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

October 2026

Version 1.1

Abstract

Fifty years after Rosen (1974), applied hedonic practice still rests on specification choices—functional form, temporal controls, spatial controls, estimation method—that are 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 with a single retransformation rule, with a municipality-clustered bootstrap interval and a Diebold–Mariano-type test on every comparison. Four lessons emerge. (1) Functional form is second-order: from linear to splines the median absolute error spans 17.6–19.2%, and the data-driven Box–Cox exponent gains half a point. (2) Controls are first-order: dropping spatial fixed effects costs ten error points; a $\sim$1 km grid used alone overfits, but nested within municipalities it becomes the most accurate linear specification—the bias–variance trade-off in absorbing location is resolved by hierarchy, not coarseness. (3) Under random validation the estimation method dominates: gradient boosting reaches 14.4%, an 18% improvement over the best linear model (95% interval 1.9–4.7 points) that survives holdouts grouped by parcel and by building and comes almost entirely from a flexible surface over space. (4) Under a forward-in-time holdout (train $<$ 2025, test 2025–26) every model degrades and the gap shrinks to half a point and to insignificance. Decomposing the error shows that relative pricing skill transfers forward intact (re-levelled, boosting leads by 2.9 points) while every frozen model under-prices the typical future sale by 10–20%, the ensembles slightly more. Random cross-validation hides this staleness cost; it does not, as a previous version of this paper claimed on the basis of an asymmetric retransformation, reverse the ranking. Constant-quality indices agree across functional forms to within 2.5 points but a boosted imputation index rises nine points less, and the boosted model's implicit age and area profiles share the shape of the quadratic OLS gradients with a milder depreciation path.

Keywords: hedonic price function, functional form, spatial fixed effects, machine learning, automated valuation model, forward-in-time holdout, house price index

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; Khoshnoud et al., 2023), the classic functional-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 at once, so that it is 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, with a single retransformation rule for every model whose dependent variable is a transformed price, and 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. Because differences of a few tenths of a point between models evaluated on the same sales can be noise, every comparison comes with a municipality-clustered bootstrap interval and a Diebold–Mariano-type test of equal accuracy.

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 17.6% to 19.2% under the random holdout. Splines beat the textbook semi-log by 1.4 points; the data-driven Box–Cox transformation, the great hope of the 1980s debate, gains half a point on the median error and nothing on the mean error, 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 (28.8% against 19.2%), six times the entire functional-form range; a $\sim$5.5 km grid is nominally but not significantly better than municipality effects, and a $\sim$1.1 km grid used on its own overfits (21.0%), its 30,548 cells holding three sales at the median. Nested within municipalities, so that thin and unseen cells inherit their municipality's level, the same fine grid becomes the most accurate linear specification in the paper (17.3%): the bias–variance trade-off in absorbing location is real,

and it is resolved by hierarchy rather than by coarseness. Time effects show the same shape in miniature: anything beats nothing (three points under the random split, six under the forward split), and 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 17.6% for the best linear model, an 18% improvement that is statistically unambiguous (-3.2 points, 95% interval -4.7 to -1.9). Stripping the coordinates from the boosting model degrades it to 23.2%, so what the learner mainly buys is a flexible price surface over space (Bourassa et al., 2010; Pace and Hayunga, 2020); the naïve k-NN comparables rule (21.2%) shows that raw proximity without attribute adjustment does not suffice either. The advantage is not an artefact of repeat sales or of condominium units geocoded to the same building: it survives holdouts grouped by parcel and by coordinate pair essentially unchanged.

Fourth—our headline—the advantage is a within-period statement. Under the forward-in-time split every model degrades, the ensembles most: gradient boosting loses 6.8 points (to 21.2%) and the spline hedonic 4.1 (to 21.7%), so the three-point gap shrinks to half a point and loses statistical significance, while the ensembles keep a modest lead on the mean error and on log-scale fit. Decomposing the forward error shows why. The 2025–26 sales are not harder to price: evaluated on the same sales, a boosting model that has seen their period posts 13.7%. What a frozen model loses is the *level*: every forward-split model under-prices the typical 2025–26 sale by 10–20%, the ensembles slightly more than the linear models because their terminal leaf in time averages the last training months. Once each model's level is corrected, the boosting advantage over the spline hedonic is 2.9 points, essentially what the random split measures. Relative pricing skill transfers forward almost intact; the price level does not, and random cross-validation hides a staleness cost of three to seven points of median error within eighteen months that any deployed model, linear or boosted, must pay unless it is re-levelled by an index or retrained. A previous version of this paper reported that the ranking reversed in the forward split; the reversal was an artefact of an asymmetric retransformation and is corrected here.

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 agree to within 2.5 index points across the largest housing cycle in recent Canadian history, whereas the imputation index from the boosting model rises nine points less over five years: form does not move the index, the estimator can. The boosting model's partial-dependence profiles in age and floor area share the shape of the quadratic OLS gradients, with a uniformly milder depreciation path, so the machine disagrees with the economist about the residual surface rather than about the main gradients. 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 machine-learning

valuation literatures. Section 3 describes the data, Section 4 the experimental design, the scoreboard and the inference procedures. Section 5 reports the horse race and decomposes the generalization gap, Section 6 the index, implicit-price, learning-curve and segment extensions, Section 7 the robustness analyses (retransformation rule, leakage-free holdouts, nested spatial effects, Box–Cox grid, boosting hyper-parameters), Section 8 discusses implications, and Section 9 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 bid and offer functions, so nothing structural pins down its curvature: the form is an empirical choice. The classical exchange settled into a three-cornered debate. Halvorsen and Pollakowski (1981) advocated letting the data choose the transformation through the Box–Cox family (Box and Cox, 1964); Cassel and Mendelsohn (1985) cautioned that flexible transformations optimise in-sample fit while degrading the implicit prices researchers actually use; and Cropper et al. (1988), in a simulation study that remains the standard reference, found that simple forms, linear and semi-log, are the most robust to the misspecification and omitted variables that characterise real data. The debate has been revisited for welfare measurement rather than prediction: Kuminoff et al. (2010) show in a Monte Carlo design calibrated to actual housing data that the choice of control variables (notably spatial fixed effects) matters more for recovering marginal willingness to pay than the choice of functional form, and Bishop et al. (2020) codify the resulting best practice. Fifty years on, the semi-log remains the profession's default (Malpezzi, 2003; Sirmans et al., 2005; Khoshnoud et al., 2023), but systematic head-to-head evidence on large modern samples remains scarce, and the classic studies predate the out-of-sample validation culture. Two technical strands feed our design. Retransformation bias in log models (Duan, 1983; Manning and Mullahy, 2001) determines how predictions in logarithms are converted into price levels, a step that turns out to matter for the comparison between linear models and tree ensembles. Forecast-evaluation methodology, from the equal-accuracy tests of Diebold and Mariano (1995) and Giacomini and White (2006) to their application to house prices (Clapp and Giaccotto, 2002; Steurer et al., 2021), supplies the inferential framework we use to decide whether differences on the scoreboard are real.

2.2. Space and time in hedonic models

A second literature concerns what the covariates cannot capture. House prices are spatially auto-correlated at fine scale (Anselin, 1988; Dubin, 1998; Pace et al., 1998), and the practical remedies range from submarket segmentation (Goodman and Thibodeau, 1998, 2003; Bourassa et al., 2003)

to spatial econometric processes, semiparametric location surfaces (Clapp, 2003; Fik et al., 2003) and simple location fixed effects. Comparative studies generally find that flexibly absorbing location buys more predictive accuracy than any parametric spatial process, and that the gain depends on how submarkets are drawn (Pace and Gilley, 1997; Case et al., 2004; Bourassa et al., 2007, 2010). Pace and Hayunga (2020) show with tree ensembles that the residuals of standard hedonic and spatial models still contain exploitable spatial information, which is precisely the margin our coordinate-based learners target. In the Quebec context, Des Rosiers et al. (2000) document the importance of access and neighbourhood attributes in Quebec City, and Dubé and Legros (2013) stress that sales pooled over time are not exchangeable: a sale can inform the price of a later neighbour but not of an earlier one, which is the asymmetry a forward-in-time holdout respects and a random split ignores. On the time dimension, the time-dummy versus imputation debate in hedonic price-index theory is surveyed by Hill (2013) and codified in the RPPI handbook (Eurostat, ILO, IMF, OECD, UNECE, World Bank, 2013; Silver, 2016); Hill et al. (2009) and Hill and Scholz (2018) show that imputation indices built from flexible, spline-based hedonic models track a boom and bust more faithfully, and Diewert et al. (2015) and Diewert and Shimizu (2016) develop the hedonic decomposition of prices into land and structure components. A by-product of our month-effects models is a direct test of how sensitive a constant-quality index is to the underlying functional form and to the estimator.

2.3. Machine learning and automated valuation

The third strand is the rapid adoption of machine learning in mass valuation. Mullainathan and Spiess (2017) frame prediction as the comparative advantage of machine learning, with hedonic pricing as their running example; Athey and Imbens (2019) survey the toolkit. The applied comparisons, from early neural networks and random forests (Zurada et al., 2011; Antipov and Pokryshevskaya, 2012; McCluskey et al., 2013) to gradient boosting (Kok et al., 2017; Mayer et al., 2019; Pérez-Rave et al., 2019; Hong et al., 2020), consistently report error reductions of 20–40% over linear hedonics, with accuracy degrading in thin rural markets (Bogin and Shui, 2020); Gençay and Yang (1996) anticipated the pattern with semiparametric estimators, and Lorenz et al. (2023) show that the gains come with interpretable partial-dependence profiles that economists can read as implicit prices. Mayer et al. (2019) is the closest antecedent to our forward-in-time design: on Swiss data they compare estimation methods together with updating rules (moving windows, weighting) and find that the choice of updating matters as much as the estimator; we isolate the same question by scoring frozen models on a future period. Three gaps motivate our design. First, most comparisons change several ingredients at once (features, sample, validation scheme), so the sources of the machine-learning advantage—curvature, interactions, spatial flexibility—are rarely isolated; our axis design holds the features constant and varies one choice at a time (compare the specification exercise of Schulz et al., 2014). Second, almost all published comparisons use

random cross-validation, which leaks the future price level into training, and, when properties sell repeatedly or share a building, lets a coordinate-based learner recall a property's own earlier price; the forecasting and ecology literatures have long documented how such temporal and spatial dependence inflates cross-validated accuracy (Bergmeir and Benítez, 2012; Roberts et al., 2017). We score every model under a forward-in-time split and under parcel- and building-grouped splits as well. Third, comparisons seldom examine what the flexible models *imply*—indices, implicit prices—which is what economists need; Section 6 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 same snapshot underlies our companion paper (UQO WP10). The roll contributes the structural descriptors that are typically 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 standardised 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, and the attribute set is the one a mass-appraisal office actually holds.

3.2. Sample

We keep residential use codes (dwellings, cottages, mobile homes), require a high-confidence roll match (distance ≤ 50 m, match score ≥ 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. Building age is the sale year minus the year of construction, bounded between 0 and 150; storeys are bounded between 1 and 6 and dwelling units between 1 and 12; lot areas above the 99.5th percentile, which are implausible for residential parcels, are set to zero and treated as units without an exclusive lot, together with condominiums. The estimation sample contains **514,212 sales**: 348,253 single-family homes, 75,734 condominiums, 74,909 plexes (two to five units), 10,349 cottages, 4,881 mobile homes and 86 dwellings of undetermined class, spread over 1,115 municipalities, 5,255 grid cells of $\sim$5.5 km and 30,548 cells of $\sim$1.1 km. Sales are spread across the window (97,868 in 2021, 90,733 in 2022, 76,411 in 2023, 90,812 in 2024, 103,830 in 2025 and 54,558 in January–July 2026). Table 1 reports the descriptive statistics. Prices

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 (and for the 0.5% of lots above the 99.5th percentile, see text).

average $440,207 (median $397,375), floor areas 132 m^2 and building ages 44 years; the window covers the tail of the post-pandemic boom, the 2022–23 rate correction and the 2024–26 recovery, during which the constant-quality price level rose by roughly 70% (Section 6). This is a demanding environment for any static price model, which the forward-in-time evaluation exploits. One fifth of the sales (105,814) concern a parcel that sells more than once in the window and 30% share their exact coordinates with another sale; Section 4.4 explains why these repeat and co-located sales matter for validation.

4. Methodology: the design of the horse race

The comparison is organised so that one design choice varies at a time while everything else is held fixed. All twenty models price the same sales from the same attribute vector; only the mapping from attributes to price changes. This section defines the models, the retransformation rule that puts them on a common scoreboard, the two holdouts and the inference procedures.

4.1. Notation and the baseline hedonic function

For sale i let P_i denote the price, $\mathbf{x}_i$ the vector of structural attributes supplied by the assessment roll (floor area a_i in m^2, lot area l_i in m^2 with an indicator $h_i = \mathbb{1}[l_i > 0]$, building age g_i in years, storeys, dwelling units, and dummies for property class and physical configuration), $m(i)$ the municipality, $c_5(i)$ and $c_1(i)$ the $\sim$5.5 km and $\sim$1.1 km grid cells containing the parcel, and $y(i)$, $q(i)$, $\tau(i)$ the year, quarter and month of sale. A linear hedonic model with a transformation $g(\cdot)$ of the price, a set $\mathscr{S}$ of spatial fixed effects and a set $\mathscr{T}$ of time fixed effects is

$$g(P_i) \;=\; \phi(\mathbf{x}_i)'\beta \;+\; \mu_{s(i)} \;+\; \delta_{t(i)} \;+\; \varepsilon_i, \qquad s(i) \in \mathscr{S},\, t(i) \in \mathscr{T}, \tag{1}$$

where $\phi(\cdot)$ is a basis expansion of the attributes (identity, logarithms, quadratics or splines), μ_s absorbs location and δ_t the market level. Following Rosen (1974), nothing in theory restricts g or ϕ; the axes below make those choices explicit. The baseline, used as the backbone of axes B and C, takes $g = \ln$, quadratics in area and age, $\mathscr{S}$ = municipalities and $\mathscr{T}$ = quarters.

4.2. The four axes

A. Functional form. With municipality and quarter fixed effects held fixed, the six forms differ only in g and ϕ:

$$\text{A1 (linear):} \quad P_i = \mathbf{x}_i'\beta + \mu_{m(i)} + \delta_{q(i)} + \varepsilon_i; \tag{2}$$

$$\text{A2 (semi-log):} \quad \ln P_i = \mathbf{x}_i'\beta + \mu_{m(i)} + \delta_{q(i)} + \varepsilon_i; \tag{3}$$

$$\text{A3 (log-log):} \quad \ln P_i = \beta_a \ln a_i + \beta_l h_i \ln l_i + \beta_h h_i + \tilde{\mathbf{x}}_i'\tilde{\beta} + \mu_{m(i)} + \delta_{q(i)} + \varepsilon_i; \tag{4}$$

$$\text{A4 (Box–Cox):} \quad \frac{P_i^{\lambda} - 1}{\lambda} = \mathbf{x}_i'\beta + \mu_{m(i)} + \delta_{q(i)} + \varepsilon_i; \tag{5}$$

$$\text{A5 (quadratics):} \quad \ln P_i = \mathbf{x}_i'\beta + \beta_{aa}\left(\tfrac{a_i}{100}\right)^2 + \beta_{gg}\left(\tfrac{g_i}{10}\right)^2 + \mu_{m(i)} + \delta_{q(i)} + \varepsilon_i; \tag{6}$$

$$\text{A6 (splines):} \quad \ln P_i = \sum_{v\in\{a,l,g\}} \sum_{k=1}^{8} \beta_{vk} B_{vk}(v_i) + \tilde{\mathbf{x}}_i'\tilde{\beta} + \mu_{m(i)} + \delta_{q(i)} + \varepsilon_i, \tag{7}$$

where $\tilde{\mathbf{x}}_i$ collects the remaining attributes and B_{vk} are cubic B-spline basis functions with six uniformly spaced knots over the training range of each variable (eight basis functions per variable). In the Box–Cox model (Box and Cox, 1964; Halvorsen and Pollakowski, 1981) the exponent is chosen on the training sample by maximising the profile log-likelihood

$$\ell(\lambda) \;=\; -\frac{n}{2}\ln\!\left(\frac{\mathrm{SSE}(\lambda)}{n}\right) \;+\; (\lambda - 1)\sum_{i=1}^{n} \ln P_i, \tag{8}$$

where $\mathrm{SSE}(\lambda)$ is the residual sum of squares of (5) at λ and the second term is the Jacobian of the transformation. The main grid is $\lambda \in \{-0.5, -0.25, \ldots, 1\}$; Section 7 refines it to steps of 0.05.

B. Time effects. On the A5 backbone with municipality fixed effects, $\mathscr{T}$ runs through the granularity ladder of the time-dummy index literature (Hill, 2013): no time controls (B1), year (B2), quarter (B3 = A5) and month fixed effects (B4).

C. Spatial controls. On the A5 backbone with quarter fixed effects, $\mathscr{S}$ runs from no spatial controls (C1) to municipality fixed effects (C2 = A5; 1,115 cells, median 79 sales per cell), a 0.05° grid (C3; 5,255 cells of roughly 5.5 km, median 10 sales) and a 0.01° grid (C4; 30,548 cells

of roughly 1.1 km, mean 17 and median 3 sales per cell). The ladder brackets the bias–variance trade-off in absorbing location (Case et al., 2004; Bourassa et al., 2010). Because the dummies are one-hot encoded with a dropped reference category, a test sale in a cell absent from the training set is priced at the reference cell's level; this affects 0.2% of random-holdout sales under C3 and 2.5% under C4. Section 7 reports nested variants (municipality plus grid effects) in which an unseen cell falls back on its municipality instead.

D. Estimation method. Holding the attribute set fixed, D1 is the semi-log quadratic OLS with municipality and month fixed effects (identical to B4); D2 replaces least squares by ridge regression on the same standardised design,

$$\hat{\beta}_\alpha = \arg\min_\beta \sum_i \left(\ln P_i - \mathbf{z}_i'\beta\right)^2 + \alpha\|\beta\|_2^2, \qquad \alpha \in \{10^{-6}, 10^{-5.5}, \ldots, 10^{2}\}, \tag{9}$$

with α selected by three-fold cross-validation within the training set. D3 is a random forest (Breiman, 2001) and D4 a histogram-based gradient-boosting machine (Friedman, 2001; Ke et al., 2017), both fitted to $\ln P_i$ on the attribute vector augmented with the raw coordinates $(\mathrm{lat}_i, \mathrm{lng}_i)$ and a continuous month counter t_i (months since January 2021), with property class and physical configuration entered as integer codes. D5 is the boosting machine without the coordinates, which prices what spatial flexibility is worth to the learner. D6 is a spatial k-nearest-neighbour comparables rule, the naïve appraiser benchmark:

$$\widehat{\ln P_i} \;=\; \ln a_i \;+\; \frac{\sum_{j\in N_{10}(i)} w_{ij}\left(\ln P_j - \ln a_j\right)}{\sum_{j\in N_{10}(i)} w_{ij}}, \qquad w_{ij} = d_{ij}^{-1}, \tag{10}$$

where $N_{10}(i)$ contains the ten training sales nearest to i in kilometre-scaled coordinates and d_{ij} is the planar distance. Hyper-parameters are fixed a priori at library-conventional values: the forest uses 120 trees, a minimum leaf size of 5 and half the features per split; the boosting machine uses 600 iterations, a learning rate of 0.08, at most 63 leaves per tree, a minimum leaf size of 20 and an ℓ_2 penalty of 0.01. The ridge is the only method tuned on the data, and only within the training set. The exercise therefore compares model *classes* as commonly used, not tuned frontier performance; Section 7 shows that the ranking is insensitive to materially different boosting configurations, including one tuned on a time-ordered validation set.

4.3. One scoreboard: price levels

Comparing models whose dependent variables differ requires every model to be scored on the same object. We score *predicted price levels*. Let $\hat{z}_i$ be the prediction of model M on the transformed scale ($g(P)$ for the linear forms, $\ln P$ for the tree ensembles and the k-NN rule) on a test sale i. Two

inversion rules are available. The *median rule* inverts the fitted value as is,

$$\hat{P}_i^{\text{med}} = g^{-1}(\hat{z}_i) = \begin{cases} \exp(\hat{z}_i) & g = \ln, \\ (\lambda \hat{z}_i + 1)^{1/\lambda} & g = \text{Box–Cox}, \end{cases} \tag{11}$$

which under symmetric transformed-scale errors is the conditional median price and is the quantity the tree ensembles deliver natively. The *mean rule* applies Duan's (1983) nonparametric smearing estimator of the conditional mean,

$$\hat{P}_i^{\text{mean}} = \frac{1}{n}\sum_{j=1}^{n} g^{-1}\left(\hat{z}_i + \hat{e}_j\right) \xrightarrow{g=\ln} \exp(\hat{z}_i)\cdot \hat{s}, \qquad \hat{s} = \frac{1}{n}\sum_{j=1}^{n}\exp(\hat{e}_j), \tag{12}$$

where $\hat{e}_j$ are the training residuals on the transformed scale. The smearing factor $\hat{s}$ exceeds one and corrects the downward bias of $\exp(\hat{z})$ as an estimator of $E[P \mid \mathbf{x}]$ (Manning and Mullahy, 2001); for the Box–Cox model we evaluate (12) on a random subsample of 1,000 training residuals. The main scoreboard uses the median rule for every model, for two reasons. First, it is the only rule that treats all twenty models identically without a model-specific calibration constant: the training residuals of tree ensembles are shrunk by overfitting, so a smearing factor estimated from them is not comparable to one estimated from least-squares residuals. Second, our headline metric is a median error, whose natural target is the conditional median, and the log-scale metrics below are only meaningful on uncorrected log predictions. Section 7 reports the full scoreboard under the mean rule as well. Version 1.0 of this paper smeared the linear log models but not the tree ensembles; the asymmetry affected the comparison between the two families and is corrected here.

Predictions are floored at \$20,000. With $\hat{P}_i$ the level prediction and N the number of test sales, the four metrics are

$$\text{APE}_i = \frac{|\hat{P}_i - P_i|}{P_i}, \qquad \text{MdAPE} = \text{median}_i\,\text{APE}_i, \qquad \text{MAPE} = \frac{1}{N}\sum_i \text{APE}_i, \tag{13}$$

$$\text{RMSE}_{\ln} = \left(\frac{1}{N}\sum_i \left(\ln P_i - \ln \hat{P}_i\right)^2\right)^{1/2}, \qquad R^2_{\ln} = 1 - \frac{\sum_i (\ln P_i - \ln \hat{P}_i)^2}{\sum_i (\ln P_i - \overline{\ln P})^2}. \tag{14}$$

MdAPE and MAPE are the accuracy standards of the automated-valuation industry and of ratio studies (International Association of Assessing Officers, 2013; Steurer et al., 2021); $R^2_{\ln}$ is an out-of-sample R^2 on the log scale, well defined for every model whatever its dependent variable.

4.4. Two holdouts, and leakage-free variants

Each model is evaluated under two splits. The *random* split assigns every sale independently to the test set with probability 0.2 (seeded; 411,283 training and 102,929 test sales). The *forward-in-time*

split trains on sales dated before 1 January 2025 (355,824) and tests on 2025–26 (158,388), the deployment situation of any valuation model that is estimated once and applied to subsequent sales. Static models carry no forecast of the price level, so we apply the standard carry-forward rule: a test sale whose time category (year, quarter or month) was never observed in training is priced at the last category observed in training, $\hat{\delta}_t = \hat{\delta}_T$ for $t > T$. The tree ensembles are piecewise constant in the month counter, so a test month beyond the training range is priced by the terminal leaf in t, the tree analogue of the same rule. Random cross-validation lets every model interpolate the price level of its own test period; the forward split reveals how much of measured accuracy depends on that interpolation (Bergmeir and Benítez, 2012; Roberts et al., 2017).

The random split has a second, less obvious leakage channel. One fifth of the sales in the sample are repeat sales of a parcel that sells again within the window, and 30% share their exact geocoded coordinates with another sale, 92% among condominiums, whose units are geocoded to the building. Under a random split, a model that reads the coordinates can learn a parcel's or a building's price from its other sales. We therefore also evaluate the full race under two grouped holdouts in which 20% of *parcels*, respectively 20% of *coordinate pairs*, are assigned to the test set, so that no property (respectively no building) appears on both sides (Roberts et al., 2017).

4.5. Inference

Differences of a few tenths of a point between models evaluated on the same 100,000 sales can be either meaningful or noise, and the sales are not independent: they cluster in municipalities whose local shocks move all errors together. We report two complementary tools. First, a *cluster bootstrap* for MdAPE: municipalities are resampled with replacement ($B = 500$), each draw contributing all its test sales, and the median APE of every model and the difference between two models' MdAPEs are recomputed; we report percentile 95% intervals. Second, a test of equal predictive accuracy in the spirit of Diebold and Mariano (1995) and Giacomini and White (2006). For models A and B and a loss L, let $d_i = L(\hat{P}_i^A, P_i) - L(\hat{P}_i^B, P_i)$; the statistic

$$t_{\mathrm{DM}} = \frac{\bar{d}}{\widehat{\mathrm{se}}(\bar{d})}, \qquad \widehat{\mathrm{se}}(\bar{d})^2 = \frac{G}{G-1}\frac{1}{N^2}\sum_{g=1}^{G}\Big(\sum_{i\in g}(d_i - \bar{d})\Big)^2, \tag{15}$$

uses standard errors clustered by municipality ($G = 1{,}115$), the cross-sectional analogue of the HAC variance in the time-series version. We use the absolute percentage error (the MAPE loss) and the squared log error (the $\mathrm{RMSE}_{\ln}$ loss); a positive $\bar{d}$ means that B is more accurate. The mean of the APE loss differential tests MAPE, not MdAPE; the bootstrap covers the median.

Finally, because the forward-in-time holdout confounds the difficulty of the test period with the staleness of a frozen model, we decompose forward error into an aggregate level error and relative

pricing skill. For each model, $\hat{b} = \text{median}_i(\ln P_i - \ln \hat{P}_i)$ on the test set measures the common level error; the *re-levelled* MdAPE, computed with $\hat{P}_i e^{\hat{b}}$, is the accuracy the model would attain if the aggregate price level of the test period were known, the time adjustment practitioners obtain from a published index (Thibodeau, 2003). We also compare the random-split models on the 2025–26 sales of their own test set with the forward-split models on the full 2025–26 test set, which holds the test period fixed and varies only whether that period was seen in training.

5. The horse race

Table 2 reports the full scoreboard under the random holdout (411,283 training sales, 102,929 test sales) and Table 3 under the forward-in-time split; Table 4 adds the cluster-bootstrap intervals and the equal-accuracy tests against the spline hedonic (A6), and Table 5 the targeted comparisons along each axis. Figures 1–4 visualise the four axes under the random holdout. All numbers use the median rule of Section 4.3; Section 7 shows how the Duan-smeared alternative moves them.

5.1. Axis A: functional form is second-order

Under the random split (Figure 1) the six forms span 17.6–19.2% MdAPE, a range of 1.5 points. The spline specification wins (17.6%), the log-log is second (18.0%)—floor area enters multiplicatively, as intuition suggests—and the linear (18.2%), Box–Cox (18.6%), semi-log (19.0%) and quadratic (19.2%) forms follow within a point of each other. The spline gain over the textbook semi-log, -1.4 points $[-1.8, -1.1]$, is precisely estimated and holds on every metric (Table 5). The Box–Cox transformation selects $\hat{\lambda} = 0.25$ by profile likelihood (Table 6), so the data reject both the linear ($\lambda = 1$) and the logarithmic ($\lambda = 0$) specification, yet the model gains only half a point on the semi-log's median error, loses on the mean error (DM $t = 9.0$ in favour of the semi-log) and remains a full point behind the splines. Optimising the curvature of the *price* in sample buys little out of sample, the warning of Cassel and Mendelsohn (1985) and Cropper et al. (1988); curvature in the *attributes*, which splines supply cheaply, is worth more.

Under the forward-in-time split (Table 3) the range widens to 20.4–23.3% and the ordering changes: the linear model in levels posts the lowest median error (20.4%), followed by the Box–Cox (21.0%), spline (21.7%) and log-log (21.8%) forms, with the semi-log and quadratic forms last (23.2–23.3%). The linear model's forward-split advantage is a bias coincidence rather than a property of the form: being a conditional-mean predictor in levels, its log prediction over-predicts the typical sale by about 5% under the random split (median log residual $+0.004$ against $+0.06$ for the log models), and in a rising market that upward tilt offsets part of the staleness of a frozen price level. Once the common level error is removed (Section 5.5), the spline form is again the most accurate linear specification (16.8% against 18.0% for the linear model). Two lessons survive

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	19.0	32.7	0.391	0.546	1
A3 Log-log	18.0	31.2	0.376	0.580	1
A4 Box-Cox	18.6	33.0	0.390	0.548	5
A5 Semi-log + quadratics	19.2	32.9	0.391	0.546	1
A6 Semi-log + splines	**17.6**	30.8	0.372	0.590	3
B. Time effects					
B1 No time effects	22.0	35.9	0.416	0.486	0
B2 Year FE	**19.1**	32.9	0.391	0.545	1
B3 Quarter FE	19.2	32.9	0.391	0.546	1
B4 Month FE	19.3	33.0	0.393	0.543	1
C. Spatial controls					
C1 No spatial controls	28.8	43.7	0.483	0.307	0
C2 Municipality FE	19.2	32.9	0.391	0.546	1
C3 Grid-cell FE ($\sim$5.5 km)	**18.7**	32.5	0.387	0.557	1
C4 Grid-cell FE ($\sim$1.1 km)	21.0	35.6	0.412	0.496	1
D. Estimation method					
D1 OLS (muni + month FE)	19.3	33.0	0.393	0.543	1
D2 Ridge (cross-validated)	17.7	30.9	0.373	0.587	11
D3 Random forest	14.6	28.2	0.353	0.631	12
D4 Gradient boosting	**14.4**	27.8	0.347	0.642	47
D5 Gradient boosting, no coordinates	23.2	38.3	0.440	0.426	17
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, eq. (13); transformed-scale predictions inverted with the median rule, eq. (11). $\text{RMSE}_{\ln}$ and $R^2_{\ln}$: eq. (14). Fit: wall-clock seconds on an 18-core workstation. Backbone rows repeat across axes by construction (A5 = B3 = C2; B4 = D1). Bold: lowest MdAPE within the 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	23.2	32.9	0.408	0.426	0
A3 Log-log	21.8	31.0	0.388	0.481	1
A4 Box-Cox	21.0	31.9	0.390	0.476	4
A5 Semi-log + quadratics	23.3	32.3	0.402	0.442	1
A6 Semi-log + splines	21.7	30.9	0.384	0.491	2
B. Time effects					
B1 No time effects	29.3	34.7	0.450	0.300	0
B2 Year FE	23.7	32.7	0.406	0.431	1
B3 Quarter FE	23.3	32.3	0.402	0.442	1
B4 Month FE	**22.9**	32.4	0.401	0.445	1
C. Spatial controls					
C1 No spatial controls	29.0	38.9	0.469	0.242	0
C2 Municipality FE	**23.3**	32.3	0.402	0.442	1
C3 Grid-cell FE ($\sim$5.5 km)	24.3	33.0	0.409	0.423	1
C4 Grid-cell FE ($\sim$1.1 km)	25.3	35.5	0.429	0.364	1
D. Estimation method					
D1 OLS (muni + month FE)	22.9	32.4	0.401	0.445	1
D2 Ridge (cross-validated)	22.3	31.1	0.388	0.480	8
D3 Random forest	21.3	29.8	0.373	0.519	11
D4 Gradient boosting	**21.2**	29.7	0.372	0.521	41
D5 Gradient boosting, no coordinates	26.2	35.7	0.447	0.310	22
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 2025–26 (158,388). Unseen time categories are priced at the last training period (carry-forward); the tree ensembles are piecewise constant in the month counter and therefore apply the same rule through their terminal leaf. Other conventions as in Table 2.

Table 4. Inference: bootstrap intervals and equal-accuracy tests against the spline hedonic

	Random holdout			Forward-in-time holdout		
Model	MdAPE [95% CI]	Δ vs. A6 [95% CI]	t_{DM}	MdAPE [95% CI]	Δ vs. A6 [95% CI]	t_{DM}
A. Functional form						
A1 Linear	18.2 [16.9, 19.6]	+0.6 [+0.1, +1.1]	12.3***	20.4 [19.1, 21.7]	-1.4 [-2.3, -0.4]	6.3***
A2 Semi-log	19.0 [17.9, 20.2]	+1.4 [+1.1, +1.8]	10.2***	23.2 [21.7, 24.7]	+1.4 [+0.8, +2.3]	7.8***
A3 Log-log	18.0 [17.1, 19.0]	+0.4 [+0.3, +0.5]	4.6***	21.8 [20.5, 23.2]	+0.0 [-0.0, +0.2]	1.8*
A4 Box–Cox	18.6 [17.5, 19.7]	+0.9 [+0.7, +1.3]	10.7***	21.0 [19.7, 22.2]	-0.8 [-1.4, -0.1]	4.3***
A5 Quadratics	19.2 [18.2, 20.2]	+1.5 [+1.1, +2.1]	5.5***	23.3 [21.9, 24.9]	+1.6 [+1.1, +2.2]	6.1***
A6 Splines	17.6 [16.7, 18.6]	—	—	21.7 [20.5, 23.1]	—	—
B. Time effects						
B1 No time FE	22.0 [21.1, 23.3]	+4.4 [+3.4, +5.6]	8.2***	29.3 [27.3, 31.5]	+7.6 [+6.0, +9.0]	5.3***
B2 Year FE	19.1 [18.1, 20.2]	+1.5 [+1.1, +2.2]	6.1***	23.7 [22.2, 25.3]	+1.9 [+1.3, +2.8]	6.2***
B3 Quarter FE	19.2 [18.2, 20.2]	+1.5 [+1.1, +2.1]	5.5***	23.3 [21.9, 24.9]	+1.6 [+1.1, +2.2]	6.1***
B4 Month FE	19.3 [18.4, 20.4]	+1.7 [+1.2, +2.5]	5.7***	22.9 [21.5, 24.4]	+1.1 [+0.7, +1.8]	6.0***
C. Spatial controls						
C1 No spatial FE	28.8 [27.3, 30.1]	+11.2 [+10.1, +12.3]	7.7***	29.0 [27.4, 30.3]	+7.3 [+5.0, +9.2]	11.1***
C2 Municipality FE	19.2 [18.2, 20.2]	+1.5 [+1.1, +2.1]	5.5***	23.3 [21.9, 24.9]	+1.6 [+1.1, +2.2]	6.1***
C3 Grid 5.5 km	18.7 [17.8, 20.0]	+1.1 [+0.2, +2.3]	2.5**	24.3 [22.5, 26.4]	+2.5 [+1.6, +3.5]	4.5***
C4 Grid 1.1 km	21.0 [20.1, 21.9]	+3.3 [+2.9, +3.8]	6.8***	25.3 [24.3, 26.3]	+3.6 [+2.1, +4.9]	12.4***
D. Estimation method						
D1 OLS	19.3 [18.4, 20.4]	+1.7 [+1.2, +2.5]	5.7***	22.9 [21.5, 24.4]	+1.1 [+0.7, +1.8]	6.0***
D2 Ridge	17.7 [16.8, 18.7]	+0.1 [-0.0, +0.2]	2.6***	22.3 [21.0, 23.7]	+0.5 [+0.5, +0.6]	6.6***
D3 Random forest	14.6 [13.7, 16.2]	-3.0 [-4.6, -1.5]	-2.9***	21.3 [19.9, 23.0]	-0.4 [-1.5, +0.6]	-1.6
D4 Boosting	14.4 [13.7, 15.7]	-3.2 [-4.7, -1.9]	-3.6***	21.2 [19.7, 22.9]	-0.5 [-1.7, +0.6]	-1.6
D5 Boosting, no coord.	23.2 [22.1, 25.1]	+5.6 [+3.9, +7.5]	4.5***	26.2 [25.2, 27.5]	+4.5 [+2.9, +6.2]	5.9***
D6 *k*-NN comparables	21.2 [19.4, 23.5]	+3.6 [+1.0, +5.9]	4.8***	25.6 [23.1, 28.3]	+3.9 [+1.8, +5.6]	4.0***

Notes: MdAPE in percent with a 95% percentile interval from 500 cluster-bootstrap replicates (municipalities resampled with replacement). Δ vs. A6: MdAPE of the row model minus MdAPE of A6 on the same test sales, with its bootstrap interval; negative values favour the row model. t_{DM}: clustered Diebold–Mariano statistic, eq. (15), for the absolute-percentage-error loss differential (row model minus A6); positive values mean A6 is more accurate on the mean loss. *, **, ***: $p < 0.10, 0.05, 0.01$ (two-sided). 1,115 municipality clusters.

Table 5. Targeted comparisons along each axis

			Random holdout		Forward-in-time	
Comparison		MdAPE	ΔMdAPE [95% CI]	t_{DM} APE / sq. log	ΔMdAPE [95% CI]	t_{DM} APE / sq. log
A2 Semi-log vs.	A1 Linear	19.0 / 18.2	+0.79 [+0.42, +1.22]	-10.7*** / -4.2***	+2.79 [+1.78, +3.88]	0.1 / 3.5***
A6 Splines vs.	A2 Semi-log	17.6 / 19.0	-1.41 [-1.83, -1.09]	-10.2*** / -8.9***	-1.43 [-2.30, -0.77]	-7.8*** / -7.2***
A4 Box–Cox vs.	A2 Semi-log	18.6 / 19.0	-0.47 [-0.59, -0.35]	9.0*** / -0.7	-2.20 [-3.15, -1.46]	-5.3*** / -7.0***
A3 Log-log vs.	A2 Semi-log	18.0 / 19.0	-1.03 [-1.45, -0.72]	-7.9*** / -10.1***	-1.39 [-2.24, -0.74]	-7.1*** / -6.3***
B2 Year FE vs.	B1 No time FE	19.1 / 22.0	-2.91 [-3.84, -2.06]	-8.1*** / -7.4***	-5.65 [-6.51, -4.52]	-3.8*** / -10.4***
B4 Month FE vs.	B3 Quarter FE	19.3 / 19.2	+0.18 [+0.01, +0.37]	2.9*** / 2.9***	-0.43 [-0.95, +0.05]	0.7 / -0.9
C3 Grid 5.5 km vs.	C2 Municipality FE	18.7 / 19.2	-0.45 [-1.07, +0.37]	-1.1 / -1.2	+0.96 [+0.16, +1.81]	2.3** / 1.6
C4 Grid 1.1 km vs.	C2 Municipality FE	21.0 / 19.2	+1.81 [+1.24, +2.32]	7.1*** / 7.3***	+2.02 [+0.15, +3.59]	8.1*** / 4.8***
D2 Ridge vs.	D1 OLS	17.7 / 19.3	-1.66 [-2.37, -1.13]	-5.5*** / -6.0***	-0.61 [-1.26, -0.17]	-5.6*** / -4.7***
D4 Boosting vs.	D3 Random forest	14.4 / 14.6	-0.24 [-0.46, -0.05]	-4.7*** / -6.7***	-0.13 [-0.37, +0.15]	-1.0 / -1.3
D4 Boosting vs.	D5 Boosting, no coord.	14.4 / 23.2	-8.84 [-9.79, -8.11]	-10.6*** / -16.1***	-4.99 [-6.50, -3.12]	-13.9*** / -11.1***
D4 Boosting vs.	A1 Linear	14.4 / 18.2	-3.86 [-5.57, -2.38]	-10.4*** / -11.9***	+0.84 [-1.19, +2.59]	-3.5*** / -3.9***
A6 Splines vs.	A1 Linear	17.6 / 18.2	-0.63 [-1.13, -0.13]	-12.3*** / -7.5***	+1.35 [+0.43, +2.28]	-6.3*** / -6.2***

Notes: For each pair, the MdAPE difference (first model minus second) with its 95% cluster-bootstrap interval and the clustered DM statistics for the APE and squared-log losses (positive: the second model is more accurate on the mean loss). Stars as in Table 4.

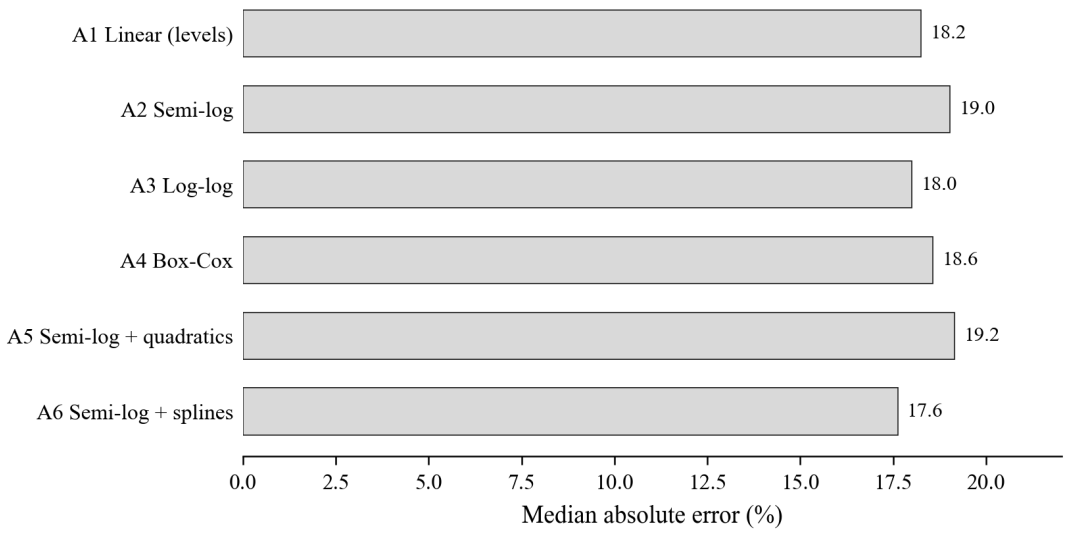


Figure 1. Axis A — functional form. MdAPE on the random holdout; municipality and quarter fixed effects and the base attribute set held fixed throughout.

Table 6. 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: †Profile log-likelihood, eq. (8), in thousands relative to the maximum (zero at the chosen λ). Estimated on the random-split training sample with the axis-A design. A finer grid (steps of 0.05, Section 7) selects $\hat{\lambda} = 0.35$ and lowers the random-holdout MdAPE from 18.6% to 18.4%.

both splits: no functional-form choice moves the median error by more than about two points, and splines dominate the semi-log everywhere.

5.2. Axis B: time controls — anything beats nothing

Omitting time controls entirely costs three points (22.0% against 19.2%, difference -2.9 [-3.8, -2.1] for year effects against none) in a window containing a boom, a correction and a recovery (Figure 2). The granularity ladder then flattens at once: year, quarter and month effects lie within 0.2 points of each other under the random split, and month effects are, if anything, marginally worse than quarter effects ($+0.19$ [$+0.01$, $+0.37$]), the first sign that finer within-sample time resolution buys nothing for prediction. Under the forward-in-time split the cost of omitting time controls rises to six points (29.3% against 23.3%), because a model without time effects prices 2025–26 sales at the *average* 2021–24 level rather than at the last observed level, and month effects with carry-forward (22.9%) beat quarter effects (23.3%) by a statistically fragile margin (-0.43 [-0.95, $+0.05$]). Once the level must be extrapolated, fine time resolution has little to offer; what matters is

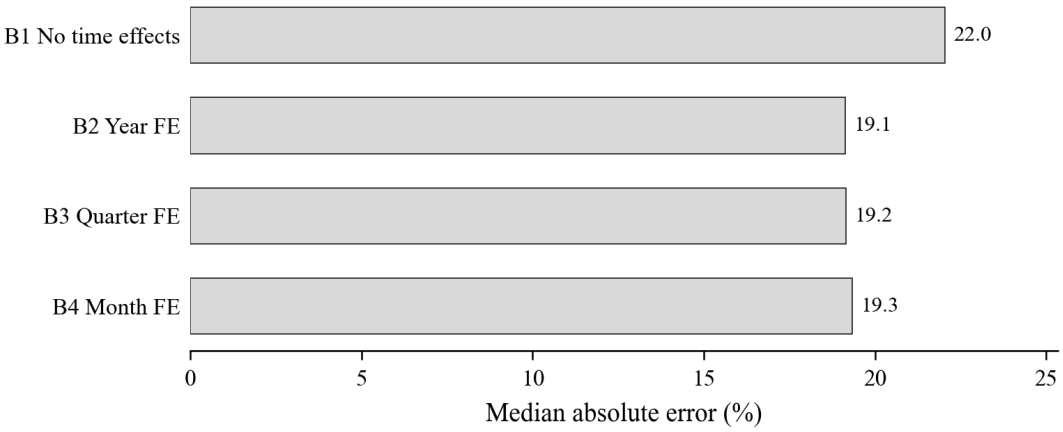


Figure 2. Axis B — time effects (quadratic semi-log backbone, municipality FE). MdAPE on the random holdout.

that the carried level be the most recent one.

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 (28.8% against 19.2% with municipality effects, 11.2 points [10.1, 12.3] relative to the spline hedonic), six times the entire functional-form range. The granularity ladder, however, is non-monotonic. The ∼5.5 km grid (5,255 cells) is nominally better than municipality effects (18.7% against 19.2%), but the difference, −0.45 [−1.07, +0.37], is within bootstrap noise and the DM tests do not reject equal accuracy; the ∼1.1 km grid (30,548 cells) is clearly worse (+1.8 [+1.2, +2.3]). Its cells contain 17 sales on average and 3 at the median, so most cell effects are estimated from a handful of sales, and the 2.5% of test sales that fall in a cell with no training sale are priced at an arbitrary reference cell. Under the forward-in-time split the ordering becomes monotonic in the wrong direction: finer cells are worse at every step (23.3%, 24.3%, 25.3%), because a fine spatial configuration estimated on 2021–24 prices transfers less well than a coarse one. The optimal spatial resolution is an interior solution that depends on market thickness, echoing Goodman and Thibodeau (1998) and Bourassa et al. (2007) on submarket definition: finer is better only while cells remain thick enough to estimate, and the relevant thickness is larger when the model must be carried forward in time. Section 7 shows that nesting the grid within municipalities, so that an unseen cell falls back on its municipality, repairs part but not all of the fine grid's deficit.

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

Under the random split (Figure 4) gradient boosting posts 14.4% [13.7, 15.7] and the random forest 14.6%, against 17.6% for the best linear model and 19.3% for the textbook OLS with municipality

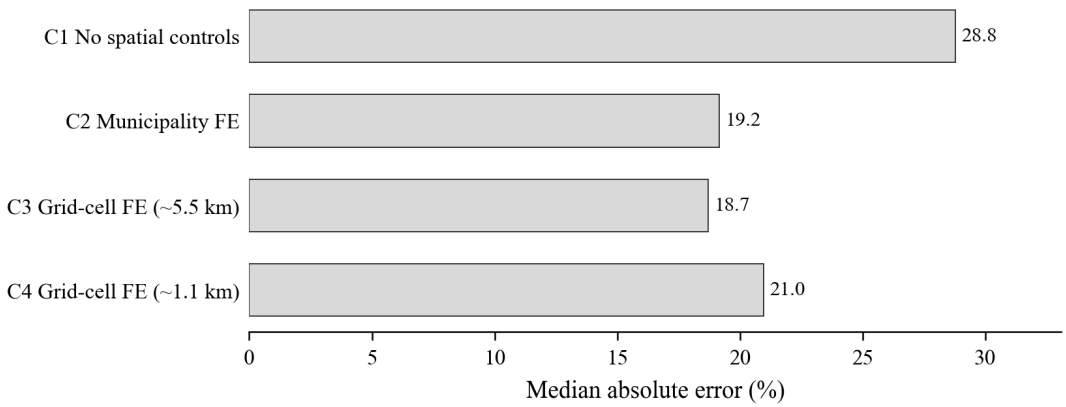


Figure 3. Axis C — spatial controls (quadratic semi-log backbone, quarter FE). MdAPE on the random holdout.

and month effects: an 18–25% reduction in the median error, squarely in the range reported by the automated-valuation literature (Mayer et al., 2019; Hong et al., 2020; Pérez-Rave et al., 2019). The boosting advantage over the spline hedonic, -3.2 points [-4.7, -1.9], is statistically unambiguous on every metric (DM $t = -3.6$ on the APE loss, -3.0 on the squared-log loss), and boosting edges the forest by a quarter point (-0.24 [-0.46, -0.05]). The decomposition rows identify the source. Boosting *without coordinates* collapses to 23.2%, worse than plain OLS with municipality dummies, so the learner's edge is overwhelmingly a flexible surface over space rather than exotic structure–attribute interactions, consistent with the residual-information results of Pace and Hayunga (2020). Pure proximity without structure fails too: the k-NN comparables rule (21.2%) is three and a half points worse than the spline hedonic. It is the *combination* of spatial flexibility with attribute adjustment that wins. The cross-validated ridge (17.7%) is a further lesson: on the identical design, shrinking the 1,115 municipality effects and the month effects toward zero improves on OLS by 1.7 points [1.1, 2.4] and ties the spline hedonic (difference $+0.06$ [-0.02, $+0.16$]). Regularising the thin fixed effects is as valuable as adding curvature; neither manufactures the spatial surface the trees learn.

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

Table 7 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 4.1 (17.6 $\rightarrow$ 21.7), the ridge 4.6 and the linear model in levels 2.1. The ranking compresses: the five best forward-split models, linear (20.4%), Box–Cox (21.0%), boosting (21.2%), forest (21.3%) and splines (21.7%), lie within 1.3 points and are statistically indistinguishable on

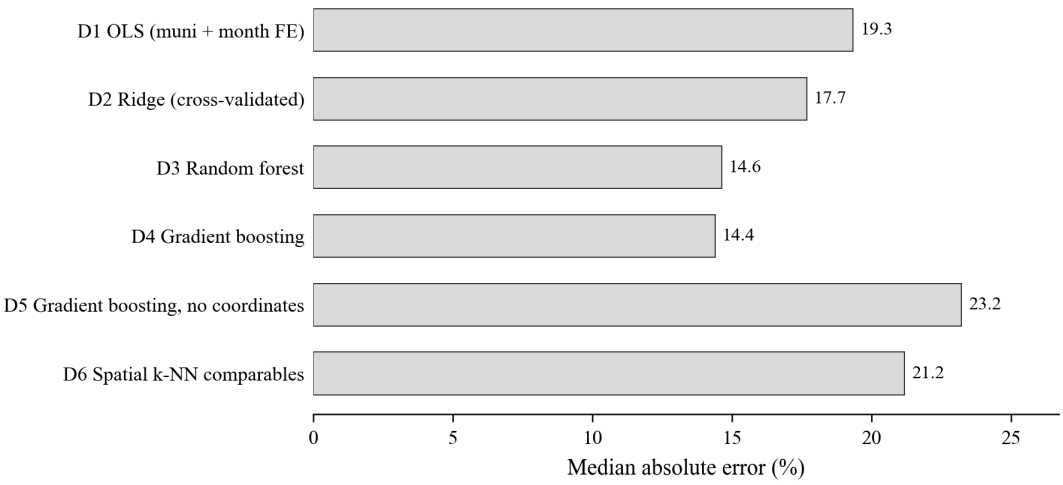


Figure 4. Axis D — estimation methods on the identical attribute set. MdAPE on the random holdout.

Table 7. Random versus forward-in-time evaluation

	Random holdout		Forward-in-time	
Model	MdAPE (%)	$R^2_{\ln}$	MdAPE (%)	$R^2_{\ln}$
A5 Semi-log + quadratics	19.2	0.546	23.3	0.442
B4 Month FE	19.3	0.543	22.9	0.445
C4 Grid-cell FE ($\sim$1.1 km)	21.0	0.496	25.3	0.364
D1 OLS (muni + month FE)	19.3	0.543	22.9	0.445
D2 Ridge (cross-validated)	17.7	0.587	22.3	0.480
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.

the median error—boosting against splines gives -0.51 [-1.68, $+0.55$], boosting against the linear model $+0.84$ [-1.19, $+2.59$]. The ensembles keep a clear lead on the mean error (MAPE 29.7% against 30.9% for splines and 32.9% for the linear model; DM $t = -3.5$ against the linear model) and on the log scale ($R^2_{\ln}$ 0.52 against 0.49 and 0.46). Version 1.0 of this paper reported that the ranking *reversed*; that reversal was an artefact of smearing the linear log models but not the tree ensembles (Section 7). The corrected statement is weaker but still consequential: the three-point advantage that random validation attributes to gradient boosting shrinks to half a point, and to insignificance, when the test set lies in the future.

Where does the gap come from? Three further cuts of the data answer the question (Table 8). First, holding the test period fixed: evaluated on the 31,465 sales of 2025–26 that happen to fall in the random test set, the random-split boosting model posts 13.7%, so the 2025–26 sales are not

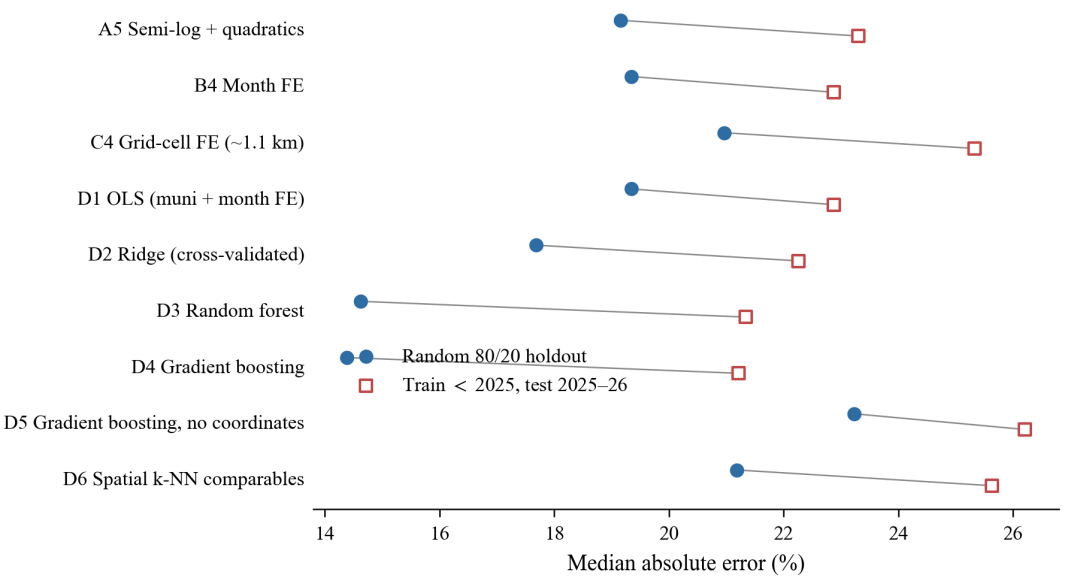


Figure 5. The generalization gap. MdAPE under the random holdout (filled circles) and the forward-in-time holdout (open squares); the grey segment joins the two evaluations of the same model.

intrinsically harder to price; the entire degradation to 21.2% comes from not having seen their period. The *staleness cost*, defined as the forward-split minus the same-period random-split MdAPE, is 7.5 points for boosting and 7.5 for the forest, against 4.7 for the spline hedonic, 5.2 for the ridge and 3.3 for the linear model. Second, the level decomposition of Section 4.5: the median log residual of every forward-split model is positive and large—+0.16 for the spline hedonic, +0.18 for the two ensembles, +0.10 for the linear model—so every frozen model under-prices the typical 2025–26 sale by 10–20%, the ensembles slightly more because their terminal leaf in the month counter averages the last few training months rather than the last one. Once each model's predictions are re-levelled by its own median ratio, the forward-split MdAPE falls to 13.9% for boosting and 16.8% for the spline hedonic: a gap of 2.9 points, essentially the 3.5-point gap measured under the random split with the same correction (13.0% against 16.5%). Third, the horizon profile (Table 9, Figure 6): in the first quarter after the training window boosting still leads the spline hedonic by 1.5 points (17.1% against 18.6%); the lead shrinks every quarter and vanishes by mid-2026, when both models post about 25%.

The picture is therefore not that the learners have memorised something that fails to transfer. Their *relative* pricing skill—which sales are dear and which are cheap within a period—transfers forward almost intact. What does not transfer is the *level*, and a frozen boosted model carries a slightly staler level than a frozen linear model with carry-forward month effects. The forward-in-time holdout thus quantifies two things that random cross-validation hides: a level staleness of 3–7 points of median error within eighteen months, common to every static model and largest for the trees, and the fact that the headline gap between learners and linear hedonics is a within-period

Table 8. Decomposing the generalization gap: same-period comparison and level error

	Random split		Forward split	Staleness	Level error	Re-levelled
Model	All test sales	2025–26 sales	2025–26 sales	cost	$\hat{b}$	MdAPE
A1 Linear	18.2	17.1	20.4	+3.3	+0.102	18.0
A5 Quadratics	19.2	18.5	23.3	+4.8	+0.171	18.3
A6 Splines	17.6	17.0	21.7	+4.7	+0.161	16.8
B4 Month FE	19.3	18.9	22.9	+4.0	+0.157	18.7
C3 Grid 5.5 km	18.7	18.3	24.3	+6.0	+0.182	18.8
C4 Grid 1.1 km	21.0	19.8	25.3	+5.5	+0.171	20.5
D2 Ridge	17.7	17.1	22.3	+5.2	+0.170	17.0
D3 Random forest	14.6	13.9	21.3	+7.5	+0.184	13.9
D4 Boosting	14.4	13.7	21.2	+7.5	+0.183	13.9
D5 Boosting, no coord.	23.2	21.0	26.2	+5.2	+0.188	21.3
D6 k-NN comparables	21.2	22.6	25.6	+3.1	+0.221	19.2
Test sales	102,929	31,465	158,388			

Notes: MdAPE in percent. Columns 2–3: models estimated on the random-split training sample, evaluated on all random test sales and on the 2025–26 subset of those sales. Column 4: models estimated on sales before 2025 and evaluated on all 2025–26 sales. Staleness cost: column 4 minus column 3. Column 5: median of $\ln P_i - \ln \hat{P}_i$ on the forward test set (common level error). Column 6: forward-split MdAPE after multiplying each model's predictions by $e^{\hat{b}}$ (Section 4.5).

Table 9. Forward-in-time accuracy by test quarter

Test quarter	n	A5 Quadratics	A6 Splines	D2 Ridge	D3 Random forest	D4 Boosting
2025Q1	17,031	20.2	18.6	19.0	17.2	17.1
2025Q2	34,546	21.1	19.5	20.1	19.2	19.1
2025Q3	28,097	22.8	21.3	21.8	20.6	20.6
2025Q4	24,156	23.8	22.2	22.7	21.7	21.6
2026Q1	15,611	24.8	23.1	23.6	22.8	22.7
2026Q2	30,950	25.8	24.5	25.1	24.4	24.3
2026Q3	7,997	26.0	24.9	25.2	25.0	24.9

Notes: MdAPE (%) of models trained on sales before January 2025, by quarter of the test sale. 2026Q3 covers July 2026 only.

statement. For models whose job is to price the future, random cross-validation is the wrong experiment (Bergmeir and Benítez, 2012; Mullainathan and Spiess, 2017); the right one separates the level, which an index or a retraining schedule must supply, from the relative accuracy the estimator contributes.

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

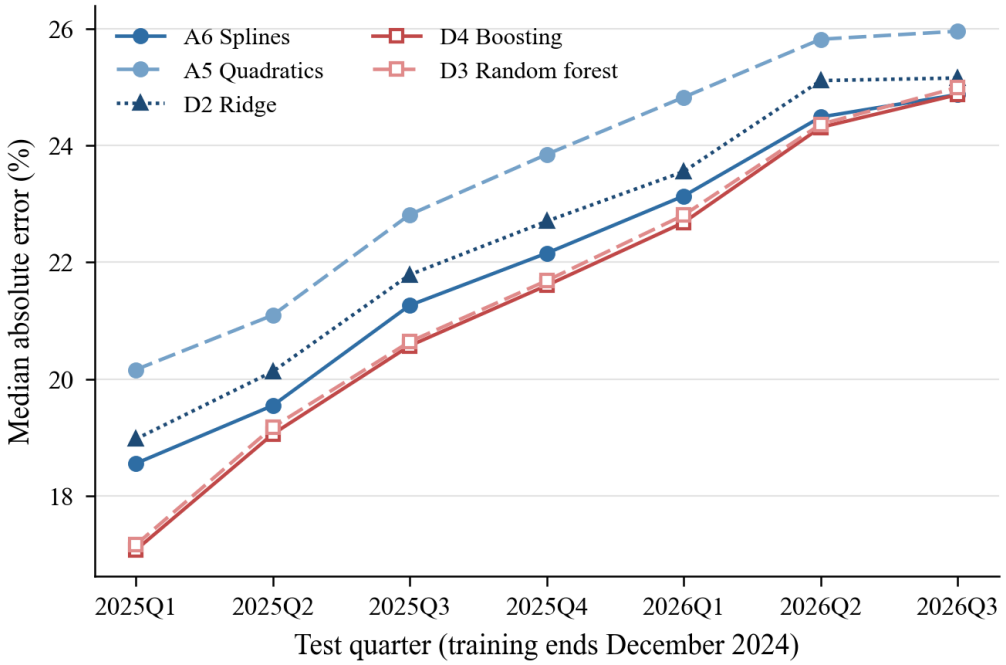


Figure 6. Forward-in-time accuracy by horizon. MdAPE by quarter of the test sale for models trained on sales before January 2025.

most often built to deliver, constant-quality price indices and implicit attribute prices, and then characterises where and when the learners' advantage materialises.

6.1. Price indices: form-robust, estimator-sensitive

Figure 7 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, a hedonic imputation index in the sense of Hill (2013) and Hill and Scholz (2018). All five track the same cycle: the 2021–22 boom, the 2022–23 correction and the 2024–26 recovery, over which the constant-quality level rises by about 70%. The four linear forms agree to within 2.5 index points at every month, so the choice that generates the most methodological discussion barely perturbs the index, consistent with the RPPI handbook's pragmatism about form (Eurostat, ILO, IMF, OECD, UNECE, World Bank, 2013; Silver, 2016). The estimator is a different matter: the boosting index rises less, ending in July 2026 at 160.5 against 168.6–171.0 for the linear forms, a gap that opens gradually from 2024 and reaches nine to eleven points. Two mechanisms contribute. The imputation index reprices a fixed portfolio, so its weights differ from the time-dummy index's implicit weights when appreciation is heterogeneous across segments; and the boosting model's month counter enters through tree splits that interact with location and attributes, so its "time effect" is an average of locally varying paths rather than a single common shift. Statistical agencies choosing between a time-dummy and an

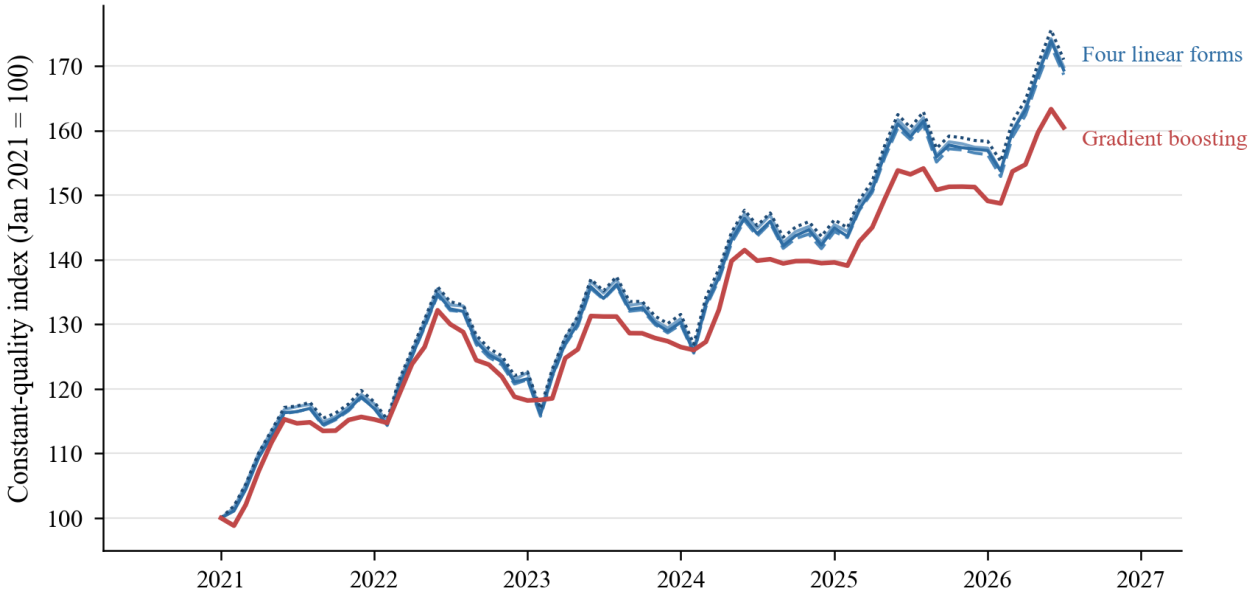


Figure 7. Constant-quality monthly price indices implied by competing specifications (January 2021 = 100). Linear forms: exponentiated month fixed effects. Gradient boosting: mean repriced value of a fixed 20,000-sale reference portfolio.

imputation index should expect this order of disagreement when the imputation model is a flexible learner; the agreement across linear forms is reassuring, the agreement across estimators is not automatic.

6.2. Implicit prices: the machine broadly agrees with the quadratic

Figure 8 compares the log-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; Lorenz et al., 2023), each normalised to its leftmost point. The two agree on the shape: steep initial depreciation that flattens with age, and strongly concave returns to floor space. On floor area the curves sit within 0.05 log points of each other over the whole 50–400 m^2 range. On age the boosting profile is uniformly flatter: it shows a small new-construction premium (the profile peaks at five years, +0.06 log points above a new building, where the quadratic can only decline), and from twenty years onward its cumulative depreciation is 0.10–0.13 log points smaller than the quadratic's (-0.20 against -0.31 at forty years, -0.28 against -0.41 at sixty), with a milder gradient for very old stock, a vintage effect the quadratic cannot bend to accommodate. These are differences of degree in a shared shape, not of sign or ordering. The boosting model's accuracy advantage therefore does not come from a different view of the main attribute gradients; it comes from higher-order interactions and, above all, from the spatial surface (Section 5). The interpretable quadratic remains a faithful, if slightly steeper, summary of the age–price relation even where it

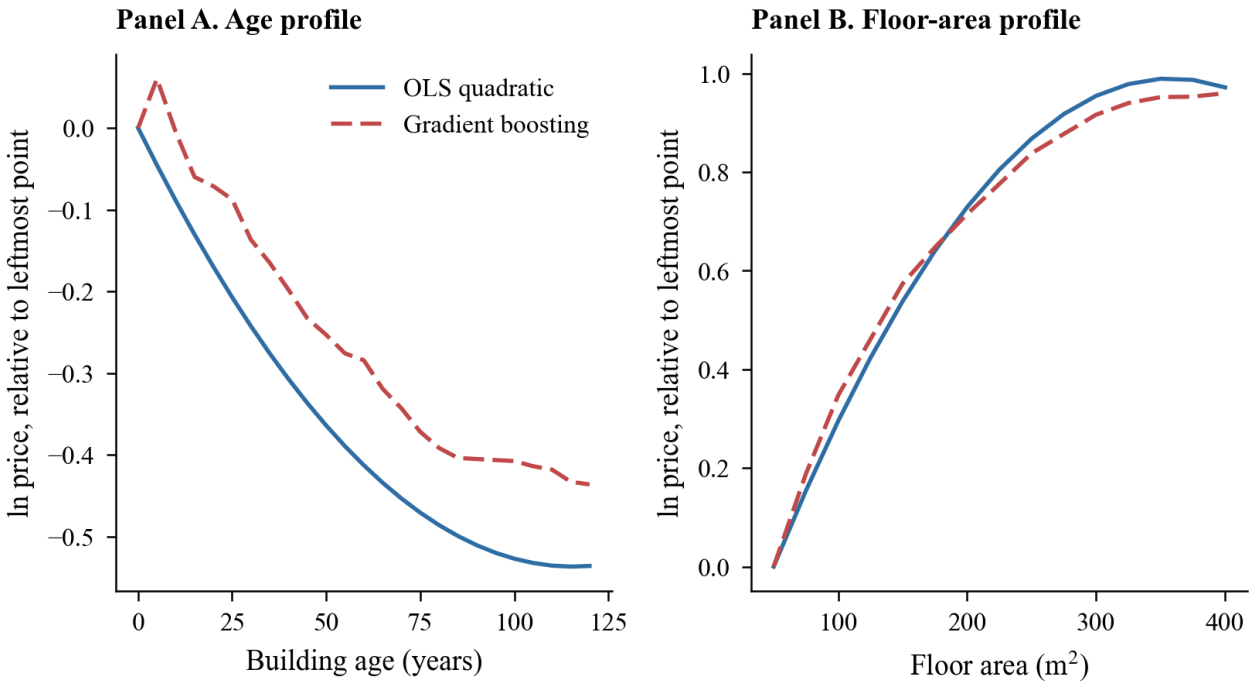


Figure 8. Implicit log-price profiles: quadratic OLS coefficients versus gradient-boosting partial dependence, each normalised to its leftmost point. Panel A: building age. Panel B: floor area.

loses the prediction race (Mullainathan and Spiess, 2017; Athey and Imbens, 2019).

6.3. Learning curves: data beat flexibility only at scale

Figure 9 and Table 10 trace accuracy against training size from 10,000 to 400,000 sales on a fixed random holdout. The OLS quadratic is flat from the very first step (19.1% at $n = 10{,}000$, 19.2% at $n = 400{,}000$): its bias floor binds immediately, and more data cannot lower it. The boosting model improves monotonically through the full sample (17.6% $\rightarrow$ 14.5%), so its advantage grows from 1.6 points at $n = 10{,}000$ to 4.7 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 learner's edge is modest, one reason thin-market automated valuation underperforms (Bogin and Shui, 2020), and a reason the provincial scale of the present data matters for the size of the gap we measure.

6.4. Where the machine wins: segments

Table 11 and Figure 10 decompose the random-holdout MdAPE by property class and municipality size, and the pattern locates the learner's edge: it is spatial resolution. The gap is largest for condominiums (9.1% against 18.2%) and in the largest cities (11.9% against 17.7%), segments where structure is homogeneous but *micro*-location varies greatly within a municipality, exactly the variation that municipality dummies cannot see and raw coordinates can. Once location is resolved,

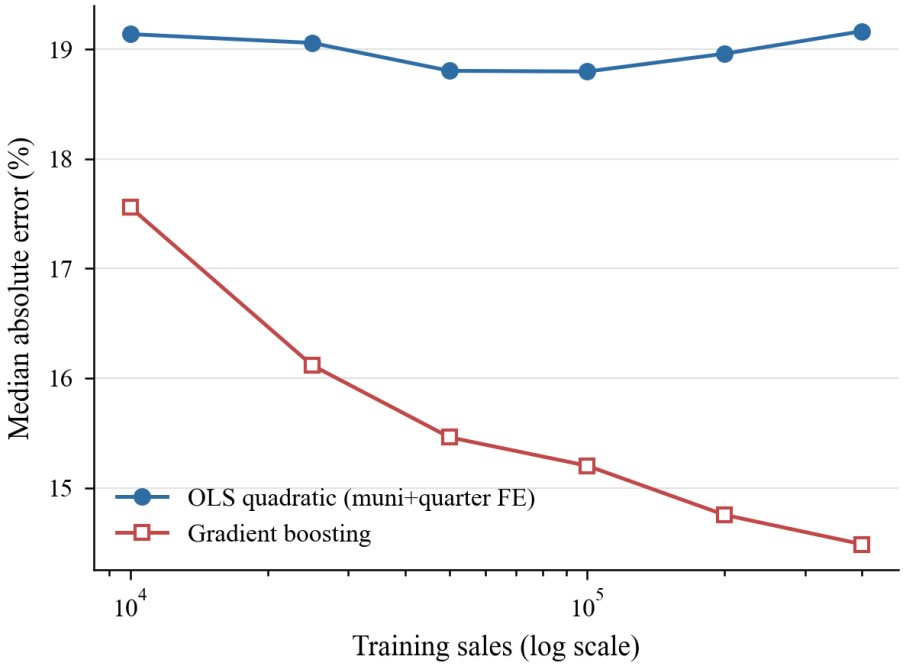


Figure 9. Learning curves: median absolute error against training-set size (random holdout fixed), OLS quadratic versus gradient boosting.

Table 10. Learning curves

Training sales	Gradient boosting	OLS quadratic (muni+quarter FE)
10,000	17.6	19.1
25,000	16.1	19.1
50,000	15.5	18.8
100,000	15.2	18.8
200,000	14.8	19.0
400,000	14.5	19.2

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

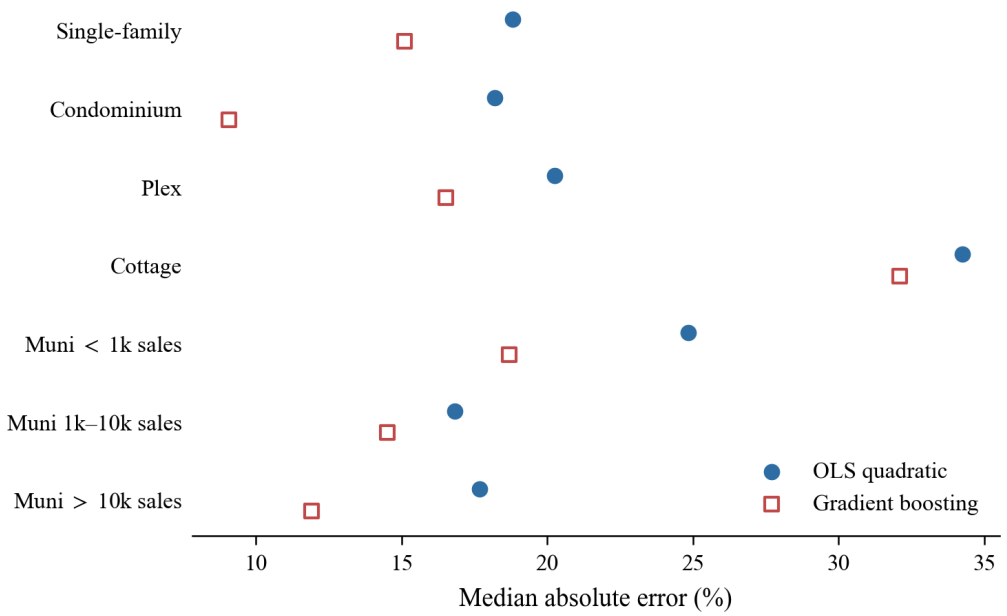


Figure 10. Median absolute error by market segment (random holdout): OLS quadratic versus gradient boosting.

a condominium is the most predictable object in the market. The gap is also large in the thinnest municipalities (18.7% against 24.8%), where a municipality effect is estimated from few sales and the learner can borrow strength from neighbouring municipalities through the coordinates. At the other extreme, cottages defeat both models (32.1% against 34.3%): idiosyncratic, waterfront-driven stock is hard for everyone, and a spatial surface cannot rescue attributes it does not observe. The complement with 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. Section 7 examines how much of the condominium advantage reflects genuine micro-location rather than recall of sales in the same building.

7. Robustness

This section examines the five choices in the design that a sceptical reader would most want varied: the rule that converts transformed-scale predictions into price levels, the leakage channels of the random holdout, the implementation of fine spatial effects, the resolution of the Box–Cox grid and the hyper-parameters of the boosting machine.

Table 11. Accuracy by market segment

Segment	n (test)	OLS quadratic	Gradient boosting
Condo	14,848	18.2	9.1
Cottage	2,109	34.3	32.1
Plex	15,115	20.2	16.5
Single family	69,835	18.8	15.1
Muni 1k–10k sales	37,419	16.8	14.5
Muni $<$ 1k sales	29,253	24.8	18.7
Muni $>$ 10k sales	36,257	17.7	11.9

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

7.1. The retransformation rule

Table 12 reports the scoreboard for the transformed-scale models under the median rule used throughout (eq. 11) and under Duan's mean rule (eq. 12). The smearing factors $\hat{s}$ lie between 1.06 and 1.07 for the least-squares models and between 1.03 and 1.05 for the ensembles, whose in-sample residuals are shrunk by overfitting. Two regularities stand out. First, the mean rule lowers the median error of *every* log-scale model under the random split, by about one point for the linear models and 1.3 points for the ensembles. This is not what a conditional-mean correction should do to a median-based metric; it happens because the log residuals are left-skewed (skewness -1.4 to -1.8), so that their median is $+0.06$ while their mean is zero, and $\exp(\hat{z})$ under-prices the typical sale by about 6%—which is, by coincidence, the size of the smearing factor. Second, under the forward-in-time split the mean rule lowers median errors by 2.5–3 points for all models, because the market rose after the training window and any upward factor offsets part of the staleness. Neither effect changes a ranking when the rule is applied uniformly: boosting leads the spline hedonic by 3.4 points under the random split and by 0.7 points under the forward split with the mean rule, against 3.2 and 0.5 with the median rule. What does change a ranking is applying the rule to one family only. Version 1.0 of this paper smeared the linear log models and not the ensembles; the result was a forward-split comparison of 19.1% for the smeared spline hedonic against 21.2% for the unsmeared boosting model, reported as a reversal. The comparable numbers are 21.7% against 21.2% (median rule) or 19.1% against 18.4% (mean rule). The log-scale metrics are affected in the same direction: smearing adds $\ln \hat{s} \approx 0.06$ to every log prediction, which inflates $\text{RMSE}_{\ln}$ and depresses $R^2_{\ln}$ for the smeared family only, so that the "identical $R^2_{\ln}$" of the first version (0.52 for both) becomes 0.49 against 0.52 in favour of boosting under the uniform rule.

Table 12. Median rule versus Duan smearing

		Random holdout				Forward-in-time holdout			
		MdAPE (%)		$R^2_{\ln}$		MdAPE (%)		$R^2_{\ln}$	
Model	$\hat{s}$	Median	Mean	Median	Mean	Median	Mean	Median	Mean
A2 Semi-log	1.070	19.0	18.2	0.546	0.533	23.2	20.6	0.426	0.458
A3 Log-log	1.064	18.0	17.0	0.580	0.569	21.8	19.2	0.481	0.513
A4 Box–Cox	—	18.6	17.9	0.548	0.536	21.0	19.6	0.476	0.490
A5 Quadratics	1.070	19.2	18.1	0.546	0.533	23.3	20.5	0.442	0.479
A6 Splines	1.062	17.6	16.5	0.590	0.579	21.7	19.1	0.491	0.522
D2 Ridge	1.063	17.7	16.7	0.587	0.577	22.3	19.5	0.480	0.515
D3 Random forest	1.032	14.6	13.7	0.631	0.629	21.3	19.4	0.519	0.544
D4 Boosting	1.048	14.4	13.1	0.642	0.636	21.2	18.4	0.521	0.556

Notes: Median: transformed-scale prediction inverted as is, eq. (11) (main scoreboard). Mean: Duan smearing, eq. (12); $\hat{s}$ is the smearing factor estimated on training residuals (random split), undefined for the Box–Cox model, whose mean prediction averages the inverse transformation over 1,000 training residuals. $R^2_{\ln}$ under the mean rule is computed on the smeared log prediction.

7.2. Leakage-free holdouts

Table 13 re-runs the full race under the two grouped holdouts of Section 4.4, and Table 14 decomposes the random holdout by whether the test sale's parcel, or its exact coordinates, also appear in training. Repeat sales are irrelevant: the parcel-grouped race reproduces the random race to within 0.1–0.4 points for every model, and within the random holdout the boosting model prices test sales whose parcel sold before in training (14.5%) no better than the others (14.4%). Shared coordinates matter more, but for a specific reason. Test sales that share coordinates with a training sale are priced at 12.0% by boosting against 15.3% for the rest, and at 16.7% against 22.9% by the k-NN rule, whose nearest neighbour is then literally the same building; the linear models are unaffected (17.8% against 17.6% for the splines). Condominiums account for most of these co-located sales, and the boosting model prices them at 8.9% when their building is in training and 10.2% when it is not, against 18.0% and 20.3% for the quadratic OLS. The condominium advantage of Section 6 is therefore genuine micro-location, with a modest recall component. The building-grouped race, in which no coordinate pair straddles the split, confirms it: boosting posts 14.6% against 17.8% for the spline hedonic, a 3.2-point gap identical to the random race, and the only model that loses materially is the k-NN rule (22.3% against 21.2%).

7.3. Nested spatial fixed effects

Section 5 attributes the poor showing of the ∼1.1 km grid to thin cells and to the arbitrary pricing of unseen cells. Table 15 adds two specifications in which the grid effects enter *together with* municipality effects. Because the two sets of dummies are nearly collinear, the vanishing-penalty ridge returns the minimum-norm least-squares solution, which assigns the component common to a

Table 13. The race under random, parcel-grouped and building-grouped holdouts

	Random 80/20		Parcel-grouped		Building-grouped	
Model	MdAPE	$R^2_{\ln}$	MdAPE	$R^2_{\ln}$	MdAPE	$R^2_{\ln}$
A. Functional form						
A1 Linear	18.2	0.477	18.4	0.469	18.4	0.473
A2 Semi-log	19.0	0.546	19.4	0.528	19.5	0.530
A3 Log-log	18.0	0.580	18.0	0.572	18.1	0.573
A4 Box–Cox	18.6	0.548	18.2	0.534	18.3	0.535
A5 Quadratics	19.2	0.546	19.2	0.539	19.0	0.552
A6 Splines	17.6	0.590	17.6	0.581	17.8	0.582
B. Time effects						
B1 No time FE	22.0	0.486	21.6	0.489	22.0	0.482
B2 Year FE	19.1	0.545	19.0	0.545	19.2	0.539
B3 Quarter FE	19.2	0.546	19.2	0.539	19.0	0.552
B4 Month FE	19.3	0.543	19.3	0.537	19.4	0.538
C. Spatial controls						
C1 No spatial FE	28.8	0.307	28.9	0.301	28.7	0.306
C2 Municipality FE	19.2	0.546	19.2	0.539	19.0	0.552
C3 Grid 5.5 km	18.7	0.557	18.7	0.550	18.8	0.551
C4 Grid 1.1 km	21.0	0.496	21.1	0.491	20.2	0.510
D. Estimation method						
D1 OLS	19.3	0.543	19.3	0.537	19.4	0.538
D2 Ridge	17.7	0.587	17.7	0.580	17.9	0.580
D3 Random forest	14.6	0.631	14.6	0.620	14.9	0.621
D4 Boosting	14.4	0.642	14.5	0.632	14.6	0.634
D5 Boosting, no coord.	23.2	0.426	23.4	0.417	23.3	0.420
D6 k-NN comparables	21.2	0.434	21.3	0.442	22.3	0.425
Test sales	102,929		102,843		102,949	

Notes: MdAPE in percent. Parcel-grouped: 20% of assessment-roll parcels assigned to the test set, so that repeat sales of a property never straddle the split. Building-grouped: 20% of exact coordinate pairs assigned to the test set, so that units geocoded to the same building or complex never straddle the split.

Table 14. Random holdout: accuracy on test sales seen and unseen through their parcel or coordinates

	Same parcel in training		Same coordinates in training	
Model	Yes	No	Yes	No
All test sales				
A5 Quadratics	19.8	19.1	19.2	19.2
A6 Splines	18.3	17.5	17.8	17.6
D3 Random forest	14.6	14.6	11.7	15.7
D4 Boosting	14.5	14.4	12.0	15.3
D5 Boosting, no coord.	22.9	23.3	20.0	24.5
D6 k-NN comparables	22.2	21.0	16.7	22.9
Test sales	17,219	85,710	27,475	75,454
Condominiums only				
A5 Quadratics	17.4	18.4	18.0	20.3
D4 Boosting	9.0	9.1	8.9	10.2
Test sales	3,374	11,474	13,261	1,587

Notes: MdAPE (%) on the random holdout. "Same parcel in training": the parcel of the test sale sold at least once in the training sample. "Same coordinates in training": at least one training sale shares the exact geocoded coordinates of the test sale.

municipality to the municipality effect and only the deviation to the cell effect—a crude form of partial pooling—and a test sale in an unseen cell inherits its municipality's level. The effect is large. Nested within municipalities, the 5.5 km grid improves from 18.7% to 17.7% and the 1.1 km grid from 21.0% to 17.3%, the lowest median error of any linear specification in the paper, below the spline hedonic (17.6%) with the same R^2_{ln} (0.59). Under the forward-in-time split the nested fine grid posts 21.6% against 23.3% for municipality effects alone and 25.3% for the fine grid alone. The conclusion for axis C is therefore sharper than "finer overfits": fine spatial effects are valuable when they are anchored to a coarser level that supplies the prior for thin and unseen cells, and harmful when they are not. A hierarchical or kriged location effect (Case et al., 2004; Hill and Scholz, 2018) would push the linear frontier further; we leave its estimation to future work.

7.4. A finer Box–Cox grid

The main grid steps λ by 0.25. Refining it to steps of 0.05 over $[-0.25, 0.75]$ moves the profile-likelihood maximum to $\hat{\lambda} = 0.35$, with a profile that is flat between 0.25 and 0.55 (differences of under one thousandth of the log-likelihood, within the tolerance of the sparse iterative solver) and falls steeply on either side. The Box–Cox model at $\hat{\lambda} = 0.35$ posts a random-holdout MdAPE of 18.4% against 18.6% at 0.25: the exponent is imprecisely identified in a range that excludes both the linear and the logarithmic case, and its exact value is immaterial for prediction.

Table 15. Nested spatial fixed effects

	Random holdout		Forward-in-time	
Spatial controls	MdAPE (%)	$R^2_{\ln}$	MdAPE (%)	$R^2_{\ln}$
C2 Municipality FE (1,115 cells)	19.2	0.546	23.3	0.442
C3 Grid 5.5 km (5,255 cells)	18.7	0.557	24.3	0.423
C4 Grid 1.1 km (30,548 cells)	21.0	0.496	25.3	0.364
C3′ Municipality + grid 5.5 km	17.7	0.582	21.5	0.481
C4′ Municipality + grid 1.1 km	17.3	0.589	21.6	0.486

Notes: Quadratic semi-log backbone with quarter fixed effects. C3′ and C4′ add the grid effects to the municipality effects; with a vanishing ridge penalty the estimator is the minimum-norm least-squares solution and unseen cells fall back on their municipality.

Table 16. Gradient-boosting hyper-parameter sensitivity

	Random holdout		Forward-in-time	
Gradient-boosting configuration	MdAPE (%)	$R^2_{\ln}$	MdAPE (%)	$R^2_{\ln}$
Gradient boosting (baseline)	14.4	0.642	21.2	0.521
Shallow (15 leaves, lr 0.05, 1500 iter)	14.7	0.636	21.4	0.521
Deep (255 leaves, lr 0.05, 600 iter)	14.3	0.645	21.6	0.518
Regularised (min leaf 200, L2 = 1)	14.3	0.643	21.5	0.522
Early stopping on last 10% of training months [700 / 1000 it.]	14.5	0.641	21.2	0.524

Notes: All configurations use the D4 feature set (attributes, coordinates, month counter). Baseline: 600 iterations, learning rate 0.08, 63 leaves, minimum leaf 20, $\ell_2 = 0.01$. Early stopping: iterations added in blocks of 100 (learning rate 0.05) until the squared log error on the last 10% of training sales in date order stops improving; the selected number of iterations is shown in brackets for the random and forward splits.

7.5. Boosting hyper-parameters

Table 16 re-estimates the boosting model under four materially different configurations: shallow trees with more iterations, deep trees, heavy regularisation (minimum leaf 200, $\ell_2 = 1$), and a configuration whose number of iterations is chosen by early stopping on the last 10% of training sales in date order, the time-honest tuning rule a careful practitioner would use for a forward-looking model. Under the random holdout the four alternatives span 14.3–14.7% MdAPE against 14.4% for the baseline, and under the forward-in-time holdout 21.2–21.6% against 21.2%; $R^2_{\ln}$ moves by at most 0.006. The time-honest early-stopping rule selects 700 iterations on the random training sample and 1,000 on the pre-2025 sample and reproduces the baseline to within 0.1 point under both splits. The conclusions of Sections 5 and 5.5 do not depend on the particular configuration we fixed a priori, and tuning on a time-ordered validation set does not remove the level staleness that dominates the forward-split error, because no choice of hyper-parameters lets a static tree extrapolate a price level it has not seen.

8. Discussion

8.1. A hierarchy of specification choices

Read jointly, the four axes imply a hierarchy of returns to specification effort. The first and largest return is to location: any spatial absorption beats none by about ten points of median error, a margin that dwarfs every other choice, and the granularity question is one of implementation rather than of resolution. A fine grid used alone overfits because thin cells are estimated from a handful of sales and unseen cells are priced arbitrarily; the same grid nested within municipalities, which pools thin cells toward their municipality, is the most accurate linear specification we estimate. The second return is to time controls: in a moving market a model without time effects mistakes appreciation for attributes, at a cost of three points under random validation and six when the level must be carried forward, while granularity beyond quarters is cosmetic. The third return is to the estimator, and its size depends on the question asked. Under random validation gradient boosting buys three points over the best linear model, a gain that is statistically unambiguous and that grows with the training sample; in deployment the gain on the median error shrinks to half a point and loses significance, because the ensembles carry a slightly staler price level, while their relative pricing skill is intact. The fourth and smallest return is to functional form: splines are a cheap and reliable upgrade over the semi-log, and the Box–Cox machinery is not worth its complexity, exactly as Cropper et al. (1988) concluded from simulations four decades ago.

8.2. Why the forward-in-time split compresses the ranking

The decomposition of Section 5.5 clarifies what the forward-in-time holdout measures. The boosting model's random-split advantage comes almost entirely from a flexible spatial surface (removing the coordinates costs it 8.8 points), and that surface transfers to 2025–26 sales: re-levelled, the model prices them with a 13.9% median error, against 16.8% for the spline hedonic. What does not transfer is the aggregate level. Trees are piecewise constant in the month counter and the terminal leaf in time averages the last training months rather than the last one, so a frozen boosted model enters the test period with a slightly staler level than a linear model whose last month effect is carried forward; by mid-2026 both are about 25% below the market and indistinguishable. The lesson is not that machine learning fails out of sample. It is that the forward-in-time holdout conflates two things—the common level error that any static model incurs, and the relative accuracy the estimator contributes—and that the first dominates the second within a few quarters. Retrained monthly, or re-levelled with a published index as practitioners do (Thibodeau, 2003; Mayer et al., 2019), the boosted model would presumably keep its interpolation advantage; the forward split prices the *retraining or re-levelling requirement* that random cross-validation hides, and it matches the

practitioner evidence that automated-valuation accuracy decays quickly outside the sample period (Bogin and Shui, 2020; Kok et al., 2017). It also explains why our first version, which smeared the linear models but not the ensembles, found a reversal: in a rising market an upward retransformation factor applied to one family only offsets part of that family's staleness.

8.3. Implications

For research. Hedonic indices are robust to functional form and implicit-price gradients are robust to the estimator in shape, if not in slope (Section 6). Researchers using hedonic models to *measure*—indices, implicit prices, capitalization effects—can keep simple forms with good, preferably hierarchical, spatial controls. When prediction claims are made, they should report a forward-in-time split alongside random cross-validation, and they should report intervals: several of the differences that organise the applied debate—a 5.5 km grid against municipality effects, month against quarter effects, boosting against splines in deployment—are not distinguishable from zero on half a million sales.

For mass appraisal. Quebec's assessors face exactly this design problem every three years. Our results suggest that the largest accuracy gains lie in spatial resolution implemented hierarchically, that flexible learners add a further three points within the period they are estimated on, and that any model, linear or boosted, frozen at a reference date under-prices the typical sale by 10–20% within eighteen months in a market like 2025–26. 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 order as the inequities measured there, and it is a level problem that an index adjustment addresses before any change of estimator.

For the automated-valuation literature. Every comparison should report a forward-in-time split, a leakage-free grouped split when properties sell repeatedly or share coordinates, and a single retransformation rule applied to all models (Steurer et al., 2021). On our data the first changes the size and significance of the headline gap, the second changes nothing, and the third, applied asymmetrically, is enough to manufacture a reversal.

8.4. Limitations

Our attribute set, though assessor-grade, omits listing-level quality (renovations, finishes, views); richer features would likely raise the learners' within-period advantage. Hyper-parameters were deliberately conventional, and Section 7 shows that materially different boosting configurations, including one tuned on a time-ordered validation set, leave the ranking unchanged; a spatially smarter linear model (kriged residuals, Case et al., 2004; semiparametric location surfaces, Clapp,

2003) would narrow the gap from the other side and is not estimated here. The forward split covers one specific regime (the 2025–26 recovery), in which prices rose by about 15% in eighteen months; a flat or falling period would produce a smaller level error and possibly a different ordering of staleness across model families. The grouped holdouts remove leakage through repeat sales and shared coordinates but not through neighbouring parcels in the same street, which a spatially blocked holdout would address at the cost of changing the question to spatial extrapolation. Finally, Quebec's institutional homogeneity aids transferability of results across municipalities but may limit it across countries.

9. 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, and a bootstrap interval on every comparison. 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 to absorb space and time—move it by three to ten, and the way fine spatial effects are implemented matters more than their resolution. Machine-learning estimators dominate the standard scoreboard by three points, a gain that is real, grows with the sample and comes almost entirely from a flexible surface over space. When the test set lies in the future, as it always does in deployment, that gain shrinks to half a point and loses significance, not because the learned surface fails to transfer—re-levelled, it prices future sales as well as it prices contemporaneous ones—but because every frozen model carries a stale price level, and a boosted model carries a slightly staler one than a linear model with carry-forward time effects.

The objects economists harvest from hedonic models prove stable along the axis that generates the most debate and less so along the axis that generates the least: constant-quality indices agree across functional forms to within 2.5 points over a full boom–bust–recovery cycle but differ by nine points between a time-dummy index and a boosted imputation index, and the machine's implicit age and area gradients share the shape of the quadratic OLS with a milder depreciation path. The practical synthesis is simple. A semi-log with splines, quarter effects and spatial fixed effects nested from municipality down to a fine grid is within three points of the frontier under random validation and indistinguishable from it in deployment—transparent, auditable and fast. Where the extra within-period accuracy of learning methods is worth a retraining cadence—high-frequency automated valuation, collateral monitoring—our decomposition says to spend the flexibility budget on the spatial surface, to supply the price level from an index or a frequent refit rather than from the model, 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.

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.

Bergmeir, Christoph and José M. Benítez, "On the Use of Cross-Validation for Time Series Predictor Evaluation," *Information Sciences*, 2012, *191*, 192–213.

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.

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), 139–159.

_ , Martin Hoesli, and Vincent S. Peng, "Do Housing Submarkets Really Matter?," *Journal of Housing Economics*, 2003, *12* (1), 12–28.

Box, George E. P. and David R. Cox, "An Analysis of Transformations," *Journal of the Royal Statistical Society, Series B*, 1964, *26* (2), 211–243.

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., "A Semiparametric Method for Valuing Residential Locations: Application to Automated Valuation," *Journal of Real Estate Finance and Economics*, 2003, *27* (3), 303–320.

_ **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.

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

Diewert, W. Erwin and Chihiro Shimizu, "Hedonic Regression Models for Tokyo Condominium Sales," *Regional Science and Urban Economics*, 2016, *60*, 300–315.

_ **, Jan de Haan, and Rens Hendriks**, "Hedonic Regressions and the Decomposition of a House Price Index into Land and Structure Components," *Econometric Reviews*, 2015, *34* (1–2), 106–126.

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

Dubé, Jean and Dièg̀o Legros, "Dealing with Spatial Data Pooled over Time in Statistical Models," *Letters in Spatial and Resource Sciences*, 2013, *6* (1), 1–18.

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.

Fik, Timothy J., David C. Ling, and Gordon F. Mulligan, "Modeling Spatial Variation in Housing Prices: A Variable Interaction Approach," *Real Estate Economics*, 2003, *31* (4), 623–646.

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.

Giacomini, Raffaella and Halbert White, "Tests of Conditional Predictive Ability," *Econometrica*, 2006, *74* (6), 1545–1578.

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.

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.

_ **and Michael Scholz**, "Can Geospatial Data Improve House Price Indexes? A Hedonic Imputation Approach with Splines," *Review of Income and Wealth*, 2018, *64* (4), 737–756.

_ **, Daniel Melser, and Iqbal Syed**, "Measuring a Boom and Bust: The Sydney Housing Market 2001–2006," *Journal of Housing Economics*, 2009, *18* (3), 193–205.

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," in I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Curran Associates 2017, pp. 3146–3154.

Khoshnoud, Mahsa, G. Stacy Sirmans, and Emily N. Zietz, "The Evolution of Hedonic Pricing Models," *Journal of Real Estate Literature*, 2023, *31* (1), 1–47.

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.

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.

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.

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.

Manning, Willard G. and John Mullahy, "Estimating Log Models: To Transform or Not to Transform?," *Journal of Health Economics*, 2001, *20* (4), 461–494.

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.

McCluskey, William J., Michael McCord, Peadar T. Davis, Martin Haran, and David McIlhatton, "Prediction Accuracy in Mass Appraisal: A Comparison of Modern Approaches," *Journal of Property Research*, 2013, *30* (4), 239–265.

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

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.

_ **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.

_ **, Ronald Barry, John M. Clapp, and Mauricio Rodriquez**, "Spatiotemporal Autoregressive Models of Neighborhood Effects," *Journal of Real Estate Finance and Economics*, 1998, *17* (1), 15–33.

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.

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.

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.

Schulz, Rainer, Martin Wersing, and Axel Werwatz, "Automated Valuation Modelling: A Specification Exercise," *Journal of Property Research*, 2014, *31* (2), 131–153.

Silver, Mick, "How to Better Measure Hedonic Residential Property Price Indexes," IMF Working Paper 16/213, International Monetary Fund, Washington, DC 2016.

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.

Thibodeau, Thomas G., "Marking Single-Family Property Values to Market," *Real Estate Economics*, 2003, *31* (1), 1–22.

Zurada, Jozef, Alan S. Levitan, and Jian Guan, "A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context," *Journal of Real Estate Research*, 2011, *33* (3), 349–388.