UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

DÉPARTEMENT DES SCIENCES ADMINISTRATIVES

WORKING PAPER NO. 11

Half a Million Prices, Twenty Models

A Systematic Assessment of Hedonic Specifications and Estimation Methods for the Quebec Housing Market, 2021–2026

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

August 2026

Version 1.0

Abstract

Fifty years after Rosen (1974), applied hedonic practice still rests on a series of specification choices — functional form, temporal controls, spatial controls, estimation method — that are usually made by convention and rarely stress-tested jointly. Using 514,212 Quebec residential sales (2021–2026) whose structural descriptors come from the provincial assessment roll, we run a controlled horse race of twenty hedonic models organised along those four axes, scoring every model on identical holdouts and on the same price-level scoreboard (log models are Duan-retransformed, Box–Cox inverted). Four lessons emerge. (1) Functional form is a second-order choice: from linear to splines, the median absolute error spans 16.5–18.6%, and the data-driven Box–Cox exponent ($\hat{\lambda} = 0.25$) buys nothing out of sample. (2) Controls are first-order: dropping spatial fixed effects costs ten error points — location granularity matters far more than curvature, but only up to the point where cells thin out (a $\sim$1 km grid *overfits*). (3) Under random validation, estimation method matters most: gradient boosting reaches 14.4% — a 13% improvement over the best linear model — with random forests close behind, and coordinates alone reproduce what thousands of dummies buy. (4) But the ranking is an artifact of the split: under a forward-in-time holdout (train $<$ 2025, test 2025–26) every model degrades and the machine-learning advantage *reverses* — the spline hedonic posts a lower median error than gradient boosting and an identical log-scale R^2 — a stark warning against evaluating valuation models with random cross-validation alone. Constant-quality monthly indices implied by the competing forms agree to within a few index points, and the boosted model's implicit age and floor-area profiles track the quadratic OLS closely — interpretation and prediction disagree less than the random-validation gap suggests.

Keywords: hedonic pricing, functional form, house price indexes, machine learning, automated valuation, out-of-sample prediction

JEL Classification: R31, C21, C52, C53

1. Introduction

Every applied hedonic study begins with a series of quiet decisions. Log the price or not? Dummies for years or for months? Neighbourhood fixed effects, and how fine? Least squares or, increasingly, a gradient-boosted ensemble? Fifty years after Rosen (1974) established that theory places no restriction on the shape of the price surface, these choices are still made largely by convention — the semi-log survives as the default (Sirmans et al., 2005), the classic form comparisons predate the out-of-sample era (Halvorsen and Pollakowski, 1981; Cropper et al., 1988), and the machine-learning comparisons that have multiplied since Mullainathan and Spiess (2017) typically change the sample, the features and the validation scheme all at once, making it impossible to say *which* ingredient buys the reported gains.

This paper runs the comparison the way one would design an experiment. From 514,212 Quebec residential sales (2021–2026) matched at the parcel level to the provincial assessment roll — so that every model sees the same assessor-grade attribute set: floor area, lot area, age, storeys, units, property class, physical configuration — we estimate **twenty models organised along four axes**, varying one design choice at a time: six functional forms (linear to Box–Cox to splines), a ladder of time controls (none to month fixed effects), a ladder of spatial controls (none to a $\sim$1 km grid), and six estimation methods (OLS to random forests, gradient boosting, and a k-nearest-neighbour comparables rule). Every model is scored on the same price-level scoreboard — log models retransformed with Duan smearing, Box–Cox inverted — and, crucially, under **two holdouts**: the standard random 80/20 split, and a forward-in-time split (train before 2025, test 2025–26) that mimics how a valuation model is actually deployed.

Four results organise the paper.

First, functional form is a second-order choice. Across linear, semi-log, log-log, profile-likelihood Box–Cox ($\hat{\lambda} = 0.25$), quadratic and spline specifications — holding controls fixed — the median absolute error spans 16.5% to 18.6%. Splines beat the textbook semi-log by 1.6 points; the data-driven Box–Cox transformation, the great hope of the 1980s debate, buys nothing out of sample, vindicating Cropper et al. (1988) and Cassel and Mendelsohn (1985) on modern data at scale.

Second, controls are first-order — with a twist. Removing all spatial controls costs ten error points (27.9% versus 17.5%), an order of magnitude more than any curvature decision; a $\sim$5.5 km grid beats both municipality effects and, notably, a $\sim$1.1 km grid, whose 30,548 cells overfit — the bias–variance trade-off in absorbing location is real and non-monotonic. Time effects show the same shape in miniature: anything beats nothing (three points), granularity beyond quarters buys nothing.

Third, under random validation the estimation method dominates. Gradient boosting

reaches a median error of 14.4% and random forests 14.6%, against 16.5% for the best linear model — a 13% improvement — and stripping the coordinates from the boosting model degrades it to 23.2%, showing that what the machine mainly learns is a flexible price surface over space (Bourassa et al., 2010). The naïve k-NN comparables rule (21.2%) confirms that neither raw proximity nor raw flexibility suffices: the gains come from their combination.

Fourth — our headline — the ranking is an artifact of the validation scheme. Under the forward-in-time split every model degrades, but not equally: the machine-learning advantage *reverses* on the median error (gradient boosting 21.2% versus 19.1% for the spline hedonic) and equalizes exactly on log-scale R^2 (0.52 for both). Random cross-validation lets flexible models interpolate the price level of their own test period; when the test period lies in the future — the only case that matters for deployment — that advantage is leakage, not skill. Comparisons of valuation models that report random cross-validation alone, i.e. most of the applied ML literature, systematically overstate the practical value of model flexibility (Clapp and Giaccotto, 2002; Steurer et al., 2021).

The paper then asks what the specification choices do to the objects economists actually extract from hedonic models. Constant-quality monthly indices implied by the month effects of four linear forms — and by repricing a fixed portfolio with the boosting model — agree to within a few index points across the largest housing cycle in recent Canadian history. And the boosting model's partial-dependence profiles in age and floor area track the quadratic OLS closely: the machine agrees with the economist about the gradients and disagrees mainly about the residual surface. Interpretation and prediction, in other words, conflict far less than the accuracy scoreboard suggests.

Section 2 situates the exercise in the functional-form, spatial-hedonic and ML-valuation literatures. Section 3 describes the data, Section 4 the experimental design. Section 5 reports the horse race, Section 6 the index, implicit-price, learning-curve and segment extensions, Section 7 discusses implications, and Section 8 concludes.

2. Related literature

2.1. The functional-form debate

Hedonic theory is silent about functional form. Rosen (1974) showed that the price surface is an equilibrium envelope of bids and offers, so nothing structural pins down its curvature — form is an empirical choice. The classical exchange settled into a three-cornered debate: Halvorsen and Pollakowski (1981) advocated data-driven Box–Cox transformations (Box and Cox, 1964); Cassel and Mendelsohn (1985) cautioned that flexible transformations optimize in-sample fit while degrading the implicit prices researchers actually use; and Cropper et al. (1988), in an influential simulation study, found that simple forms — linear and semi-log — are the most robust to the

misspecification and omitted variables that characterize real data. Fifty years on, the semi-log remains the profession's default (Malpezzi, 2003; Sirmans et al., 2005), but systematic head-to-head evidence on modern samples remains scarce, and the classic studies predate the out-of-sample validation culture. Two further technical strands feed our design: retransformation bias in log models (Duan, 1983), which we correct with the smearing factor so that every form competes in levels; and forecast-evaluation methodology for house prices (Clapp and Giaccotto, 2002; Steurer et al., 2021).

2.2. Space and time in hedonic models

A second literature concerns what the covariates cannot capture. House prices are spatially autocorrelated at fine scale (Dubin, 1998; Anselin, 1988), and the practical remedies range from submarket segmentation (Goodman and Thibodeau, 1998) to spatial econometrics and simple location fixed effects; comparative studies generally find that flexibly absorbing location buys more predictive accuracy than any parametric spatial process (Case et al., 2004; Bourassa et al., 2010; Pace and Gilley, 1997). In the Quebec context, Des Rosiers et al. (2000) document the importance of access and neighbourhood attributes in Quebec City. On the time dimension, hedonic price-index theory — the time-dummy versus imputation debate — is surveyed by Hill (2013) and codified in the RPPI handbook (Eurostat, ILO, IMF, OECD, UNECE, World Bank, 2013; Silver, 2018); a byproduct of our month-FE models is a direct test of how sensitive a constant-quality index is to the underlying form.

2.3. Machine learning and automated valuation

The third strand is the rapid colonization of mass valuation by machine learning. Mullainathan and Spiess (2017) frame prediction as ML's comparative advantage, with hedonic pricing as their running example; Athey and Imbens (2019) survey the toolkit. Applied comparisons — random forests (Breiman, 2001), gradient boosting (Friedman, 2001; Ke et al., 2017) — consistently report 20–40% error reductions over linear hedonics (Mayer et al., 2019; Hong et al., 2020; Kok et al., 2017), with accuracy degrading in thin rural markets (Bogin and Shui, 2020); Gençay and Yang (1996) anticipated the pattern with semiparametric estimators. Three gaps motivate our design. First, most comparisons change several ingredients at once (features, sample, validation scheme), so the *sources* of the ML advantage — curvature? interactions? spatial flexibility? — are rarely isolated; our axis design holds features constant and varies one choice at a time. Second, almost all published comparisons use random cross-validation, which leaks future price levels into training and flatters any flexible model; we score every model under a forward-in-time split as well. Third, comparisons seldom examine what the flexible models *imply* — indices, implicit prices — which is

what economists need; our extensions section does exactly that.

3. Data

3.1. Sources

We use a registry of 745,119 Quebec residential-market transactions recorded between January 2021 and July 2026 — sale price, date, address and coordinates — each matched at the parcel level to the municipal assessment roll in force at the sale date (median match distance 0.6 metres, validated against the roll's recorded value). The roll contributes the structural descriptors that are notoriously missing from transaction registries: total floor area, lot area, year of construction, number of dwelling units, number of storeys, physical configuration (detached, semi-detached, row, apartment-linked) and the standardized use code. This combination — market prices with assessor-grade structure for half a million sales across an entire province — is what makes a controlled specification race feasible: every model sees exactly the same attributes.

3.2. Sample

We keep residential use codes (dwellings, cottages, mobile homes), require a high-confidence roll match (distance $\leq$ 50 m, score $\geq$ 150), a price of at least \$50,000, a plausible floor area (20–2,000 m^2) and a known year of construction, and trim prices and floor areas at the 1st and 99th percentiles within each sale year. The estimation sample contains **514,212 sales**: 348,253 single-family homes, 75,734 condominiums, 74,909 plexes (2–5 units), 10,349 cottages and 4,881 mobile homes, spread over 1,115 municipalities, 5,255 grid cells of $\sim$5.5 km and 30,548 cells of $\sim$1.1 km. Table 1 reports the descriptives. Prices average \$462,000 (median \$399,000); the window covers the tail of the post-pandemic boom, the 2022–2023 rate correction and the subsequent recovery — a demanding environment for any static price model, which we exploit in the forward-in-time evaluation.

4. Methodology: the design of the horse race

The comparison is organised so that each design choice varies while everything else is held fixed. All models price the same sales from the same attribute set; only the functional mapping changes.

4.1. The four axes

A. Functional form. With municipality and quarter fixed effects and the base attributes (floor area, lot area, building age, storeys, units, property class, physical configuration), we estimate: (A1)

Table 1. Summary statistics, estimation sample

Variable	n	Mean	SD	P10	Median	P90
Sale price ($)	514,212	440,207	245,716	170,000	397,375	759,000
Floor area (m^2)	514,212	132.36	61.54	76.90	111.70	217.40
Lot area (m^2)	514,212	1,060	1,783	114	557	2,474
Building age (years)	514,212	43.90	29.78	10.00	40.00	80.00
Storeys	514,212	1.36	0.51	1.00	1.00	2.00
Dwelling units	514,212	1.24	0.69	1.00	1.00	2.00

Notes: 514,212 residential sales, January 2021 – July 2026. Structural descriptors from the assessment roll in force at the sale date. Lot area is zero for condominiums and other units without an exclusive lot.

fully linear in levels; (A2) semi-log, $\ln P$ on levels (Malpezzi, 2003); (A3) log-log, $\ln P$ on logged areas; (A4) Box–Cox on the price, with λ chosen by profile likelihood on the training set over a grid (Box and Cox, 1964; Halvorsen and Pollakowski, 1981); (A5) semi-log with quadratics in area and age; and (A6) semi-log with cubic B-splines (six knots) in area, age and lot.

B. Time effects. On the A5 backbone: no time controls, year fixed effects, quarter fixed effects, month fixed effects — the granularity ladder of the time-dummy index literature (Hill, 2013).

C. Spatial controls. On the A5 backbone with quarter FE: no spatial controls, municipality FE (1,115), a ∼5.5 km grid (5,255 cells), and a ∼1.1 km grid (30,548 cells) — a granularity ladder that brackets the bias–variance trade-off in absorbing location (Case et al., 2004; Bourassa et al., 2010).

D. Estimation method. Holding the attribute set fixed: (D1) OLS with municipality and month FE; (D2) cross-validated ridge on the same standardized design; (D3) a random forest (Breiman, 2001) and (D4) gradient boosting (Friedman, 2001) on the attributes plus raw coordinates and a continuous month counter; (D5) gradient boosting *without* coordinates, to price what spatial flexibility is worth to the machine; and (D6) a spatial k-nearest-neighbour comparables rule — the distance-weighted price per square metre of the ten nearest training sales — as the naïve appraiser benchmark.

4.2. One scoreboard: price levels

Comparing forms whose dependent variables differ requires care. Every model is scored on *predicted price levels*: log models are retransformed with Duan's smearing factor estimated on training residuals (Duan, 1983); Box–Cox predictions are analytically inverted; level predictions are floored at $20,000. We report the median and mean absolute percentage errors (MdAPE, MAPE)

— the automated-valuation industry's standards (Steurer et al., 2021; International Association of Assessing Officers, 2013) — together with the RMSE and R^2 of log predicted versus log actual prices, which are well-defined for every model.

4.3. Two holdouts

Each model is evaluated under two splits. The *random* split holds out 20% of sales (seeded). The *forward-in-time* split trains on sales before January 2025 (355,824) and tests on 2025–2026 (158,388) — the deployment situation of any valuation model. Static models carry no forecast of the price level, so we apply the standard carry-forward rule: unseen time categories in the test period are priced at the last level observed in training; the machine-learning models receive the same information through the clamped month counter. Random cross-validation lets every model interpolate the price level of its own test period — the forward split reveals how much of measured accuracy is such leakage (Mullainathan and Spiess, 2017; Clapp and Giaccotto, 2002).

4.4. Estimation details

All linear models are estimated as sparse ridge regressions with a vanishing penalty ($\alpha = 10^{-6}$) — numerically OLS, but robust to collinear dummies, and unseen categories at prediction time are priced at the reference level. Dummies are one-hot with a dropped reference; the ridge (D2) standardizes columns and cross-validates $\alpha \in [10^{-6}, 10^{2}]$. The forest uses 120 trees (minimum leaf 5, half the features per split); the boosting machine uses 600 trees of at most 63 leaves (learning rate 0.08, minimum leaf 20). Hyper-parameters were fixed a priori at library-conventional values — the exercise compares model *classes* as used in practice, not tuned frontier performance.

5. The horse race

Table 2 reports the full scoreboard under the random holdout (411,283 training sales, 102,929 test sales); Table 3 repeats it under the forward-in-time split. Figures 1–4 visualize each axis.

5.1. Axis A: functional form is second-order

Under the random split (Figure 1), the six forms span 16.5–18.6% MdAPE. The spline specification wins (16.5%), the log-log is second (17.0%) — floor area enters multiplicatively, as intuition suggests — and the linear, semi-log and quadratic forms cluster within half a point of each other. The Box–Cox transformation selects $\hat{\lambda} = 0.25$ by profile likelihood (Table 4), i.e. the data reject both the linear ($\lambda = 1$) and the log ($\lambda = 0$) — and then *underperforms* the simple semi-log out of sample (18.6%), because optimizing curvature in-sample does not survive prediction, precisely the

Table 2. The scoreboard, random 80/20 holdout

Model	MdAPE (%)	MAPE (%)	$\text{RMSE}_{\ln}$	$R^2_{\ln}$	Fit (s)
A. Functional form					
A1 Linear (levels)	18.2	35.6	0.420	0.477	0
A2 Semi-log	18.2	34.5	0.397	0.533	1
A3 Log-log	17.0	32.7	0.381	0.569	1
A4 Box-Cox	18.6	33.0	0.390	0.548	4
A5 Semi-log + quadratics	18.1	34.6	0.397	0.533	1
A6 Semi-log + splines	**16.5**	32.1	0.376	0.579	3
B. Time effects					
B1 No time effects	21.1	38.4	0.423	0.468	0
B2 Year FE	18.1	34.7	0.397	0.532	1
B3 Quarter FE	**18.1**	34.6	0.397	0.533	1
B4 Month FE	18.3	34.8	0.398	0.529	1
C. Spatial controls					
C1 No spatial controls	27.9	48.4	0.494	0.274	0
C2 Municipality FE	18.1	34.6	0.397	0.533	1
C3 Grid-cell FE ($\sim$5.5 km)	**17.5**	34.0	0.392	0.544	1
C4 Grid-cell FE ($\sim$1.1 km)	19.4	37.5	0.418	0.481	1
D. Estimation method					
D1 OLS (muni + month FE)	18.3	34.8	0.398	0.529	1
D2 Ridge (cross-validated)	16.7	32.3	0.378	0.577	11
D3 Random forest	14.6	28.2	0.353	0.631	8
D4 Gradient boosting	**14.4**	27.8	0.347	0.642	42
D5 Gradient boosting, no coordinates	23.2	38.3	0.440	0.426	18
D6 Spatial k-NN comparables	21.2	36.7	0.437	0.434	0

Notes: All metrics on the held-out 20% (102,929 sales). MdAPE/MAPE: median/mean absolute percentage error on price levels (log models retransformed with Duan smearing; Box–Cox inverted). $\text{RMSE}_{\ln}$ and $R^2_{\ln}$: computed on log predicted vs. log actual prices. Backbone rows repeat across axes by construction (A5 = B3 = C2; B4 = D1). Bold: best in axis.

Table 3. The scoreboard, forward-in-time holdout

Model	MdAPE (%)	MAPE (%)	$\text{RMSE}_{\ln}$	$R^2_{\ln}$	Fit (s)
A. Functional form					
A1 Linear (levels)	20.4	32.9	0.397	0.455	0
A2 Semi-log	20.6	33.1	0.396	0.458	0
A3 Log-log	19.2	30.7	0.376	0.513	1
A4 Box-Cox	21.0	31.9	0.390	0.476	4
A5 Semi-log + quadratics	20.5	32.1	0.388	0.479	1
A6 Semi-log + splines	**19.1**	30.5	0.372	0.522	2
B. Time effects					
B1 No time effects	25.4	33.2	0.418	0.397	0
B2 Year FE	21.0	32.6	0.393	0.468	1
B3 Quarter FE	20.5	32.1	0.388	0.479	1
B4 Month FE	**20.3**	32.5	0.390	0.476	1
C. Spatial controls					
C1 No spatial controls	26.5	41.1	0.457	0.279	0
C2 Municipality FE	**20.5**	32.1	0.388	0.479	1
C3 Grid-cell FE ($\sim$5.5 km)	21.3	32.7	0.394	0.465	1
C4 Grid-cell FE ($\sim$1.1 km)	22.1	36.0	0.417	0.400	1
D. Estimation method					
D1 OLS (muni + month FE)	20.3	32.5	0.390	0.476	1
D2 Ridge (cross-validated)	**19.5**	30.6	0.375	0.515	9
D3 Random forest	21.3	29.8	0.373	0.519	7
D4 Gradient boosting	21.2	29.7	0.372	0.521	42
D5 Gradient boosting, no coordinates	26.2	35.7	0.447	0.310	21
D6 Spatial k-NN comparables	25.6	34.2	0.465	0.253	0

Notes: Training on sales before January 2025 (355,824); testing on 2025–2026 (158,388). Unseen time categories are priced at the last training period (carry-forward); the machine-learning models receive the same information through the clamped month counter.

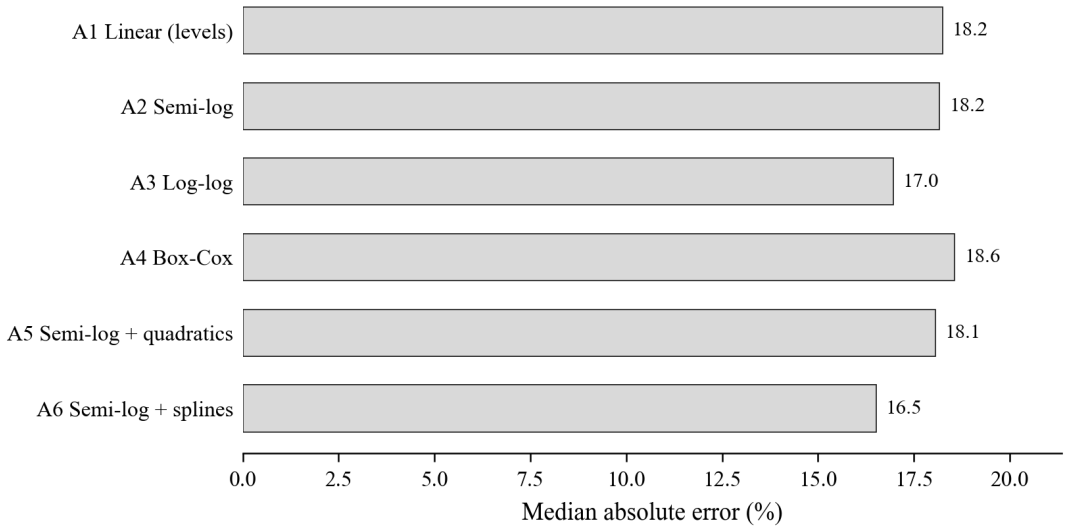


Figure 1. Axis A — functional form. MdAPE on the random holdout; municipality and quarter fixed effects and the base attribute set held fixed throughout.

Table 4. Box–Cox profile likelihood

λ	-0.50	-0.25	0.00	0.25	0.50	0.75	1.00
Profile log-likelihood[†]	-106.9	-59.7	-17.0	0.0	-3.7	-14.3	-45.8

Notes: [†]Thousands, relative to the maximum (zero = chosen λ). Estimated on the random-split training sample with the A-axis design.

warning of Cassel and Mendelsohn (1985) and Cropper et al. (1988). The ordering is preserved under the temporal split. Two free lessons: curvature in *attributes* (splines) is worth more than curvature in the *price* (Box–Cox); and no form choice moves the needle by more than two points.

5.2. Axis B: time controls — anything beats nothing

Omitting time controls entirely costs three points (21.1% versus 18.1%) in a window containing a boom, a correction and a recovery (Figure 2). But the granularity ladder flattens immediately: year, quarter and month effects are indistinguishable to within a quarter point under the random split. Under the temporal split the ordering inverts at the bottom — *no* time effects (25.4%) is the worst choice again, but month effects with carry-forward (20.3%) beat quarter effects (20.5%) only marginally: once the level must be extrapolated, fine within-sample time resolution has little to offer.

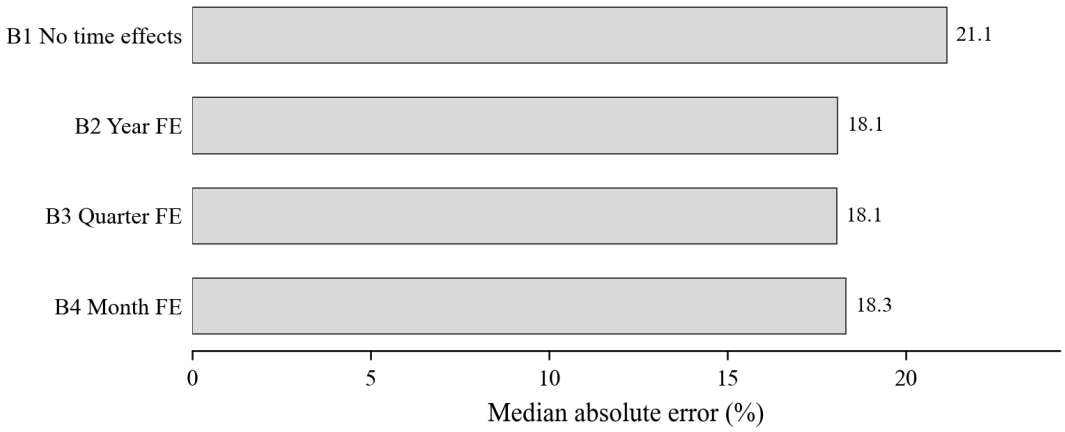


Figure 2. Axis B — time effects (quadratic semi-log backbone, municipality FE). MdAPE on the random holdout.

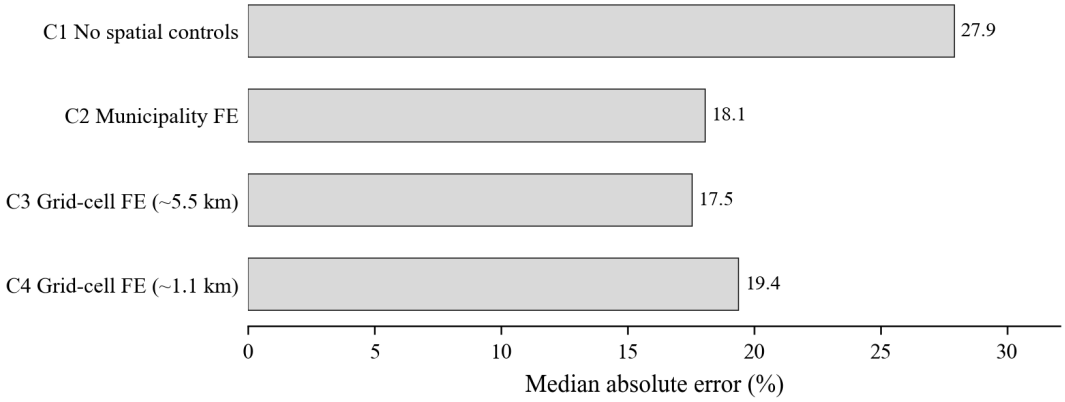


Figure 3. Axis C — spatial controls (quadratic semi-log backbone, quarter FE). MdAPE on the random holdout.

5.3. Axis C: spatial controls are first-order — up to a point

Location dominates every other design margin (Figure 3). Stripping all spatial controls costs *ten* points (27.9% versus 17.5%) — five times the entire functional-form range. The granularity ladder, however, is non-monotonic: the ∼5.5 km grid (5,255 cells, 17.5%) beats municipality effects (18.1%), but the ∼1.1 km grid (30,548 cells, 19.4%) *overfits* — its cells average seventeen K sales each, unseen cells in the holdout revert to the baseline, and estimated effects are noisy. The optimal spatial resolution is an interior solution, echoing Goodman and Thibodeau (1998) on submarket definition: finer is better only while cells remain thick enough to estimate.

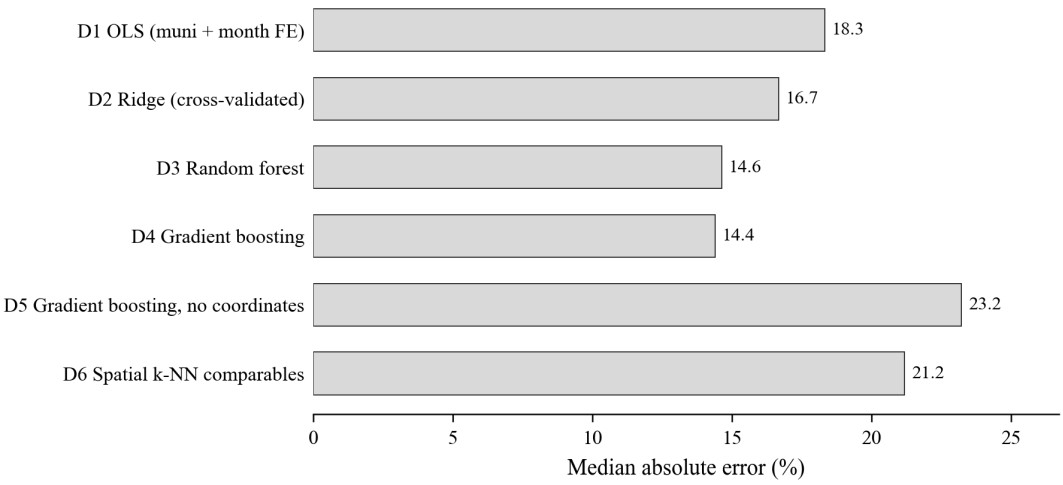


Figure 4. Axis D — estimation methods on the identical attribute set. MdAPE on the random holdout.

5.4. Axis D: the machines — and what they actually buy

Under the random split (Figure 4), gradient boosting posts 14.4% and the random forest 14.6%, against 16.5% for the best linear model and 18.3% for the textbook OLS: a 13–21% error reduction, squarely in the range reported by the AVM literature (Mayer et al., 2019; Hong et al., 2020). The decomposition rows identify the source. Boosting *without coordinates* collapses to 23.2% — worse than plain OLS with municipality dummies — so the machine's edge is overwhelmingly a flexible surface over space, not exotic structure–attribute interactions. Pure proximity without structure (the k-NN comparables rule, 21.2%) fails too: it is the *combination* of spatial flexibility with attribute adjustment that wins. The cross-validated ridge (16.7%) confirms that regularizing the high-dimensional dummy design buys a little; it cannot manufacture the missing interactions.

5.5. The generalization gap: when the test set is the future

Table 5 and Figure 5 contain the paper's central result. Moving from the random to the forward-in-time split degrades every model — the price level of 2025–26 must be carried forward, not interpolated — but the degradation is systematically larger for the flexible methods. Gradient boosting loses 6.8 points of MdAPE (14.4 $\rightarrow$ 21.2) and the random forest 6.7, while the spline hedonic loses 2.6 (16.5 $\rightarrow$ 19.1) and the log-log 2.2. The ranking *reverses*: the best linear models beat the machines on the median error in the deployment scenario, and tie them exactly on $R^2_{\ln}$ (≈ 0.52). The machines remain better on MAPE (29.7 versus 30.5), i.e. in the tails, but the headline random-validation gap — the number the ML-valuation literature reports — overstates their deployable advantage entirely. The lesson is methodological and general: for models whose job is to price the future, random cross-validation is the wrong experiment (Mullainathan and Spiess,

Table 5. Random versus forward-in-time evaluation

	Random holdout		Forward-in-time	
Model	MdAPE (%)	$R^2_{\ln}$	MdAPE (%)	$R^2_{\ln}$
A5 Semi-log + quadratics	18.1	0.533	20.5	0.479
B4 Month FE	18.3	0.529	20.3	0.476
C4 Grid-cell FE (∼1.1 km)	19.4	0.481	22.1	0.400
D1 OLS (muni + month FE)	18.3	0.529	20.3	0.476
D2 Ridge (cross-validated)	16.7	0.577	19.5	0.515
D3 Random forest	14.6	0.631	21.3	0.519
D4 Gradient boosting	14.4	0.642	21.2	0.521
D5 Gradient boosting, no coordinates	23.2	0.426	26.2	0.310
D6 Spatial k-NN comparables	21.2	0.434	25.6	0.253

Notes: Selected models (the backbone and the full D axis) under both splits.

2017; Clapp and Giaccotto, 2002).

6. Beyond the scoreboard: what the models imply

Accuracy is not the only output economists take from a hedonic model. This section asks whether the specification choices that move the scoreboard also move the two objects hedonic models are most often built to deliver: constant-quality price indices and implicit attribute prices.

6.1. Price indices are form-robust

Figure 6 plots the constant-quality monthly index implied by the month fixed effects of four linear forms — semi-log, log-log, quadratic and spline — together with a gradient-boosting index obtained by repricing a fixed 20,000-sale reference portfolio at each month (an imputation *à la* Hill, 2013). All five track the same cycle — the 2021–22 boom, the 2022–23 correction, the 2024–26 recovery — and the four linear forms agree to within a few index points throughout. The choice that dominates the accuracy scoreboard barely perturbs the index: reassuring news for statistical agencies, and consistent with the RPPI handbook's pragmatism about form (Eurostat, ILO, IMF, OECD, UNECE, World Bank, 2013; Silver, 2018).

6.2. Implicit prices: the machine agrees with the quadratic

Figure 7 compares the ln-price profiles in building age and floor area implied by the quadratic OLS coefficients with the partial-dependence profiles of the boosting model (Friedman, 2001). The two agree over the bulk of the data: steep initial depreciation that flattens with age, and strongly concave returns to floor space. The visible divergences are where the quadratic *cannot* bend — the machine

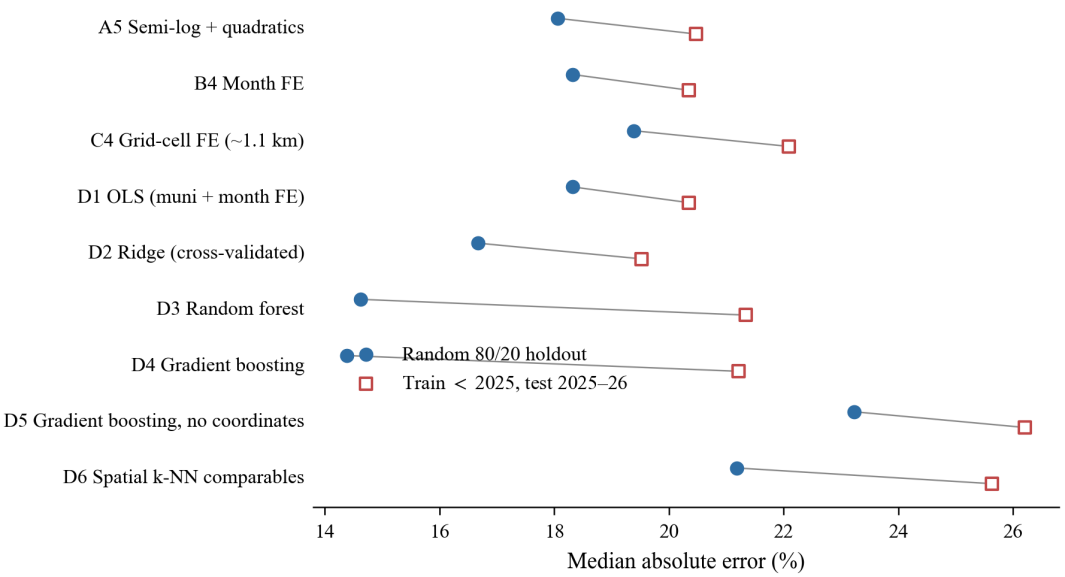


Figure 5. The generalization gap. MdAPE under the random holdout (filled circles) and the forward-in-time holdout (open squares); the grey segment joins the two evaluations of the same model.

sees a milder gradient for very old stock (a vintage effect) and a small new-construction premium spike — but through the ages and areas where 90% of sales live, the curves sit within a tenth of a log point. The boosting model's advantage evidently comes from higher-order interactions and spatial flexibility, *not* from a different view of the main attribute gradients — so the interpretable quadratic form remains a faithful summary of the price surface even where it loses the prediction race (Mullainathan and Spiess, 2017; Athey and Imbens, 2019).

6.3. Learning curves: data beat flexibility only at scale

Figure 8 and Table 6 trace accuracy against training size from 10,000 to 400,000 sales. The OLS quadratic is essentially flat from the very first step (18.4% at $n = 10{,}000$, 18.1% at $n = 400{,}000$) — its bias floor binds almost immediately — while the boosting model improves monotonically through the full sample (17.6% $\rightarrow$ 14.5%): the machine's advantage grows from under one error point at $n = 10{,}000$ to 3.6 points at $n = 400{,}000$. Flexible methods are not a free lunch in small markets: below a few tens of thousands of training sales, the machine's edge is modest, one reason rural and thin-market AVMs underperform (Bogin and Shui, 2020).

6.4. Where the machine wins: segments

Table 7 and Figure 9 decompose the random-holdout MdAPE by property class and municipality size, and the pattern identifies the machine's edge with striking precision: it is spatial resolution. The gap is enormous for condominiums (9.1% versus 20.4%) and in the largest cities (11.9%

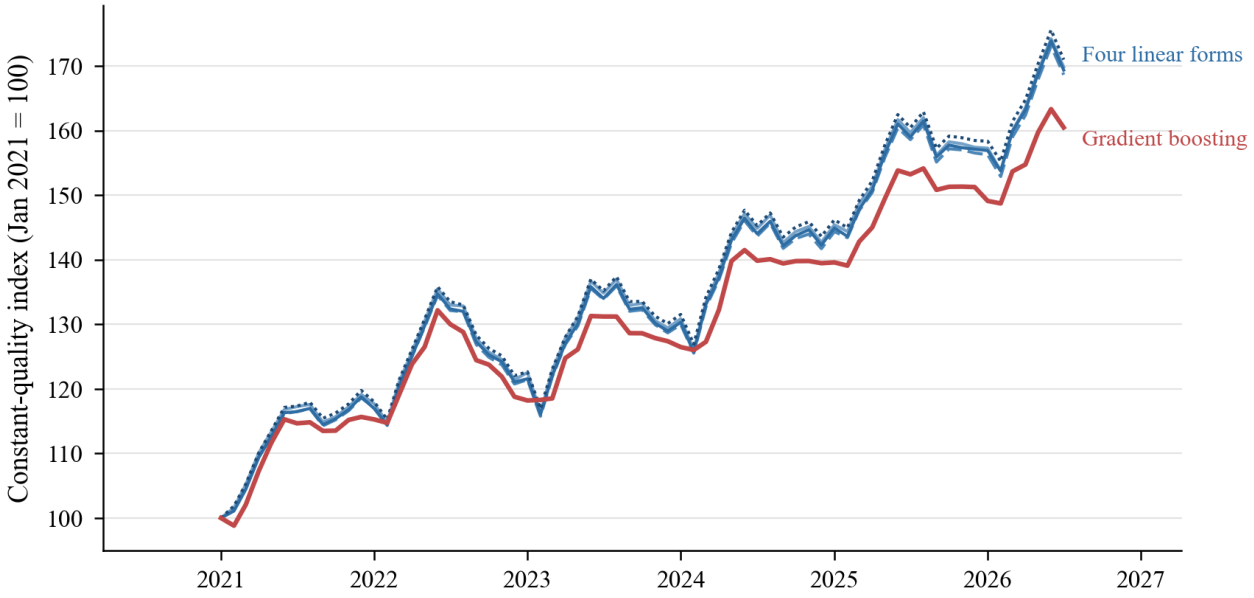


Figure 6. Constant-quality monthly price indices implied by competing specifications (January 2021 = 100). Linear forms: exponentiated month fixed effects. Gradient boosting: mean repriced value of a fixed 20,000-sale reference portfolio.

Table 6. Learning curves

Training sales	Gradient boosting	OLS quadratic (muni+quarter FE)
10,000	17.6	18.4
25,000	16.1	18.1
50,000	15.5	17.9
100,000	15.2	17.9
200,000	14.8	17.9
400,000	14.5	18.1

Notes: MdAPE (%) on the fixed random holdout, training on seeded subsamples of the indicated size.

versus 17.8%) — segments where structure is homogeneous but *micro*-location varies hugely within a municipality, exactly the variation that municipality dummies cannot see and raw coordinates can. Once location is resolved, a condo is the most predictable object in the market. At the other extreme, cottages defeat both models (32.1% versus 33.9%): idiosyncratic, waterfront-driven stock is hard for everyone, and the machine's spatial surface cannot rescue attributes it does not observe. The complement of the assessment-inequity anatomy in our companion paper (UQO WP10) is instructive: the segments where assessors are most regressive (heterogeneous, land-heavy stock) are also those where *no* standard technology, linear or boosted, prices accurately from roll attributes alone.

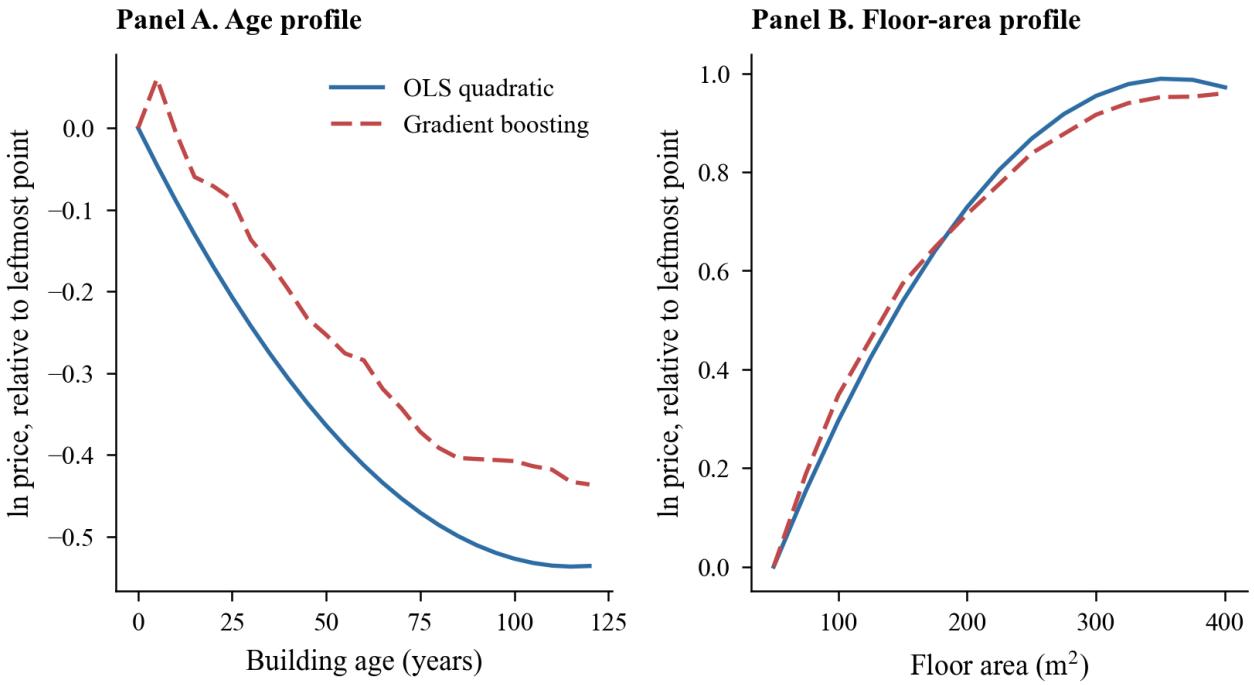


Figure 7. Implicit ln-price profiles: quadratic OLS coefficients versus gradient-boosting partial dependence, each normalized to its leftmost point. Panel A: building age. Panel B: floor area.

7. Discussion

7.1. A hierarchy of specification choices

Read jointly, the four axes imply a clear hierarchy of returns to specification effort, useful to anyone building a hedonic model:

1. **Get location right** ($\sim$10 points of median error): any spatial absorption beats none by a margin that dwarfs every other choice; the optimal granularity is interior — around 5 km here — because finer cells trade bias for variance.
2. **Include some time control** ($\sim$3 points): in a moving market, a model without time effects mistakes appreciation for attributes; granularity beyond quarters is cosmetic.
3. **Choose the estimator for the job** ($\sim$2 points deployed, $\sim$4 under random validation): boosting buys real accuracy in interpolation-heavy uses (filling gaps within a period, mass appraisal with contemporaneous comparables) and much less when the target is the future.
4. **Stop worrying about functional form** ($\lesssim$2 points): splines are a cheap upgrade; the Box–Cox machinery is not worth its complexity, exactly as Cropper et al. (1988) concluded from simulations four decades ago.

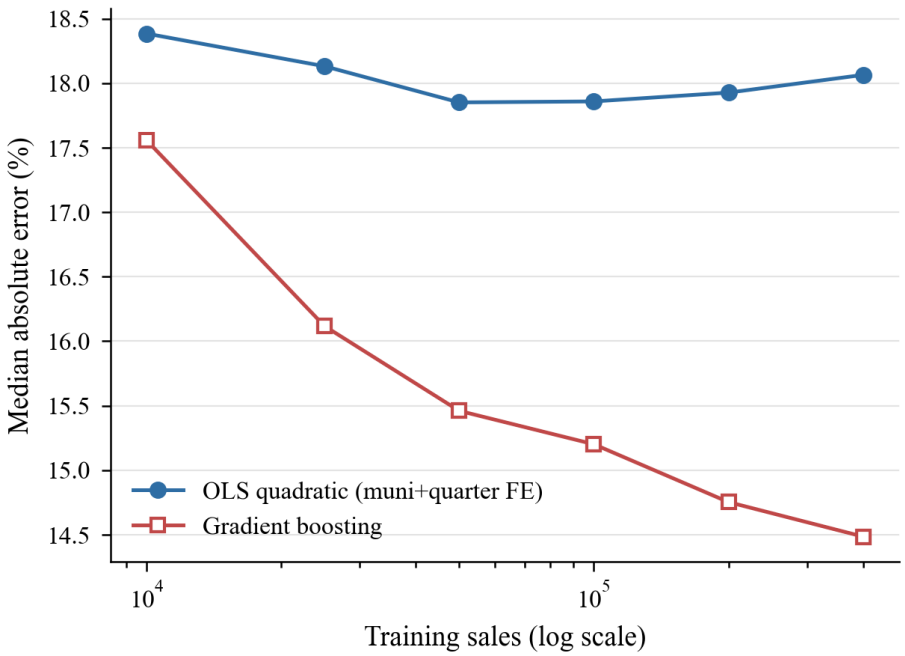


Figure 8. Learning curves: median absolute error against training-set size (random holdout fixed), OLS quadratic versus gradient boosting.

7.2. Why the machines lose their edge out of time

The decomposition rows explain the generalization gap. The boosting model's random-split advantage comes almost entirely from a flexible spatial surface (removing coordinates costs it 8.8 points). That surface is estimated *jointly* with the time path, and trees clamp: beyond the last training month the model prices 2025–26 sales at a frozen level, exactly like the carry-forward linear models, while its fine-grained spatial fit — calibrated on the 2021–24 price configuration — partially decays as relative prices shift. The linear models, coarser but more rigid, carry a structure that transfers better. This is not an indictment of machine learning — retrained monthly, the boosting model would presumably keep its interpolation advantage — but it prices the *retraining requirement* that random cross-validation hides, and it matches the practitioner evidence that AVM accuracy decays quickly out of sample period (Bogin and Shui, 2020; Kok et al., 2017).

7.3. Implications

For research. Hedonic coefficients and indices are robust objects: form and even estimator perturb them little (Section 6). Researchers using hedonics to *measure* — indices, implicit prices, capitalization effects — can keep simple forms with good controls and report validation under a temporal split when prediction claims are made.

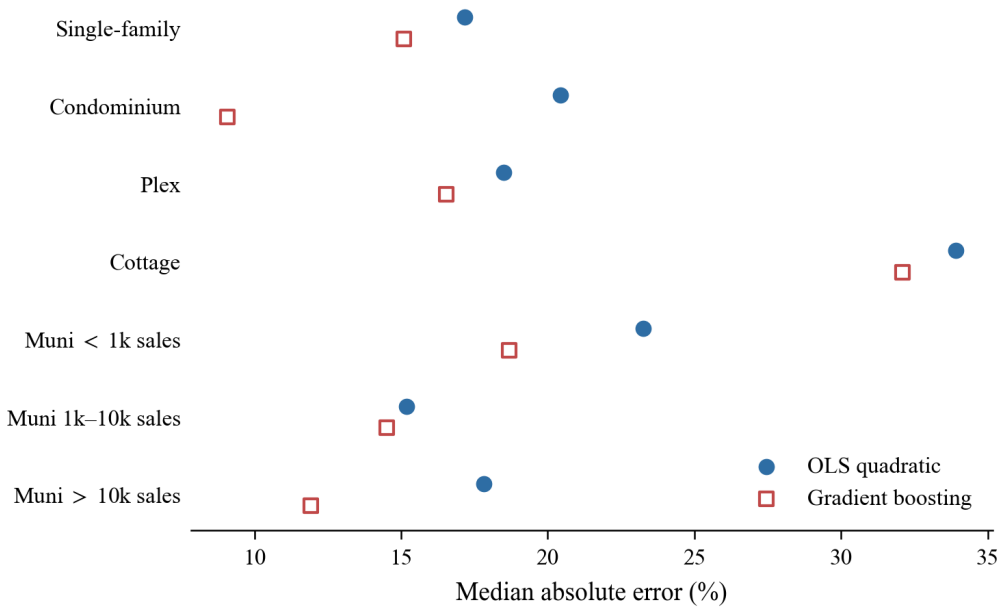


Figure 9. Median absolute error by market segment (random holdout): OLS quadratic versus gradient boosting.

For mass appraisal. Quebec's assessors face exactly this design problem every three years. Our results suggest the largest accuracy gains lie not in exotic estimators but in spatial resolution chosen by market thickness — and that any model, linear or boosted, frozen at a reference date degrades by 3–7 points of median error within two years. This quantifies the staleness mechanism behind the assessment inequities documented in our companion paper (UQO WP10): the valuation technology's out-of-time decay is of the same magnitude as the inequities measured there.

For the AVM literature. Every comparison should report a forward-in-time split alongside random cross-validation (Steurer et al., 2021). On our data, the choice of split changes not just magnitudes but the *winner*.

7.4. Limitations

Our attribute set, though assessor-grade, omits listing-level quality (renovations, finishes, views); richer features would likely raise the machines' interpolation advantage. Hyper-parameters were deliberately conventional — a tuned boosted model would gain somewhat, as would a spatially smarter linear model (e.g. kriging residuals, Case et al., 2004). The forward split covers one specific regime change (the 2025–26 recovery); other windows would shift magnitudes. And Quebec's institutional homogeneity aids transferability of results across municipalities but may limit it across countries.

Table 7. Accuracy by market segment

Segment	n (test)	OLS quadratic	Gradient boosting
Condo	14,848	20.4	9.1
Cottage	2,109	33.9	32.1
Plex	15,115	18.5	16.5
Single family	69,835	17.2	15.1
Muni 1k–10k sales	37,419	15.2	14.5
Muni $<$ 1k sales	29,253	23.3	18.7
Muni $>$ 10k sales	36,257	17.8	11.9

Notes: MdAPE (%) on the random holdout, by property class and by the number of 2021–2026 sales in the municipality.

8. Conclusion

We ran the hedonic specification debate as a controlled experiment: twenty models, four design axes, one attribute set, one scoreboard, two holdouts, half a million sales. The verdict inverts the profession's implicit priorities. The choices that consume the most methodological attention — functional form, transformation parameters — move out-of-sample accuracy by at most two points of median error, while the choices often made by default — whether and how finely to absorb space and time — move it by three to ten. Machine-learning estimators dominate the standard scoreboard, but the standard scoreboard asks the wrong question: when the test set lies in the future, as it always does in deployment, the boosted ensemble's advantage reverses against a spline hedonic with good controls, because most of what it learned was a spatially fine price configuration that does not transfer across the market cycle.

Meanwhile the objects economists actually harvest from hedonic models prove reassuringly stable: constant-quality indices agree across forms to within a few points over a full boom–bust–recovery cycle, and the machine's implicit age and area gradients replicate the quadratic OLS. The practical synthesis is almost embarrassingly simple: a semi-log with splines, quarter effects, and spatial fixed effects at a granularity matched to market thickness is within two points of the frontier in deployment conditions — transparent, auditable, and fast. Where the extra accuracy of learning methods is worth its retraining cadence — high-frequency AVMs, collateral monitoring — our decomposition says to spend the flexibility budget on the spatial surface, and to validate, always, against the future rather than against a shuffle of the past.

References

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

Athey, Susan and Guido W. Imbens, "Machine Learning Methods That Economists Should Know About," *Annual Review of Economics*, 2019, *11*, 685–725.

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), 40–52.

Bourassa, Steven C., Eva Cantoni, and Martin Hoesli, "Predicting House Prices with Spatial Dependence: A Comparison of Alternative Methods," *Journal of Real Estate Research*, 2010, *32* (2), 139–159.

Box, George E. P. and David R. Cox, "An Analysis of Transformations," *Journal of the Royal Statistical Society, Series B*, 1964, *26* (2), 211–252.

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

Case, Bradford, John Clapp, Robin Dubin, and Mauricio Rodriguez, "Modeling Spatial and Temporal House Price Patterns: A Comparison of Four Models," *Journal of Real Estate Finance and Economics*, 2004, *29* (2), 167–191.

Cassel, Eric and Robert Mendelsohn, "The Choice of Functional Forms for Hedonic Price Equations: Comment," *Journal of Urban Economics*, 1985, *18* (2), 135–142.

Clapp, John M. and Carmelo Giaccotto, "Evaluating House Price Forecasts," *Journal of Real Estate Research*, 2002, *24* (1), 1–26.

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.

Duan, Naihua, "Smearing Estimate: A Nonparametric Retransformation Method," *Journal of the American Statistical Association*, 1983, *78* (383), 605–610.

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

Eurostat, ILO, IMF, OECD, UNECE, World Bank, "Handbook on Residential Property Prices Indices (RPPIs)," Technical Report, Eurostat, Luxembourg 2013.

Friedman, Jerome H., "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, 2001, *29* (5), 1189–1232.

Gençay, Ramazan and Xian Yang, "A Forecast Comparison of Residential Housing Prices by Parametric versus Semiparametric Conditional Mean Estimators," *Economics Letters*, 1996, *52* (2), 129–135.

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

Halvorsen, Robert and Henry O. Pollakowski, "Choice of Functional Form for Hedonic Price Equations," *Journal of Urban Economics*, 1981, *10* (1), 37–49.

Hill, Robert J., "Hedonic Price Indexes for Residential Housing: A Survey, Evaluation and Taxonomy," *Journal of Economic Surveys*, 2013, *27* (5), 879–914.

Hong, Jengei, Heeyoul Choi, and Woo sung Kim, "A House Price Valuation Based on the Random Forest Approach: The Mass Appraisal of Residential Property in South Korea," *International Journal of Strategic Property Management*, 2020, *24* (3), 140–152.

International Association of Assessing Officers, *Standard on Ratio Studies*, Kansas City, MO: IAAO, 2013.

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," *Advances in Neural Information Processing Systems*, 2017, *30*, 3146–3154.

Kok, Nils, Eija-Leena Koponen, and Carmen Adriana Martínez-Barbosa, "Big Data in Real Estate? From Manual Appraisal to Automated Valuation," *Journal of Portfolio Management*, 2017, *43* (6), 202–211.

Malpezzi, Stephen, "Hedonic Pricing Models: A Selective and Applied Review," in Tony O'Sullivan and Kenneth Gibb, eds., *Housing Economics and Public Policy*, Oxford: Blackwell, 2003, pp. 67–89.

Mayer, Michael, Steven C. Bourassa, Martin Hoesli, and Donato Scognamiglio, "Estimation and Updating Methods for Hedonic Valuation," *Journal of European Real Estate Research*, 2019, *12* (1), 134–150.

Mullainathan, Sendhil and Jann Spiess, "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 2017, *31* (2), 87–106.

Pace, R. Kelley and Otis W. Gilley, "Using the Spatial Configuration of the Data to Improve Estimation," *Journal of Real Estate Finance and Economics*, 1997, *14* (3), 333–340.

Rosen, Sherwin, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, 1974, *82* (1), 34–55.

Rosiers, François Des, Marius Thériault, and Paul-Y. Villeneuve, "Sorting Out Access and Neighbourhood Factors in Hedonic Price Modelling," *Journal of Property Investment & Finance*, 2000, *18* (3), 291–315.

Silver, Mick, "How to Measure Hedonic Property Price Indexes Better," *EURONA — Eurostat Review on National Accounts and Macroeconomic Indicators*, 2018, *1*, 35–66.

Sirmans, G. Stacy, David A. Macpherson, and Emily N. 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.