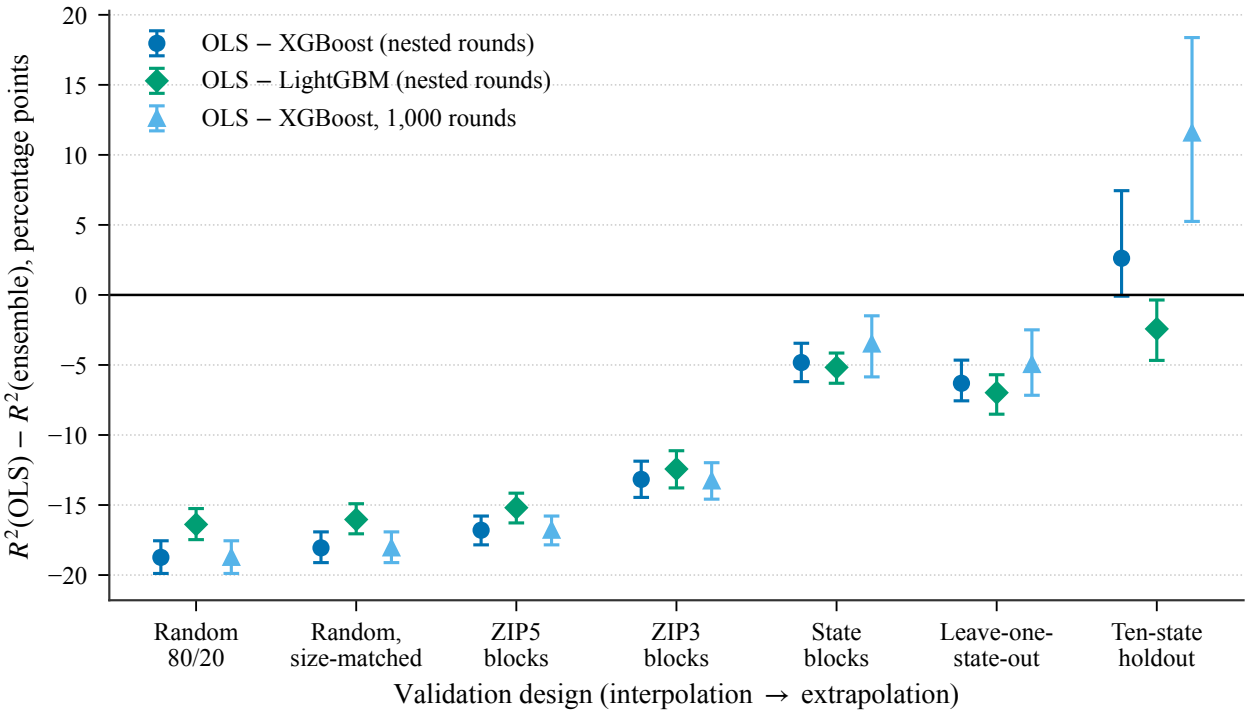


Figure 7. The OLS-versus-ensemble gap by validation design. Points are R^2(OLS) − R^2(ensemble) in percentage points on the pooled out-of-fold predictions of the trimmed v2 sample (Table 19); bars are 95% percentile intervals from 1,000 ZIP3-cluster bootstrap draws of the test listings. XGBoost and LightGBM with rounds selected on the training groups (nested) and XGBoost with 1,000 fixed rounds (the v1.1 comparator). Negative values favour the ensemble. The ensembles are more accurate under every design except the single ten-state partition, where the interval for the nested XGBoost comparator covers zero.

Table 19. Every Model Under Every Validation Design on the v2 Sample: Pooled Out-of-Fold R^2, Oracle Within-State R^2, Dollar Error and Selected Rounds

Design and model	N_{test}	R^2	(SD)	R^2_{within}	MdAPE (%)	Rounds	OLS − model, 95% CI
Random 80/20							
OLS (62 regressors + lot indicator)	156,770	0.646	—	0.615	26.5	—	—
OLS, no region dummies	156,770	0.619	—	0.600	27.4	—	—
OLS + state fixed effects	156,770	0.690	—	0.632	24.5	—	—
XGBoost, nested rounds	156,770	0.833	(0.0002)	0.802	16.4	1,000	[−0.199, −0.176]
XGBoost, 1,000 rounds	156,770	0.833	(0.0002)	0.802	16.4	—	[−0.199, −0.176]
XGBoost, 100 rounds	156,770	0.770	(0.0004)	0.732	20.0	—	[−0.135, −0.114]
XGBoost + missingness indicators, nested	156,770	0.835	(0.0001)	0.805	16.3	1,000	[−0.201, −0.178]
LightGBM, nested rounds	156,770	0.809	—	0.775	18.1	1,000	[−0.175, −0.153]
Random, geographic training size							
OLS (62 regressors + lot indicator)	435,574	0.649	—	0.618	26.4	—	—
OLS, no region dummies	435,574	0.622	—	0.603	27.3	—	—
OLS + state fixed effects	435,574	0.693	—	0.636	24.5	—	—
XGBoost, nested rounds	435,574	0.829	(0.0003)	0.798	16.6	1,000	[−0.191, −0.169]
XGBoost, 1,000 rounds	435,574	0.829	(0.0003)	0.798	16.6	—	[−0.191, −0.169]
XGBoost, 100 rounds	435,574	0.771	(0.0005)	0.733	20.0	—	[−0.132, −0.112]
XGBoost + missingness indicators, nested	435,574	0.831	(0.0002)	0.800	16.5	1,000	[−0.193, −0.171]
LightGBM, nested rounds	435,574	0.809	—	0.774	18.1	1,000	[−0.171, −0.149]
Ten-state holdout							
OLS (62 regressors + lot indicator)	435,553	0.541	—	0.573	27.6	—	—
OLS, no region dummies	435,553	0.533	—	0.559	29.7	—	—
XGBoost, nested rounds	435,553	0.515	(0.0052)	0.637	27.6	175	[−0.001, +0.074]
XGBoost, 1,000 rounds	435,553	0.425	(0.0146)	0.628	29.7	—	[+0.052, +0.184]
XGBoost, 100 rounds	435,553	0.538	(0.0070)	0.631	27.2	—	[−0.013, +0.037]
XGBoost + missingness indicators, nested	435,553	0.523	(0.0104)	0.640	26.9	300	[−0.022, +0.032]
LightGBM, nested rounds	435,553	0.565	—	0.643	26.6	300	[−0.047, −0.004]

Notes: Continued on the next page; notes follow the last part.

are predicted worst. XGBoost is ahead of OLS in 44 of 51 states (sign-test p <0.0001, Wilcoxon p <0.0001), by a median of 0.082 points of R^2; LightGBM is ahead in 44. Level bias is the dominant error for both models: the median absolute mean error across states is 0.105 log points for XGBoost and 0.127 for OLS, and the states predicted worst are those whose price level lies farthest from what the rest of the country implies for homes with their attributes—California (R^2 = 0.22, mean error −0.42 log points for XGBoost; 0.14 for OLS) and Hawaii (0.25, −0.45) are under-priced, Florida (0.50, −0.31) over-priced. The nested and fixed-100 rules give nearly identical state distributions; the 1,000-round rule has a higher median (0.659) but a lower listing-weighted mean (0.565), consistent with the additional capacity helping small states and hurting the large ones. The re-specified random forest reaches 0.583 under the ten-state holdout, between OLS and the boosters.

5.7. Where the Out-of-State Error Comes From

5.7.1. Level Versus Within-Market Error

The exact decomposition of the held-out squared error into within-state and between-state components (Table 39 in Appendix C) shows that the geographic designs differ from the random split mainly in the level component. Under the random split, the between-state share of the test variance of log prices is 0.0% for XGBoost and 3.0% for OLS, so the oracle within-state R^2 (0.802 and 0.615) is

Table 19. Every Model Under Every Validation Design on the v2 Sample: Pooled Out-of-Fold R^2, Oracle Within-State R^2, Dollar Error and Selected Rounds (continued)

Design and model	N_{test}	R^2	(SD)	R^2_{within}	MdAPE (%)	Rounds	OLS − model, 95% CI
GroupKFold(5) by state							
OLS (62 regressors + lot indicator)	783,872	0.580	—	0.587	28.7	—	—
OLS, no region dummies	783,872	0.568	—	0.576	29.2	—	—
XGBoost, nested rounds	783,872	0.628	(0.0013)	0.651	26.1	155	[−0.062, −0.035]
XGBoost, 1,000 rounds	783,872	0.615	(0.0027)	0.659	26.1	—	[−0.059, −0.015]
XGBoost, 100 rounds	783,872	0.635	(0.0023)	0.649	25.9	—	[−0.066, −0.041]
XGBoost + missingness indicators, nested	783,872	0.634	(0.0022)	0.655	26.0	135	[−0.069, −0.046]
LightGBM, nested rounds	783,872	0.631	—	0.651	26.2	235	[−0.063, −0.041]
GroupKFold(5) by ZIP3							
OLS (62 regressors + lot indicator)	783,872	0.635	—	0.606	26.9	—	—
OLS, no region dummies	783,872	0.609	—	0.593	27.8	—	—
OLS + state fixed effects	783,872	0.673	—	0.616	25.4	—	—
XGBoost, nested rounds	783,872	0.767	(0.0002)	0.727	19.8	820	[−0.145, −0.119]
XGBoost, 1,000 rounds	783,872	0.768	(0.0003)	0.729	19.7	—	[−0.146, −0.120]
XGBoost, 100 rounds	783,872	0.736	(0.0005)	0.696	21.6	—	[−0.114, −0.089]
XGBoost + missingness indicators, nested	783,872	0.768	(0.0001)	0.729	19.8	860	[−0.146, −0.120]
LightGBM, nested rounds	783,872	0.759	—	0.719	20.4	1,000	[−0.138, −0.111]
GroupKFold(5) by ZIP5							
OLS (62 regressors + lot indicator)	783,872	0.647	—	0.616	26.4	—	—
OLS, no region dummies	783,872	0.620	—	0.601	27.4	—	—
OLS + state fixed effects	783,872	0.691	—	0.633	24.6	—	—
XGBoost, nested rounds	783,872	0.815	(0.0001)	0.781	17.4	1,000	[−0.178, −0.158]
XGBoost, 1,000 rounds	783,872	0.815	(0.0001)	0.781	17.4	—	[−0.178, −0.158]
XGBoost, 100 rounds	783,872	0.764	(0.0004)	0.725	20.3	—	[−0.126, −0.107]
XGBoost + missingness indicators, nested	783,872	0.817	(0.0002)	0.783	17.3	1,000	[−0.181, −0.160]
LightGBM, nested rounds	783,872	0.799	—	0.762	18.5	1,000	[−0.163, −0.142]
Leave-one-state-out (51)							
OLS (62 regressors + lot indicator)	783,872	0.573	—	0.592	28.8	—	—
OLS, no region dummies	783,872	0.564	—	0.577	29.2	—	—
XGBoost, nested rounds	783,872	0.637	(0.0012)	0.657	25.9	153	[−0.076, −0.047]
XGBoost, 1,000 rounds	783,872	0.623	(0.0023)	0.666	25.6	—	[−0.072, −0.025]
XGBoost, 100 rounds	783,872	0.639	(0.0008)	0.653	25.7	—	[−0.077, −0.052]
XGBoost + missingness indicators, nested	783,872	0.632	(0.0020)	0.661	25.9	150	[−0.071, −0.040]
LightGBM, nested rounds	783,872	0.643	—	0.659	25.7	261	[−0.085, −0.057]

Notes: One row per model and design (a condensed one-row-per-design version is generated as `paper/tables/v2_headline.tex`). R^2_{within} is the oracle within-state R^2 of Eq. (8), which removes the state means of actual and predicted log prices in the test set (it uses the held-out state's actual mean and is therefore not a feasible forecast; Table 23 gives the feasible version). "Rounds" is the mean number of boosting rounds selected by the inner GroupKFold(3) rule; fixed-round rows reproduce the v1.1 comparators (the 1,000-round ten-state value of v1.1, 0.425, is reproduced). SD across seeds 42–46 for XGBoost; LightGBM and OLS are single fits. The ZIP3-cluster bootstrap interval refers to $R^2(\text{OLS}) - R^2(\text{model})$.

Table 20. Blocked Five-Fold Cross-Validation at Three Spatial Scales: R^2 by Fold and Pooled

Design and model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Pooled R^2	R^2_{within}
GroupKFold(5) by ZIP5							
OLS	0.647	0.646	0.652	0.651	0.638	0.647	0.616
XGBoost, nested rounds	0.819	0.815	0.823	0.808	0.810	0.815	0.781
XGBoost, 1,000 rounds	0.819	0.815	0.823	0.808	0.810	0.815	0.781
GroupKFold(5) by ZIP3							
OLS	0.607	0.618	0.658	0.624	0.658	0.635	0.606
XGBoost, nested rounds	0.757	0.759	0.777	0.754	0.780	0.767	0.727
XGBoost, 1,000 rounds	0.760	0.759	0.777	0.756	0.780	0.768	0.729
GroupKFold(5) by state							
OLS	0.456	0.598	0.617	0.638	0.563	0.580	0.587
XGBoost, nested rounds	0.510	0.592	0.668	0.717	0.623	0.628	0.651
XGBoost, 1,000 rounds	0.439	0.625	0.622	0.714	0.651	0.615	0.659

Notes: Trimmed v2 sample; each model is trained on four folds and scored on the fifth; "Pooled" concatenates the out-of-fold predictions. XGBoost values are means over seeds 42–46. The ensemble advantage shrinks as the block widens (Table 19). Generated for items WP03-ADVE-03 and ECON-07.

Table 21. Leave-One-State-Out Validation on the v2 Sample: Distribution of Held-Out-State Accuracy Across the 51 Jurisdictions

Model	Median	Weighted mean	Mean	p_{10}	p_{90}	Min	Max	$R^2 > 0.5$	Median |bias|
OLS	0.551	0.505	0.516	0.376	0.623	0.129	0.711	33	0.127
XGBoost, nested rounds	0.634	0.578	0.603	0.441	0.727	0.220	0.797	42	0.105
XGBoost, 1,000 rounds	0.659	0.565	0.604	0.407	0.741	0.001	0.767	43	0.101
XGBoost, 100 rounds	0.629	0.581	0.600	0.448	0.726	0.224	0.772	42	0.112
XGBoost + missingness, nested	0.649	0.573	0.605	0.434	0.737	0.122	0.793	42	0.096
LightGBM, nested rounds	0.641	0.583	0.599	0.371	0.730	0.094	0.794	43	0.102

Notes: 51 models per specification, each trained on every listing outside the held-out jurisdiction (trimmed v2 sample); XGBoost rows are means over seeds 42–46. "Weighted mean" weights states by their number of listings (the pooled-R^2 view of Table 19 is also listing-weighted); "Median |bias|" is the median across states of the absolute mean error of predicted minus actual log price, i.e. the level error. Paired state-level tests are in Table 22. No state has a negative R^2 for any model. The v1.1 full model (uncorrected design, 1,000 rounds) had a median of 0.653 and a listing-weighted mean of 0.564 with 43 states above 0.5 (the text said 44). State-level values in Table 41 and Figure 8; housing-unit weights (ACS) were not computed (Section 8).

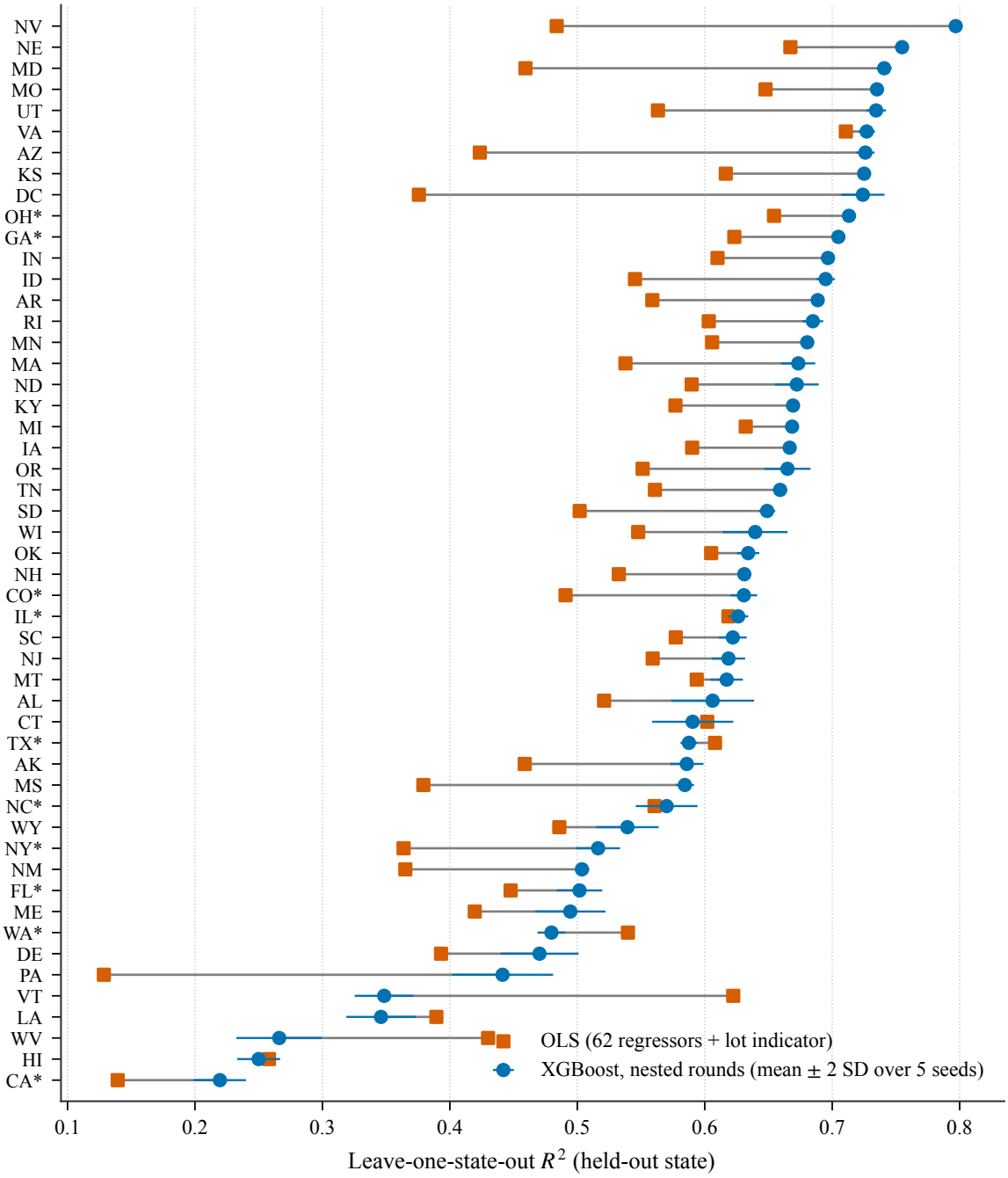


Figure 8. Leave-one-state-out R^2 by held-out state, OLS versus XGBoost with nested round selection (trimmed v2 sample; each state is predicted by models trained on the other fifty jurisdictions). States are sorted by the XGBoost value; a horizontal segment joins the two models; XGBoost bars are ±2 seed standard deviations (seeds 42–46; OLS is deterministic). Starred states form the ten-state holdout of v1.1. XGBoost is more accurate in 44 of 51 states (sign test $p = 1.2e-07$; Table 21); state-level values in Table 41.

Table 22. Paired State-Level Comparison of OLS and the Tree Ensembles Across the 51 Leave-One-State-Out Fits

	R^2(OLS) − R^2(comparator)			States	p-value		
Comparator	Mean	Median	Weighted	OLS better	Sign	Wilcoxon	Paired t
XGBoost, nested rounds	−0.087	−0.082	−0.073	7 of 51	1.2e-07	5.1e-07	6.3e-07
XGBoost, 1,000 rounds	−0.088	−0.099	−0.060	11 of 51	5.7e-05	1.4e-05	1.0e-06
XGBoost, 100 rounds	−0.083	−0.076	−0.076	5 of 51	2.3e-09	3.1e-07	1.2e-06
XGBoost + missingness, nested	−0.089	−0.084	−0.068	6 of 51	1.8e-08	4.2e-07	1.2e-06
LightGBM, nested rounds	−0.083	−0.082	−0.078	7 of 51	1.2e-07	1.2e-05	4.9e-06

Notes: Differences of held-out-state R^2 between OLS and each comparator over the 51 jurisdictions of Table 41 (XGBoost values are means over seeds 42–46); "Weighted" weights states by their number of listings. "States OLS better" counts jurisdictions where OLS has the higher R^2. p-values: exact two-sided sign test, Wilcoxon signed-rank test and paired t-test over the states; the states are treated as independent draws, which ignores the overlap of their training sets.

close to the pooled R^2. Under the ten-state holdout the fully grown 1,000-round XGBoost model leaves 25.0% of the variance as state-level error and 32.5% as within-state error, against 8.5% and 37.4% for OLS: the ensemble predicts within-state structure better than OLS (oracle within-state R^2 of Eq. (8), 0.628 against 0.573) and the level of the withheld states much worse. Selecting the rounds inside the training states reduces the ensemble's level share to 17.6%, and under leave-one-state-out, where each state is predicted from fifty others, the level shares of XGBoost and OLS are 7.7% and 8.3%. Within-market accuracy is not constant across designs: the oracle within-state R^2 of XGBoost is 16.5 points lower under the ten-state holdout than under the random split, so the statement that the ensemble transfers within-market structure intact is not supported; it transfers it partly, and the level is what it gets most wrong.

Because the dominant error is a level, it is cheaply correctable. Table 23 reports the feasible level correction of Section 4.5. Under leave-one-state-out, estimating each held-out state's mean error from five random local listings raises XGBoost's pooled R^2 from 0.635 (the seed-42 fit on which the correction is computed; the five-seed mean is 0.637) to 0.658, twenty listings to 0.699, fifty to 0.706, and two hundred to 0.711, against an oracle of 0.712; twenty listings recover 83% of the oracle gain. OLS gains similarly (0.573 to 0.641 with twenty listings), and under the ten-state holdout the fully grown ensemble moves from 0.425 to 0.661. An AVM deployed in a new market with a handful of local listings therefore recovers most of its within-market accuracy, and the ordering of models after the correction is the ordering of their within-market accuracy, which favours the ensembles; Figure 9 plots the recovery curves against k.

Table 23. A Feasible Level Correction: Held-Out-State R^2 After Re-Centring Predictions on k Observed Listings of the Held-Out State

	Raw	Feasible, k calibration listings (median over draws; [p_{25}, p_{75}])					Oracle
Design and model	R^2	$k = 5$	$k = 20$	$k = 50$	$k = 100$	$k = 200$	R^2
Leave-one-state-out (51)							
OLS	0.573	0.598 [0.580, 0.610]	0.641 [0.637, 0.643]	0.650 [0.648, 0.651]	0.653 [0.652, 0.654]	0.655 [0.654, 0.655]	0.656
XGBoost, nested rounds	0.635	0.658 [0.644, 0.672]	0.699 [0.695, 0.702]	0.706 [0.705, 0.707]	0.709 [0.709, 0.710]	0.711 [0.710, 0.711]	0.712
XGBoost, 1,000 rounds	0.623	0.668 [0.657, 0.680]	0.706 [0.702, 0.709]	0.715 [0.713, 0.716]	0.718 [0.717, 0.718]	0.719 [0.718, 0.719]	0.720
Ten-state holdout							
OLS	0.541	0.564 [0.541, 0.584]	0.610 [0.602, 0.616]	0.619 [0.616, 0.622]	0.623 [0.621, 0.624]	0.624 [0.624, 0.625]	0.626
XGBoost, nested rounds	0.506	0.629 [0.601, 0.652]	0.668 [0.659, 0.672]	0.677 [0.673, 0.679]	0.680 [0.679, 0.681]	0.681 [0.680, 0.681]	0.682
XGBoost, 1,000 rounds	0.425	0.622 [0.598, 0.641]	0.661 [0.653, 0.665]	0.669 [0.667, 0.671]	0.672 [0.671, 0.673]	0.673 [0.673, 0.674]	0.675

Notes: The level error of an out-of-state model is corrected by adding the mean residual on k listings drawn at random from the held-out state (listings excluded from the scored set); 100 random draws per k, median and interquartile range of the resulting pooled R^2. "Oracle" re-centres on the full held-out state (Eq. (8), not feasible). With $k = 20$ listings most of the oracle gain is recovered for every model, so the level channel is a practical, not a structural, limit. Generated for items WP03-THEO-01, THEO-02 and ECON-05.

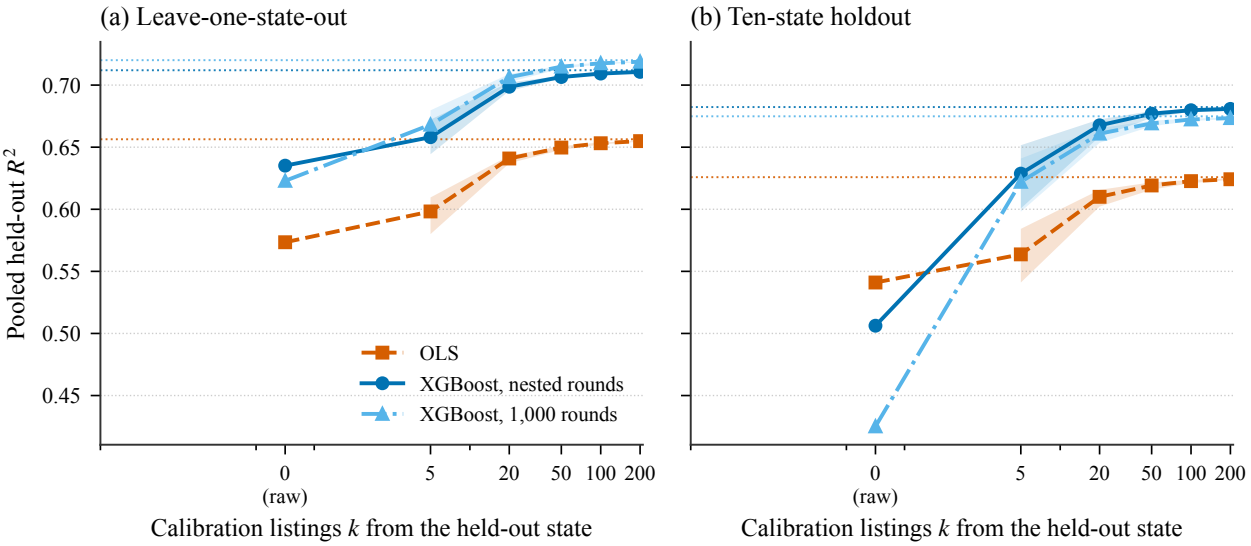


Figure 9. Feasible level correction of out-of-state predictions. Each curve is the pooled held-out R^2 after adding to every prediction the mean residual on k listings drawn at random from the held-out state (median over 100 draws; shaded interquartile range; $k = 0$ is the raw model); dotted horizontal lines are the oracle corrections that use the full held-out state (Table 23). The x-axis is symmetric-log. Twenty observed listings recover most of the oracle gain for every model under both designs.

5.7.2. Boosting Rounds and Region Dummies

Figure 10 and Table 24 trace held-out-state R^2 of the full XGBoost model as a function of boosting rounds under the ten-state holdout, for five seeds. Accuracy on the withheld states peaks after 65–70 rounds at a mean of 0.545 and declines to 0.425 (seed standard deviation 0.015) at 1,000 rounds, while R^2 on the early-stopping subset of the training states and on the random test set keep rising. Beyond roughly 100 rounds every additional tree improves the fit within the training states and degrades it outside them; early stopping on a random subset of the training data cannot detect this because the evaluation set shares the training states' price levels. The peak itself is a test-selected number: the nested rule, which sees only the training states, selects 175 rounds and reaches 0.515, 3.0 points below the peak. The pattern is the national-scale counterpart of the observation of Deppner and Cajias (2024) that random resampling selects overly complex boosters, and of the spatial-tuning problem of Schratz et al. (2019).

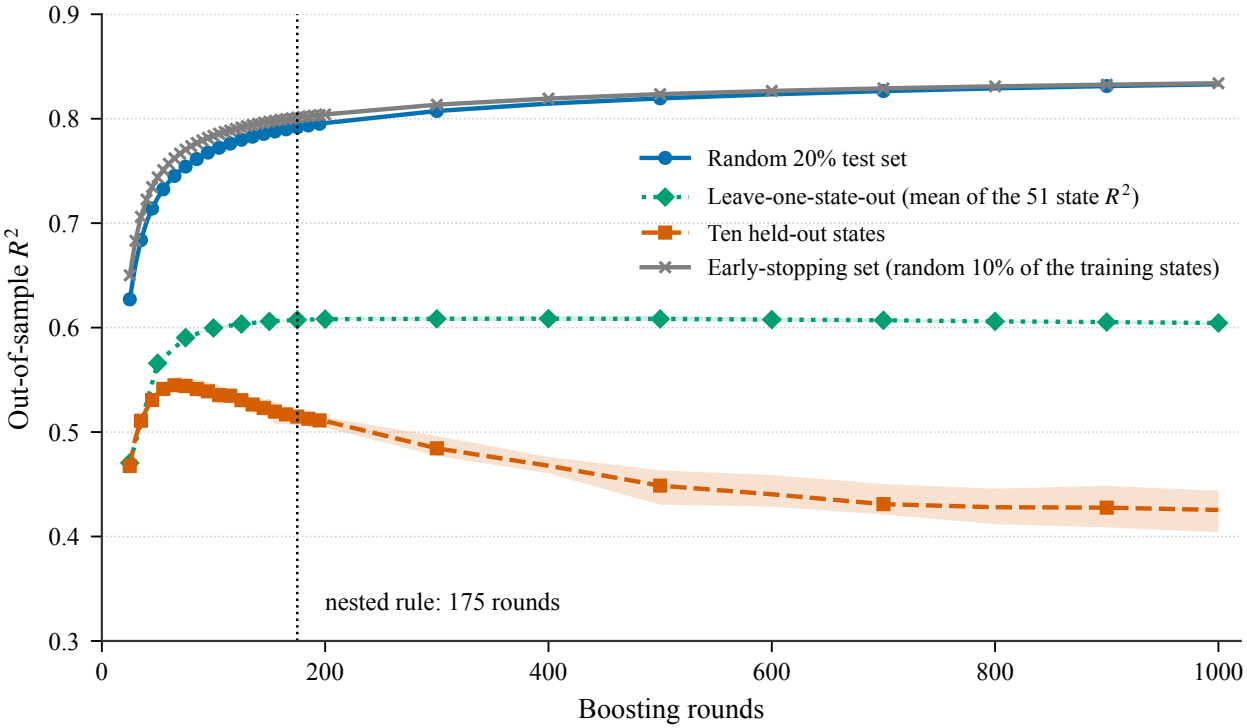


Figure 10. Out-of-sample R^2 of the full XGBoost model as a function of the number of boosting rounds (trimmed v2 sample). Lines are means over seeds 42–46 (for leave-one-state-out, the unweighted mean of the 51 held-out-state R^2 values, not the pooled out-of-fold R^2) and shaded bands the min–max range across seeds; the dotted vertical line marks the number of rounds (175) chosen by the inner GroupKFold(3) rule on the training states for the ten-state design. The held-out-state curve peaks between 65 and 70 rounds depending on the seed and declines to 0.425 at 1,000 rounds, while the early-stopping set drawn from the training states keeps rising; the peak is not observable without the test states (Table 24).

The capacity effect is not a property of boosting alone; it operates through the region dummies.

Table 24. Out-of-Sample R^2 of the Full XGBoost Model by Number of Boosting Rounds: Random Test Set, Early-Stopping Set, Leave-One-State-Out and Ten Held-Out States (Five Seeds)

	Random	Early-stopping	LOSO (mean of	Ten held-out states		
Rounds	test set	set	51 state R^2)	Mean	(SD)	Min–max
25	0.627	0.650	0.470	0.468	(0.0029)	0.464–0.472
50	0.724	0.744	0.566	0.536	(0.0036)	0.531–0.541
75	0.754	0.771	0.590	0.544	(0.0057)	0.538–0.552
100	0.770	0.784	0.600	0.538	(0.0070)	0.531–0.546
125	0.780	0.792	0.603	0.531	(0.0053)	0.525–0.537
150	0.787	0.797	0.606	0.522	(0.0073)	0.513–0.532
175	0.792	0.801	0.607	0.515	(0.0052)	0.506–0.519
200	0.796	0.804	0.608	0.511	(0.0038)	0.504–0.514
300	0.807	0.813	0.608	0.484	(0.0072)	0.477–0.496
500	0.819	0.824	0.608	0.449	(0.0138)	0.430–0.463
700	0.826	0.829	0.607	0.431	(0.0113)	0.421–0.450
1000	0.833	0.834	0.604	0.425	(0.0146)	0.404–0.444

Notes: Trimmed v2 sample, seeds 42–46 (SD and range across seeds for the held-out states). The random-test and early-stopping columns come from the model trained on the random 80% split (the early-stopping set is a random 10% of the training listings, drawn from the training states); the LOSO column is the unweighted mean, over the 51 leave-one-state-out fits, of the held-out state's R^2 (not the pooled out-of-fold R^2 of Table 19, which weights states by their listings); the held-out-state columns come from the model trained on the forty remaining states. Across seeds the held-out-state curve peaks between 65 and 70 rounds (mean peak R^2 0.545), a value selected on the test states; the nested rule applied to the training states selects 175 rounds and reaches 0.515 (SD 0.005), so the v1.1 "peak at 75 rounds, 0.550" was a test-selected number (Appendix F). Figure 10 plots the curves.

Table 25 crosses the rounds rule with the removal of the region dummies and of the school variables under each geographic design. Without the region dummies, ten-state R^2 is essentially flat across rounds rules (0.527, 0.520, 0.521; a range of 0.8 points, within the seed standard deviation of the full model at 1,000 rounds), whereas with them the full model falls from 0.538 at 100 rounds to 0.425 at 1,000. Under leave-one-state-out the no-region curve is again flat (range 1.1 points) while the full model loses 1.3 points from the nested to the 1,000-round rule. Trees use the region dummies to partition the training states' price levels, and additional rounds refine that partition in a way that is wrong by construction for a held-out state. The school variables behave differently: removing them lowers out-of-state R^2 under every rule (ten-state nested 0.490 against 0.515; leave-one-state-out 0.622 against 0.637), except that at 100 rounds under the ten-state holdout the difference (0.536 against 0.538, seed SD 0.007) is within seed noise; so they carry transferable information despite being state medians for most listings. The earlier reading that the imputed school variables *cause* the extrapolation failure is therefore not supported by the design's own ablation; what they do is inflate random-split accuracy, as the next subsection shows.

Table 25. Region Dummies, School Variables and Boosting Capacity Are One Interaction: XGBoost Variants by Number of Rounds Under Three Geographic Designs

	Nested rounds		100 rounds		1,000 rounds		R^2_{within}
Design and variant	R^2	(SD)	R^2	(SD)	R^2	(SD)	(nested)
Ten-state holdout							
Full model (region dummies, school variables)	0.515	(0.005)	0.538	(0.007)	0.425	(0.015)	0.637
Without region dummies	0.527	—	0.520	—	0.521	—	0.627
Without school variables	0.490	—	0.536	—	0.407	—	0.621
Without school variables and region dummies	0.516	—	0.513	—	0.488	—	0.607
Full + 6 missingness indicators	0.523	(0.010)	0.560	(0.007)	0.458	(0.019)	0.640
GroupKFold(5) by state							
Full model (region dummies, school variables)	0.628	(0.001)	0.635	(0.002)	0.615	(0.003)	0.651
Without region dummies	0.623	—	0.614	—	0.630	—	0.656
Without school variables	0.622	—	0.626	—	0.589	—	0.652
Without school variables and region dummies	0.593	—	0.591	—	0.577	—	0.640
Full + 6 missingness indicators	0.634	(0.002)	0.640	(0.001)	0.620	(0.004)	0.655
Leave-one-state-out (51)							
Full model (region dummies, school variables)	0.637	(0.001)	0.639	(0.001)	0.623	(0.002)	0.657
Without region dummies	0.624	—	0.614	—	0.624	—	0.658
Without school variables	0.622	—	0.622	—	0.598	—	0.654
Without school variables and region dummies	0.593	—	0.593	—	0.576	—	0.647
Full + 6 missingness indicators	0.632	(0.002)	0.643	(0.001)	0.614	(0.009)	0.661

Notes: Trimmed v2 sample; mean (SD) over seeds 42–46 where several seeds were run, single seed otherwise. "School variables" are the average school rating, school count and nearest-school distance (state medians for most listings); "Nested" selects the rounds by inner GroupKFold(3) on the training groups. Without the region dummies the R^2 no longer falls with the number of rounds (the curve is flat within seed noise), so the capacity loss of v1.1 operates through the region dummies; removing the school variables lowers held-out accuracy under every design. Generated for items WP03-THEO-02, THEO-04 and EDIT-03.

5.7.3. Ablation: What Drives the Random-Split Gain?

Table 26 reports the nested ablation of XGBoost under both designs with five seeds. Eight structural variables alone give $R^2 = 0.497$ under random validation and 0.344 in the held-out states, so basic size and room counts carry predictive content that transfers across markets. The neighbourhood block (Walk, Bike and Transit Scores, the three school variables and the tax rate) contributes by far the largest random-validation gain, 16.0 points, but only 3.8 points under the ten-state design. The split of that block identifies the state-level encodings. Added to the amenity model on their own, the tax rate—which takes 243 distinct values across the country—is worth 8.6 points in sample and −7.2 out of state; fifty state dummies are worth 10.2 and −0.4; and latitude and longitude, the most flexible location encoding, are worth 20.0 and −5.1. Replacing the three school variables in the neighbourhood block by a bare state identifier leaves random accuracy nearly unchanged but lowers ten-state accuracy by 5.1 points, and removing Walk, Bike and Transit Scores lowers both. The school block therefore supplies something a state identifier does not: state median ratings are correlated with state price levels, so a held-out state's imputed rating is routed to the leaf of the training state with the nearest rating and produces a crude but informative guess of the unseen level. Finally, the full model loses 8.4 points of ten-state R^2 relative to stage 5 at 1,000 rounds while gaining 0.9 under random validation; the region dummies are responsible, as Table 25 shows, and the full model with coordinates reaches 0.871 in sample but only 0.477 out of state. The state-level information in the features thus helps in sample in proportion to how precisely it identifies the state, and transfers out of state only when it is a continuous proxy for the level rather than a partition; Figure 11 shows the cascade under both designs with the seed ranges.

5.8. SHAP Analysis and Cross-Model Stability

Table 27 reports the twenty features with the largest mean absolute SHAP values in the XGBoost model, with the corresponding values for LightGBM and the random forest. Living area (0.182) and bathrooms (0.157) dominate; school rating (0.087), lot size (0.076), the West dummy (0.073), the tax rate (0.070) and the luxury score (0.063) form a second tier. The ranking must be read with Table 4 in hand: school rating, the region dummies, nearest-school distance and the tax rate are location identifiers rather than property attributes, and together they are the channel through which the model encodes state price levels. The 3rd-ranked "school rating" ranks where it does because it is, for 93.0% 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 12 shows the distribution of the contributions of the leading features and Figure 13 how six of them vary with the feature value: the concave living-area and saturating bathroom contributions are non-linearities the

Table 26. XGBoost Feature Ablation on the v2 Design with Seed Dispersion: Cumulative Cascade and Decomposition of the Neighbourhood Block

		Random R^2		Ten-state R^2		Δ (pp)	
Specification	Features	Mean	(SD)	Mean	(SD)	Random	Ten-state
Panel A: cumulative cascade							
1. Structural only	8	0.497	(0.000)	0.344	(0.001)	—	—
2. + Lot (recoded + indicator)	10	0.574	(0.000)	0.365	(0.001)	7.7	2.1
3. + Amenities	21	0.654	(0.000)	0.447	(0.002)	8.0	8.2
4. + Neighbourhood	28	0.814	(0.000)	0.486	(0.005)	16.0	3.8
5. + Market status	33	0.824	(0.000)	0.510	(0.005)	1.0	2.4
6. Full (+ interactions, categorical, region)	63	0.833	(0.000)	0.425	(0.015)	0.9	−8.4
Panel B: decomposition of the neighbourhood block (4a–4e, Δ relative to stage 3) and of the full model (6a–6d, Δ relative to stage 6)							
4a. Stage 3 + tax rate only	22	0.740	(0.000)	0.375	(0.004)	8.6	−7.2
4b. Stage 3 + latitude and longitude	23	0.855	(0.000)	0.396	(0.026)	20.0	−5.1
4c. Stage 3 + 50 state dummies	71	0.756	(0.000)	0.444	(0.001)	10.2	−0.4
4d. Stage 4, school block → state identifier	26	0.820	(0.000)	0.435	(0.004)	16.6	−1.3
4e. Stage 4 without Walk/Bike/Transit	25	0.797	(0.001)	0.448	(0.006)	14.3	0.1
6a. Full without region dummies	60	0.830	(0.000)	0.528	(0.006)	−0.3	10.3
6b. Full without school variables	60	0.818	(0.000)	0.410	(0.004)	−1.1	−11.8
6c. Full without school and region	57	0.799	(0.000)	0.489	(0.003)	−1.9	7.9
6d. Full + latitude and longitude	65	0.871	(0.000)	0.477	(0.019)	7.1	−1.1

Notes: XGBoost with 1,000 rounds and early stopping on a random 10% of the training set (the v1.1 protocol), mean and SD over seeds 42–46; random 80/20 split and ten-state holdout of the trimmed v2 sample. Panel A adds one block per row; Δ is the change from the previous row. The neighbourhood block (Walk, Bike and Transit Scores, three school variables, tax rate) adds +16.0 pp under random validation and +3.8 pp out of state. Panel B isolates its components: each 4x row is stage 3 plus one block (or stage 4 with one substitution) and Δ is measured from stage 3; rows 6a–6d modify the full model and Δ is measured from stage 6. The 100-round variants of every specification are in `results/geo/ablation.csv`. Generated for items WP03-EDIT-03, THEO-04 and ADVE-21.

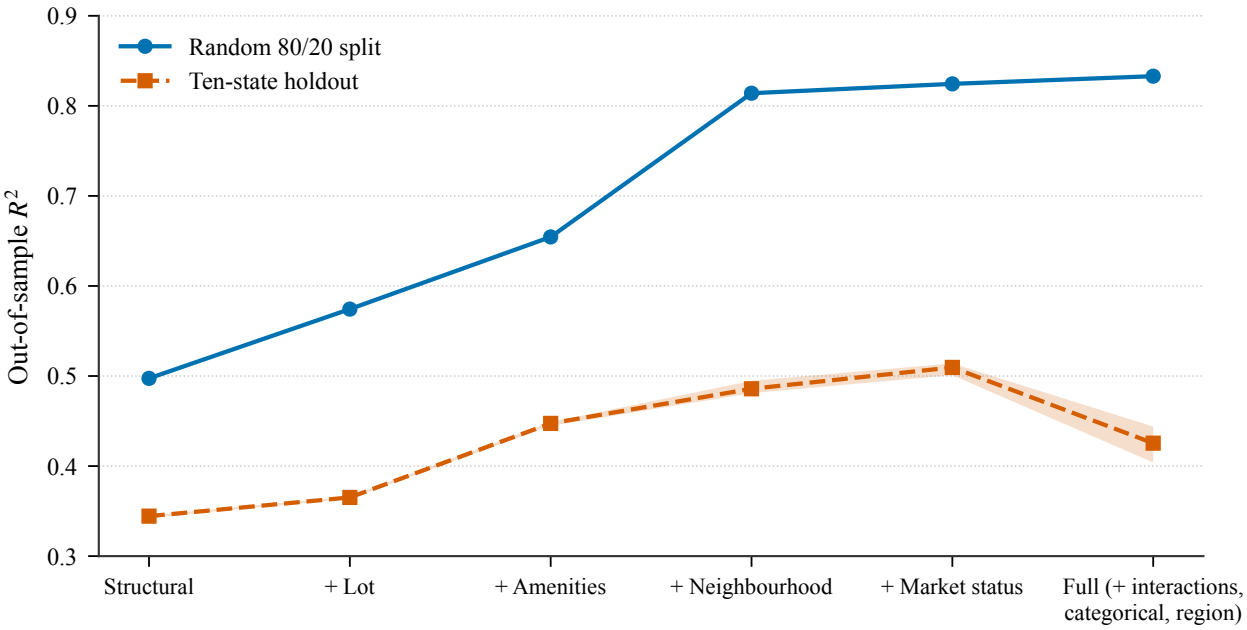


Figure 11. Cumulative feature-ablation cascade for XGBoost (1,000 rounds with early stopping, trimmed v2 sample): mean out-of-sample R^2 over seeds 42–46 with the min–max range shaded (Table 26). The neighbourhood block adds 16.0 pp under random validation but 3.8 pp out of state, and the full model loses out-of-state accuracy when the region dummies enter; the seed ranges are too narrow to be visible under the random split.

semi-log form cannot capture, while the step pattern of the school rating reflects its state-median support.

Table 27. Twenty Largest Mean Absolute SHAP Values of XGBoost on the v2 Design, with the Same Features in LightGBM and the Random Forest

		XGBoost	LightGBM		Random forest	
Rank	Feature	Mean $\|\phi\|$	Mean $\|\phi\|$	Rank	Mean $\|\phi\|$	Rank
1	ln(living area)	0.1817	0.1721	1	0.1570	2
2	Bathrooms	0.1567	0.1520	2	0.2195	1
3	Avg. school rating¶	0.0875	0.0860	3	0.1090	3
4	ln(lot size)	0.0762	0.0741	5	0.0546	7
5	Region: West	0.0725	0.0780	4	0.0911	5
6	Property tax rate (%)	0.0695	0.0642	7	0.0488	8
7	Luxury score (0–7)	0.0625	0.0704	6	0.1056	4
8	Nearest-school distance¶	0.0444	0.0494	8	0.0322	13
9	Region: Northeast	0.0428	0.0474	9	0.0218	21
10	Age (years)	0.0402	0.0389	12	0.0253	17
11	ln(1 + HOA fee)	0.0395	0.0390	11	0.0456	9
12	Bike Score	0.0393	0.0417	10	0.0323	12
13	Garage	0.0372	0.0386	13	0.0299	16
14	Hardwood floors	0.0319	0.0359	14	0.0188	25
15	Region: South	0.0306	0.0186	20	0.0327	11
16	Waterfront × ln(living area)	0.0277	0.0281	16	0.0301	15
17	Living area per bedroom	0.0275	0.0327	15	0.0665	6
18	Central air	0.0269	0.0280	17	0.0215	22
19	Walk Score	0.0266	0.0258	18	0.0319	14
20	Transit Score	0.0227	0.0214	19	0.0188	26

Notes: TreeSHAP values (log-price units) on the 9,937 stored explanation listings of the random test set that survive the v2 trimming; XGBoost and LightGBM exact TreeSHAP, random forest with the Saabas approximation. Mean $|\phi|$ is the average absolute contribution of the feature to the deviation of the predicted log price from its mean; these are attributions of a fitted prediction function, not implicit prices. ¶State median for most listings (Table 4), so the feature identifies the state rather than school quality. Rank correlations over all 63 features are in Table 28. The v1.1 ranking (uncorrected design) had the same top three.

Table 28 reports pairwise Spearman rank correlations of the mean absolute SHAP values across the three ensembles. XGBoost and LightGBM produce nearly identical rankings ($\rho = 0.987$); the random forest correlates at 0.905 and 0.912 with the two boosters, and the same leading features rank at the top for all three. Stability across algorithms establishes that the ranking describes the predictive structure of the data as encoded in these features; it does not convert any entry into an implicit price, and in the case of the school variables it reflects the stability of a state identifier that every algorithm finds useful.

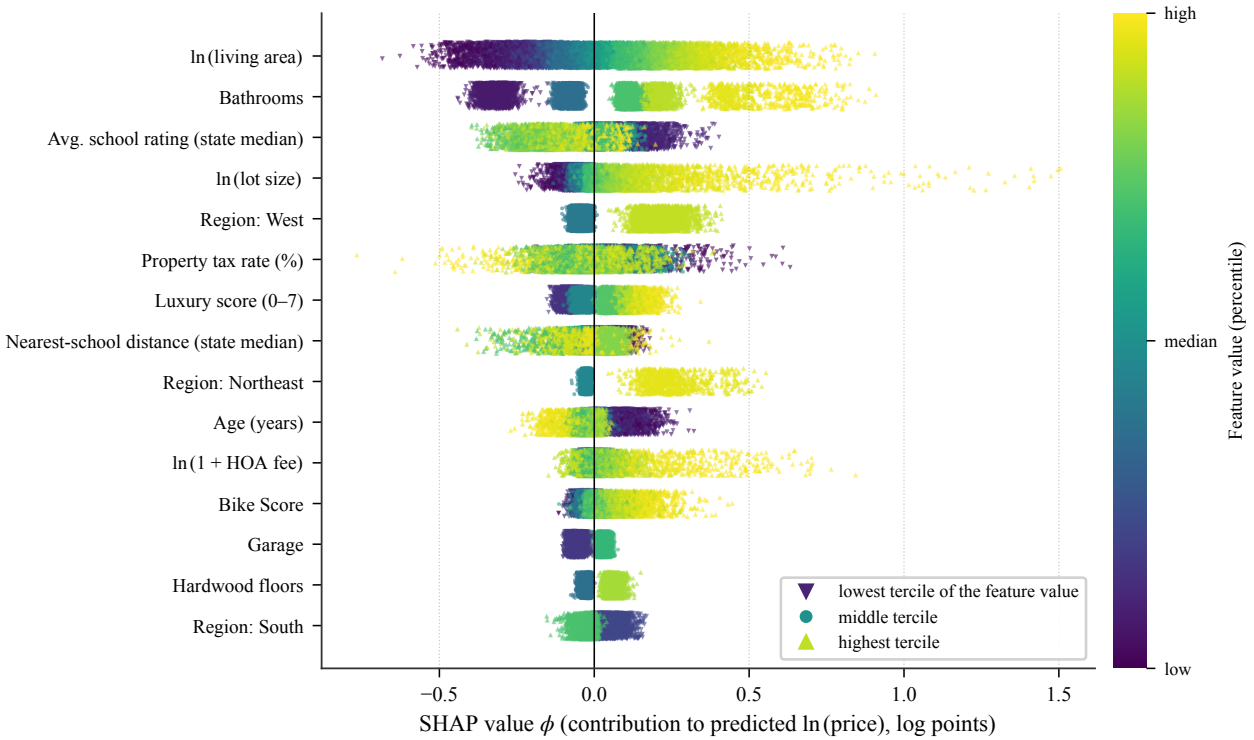


Figure 12. SHAP summary plot of the XGBoost model on the v2 design (TreeSHAP on the 9,937 stored explanation listings of the random test set; the fifteen features with the largest mean $|\phi|$, Table 27). Each point is one listing; its horizontal position is the feature's contribution to the predicted log price; the percentile rank of the feature value is encoded twice, by colour (viridis: dark = low, yellow = high) and by marker shape (triangle down = lowest tercile, circle = middle, triangle up = highest), so that the contrast is legible in grayscale. Contributions describe the fitted prediction function, not implicit prices; the school rating is a state median for most listings and acts as a state identifier.

Table 28. Cross-Model Stability of SHAP Importance: Spearman Rank Correlations of the 63 Mean Absolute SHAP Values

Pair of models	Spearman ρ
XGBoost vs. LightGBM	0.987
XGBoost vs. random forest	0.905
LightGBM vs. random forest	0.912

Notes: Spearman rank correlation of the mean absolute SHAP values of the 63 features between pairs of tree ensembles, computed on the same 9,937 explanation listings as Table 27 (v2 design; random forest with the Saabas approximation). The v1.1 values on the uncorrected design were 0.990, 0.906, 0.892.

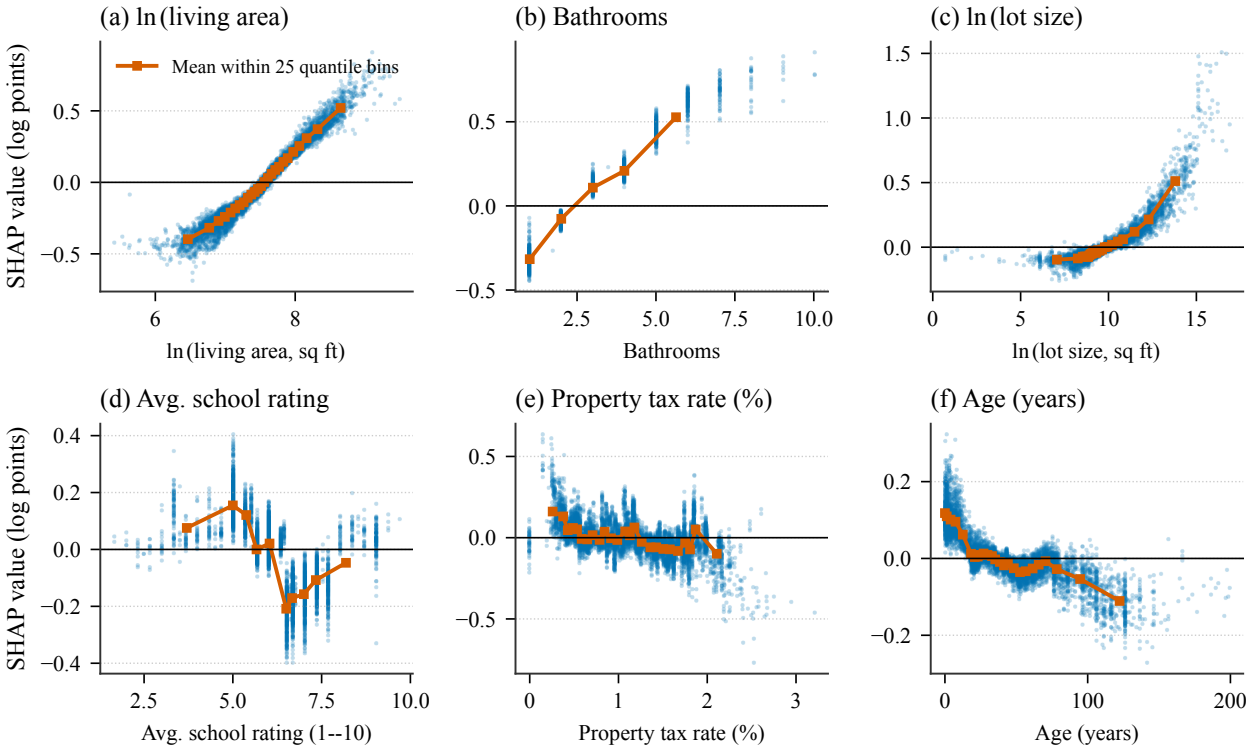


Figure 13. SHAP dependence of the XGBoost model on six features (v2 design, 9,937 explanation listings): each point is one listing's contribution against the feature value in natural units; the line joins the mean contribution within 25 quantile bins. The concave living-area and saturating bathroom contributions are non-linearities the semi-log form cannot capture; the step pattern of the school rating reflects its state-median support (Table 4).

6. Robustness Checks

6.1. Lot-Size Coding and Winsorisation

Table 29 compares five codings of lot size in the natural-unit baseline, on the full and on the trimmed sample. Under the coding that set ln(lot) to zero for the 130,271 listings whose lot was missing, zero, one square foot or negative, the lot elasticity on the full sample is 0.018 and R^2 is 0.634; replacing the zeros by the log of the state-median lot raises the elasticity to 0.096 and R^2 to 0.648, adding the lot-missing indicator leaves the elasticity unchanged and attaches a discount of about 2.1% to listings without a reported lot, restricting the sample to listings with a reported positive lot gives 0.097, and winsorising the recoded lot at its 99.5th percentile gives 0.097. The living-area elasticity falls from the uncorrected to the corrected coding because part of the lot gradient had been loading on living area. The mechanism is not imputation but measurement: 130,271 observations placed far below the mass of the data acted as a missing-value dummy and pulled the slope toward zero. The corrected coding is the design matrix used by every model in the paper, including the quantile regressions of Section 5.4, whose lot path under the corrected coding rises from 0.101 to 0.147; the earlier quantile estimates on the uncorrected design, which also carried the corrupt parking column, are reproduced and documented in Appendix F.

Table 29. Lot-Size Coding and the Lot-Size Elasticity of the Asking Price (Natural-Unit OLS)

Coding	N	R^2	$\hat{\beta}_{\ln \text{lot}}$	(SE)	$\hat{\beta}_{\ln \text{sqft}}$	Indicator
Panel A: full sample						
v1.1: ln(lot) = 0 if missing, zero, 1 sq ft or negative	788,842	0.634	0.0182	(0.0003)	0.628	—
v2: ln(state-median lot) when missing, zero or ≤ 1	788,842	0.648	0.0958	(0.0007)	0.585	—
v2 + lot-missing indicator (design used throughout)	788,842	0.648	0.0959	(0.0007)	0.587	−0.021
Complete cases (raw lot > 1 sq ft)	658,571	0.660	0.0974	(0.0007)	0.545	—
v2 + indicator, lot winsorised at $p_{99.5}$	788,842	0.648	0.0971	(0.0007)	0.585	−0.022
Panel B: trimmed sample (implausible listings removed)						
v1.1: ln(lot) = 0 if missing, zero, 1 sq ft or negative	783,872	0.636	0.0176	(0.0002)	0.619	—
v2: ln(state-median lot) when missing, zero or ≤ 1	783,872	0.649	0.0967	(0.0007)	0.585	—
v2 + lot-missing indicator (design used throughout)	783,872	0.649	0.0968	(0.0007)	0.587	−0.023
Complete cases (raw lot > 1 sq ft)	655,220	0.661	0.0982	(0.0007)	0.536	—
v2 + indicator, lot winsorised at $p_{99.5}$	783,872	0.648	0.0961	(0.0007)	0.586	−0.023

Notes: 62-regressor natural-unit specification; HC3 standard errors in parentheses (state-clustered errors for the design used throughout are in Table 9). "Missing, zero, 1 sq ft or negative" covers 130,271 listings (16.5%, Table 5); state medians are computed from reported lots above one square foot. "Indicator" is the coefficient on the lot-missing dummy. The v1.1 elasticity is a coding artefact: the 130,271 observations placed about eight log units below the mass of the data acted as a missing-value dummy.

6.2. Quantile-Regression Subsample Stability

Table 30 repeats the median regression on ten random subsamples of 150,000 listings of the trimmed sample and reports the coefficient of variation and sign stability of the key coefficients; all ten fits

converged for every variable. Of the 15 variables tracked, 13 have a coefficient of variation below 20% and the same sign in every subsample. The exceptions are waterfront (coefficient of variation 36.2%) and age (179%, with the sign of its mean in 60% of subsamples). The subsamples are random draws from the national sample with the same geographic composition, so the instability is a precision problem rather than geographic heterogeneity: the median-regression age coefficient is near zero (0.069) because age enters together with its square and two interactions that are nearly collinear once 18.9% of ages are set to a state median, and a coefficient of that size will change sign in a substantial share of random subsamples. The waterfront coefficient is imprecise for the reason Section 5.1.4 makes explicit: the attribute is concentrated in a few markets, and its state-cluster t-statistic in the OLS model is −0.53. The stability table thus agrees with the clustered inference on which coefficients the data pin down.

Table 30. Median-Regression ($\tau = 0.50$) Coefficient Stability Across Ten Random Subsamples of the v2 Design

Variable	Mean	SD	CV (%)	Sign stability (%)	Converged
ln(lot size)	0.123	0.001	1.1	100	10/10
Region: West	0.384	0.005	1.2	100	10/10
Region: Northeast	0.447	0.007	1.5	100	10/10
Bathrooms	0.246	0.005	2.0	100	10/10
Garage	0.106	0.002	2.0	100	10/10
ln(living area)	0.253	0.009	3.6	100	10/10
Foreclosure	−0.348	0.021	6.0	100	10/10
Parking spaces	0.024	0.002	6.2	100	10/10
Age^2	0.049	0.003	7.0	100	10/10
Bedrooms	−0.056	0.007	11.6	100	10/10
Pool	0.057	0.007	12.7	100	10/10
Lot missing (indicator)	−0.033	0.005	14.6	100	10/10
Luxury score (0–7)	0.013	0.003	19.7	100	10/10
Waterfront	−0.302	0.109	36.2	100	10/10
Age (years)	0.016	0.028	178.8	60	10/10

Notes: Ten random subsamples of 150,000 listings (seeds in `src/wp3/config.py`), standardised regressors. CV is the standard deviation divided by the absolute mean across subsamples; sign stability is the share of subsamples in which the coefficient has the sign of its mean. The age coefficient at the median is near zero (0.016), so its CV is large for a precision reason (Section 6.2), not because the subsamples differ geographically.

6.3. Duplicates Across the Random Split

Exact duplicates—builder plan listings of the same model home on different lots and relisted properties—are a non-spatial channel through which information can leak from training to test observations and inflate random-split accuracy, plausibly more for trees than for OLS. Table 31 removes from the random test set the 1,119 listings (0.71%) that have an exact duplicate in the training set. R^2 moves from 0.6456 to 0.6459 for OLS and from 0.8328 to 0.8326 for XGBoost;

the duplicates are predicted better than average by both models, but they are too few to affect the comparison.

Table 31. Exact Duplicate Listings Across the Random Split and Their Effect on Test-Set R^2

Model	N_{test}	With train duplicate	R^2, all	R^2, without	R^2, duplicates only
OLS	156,770	1,119	0.6456	0.6459	0.5931
XGBoost, nested rounds	156,770	1,119	0.8329	0.8326	0.8854
XGBoost, 1,000 rounds	156,770	1,119	0.8329	0.8326	0.8854

Notes: Exact duplicates on ZIP, living area, bedrooms, bathrooms, year built and price (6,787 listings in 3,152 groups, 0.86% of the sample; builder plan listings and re-listings). "With train duplicate" counts test listings that have an exact duplicate in the training set of the seed-42 random split of the trimmed v2 sample. Removing them changes R^2 by less than 0.001 for every model. Generated for item WP03-THEO-15.

6.4. Seed Dispersion of the Small Differences

Several differences between machine-learning variants are a few tenths of a point of R^2. Table 32 compares each such difference with the standard deviation of the difference implied by the five seeds. The training-size effect of 0.4 points between the random and the size-matched random split exceeds two seed standard deviations, as does the gain from the missingness indicators under random validation; the gains from the missingness indicators under the ten-state holdout (nested and 1,000 rounds) do not, and are therefore not interpreted.

Table 32. Are the Small Differences Quoted in v1.1 Larger Than Seed Noise? Differences in R^2 with Their Seed Standard Deviation

Difference	Value	SD of difference	Exceeds 2 SD
Size-matched minus random (1,000 rounds)	−0.0036	0.0004	yes
+ missingness minus full, ten-state (1,000 rounds)	0.0326	0.0236	no
+ missingness minus full, ten-state (nested)	0.0077	0.0116	no
+ missingness minus full, random (1,000 rounds)	0.0022	0.0003	yes

Notes: SD of the difference is $\sqrt{s_1^2 + s_2^2}$ from the seed standard deviations of the two means (five seeds each, trimmed v2 sample). Differences below two standard deviations are within seed noise and are not interpreted in the text. Generated for items WP03-EDIT-06, ECON-06 and THEO-13.

6.5. Summary

Table 33 collects the robustness evidence of the paper, with the section in which each check is reported; it is a cross-reference table whose every cell is an inline macro drawn from the generated exhibits, not a separately generated table. The additional inference checks—Conley spatial-HAC standard errors, the Kennedy correction under clustered variances, the reassignment of the listings without a ZIP code, and the ZIP3-cluster standard errors of the quantile coefficients—are in

Appendix B; the fold membership of the grouped designs and the full decomposition and gap tables are in Appendix C.

7. Discussion

The results support four messages, one per modelling framework, and a fifth that cuts across them. For each the estimates are read against the literature, with the evidence in this paper and its limits stated.

7.1. The Pooled Hedonic Regression Describes Averages, Not Markets

The semi-log OLS model explains 64.9% of the variation in log asking prices with 62 regressors, within the range 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 0.59, a bathroom differential of 24.8%, a negative conditional bedroom gradient, a lot elasticity of 0.10 once the lot variable is coded as intended, and a foreclosure discount of 32.3%, above the forced-sale discount of Campbell et al. (2011) as expected for listings positioned below market before negotiation. Regularised linear benchmarks reproduce OLS, so the specification is not overfit.

Two results 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 (median state-cluster ratio 11.7), and because the 51 state clusters are very unequal, cluster-robust standard errors with normal critical values over-reject as well: the wild cluster bootstrap leaves 23 of 63 slopes significant where HC3 left 58 and state clustering 33. The gradients that survive are the structural ones—living area, bathrooms, garage, hardwood, central air, foreclosure, the Northeast and West differentials—and the ones that do not are 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. The evidence concerns this design; it suggests, without establishing, that conventional robust standard errors on other national scraped-data hedonic regressions deserve the same scrutiny.

Second, aggregation. The state-by-state estimates show that the pooled coefficients are averages of markedly different local price functions: the living-area elasticity has an inter-decile range of 0.32–0.72 across jurisdictions, a cluster-robust test rejects slope homogeneity for every attribute, and 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. Sign reversals that survive inference are rarer than point estimates suggest (between 3 and 12 states for the four attributes with the most reversals), but the heterogeneity in magnitude 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):

Table 33. Robustness Summary

Check	Baseline	Alternative	Finding
OLS specification and inference			
Region FE → state FE (Section 5.1.3)	$R^2 = 0.646$	0.690	+4.5 pp out of sample
Region FE → ZIP3 FE	$R^2 = 0.646$	0.762	+11.6 pp out of sample
State FE → state-by-state (Section 5.3)	0.690	0.752	Slopes differ across markets (+6.1 pp)
HC3 → state cluster → wild bootstrap (Section 5.1.4)	58 of 63 slopes	33; wild 23	Median SE ratio 11.7; CR1 not conservative
Lot-size coding (Section 6.1)	$\hat{\beta}_{\ln \text{lot}} = 0.018$	0.097	Zero-coding artefact
Winsorisation of the recoded lot (p99.5)	0.097	±0.001	Immaterial
Spatial diagnostics			
Moran's I, OLS vs XGBoost residuals (Section 5.2)	0.471	0.236	Trees absorb about half of the spatial structure
Conley HAC 25–100 km (Appendix B)	HC0	×2.4–3.5	Between HC and state cluster (×4.2)
Quantile regression			
Stability over ten subsamples (Section 6.2)	13 of 15 CV < 20%	Age CV 179%	Imprecision, not heterogeneity
Inter-quantile tests, multiplicity (Section 5.4)	13 of 14 (indep.)	10 (ZIP3 CI); 8 (RW/Holm)	SE ratio 2.3
Machine learning			
Random vs ZIP5 / ZIP3 / state blocks (Section 5.6.1)	gap −18.7 pp	−16.8 / −13.2 / −4.8	Advantage shrinks with block scale
Leave-one-state-out (Section 5.6.3)	—	gap −6.3 pp	XGBoost better in 44 of 51 states
Ten-state holdout, nested vs 1,000 rounds (Section 5.7.2)	0.515	0.425	Capacity effect runs through region dummies
Size-matched random split (Section 6.4)	0.833	0.829	Training size explains 0.4 pp
Feasible level correction, $k = 20$ (Section 5.7.1)	LOSO 0.635 (seed-42 fit)	0.699	Level error is cheaply correctable
Remove region dummies, ten-state, 1,000 r. (Section 5.7.2)	0.425	0.521	Curve flat without region dummies
Remove school variables, ten-state, nested	0.515	0.490	School block transfers partly
Add missingness indicators (Section 6.4)	0.515	0.523	Within seed noise out of state
Exact duplicates across the split (Section 6.3)	0.833	0.833	Immaterial
SHAP stability (Section 5.8)			
XGBoost vs LightGBM / random forest	—	$\rho = 0.987$ / 0.905	Same leading features; rank 3 is a state identifier

Notes: "pp" = percentage points of R^2. Unless stated, estimates use the trimmed sample ($N = 783,872$) and the corrected design; booster values are means over five seeds. Every entry is drawn from the tables cited in the first column.

a national hedonic equation summarises average capitalisation gradients in asking prices and is not the price function of any market.

7.2. Quantile Regression Adds Distributional Insight, with Wider Uncertainty than Independence Implies

Quantile regression on the corrected design reveals heterogeneity that the mean masks, and its shape matches the housing literature. The declining living-area gradient (0.363 at $\tau = 0.10$ against 0.240 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, 11.5 times larger at the bottom decile than at 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). The inference is wider than independence-based statistics would suggest: the paired ZIP3-cluster bootstrap standard errors of the tail differences are 2.3 times the independence values on median, and 8 of 14 differences survive Romano–Wolf or Holm control against 13 under independence. The structural gradients carry the distributional story; the tail differences of pool, waterfront and the West dummy do not survive.

7.3. The Machine-Learning Advantage Depends on the Validation Design

XGBoost improves on OLS by 18.7 points of R^2 under random validation, 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 seven designs show that this number answers the interpolation question only. The advantage declines with the spatial scale of the validation blocks—16.8 points under ZIP5 blocks, 13.2 under ZIP3 blocks, 4.8 under state blocks, 6.3 under leave-one-state-out—and vanishes only under the single ten-state partition, where the interval of the OLS–XGBoost difference includes zero and LightGBM remains ahead. The earlier reading of that partition, that the semi-log regression is as accurate out of state as the best tree ensemble, is therefore a statement about one unfavourable draw, not about extrapolation in general: the ensembles are better out of state, by less than in sample, and by an amount that depends on how far the held-out market is from the training markets.

The mechanics are in the level. Out-of-state error is dominated by the price level of the withheld market, the ensembles mispredict levels more than OLS when the training set loses the largest markets at once, and the level error is cheaply correctable: twenty local listings recover 83% of the gain an oracle level correction would deliver under leave-one-state-out, after which the ordering of

models is the ordering of their within-market accuracy, which favours the ensembles. Three features of the data and the method generate the level error. The region dummies partition the training states' levels, and additional boosting rounds refine that partition in a way that is wrong by construction for a held-out state; without the region dummies the rounds curve is flat, so the capacity effect and the region-dummy effect are one interaction, not two channels. State-level encodings in the features—the tax rate, state dummies and coordinates most of all—inflate random-split accuracy in proportion to how precisely they identify the state and transfer out of state only when they are continuous proxies for the level; the imputed school variables, which the earlier version of this analysis blamed for the extrapolation failure, in fact transfer partly, and removing them lowers out-of-state accuracy. And the nested selection of rounds inside the training groups, which is the only selection rule a deployment can use, does not reach the test-selected peak of the rounds curve. Deppner and Cajias (2024) report the same direction of the tuning effect at city scale; the present results locate it in the region-dummy interaction and quantify it across seven designs with intervals.

The scope of the conclusion follows the debate opened by Wadoux et al. (2021) and Milà et al. (2022). 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 ensemble advantage is real, as the Moran statistics of the tree residuals (0.236 against 0.471 for OLS) show. For extrapolation to unrepresented markets, blocked validation at the relevant scale is the benchmark (Roberts et al., 2017; Meyer and Pebesma, 2021), the gap is smaller, the ranking of feature sets changes, and the optimal capacity changes. The practical rules for AVM deployment that the evidence supports are: match the validation protocol to the deployment question; report out-of-market error in its level and within-market parts; calibrate the level from local listings where any are available, which recovers most of the within-market accuracy; and select capacity on spatially blocked, not random, training folds. The evidence does not support a blanket rule of dropping geographic features for new-market deployment: without the region dummies the ensemble gains out of state when capacity is left unchecked at 1,000 rounds, but under the nested and 100-round rules it loses under state blocks and leave-one-state-out; and without the school variables it is worse under every rule except the 100-round ten-state case, where the difference is within seed noise (Section 5.7.2). Bogin and Shui (2020) reach a related conclusion for rural areas where AVMs lack comparable data.

7.4. SHAP Interprets Prediction, Not Economics

SHAP values identify living area, bathrooms, school rating, lot size, region, the tax rate and the luxury score as the dominant predictive contributors, and this ranking is stable across three tree ensembles (ρ = 0.905–0.987), which addresses the concern that importance rankings are artefacts of one fitted model (Ribeiro et al., 2016). But stability is not structure: the 3rd most important feature is, for 93.0% of listings, the state, and 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 documented here supports SHAP as a description of this prediction technology, not as a measurement of market valuation. The housing applications that interpret ensembles with SHAP (Krämer et al., 2023; Iban, 2022; Potrawa and Tetereva, 2022) would benefit from the same audit of what their leading features encode.

7.5. Data Documentation Is Part of the Method

Three of the paper's most consequential findings—the school variables as state identifiers, the zero-coded lot size and the corrupt parking and story counts behind an out-of-state R^2 of $-1{,}956{,}000$—are data-documentation findings, not modelling findings. None was visible in a coefficient table; 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, and the data practices that Nolte et al. (2024) set out for nationwide property data—documenting missingness, mismeasurement and property-type identification before estimation—extend to listing extracts. A table like Table 4, reporting for every regressor the share of imputed values and the rule actually applied in the code, together with a sample-construction script that reproduces the analysis data from the raw extract, is a necessary part of a 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 wild-cluster 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 cluster bootstrap with multiplicity control for most structural attributes but not for the rarer amenities; boosting delivers interpolation gains that are large and extrapolation gains that are smaller, design-dependent and concentrated in within-market structure, with the level error traceable to regional partitions and state-level encodings and cheaply corrected from local data; 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 by an amount that depends on the spatial scale of deployment—is the paper's most transferable lesson for the hedonic and AVM literatures.

8. Limitations

The limitations fall into five groups: the dependent variable and the sample, the regressors, the inference, the validation designs, and what was not done. For each, the direction in which it could move the results is stated where it can be.

Dependent variable and sample. 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 Kolbe et al. (2021) show to be attribute-dependent; listing-price gradients are therefore not transaction-price implicit prices, and the cross-framework comparisons are comparisons of models of asking prices, which could rank differently on sale prices if the wedge is more non-linear in attributes than the semi-log form allows. The extract contains no sale prices for the listed homes, so a re-estimation on transactions was not possible; the Zestimate, available for 55.0% of listings, is re-anchored to the list price and does not bound the wedge. The sample is a single snapshot of active listings completed on 8 January 2026, with the South over-represented (56.6% of listings) and no time dimension or repeat-sales control; no reweighting to Census housing counts was performed, so the listing-weighted summaries describe the population of listings, not of homes; a housing-unit-weighted leave-one-state-out summary would weight the states differently from both the median and the listing-weighted mean reported here, and its value was not computed. Climate-risk variables, which the capitalisation 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.

Regressors. The regressors inherit the incompleteness of scraped data. The three school variables are state medians for 93.0% of listings, the tax rate varies at the county level, and year built, stories, parking, garage and Transit Score are imputed for 18.9–70.2% of listings by rules that make them partly indicators of reporting completeness; the free-text categorical fields are mapped by a frozen lookup whose keyword logic cannot be recovered from first principles. These variables are treated as what they are—state-level or reporting-level encodings—rather than as the attributes they are named after, and the ablation shows that their contribution to in-sample accuracy is largely geographic. Exact duplicates (0.86% of listings) are kept; removing those that straddle the random split changes R^2 by less than one thousandth.

Inference. No estimate 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 average local price functions that differ in magnitude and sometimes in sign. The state-cluster inference rests on 51 clusters whose effective number is 14.2 at the median and 3.7 at the minimum; the wild cluster bootstrap is the reference, but with so few effective clusters its p-values for the state-level regressors (region dummies, tax rate, school rating) remain approximate, and a reader should treat those coefficients as imprecisely estimated rather than as zero. Spatial lag, error and geographically weighted models (Anselin, 1988; LeSage and Pace, 2009) are not estimated; the Conley estimator is reported only as a check. The quantile regressions use a 150,000-listing subsample for computational reasons, their bootstrap resamples that subsample's ZIP3 areas, and the standardised age coefficient is too imprecise to be interpreted at any quantile. The state-by-state regressions in the smallest jurisdictions are near-singular even after dropping constant columns (maximum condition number 1.2×10^{21}); their point estimates are reported with intervals and should not be read individually.

Validation designs. The ten-state holdout withholds 55.6% of the listings including the five largest states and is the harshest of the seven designs; the blocked and leave-one-state-out designs locate it within the range of outcomes rather than replace it. Leave-one-state-out 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 is the appropriate benchmark. Blocking finer than ZIP5 was not attempted; county identifiers are missing for 97.0% of listings, but the ZIP-based blocks already span the neighbourhood-to-state range. The feasible level correction draws calibration listings at random from the held-out state; the listings a deployment actually observes may be less representative, so the reported recoveries assume random local coverage and may overstate what non-random coverage achieves, while the oracle bounds any level correction. Hyperparameters other than the number of rounds were not tuned, by random or by blocked resampling; the capacity analysis shows that tuning rounds under random resampling would select the fully grown model, but it does not establish how much of the extrapolation gap a full spatially blocked tuning of depth, learning rate and subsampling would recover. The seed dispersion of the boosters is small under the blocked designs and larger under the ten-state holdout (standard deviation 0.015 at 1,000 rounds), and differences between variants that fall within two seed standard deviations are not interpreted.

What was not done. The random forest, Ridge, Lasso and Elastic Net benchmarks of an earlier version of this analysis cannot be reproduced because their code was lost; the versions reported here are re-specified with declared hyperparameters and labelled as such, and no claim rests on them. The quantile estimates of that earlier version on the uncorrected design reproduce with a median absolute difference of <0.0001 at every quantile; at $\tau = 0.75$ and 0.90 the parking coefficient,

which depended on a column with values up to 10^{11} and is solver-dependent, and through it the largest remaining difference (0.11 and 0.59) do not reproduce; those estimates are documented in Appendix F and superseded. The SHAP values of the random forest use the Saabas approximation, which is not exactly additive, so the cross-model correlations involving the forest are approximate. Finally, every result is conditional on one extract from one platform; replication on another listing source or on transaction data is the test that would most directly overturn the conclusions.

9. Conclusion

This paper asked how much of the machine-learning advantage over a hedonic regression survives when the model must price homes in markets it has not seen, and what produces the part that does not. On 783,872 active U.S. Zillow listings, with semi-log OLS, quantile regression and gradient-boosted ensembles estimated on one corrected design matrix and compared under seven validation designs with cluster-bootstrap intervals, the answer is that the advantage is real and design-dependent. It is 18.7 points of out-of-sample R^2 under random validation, 13.2 under ZIP3 blocks, 4.8 under state blocks and 6.3 under leave-one-state-out, where XGBoost is ahead in 44 of 51 states; it is indistinguishable from zero only under a single ten-state partition that the other designs identify as an unfavourable draw. The part that does not survive is the prediction of market price levels: the ensembles mispredict the level of withheld states more than OLS when the largest markets are withheld together, the error is traceable to the region dummies, to state-level encodings in the features and to boosting capacity acting through the regional partition, and it is cheaply correctable—twenty local listings recover 83% of the oracle level correction under leave-one-state-out, after which the ensembles' within-market advantage reasserts itself.

The parametric and distributional results discipline the comparison. The semi-log regression yields average listing-price gradients that align with the meta-analytic record, with a lot elasticity of 0.10 once a coding error is corrected; cluster-robust and wild-bootstrap inference leave 23 of 63 slopes significant, the casualties being the spatially aggregated proxies and the rare, concentrated amenities; and the pooled coefficients average 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. Quantile regression shows that the garage gradient is an order of magnitude larger at the 10th than at the 90th conditional percentile, that the pool and lot-size gradients rise across the distribution and that the living-area gradient declines, with 8 of 14 tail differences surviving a paired cluster bootstrap with multiplicity control. SHAP rankings are stable across three ensembles, and the third-ranked feature is, for most listings, a state identifier.

The contribution to the hedonic and AVM literatures is methodological. Model evaluation on spatially dependent housing data hinges on validation design, and the random-versus- blocked

distinction developed in geostatistics and ecology (Brenning, 2012; Roberts et al., 2017; Milà et al., 2022) transfers to national housing data with three additions that this paper makes concrete: the OLS–ensemble gap should be reported as a function of the spatial scale of the blocks, with intervals; out-of-market error should be decomposed into level and within-market parts, and the feasible level correction reported alongside the oracle; and capacity should be selected on spatially blocked training folds, because random early stopping selects the fully grown model whose regional partition does not transfer. Studies that compare machine learning to hedonic regression on random splits alone will overstate the machine-learning advantage for applications involving new markets, by an amount that this paper shows to depend on how far the deployment market lies from the training markets; studies that report conventional robust standard errors on national scraped data will overstate the precision of their gradients.

Several extensions follow. Replication on transaction prices would test whether the cross-framework ranking is invariant to the listing wedge that Kolbe et al. (2021) document; a full spatially blocked tuning of the ensembles would establish how much of the remaining extrapolation gap is recoverable; spatial econometric estimation would model the dependence that this paper diagnoses; climate-risk variables are a first-order omission; and an observed rather than imputed school-quality measure would allow the capitalisation question to be asked at all. Leave-region-out validation reported alongside random validation, decomposed into level and within-region accuracy and with a feasible level correction, is a reasonable default for future comparisons of machine-learning and hedonic models.

References

Ahmed, Usman and Jason Hawkins, "Residential price modeling using spatially validated machine learning methods: A comparison across geographical contexts," *Transportation Research Record*, 2026. Published online; volume and pages not yet assigned.

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), 3–26.

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.

Benjamini, Yoav and Yosef Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, 1995, *57* (1), 289–300.

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.

Brenning, Alexander, "Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest," in "2012 IEEE International Geoscience and Remote Sensing Symposium" 2012, pp. 5372–5375.

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

_ , Jonah B. Gelbach, and Douglas L. Miller, "Bootstrap-based improvements for inference with clustered errors," *The Review of Economics and Statistics*, 2008, *90* (3), 414–427.

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.

Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald, "Asymptotic behavior of a *t*-test robust to cluster heterogeneity," *The Review of Economics and Statistics*, 2017, *99* (4), 698–709.

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.

Deppner, Juergen and Marcelo Cajias, "Accounting for spatial autocorrelation in algorithm-driven hedonic models: A spatial cross-validation approach," *The Journal of Real Estate Finance and Economics*, 2024, *68* (2), 235–273.

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.

Geerts, Margot, Seppe vanden Broucke, and Jochen De Weerdt, "A survey of methods and input data types for house price prediction," *ISPRS International Journal of Geo-Information*, 2023, *12* (5), 200.

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.

Holm, Sture, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, 1979, *6* (2), 65–70. No Crossref DOI exists for this article; record verified through OpenAlex W2121044470.

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.

Iban, Muzaffer Can, "An explainable model for the mass appraisal of residences: The application of tree-based machine learning algorithms and interpretation of value determinants," *Habitat International*, 2022, *128*, 102660.

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.

Karasiak, Nicolas, Jean-François Dejoux, Claude Monteil, and David Sheeren, "Spatial dependence between training and test sets: Another pitfall of classification accuracy assessment in remote sensing," *Machine Learning*, 2022, *111* (7), 2715–2740.

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," Vol. 30 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: Causes and effects of listing price changes," *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.

Kolbe, Jens, Rainer Schulz, Martin Wersing, and Axel Werwatz, "Real estate listings and their usefulness for hedonic regressions," *Empirical Economics*, 2021, *61* (6), 3239–3269.

Krämer, Bastian, Cathrine Nagl, Moritz Stang, and Wolfgang Schäfers, "Explainable AI in a real estate context: Exploring the determinants of residential real estate values," *Journal of Housing Research*, 2023, *32* (2), 204–245.

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," Vol. 30 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.

_ **and Matthew D. Webb**, "Wild bootstrap inference for wildly different cluster sizes," *Journal of Applied Econometrics*, 2017, *32* (2), 233–254.

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. Crossref records the print date as 2002-10-18; the volume's copyright year is 2003.

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.

Milà, Carles, Jorge Mateu, Edzer Pebesma, and Hanna Meyer, "Nearest neighbour distance matching leave-one-out cross-validation for map validation," *Methods in Ecology and Evolution*, 2022, *13* (6), 1304–1316.

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.

Nolte, Christoph, Kevin J. Boyle, Anita M. Chaudhry, Christopher Clapp, Dennis Guignet, Hannah Hennighausen, Ido Kushner, Yanjun Liao, Saleh Mamun, Adam Pollack, Jesse Richardson, Shelby Sundquist, Kristen Swedberg, and Johannes H. Uhl, "Data practices for studying the impacts of environmental amenities and hazards with nationwide property data," *Land Economics*, 2024, *100* (1), 200–221.

Oates, Wallace E., "The effects of property taxes and local public spending on property values: An empirical study of tax capitalization and the Tiebout hypothesis," *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.

Pohjankukka, Jonne, Tapio Pahikkala, Paavo Nevalainen, and Jukka Heikkonen, "Estimating the prediction performance of spatial models via spatial k-fold cross validation," *International Journal of Geographical Information Science*, 2017, *31* (10), 2001–2019.

Potrawa, Tomasz and Anastasija Tetereva, "How much is the view from the window worth? Machine learning-driven hedonic pricing model of the real estate market," *Journal of Business Research*, 2022, *144*, 50–65.

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.

Rico-Juan, Juan Ramón and Paloma Taltavull de La Paz, "Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain," *Expert Systems with Applications*, 2021, *171*, 114590.

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, spatial, hierarchical, or phylogenetic structure," *Ecography*, 2017, *40* (8), 913–929.

Romano, Joseph P. and Michael Wolf, "Stepwise multiple testing as formalized data snooping," *Econometrica*, 2005, *73* (4), 1237–1282.

Roodman, David, Morten Ørregaard Nielsen, James G. MacKinnon, and Matthew D. Webb, "Fast and wild: Bootstrap inference in Stata using boottest," *The Stata Journal*, 2019, *19* (1), 4–60.

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.

Webb, Matthew D., "Reworking wild bootstrap-based inference for clustered errors," *Canadian Journal of Economics*, 2023, *56* (3), 839–858.

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. Full OLS Coefficient Table

Table 34 reports the complete set of OLS coefficients on the trimmed sample in standardised and natural units with HC3 standard errors, including all categorical controls. Cluster-robust and wild-bootstrap inference for the main coefficients is in Table 9; across all 63 slopes, ZIP3-cluster standard errors are on median 7.8 times the HC3 values and state-cluster standard errors 11.7 times.

Table 34. Full OLS Regression on the v2 Design: All Standardised Coefficients with HC3 and State-Clustered Standard Errors

Variable	Coef.	(HC3)	(State)	Variable	Coef.	(HC3)	(State)
ln(living area)	0.291***	(0.003)	(0.033)	ln(living area) × age	−0.158	(0.014)	(0.108)
Bedrooms	−0.070**	(0.002)	(0.028)	Pool × South	0.005	(0.001)	(0.030)
Bathrooms	0.256***	(0.003)	(0.023)	Waterfront × ln(living area)	0.171	(0.009)	(0.124)
Age (years)	0.012	(0.014)	(0.115)	Basement × North	−0.023	(0.001)	(0.019)
Age^2	0.044*	(0.002)	(0.021)	Condominium × Walk Score	0.037	(0.001)	(0.017)
Stories	0.000	(0.001)	(0.008)	Age × luxury score	0.037**	(0.001)	(0.015)
Bathrooms per bedroom	−0.027	(0.002)	(0.019)	Roof: Metal	0.018	(0.008)	(0.032)
Living area per bedroom	−0.009	(0.002)	(0.008)	Roof: Other	0.026	(0.008)	(0.041)
ln(lot size)	0.128***	(0.001)	(0.017)	Roof: Shingle/Comp	−0.087	(0.008)	(0.058)
Pool	0.068	(0.003)	(0.052)	Roof: Slate	−0.001	(0.017)	(0.070)
Spa	0.062*	(0.003)	(0.022)	Roof: Tile	−0.107	(0.008)	(0.066)
Basement	−0.037	(0.002)	(0.042)	Roof: Unknown	−0.025	(0.008)	(0.071)
Fireplace	0.024	(0.002)	(0.017)	Construction: Concrete/Block	0.239***	(0.003)	(0.032)
Garage	0.115***	(0.001)	(0.016)	Construction: Other	0.066***	(0.002)	(0.021)
Waterfront	−0.240	(0.033)	(0.450)	Construction: Steel	0.005	(0.013)	(0.069)
Central air	0.083***	(0.001)	(0.022)	Construction: Stone	0.133***	(0.004)	(0.018)
Forced-air heating	−0.013	(0.002)	(0.030)	Construction: Stucco	0.152**	(0.002)	(0.033)
Hardwood floors	0.123***	(0.001)	(0.018)	Construction: Unknown	0.171***	(0.002)	(0.044)
Parking spaces	0.021***	(0.001)	(0.006)	Construction: Vinyl	0.005	(0.002)	(0.021)
Luxury score (0–7)	0.004	(0.002)	(0.009)	Construction: Wood/Frame	0.090**	(0.002)	(0.032)
Walk Score	0.037*	(0.001)	(0.019)	Foundation: Concrete	0.210**	(0.005)	(0.045)
Bike Score	0.079**	(0.001)	(0.027)	Foundation: Crawlspace	0.173*	(0.006)	(0.062)
Transit Score	0.054***	(0.001)	(0.012)	Foundation: Other	0.142*	(0.006)	(0.054)
Avg. school rating¶	−0.048	(0.001)	(0.024)	Foundation: Pier/Raised	0.157	(0.006)	(0.075)
School count¶	−0.001	(0.001)	(0.003)	Foundation: Slab	0.128	(0.005)	(0.057)
Nearest-school distance¶	−0.014	(0.001)	(0.013)	Foundation: Stone	0.021	(0.010)	(0.130)
Property tax rate (%)	−0.019	(0.001)	(0.027)	Foundation: Unknown	0.165*	(0.005)	(0.047)
Condominium	−0.093	(0.004)	(0.054)	Region: Northeast	0.477**	(0.003)	(0.124)
HOA member	−0.066	(0.002)	(0.029)	Region: South	0.035	(0.003)	(0.062)
ln(1 + HOA fee)	0.058***	(0.001)	(0.019)	Region: West	0.419***	(0.003)	(0.104)
New construction	0.071**	(0.003)	(0.030)	Lot missing (indicator)	−0.023	(0.002)	(0.020)
Foreclosure	−0.390***	(0.010)	(0.036)	Constant	12.491	(0.010)	(0.138)

$N = 783,872$; $R^2 = 0.649$; adjusted $R^2 = 0.649$; RMSE = 0.478; $F = 18,943$; residual skewness 0.12, kurtosis 5.38

Notes: Same regression as Table 9; every regressor shown. Standardised continuous regressors; indicators unstandardised. "(HC3)": heteroskedasticity-robust standard error; "(State)": cluster-robust by state (51 clusters). Stars from the wild cluster bootstrap by state (Webb weights): *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Reference categories: roof = flat; construction = brick; foundation = basement; region = Midwest. ¶State median for most listings (Table 4).

B. Additional Inference Checks

Table 35. Spatial HAC (Conley) Standard Errors Compared with HC0 and State-Clustered Standard Errors

			Conley, Bartlett kernel			
Variable	Coef.	HC0	25 km	50 km	100 km	State
ln(living area)	0.300	0.0082	0.0137	0.0159	0.0181	0.0344
Bedrooms	−0.072	0.0059	0.0125	0.0162	0.0201	0.0306
Bathrooms	0.253	0.0071	0.0155	0.0200	0.0248	0.0258
Age (years)	0.048	0.0386	0.0867	0.0997	0.1090	0.1229
Age^2	0.046	0.0055	0.0097	0.0106	0.0118	0.0221
ln(lot size)	0.130	0.0025	0.0060	0.0073	0.0085	0.0172
Lot missing (indicator)	−0.029	0.0066	0.0255	0.0310	0.0350	0.0212
Pool	0.060	0.0072	0.0241	0.0341	0.0461	0.0556
Garage	0.113	0.0041	0.0097	0.0110	0.0121	0.0159
Waterfront	−0.203	0.0921	0.2181	0.2764	0.3258	0.4898
Central air	0.080	0.0038	0.0109	0.0138	0.0165	0.0225
Hardwood floors	0.124	0.0039	0.0087	0.0106	0.0124	0.0179
Luxury score (0–7)	−0.001	0.0052	0.0090	0.0107	0.0130	0.0115
Walk Score	0.036	0.0033	0.0103	0.0119	0.0139	0.0198
Bike Score	0.081	0.0028	0.0110	0.0127	0.0145	0.0279
Transit Score	0.054	0.0023	0.0123	0.0119	0.0115	0.0127
Avg. school rating[¶]	−0.046	0.0020	0.0077	0.0099	0.0130	0.0244
Property tax rate (%)	−0.016	0.0022	0.0115	0.0143	0.0162	0.0261
Condominium	−0.095	0.0105	0.0406	0.0426	0.0425	0.0481
HOA member	−0.057	0.0067	0.0225	0.0249	0.0256	0.0293
New construction	0.071	0.0082	0.0254	0.0288	0.0311	0.0317
Foreclosure	−0.387	0.0244	0.0288	0.0310	0.0338	0.0407
Region: Northeast	0.485	0.0084	0.0610	0.0835	0.1073	0.1321
Region: South	0.040	0.0075	0.0331	0.0385	0.0428	0.0622
Region: West	0.431	0.0085	0.0417	0.0489	0.0561	0.1053
Median ratio to HC0 (all slopes)		1.0	2.4	3.1	3.5	4.2

Notes: Standardised v2 regression re-estimated on a random subsample of 100,000 listings (the HAC kernel is quadratic in N); Conley standard errors use a Bartlett kernel in great-circle distance with the stated cut-off; HC0 and state-clustered (CR1) errors on the same subsample. Conley errors lie between HC0 and the state-clustered values for every slope, so clustering by state remains the conservative choice among the candidate procedures. Generated for item WP03-ECON-02.

C. Geographic Designs: Intervals, Decomposition, Fold Membership and State Detail

D. The Zestimate: Availability and Selection

Zillow's own valuation, the Zestimate, is present for 55.0% of listings. Because Zillow re-anchors its valuation to the list price once a home is listed, the ratio of asking price to Zestimate (median close to one in every region) is informative about neither the size nor the direction of the asking-to-transaction wedge, and the paper does not use it as a bound; the ratio table of the earlier version is therefore not reproduced. Table 42 compares the observables of listings with and without a Zestimate.

Table 36. Percentage Effects of Indicator Variables with the Kennedy Correction Under Three Variance Estimators

		$100(e^{\beta}-1)$	Kennedy, by variance estimator			t
Indicator	β	(naive)	HC3	ZIP3	State	(state)
Waterfront	−0.2405	−21.4	−21.4	−23.7	−28.9	−0.53
Pool	0.0685	7.1	7.1	7.0	6.9	1.31
Garage	0.1148	12.2	12.2	12.2	12.1	7.15
Central air	0.0835	8.7	8.7	8.7	8.7	3.76
Hardwood floors	0.1233	13.1	13.1	13.1	13.1	6.99
Condominium	−0.0933	−8.9	−8.9	−9.0	−9.0	−1.72
Foreclosure	−0.3895	−32.3	−32.3	−32.3	−32.3	−10.73
New construction	0.0711	7.4	7.4	7.3	7.3	2.35
HOA member	−0.0658	−6.4	−6.4	−6.4	−6.4	−2.29
Basement	−0.0366	−3.6	−3.6	−3.6	−3.7	−0.88
Spa	0.0619	6.4	6.4	6.4	6.4	2.88
Fireplace	0.0244	2.5	2.5	2.5	2.4	1.42
Lot missing (indicator)	−0.0230	−2.3	−2.3	−2.3	−2.3	−1.14

Notes: Natural-unit v2 regression of Table 9. The Kennedy (1981) correction $100[\exp(\beta - \frac{1}{2}\hat{V}(\beta)) - 1]$ is evaluated with the HC3, ZIP3-clustered and state-clustered variance of β (Eq. (2)); with the state-clustered variance the correction is visible for the imprecisely estimated waterfront and foreclosure indicators only. Generated for items WP03-THEO-08 and ECON-10.

Table 37. Quantile-Regression Standard Errors at $\tau = 0.10$ and 0.90: Kernel (Independence) versus ZIP3-Cluster Bootstrap

	$\tau = 0.10$				$\tau = 0.90$			
Variable	Coef.	(kernel)	(cluster)	Ratio	Coef.	(kernel)	(cluster)	Ratio
ln(living area)	0.363	(0.0064)	(0.0167)	2.6	0.240	(0.0088)	(0.0266)	3.0
Bedrooms	−0.100	(0.0054)	(0.0172)	3.2	−0.050	(0.0065)	(0.0165)	2.5
Bathrooms	0.188	(0.0066)	(0.0208)	3.1	0.290	(0.0072)	(0.0186)	2.6
Age (years)	−0.252	(0.0302)	(0.0973)	3.2	0.089	(0.0413)	(0.0774)	1.9
ln(lot size)	0.101	(0.0019)	(0.0081)	4.2	0.147	(0.0033)	(0.0091)	2.8
Lot missing (indicator)	−0.024	(0.0067)	(0.0206)	3.1	−0.040	(0.0083)	(0.0292)	3.5
Pool	0.034	(0.0073)	(0.0220)	3.0	0.107	(0.0086)	(0.0428)	5.0
Garage	0.210	(0.0038)	(0.0128)	3.4	0.018	(0.0050)	(0.0143)	2.8
Luxury score (0–7)	−0.030	(0.0054)	(0.0094)	1.7	0.035	(0.0066)	(0.0142)	2.2
Foreclosure	−0.439	(0.0234)	(0.0605)	2.6	−0.299	(0.0293)	(0.0327)	1.1
Waterfront	−0.068	(0.0787)	(0.2558)	3.2	−0.559	(0.0902)	(0.2943)	3.3
Region: Northeast	0.342	(0.0079)	(0.0466)	5.9	0.611	(0.0097)	(0.0654)	6.7
Region: West	0.412	(0.0084)	(0.0384)	4.6	0.481	(0.0108)	(0.0567)	5.3

Notes: Same subsample and specification as Table 14. "(cluster)" is the standard deviation of the coefficient across the 999 paired ZIP3-cluster bootstrap draws of Table 15; "Ratio" divides it by the kernel standard error. Median ratio across the 13 variables: 3.2 at $\tau = 0.10$ and 2.8 at $\tau = 0.90$.

Table 38. ZIP3-Cluster Bootstrap Inference on R^2(OLS) − R^2(Comparator) Under Every Validation Design and Comparator

Comparator	OLS R^2	Comp. R^2	(SD, seeds)	Gap	Boot. SE	95% CI
Random 80/20 (N_{test} = 156, 770, 885 ZIP3 clusters)						
XGBoost, rounds selected on training groups	0.646	0.833	(0.0002)	−0.187	0.006	[−0.199, −0.176]
XGBoost, 1,000 rounds (fixed)	0.646	0.833	(0.0002)	−0.187	0.006	[−0.199, −0.176]
XGBoost, 100 rounds (fixed)	0.646	0.770	(0.0004)	−0.124	0.005	[−0.135, −0.114]
XGBoost + 6 missingness indicators, nested rounds	0.646	0.835	(0.0001)	−0.190	0.006	[−0.201, −0.178]
LightGBM, nested rounds	0.646	0.809	—	−0.164	0.006	[−0.175, −0.153]
Random, geographic training size (N_{test} = 435, 574, 885 ZIP3 clusters)						
XGBoost, rounds selected on training groups	0.649	0.829	(0.0003)	−0.181	0.006	[−0.191, −0.169]
XGBoost, 1,000 rounds (fixed)	0.649	0.829	(0.0003)	−0.181	0.006	[−0.191, −0.169]
XGBoost, 100 rounds (fixed)	0.649	0.771	(0.0005)	−0.122	0.005	[−0.132, −0.112]
XGBoost + 6 missingness indicators, nested rounds	0.649	0.831	(0.0002)	−0.183	0.006	[−0.193, −0.171]
LightGBM, nested rounds	0.649	0.809	—	−0.160	0.005	[−0.171, −0.149]
GroupKFold(5) by ZIP5 (N_{test} = 783, 872, 886 ZIP3 clusters)						
XGBoost, rounds selected on training groups	0.647	0.815	(0.0001)	−0.168	0.005	[−0.178, −0.158]
XGBoost, 1,000 rounds (fixed)	0.647	0.815	(0.0001)	−0.168	0.005	[−0.178, −0.158]
XGBoost, 100 rounds (fixed)	0.647	0.764	(0.0004)	−0.117	0.005	[−0.126, −0.107]
XGBoost + 6 missingness indicators, nested rounds	0.647	0.817	(0.0002)	−0.170	0.005	[−0.181, −0.160]
LightGBM, nested rounds	0.647	0.799	—	−0.152	0.005	[−0.163, −0.142]
GroupKFold(5) by ZIP3 (N_{test} = 783, 872, 886 ZIP3 clusters)						
XGBoost, rounds selected on training groups	0.635	0.767	(0.0002)	−0.132	0.007	[−0.145, −0.119]
XGBoost, 1,000 rounds (fixed)	0.635	0.768	(0.0003)	−0.133	0.007	[−0.146, −0.120]
XGBoost, 100 rounds (fixed)	0.635	0.736	(0.0005)	−0.102	0.006	[−0.114, −0.089]
XGBoost + 6 missingness indicators, nested rounds	0.635	0.768	(0.0001)	−0.133	0.007	[−0.146, −0.120]
LightGBM, nested rounds	0.635	0.759	—	−0.124	0.007	[−0.138, −0.111]
GroupKFold(5) by state (N_{test} = 783, 872, 886 ZIP3 clusters)						
XGBoost, rounds selected on training groups	0.580	0.628	(0.0013)	−0.048	0.007	[−0.062, −0.035]
XGBoost, 1,000 rounds (fixed)	0.580	0.615	(0.0027)	−0.035	0.011	[−0.059, −0.015]
XGBoost, 100 rounds (fixed)	0.580	0.635	(0.0023)	−0.055	0.006	[−0.066, −0.041]
XGBoost + 6 missingness indicators, nested rounds	0.580	0.634	(0.0022)	−0.055	0.006	[−0.069, −0.046]
LightGBM, nested rounds	0.580	0.631	—	−0.052	0.006	[−0.063, −0.041]
Leave-one-state-out (51) (N_{test} = 783, 872, 886 ZIP3 clusters)						
XGBoost, rounds selected on training groups	0.573	0.637	(0.0012)	−0.063	0.007	[−0.076, −0.047]
XGBoost, 1,000 rounds (fixed)	0.573	0.623	(0.0023)	−0.050	0.012	[−0.072, −0.025]
XGBoost, 100 rounds (fixed)	0.573	0.639	(0.0008)	−0.066	0.006	[−0.077, −0.052]
XGBoost + 6 missingness indicators, nested rounds	0.573	0.632	(0.0020)	−0.058	0.008	[−0.071, −0.040]
LightGBM, nested rounds	0.573	0.643	—	−0.070	0.007	[−0.085, −0.057]
Ten-state holdout (N_{test} = 435, 553, 316 ZIP3 clusters)						
XGBoost, rounds selected on training groups	0.541	0.515	(0.0052)	0.026	0.019	[−0.001, +0.074]
XGBoost, 1,000 rounds (fixed)	0.541	0.425	(0.0146)	0.116	0.033	[+0.052, +0.184]
XGBoost, 100 rounds (fixed)	0.541	0.538	(0.0070)	0.003	0.013	[−0.013, +0.037]
XGBoost + 6 missingness indicators, nested rounds	0.541	0.523	(0.0104)	0.018	0.014	[−0.022, +0.032]
LightGBM, nested rounds	0.541	0.565	—	−0.024	0.011	[−0.047, −0.004]

Notes: 1,000 bootstrap draws resample the ZIP3 areas of the test listings with replacement and recompute both R^2 values on the stored out-of-fold predictions; "Gap" is the point estimate (mean over seeds for the comparator) and the interval is the percentile interval of the draws. Across the seven designs the OLS − XGBoost (nested) gap ranges from −0.187 (random) to +0.026 (ten-state holdout); the interval excludes zero in 6 of 7 designs for XGBoost (nested) and in 7 of 7 for LightGBM.

Table 39. Where the Out-of-Sample Error Comes From: Within-State versus Between-State (Level) Shares of the Prediction Error by Design

		Oracle	Share of SST (%)		Oracle	Between
Design and model	R^2	$R^2_{\text{within state}}$	within	between	$R^2_{\text{within block}}$	block (%)
Random 80/20						
OLS	0.646	0.615	32.4	3.0	—	—
OLS, no region dummies	0.619	0.600	33.7	4.5	—	—
OLS + state fixed effects	0.690	0.632	31.0	0.0	—	—
XGBoost, nested rounds	0.833	0.802	16.7	0.0	—	—
XGBoost, 1,000 rounds	0.833	0.802	16.7	0.0	—	—
XGBoost, 100 rounds	0.770	0.733	22.5	0.5	—	—
XGBoost + missingness, nested	0.835	0.805	16.5	0.0	—	—
LightGBM, nested rounds	0.809	0.775	19.0	0.1	—	—
Random, geographic training size						
OLS	0.649	0.618	32.1	3.0	—	—
OLS, no region dummies	0.622	0.603	33.4	4.4	—	—
OLS + state fixed effects	0.693	0.636	30.6	0.0	—	—
XGBoost, nested rounds	0.829	0.798	17.0	0.0	—	—
XGBoost, 1,000 rounds	0.829	0.798	17.0	0.0	—	—
XGBoost, 100 rounds	0.771	0.734	22.4	0.5	—	—
XGBoost + missingness, nested	0.832	0.800	16.8	0.0	—	—
LightGBM, nested rounds	0.809	0.774	19.0	0.1	—	—
GroupKFold(5) by ZIP5						
OLS	0.647	0.616	32.3	3.0	0.593	16.3
OLS, no region dummies	0.620	0.601	33.6	4.4	0.559	17.4
OLS + state fixed effects	0.691	0.633	30.9	0.0	0.619	13.2
XGBoost, nested rounds	0.815	0.781	18.5	0.0	0.726	5.7
XGBoost, 1,000 rounds	0.815	0.781	18.5	0.0	0.726	5.7
XGBoost, 100 rounds	0.764	0.725	23.2	0.4	0.693	9.3
XGBoost + missingness, nested	0.817	0.783	18.3	0.0	0.729	5.7
LightGBM, nested rounds	0.799	0.762	20.0	0.1	0.714	6.8
GroupKFold(5) by ZIP3						
OLS	0.635	0.606	33.2	3.4	0.617	10.0
OLS, no region dummies	0.609	0.593	34.3	4.8	0.592	10.9
OLS + state fixed effects	0.673	0.616	32.3	0.4	0.638	7.7
XGBoost, nested rounds	0.767	0.727	23.0	0.4	0.730	4.7
XGBoost, 1,000 rounds	0.768	0.729	22.8	0.4	0.731	4.6
XGBoost, 100 rounds	0.737	0.696	25.6	0.7	0.708	6.1
XGBoost + missingness, nested	0.768	0.729	22.8	0.4	0.733	4.7
LightGBM, nested rounds	0.759	0.719	23.7	0.4	0.724	5.0

Notes: Continued on the next page; notes follow the last part.

Table 39. Where the Out-of-Sample Error Comes From: Within-State versus Between-State (Level) Shares of the Prediction Error by Design (continued)

		Oracle	Share of SST (%)		Oracle	Between
Design and model	R^2	$R^2_{\text{within state}}$	within	between	$R^2_{\text{within block}}$	block (%)
GroupKFold(5) by state						
OLS	0.580	0.587	34.8	7.3	—	—
OLS, no region dummies	0.568	0.576	35.7	7.5	—	—
XGBoost, nested rounds	0.628	0.649	29.6	7.6	—	—
XGBoost, 1,000 rounds	0.617	0.657	28.8	9.4	—	—
XGBoost, 100 rounds	0.633	0.646	29.8	6.9	—	—
XGBoost + missingness, nested	0.637	0.655	29.1	7.3	—	—
LightGBM, nested rounds	0.631	0.651	29.4	7.5	—	—
Leave-one-state-out (51)						
OLS	0.573	0.592	34.4	8.3	—	—
OLS, no region dummies	0.564	0.577	35.6	8.0	—	—
XGBoost, nested rounds	0.635	0.658	28.8	7.7	—	—
XGBoost, 1,000 rounds	0.623	0.668	28.0	9.7	—	—
XGBoost, 100 rounds	0.638	0.653	29.2	7.0	—	—
XGBoost + missingness, nested	0.630	0.660	28.7	8.3	—	—
LightGBM, nested rounds	0.643	0.659	28.7	7.0	—	—
Ten-state holdout						
OLS	0.541	0.573	37.4	8.5	—	—
OLS, no region dummies	0.533	0.559	38.7	8.1	—	—
XGBoost, nested rounds	0.506	0.638	31.8	17.6	—	—
XGBoost, 1,000 rounds	0.425	0.629	32.5	25.0	—	—
XGBoost, 100 rounds	0.531	0.631	32.4	14.6	—	—
XGBoost + missingness, nested	0.538	0.638	31.7	14.5	—	—
LightGBM, nested rounds	0.565	0.643	31.3	12.2	—	—

Notes: Exact decomposition SSE = $\text{SSE}_{\text{within}}$ + $\text{SSE}_{\text{between}}$ of the squared prediction error on the test listings into deviations from state means and state-mean (level) errors, each expressed as a share of the total sum of squares of log prices, so that $1 - R^2$ equals the sum of the two shares. "Oracle $R^2_{\text{within state}}$" is Eq. (8) (state means of the held-out state taken from the test data). The between-state component is 15.8% of the variance of log prices in the random test set. The last two columns repeat the decomposition at the block level (ZIP5 or ZIP3) for the blocked designs. XGBoost rows are seed-42 fits. Generated for items WP03-THEO-01 and THEO-02.

Table 40. Composition of the Geographic Validation Folds

Design	Fold	States	N_{train}	N_{test}	Held-out states (or groups)
State blocks	Fold 1	7	626,868	157,004	AK, FL, MT, NH, OR, UT, WY
State blocks	Fold 2	9	626,685	157,187	HI, LA, MD, NE, NJ, RI, TX, VT, WA
State blocks	Fold 3	11	627,528	156,344	AL, AR, AZ, CA, DE, ID, KS, MI, MN, MO, SD
State blocks	Fold 4	12	627,142	156,730	CO, CT, GA, MS, ND, NV, OH, OK, TN, VA, WI, WV
State blocks	Fold 5	12	627,265	156,607	DC, IA, IL, IN, KY, MA, ME, NC, NM, NY, PA, SC
Ten-state holdout	Held out	10	348,319	435,553	CA, CO, FL, GA, IL, NC, NY, OH, TX, WA
ZIP3 blocks	5 folds	44–49	627,098	156,774	177–178 ZIP3 prefixes per fold
ZIP5 blocks	5 folds	51–51	627,098	156,774	5,279–5,281 ZIP codes per fold

Notes: Folds formed by `GroupKFold` on the state (five folds), the ZIP3 prefix or the five-digit ZIP code; the ten-state holdout is the single partition of v1.1. Leave-one-state-out fits one model per jurisdiction (51 fits; test sizes in Table 41). Generated for item WP03-EDIT-13.

Table 41. Leave-One-State-Out Accuracy by Held-Out State: OLS and Tree Ensembles on the v2 Sample

		R^2					Mean error (log points)	
State	N_{test}	OLS	XGB nested	XGB 1,000	XGB 100	LGB nested	OLS	XGB nested
NV	10,760	0.484	0.797	0.763	0.772	0.794	0.287	−0.024
NE	3,925	0.667	0.755	0.758	0.746	0.742	−0.037	0.016
MD	9,666	0.459	0.741	0.702	0.741	0.744	−0.367	0.039
MO	14,784	0.648	0.735	0.741	0.726	0.726	0.031	0.050
UT	6,624	0.563	0.734	0.767	0.729	0.748	0.025	−0.029
VA	15,719	0.711	0.727	0.757	0.726	0.750	−0.134	−0.161
AZ	23,871	0.424	0.726	0.738	0.717	0.730	0.331	0.123
KS	5,974	0.616	0.725	0.754	0.720	0.711	0.106	0.079
DC	1,409	0.376	0.724	0.675	0.725	0.722	−0.458	−0.095
OH	19,408	0.654	0.713	0.708	0.706	0.709	0.028	−0.010
GA	31,988	0.623	0.705	0.728	0.693	0.691	0.103	0.047
IN	14,338	0.610	0.697	0.709	0.686	0.691	0.099	0.045
ID	4,532	0.545	0.695	0.737	0.685	0.691	0.052	−0.011
AR	10,665	0.559	0.689	0.693	0.674	0.685	0.189	0.102
RI	1,040	0.603	0.685	0.716	0.671	0.690	−0.128	−0.114
MN	8,710	0.606	0.680	0.705	0.677	0.682	−0.083	−0.076
MA	5,301	0.538	0.673	0.701	0.636	0.652	−0.204	−0.144
ND	1,339	0.590	0.672	0.678	0.666	0.687	0.042	0.070
KY	10,637	0.577	0.669	0.683	0.668	0.671	0.034	−0.031
MI	19,190	0.632	0.669	0.690	0.663	0.669	−0.058	0.017
IA	8,265	0.590	0.667	0.654	0.661	0.672	0.011	−0.004
OR	9,485	0.551	0.665	0.674	0.666	0.633	0.078	0.028
TN	24,597	0.561	0.659	0.668	0.659	0.647	−0.127	−0.127
SD	2,051	0.502	0.649	0.678	0.642	0.654	−0.189	0.032
WI	9,114	0.548	0.640	0.703	0.627	0.668	−0.180	0.105
OK	12,246	0.605	0.634	0.599	0.640	0.643	0.125	0.221
NH	2,289	0.533	0.631	0.624	0.628	0.623	−0.100	−0.062
CO	17,480	0.491	0.631	0.647	0.625	0.608	0.058	0.018
IL	19,170	0.619	0.626	0.619	0.629	0.614	0.041	0.021
SC	20,322	0.577	0.622	0.659	0.618	0.608	−0.056	−0.037
NJ	7,321	0.559	0.619	0.530	0.622	0.641	−0.026	−0.138
MT	3,959	0.594	0.617	0.638	0.603	0.622	0.104	0.047
AL	16,919	0.521	0.606	0.616	0.588	0.612	0.185	0.152
CT	4,127	0.602	0.591	0.572	0.593	0.575	−0.118	−0.258
TX	101,669	0.608	0.588	0.679	0.611	0.625	0.122	0.190
AK	808	0.459	0.586	0.583	0.586	0.598	0.221	−0.004
MS	6,773	0.379	0.584	0.613	0.547	0.601	0.324	0.182
NC	29,245	0.561	0.570	0.583	0.577	0.568	−0.146	−0.213
WY	1,653	0.486	0.539	0.584	0.534	0.505	0.148	0.180
NY	24,823	0.364	0.516	0.583	0.516	0.601	−0.381	−0.248
NM	4,704	0.365	0.504	0.530	0.502	0.527	0.374	0.295
FL	132,186	0.448	0.502	0.420	0.513	0.530	−0.276	−0.308
ME	3,266	0.420	0.494	0.525	0.483	0.504	0.173	0.105
WA	11,934	0.540	0.480	0.407	0.480	0.455	−0.090	−0.234
DE	1,998	0.393	0.470	0.528	0.448	0.466	−0.164	−0.170
PA	15,127	0.129	0.441	0.326	0.488	0.357	0.567	0.372
VT	1,685	0.622	0.348	0.465	0.356	0.371	0.014	−0.383
LA	14,934	0.389	0.346	0.284	0.367	0.267	0.317	0.367
WV	3,179	0.430	0.266	0.241	0.260	0.261	0.324	0.459
HI	5,013	0.258	0.250	0.188	0.261	0.094	−0.384	−0.447
CA	47,650	0.139	0.220	0.001	0.224	0.208	−0.427	−0.418

Notes: Each row is a separate set of models trained on all listings outside the state (trimmed v2 sample); XGBoost values are means over seeds 42–46. Sorted by the R^2 of XGBoost with nested rounds. Mean error is the average of predicted minus actual log price in the held-out state (positive: the state is over-priced by the out-of-state model).

Table 42. The Zestimate Is Available for About Half of the Listings: Asking-Price-to-Zestimate Ratios by Region and Observables by Zestimate Availability

	All	Northeast	Midwest	South	West
Panel A: ratio of asking price to Zestimate, listings with a Zestimate					
p_{10}	1.001	0.988	0.999	1.003	0.999
p_{25}	1.008	1.002	1.007	1.011	1.005
Median	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 of ratios above 1 (%)	91.4	78.2	88.1	95.1	88.0
Listings with a Zestimate	433,738	32,969	58,622	251,274	90,873
Share of listings with a Zestimate (%)	—	50.0	46.3	56.3	60.8
Panel B: mean observables of listings with and without a Zestimate					
Variable	With	Without	Std. diff.		
Listing price ($)	617,174	627,313	−0.01		
Living area (sq ft)	2,115	2,065	0.04		
Bedrooms	3.30	3.20	0.09		
Bathrooms	2.57	2.47	0.09		
Age (years)	41.06	43.59	−0.08		
Lot size (sq ft, imputed field)	232,519	277,290	0.00		
Condominium	0.14	0.15	−0.02		
Pool	0.44	0.35	0.19		
New construction	0.03	0.09	−0.25		
Foreclosure	0.00	0.01	−0.07		
Walk Score	25.30	25.87	−0.02		
Avg. school rating[¶]	6.02	6.03	−0.02		
Luxury score (0–7)	1.44	1.25	0.15		
Days on Zillow	87.12	126.38	−0.21		

Notes: Panel A: listings with a non-missing Zestimate (ratios outside $[0.2, 5]$ excluded), computed from the stored sample by `scripts/06_revision_inference.py`; the Zestimate is Zillow's automated valuation at extraction and is re-anchored to the list price once a home is listed, so the ratio describes the extract and is not a measure of an asking-price premium (Section 3). Panel B: standardised mean difference (with minus without, divided by the pooled SD) of the main observables; availability is 55.2% overall (`scripts/10_stored_checks.py`). Generated for items WP03-THEO-10, ADVE-23 and EDIT-12.

E. Data Availability and Reproducibility

Data. The raw extract (839,313 active Zillow listings, snapshot completed 8 January 2026) was collected by the author from public listing pages into a DuckDB database. Zillow's terms of use do not permit its redistribution; the database and the derived analysis files are available from the author for verification purposes, and their sizes and SHA-256 checksums are recorded in the repository (`data/DATA_AVAILABILITY.md`). The paper reports only aggregate statistics and model estimates.

Code and environment. The repository contains the sample-construction script that rebuilds the analytical sample from the raw table and reproduces the earlier sample bit for bit, the design-matrix script that holds both the original and the corrected codings, the estimation scripts for every model and validation design, the table and figure generator that writes every exhibit and the inline numbers of this paper from stored CSV outputs, a single entry point (`run_all.sh`) that chains them, and a test suite that asserts the rebuild against the stored artefacts. Package versions are pinned for the two machines used (a laptop for the sample, inference and quantile stages; a sixteen-core node for the geographic designs and primary models); seeds are fixed throughout (splits and subsets 42, boosters 42–46, bootstrap draws $1000 + i$). Table 43 summarises the status of each item.

F. Corrections Relative to Version 1.2

This version supersedes version 1.2 of this working paper (October 2026). Table 44 lists, statement by statement, what that version said, where, and the corrected value or wording of this version; every correction is produced by the scripts listed in Appendix E. The largest are the headline comparison (parity held only for the single ten-state partition against a test-selected comparator; Tables 19 and 38), the design matrix (the lot recoding had been applied to OLS only, and the corrupt parking column had entered every other model), the quantile pseudo-R^2 row (which reported $1 - R^1(\tau)$), the significance tallies (the intercept was counted and normal critical values used; the wild cluster bootstrap gives 23 of 63 slopes), the mechanism statements about the school variables and the capacity channel (withdrawn, Sections 5.7.2 and 5.7.3), and the oracle nature of the within-state R^2. Table 45 documents the reproduction of the version 1.2 quantile regressions on the original design, and Table 46 the reproduction of the version 1.1 random-validation comparison from the stored predictions, with the rows whose code was lost marked as not reproducible.

Table 43. Reproducibility Checklist

Item	Status	Where
Data		
Data source, collection window, licence	Documented	Section 3.1; `data/DATA_AVAILABILITY.md`
Sample construction from the raw extract	Scripted, tested	Table 2; `scripts/00_build_sample.py`
Missing data rates and imputation rules	Reported	Table 4
Implausible values and trimmed sample	Reported	Section 3.6
Checksums of raw and derived files	Recorded	`results/build/sha256.txt`
Methodology		
One design matrix for every model	Scripted, tested	`scripts/00b_design_matrix.py`
Model specifications and hyperparameters	Documented	Section 4
Round selection, seeds, bootstrap draws	Documented	Sections 4.4 and 4.5
Validation designs and fold membership	Documented	Section 4.5; Table 40
Results		
Every table generated from stored outputs	Yes	`scripts/08_tables_figures_v2.py`
Every number in the text from a generated macro	Yes	`results/inline_numbers.json`
Cluster, wild-bootstrap and spatial-HAC inference	Yes	Tables 9 and 35
Intervals on every model comparison	Yes	Tables 19 and 38
Code and data availability		
Code	Public repository	`wp3_uqo`, commit recorded on the title page of the compiled paper
Raw and processed data	On request	Zillow terms of use; checksums published
Earlier benchmark code (random forest, regularised models)	Lost	Re-specified and labelled (Table 16)

Notes: Large intermediate artefacts (out-of-fold predictions, SHAP and bootstrap arrays) are regenerated by the scripts and not committed.

Table 44. Corrigenda: Statements of the Previous Version That the Rebuilt Pipeline Does Not Support, with the Corrected Values

Location (v1.2)	Statement	Correction (this version)
Abstract, §5.6, §9	"OLS is as accurate out of state (0.539) as the best tree ensemble"	Holds for the single ten-state partition only (OLS 0.541, XGBoost nested 0.515); under every other design the ensembles are more accurate (Table 19)
§5.6.2, Fig. 6	Held-out-state R^2 "peaks at 75 rounds (0.550)"; 100-round comparator 0.544	Test-selected: the peak varies between 65 and 70 rounds across seeds (mean 0.545); the nested rule gives 0.515 ± 0.005 (Table 24)
Tables 7, 8, 15, 16, 24–27, 33	Lot coding $\ln(\text{lot}) = 0$ for 16.5% and corrupt parking column in every non-OLS model	One corrected design matrix for every model; all exhibits re-estimated (Section 3)
Table 15 (QR)	Parking coefficients 691 / 317 / -98 / -527 at $\tau = 0.25$–0.90	Corrupt regressor; at $\tau = 0.75, 0.90$ the stored values do not reproduce (Table 45); v2 values $\lvert\beta\rvert \leq 0.029$
Table 15 (QR)	"Pseudo-R^2" row 0.583 / 0.598 / 0.601 / 0.580 / 0.544	Reported $1 - R^1(\tau)$; the Koenker–Machado values are 0.417 / 0.402 / 0.399 / 0.420 / 0.456
Abstract, Table 16	Inter-quantile differences: "10 of 13" rejected (200 draws, no multiplicity control)	10 of 13 by percentile CI on the v1.1 estimates; on the v2 design 8 of 14 after Holm and Romano–Wolf (Table 15)
Abstract, §5.1.5	"33 of 63 coefficients remain significant under state clustering"	The 63 included the intercept: 32 of 62 slopes (v1.2), 19 after Holm; v2 with the wild cluster bootstrap: 23 of 63 (Table 9)
§5.6.3	"forty-four of the 51 jurisdictions exceed 0.5"	43 states (stored v1.2 LOSO results)
§5.6.1	"204 held-out listings with corrupt parking or story counts"	1,856 held-out listings have an absolute error above 2 log points and carry 100.0% of the SSE; only 36 are among the implausible-flagged rows
Table 4 vs. Table 28	6,329 zero lots (Table 28) vs. 6,115 (Table 4)	6,115 zeros + 211 one-square-foot lots + 3 negative lots (Table 5)
Table 3	Stories "set to 1", parking "set to 0" when missing	Both are state-median imputed in the stored sample (202,087 and 191,492 listings; Table 4)
Table 7 notes	Cluster-robust standard errors "2–36 times" the HC3 values	Range of the state/HC3 ratio over the slopes was 3.7–41.2 on the v2 design (Table 9)
Tables 10, 18	885 ZIP3 clusters	886 clusters: 885 prefixes plus the code for a missing ZIP (16 listings)
Table 11	Moran z-statistics 41.8 / 43.5 / 40.5; SD 0.0076	The z-statistics are not stored in any artefact and are withdrawn; 0.0076 is the population SD (sample SD 0.0094); the full-test-set statistic replaces the subsample one (Table 11)
Table 17	Ridge 0.6301, Lasso 0.6280, elastic net 0.6145, random forest 0.7842	v1.1 code and hyperparameters lost (random-forest value traced to `extended_results.pkl`, 0.7842); rows re-specified in v2 (Table 16)
§5.6.2, README	"0.547" (v1.1 ten-state run) and README random-forest "0.542"	Traced to `extended_results.pkl` (0.5468 and 0.5423; rounds not stored); superseded by the v2 values
§7.3	School variables as "imputation-as-geography" that inflates in-sample accuracy only	Removing them lowers held-out accuracy under every geographic design (Tables 26, 25)
Table 29	Lot elasticity 0.018 (v1.1 coding) vs. 0.095 recoded	Confirmed on the rebuilt matrix: 0.0182 vs. 0.0959 (full sample, Table 29)

Notes: Every corrected value is produced by a committed script from the raw DuckDB extract or a stored artefact (`run_all.sh`); the v1.2 numbers are quoted for the record. Generated for item WP03-EDIT-10.

Table 45. Reproduction of the v1.1 Quantile Regressions from the Rebuilt Design Matrix

	"Pseudo-R^2"		$\lvert\Delta\beta\rvert$		ln(lot)		Parking spaces	
τ	stored	rebuilt R^1	median	max, excl. parking	stored	rebuilt	stored	rebuilt
0.10	0.583	0.417	0.000000	0.0002	0.0451	0.0451	0.0	−0.013
0.25	0.598	0.402	0.000001	0.0004	0.0507	0.0507	691.3	691.257
0.50	0.601	0.399	0.000013	0.0003	0.0614	0.0614	317.2	317.144
0.75	0.580	0.420	0.000024	0.1101	0.0703	0.0703	−97.8	−0.014
0.90	0.544	0.456	0.000014	0.5940	0.0860	0.0860	−527.3	−0.015

Notes: The five v1.1 quantile regressions (uncorrected design, ln(lot) = 0 coding, corrupt parking column) re-estimated from the rebuilt design matrix and compared with the stored `qr_results.pkl`. All 62 coefficients other than parking reproduce to within 0.59 (median difference below 10^{-4}); the stored "pseudo-R^2" equals $1 - R^1(\tau)$ exactly; the stored parking coefficients at $\tau = 0.75$ and 0.90 do not reproduce because the column (maximum 1.1×10^{11}) makes the interior-point path solver-dependent (Appendix F). Generated for items WP03-DATA-07 and ADVE-06.

Table 46. Reproduction of the v1.1 Random-Validation Comparison from the Stored Predictions (Uncorrected Design)

Model	R^2	RMSE	MAE	MAPE (%)	MdAPE (%)	MAE ($)
OLS	0.6301	0.4911	0.3619	39.8	27.2	240,941
Ridge (alpha = 1)	—	—	—	—	—	—
Lasso (alpha = 0.001)	—	—	—	—	—	—
Elastic Net	—	—	—	—	—	—
OLS + state FE	0.6775	0.4586	—	—	—	—
OLS + ZIP3 FE	0.7253	0.4232	—	—	—	—
Random forest	0.7842	0.3751	0.2636	—	—	—
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: Metrics recomputed from the predictions stored in `ml_results.pkl` (v1.1 design, 631,073 training and 157,769 test listings, seed 42). Cells marked "—" are not reproducible: the Ridge, Lasso and elastic-net rows of v1.2 (0.6301, 0.6280, 0.6145) have no stored predictions and the random-forest dollar metrics were never computed. These models are re-specified on the v2 design in Table 16.