UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

DÉPARTEMENT DES SCIENCES ADMINISTRATIVES

WORKING PAPER NO. 16

Does the Sales Comparison Approach Actually Work? A Large-Scale Point-in-Time Validation of Comparable Selection, Price Adjustments, and Value Reconciliation

Evidence from 391,000 Single-Family Transactions in Quebec, 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

The sales comparison approach is the dominant method of residential valuation, yet its operating rules—how many comparables to use, how far and how far back to search, which adjustments to apply, how to weight and reconcile—rest largely on professional convention. We subject these rules to a population-scale, point-in-time backtest on 390,996 arm's-length single-family transactions in Quebec (2021–2026) matched to the municipal assessment roll. For each of 189,354 target sales in 2024–2026 the price is masked and 2,256 configurations of the approach, together with hedonic, gradient-boosting, assessment-ratio and hybrid benchmarks, are applied using only information available on the valuation date; configurations are selected on 2024 and evaluated out of sample on 2025–2026, and paired differences are tested with a cluster bootstrap over municipality–quarter cells. Adjustments dominate: the median absolute error of the five nearest sales falls from 18.1% with a time adjustment only to 13.0% with a hedonic grid, and a stylised conventional grid captures most of that gain. Selecting physically similar rather than merely nearby comparables delivers the same gain without any adjustment. The optimal number of comparables is about twenty once a grid is applied, with no penalty up to fifty; short windows are dominated and the search radius matters little. Weighting and reconciliation rules matter little. The best transparent configuration chosen in 2024 reaches 11.1% out of sample, against 9.0% for a quarterly-retrained gradient-boosting model using the assessed value; the cost of transparency, 2.2 points, is precisely estimated and narrows in homogeneous submarkets, and the chosen rule ranks second among 1,370 configurations out of sample. Accuracy is governed primarily by submarket liquidity, and a wide-net rule retaining the most similar sales adapts to it on its own.

Keywords: sales comparison approach, comparable selection, adjustment grid, automated valuation models, hedonic pricing, backtesting, residential real estate, Quebec

JEL Classification: R31, R32, C52, C53, G12

1. Introduction

The sales comparison approach is the workhorse of residential property valuation. Appraisers, assessors, lenders and courts rely on it daily: a small set of recent sales of similar properties is selected, each sale price is adjusted for the differences between the comparable and the subject, and the adjusted prices are reconciled into a single opinion of value (Appraisal Institute, 2020; International Association of Assessing Officers, 2017). The approach is taught in every valuation curriculum and codified in professional standards, and its logic—that the market itself reveals what buyers pay for a bundle of characteristics—is compelling. Yet the *operating rules* that turn this logic into a number are, to a remarkable degree, matters of convention rather than evidence. How many comparables should be used? Three is the traditional answer; some standards say "at least three." How far should one search, and how far back in time? Which adjustments are warranted, and of what magnitude? Should comparables be weighted equally, by proximity, by recency, or by similarity? Should adjusted prices be averaged, or should the appraiser "reconcile" them by judgement? Each of these questions has been debated for four decades (Colwell et al., 1983; Vandell, 1991; Gau et al., 1992; Pace, 1996; Isakson, 2002; Lai et al., 2008), but the empirical evidence remains thin, dated, and almost always drawn from a few hundred sales in a single metropolitan area.

This paper provides a population-scale, out-of-sample validation of the sales comparison approach. We exploit a research database—Vrai Prix—that links the universe of registered residential transactions in Quebec to the municipal assessment roll, giving us 390,996 arm's-length single-family sales from January 2021 to September 2026 with harmonised physical characteristics, coordinates and assessed values. On this base we build a *point-in-time backtesting* engine. For each of 189,354 target sales occurring in 2024, 2025 and 2026, the sale price is masked and the property is valued as an appraiser would have had to value it on the day of the sale: only transactions closed strictly before that date are admissible as comparables; the price index used to time-adjust them is re-estimated every month from prior sales only; the adjustment grids and machine-learning benchmarks are re-estimated every quarter from prior sales only. The masked price is then compared with the estimate. Because every configuration of the method is applied to exactly the same targets, differences in accuracy are paired differences of forecast losses in the sense of Diebold and Mariano (1995), and we quantify their uncertainty with a cluster bootstrap over municipality–quarter cells (Cameron et al., 2008) that respects the spatial and temporal dependence of valuation errors. Because configurations are chosen on 2024 sales and evaluated on 2025–2026 sales, the headline figures are protected against the data-snooping bias that a search over thousands of specifications would otherwise create (White, 2000; Hansen, 2005); we verify directly that the configuration chosen on 2024 ranks 2nd among the 1,370 eligible configurations in 2025–2026.

Within this engine we run 2,256 configurations of the sales comparison approach. A first experiment maps the *error surface* $\mathrm{Error}(k,d,t)$ over the number of comparables $k \in \{1,\ldots,50\}$, the search radius $d \in \{0.5,\ldots,25\}$ km and the look-back window $t \in \{3,\ldots,24\}$ months, for three selection rules (nearest, most physically similar, and a compromise) and two adjustment regimes (time only, hedonic grid). A second experiment is a full factorial design over the *rules* of the method at three anchor configurations: four selection criteria, eight adjustment schemes (none; time only; a stylised conventional grid; a global hedonic grid; a regional grid; a locally re-estimated grid; non-linear adjustments from a gradient-boosting model; and the assessment-ratio adjustment), five weighting schemes, and two reconciliation rules. Sixteen named models, from "median of the three nearest sales" to a quarterly-retrained gradient-boosting model and a hybrid, are then placed in a horse race, and five hypotheses about the trade-offs between proximity, similarity, sample size, local estimation, weighting and market liquidity are tested directly.

Five results stand out. First, *adjustments dominate*. Averaging the five nearest time-adjusted sales yields a median absolute percentage error (MdAPE) of 18.1%; applying a hedonic adjustment grid to the same five sales cuts it to 13.0%, and a stylised conventional grid—half the comparable's price per square metre for living area, half the implied land value per square metre for lot size, half a percent per year of age—captures most of that gain (13.4%). What matters is *that* one adjusts, far more than the sophistication of the grid: regional, local and non-linear refinements add little, and the locally re-estimated grid is less accurate where sales are sparse. Second, *similarity beats proximity*. With no physical adjustment at all, selecting the five most similar sales within a wide radius rather than the five nearest yields 13.0%, essentially the same accuracy as adjusting the nearest ones. A nearby but dissimilar sale is a worse comparable than a more distant but similar one. Third, *three comparables is too few*. Without physical adjustment, accuracy improves up to about ten comparables and then deteriorates as more distant sales are added; with a grid it improves up to roughly 20 and stays flat to fifty, in sparse and dense markets alike. Three-month windows are dominated everywhere, whereas the search radius, once enough similar sales are retained, matters little. Fourth, *weighting and reconciliation matter little*: the choice between median and mean, or between equal, distance, recency and similarity weights, moves the MdAPE by a few tenths of a point, an order of magnitude less than selection and adjustment. Fifth, *the cost of transparency is small*. The best configuration of the method chosen on 2024 data reaches 11.1% out of sample in 2025–2026, against 9.0% for a gradient-boosting model retrained every quarter with access to the assessed value, and 8.9% for a hybrid that combines the two; the paired difference of 2.2 points is precisely estimated, and it falls to about a point and a half in homogeneous submarkets. Accuracy is, above all, a function of submarket liquidity: the same method is two and a half times more accurate where hundreds of sales occur within two kilometres in a year than where fewer than five do. A wide-net rule that retains the most similar sales within a large radius adapts to this on its own, and

rules tuned stratum by stratum do not improve on it.

The paper contributes to three literatures. It extends the theoretical and small-sample studies of comparable selection and weighting (Colwell et al., 1983; Vandell, 1991; Gau et al., 1992; Green, 1994; Pace, 1996; Epley, 1997; Lai et al., 2008) to a population-scale, out-of-sample, point-in-time setting, and resolves several of their open questions with confidence intervals rather than illustrations. It speaks to the literature comparing hedonic and grid-adjustment techniques (Kang and Reichert, 1991; Lipscomb and Gray, 1990; Pace and Gilley, 1998) and to the comparisons of machine learning with traditional valuation (McCluskey et al., 2013; Mayer et al., 2019; Bogin and Shui, 2020; Pace and Hayunga, 2020), by holding the information set fixed across methods and measuring the price of interpretability in a transparent method tuned to the same data. And it offers the profession and its regulators a reproducible empirical basis for rules that are currently taught as conventions. We are careful about what the evidence can and cannot say: we validate the *predictive accuracy of simulated professional rules* on observable characteristics, not the judgement of actual appraisers, whose files, comparables and adjustments we do not observe.

The remainder of the paper is organised as follows. Section 2 reviews the literature on the sales comparison approach, adjustment estimation and automated valuation. Section 3 describes the Vrai Prix database and the construction of the analysis sample. Section 4 presents the point-in-time backtesting protocol, the configuration space, the benchmark models and the inference strategy. Section 5 reports the results, Section 6 examines their stability across submarkets, time, specifications and the selection step itself, and Section 7 concludes.

2. Related literature

2.1. The sales comparison approach and its conventions

The sales comparison approach rests on the principle of substitution: a buyer will pay no more for a property than the cost of acquiring an equally desirable substitute (Appraisal Institute, 2020). Its implementation follows a sequence that has barely changed since the method was formalised: identify comparable sales, verify that they are arm's-length, adjust each sale price for differences in transaction conditions, market conditions (time), location and physical characteristics, and reconcile the adjusted indications into a point estimate. Professional texts and standards give guidance on each step but rarely quantify it. The Appraisal Institute recommends "a sufficient number" of comparables and treats three as a practical minimum; the International Association of Assessing Officers' standard on automated valuation models recommends that comparable-sales algorithms use "three to ten" sales and weight them by similarity (International Association of Assessing Officers, 2018). Adjustments are to be "market-derived," typically through paired-sales analysis, but in practice they are frequently taken from rules of thumb, published grids, or the appraiser's experience

(Diaz, 1990; Diaz and Hansz, 2001). Lipscomb and Gray (1990) compared four market-derived adjustment methods on a single-market sample and found that the choice of method materially changes the adjustments, which is itself an argument for testing them.

A long-standing academic literature has treated the approach as a statistical estimator and questioned whether its conventions are optimal. Colwell et al. (1983) showed that the familiar adjustment grid is a special case of a regression in which the coefficients are fixed a priori, so that its accuracy depends on how close the pre-set coefficients are to the implicit prices. Isakson (1986, 1988) proposed a nearest-neighbour technique in characteristic space as an alternative to the grid, and Isakson (2002) set out the linear algebra that unifies the two. Quan and Quigley (1991) modelled the appraiser as a Bayesian signal extractor who combines noisy comparable prices with a prior; in that framework the number of comparables and their weights follow from the relative variances of the signals. Vandell (1991) derived the optimal weights to assign to comparables as a function of their adjustment variance and showed that equal weighting is optimal only under restrictive conditions; Gau et al. (1992) extended the problem to the joint selection and weighting of comparables under a minimum-variance criterion, which trades off the number of comparables against their heterogeneity, and the exchange with Green (1994) and Gau et al. (1994) clarified the role of the covariance between adjustment errors. Epley (1997) noted that these optimal rules require variance estimates that practitioners rarely possess. Pace (1996) is the closest antecedent to our design: he interprets the grid estimator as a combination of a nearest-neighbour estimator and a parametric estimator and shows, analytically and by Monte Carlo simulation, that ordinary least squares dominates the grid, which dominates the nearest neighbour, under correct specification, and that the ranking is fragile under misspecification; Pace and Gilley (1998) generalise the two estimators into a single family. Lai et al. (2008) propose a "replication" estimator that selects comparables so as to reproduce the subject's characteristics in combination rather than one at a time. Lentz and Wang (1998) survey the appraisal issues raised by the lending process, including the reliance on a small number of comparables and on judgemental adjustments. These contributions are theoretical, simulated, or illustrated on a few hundred sales in one city; none evaluates the rules out of sample on a large, heterogeneous market. Our design implements their estimators (nearest, most similar, minimum-adjustment selection; equal, distance, recency and similarity weights; median and mean reconciliation) as configurations of a single engine and measures their accuracy on more than a hundred thousand masked sales.

2.2. Estimating adjustments: grids, hedonics and local models

The estimation of adjustment magnitudes is where the sales comparison approach meets hedonic price theory. Since Rosen (1974), the hedonic regression of price on characteristics has been the canonical tool for recovering implicit prices, and its coefficients are natural candidates for

adjustment factors (Kang and Reichert, 1991; Pace, 1998). Kang and Reichert (1991) compared hedonic regression with grid-adjustment techniques on sales in a single county and found hedonic adjustments superior; Cannaday and Sunderman (1986) showed how depreciation adjustments can be estimated from the same regressions; Pace (1998) and Pace and Gilley (1997) argued for non-parametric and spatial estimation of adjustments. A large literature has since emphasised spatial heterogeneity in implicit prices and the value of submarket definitions for prediction (Des Rosiers et al., 2000; Bourassa et al., 2003; Goodman and Thibodeau, 2003; Case et al., 2004; Bourassa et al., 2007, 2010), and Dubé and Legros (2013, 2014) have stressed that spatial dependence in housing data is also temporal, since only prior sales can influence a current price. Whether locally estimated adjustments improve on a global grid in practice is an empirical question that depends on the bias–variance trade-off: local models track heterogeneity but are noisy where sales are few. We estimate adjustments at four levels of locality—a global within-municipality hedonic grid, a regional grid shrunk towards the global one, a per-valuation local regression on the candidate pool, and a non-linear gradient-boosting model—and test the trade-off directly across density strata.

Time adjustments raise a separate issue. Price indices are routinely estimated with the full sample and applied retroactively, which leaks information from the future into every "historical" valuation; Clapham et al. (2006) document how much repeat-sales and hedonic indices are revised as later data arrive. Bailey et al. (1963) and Case and Shiller (1989) introduced repeat-sales indices, Hill (2013) reviews hedonic index construction and Hill and Scholz (2018) show the gains from geospatial information in hedonic imputation; we follow the hedonic route but re-estimate the regional monthly index every month using only prior sales, so that the time adjustment applied to a comparable on a given valuation date is one that could actually have been computed on that date. The seasonality of housing markets documented by Ngai and Tenreyro (2014) matters here: because the price level of the valuation month is never observed, every point-in-time method inherits a seasonal error that we report rather than remove.

The assessed value is a further source of information that the appraisal literature treats ambiguously. Clapp and Giaccotto (1992) showed that assessed values can be used to construct price indices, and Thibodeau (2003) that marking assessed values to market with a small number of sales yields accurate estimates. Assessors' sales-ratio methods rest on the same idea. We treat the assessment ratio as one adjustment scheme among eight and quantify what the assessed value adds to both the traditional and the machine-learning methods.

2.3. Automated valuation and machine learning

Automated valuation models (AVMs) have moved from mass appraisal (Gloudemans, 1999) to lending and consumer applications, and the literature comparing them with traditional valuation has grown rapidly. Shiller and Weiss (1999) set out the logic of evaluating valuation systems on

their out-of-sample forecast errors rather than on in-sample fit. Pagourtzi et al. (2003) survey the methods; Zurada et al. (2011) and McCluskey et al. (2013) compare neural networks and other learners with hedonic models for mass appraisal; Kok et al. (2017) document the move from manual appraisal to automated valuation in institutional practice; Mayer et al. (2019) evaluate several machine-learning algorithms on a large Swiss dataset and find that gradient boosting and random forests dominate linear hedonics, with mixed-effects and spatial models close behind; Pace and Hayunga (2020) show that tree ensembles extract information from the residuals of hedonic and spatial models; Bogin and Shui (2020) find that both AVMs and appraisers are less accurate in rural areas, where comparables are scarce. Steurer et al. (2021) review the metrics used to evaluate AVMs and argue for reporting medians, shares within tolerance bands and coverage together, which is the convention we follow; Krause et al. (2020) compare error-based and model-based measures of AVM uncertainty, a question we take up when calibrating value intervals; Lorenz et al. (2023) show how interpretable machine learning can recover implicit prices from tree ensembles, which narrows but does not close the gap between opaque and transparent methods. A consistent finding is that tree ensembles with location features achieve median absolute errors of roughly 8–15% on residential sales, depending on market and data richness. Comparable-sales algorithms are themselves a major class of AVM (International Association of Assessing Officers, 2018), and hybrid approaches that use a model to adjust comparables are common in industry but rarely evaluated in the academic literature.

Two features of this literature motivate our design. First, machine-learning benchmarks are usually trained on random splits of the data, which overstate accuracy relative to a temporal split (Mayer et al., 2019); we use strict chronological separation and quarterly retraining, which is how such models are deployed. Second, the information set is rarely held constant: AVMs often exploit variables (assessed values, listing information) that the traditional method in the same study does not. We give every method the same characteristics and the same point-in-time index, and we separately quantify the value of the assessed value as an input to both the traditional and the machine-learning methods. Our inference draws on the forecast-evaluation literature: paired comparisons of losses on common targets (Diebold and Mariano, 1995; Giacomini and White, 2006), the median as the loss-minimising point forecast under absolute error (Gneiting, 2011; Hyndman and Koehler, 2006), the protection against data snooping that a separate selection sample affords when many specifications are searched (White, 2000; Hansen, 2005), and distribution-free calibration of prediction intervals on a hold-out sample (Lei et al., 2018).

2.4. Appraiser behaviour and the limits of simulation

Finally, a behavioural literature studies how appraisers actually select and adjust comparables and documents anchoring on contract prices, prior appraisals and assessed values (Northcraft and

Neale, 1987; Diaz, 1990; Diaz and Hansz, 2001); on administrative data, Eriksen et al. (2019) and Calem et al. (2021) show that purchase appraisals cluster at or just above the contract price. We do not observe appraisers' files, and our results concern the predictive accuracy of *rules*, not of practitioners. The distinction matters in both directions: a practitioner may do better than any rule by using information we do not observe (condition, renovations, listing history), and worse by anchoring. Our contribution is to establish what the observable record supports, so that professional rules and training can be benchmarked against it.

3. Data

3.1. The Vrai Prix transaction database

Our data come from Vrai Prix, a research database that assembles every residential transfer registered in the Quebec land register (*Registre foncier du Québec*) since January 2021 and links each transfer to the corresponding unit of the municipal assessment roll (*rôle d'évaluation foncière*). The snapshot used here was taken on 10 October 2026 and contains 766,596 transactions and 3.75 million assessment units covering the whole province. For each transaction we observe the registration date, the price, the civic address and coordinates, the property type, and—through the link to the roll—the physical characteristics recorded by the municipal assessor: living area (sum of floor areas), lot area, year of construction, number of storeys, physical link (detached, semi-detached, row, integrated), construction type, the neighbourhood unit used by the assessor, and the assessed value in force at the date of sale together with its land and building components. Assessed values are revised on a three-year cycle that differs across municipalities; we record, for each sale, whether the assessed value in force belongs to the roll that was current at the snapshot date, to the previous roll, or to an older one, and we call this the roll *vintage*.

Two features of this database are essential for our purpose. First, it is a population, not a sample: every registered sale is present, so that the comparable pool available on any valuation date is the one an appraiser with full access to the register would have had. We date each sale by its registration date and ignore the lag between the deed and its publication in the register, which slightly favours every method in the same way. Second, the characteristics are those of the assessment roll, harmonised across the province by the *Manuel d'évaluation foncière du Québec*, rather than listing descriptions whose coding varies by broker. The price of this harmonisation is that some characteristics commonly adjusted for in practice—garage, pool, number of bedrooms and bathrooms, interior quality, condition, renovations—are not recorded in the roll extract and are therefore unavailable. We return to this limitation in Section 7.

3.2. Sample construction

We restrict attention to single-family dwellings, the segment for which the sales comparison approach is the reference method. Starting from 456,950 transactions coded as single-family, we apply a sequence of explicit exclusion rules, each logged with the number of observations it removes (Table 1). The rules remove transactions that cannot be matched to the roll; duplicates; units whose roll record is not a single dwelling; implausible or missing living area, lot area or year of construction; sales closed before the dwelling was built (which are land sales) or in the year of construction (whose prices are, on inspection, those of lots or unfinished buildings: a median of $346,108 and a price-to-assessment ratio of 1.15, against $450,000–480,000 and 1.30 for dwellings aged one to four years); prices outside $50,000–$3,000,000; and resales within 180 days with a price change above 40%, which signal renovations or flips for which the recorded characteristics no longer describe the dwelling.

The most consequential rule is the exclusion of sales whose price-to-assessment ratio lies outside $[0.33, 3]$. The register does not flag non-arm's-length transfers (intra-family sales, transfers at nominal prices, distress sales), and an extreme ratio is the only observable indicator. An appraiser would screen such sales out of the comparable pool; we therefore exclude them both as targets and as comparables, and report the count. Because the rule uses the price, it also removes some genuine but unusual sales, and Section 6 examines how the results depend on its bounds. All other rules are standard. The final sample comprises 390,996 sales between 2021-01-05 and 2026-09-29, with a median price of $400,000.

Table 1. Sample construction

Step	n
Single-family transactions, raw	456,950
Not matched to the assessment roll	9,255
Duplicates (same unit, same date)	355
Roll unit not a single dwelling or more than one dwelling	29,509
Coordinates missing or outside Quebec	0
Living area missing or outside [40, 600] m^2	3,797
Lot area missing or outside [80, 20,000] m^2	4,278
Year built missing, before 1850 or after the sale (land sale)	2,393
Sale in the year of construction (lot or unfinished building)	5,427
Price outside [$50,000, $3,000,000]	528
Assessed value missing or price/assessment ratio outside [0.33, 3]	8,074
Resale within 180 days with price change above 40% (flip)	794
Physical link or construction type missing	1,512
Municipality absent from the MAMH directory	32
Final sample	390,996

3.3. Chronological partition and descriptive statistics

The sample is partitioned chronologically. Sales from January 2021 to December 2023 (201,642 observations) serve to develop the adjustment grids and benchmark models. Sales in 2024 (64,795 observations) are the *validation* targets, on which configurations and hyperparameters are chosen. Sales from January 2025 to September 2026 (124,559 observations) are the *test* targets, on which all results reported as out of sample are computed. Every sale in the sample, whatever its period, is eligible as a comparable for any later target, and the grids and benchmark models are re-estimated each quarter on all prior sales, including those of the validation period once it has elapsed; what distinguishes the periods is the role of the sale as a target.

Table 2 reports descriptive statistics by period and Figure 1 shows the hedonic price index by region over the sample. Prices rose steeply until mid-2022, fell by about 13% to a trough in early 2023, and rose again from 2024 onwards; over the whole period the provincial index gained 79%. The recovery was strongly heterogeneous across regions (Table 26 in the Appendix): between December 2024 and September 2026 prices rose by about 25% in the Capitale-Nationale and 14% in the Montérégie, against 4% on the island of Montréal, whose cumulative gain since 2021 (33%) is less than half the provincial one, while the Estrie more than doubled (112%). This heterogeneity

is a feature, not a nuisance: it is precisely the environment in which the quality of time adjustments and the locality of comparables matter. The median price-to-assessment ratio also falls from 1.39 in 2021–2023 to 1.20 in 2025–2026 as the 2025 and 2026 rolls incorporate the boom, which is why the assessment-ratio adjustment must be restricted to comparables on the same roll (Section 4.3).

Table 2. Descriptive statistics by period

Statistic	2021–23 (dev.)	2024 (val.)	2025–26 (test)
Sales	201,642	64,795	124,559
Median price ($)	350,000	415,000	470,000
Mean price ($)	401,267	463,665	520,101
Median living area (m^2)	112.0	112.7	113.3
Median lot area (m^2)	650	645	650
Median age (years)	41.0	42.0	43.0
Detached (%)	81.6	80.9	81.3
Median price/assessment ratio	1.39	1.35	1.20
Median sales within 2 km, 12 mo.	–	92	108
Median candidate pool (25 km, 24 mo.)	–	6,188	6,788

Note: "Sales within 2 km, 12 mo." is the number of sample sales within 2 km of the target in the twelve months preceding its date (the density measure of Section 3.4); "candidate pool" is the number of prior sales within 25 km and 24 months. Both are defined for targets only.

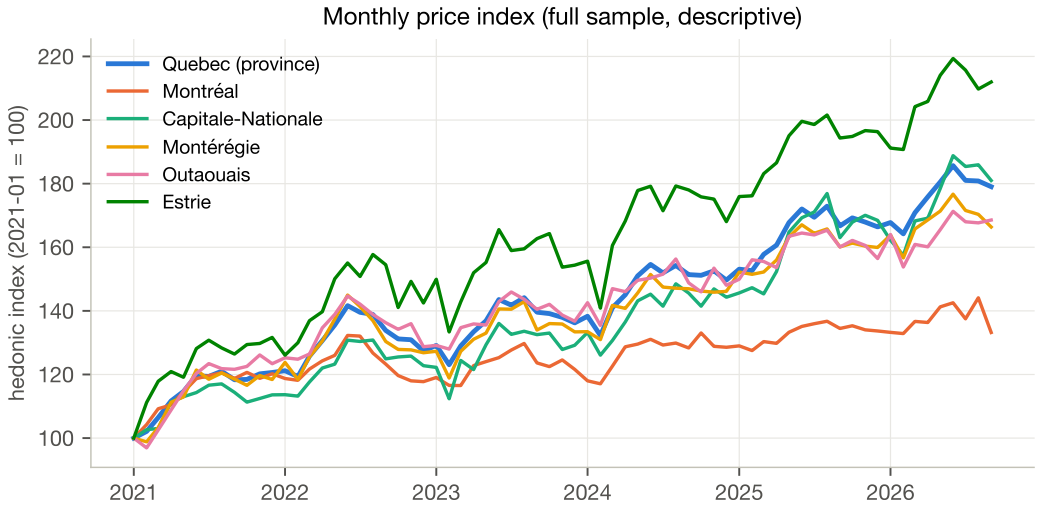

Figure 1. Hedonic price index, province and selected regions, full sample (descriptive). Monthly effects from a within-municipality regression of log price on characteristics, January 2021 = 100. The point-in-time indices used in the backtest are re-estimated every month from prior sales only and are not shown.

3.4. Market density

Because the accuracy of any comparable-based method depends on the supply of comparables, we characterise each target by the number of sample sales that occurred within 2 km during the twelve months preceding its date, excluding prior sales of the subject itself. This *density* measure is computed from the target's own comparable pool and is therefore available at valuation time. We group targets into five strata—fewer than 5, 5–19, 20–49, 50–199 and 200 or more sales—which range from rural municipalities and small towns to the dense suburbs of Montréal and Quebec City. Table 3 gives the distribution of test targets across strata; a fifth of the test targets are in markets with fewer than twenty sales within 2 km in a year.

Table 3. Distribution of test targets by market density

Density (sales within 2 km, 12 months)	Test targets	Share (%)
< 5	9,426	7.6
5–19	15,149	12.2
20–49	15,553	12.5
50–199	56,442	45.3
≥ 200	27,989	22.5

4. Methodology

4.1. Point-in-time backtesting

The core of the design is a backtest that reproduces, for each target sale i with date τ_i and price P_i, the information set available to a valuer on τ_i. Let $\mathscr{S}$ denote the clean sample. The *candidate pool* of i is

$$\mathscr{C}_i = \{\, j \in \mathscr{S} : \tau_j < \tau_i,\ \tau_i - \tau_j \leq 24 \text{ months},\ d_{ij} \leq 25 \text{ km},\ u_j \neq u_i \,\}, \tag{1}$$

where d_{ij} is the distance between the two properties and u_j the assessment unit; the last condition excludes prior sales of the subject itself, so that the method is evaluated as a comparable-sales method rather than a repeat-sales one. Every quantity entering a valuation on τ_i is computed from sales with $\tau_j < \tau_i$: the regional price index used to time-adjust comparables is re-estimated at the start of every month m from all sales in months before m (Section 4.2); adjustment coefficients and benchmark models are re-estimated at the start of every quarter q from all sales in quarters before q (Sections 4.3 and 4.4). No configuration, hyperparameter or weight is chosen on test-period data: all are selected on the 2024 validation targets and frozen.

A subtle but important consequence is that the price level of the target's own month is never

observed. All methods, traditional and machine-learning alike, work at the level of the last fully observed month, exactly as a valuer must. This introduces a small, common and seasonal bias that we report rather than remove (Section 6.1).

4.2. Point-in-time price index

For each evaluation month m and administrative region r we estimate, on all sales of region r in months $0,\dots,m-1$, the within-municipality regression

$$\ln P_j = \alpha_{\mu(j)} + \mathbf{x}_j'\beta + \sum_{s\le m-2} \delta_s^{(r,m)} \mathbb{1}[\text{month}_j = s] + \varepsilon_j, \tag{2}$$

where α_μ is a municipality fixed effect, $\mathbf{x}_j$ the characteristics (log living area, log lot area, age and its square, physical link, construction type, storeys, new-build indicator) and $\delta_s^{(r,m)}$ the effect of month s relative to the last observed month $m-1$, whose effect is normalised to zero. The municipality fixed effects prevent the index from being contaminated by shifts in the composition of sales between expensive and inexpensive municipalities within a region. For regions with fewer than 200 sales the provincial index is used; otherwise the monthly effects are shrunk towards the provincial estimate with weight $n_s/(n_s+40)$, where n_s is the number of regional sales in month s. A comparable sold in month $s<m$ and valued in month m is time-adjusted by the factor $\exp(-\delta_s^{(r,m)})$, i.e. brought to the level of month $m-1$; comparables sold in the same month as the target, for which no effect can be estimated, are not adjusted. The index is thus a three-dimensional object (evaluation month, region, sale month), and the estimate of any historical month is allowed to change as later data arrive, as it does in reality (Clapham et al., 2006).

4.3. Configurations of the sales comparison approach

A configuration is a tuple $c = (\text{selection}, d, t, k, \text{adjustment}, \text{weighting}, \text{reconciliation})$. Given the pool $\mathscr{C}_i$, the method proceeds in four steps.

Selection. Candidates satisfying $d_{ij} \le d$ and $\tau_i - \tau_j \le t$ are ranked by a score and the k best are retained. Physical dissimilarity is measured by

$$D_{ij} = \left[\left(\frac{\Delta \ln A_{ij}}{0.25}\right)^2 + \left(\frac{\Delta \ln L_{ij}}{0.6}\right)^2 + \left(\frac{\Delta \text{age}_{ij}}{15}\right)^2 + \mathbb{1}[\text{link}_i \neq \text{link}_j] + 0.5\,\mathbb{1}[\text{storeys}_i \neq \text{storeys}_j]\right]^{1/2}, \tag{3}$$

where A is living area, L lot area and Δ denotes the difference between subject and comparable; the denominators are typical within-market standard deviations, so that D_{ij} is in standard-deviation units. Four scores are implemented: *nearest*, $s_{ij}^{\text{geo}} = d_{ij}$; *most similar*, $s_{ij}^{\text{sim}} = D_{ij} + d_{ij}/25$, in which distance (in km) acts only as a tie-breaker; *compromise*, $s_{ij}^{\text{mix}} = D_{ij} + d_{ij}/2$, which trades one standard

deviation of dissimilarity against two kilometres; and *least adjusted*, $s_{ij}^{\text{adj}} = \sum_m |\hat{\beta}_m \Delta x_{ijm}| + d_{ij}/25$, the gross hedonic adjustment, which operationalises the professional advice to prefer comparables requiring the fewest adjustments.

Adjustment. Let P_j^T denote the time-adjusted price of comparable j. Each retained comparable is converted into an indication of the subject's value $\hat{P}_{i|j}$ under one of eight schemes:

(a) *none*: $\hat{P}_{i|j} = P_j$ (not even time-adjusted);

(b) *time only*: $\hat{P}_{i|j} = P_j^T$;

(c) *conventional grid*: dollar adjustments in the spirit of professional practice,

$$\begin{aligned}\hat{P}_{i|j} = P_j^T + \tfrac{1}{2}\tfrac{P_j^T}{A_j}(A_i - A_j) + \left[\tfrac{1}{2}\tfrac{\lambda_j P_j^T}{L_j}(L_i - L_j)\right]_{\pm 0.15 P_j^T} \\ + \left[-0.005\, P_j^T (\text{age}_i - \text{age}_j)\right]_{\pm 0.20 P_j^T} \pm 0.05\, P_j^T \mathbb{1}[\text{link differs}],\end{aligned} \tag{4}$$

where λ_j is the land share of the comparable's assessed value, $[\cdot]_{\pm b}$ denotes capping at $\pm b$, and the link adjustment is positive when the subject is detached and the comparable attached; the total adjustment is floored at -60% of P_j^T;

(d) *hedonic grid*: $\hat{P}_{i|j} = P_j^T \exp\big(\hat{\beta}'(\mathbf{x}_i - \mathbf{x}_j)\big)$, with $\hat{\beta}$ the within-municipality coefficients of a regression of the time-adjusted log price on $\mathbf{x}$, estimated on all prior sales and re-estimated quarterly;

(e) *regional grid*: as (d) with region-specific coefficients shrunk towards the global ones with weight $n_r/(n_r + 300)$;

(f) *local grid*: as (d) with coefficients $\hat{\mathbf{b}}_i$ re-estimated on the subject's own candidates within 10 km (at least 40 sales) by ridge regression penalised towards the regional grid $\hat{\beta}^{\text{reg}}$,

$$\hat{\mathbf{b}}_i = \arg\min_{\mathbf{b}} \sum_{j \in \mathscr{C}_i,\, d_{ij} \le 10} \big(\tilde{y}_j - \tilde{\mathbf{x}}_j' \mathbf{b}\big)^2 + \lambda \, \|\mathbf{b} - \hat{\beta}^{\text{reg}}\|^2, \qquad \lambda = 50, \tag{5}$$

where tildes denote variables centred on the local pool;

(g) *non-linear (GBM)*: $\hat{P}_{i|j} = P_j^T \exp\big(f(\mathbf{z}_i) - f(\mathbf{z}_j)\big)$, where f is the gradient-boosting model of the time-adjusted log price on characteristics and location described in Section 4.4 (the version without the assessed value);

(h) *assessment ratio*: $\hat{P}_{i|j} = V_i \cdot P_j^T / V_j$, where V is the assessed value in force at the sale, restricted to comparables in the same municipality and of the same roll vintage as the subject.

Scheme (h) is the ratio method used by assessors and, informally, by many appraisers; it uses information (the assessed value) that schemes (a)–(g) do not, and we treat it as a distinct model in the horse race.

Weighting and reconciliation. Let $v_{ij} = \ln \hat{P}_{i|j}$ and let w_{ij} be the weight of comparable j. Five weighting schemes are compared: equal weights; distance, $w_{ij} = 1/(0.25 + d_{ij})$; recency, $w_{ij} = \exp(-\text{months}_{ij}/6)$; similarity, $w_{ij} = 1/(0.1 + g_{ij})$ where g_{ij} is the gross hedonic adjustment in log points; and the product of the three. Two reconciliation rules are compared: the (weighted) median of the v_{ij}, and the dollar-weighted mean

$$\hat{P}_i = \frac{\sum_j w_{ij} \hat{P}_{i|j}}{\sum_j w_{ij}}. \tag{6}$$

The standard deviation $\hat{\sigma}_i$ of the v_{ij} across the retained comparables is recorded as a measure of uncertainty (Section 5.7).

Experiments. Experiment E1 maps the error surface over $k \in \{1,3,5,7,10,15,20,30,50\}$, $d \in \{0.5,1,2,5,10,25\}$ km and $t \in \{3,6,12,24\}$ months, for the nearest, most-similar and compromise selections and the time-only and hedonic adjustments, with equal weights and the median (1,296 configurations). Experiment E2 is the full factorial of selection (4) $\times$ adjustment (8) $\times$ weighting (5) $\times$ reconciliation (2) at three anchors $(k,d,t) = (3, 2\text{ km}, 6\text{ mo})$, $(5, 5\text{ km}, 12\text{ mo})$ and $(10, 10\text{ km}, 24\text{ mo})$ (960 configurations). All 2,256 configurations are evaluated on every target.

4.4. Benchmark models

Sixteen named models are compared (Table 4). M1–M8 and M13 are configurations of the sales comparison approach that correspond to recognisable professional practices. M10 is a direct global hedonic model: the within-municipality coefficients augmented with region, county (MRC), municipality and neighbourhood-unit effects estimated as shrunken residual means, plus a median correction estimated on the validation targets. M11 is a gradient-boosting model (LightGBM, 127 leaves, learning rate 0.05, L_1 objective, number of trees chosen by early stopping on the chronologically last 10% of the training period) of the time-adjusted log price on characteristics, coordinates, region, construction type and the month of sale; M11a excludes and M11b includes the assessed value, its land share and its roll vintage. Both are retrained at the start of every quarter on all prior sales. Because the L_1 objective targets the conditional median of the log price, the models estimate the point forecast that minimises expected absolute error (Gneiting, 2011), which is the loss the MdAPE rewards; a squared-error objective would target the conditional mean of a left-skewed distribution and produce a negative median bias (Section 6.7). M9 is an *optimised reconciliation*: a convex combination, estimated by non-negative least squares on the validation targets, of the seven comparable-based indications at the $(5,5,12)$ anchor. M12 is a hybrid combining M11b with the GBM-adjusted and hedonic-adjusted comparable estimates, with weights estimated the same way.

M0 is the single best of the 2,256 configurations on the validation targets, among those that return a value for at least 97% of targets.

Table 4. Models in the horse race

Model	Definition
M1	Median of the 3 nearest sales ($\leq$ 2 km, $\leq$ 6 months), time adjustment only
M2	Mean of the 5 nearest sales ($\leq$ 5 km, $\leq$ 12 months), time adjustment only
M3	Median of the 5 most physically similar sales ($\leq$ 5 km, $\leq$ 12 months), time adjustment only
M4	5 nearest sales, distance-weighted mean, time adjustment only
M5	5 nearest sales, recency-weighted mean, time adjustment only
M6	5 nearest sales, conventional adjustment grid, median
M7	5 nearest sales, hedonic adjustment grid, median
M8a	5 nearest sales, locally re-estimated grid, median
M8b	5 nearest sales, non-linear (GBM) adjustments, median
M9	Optimised reconciliation: NNLS combination of seven comparable-based indications (weights fitted on 2024)
M10	Direct global hedonic model with region, MRC, municipality and neighbourhood effects
M11a / M11b	Gradient boosting without / with the assessed value, retrained quarterly
M12	Hybrid: NNLS combination of M11b and GBM- and hedonic-adjusted comparables (weights fitted on 2024)
M13	Assessment ratio: $V_i\times$ median price-to-assessment ratio of the 5 nearest same-roll sales ($\leq$ 5 km, $\leq$ 12 months)
M0	Best of all 2,256 configurations on 2024 (coverage $\geq$ 97%), frozen

4.5. Accuracy measures

For each target i and method (or configuration) c that returns a value, the absolute percentage error is

$$\mathrm{APE}_i(c) = \frac{|\hat{P}_i(c) - P_i|}{P_i}, \qquad e_i(c) = \frac{\hat{P}_i(c) - P_i}{P_i}. \tag{7}$$

Over a set of targets $\mathscr{T}$, let $\mathscr{T}_c \subseteq \mathscr{T}$ be the targets for which c returns a value. We report

$$\mathrm{MdAPE}(c) = \operatorname*{med}_{i\in\mathscr{T}_c} \mathrm{APE}_i(c), \qquad \mathrm{MAPE}(c) = \frac{1}{|\mathscr{T}_c|}\sum_{i\in\mathscr{T}_c} \mathrm{APE}_i(c), \qquad \mathrm{Bias}(c) = \operatorname*{med}_{i\in\mathscr{T}_c} e_i(c), \tag{8}$$

together with the shares of estimates within ±5, ±10 and ±20% of the price and the *coverage* $|\mathscr{T}_c|/|\mathscr{T}|$ (Steurer et al., 2021; Hyndman and Koehler, 2006). The MdAPE is the headline measure

because valuation errors are heavy-tailed and the median is the natural loss of an appraiser who is judged on the typical file; the MAPE and the tolerance shares are reported alongside because they weight the tails differently. Coverage is reported because a restrictive configuration does not fail on difficult targets—it abstains—and an MdAPE computed on the targets it covers would flatter it. Configuration selection on the validation set is restricted to configurations with coverage of at least 97%.

4.6. Inference

Because all methods are evaluated on the same targets, comparisons are paired. For a pair of methods (A, B) and the common set of covered targets, the difference in MAPE is the mean of the loss differential $\Delta_i = \mathrm{APE}_i(A) - \mathrm{APE}_i(B)$, the quantity tested by Diebold and Mariano (1995) and Giacomini and White (2006), and the difference in MdAPE is the difference of the two medians. Valuation errors are correlated within local markets and within periods, because neighbouring sales share unobserved amenities and the same index estimate; we therefore treat municipality–quarter cells as clusters. Confidence intervals are obtained from a cluster bootstrap (Cameron et al., 2008) that resamples cells with replacement $B = 999$ times and recomputes both statistics on each replication, and we also report a cluster-robust t statistic for the mean differential,

$$t = \frac{\bar{\Delta}}{\left[\sum_g \left(\sum_{i \in g} (\Delta_i - \bar{\Delta})\right)^2\right]^{1/2} / N}, \tag{9}$$

where g indexes the clusters and N the number of common targets. We also report the share of targets on which A is more accurate than B. Section 6.5 shows that the intervals are insensitive to the cluster definition.

The second inferential issue is selection. Choosing the best of 2,256 configurations on the same data that are used to report its accuracy would overstate it (White, 2000; Hansen, 2005). We avoid this by selecting on the 2024 validation targets and reporting on the 2025–2026 test targets, and Section 6.2 examines how the validation ranking carries over to the test period.

4.7. Hypotheses

Five hypotheses from the professional and academic debate are tested directly. **H1** (similarity versus proximity): a comparable that is very close but physically dissimilar is less informative than one that is farther away but similar. **H2** (number of comparables): adding comparables improves accuracy up to a threshold beyond which heterogeneity dominates, and the threshold rises with market density. **H3** (local adjustments): locally estimated adjustments outperform fixed grids but become unstable in small samples. **H4** (weighting and liquidity): the optimal weighting scheme depends on the

liquidity of the submarket. **H5** (transparency): simple, transparent methods rival machine learning in homogeneous submarkets even if the latter dominates overall.

5. Results

All results in this section are computed on the 124,559 test targets of January 2025 to September 2026, with every configuration, weight and model frozen on the basis of the 64,795 validation targets of 2024. Validation-period figures are reported alongside to document the stability of the rankings. Differences between methods quoted in the text are paired differences on the common targets, computed before rounding, so that they may differ by a tenth of a point from the difference of the rounded levels.

5.1. How many comparables, how far, how far back?

Figure 2 plots the out-of-sample MdAPE against the number of comparables k for each search radius, with a twelve-month window and selection by distance, without physical adjustment (left) and with the hedonic grid (right). Figure 3 shows the best attainable MdAPE for each radius–window pair, and Table 5 reports the full surface for the hedonic grid.

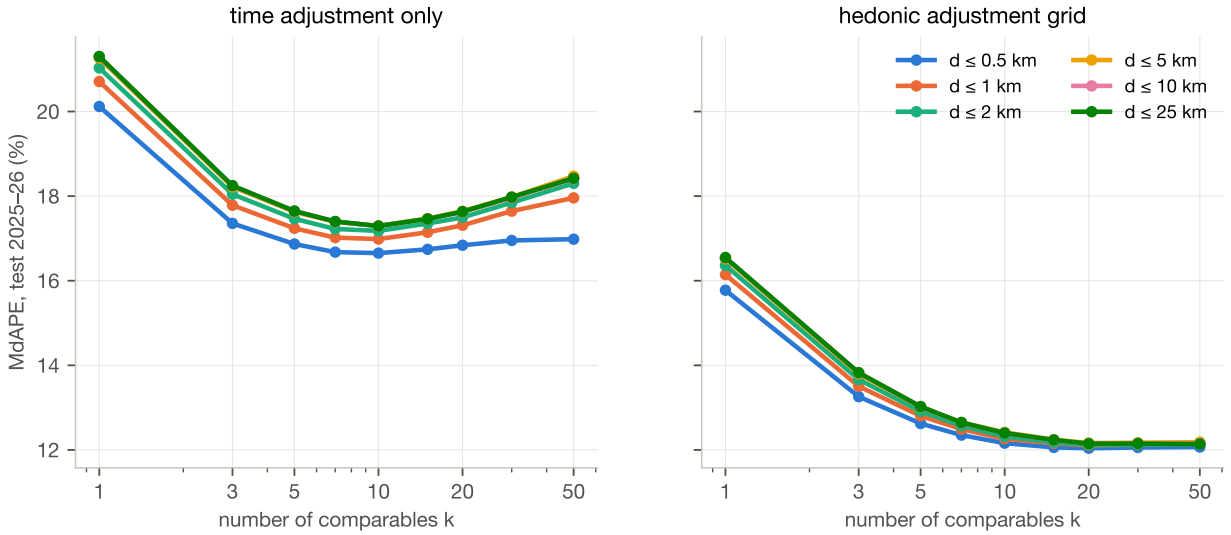

Figure 2. Out-of-sample MdAPE (2025–2026) by number of comparables and search radius; twelve-month window, selection by distance, median reconciliation. Left: time adjustment only. Right: hedonic adjustment grid.

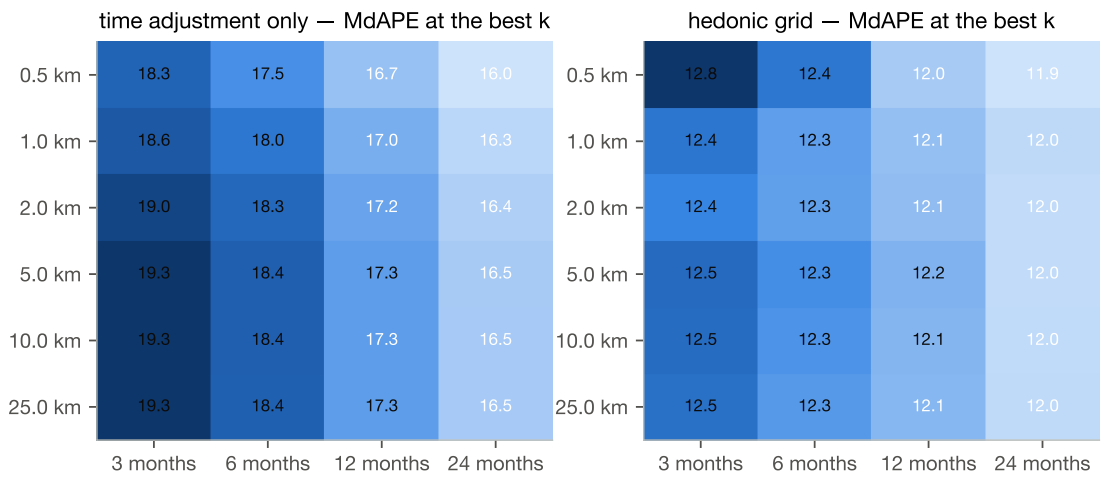

Figure 3. Best attainable out-of-sample MdAPE (over k) for each search radius and look-back window; selection by distance.

Table 5. Error surface Error(k,d): out-of-sample MdAPE (%), selection by distance, hedonic grid, twelve-month window

Radius	$k=1$	$k=3$	$k=5$	$k=7$	$k=10$	$k=15$	$k=20$	$k=30$	$k=50$
≤ 0.5 km	15.8	13.3	12.6	12.3	12.2	12.1	**12.0**	12.1	12.1
≤ 1 km	16.1	13.5	12.8	12.5	12.3	12.2	**12.1**	12.1	12.1
≤ 2 km	16.4	13.7	12.9	12.6	12.3	12.2	**12.1**	12.1	12.1
≤ 5 km	16.5	13.8	13.0	12.7	12.4	12.2	**12.2**	12.2	12.2
≤ 10 km	16.5	13.8	13.0	12.6	12.4	12.2	**12.1**	12.1	12.1
≤ 25 km	16.5	13.8	13.0	12.6	12.4	12.2	12.1	12.1	**12.1**

Note: MdAPE on the 2025–2026 test targets for which the configuration returns a value; the minimum of each row is in bold. Coverage falls below 90% for radii of 1 km and less in low-density markets; see Table 27 in the Appendix.

Three regularities emerge. First, without physical adjustment, accuracy improves with k up to about ten comparables (17.3% at $k = 10$ for a 10 km radius and twelve-month window) and then deteriorates (18.4% at $k = 50$): beyond the nearest few, additional sales are increasingly dissimilar and the median of their time-adjusted prices drifts away from the subject's value. This is the pattern that Pace (1996) predicts for a nearest-neighbour estimator whose bias grows with the number of neighbours. Second, with the hedonic grid the picture changes qualitatively. The error falls from 16.5% with a single comparable to 13.8% with three, 13.0% with five, 12.4% with ten, reaches 12.1% at $k = 20$ and stays there to $k = 50$ (12.1%). Adjustment converts dissimilar sales into usable evidence, so that the variance reduction from averaging more of them is no longer offset by their

heterogeneity. Third, the window matters and the radius does not. The three-month window is dominated by the twelve- and twenty-four-month windows at every radius (Figure 3); but at the best k the MdAPE with a 24-month window lies between 11.9 and 12.0% for every radius from 0.5 to 25 km, because k is a maximum and a wide radius only adds sales where the nearby ones run out. The conventional combination—three comparables, within a kilometre, within three months—yields an MdAPE of 13.6% with the hedonic grid and returns a value for only 89.2% of targets; the best configuration of the surface chosen on the 2024 validation set (compromise, $k = 30$, $d \leq 25$ km, $t = 24$ months, hedonic grid) yields 11.9% on validation and 11.1% on the test set.

Table 6 tests H2 by stratifying the k-curve by market density. The threshold beyond which additional comparables hurt exists only without physical adjustment (lower panel): with time adjustment alone the optimum is seven to ten comparables in the three densest strata, where the eleventh-nearest sale is already dissimilar, fifteen in the 5–19 stratum, and fifty in the sparsest one, where every sale within reach is needed. With the hedonic grid (upper panel) there is no threshold within the range studied: the error is minimised at twenty to fifty comparables in every density stratum and the curve is flat beyond twenty. H2 is therefore confirmed for the unadjusted method and rejected for the adjusted one—adjustment, not density, determines whether more comparables help.

Table 6. Optimal number of comparables by market density (H2): out-of-sample MdAPE (%)

Density	$k = 1$	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 15$	$k = 20$	$k = 30$	$k = 50$
Hedonic adjustment grid									
< 5	31.3	26.9	25.3	24.4	24.1	23.7	23.8	**23.6**	23.7
5–19	25.4	21.6	20.4	19.6	19.1	18.7	18.4	18.4	**18.2**
20–49	20.1	16.6	15.4	14.9	14.5	14.3	14.1	14.1	**13.9**
50–199	14.7	12.2	11.5	11.2	11.0	10.9	**10.9**	**10.9**	10.9
≥ 200	12.9	10.8	10.2	9.9	9.7	9.6	9.6	**9.6**	9.7
Time adjustment only									
< 5	39.2	34.7	33.2	32.2	32.1	32.1	31.7	31.7	**31.5**
5–19	33.8	29.4	28.3	27.7	27.2	**27.2**	27.4	27.6	27.8
20–49	26.6	23.0	22.3	21.9	**21.4**	21.7	21.7	22.2	22.4
50–199	18.5	15.9	15.4	**15.3**	15.3	15.6	15.7	16.1	16.6
≥ 200	16.0	13.7	13.1	13.1	**12.9**	13.0	13.2	13.7	14.3

Note: selection by distance, $d \leq 10$ km, twelve-month window, median. Density is the number of sample sales within 2 km in the twelve months preceding the target's date. The minimum of each row is in bold.

5.2. Which adjustments are warranted?

Table 7 and Figure 4 compare the eight adjustment schemes at the $(5, 5\text{ km}, 12\text{ mo})$ anchor with selection by distance, equal weights and the median. The time adjustment alone is worth only 0.3 points of MdAPE relative to raw prices with a twelve-month window and 1.4 points with a 24-month window: comparables are recent, and a monthly index estimated in real time is itself noisy. Any physical grid is worth a further 4.2 to 5.7 points. The stylised conventional grid—which uses no estimation whatsoever—recovers most of the gain of the hedonic grid (13.4% against 13.0%), consistent with the observation of Colwell et al. (1983) that the grid is a regression with pre-set coefficients that need only be roughly right. The refinements add little: the regional grid is indistinguishable from the global one, the local grid is slightly worse (13.5%), and the non-linear GBM adjustments improve on the hedonic grid by 1.1 points. The assessment-ratio adjustment stands apart: with 11.5% it is the most accurate single scheme, but it returns a value for only 90.3% of targets because it requires comparables on the same assessment roll.

Table 7. Adjustment schemes at the (5, 5 km, 12 months) anchor

Adjustment	MdAPE 2024	MdAPE 2025–26	MAPE	Bias	Within ±10%	Coverage
none	18.9	17.9	30.4	-4.2	31.0	99.8
time only	18.8	17.6	31.1	-0.8	31.7	99.8
conventional grid	14.1	13.4	26.4	0.7	39.7	99.8
hedonic grid	13.8	13.0	25.1	-0.5	40.4	99.8
regional grid	13.8	13.1	25.2	-0.5	40.4	99.8
local grid	14.3	13.5	25.7	-0.4	39.4	99.8
GBM adjustments	12.7	11.9	24.1	-0.6	43.6	99.8
assessment ratio	12.0	11.5	22.5	-0.1	40.6	90.3

Note: five nearest sales, equal weights, median. MdAPE, MAPE, bias (median signed error) and the share within ±10% in per cent, on the 2025–2026 test targets unless indicated; coverage is the share of test targets for which the scheme returns a value.

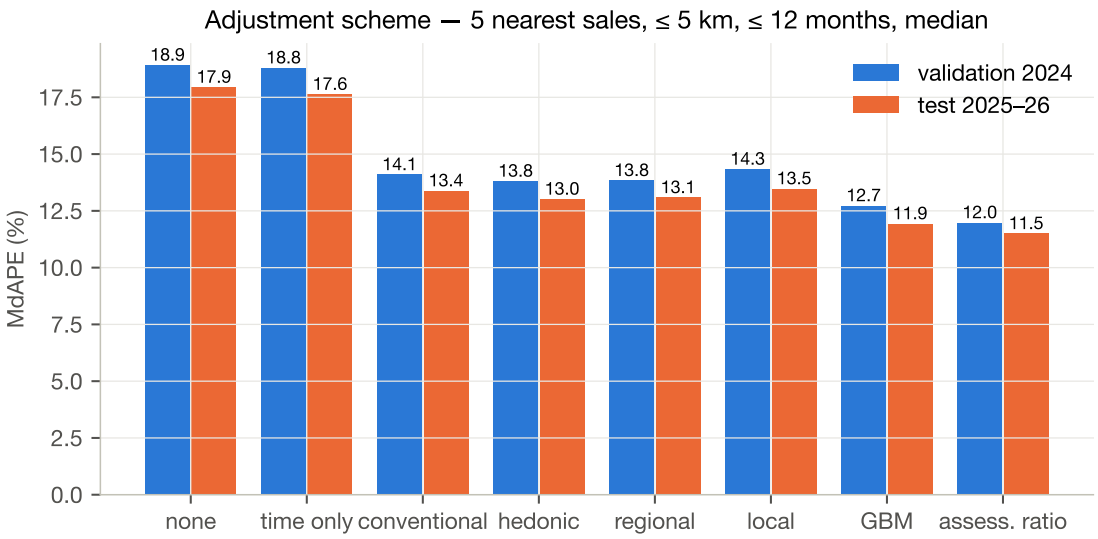

Figure 4. MdAPE by adjustment scheme, validation (2024) and test (2025–2026); five nearest sales, $\leq$ 5 km, $\leq$ 12 months, median.

Table 8 examines H3 by crossing the adjustment schemes with market density at two anchors. The local grid tracks the hedonic grid closely where sales are abundant and falls behind where they are scarce, as the bias–variance argument predicts; at the tight $(3, 2\text{ km}, 6\text{ mo})$ anchor, where the local regression is estimated on the smallest pools, the gap widens. The non-linear GBM adjustments are the most accurate scheme in every stratum, but their advantage over the hedonic grid is largest in dense markets (about a point) and smallest in the sparsest one. H3 is therefore rejected in its first part—local re-estimation does not outperform the fixed grid even where sales are plentiful—and confirmed in its second. The practical implication is that adjustment grids should be estimated on the largest pool that is homogeneous in implicit prices—in these data, the province with municipality fixed effects—rather than on the handful of sales around the subject.

Table 8. Adjustment schemes by market density (H3): out-of-sample MdAPE (%)

Density	conventional grid	hedonic grid	regional grid	local grid	GBM adjustments
Anchor $(5, 5\ km, 12\ months)$					
< 5	27.1	25.3	25.3	25.9	**24.5**
5–19	21.4	20.4	20.1	20.8	**19.7**
20–49	16.2	15.4	15.3	15.9	**14.6**
50–199	11.7	11.5	11.5	11.9	**10.4**
≥ 200	10.4	10.2	10.3	10.7	**9.2**
Anchor $(3, 2\ km, 6\ months)$					
< 5	30.0	28.6	28.6	29.5	**28.0**
5–19	23.8	22.4	21.9	22.8	**21.4**
20–49	17.7	16.9	16.6	17.4	**15.5**
50–199	12.8	12.5	12.6	12.9	**11.1**
≥ 200	11.1	10.8	11.0	11.4	**9.8**

Note: selection by distance, equal weights, median. Upper panel: $(5, 5\ \text{km}, 12\ \text{mo})$; lower panel: $(3, 2\ \text{km}, 6\ \text{mo})$. The minimum of each row is in bold.

The estimated hedonic grid itself is of independent interest (Table 9). Within a municipality, the elasticity of price with respect to living area is 0.60, that with respect to lot area is 0.07, and depreciation is 1.0% per year at the mean age with a convex profile, a magnitude in line with the single-family depreciation estimates of Cannaday and Sunderman (1986); semi-detached and row dwellings sell at discounts of 7.4% and 10.5% relative to detached dwellings of the same size. The within-municipality R^2 is 0.36: the roll characteristics explain about a third of the price variation among dwellings of the same municipality, and the rest—location within the municipality, condition, quality—is what the comparables themselves must supply. These coefficients are stable across the eleven quarterly re-estimations to the second decimal.

Table 9. Hedonic adjustment grid (within-municipality coefficients, log price)

Characteristic	Coefficient
log living area	0.605
log lot area	0.067
age (years)	-0.010
age^2/100	0.004
semi-detached	-0.074
row / integrated	-0.105
construction type: full storeys	-0.309
split-level	-0.098
mansard	0.012
storeys	-0.094
built within 3 years	-0.015

Within-R^2 = 0.362; n = 266,437 sales before 2025-01-01.

Note: coefficients of the within-municipality regression of the time-adjusted log price on characteristics, estimated on all sales before 2025 (grid in force in the first test quarter). Living and lot areas in logs; age in years; age^2/100. Omitted categories: detached, single-storey construction type.

5.3. How should comparables be selected? (H1)

Table 10 compares the four selection criteria at the $(5, 5, 12)$ anchor with the hedonic grid. Selecting by physical similarity rather than by distance lowers the MdAPE by 0.5 points when comparables are adjusted, and the compromise criterion, which uses both, is the best of the four. The "least adjusted" criterion, which embodies the professional advice to prefer comparables needing few adjustments, performs like the similarity criterion for which it is a proxy.

Table 10. Selection criteria at the (5, 5 km, 12 months) anchor, hedonic grid

Selection	MdAPE 2024	MdAPE 2025–26	Bias	Within $\pm$10%
nearest	13.8	13.0	-0.5	40.4
most similar	13.3	12.5	-0.1	42.1
compromise	12.9	12.1	-0.1	43.0
least adjusted	13.1	12.2	-0.1	42.8

The direct test of H1 is starker. Without any physical adjustment, the five nearest sales within one kilometre yield an MdAPE of 17.2% (coverage 96.0%), and the five nearest within ten kilometres 17.6%; the five most similar sales within ten kilometres yield 13.0%. Holding the radius fixed, the

selection criterion alone is worth more than four points, as much as the hedonic grid is worth to the nearest sales: the nearby-but-dissimilar comparable is worse than the distant-but-similar one by a wide margin. The reason is that, in these data, the physical characteristics explain a larger share of the within-market variance of prices than micro-location does once the municipality is fixed. With the hedonic grid the two criteria converge (13.0% against 12.7%), because adjustment does for the nearest sales what selection does for the most similar; the compromise criterion, which uses both, is best (12.1%). H1 is confirmed.

5.4. Weighting and reconciliation (H4)

Table 11 reports the ten weighting–reconciliation combinations at the $(5, 5, 12)$ anchor with the hedonic grid, and Table 12 stratifies seven of them by density. The range of the MdAPE across all ten combinations is 1.0 points. The median is more robust than the mean, as expected from the heavy tails of adjusted indications, and the mean carries a negative bias of about -1.5 points because averaging in dollars overweights the high indications in logs. The density profile does not support H4 in its stated form. The equal-weight median is the best or within a tenth of a point of the best in the three densest strata; weighting by gross adjustment improves on the equal-weight mean in dense markets (10.5% against 10.9% where 200 or more sales are available) but not in sparse ones; distance weights are neutral everywhere; and the product of the three weights with the median is the worst combination in the two sparsest strata, because it concentrates the weight on one or two sales. Weighting thus matters, if anywhere, in liquid markets, which is the opposite of what H4 predicts, and the effects are an order of magnitude smaller than those of selection and adjustment. The marginal effects over the whole E2 design (Table 13) make the hierarchy explicit: the adjustment factor spans 11.8–15.4 points of MdAPE, selection 12.9–14.4, weighting 13.1–13.6, and reconciliation 13.2 (median) against 13.6 (mean).

Table 11. Weighting and reconciliation at the $(5, 5\text{ km}, 12\text{ months})$ anchor, hedonic grid

Weighting	Reconciliation	MdAPE 2024	MdAPE 2025–26	Bias	Within $\pm 10\%$
equal	median	13.8	13.0	-0.5	40.4
equal	mean	14.6	13.7	-1.8	38.4
1/distance	median	13.9	13.1	-0.5	40.3
1/distance	mean	14.6	13.7	-1.8	38.6
recency	median	14.5	13.7	-0.5	38.8
recency	mean	14.8	13.9	-1.6	38.0
1/gross adjustment	median	13.8	12.9	-0.3	40.8
1/gross adjustment	mean	14.3	13.3	-1.5	39.7
product of the three	median	14.6	13.6	-0.3	39.2
product of the three	mean	14.5	13.5	-1.4	39.3

Note: "1/gross adjustment" weights each comparable by the inverse of the absolute hedonic adjustment it requires; "product of the three" multiplies the distance, recency and gross-adjustment weights.

Table 12. Weighting by market density (H4): out-of-sample MdAPE (%)

Density	equal/med.	equal/mean	dist./mean	rec./mean	1/adj./mean	prod./mean	prod./med.
< 5	25.3	25.1	**25.0**	25.9	25.2	25.8	28.1
5–19	20.4	20.6	20.5	21.0	**20.3**	20.6	21.8
20–49	**15.4**	16.0	16.1	16.4	16.0	16.2	16.7
50–199	**11.5**	12.3	12.3	12.5	11.9	12.0	12.0
≥ 200	**10.2**	10.9	10.9	11.0	10.5	10.6	10.5

Note: $(5, 5\text{ km}, 12\text{ mo})$ anchor, selection by distance, hedonic grid. Columns: weighting/reconciliation; "1/adj." is the inverse gross-adjustment weight, "prod." the product of the distance, recency and gross-adjustment weights. The minimum of each row is in bold.

Table 13. Marginal effects of the four rule factors in the E2 design

Factor	Level	Mean MdAPE
Selection	nearest	14.4
Selection	most similar	13.2
Selection	compromise	12.9
Selection	least adjusted	13.1
Adjustment	none	15.4
Adjustment	time only	14.7
Adjustment	conventional grid	13.2
Adjustment	hedonic grid	13.1
Adjustment	regional grid	13.1
Adjustment	local grid	13.3
Adjustment	GBM adjustments	12.6
Adjustment	assessment ratio	11.8
Weighting	equal	13.2
Weighting	1/distance	13.4
Weighting	recency	13.6
Weighting	1/gross adjustment	13.1
Weighting	product of the three	13.6
Reconciliation	median	13.2
Reconciliation	mean	13.6

Note: mean out-of-sample MdAPE (%) of all E2 configurations sharing the given level of the factor, averaged over the other factors and the three anchors.

5.5. The horse race

Table 14 and Figure 5 report the sixteen named models; Table 31 in the Appendix adds cluster-bootstrap confidence intervals for each model's MdAPE and MAPE. The ordering is informative at every level. The traditional practices M1–M5—a few nearest sales, time-adjusted, weighted in one way or another—cluster at 17.9–19.1% of MdAPE, with fewer than a third of estimates within $\pm 10\%$ of the price. Adding a grid (M6, M7) or selecting by similarity (M3) brings the error into the 13.0–13.4% range. The assessment ratio (M13) reaches 11.5% on the targets it covers. The best single transparent configuration frozen on 2024 (M0: compromise, $k = 30$, ≤ 25 km, 24 months, hedonic grid, equal weights, median) reaches 11.1% (95% interval 10.8–11.5) with 46.1% of estimates within $\pm 10\%$, and the optimised reconciliation M9 10.6%; M9 places weights of 0.59 on the assessment-ratio indication, 0.33 on the GBM-adjusted one and 0.07 on the time-only one,

and nothing on the other grids. Among the external benchmarks, the direct hedonic model M10 is the least accurate (13.8%): a global model with municipality and neighbourhood effects cannot match the local information of comparables. The gradient-boosting model without the assessed value (M11a, 10.7%) is only 0.4 points better than the best frozen comparables configuration; with the assessed value (M11b) it reaches 9.0% (8.8–9.2) and 54.1% within ±10%, and the hybrid M12 8.9%, with 0.91 of its weight on M11b and the rest on the GBM- and hedonic-adjusted comparables.

Table 14. Horse race: out-of-sample accuracy, 2025–2026

Model	MdAPE 2024	MdAPE 2025–26	MAPE	Bias	±5%	±10%	±20%	Coverage
M12 hybrid comparables + GBM	9.7	8.9	19.9	-0.1	30.7	54.4	78.8	100.0
M11b gradient boosting, with assessed value	9.8	9.0	19.9	0.1	30.5	54.1	78.7	100.0
M9 optimised reconciliation	11.3	10.6	21.8	-0.8	26.2	47.6	73.0	99.8
M11a gradient boosting, no assessed value	11.5	10.7	23.1	-0.4	26.1	47.5	72.2	100.0
M0 best configuration, chosen on 2024	11.9	11.1	23.6	0.3	25.1	46.1	71.5	100.0
M13 assessment ratio	12.0	11.5	22.5	-0.1	22.1	40.6	63.9	90.3
M8b GBM adjustments	12.7	11.9	24.1	-0.6	23.9	43.6	68.8	99.8
M3 5 most similar, time only	13.9	13.0	26.2	-0.2	22.1	40.8	65.3	99.8
M7 hedonic grid	13.8	13.0	25.1	-0.5	21.7	40.4	66.4	99.8
M6 conventional grid	14.1	13.4	26.4	0.8	21.3	39.7	65.0	99.8
M8a local grid	14.3	13.5	25.7	-0.4	21.2	39.4	64.9	99.8
M10 global hedonic model	14.7	13.8	25.9	0.4	19.8	37.9	65.0	100.0
M4 5 nearest, distance-weighted	19.1	17.9	31.4	-0.2	16.0	30.7	54.3	99.8
M2 mean of 5 nearest, time only	19.3	18.1	31.8	-0.1	15.7	30.3	53.7	99.8
M5 5 nearest, recency-weighted	19.7	18.5	32.3	-0.2	15.4	29.8	52.8	99.8
M1 median of 3 nearest, time only	20.3	19.1	32.8	-0.8	15.2	28.5	49.8	96.6

Note: MdAPE, MAPE, median signed error and the shares within ±5, ±10 and ±20% in per cent, on the test targets for which the model returns a value; coverage is the share of such targets. The validation column gives the MdAPE on the 2024 targets. Weights of M9 and M12 and the configuration of M0 are chosen on 2024 only. Bootstrap confidence intervals for each entry are in Table 31.

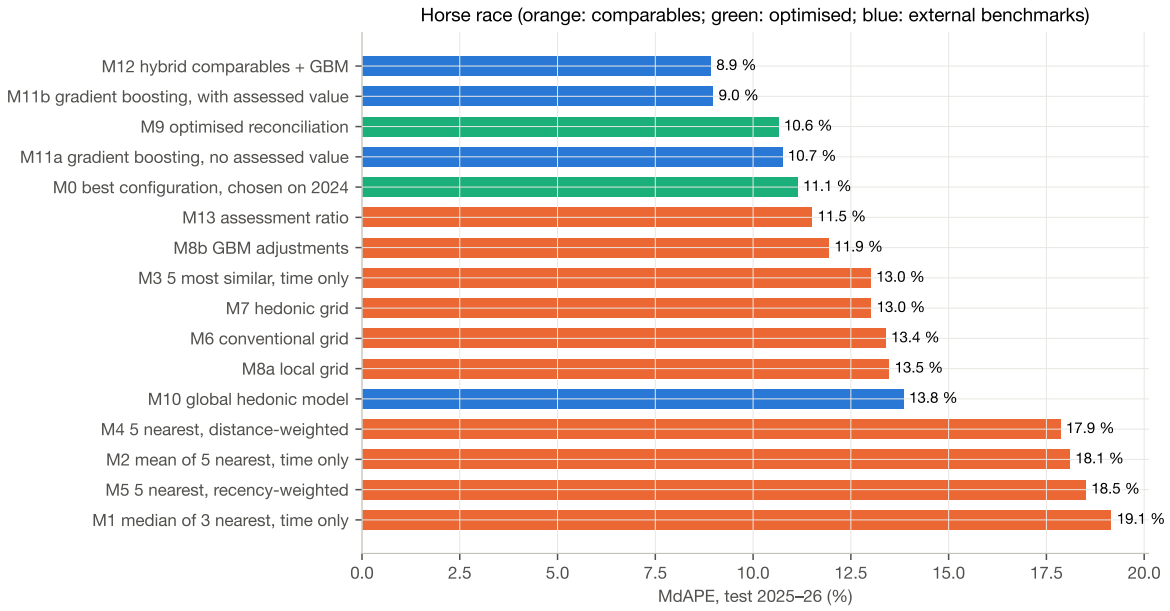

Figure 5. Out-of-sample MdAPE by model, 2025–2026. Orange: configurations of the sales comparison approach; green: empirically optimised configurations; blue: external benchmarks.

Table 15 gives the paired differences with cluster-bootstrap confidence intervals and cluster-robust t statistics for the comparisons that matter. The gain from the hedonic grid over time-only adjustment (M7 − M2, $t = -58$), the gain from similarity selection (M3 − M2), and the gain of M11b over M0 ($t = -36$) are all precisely estimated and far from zero. The difference between the conventional and hedonic grids (M7 − M6) is small but significant, as is the loss from the local grid (M8a − M7). The hybrid M12 improves on M11b by only six hundredths of a point of MdAPE, a difference that is statistically distinguishable from zero ($t = -8.3$) but economically negligible: a quarterly-retrained tree model with coordinates and the assessed value already extracts almost all of the local information that comparables carry. On the common support of the ratio method, M0 has a lower MdAPE than M13 but a higher MAPE, so that the two are not ranked by all measures.

Table 15. Paired differences in accuracy with cluster-bootstrap 95% confidence intervals

Comparison	Δ MdAPE [95% CI]	Δ MAPE [95% CI]	t	A better (%)	n	Clusters
M7 − M2	-5.08 [-5.26; -4.88]	-6.65 [-6.90; -6.43]	-57.6	61.1	124,310	6,346
M7 − M6	-0.37 [-0.45; -0.28]	-1.32 [-1.42; -1.24]	-28.4	52.3	124,310	6,346
M8a − M7	0.46 [0.33; 0.59]	0.59 [0.47; 0.72]	8.7	48.0	124,310	6,346
M8b − M7	-1.08 [-1.17; -1.00]	-1.04 [-1.11; -0.98]	-31.3	54.3	124,310	6,346
M11b − M0	-2.17 [-2.30; -2.04]	-3.69 [-3.92; -3.49]	-36.1	59.1	124,555	6,412
M11a − M0	-0.39 [-0.49; -0.32]	-0.50 [-0.57; -0.43]	-14.1	51.8	124,555	6,412
M12 − M11b	-0.06 [-0.09; -0.03]	-0.07 [-0.09; -0.06]	-8.3	52.6	124,559	6,415
M13 − M7	-1.40 [-1.58; -1.24]	-2.39 [-2.64; -2.16]	-20.2	54.5	112,423	5,986
M3 − M2	-5.10 [-5.26; -4.92]	-5.58 [-5.79; -5.38]	-52.1	59.8	124,310	6,346
M9 − M7	-2.38 [-2.51; -2.26]	-3.33 [-3.51; -3.16]	-38.3	59.7	124,310	6,346
M1 − M2	1.41 [1.26; 1.61]	1.82 [1.69; 1.97]	23.2	45.8	120,372	5,760
M0 − M13	-0.48 [-0.64; -0.32]	0.85 [0.63; 1.09]	7.4	50.1	112,423	5,986
M0 − M9	0.48 [0.36; 0.59]	1.79 [1.62; 1.96]	20.6	46.9	124,310	6,346
M0 − M7	-1.90 [-2.00; -1.80]	-1.54 [-1.64; -1.44]	-30.7	56.6	124,310	6,346
M13 − M11b	2.65 [2.51; 2.80]	2.91 [2.76; 3.07]	36.6	40.5	112,423	5,986

Note: Δ is the difference $A - B$ in MdAPE and in MAPE (percentage points) on the common test targets; a negative value means A is more accurate. Confidence intervals from 999 cluster-bootstrap replications over municipality–quarter cells; t is the cluster-robust statistic of equation (9) for the MAPE difference. "A better" is the share of targets on which A has the smaller error.

Two findings deserve emphasis. The first is the *cost of transparency*: 2.2 points of MdAPE separate the best frozen configuration of the sales comparison approach from the best opaque model with the same information plus the assessed value, and 0.4 points separate it from the same model without the assessed value. Every estimate of M0 is the median of 30 identified sales adjusted by published coefficients; every estimate of M11b is the output of several hundred trees. The second is the *value of the assessed value*. The ratio method M13 and the GBM with roll information M11b dominate their counterparts without it by 1.5 and 1.8 points respectively. The assessed value summarises characteristics that the roll extract does not expose to us—quality, condition, outbuildings—and a local price level that the assessor has already estimated; it is, in effect, a professional valuation available for every property, and the evidence says it should be used, as Thibodeau (2003) argued for marking assessed values to market.

5.6. Where does the method hold, and where does it fail?

Table 16 and Figure 6 report accuracy by market density for the main models. The dependence is steep and common to all methods: the MdAPE of the hedonic-grid method M7 is 10.2% where

two hundred or more sales occurred within 2 km in the past year and 25.3% where fewer than five did. The gap between the best transparent method and the best opaque one is widest in the sparsest markets, where the tree model's ability to borrow strength from a wider area is most valuable, and narrowest in the densest ones; the pattern is the comparable-sales counterpart of the rural accuracy penalty that Bogin and Shui (2020) document for both appraisers and AVMs. Tables 32–33 in the Appendix report the same comparison by region, price quintile, age and physical link; accuracy is highest in the dense suburban regions (Laval, Outaouais, Montérégie) and lowest in the remote regions (Gaspésie, Nord-du-Québec, Côte-Nord), in the lowest price quintile (36.7% for M7, against 10.0% in the fourth), for dwellings older than seventy years and for integrated dwellings—the segments where condition and unobserved characteristics carry the largest share of value. The lowest quintile deserves a word of caution: because quintiles are defined on the realised price, the APE of a cheap sale is mechanically inflated by any given dollar error, so part of the gradient across quintiles is a denominator effect rather than a failure of the method.

Table 16. Out-of-sample MdAPE (%) by market density

Density	n	M2	M7	M8a	M13	M0	M10	M11b	M12
< 5	9,177	32.9	25.3	25.9	19.6	22.9	23.6	15.9	16.0
5–19	15,149	27.7	20.4	20.8	16.5	17.8	18.9	13.0	13.0
20–49	15,553	22.2	15.4	15.9	13.4	13.3	15.1	10.4	10.4
50–199	56,442	16.1	11.5	11.9	10.4	9.8	12.9	8.0	8.0
≥ 200	27,989	14.0	10.2	10.7	9.3	8.5	11.3	7.4	7.3

Note: n is the number of test targets of the stratum covered by M7; in the sparsest stratum it is 9,177 of 9,426 targets because five sales within 5 km and 12 months are not always available. Each model's MdAPE is computed on the targets it covers.

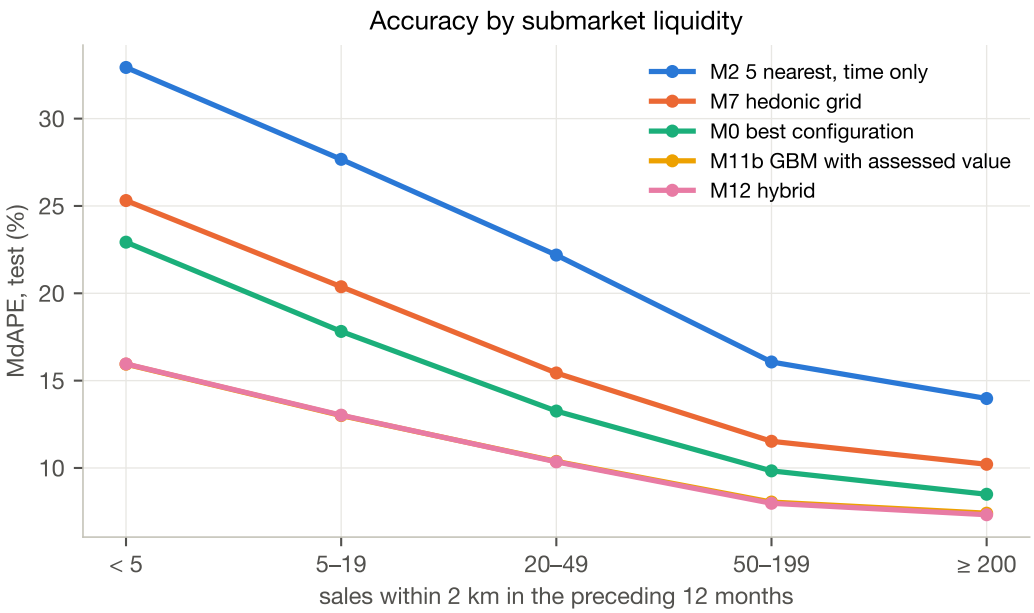

Figure 6. Out-of-sample MdAPE by market density (sales within 2 km in the preceding twelve months), selected models.

Table 17 addresses H5 by splitting the test targets into terciles of submarket heterogeneity, measured by the dispersion of the adjusted indications of the ten nearest comparables within 10 km and 24 months—a quantity available at valuation time. In the homogeneous tercile the transparent configuration M0 is within 1.4 points of M11b; in the heterogeneous tercile the gap is 3.1 points. Simple, explainable rules rival machine learning precisely where the market is informative enough for a few well-chosen sales to pin down value; machine learning earns its keep where the market is thin or heterogeneous. H5 is confirmed in direction, although the gap never closes entirely.

Table 17. Accuracy by submarket heterogeneity (H5): out-of-sample MdAPE (%)

Submarket	Median $\hat{\sigma}$	M2	M7	M8a	M13	M0	M10	M11b	M12
homogeneous	0.121	13.9	10.2	10.5	9.2	9.0	12.1	7.5	7.4
intermediate	0.252	19.0	13.9	14.3	12.1	11.9	14.3	9.5	9.5
heterogeneous	0.400	22.8	16.4	16.9	14.3	13.6	15.7	10.4	10.4

Note: terciles of the standard deviation $\hat{\sigma}$ of the log adjusted indications of the ten nearest sales within 10 km and 24 months (hedonic grid), computed on the test targets.

5.7. Value intervals

The dispersion of the adjusted indications provides an interval of value. Table 18 reports, for the time-only, hedonic and GBM adjustments at two anchors, the coverage of the interval $\hat{P}_i \exp(\pm z\,\hat{\sigma}_i)$

with $z = 1.28$ (nominal 80% under normality) and with z calibrated on the validation targets as the 80th percentile of the standardised absolute errors $|\ln P_i - \ln \hat{P}_i| / \hat{\sigma}_i$—a split-conformal calibration in the sense of Lei et al. (2018)—together with the median width of the calibrated interval. The nominal intervals under-cover (66–78% for a nominal 80%): the dispersion of five or ten adjusted sales understates the uncertainty of the estimate, because the comparables share unobserved characteristics with each other but not with the subject, which is the error-based versus model-based distinction drawn by Krause et al. (2020). Calibrated on 2024, the intervals achieve their target coverage in 2025–2026 to within half a point, but at a median width of 66.5% of value: an honest 80% interval around a comparables-based estimate spans roughly ±33%. The dispersion of the comparables is a valid but coarse measure of uncertainty, and the breadth of the calibrated interval is itself a finding about the precision the method can claim.

Table 18. Coverage of value intervals derived from the dispersion of adjusted indications

Configuration	z (calibrated on 2024)	Coverage at $z = 1.28$	Coverage, calibrated z	Median width (%)	n
time only, $k = 5$, ≤ 5 km, 12 months	1.89	67.2	80.3	90.5	123,913
time only, $k = 10$, ≤ 10 km, 24 months	1.41	77.3	80.4	85.2	124,516
hedonic grid, $k = 5$, ≤ 5 km, 12 months	1.88	67.9	79.9	65.0	123,913
hedonic grid, $k = 10$, ≤ 10 km, 24 months	1.35	78.5	80.0	66.5	124,516
GBM adjustments, $k = 5$, ≤ 5 km, 12 months	2.03	66.5	80.2	62.9	123,913
GBM adjustments, $k = 10$, ≤ 10 km, 24 months	1.44	77.3	80.5	66.7	124,516

Note: selection by distance, equal weights, median. Coverage is the share of test targets whose price lies within the interval; width is $2z\hat{\sigma}$ in per cent of value.

5.8. Towards an empirical comparable-recommendation system

The steep dependence on density suggests that the best rule might be conditional on the market. Table 19 implements the simplest version of a recommendation system: for each density stratum, the configuration (among all 2,256 that return a value for at least 95% of the stratum's validation targets) with the lowest MdAPE on 2024 is selected and applied to the 2025–2026 targets of that stratum. The selected rules do differ—the assessment-ratio adjustment is chosen in the two densest strata, the hedonic grid elsewhere, and the compromise or least-adjusted selection everywhere—but together they achieve an out-of-sample MdAPE of 11.3% with 93.4% coverage, no better than the 11.1% of the best single rule at full coverage. The reason is that the best single rule is already self-adapting: retaining the thirty most similar sales within 25 km and 24 months uses nearby sales where they exist and reaches further only where they do not. A recommendation system therefore adds value through the choice of adjustment (assessment ratio where same-roll comparables are plentiful) rather than through stratum-specific counts and radii.

Table 19. Density-conditional rules chosen on 2024 and evaluated on 2025–2026

Density	Rule	MdAPE 2024	MdAPE 2025–26	Coverage 2024	n test
< 5	compromise, $k = 10$, ≤ 10 km, 24 mo, hedonic grid, equal w., mean	24.7	23.7	99.8	9,426
5–19	compromise, $k = 30$, ≤ 10 km, 24 mo, hedonic grid, equal w., median	18.3	17.8	100.0	15,149
20–49	compromise, $k = 30$, ≤ 10 km, 24 mo, hedonic grid, equal w., median	13.9	13.3	100.0	15,553
50–199	compromise, $k = 10$, ≤ 10 km, 24 mo, assessment ratio, equal w., median	9.8	9.7	96.5	56,442
≥ 200	least adjusted, $k = 10$, ≤ 10 km, 24 mo, assessment ratio, equal w., median	8.9	8.8	95.8	27,989

6. Robustness

6.1. Stability over time and seasonality

Table 20 reports the MdAPE and bias of the main models by quarter of the test period. The rankings are stable quarter by quarter, and the validation-period (2024) and test-period (2025–2026) MdAPEs of every configuration in Tables 7–14 differ by about a point in the same direction, which indicates that the rules selected on 2024 did not overfit it; the test period was simply slightly easier for every method. The comparable-based methods display a seasonal bias: pooling the two years by calendar quarter (Table 21), the hedonic-grid method M7 is too low in the first quarter (-2.4%) and too high in the third (+1.6%), and the best configuration M0 and the direct hedonic model M10 show the same pattern (-1.6% and +2.4%; -1.2% and +2.8%). This is the signature of the point-in-time constraint: every method works at the level of the previous month, Quebec prices rise into the spring and soften after the summer (Ngai and Tenreyro, 2014), and the price level of the valuation month is not knowable on the valuation date. The gradient-boosting benchmarks do not share the pattern (M11b: +1.4% and -0.4%). The reason is not that they observe the month of sale—at prediction time their month feature is fixed at the last training month, and their level is carried forward by the same index as the comparables—and we have not identified it; it may reflect the composition of the comparable pool, which over-represents the busy spring and summer months when the target sells in winter. A seasonal correction estimated on prior years is the obvious remedy for the comparables method, and we flag it as an extension that we have not implemented.

Table 20. Stability over the test period: MdAPE and bias (%) by quarter

Quarter	n	M2	M7	M0	M11b	M12	Bias M7	Bias M11b
2025Q1	11,337	19.5	14.2	12.5	10.3	10.1	-2.4	3.0
2025Q2	24,823	17.3	12.4	10.5	8.5	8.4	-1.1	-0.3
2025Q3	20,140	18.2	13.0	11.1	8.9	8.8	1.5	-0.5
2025Q4	16,692	19.2	13.9	12.0	9.5	9.6	-0.7	-0.1
2026Q1	10,619	19.6	14.2	12.7	10.0	10.0	-2.4	-0.1
2026Q2	22,825	16.8	12.1	10.1	8.3	8.2	-1.3	0.4
2026Q3	18,123	18.3	12.9	11.1	8.8	8.8	1.7	-0.3

Table 21. Median signed error (%) by calendar quarter, test targets of 2025 and 2026 pooled

Calendar quarter	n	M2	M7	M13	M0	M10	M11a	M11b	M12
Q1	21,956	-1.1	-2.4	-2.2	-1.6	-1.2	-1.4	+1.4	+1.0
Q2	47,648	-1.2	-1.2	-0.4	-0.5	-0.9	-0.1	+0.0	-0.1
Q3	38,263	+1.7	+1.6	+1.6	+2.4	+2.8	-0.4	-0.4	-0.4
Q4	16,692	+0.2	-0.7	-1.2	+0.1	+0.7	+0.2	-0.1	-0.3

6.2. Selection stability and data snooping

The configuration M0 is the best of 1,370 configurations with at least 97% coverage on the 2024 validation targets, and its test-period accuracy is reported on sales that played no role in that choice. Two further checks establish that the choice is not fragile. First, the validation and test rankings of the 2,256 configurations are almost identical: the Spearman rank correlation between their 2024 and 2025–2026 MdAPEs is 0.996 over all configurations and 0.994 among those eligible for selection. Second, M0 ranks 2nd of 1,370 in the test period, within a hundredth of a point of the configuration that would have been chosen with hindsight (11.1%), which replaces the hedonic grid by GBM adjustments and is not transparent. Table 22 lists the ten best configurations on 2024 and their test ranks: all ten combine the compromise selection, a 24-month window and the hedonic grid, differ only in k (20 to 50) and d (2 to 25 km), and all ten rank among the twelve best out of sample. Table 23 shows that the 37 configurations within a quarter of a point of the best validation MdAPE have test MdAPEs between 11.1% and 11.4%. The accuracy attributed to M0 is therefore a property of a broad plateau of wide-net, similarity-based, adjusted configurations rather than of one lucky draw, which is what protects the headline figure from the data-snooping bias of White (2000).

Table 22. The ten best configurations on the 2024 validation targets and their out-of-sample rank

Selection	k	d (km)	t (mo)	Adjustment	Weights/recon.	MdAPE 2024	MdAPE 2025–26	Rank 2025–26
compromise	30	25	24	hedonic grid	equal/median	11.86	11.13	2
compromise	30	10	24	hedonic grid	equal/median	11.87	11.14	3
compromise	10	10	24	GBM adjustments	equal/median	11.89	11.13	1
compromise	30	5	24	hedonic grid	equal/median	11.90	11.16	5
compromise	50	10	24	hedonic grid	equal/median	11.91	11.23	12
compromise	50	25	24	hedonic grid	equal/median	11.92	11.21	11
compromise	20	25	24	hedonic grid	equal/median	11.93	11.17	7
compromise	20	5	24	hedonic grid	equal/median	11.93	11.17	6
compromise	20	10	24	hedonic grid	equal/median	11.93	11.17	8
compromise	20	2	24	hedonic grid	equal/median	11.94	11.16	4

Note: configurations with at least 97% coverage on 2024 (1,370 of 2,256). "Rank 2025–26" is the rank of the configuration's test MdAPE among the same 1,370 configurations.

Table 23. Near-optimal configurations on 2024 and the range of their out-of-sample MdAPE (%)

Validation MdAPE within (pts of best)	Configurations	Test MdAPE min	median	max
≤ 0.10	12	11.13	11.17	11.25
≤ 0.25	37	11.13	11.26	11.39
≤ 0.50	92	11.13	11.37	11.66
≤ 1.00	252	11.13	11.72	12.18

6.3. Coverage and the common-support comparison

Configurations with narrow radii or short windows abstain on a substantial share of targets, and their MdAPE on the targets they cover is not comparable to that of configurations that cover all targets. Table 27 in the Appendix reports the coverage of every point of the error surface. Our configuration selection enforces 97% coverage; relaxing the constraint selects a narrower configuration (compromise, $k = 10$, ≤ 10 km, 24 months, assessment ratio) with a lower MdAPE on its own support (10.7%) but 90.3% coverage. The conclusions of Section 5.1 hold on the common support of all configurations with at least 97% coverage, and the paired comparisons of Table 15 are always computed on the targets covered by both methods.

6.4. Exclusion rules

The exclusion of extreme price-to-assessment ratios (rule R8) is the only rule that removes a non-trivial share of sales on a criterion related to price. Its bounds affect the results through two channels: the composition of the target set, on which every method is evaluated, and the composition of the comparable pool. Table 24 examines the first channel with the existing estimates by tightening the

bounds on the targets' ratio while leaving the pool unchanged. Restricting the test targets to ratios in $[0.5, 2]$ keeps 96.1% of them and lowers every MdAPE by about half a point (M0 10.6%, M11b 8.5%); the ordering of the models and the gaps between them are unchanged, and even the drastic $[0.67, 1.5]$ window, which keeps four targets in five, preserves them. The sales removed by tighter bounds are those with the largest errors for every method, as one would expect of transactions whose price departs from any market reference. The second channel—re-running the backtest with a different pool—requires the distributed computation and was not re-executed; because the baseline bounds are wide and remove only 1.8% of the raw sales, and because an appraiser would in any case discard such sales from the comparable set, we do not expect it to alter the comparisons. The flip rule R9 removes 794 sales and cannot matter at the scale of the sample.

Table 24. Sensitivity to the price-to-assessment bounds of rule R8, test targets: MdAPE (%)

Target price/assessment ratio	n	Share (%)	M2	M7	M13	M0	M10	M11a	M11b	M12
$[0.33, 3]$ (baseline)	124,559	100.0	18.1	13.0	11.5	11.1	13.8	10.7	9.0	8.9
$[0.40, 2.5]$	123,112	98.8	17.8	12.8	11.3	11.0	13.7	10.6	8.8	8.8
$[0.50, 2.0]$	119,681	96.1	17.4	12.5	11.0	10.6	13.3	10.3	8.5	8.5
$[0.67, 1.5]$	102,333	82.2	16.3	11.5	9.8	9.7	12.4	9.4	7.5	7.4

Note: the test targets are restricted to the stated interval of the ratio of sale price to assessed value; comparable pools, grids and models are unchanged. "Share" is the share of the 124,559 test targets retained.

6.5. Alternative accuracy measures and cluster definition

The MdAPE is our headline measure because the distribution of valuation errors is heavy-tailed and the median is the appraiser's natural loss. Table 14 shows that the MAPE and the shares within ± 5, ± 10 and $\pm 20\%$ rank the models almost identically—the only inversion is between M9 and M11a, which the MAPE separates more clearly than the MdAPE—and the paired bootstrap on MAPE differences in Table 15 yields the same conclusions as on MdAPE differences. The bias column shows that no comparable-based method is materially biased out of sample once the whole test period is pooled. Table 25 re-estimates the confidence intervals of four key paired differences under alternative cluster definitions, from municipality–quarter cells (the baseline, with several thousand clusters) to region–quarter cells (about a hundred). The intervals widen as clusters become coarser, as they should when dependence at wider scales is allowed for, but none of the conclusions changes: the smallest difference that we call significant, M12 − M11b, remains distinguishable from zero even with region–quarter clusters.

Table 25. Sensitivity of the bootstrap confidence intervals to the cluster definition: 95% intervals of the MdAPE difference (points)

Cluster definition	Clusters	M7 − M2	M11b − M0	M12 − M11b	M7 − M6
municipality × quarter	6,346	[-5.28; -4.89]	[-2.29; -2.03]	[-0.09; -0.03]	[-0.45; -0.28]
MRC × quarter	605	[-5.79; -4.73]	[-2.72; -1.91]	[-0.10; -0.03]	[-0.56; -0.26]
municipality	1,079	[-5.40; -4.79]	[-2.43; -1.93]	[-0.09; -0.03]	[-0.47; -0.26]
region × quarter	119	[-5.30; -4.88]	[-2.37; -1.97]	[-0.09; -0.02]	[-0.48; -0.27]

Note: 399 cluster-bootstrap replications for each alternative definition (999 for the baseline, Table 15).

6.6. Index construction

The point-in-time index uses an expanding window of all prior sales and municipality fixed effects. Because comparables are at most 24 months old and the time adjustment contributes only 0.3–1.4 points of accuracy (Section 5.2), the comparable-based results are insensitive to the details of the index. The global benchmarks are more exposed: their level in 2025–2026 is carried entirely by the index from the last training quarter to the valuation month, and we verified that an index without municipality fixed effects, which is contaminated by the shift of sales towards less expensive municipalities, lowers their level.

6.7. Benchmark specification

The gradient-boosting benchmarks are trained with an L_1 objective so that they estimate the conditional median, which is the quantity the MdAPE rewards; the same models trained with a squared-error objective have a median bias of about -5% in-sample and out of sample, because the distribution of log prices conditional on characteristics is left-skewed, and a correspondingly higher MdAPE; the direct hedonic model M10, estimated by least squares, is for the same reason recentred on the median of its validation residuals. The number of trees chosen by early stopping on the last 10% of the training window is floored at 300 because the final, partial months of the sample make that slice unreliable. The hybrid weights of M12 and the reconciliation weights of M9 are convex combinations fitted on 2024 and applied unchanged; M9 places 0.59 of its weight on the assessment-ratio indication and 0.33 on the GBM-adjusted one, and M12 places 0.91 on M11b. Neither combination was tuned beyond the choice of its components.

7. Conclusion

The sales comparison approach is usually defended on the grounds of its logic and its transparency and criticised on the grounds of its subjectivity. This paper has asked a different question: taken as a

set of explicit rules, how accurate is it, and which rules are empirically justified? Using the universe of single-family transactions in Quebec and a backtest that respects the information available on each valuation date, we have measured the accuracy of 2,256 configurations of the method and of the natural alternatives on 124,559 out-of-sample sales, with configurations chosen on a separate year and differences tested against the spatial and temporal dependence of valuation errors.

The answers are clear enough to be taught. Adjust: a grid—even a crude one—is worth four to six points of error. Choose similar comparables, not merely near ones, and search widely in space and time. Use more comparables than three: twenty or more when a grid is applied, in sparse and dense markets alike, and about ten when it is not. Do not agonise over weights or over the choice between median and mean. Use the assessed value: the ratio method is the most accurate single adjustment scheme where it applies. And report an interval, calibrated so that it means what it says—and be prepared for it to be wide. A transparent method configured on these principles and frozen in 2024 reaches a median absolute error of 11.1% on 2025–2026 sales, within 2.2 points of the best quarterly-retrained machine-learning model given the same data, and within about a point and a half in homogeneous submarkets. The choice of that configuration is robust: the ten best configurations on 2024 differ only in the number of comparables and the radius, and all rank among the twelve best out of sample.

The results have implications for practice, for education and for regulation. For practice, they suggest that the appraiser's effort is best spent on selecting similar comparables and on estimating adjustments from the largest homogeneous pool of evidence, rather than on reconciling a handful of nearby sales by judgement. For education, they provide a set of validated rules that can be taught alongside the conventions, and a laboratory—the backtesting engine—in which students can test their own. For regulation and for the design of automated valuation models, they show that a comparable-based method can be made nearly as accurate as an opaque model in most of the market while remaining fully explainable, and that a wide-net rule retaining the most similar sales adapts to market density on its own.

Four limitations bound the conclusions. We simulate rules; we do not observe appraisers, whose files may contain information (condition, renovations, listing history) that would improve on any rule, and whose judgement may also degrade it. We observe the characteristics of the assessment roll, not those of listings or inspections, so that adjustments for garages, pools, bedrooms and quality are untested and their effects remain in the residual error, which the within-municipality R^2 of about a third makes visible. We screen non-arm's-length transfers with a price-to-assessment ratio, date sales by their registration and ignore the publication lag of the register, three choices that favour every method alike but that a user of the rules in real time would have to confront. And we study single-family dwellings in one province; condominiums, income properties and other markets may yield different optima, though we expect the qualitative hierarchy—adjustment and selection

first, number and radius second, weighting last—to carry over. Two extensions follow directly from the results and remain to be done: a seasonal correction of the comparables estimate, which Section 6.1 suggests would remove a bias of one to two points in the first and third quarters, and a re-execution of the backtest under alternative screening rules for the comparable pool. A companion project develops the recommendation system of Section 5.8 into a transparent, empirically optimised comparable-sales algorithm and integrates it into a teaching platform.

References

Appraisal Institute, *The Appraisal of Real Estate*, 15th ed., Chicago: Appraisal Institute, 2020.

Bailey, Martin J., Richard F. Muth, and Hugh O. Nourse, "A Regression Method for Real Estate Price Index Construction," *Journal of the American Statistical Association*, 1963, *58* (304), 933–942.

Bogin, Alexander N. and Jessica Shui, "Appraisal Accuracy and Automated Valuation Models in Rural Areas," *Journal of Real Estate Finance and Economics*, 2020, *60* (1–2), 40–52.

Bourassa, Steven C., Eva Cantoni, and Martin Hoesli, "Spatial Dependence, Housing Submarkets, and House Price Prediction," *Journal of Real Estate Finance and Economics*, 2007, *35* (2), 143–160.

_ , _ , and _ , "Predicting House Prices with Spatial Dependence: A Comparison of Alternative Methods," *Journal of Real Estate Research*, 2010, *32* (2), 139–160.

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

Calem, Paul S., Jeanna Kenney, Lauren Lambie-Hanson, and Leonard I. Nakamura, "Appraising Home Purchase Appraisals," *Real Estate Economics*, 2021, *49* (S1), 134–168.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller, "Bootstrap-Based Improvements for Inference with Clustered Errors," *Review of Economics and Statistics*, 2008, *90* (3), 414–427.

Cannaday, Roger E. and Mark A. Sunderman, "Estimation of Depreciation for Single-Family Appraisals," *Real Estate Economics*, 1986, *14* (2), 255–273.

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.

Case, Karl E. and Robert J. Shiller, "The Efficiency of the Market for Single-Family Homes," *American Economic Review*, 1989, *79* (1), 125–137.

Clapham, Eric, Peter Englund, John M. Quigley, and Christian L. Redfearn, "Revisiting the Past and Settling the Score: Index Revision for House Price Derivatives," *Real Estate Economics*, 2006, *34* (2), 275–302.

Clapp, John M. and Carmelo Giaccotto, "Estimating Price Indices for Residential Property: A Comparison of Repeat Sales and Assessed Value Methods," *Journal of the American Statistical Association*, 1992, *87* (418), 300–306.

Colwell, Peter F., Roger E. Cannaday, and Chunchi Wu, "The Analytical Foundations of Adjustment Grid Methods," *Real Estate Economics*, 1983, *11* (1), 11–29.

Des Rosiers, François, 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.

Diaz, Julian, "How Appraisers Do Their Work: A Test of the Appraisal Process and the Development of a Descriptive Model," *Journal of Real Estate Research*, 1990, *5* (1), 1–15.

_ **and J. Andrew Hansz**, "The Use of Reference Points in Valuation Judgment," *Journal of Property Research*, 2001, *18* (2), 141–148.

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

Dubé, Jean and Diègo Legros, "A Spatio-Temporal Measure of Spatial Dependence: An Example Using Real Estate Data," *Papers in Regional Science*, 2013, *92* (1), 19–31.

_ **and** _ , "Spatial Econometrics and the Hedonic Pricing Model: What About the Temporal Dimension?," *Journal of Property Research*, 2014, *31* (4), 333–359.

Epley, Donald R., "A Note on the Optimal Selection and Weighting of Comparable Properties," *Journal of Real Estate Research*, 1997, *14* (2), 175–182.

Eriksen, Michael D., Hamilton B. Fout, Mark Palim, and Eric Rosenblatt, "The Influence of Contract Prices and Relationships on Appraisal Bias," *Journal of Urban Economics*, 2019, *111*, 132–143.

Gau, George W., Tsong-Yue Lai, and Ko Wang, "Optimal Comparable Selection and Weighting in Real Property Valuation: An Extension," *Real Estate Economics*, 1992, *20* (1), 107–123.

_ **,** _ **, and** _ , "A Further Discussion of Optimal Comparable Selection and Weighting, and a Response to Green," *Real Estate Economics*, 1994, *22* (4), 655–663.

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

Gloudemans, Robert J., *Mass Appraisal of Real Property*, Chicago: International Association of Assessing Officers, 1999.

Gneiting, Tilmann, "Making and Evaluating Point Forecasts," *Journal of the American Statistical Association*, 2011, *106* (494), 746–762.

Goodman, Allen C. and Thomas G. Thibodeau, "Housing Market Segmentation and Hedonic Prediction Accuracy," *Journal of Housing Economics*, 2003, *12* (3), 181–201.

Green, Richard K., "Optimal Comparable Weighting and Selection: A Comment," *Real Estate Economics*, 1994, *22* (4), 647–654.

Hansen, Peter Reinhard, "A Test for Superior Predictive Ability," *Journal of Business & Economic Statistics*, 2005, *23* (4), 365–380.

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.

Hyndman, Rob J. and Anne B. Koehler, "Another Look at Measures of Forecast Accuracy," *International Journal of Forecasting*, 2006, *22* (4), 679–688.

International Association of Assessing Officers, "Standard on Mass Appraisal of Real Property," Technical Report, IAAO, Kansas City 2017.

_ , "Standard on Automated Valuation Models (AVMs)," Technical Report, IAAO, Kansas City 2018.

Isakson, Hans R., "The Nearest Neighbors Appraisal Technique: An Alternative to the Adjustment Grid Methods," *Real Estate Economics*, 1986, *14* (2), 274–286.

_ , "Valuation Analysis of Commercial Real Estate Using the Nearest Neighbors Appraisal Technique," *Growth and Change*, 1988, *19* (2), 11–24.

_ , "The Linear Algebra of the Sales Comparison Approach," *Journal of Real Estate Research*, 2002, *24* (2), 117–128.

Kang, Han-Bin and Alan K. Reichert, "An Empirical Analysis of Hedonic Regression and Grid-Adjustment Techniques in Real Estate Appraisal," *Real Estate Economics*, 1991, *19* (1), 70–91.

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.

Krause, Andy, Andrew Martin, and Miriam Fix, "Uncertainty in Automated Valuation Models: Error- versus Model-Based Approaches," *Journal of Property Research*, 2020, *37* (4), 308–339.

Lai, Tsong-Yue, Kerry Vandell, Ko Wang, and Gerd Welke, "Estimating Property Values by Replication: An Alternative to the Traditional Grid and Regression Methods," *Journal of Real Estate Research*, 2008, *30* (4), 441–460.

Lei, Jing, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman, "Distribution-Free Predictive Inference for Regression," *Journal of the American Statistical Association*, 2018, *113* (523), 1094–1111.

Lentz, George H. and Ko Wang, "Residential Appraisal and the Lending Process: A Survey of Issues," *Journal of Real Estate Research*, 1998, *15* (1), 11–39.

Lipscomb, Joseph B. and J. Brian Gray, "An Empirical Investigation of Four Market-Derived Adjustment Methods," *Journal of Real Estate Research*, 1990, *5* (1), 53–66.

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.

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.

Ngai, L. Rachel and Silvana Tenreyro, "Hot and Cold Seasons in the Housing Market," *American Economic Review*, 2014, *104* (12), 3991–4026.

Northcraft, Gregory B. and Margaret A. Neale, "Experts, Amateurs, and Real Estate: An Anchoring-and-Adjustment Perspective on Property Pricing Decisions," *Organizational Behavior and Human Decision Processes*, 1987, *39* (1), 84–97.

Pace, R. Kelley, "Relative Performance of the Grid, Nearest Neighbor, and OLS Estimators," *Journal of Real Estate Finance and Economics*, 1996, *13* (3), 203–218.

_ , "Appraisal Using Generalized Additive Models," *Journal of Real Estate Research*, 1998, *15* (1), 77–99.

_ **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.

_ **and** _ , "Generalizing the OLS and Grid Estimators," *Real Estate Economics*, 1998, *26* (2), 331–347.

Pagourtzi, Elli, Vassilis Assimakopoulos, Thomas Hatzichristos, and Nick French, "Real Estate Appraisal: A Review of Valuation Methods," *Journal of Property Investment & Finance*, 2003, *21* (4), 383–401.

Quan, Daniel C. and John M. Quigley, "Price Formation and the Appraisal Function in Real Estate Markets," *Journal of Real Estate Finance and Economics*, 1991, *4* (2), 127–146.

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

Shiller, Robert J. and Allan N. Weiss, "Evaluating Real Estate Valuation Systems," *Journal of Real Estate Finance and Economics*, 1999, *18* (2), 147–161.

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.

Vandell, Kerry D., "Optimal Comparable Selection and Weighting in Real Property Valuation," *Real Estate Economics*, 1991, *19* (2), 213–239.

White, Halbert, "A Reality Check for Data Snooping," *Econometrica*, 2000, *68* (5), 1097–1126.

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.

Appendices

A. Additional tables

Table 26. Hedonic price index (descriptive, full sample): cumulative growth by region (%)

Region	2021-01 to 2026-09 (%)	Peak	Trough	Peak to trough (%)	2024-12 to 2026-09 (%)
Quebec (province)	79.0	2022-06	2023-02	-13.1	19.6
Montréal	33.2	2022-06	2023-03	-11.9	3.6
Capitale-Nationale	81.0	2022-08	2023-02	-14.1	25.4
Montérégie	66.4	2022-06	2023-02	-18.0	13.9
Laval	48.3	2022-06	2023-02	-16.2	12.6
Outaouais	68.6	2022-06	2023-02	-11.5	13.9
Laurentides	73.6	2022-06	2023-02	-14.3	16.1
Lanaudière	88.0	2022-06	2023-02	-15.4	20.1
Estrie	112.0	2022-08	2023-02	-15.4	26.1

Note: monthly effects of a within-municipality hedonic regression estimated on the full sample (Figure 1); the peak is the highest month before 2023 and the trough the lowest month between the peak and December 2023.

Table 27. Coverage of the error surface: share of test targets (%) for which a value is returned, by radius and window

Radius	3 months	6 months	12 months	24 months
≤ 0.5 km	80.0	87.1	92.2	95.0
≤ 1 km	89.2	93.0	96.0	97.5
≤ 2 km	94.2	96.6	98.3	99.1
≤ 5 km	98.5	99.4	99.8	99.9
≤ 10 km	99.7	99.9	100.0	100.0
≤ 25 km	100.0	100.0	100.0	100.0

Note: selection by distance, $k = 5$, hedonic grid; coverage depends only on the radius and the window.

Table 28. Error surface without physical adjustment: out-of-sample MdAPE (%), selection by distance, time adjustment only, twelve-month window

Radius	$k=1$	$k=3$	$k=5$	$k=7$	$k=10$	$k=15$	$k=20$	$k=30$	$k=50$
≤ 0.5 km	20.1	17.4	16.9	16.7	**16.7**	16.7	16.8	17.0	17.0
≤ 1 km	20.7	17.8	17.2	17.0	**17.0**	17.1	17.3	17.6	18.0
≤ 2 km	21.0	18.0	17.5	17.2	**17.2**	17.4	17.5	17.8	18.3
≤ 5 km	21.3	18.2	17.6	17.4	**17.3**	17.5	17.6	18.0	18.5
≤ 10 km	21.3	18.2	17.6	17.4	**17.3**	17.5	17.6	18.0	18.4
≤ 25 km	21.3	18.2	17.6	17.4	**17.3**	17.5	17.6	18.0	18.4

Table 29. Error surface with selection by similarity: out-of-sample MdAPE (%), hedonic grid, twelve-month window

Radius	$k=1$	$k=3$	$k=5$	$k=7$	$k=10$	$k=15$	$k=20$	$k=30$	$k=50$
≤ 0.5 km	14.8	12.6	12.0	11.8	**11.7**	11.8	11.9	12.0	12.1
≤ 1 km	15.2	12.7	12.0	11.8	**11.6**	11.6	11.6	11.7	11.9
≤ 2 km	15.5	12.9	12.1	11.8	11.5	11.5	**11.4**	11.4	11.6
≤ 5 km	16.1	13.3	12.5	12.1	11.7	11.6	**11.6**	11.6	11.6
≤ 10 km	16.4	13.6	12.7	12.3	12.0	11.8	11.7	**11.7**	11.8
≤ 25 km	16.6	13.7	12.8	12.4	12.1	12.0	11.9	**11.9**	12.0

Table 30. Best configuration of the error surface by selection and adjustment (chosen on 2024, evaluated on 2025–2026)

Selection	Adjustment	k	d (km)	t (months)	MdAPE 2024	MdAPE 2025–26	Coverage
nearest	hedonic grid	30	1	24	12.6	12.0	97.5
nearest	time only	10	1	24	17.3	16.3	97.5
compromise	hedonic grid	30	25	24	11.9	11.1	100.0
compromise	time only	10	25	24	12.8	12.0	100.0
most similar	hedonic grid	20	2	24	12.0	11.2	99.1
most similar	time only	20	25	24	12.7	11.8	100.0

Note: best configuration among those with at least 97% coverage on the validation targets.

Table 31. Horse race with cluster-bootstrap 95% confidence intervals, 2025–2026

Model	MdAPE	95% CI	MAPE	95% CI	*n*
M12	8.91	[8.73; 9.13]	19.85	[19.40; 20.37]	124,559
M11b	8.97	[8.77; 9.20]	19.93	[19.46; 20.45]	124,559
M9	10.64	[10.38; 10.92]	21.79	[21.28; 22.33]	124,310
M11a	10.74	[10.43; 11.04]	23.12	[22.53; 23.76]	124,559
M0	11.13	[10.84; 11.45]	23.62	[23.04; 24.25]	124,555
M13	11.50	[11.22; 11.80]	22.48	[21.98; 23.03]	112,423
M8b	11.93	[11.63; 12.28]	24.08	[23.49; 24.70]	124,310
M3	13.00	[12.65; 13.40]	26.19	[25.52; 26.90]	124,310
M7	13.01	[12.71; 13.35]	25.12	[24.52; 25.75]	124,310
M6	13.38	[13.06; 13.74]	26.44	[25.78; 27.16]	124,310
M8a	13.47	[13.12; 13.86]	25.71	[25.12; 26.33]	124,310
M10	13.84	[13.46; 14.31]	25.90	[25.28; 26.54]	124,559
M4	17.85	[17.44; 18.32]	31.39	[30.71; 32.18]	124,310
M2	18.09	[17.67; 18.55]	31.77	[31.07; 32.57]	124,310
M5	18.50	[18.06; 18.99]	32.29	[31.58; 33.09]	124,310
M1	19.15	[18.69; 19.68]	32.80	[32.12; 33.58]	120,372

Note: MdAPE and MAPE in per cent on the test targets covered by each model; intervals from 999 cluster-bootstrap replications over municipality–quarter cells. Intervals of different models overlap even where their paired difference is significant, because the paired comparison removes the target-specific component of the error.

Table 32. Out-of-sample MdAPE (%) by administrative region

Region	*n*	M2	M7	M8a	M13	M0	M10	M11b	M12
Montérégie	25,133	16.1	11.3	11.9	10.5	9.5	11.7	8.0	7.9
Laurentides	13,629	18.0	12.8	12.8	11.4	10.6	12.9	8.5	8.5
Capitale-Nationale	11,711	18.7	13.4	14.2	11.4	11.6	13.8	9.1	9.1
Lanaudière	11,377	17.2	12.9	12.6	11.6	10.6	13.0	8.6	8.6
Montréal	9,496	16.0	12.0	13.5	11.2	10.7	21.4	9.2	9.1
Estrie	9,437	21.0	14.6	15.2	12.7	12.5	14.7	9.7	9.6
Chaudière-Appalaches	8,135	22.0	15.5	16.6	13.4	13.6	14.7	10.4	10.4
Outaouais	7,426	14.3	10.6	9.8	9.4	9.1	13.0	7.6	7.5
Laval	5,243	14.5	9.6	10.3	8.1	8.1	10.6	6.9	7.0
Saguenay–Lac-Saint-Jean	5,030	20.7	16.0	16.0	12.8	14.2	15.6	10.2	10.2
Centre-du-Québec	4,681	20.5	13.7	14.0	12.4	12.1	13.1	9.6	9.6
Mauricie	4,670	21.6	16.3	16.7	14.3	14.2	16.4	11.4	11.3
Bas-Saint-Laurent	3,260	23.3	18.2	18.7	15.4	15.9	16.5	12.3	12.2
Abitibi-Témiscamingue	2,412	22.2	17.0	16.6	12.4	15.3	17.1	9.8	10.0
Côte-Nord	1,227	23.9	19.3	19.5	16.9	17.7	19.1	13.7	13.6
Gaspésie–Îles-de-la-Madeleine	1,157	29.0	24.0	23.5	22.6	21.5	22.9	17.1	17.4
Nord-du-Québec	286	27.1	21.8	22.4	15.7	20.7	22.0	13.9	13.6

Table 33. Out-of-sample MdAPE (%) by price quintile, age and physical link

Price quintile	n	M2	M7	M8a	M13	M0	M10	M11b	M12
Q1 (low)	21,659	50.4	36.7	37.0	28.7	35.8	36.7	26.4	26.2
Q2	24,680	17.5	13.7	14.2	12.7	11.7	13.5	9.4	9.4
Q3	23,834	13.6	10.4	10.6	9.7	9.0	11.5	7.1	7.1
Q4	26,783	13.3	10.0	10.1	8.9	8.2	11.3	6.7	6.7
Q5 (high)	27,354	19.3	11.6	12.4	9.7	10.1	13.2	8.2	8.2
Age (years)	n	M2	M7	M8a	M13	M0	M10	M11b	M12
≤ 3 years	2,373	20.1	11.9	12.1	9.8	9.8	12.9	8.0	7.8
4–15	17,344	17.0	9.9	10.1	8.8	8.4	12.1	6.7	6.6
16–40	39,475	16.6	11.1	11.3	9.8	9.2	11.9	7.5	7.4
41–70	48,592	17.7	14.4	14.9	13.0	12.5	14.5	10.2	10.2
> 70	16,526	25.1	19.7	21.6	16.9	18.0	20.9	14.2	14.2
Physical link	n	M2	M7	M8a	M13	M0	M10	M11b	M12
detached	100,999	19.6	14.3	14.7	12.4	12.2	14.2	9.8	9.7
semi-detached	15,936	12.9	9.1	9.4	8.6	7.7	11.8	6.6	6.4
row	7,297	11.9	8.2	8.6	8.0	7.8	14.5	6.2	6.1
integrated	78	34.3	22.3	18.0	10.3	15.3	31.0	10.9	9.9

Note: price quintiles are defined on the realised sale price of the test targets; see Section 5.6 on the denominator effect in the lowest quintile.

B. Implementation details

Computation. The backtest evaluates 2,256 configurations on 189,354 targets. For each target the candidate pool is retrieved with a k-d tree on projected coordinates, the time, conventional, hedonic, regional, local, GBM and ratio adjustments are computed once for every candidate, and the selection–window–count–weighting–reconciliation combinations are evaluated as operations on the resulting vectors. The computation is embarrassingly parallel over targets and was distributed over five Apple-silicon nodes of a local cluster; the backtest itself takes under ten minutes once the index and the quarterly models are estimated, and the full chain (sample construction, index, eleven quarterly re-estimations of the grids and gradient-boosting models, backtest, analysis) about half an hour. All code is in Python (pandas, NumPy, SciPy, LightGBM) and is available from the author; the analyses of Sections 6.2, 6.4 and 6.5, the seasonal decomposition and the per-model confidence intervals are computed from the stored per-target estimates by a separate script that runs in about a minute.

Local grid. The local regression (adjustment scheme f) estimates, on the subject's candidates within 10 km and 24 months, the coefficients of log living area, log lot area, age, the two attached-dwelling indicators and the number of storeys, by ridge regression with penalty $\lambda = 50$ towards the regional grid (equation 5); with fewer than 40 candidates the regional grid is used.

Gradient boosting. LightGBM with 127 leaves, minimum 40 observations per leaf, learning rate 0.05, feature and bagging fractions 0.8, L_2 regularisation 1, L_1 objective, number of trees by early stopping (patience 100) on the chronologically last 10% of the training sample with a floor of 300 trees, then refit on the whole training sample. Features: log living area, log lot area, age, age^2, physical-link and construction-type indicators, storeys, new-build indicator, latitude, longitude, projected coordinates, region, building type, month of sale (set to the last training month when predicting), and for M11b the log assessed value, its land share and its roll vintage. The target is the log price brought to the level of the last training month by the point-in-time index; predictions are carried to the level of the month preceding the valuation by the index known at valuation time. The GBM adjustments of scheme (g) use the model without the assessed value.

Cluster bootstrap. Municipality–quarter cells of the test period (several thousand cells) are resampled with replacement 999 times. For each replication, the resampled cells define integer multiplicities for every target; the mean of the paired APE differences and the two medians are recomputed with these multiplicities as weights, so that all models and all pairs are evaluated on the same replications. The cluster-robust t statistic of equation (9) uses the same cells. The sensitivity analysis of Table 25 repeats the procedure with MRC–quarter, municipality and region–quarter cells and 399 replications.