Beyond Hedonic Models: Spatiotemporal, Hierarchical and Probabilistic Architectures for Residential Real Estate Price Forecasting

A pseudo-real-time evaluation on 632,000 Quebec transactions and 82 local market indices, 2021–2026

[Author identity withheld for review]

October 2026

Abstract

This paper asks whether new forecasting architectures improve residential price forecasts over strong conventional methods, and under which conditions. Using the audited Quebec land-registry sample of UQO Working Paper 22 (632,006 sales, 2021–2026) and its real-time quality-adjusted index vintages, we define three tasks in a strict pseudo-real-time design: the price of a property sold one to twelve months after a forecast origin (Task A), the one- to twelve-month change of municipal, regional and provincial quality-adjusted indices (Task B), and calibrated predictive distributions for both (Task C). We implement five candidate architectures (a hierarchical spatiotemporal hedonic network, a dynamic spatial graph forecasting network, a regime-adaptive mixture of experts, a Bayesian hierarchical probabilistic forecaster and a quality-adjusted neural state-space model) and an integrated stacked architecture, against 14 benchmarks ranging from no-change and momentum rules to gradient boosting, recurrent, convolutional and attention networks. Evaluation uses Diebold–Mariano and Giacomini–White tests, model confidence sets and block bootstraps, with Monte Carlo experiments on synthetic markets whose latent dynamics range from unforecastable random walks to regime switching and regional spillovers. For the price of a future sale, gradient boosting on hedonic and real-time local market features is the most accurate model at every horizon and the hierarchical network does not improve on it; for index changes, one- to six-month movements are forecastable through the cell's own momentum and seasonality (MASE 0.75 at three months for a pooled tree model), spatial, regime and hierarchical components add nothing that survives ablation, and at twelve months nothing beats an autoregression; the state-space forecaster is the only model with calibrated intervals at every horizon. Every component of the architectures exists in the literature; the contribution is the measured answer, on transaction-level data, to the question of when complexity pays.

Keywords: house price forecasting, machine learning, graph neural networks, state-space models, hierarchical Bayes, regime switching, probabilistic forecasting, Quebec

JEL Classification: R31, C53, C45, C11, C21, C55

1. Introduction

Forecasting residential prices is a different problem from measuring them. A price index answers how much the price of a constant-quality dwelling has changed; a forecast answers how much it will change, or what a given dwelling will sell for, using only what is known at the forecast origin. The two problems share the data and the quality-adjustment apparatus, but they differ in what can go wrong: an index can be revised, a forecast cannot be, and an index is judged against a latent truth while a forecast is judged against an outcome that arrives later. This paper takes the transaction database, the geography and the real-time index vintages constructed in UQO Working Paper 22 (Boucher, 2026) and asks whether forecasting architectures that combine property-level characteristics, spatial dependence, hierarchical pooling, regime adaptation and probabilistic output improve on conventional forecasts, and when.

The recent literature offers many candidate mechanisms and few controlled comparisons. Gradient boosting dominates cross-sectional valuation contests (Mayer et al., 2019; Hong et al., 2020), but valuation is not forecasting; cross-regional spatial dependence of price dynamics is well documented (Pace et al., 2000; Holly et al., 2010; Brady, 2011), and graph neural networks designed for traffic forecasting (Li et al., 2018; Wu et al., 2019; Bai et al., 2020) offer a flexible way to learn it, but the housing applications so far value properties cross-sectionally (Zhang et al., 2021; Das et al., 2021); regime switching in house price dynamics is classical (Hamilton, 1989; Capozza et al., 2004); hierarchical Bayesian pooling of local indices is established (Francke and van de Minne, 2017); neural state-space models exist in the forecasting literature (Rangapuram et al., 2018); and simple linear models have been shown to beat elaborate transformers on many benchmarks (Zeng et al., 2023). What is missing is an evaluation of these mechanisms on the same transaction-level data, in a design that uses only information available at each origin, with tests that account for overlapping horizons and serial dependence, and with a simulation study that explains why a mechanism helps when it does.

We make four contributions. First, we define three forecasting tasks on an audited registry sample and construct, for every monthly origin from December 2022 to September 2025, the information set actually available then: property characteristics, real-time municipal market features computed from past sales, and the real-time quality-adjusted index vintages of WP22, which were re-estimated at every publication date so that no full-sample index enters a predictor. Second, we implement five candidate architectures and an integrated architecture alongside 14 benchmarks, from no-change and momentum rules through ARIMA-type, factor and state-space models to gradient boosting and recurrent, convolutional and attention networks, each with a quantile or distributional output. Third, we evaluate them with forward-chaining experiments and the inferential toolkit of the forecasting literature: Diebold–Mariano tests with horizon-adjusted long-run variances (Diebold

and Mariano, 1995; Harvey et al., 1997), Giacomini–White conditional tests (Giacomini and White, 2006), model confidence sets (Hansen et al., 2011), stationary block bootstraps (Politis and Romano, 1994), scale-free accuracy (Hyndman and Koehler, 2006), and proper scoring rules for the predictive distributions (Gneiting and Raftery, 2007; Bracher et al., 2021). Fourth, we run the same forecasters on synthetic markets whose latent dynamics are known, from unforecastable random walks to persistent momentum, Markov-switching regimes and regional spillovers, so that the empirical ranking can be read against the mechanisms that generate it.

We are explicit about novelty, following the verified literature review of Section 2: each architecture is a combination of known components, and we classify each as an application, an adaptation or a new combination. The answer to the title question is largely negative: gradient boosting on hedonic and real-time market features dominates the property-level task; index changes are forecastable at one to six months through the cell's own momentum and seasonality, which a pooled tree model or an autoregression captures, while the spatial graph, regime gating and hierarchical intercepts add nothing that survives ablation; the neural sequence models are no better than the no-change forecast in a 45-origin sample and vary across seeds as much as across architectures; and at twelve months every pooled model inherits the bias of the 2021–2023 boom. Two components earn their place: the quality-adjusted state-space forecaster, stable and calibrated, and an equal-weight combination of components, which beats a learned stack.

The paper proceeds as follows. Section 2 reviews the literature and states the gap. Section 3 describes the data and the institutional context. Section 4 defines the forecasting targets and information sets. Section 5 presents the architectures and Section 6 the benchmarks. Section 7 reports the simulation study. Section 8 sets out the empirical forecasting framework. Sections 9 to 13 report the main results, spatial and regional heterogeneity, probabilistic forecasting and calibration, interpretation, and ablations and robustness. Section 14 concludes with limitations.

2. Related literature

Hedonic models, repeat sales and index construction. Quality adjustment rests on hedonic regression (Rosen, 1974) and on repeat-sales methods (Bailey et al., 1963; Case and Shiller, 1989); the index methods relevant here, including rolling time dummies (Hill et al., 2022), multilateral time-series smoothing (Ishaak et al., 2024) and hierarchical structural time-series models for thin markets (Francke, 2010; Francke and van de Minne, 2017), are reviewed in Hill (2013), the Eurostat et al. handbook (Eurostat et al., 2013) and WP22. Sijp and Francke (2026) model spatiotemporal random effects in hedonic and repeat-sales models through graph-Laplacian priors, the parametric counterpart of the spatial sharing that the neural architectures below attempt to learn. Selection into transaction is the subject of Heckman (1979), Gatzlaff and Haurin (1997, 1998), Jud and Seaks

(1994), Munneke and Slade (2000), Clapp et al. (1991) and Hwang and Quigley (2004).

Forecasting housing markets. Predictability of house price changes is documented by Case and Shiller (1989, 1990); momentum and mean reversion by Capozza et al. (2004), Beracha and Skiba (2011), Titman et al. (2014) and, theoretically, Glaeser et al. (2014), Head et al. (2014) and Guren (2018). Regional index forecasting with cross-sectional information includes Rapach and Strauss (2009), Gupta et al. (2011b,a) (factor-augmented and large Bayesian VARs), Bork and Møller (2015) (dynamic model averaging), Kouwenberg and Zwinkels (2014), Plakandaras et al. (2015), Milunovich (2020) (time series against machine learning for Australia) and Yeo et al. (2025) (dynamic-factor VAR with deep learning); Ghysels et al. (2013) survey the field. Spatial and spatiotemporal dependence of prices is modelled by Pace et al. (1998, 2000), Holly et al. (2010, 2011) and Brady (2011); the spatial-econometric toolkit is Anselin (1988), LeSage and Pace (2009), Kelejian and Prucha (2010) and Elhorst (2014).

Machine learning and deep learning for property prices. Tree ensembles (Breiman, 2001; Friedman, 2001; Chen and Guestrin, 2016; Ke et al., 2017; Prokhorenkova et al., 2018) dominate valuation comparisons (Mayer et al., 2019; Hong et al., 2020; Baldominos et al., 2018; Kok et al., 2017); Pace and Hayunga (2020) examine what trees extract from hedonic residuals; Deppner and Cajias (2024) show that random cross-validation is biased by spatial dependence; interpretability is treated by Lorenz et al. (2023) and Lenaers and De Moor (2023). Graph neural networks (Kipf and Welling, 2017; Veličković et al., 2018; Wu et al., 2021) have been applied to valuation with transaction, point-of-interest and road graphs (Viana and Barbosa, 2021; Zhang et al., 2021; Das et al., 2021; Guo and vanden Broucke, 2026); spatiotemporal graph forecasting models (Li et al., 2018; Yu et al., 2018; Wu et al., 2019; Bai et al., 2020) are surveyed by Jin et al. (2024a,b). Deep time-series forecasters include LSTM and GRU (Hochreiter and Schmidhuber, 1997; Cho et al., 2014), temporal convolutions (Bai et al., 2018), N-BEATS (Oreshkin et al., 2020), DeepAR (Salinas et al., 2020), the Temporal Fusion Transformer (Lim et al., 2021) and PatchTST (Nie et al., 2023); Zeng et al. (2023) and the M4 and M5 competitions (Makridakis et al., 2020, 2022a,b) are the cautionary references.

Regimes, state space, hierarchy and probability. Markov-switching models (Hamilton, 1989; Kim and Nelson, 1999), mixtures of experts (Jacobs et al., 1991; Jordan and Jacobs, 1994; Shazeer et al., 2017), deep state-space models (Rangapuram et al., 2018; Krishnan et al., 2017; de Bézenac et al., 2020), hierarchical Bayesian modelling (Gelman, 2006; Hoffman and Gelman, 2014; Blei et al., 2017; Rue et al., 2009), forecast reconciliation (Hyndman et al., 2011; Wickramasuriya et al., 2019; Athanasopoulos et al., 2024), forecast combination (Bates and Granger, 1969; Timmermann,

2006), quantile and conformal methods (Koenker and Bassett, 1978; Taylor, 2000; Romano et al., 2019; Vovk et al., 2005) and proper scoring rules (Gneiting and Raftery, 2007; Gneiting et al., 2007; Gneiting and Katzfuss, 2014; Bracher et al., 2021) supply the components. Forecast evaluation follows Diebold and Mariano (1995), Harvey et al. (1997), West (1996), Clark and McCracken (2001, 2013), Giacomini and White (2006), Hansen et al. (2011), Politis and Romano (1994), Künsch (1989), Hyndman and Koehler (2006) and Diebold (2015); dynamic factor and Bayesian VAR benchmarks follow Stock and Watson (2002b,a), Litterman (1986) and Bańbura et al. (2010); interpretation follows Lundberg and Lee (2017); Lundberg et al. (2020), Apley and Zhu (2020) and Molnar (2025).

The gap and the classification of contributions. Table 1 classifies the six architectures against their closest predecessors (full detail in the repository's novelty scan). None is unprecedented. The hierarchical hedonic network is a neural parameterisation of mixed-effects and graph-Laplacian hedonics; the graph network is an adaptation of traffic-forecasting models to municipal nodes; the regime model is a gated mixture of experts with origin-only regime inference; the Bayesian forecaster is a conjugate hierarchical model in the small-area tradition; the neural state-space model is a deep state-space model whose observation equation is the WP22 quality-adjusted index; the integrated architecture is a constrained forecast combination. The gap this paper fills is evaluative: a controlled, leakage-audited comparison of these mechanisms on transaction-level data over a geographic hierarchy, with inference that respects overlapping horizons, and a simulation study that attributes performance to mechanisms.

3. Data and institutional context

The data are those of WP22, used read-only: the Vrai Prix snapshot of 21 October 2026 of Quebec land-registry deeds republished by the Journal de Québec / Journal de Montréal feed, geocoded and matched to the provincial assessment roll (769,995 registered sales from 4 January 2021 to 7 October 2026), reduced by the WP22 exclusion ledger to a headline population of 632,006 house, condominium and plex sales with roll characteristics. The institutional facts that matter for forecasting are three. Deeds are published with a lag of ten to thirteen days and never backfilled (WP22 measured this on four earlier extractions), so the information set at an origin is the set of sales dated at least ten days before the publication date; the last complete month is September 2026 and October is partial. The roll is a single current vintage, so characteristics are post-sale for most dwellings; this affects Task A in the same way for every model. The provider collects no sale below $50,000.

Table 2 lists the variables actually available after inspection. The quality variables are floor

Table 1. Classification of the proposed architectures against the verified literature

Architecture	Closest predecessors	Classification	What is tested here
HSHN	Mixed-effects / hierarchical hedonics (Mayer et al., 2019; Gelman and Hill, 2007); graph-Laplacian hedonics (Sijp and Francke, 2026); MugRep (Zhang et al., 2021)	Adaptation (neural partial pooling with a horizon input)	Out-of-time gains over trees and MLP; partial vs full vs no pooling in thin municipalities
DSGFN	DCRNN, Graph WaveNet, AGCRN (Li et al., 2018; Wu et al., 2019; Bai et al., 2020); STAR / FAVAR (Pace et al., 2000; Gupta et al., 2011b); Deep-DFVAR (Yeo et al., 2025)	Application of adaptive-graph ST models to municipal indices	Gains over spatial-lag and parent-momentum benchmarks; edge-type ablation
RAREF	Markov switching (Hamilton, 1989); mixtures of experts (Jacobs et al., 1991); DMA (Bork and Møller, 2015)	Adaptation (origin-only HMM gating of ridge experts)	Gains around regime transitions; stability with 45 origins
BHPF	Hierarchical repeat sales (Francke and van de Minne, 2017); Fay–Herriot (Fay and Herriot, 1979)	Application (conjugate hierarchical Gaussian model)	Calibration and sharpness vs quantile networks and conformal methods
QANSS	Structural time-series indices (Francke, 2010); deep state space (Rangapuram et al., 2018)	New combination (WP22 quality-adjusted state + learned drift transition)	Gains over the local-linear-trend and AR benchmarks on the same vintages
QREFA	Forecast combination (Bates and Granger, 1969; Timmermann, 2006); TFT (Lim et al., 2021)	Constrained combination of the five	Whether integration beats the best component and an equal-weight combination

area, lot area, year built, storeys, building genre, physical link (detached, semi-detached, row, integrated), number of dwellings for plexes, coordinates, municipality, MRC and region, and the assessment in force at the sale; bedrooms, bathrooms, parking, basement, condition, renovations, quality grade and listing attributes are not observed, and no image or text features are used. WP22 provides, in addition, the nested metric location cells and the feature engineering (imputation with time-stable missingness indicators, centred log areas, age bins) that we reuse unchanged, and two index products: the retrospective SHARP-HPI panel (province, 17 regions, 103 MRCs, 1,125 municipalities $\times$ three classes $\times$ 69 months), which serves only as the *target* of Task B, and the real-time vintage database (monthly publication dates from July 2022, filtered and smoothed modes, province, regions and the 64 municipalities with at least 1,500 sales), which is the only index information admitted as a *predictor*.

Table 3 summarises coverage by region and class and the evaluation samples of each task. Figure 1 maps municipal sales density. The sample is short for monthly forecasting: 69 months in total, of which the first 24 are history before the first origin; this constrains every model and is discussed throughout.

Table 5 quantifies selection into transaction directly: the annual sale rate of the roll stock is 4.8% for the newest quintile of houses against 3.3 to 3.8% for older ones, and 4.3% for the smallest

Table 2. Variables available and their use

Variable	Source	Treatment	Used in
Floor area of storeys (m²)	roll	continuous; cubic in centred log	all tasks
Lot area (m²)	roll	quadratic in centred log	all tasks
Year built / age	roll	nine age bins; missingness indicator	all tasks
Storeys, building genre, physical link, dwellings (plex)	roll	categorical	Task A
Coordinates, municipality, MRC, region	feed / SDA V2026-07	location; nested metric cells (WP22)	all tasks
Assessment in force at sale	feed	not used as a predictor (post-dated for Task A test sales would be fine, but excluded for comparability)	—
Municipal / regional market features at origin	computed from sales $\leq$ origin	median log price and quality-stripped level (3/6/12/24 m), momentum, volumes, dispersion	Task A
10 nearest past sales ($\leq$ 12 m)	computed	mean quality-stripped log price, mean distance, dispersion	Task A
Real-time SHARP-HPI vintage (13-month history)	WP22 index_vintages (smoothed mode)	1/3/6/12-m changes, volatilities, acceleration, counts, s.e.; parent (region, province) momentum	Task B
Bedrooms, bathrooms, parking, basement, condition, renovations, quality, listing, images	—	not available	—

Table 3. Headline sales by administrative region and class (WP22 sample reused)

Region	Houses	Condominiums	Plexes
01	11,607	526	1,546
02	17,151	754	3,876
03	40,660	16,479	7,621
04	16,644	1,187	4,527
05	32,109	4,019	5,678
06	29,359	46,957	22,376
07	26,740	4,851	4,689
08	7,994	186	1,631
09	4,509	52	669
10	945	0	117
11	4,505	6	630
12	28,708	1,674	3,541
13	16,910	7,490	2,326
14	38,606	6,996	4,961
15	46,363	7,864	8,990
16	83,868	23,340	10,825
17	15,980	608	2,356

Table 4. Transaction selection: medians of sold dwellings (2021–2026) versus the assessment-roll stock

Class	Floor area sold	Floor area stock	Age sold	Age stock	Lot sold	Lot stock	Sales	Stock units
house	114	116	41	45	653	699	422,658	1,899,469
condo	90	89	17	20	104	87	122,989	344,399
plex	194	204	65	66	507	488	86,359	336,283

Stock = residential use code 1000 in the 2026 roll, classified with the WP22 rules; age of the stock as of 2026.

quintile against 3.5% for the largest. Sold dwellings are newer and smaller than the stock; the Task A evaluation population is the dwellings that come to market, and every model is evaluated on that population.

Table 5. Transaction selection — annual sale rate of the roll stock (% of units sold per year, 2021–2026) by quintile of floor area and of age

Class and dimension	Q1	Q2	Q3	Q4	Q5
house, by floor area quintile (1 = smallest / newest)	4.3	3.9	3.9	3.8	3.5
house, by age quintile (1 = smallest / newest)	4.8	3.7	3.3	3.9	3.8
condo, by floor area quintile (1 = smallest / newest)	5.8	6.4	6.7	6.8	5.5
condo, by age quintile (1 = smallest / newest)	5.6	7.2	6.3	6.1	6.1
plex, by floor area quintile (1 = smallest / newest)	5.6	4.8	4.2	4.0	3.8
plex, by age quintile (1 = smallest / newest)	4.6	4.0	4.1	4.7	4.9

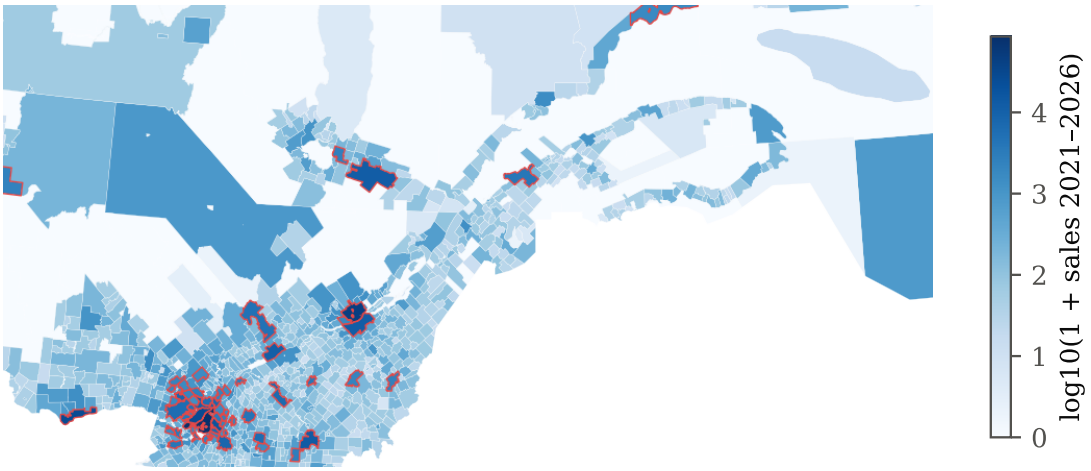

Figure 1. Municipal sales density (headline sample, 2021–2026) and the 64 municipalities with real-time index vintages (outlined).

4. Forecasting targets and information sets

4.1. Task A: the price of a future sale

For a forecast origin o (a month) and a sale i dated in month $o+h$, $h \in \{1,\ldots,12\}$, the target is $\log P_i$ and the information set is $\mathscr{I}_o = \{$characteristics of i (known from the roll), its location, the market features of its municipality and region computed from sales dated $\leq o$, the quality-stripped prices of its ten nearest sales dated in $(o-12, o]$, and $h\}$. The characteristics of the sold dwelling are treated as known at the origin, which corresponds to the use case of valuing a specific dwelling h months ahead; the identity of the dwellings that will sell is not. Models are estimated on sales dated $\leq o$; every training sale is assigned a horizon h drawn from $\{1,3,6,12\}$ and receives the market features of the origin h months before its sale, so that horizon dependence is learned from the same kind of pairs that are evaluated. Origins are quarterly from December 2022 to September 2025; test sales are grouped into horizon bands 1–3, 4–6 and 7–12 months. The transaction-selection problem is explicit: the evaluation population is the dwellings that transacted, which differ from the stock (Table 4 compares the characteristics of sold dwellings with the roll stock by class), so Task A measures the accuracy of forecasts for dwellings that come to market, not for the stock.

4.2. Task B: the change of a quality-adjusted index

For cell g (province, region or one of the 64 municipalities), class k, origin o and horizon $h \in \{1,3,6,12\}$, the target is $y^{\text{final}}_{g,k,o+h} - y^{\text{final}}_{g,k,o}$, the change of the retrospective smoothed SHARP-HPI log index, the best available estimate of the true quality-adjusted market movement. The predictors are the thirteen-month history of the real-time index published for o (the vintage estimated on sales $\leq o-10$ days, smoothed mode), summarised as 1-, 3-, 6- and 12-month changes, 6- and 12-month volatilities, acceleration, monthly sale counts and their ratio, the published standard error, calendar month, and the same summaries for the cell's region and the province (the parent momentum). Using the final series as the target and the real-time series as predictor means that the target embeds the revision between the two; we report the decomposition of forecast error into the revision and the genuine future change. Origins are monthly from December 2022; for horizon h the last origin is September 2026 minus h, so that every evaluated forecast has an observed target. Models are re-estimated quarterly on the rows whose targets were observed by the origin and issue forecasts for the three following origins.

4.3. Task C: predictive distributions

Every model produces the 5%, 25%, 50%, 75% and 95% quantiles of its target, either natively (quantile heads, Gaussian posterior predictive, mixture outputs) or from the empirical quantiles of

its training residuals (homoskedastic fallback, labelled as such). We score them with the pinball loss, a quantile-grid approximation of the CRPS, the coverage and width of the 50% and 90% intervals and the weighted interval score (Bracher et al., 2021), and we report calibration as the share of outcomes below each quantile. Probabilities of decline and of appreciation above thresholds are read off the quantile forecasts by interpolation.

4.4. Leakage audit

Every predictor is computed from data dated on or before the origin; the index vintages are those WP22 published for the origin, re-estimated at each publication date with their own location partition and empirical-Bayes priors; one-hot categories, standardisation constants and neighbour graphs are built on training data only; hyper-parameters are fixed in advance (Appendix B) rather than tuned on test origins; and random splits are never used. The audit list of the protocol (hedonic estimation, clustering, normalisation, feature selection, imputation, target encoding, graph estimation, tuning, index revisions) is checked item by item in Appendix C.

5. Proposed forecasting architectures

All architectures are implemented in the accompanying repository (PyTorch for the neural ones) with fixed hyper-parameters (Appendix B). Each produces the five quantiles of Task C; the point forecast is the median unless stated.

5.1. HSHN: hierarchical spatiotemporal hedonic network (Task A)

The network maps the standardised hedonic and market features x_i, three learned embeddings $g_i = e_{\mathrm{reg}(i)} + e_{\mathrm{mrc}(i)} + e_{\mathrm{mun}(i)}$ (dimension 8) and a horizon embedding e_h (dimension 4) to five quantiles of $\log P_i$: an encoder of two GELU layers (256, 128) on $[x_i, g_i, e_h]$, an interaction block that combines the encoded representation with $[g_i, e_h]$, and a linear head. Partial pooling is implemented by weight decay on the municipality embedding only (coefficient 5×10^{-3} against 10^{-4} elsewhere), so that a municipality's embedding is a penalised deviation from its MRC and region embeddings and thin municipalities fall back on their parents; *full pooling* removes the municipality embedding, *no pooling* removes the penalty. Training minimises the mean pinball loss over the five quantiles with AdamW and a one-cycle schedule (8 epochs, batch 4,096). The design is a neural counterpart of mixed-effects hedonics with a horizon input; the test is whether the learned pooling beats trees and an unpooled MLP out of time, especially in thin municipalities.

5.2. DSGFN: dynamic spatial graph forecasting network (Task B)

Nodes are the cells of one class at one origin; each node is encoded from its standardised features and its thirteen-month relative history by a two-layer network (dimension 32). A single graph-attention layer aggregates neighbours with logits $q_u^\top k_v/\sqrt{d} + \omega \log A_{uv}$, where A is a prior adjacency that multiplies a Gaussian distance kernel between municipal centroids (bandwidth 60 km), a co-movement term $\frac{1}{2} + \frac{1}{2}\,\mathrm{corr}$ of the two nodes' monthly real-time changes over the history window, and an indicator that the nodes share the level and class; self-attention is excluded so that the aggregated vector is a neighbourhood signal, and ω is learned. The head reads the node's own encoding and the neighbourhood vector. Edges are re-computed at every origin from information available then (dynamic graph). Ablations drop the distance kernel, the co-movement term or the whole neighbourhood vector.

5.3. RAREF: regime-adaptive mixture of experts (Task B)

At each origin a three-state Gaussian hidden Markov model is estimated by Baum–Welch on the sequence of provincial real-time monthly growth rates known at that origin (one value per past origin, each taken from its own vintage, so that no smoothed full-sample growth enters), and its filtered state probabilities at the origin gate three ridge experts estimated on the training rows weighted by their own origins' state probabilities. The forecast is the probability-weighted expert forecast; the predictive quantiles are the mixture's. Regimes are therefore inferred only from origin-dated information and can be mapped to calendar episodes.

5.4. BHPF: Bayesian hierarchical probabilistic forecaster (Task B)

The target is modelled as $y = \beta_0 + x^\top \beta + u_{\text{region}} + u_{\text{cell}} + \varepsilon$ with Gaussian priors $\beta \sim N(0, \tau_b^2 I)$, $u_{\text{region}} \sim N(0, \tau_r^2)$, $u_{\text{cell}} \sim N(0, \tau_c^2)$ and $\varepsilon \sim N(0, \sigma^2)$. The posterior is conjugate and exact; the variance components are updated by fifteen iterations of an EM-type scheme (posterior second moments of the random effects). The predictive distribution is Gaussian with the posterior-predictive variance, which separates parameter uncertainty from residual variance; cells unseen in training receive the region effect only. This is a small-area model in the Fay–Herriot tradition applied to index changes; it is the natural null for the neural hierarchical claims.

5.5. QANSS: quality-adjusted neural state-space forecaster (Task B)

WP22 already separates the quality component from the latent market component: its real-time index is the filtered latent state. QANSS forecasts the latent state by a learned transition: a two-layer network (32 tanh units) maps the cell's state at the origin (recent drift, volatility, volume ratio, provincial and regional drifts, last two monthly changes) to the mean monthly drift over the horizon

and the log standard deviation of the h-month change, estimated by maximising the Gaussian likelihood of the training targets; the predictive variance adds the published filter variance of the origin level. The model is a deep state-space model in the sense of Rangapuram et al. (2018) whose observation equation is the WP22 hedonic state-space index; the question is whether the learned drift transition beats the local-linear-trend and autoregressive benchmarks on the same vintages.

5.6. QREFA: integrated architecture

The integrated architecture stacks the components by a non-negative least-squares combination of their in-sample forecasts, normalised to sum to one, estimated on the training origins; its quantiles are the weighted quantiles. Each component can be removed from the stack, which is the ablation design of the protocol. The null hypotheses are the best single component and the equal-weight combination.

6. Econometric and machine learning benchmarks

Task A benchmarks are a hedonic OLS, ridge and elastic net on the same features, a random forest (150 trees, depth 18), XGBoost (600 trees, depth 8), LightGBM (800 trees, 63 leaves, with five quantile models), support vector regression on a 15,000-sale subsample, and a quantile MLP (two GELU layers, 256 and 128 units, pinball loss) without embeddings; the naive reference is the municipality's median log price over the three months before the origin. Task B benchmarks are the no-change forecast, the cell's historical mean h-month growth, momentum (the last three-month change scaled to the horizon with a fitted slope), random walk with the twelve-month drift, seasonal naive (the same-calendar change of the previous year), an AR(2) on the cell's monthly changes, a local linear trend (Kalman slope extrapolation), a pooled ridge regression on the panel features, LightGBM on the panel (with quantile models), a panel MLP, LSTM and GRU on the thirteen-month history with the features, a temporal convolutional network, and a TFT-lite with gated variable selection and self-attention over the history. Factor and VAR models are represented by the parent-momentum features (provincial and regional changes enter every pooled model) and by the AR benchmark; a full Bayesian VAR on 82 series with 45 origins is not estimable with any reliability and is not reported. All neural benchmarks output the five quantiles with the pinball loss; the others use residual quantiles. Hyper-parameters are listed in Appendix B; none was tuned on test origins.

7. Simulation design and results

7.1. Design

The simulation reuses the WP22 synthetic markets (60 municipalities nested in 12 MRCs and 4 regions, log-normal municipality sizes with a median of 3.3 sales a month, a nonlinear hedonic surface, spatially correlated cell quality, Student-t noise, seasonal volumes, resales and missing characteristics) and replaces the latent price paths by five data-generating processes that differ in forecastability. In *rw* the provincial path is a random walk with a constant drift of 0.3% a month and the regional and municipal deviations are pure random walks: the drift is learnable, nothing else is. In *momentum* the drifts at every level follow AR(1) processes (persistence 0.85 at the province, 0.8 for regions, 0.7 for municipalities) with small innovations. In *regimes* the provincial drift switches between $+1.2\%$ and -0.6% a month by a Markov chain with persistence 0.93 and 0.90. In *spillover* each municipality's drift loads with weight 0.6 on its region's lagged drift, so that regional information predicts municipal changes. *Sparse* is *momentum* with 35% of the sales. For every market, the real-time index at each origin is the WP22 SHARP-HPI filtered on the synthetic sales dated up to the origin (so that estimation noise and pooling are those of the real pipeline), the Task B features are built from it exactly as on the real data, and the target is the true latent change. Ten replications of 72 months are run for the classic and tree models and six for the neural models; forecasts are issued at horizons 1, 3 and 6 from origins 24 to 71 with re-estimation every six origins. Tables 6 to 32 report the three-month MAE relative to the no-change forecast with Monte Carlo standard errors, and Figure 2 the municipal results.

7.2. Results

The simulation explains the empirical ranking by separating two effects: the forecastability of the latent dynamics and the noise of the real-time index. When the predictable component is a constant drift (*rw*) or a large persistent switching drift (*regimes*), the simple rules recover it: the AR(2) and mean-growth forecasts reach 0.76 and 0.78 of the no-change MAE at the municipal level in *rw*, 0.78 and 0.82 in *regimes*, and the pooled LightGBM 0.87 and 0.79. When the predictable component is a small persistent drift (*momentum*, *sparse*), it is buried in the estimation noise of the municipal real-time index and no model beats the no-change forecast; the feature-rich pooled models are markedly *worse* than it (ridge 1.53, BHPF 1.61, LightGBM 1.16), because they fit the noise of nineteen noisy predictors on a few hundred training rows, and the same holds at the regional and provincial levels. Regional spillovers of the size simulated (*spillover*) are only weakly learnable: the AR(2) on the cell's own history already captures most of the lagged regional signal (0.91), and the pooled models with explicit parent features do not improve on it. The neural models (Table 6,

six replications) do no better than the pooled models and the two proposed architectures do worse: DSGFN reaches 1.32 in *momentum* and 0.95 in *regimes*, and does not exploit the spillover structure it was designed for (1.29 against 0.91 for the AR(2)), because with 60 nodes and 48 origins the attention cannot separate the regional signal from the noise; QANSS, whose learned drift transition reads the noisy recent changes of the real-time index, is the worst model in every scenario (2.04 in *rw*, 1.51 in *momentum*), although its intervals remain the best calibrated of the neural models (84% against 53% for DSGFN). The TFT-lite, which is the weakest neural model on the real data, is the best of them in the simulation; the ranking of the neural models is not stable across data-generating processes, which is the simulation counterpart of the seed variability of Section 13.

Two lessons carry over to the real data. First, the gains of the pooled tree model on the Quebec panel (MASE 0.75 at three months) are larger than anything the simulation produces from latent momentum alone; the simulated latent paths have no seasonal component, whereas the Quebec index has a pronounced seasonal cycle (Figure 6) that the calendar-month feature and the cell's own history capture. The real-data forecastability is therefore mostly seasonal composition in the quality-adjusted index, not momentum in latent prices, which is why the seasonal naive is the best provincial forecast and why the twelve-month horizon, where the seasonal cycle cancels, is unforecastable. Second, feature-rich models on noisy real-time indices can be worse than no change; the interval coverage of the pooled models in the simulation (56% for the ridge, 58% for LightGBM, for a nominal 90%) shows the same under-coverage as on the real data, while the no-change residual quantiles (81%) are close to nominal.

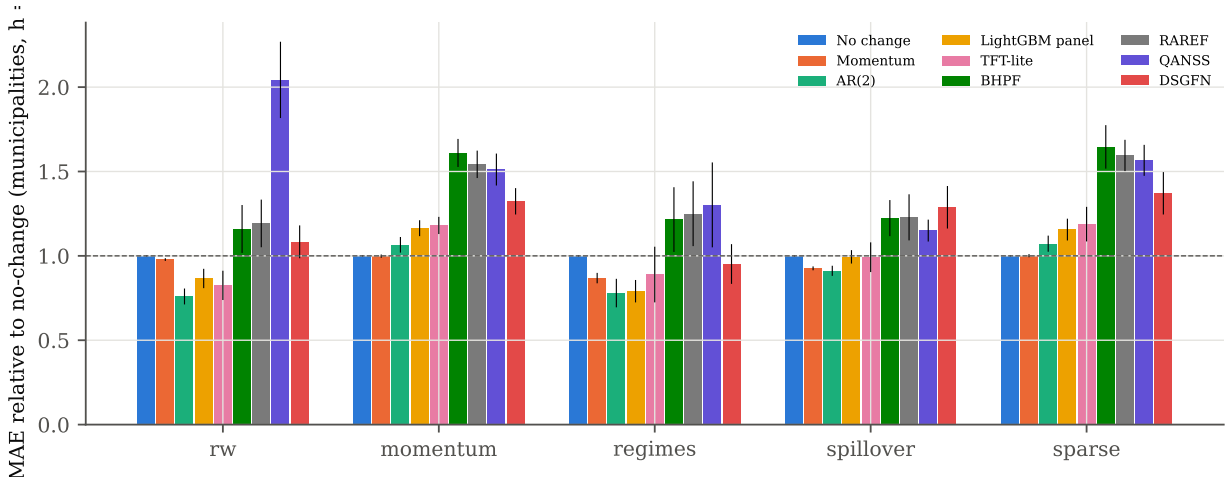

Figure 2. Simulation: three-month MAE relative to the no-change forecast at the municipality level, by scenario and model (whiskers: Monte Carlo standard errors).

Table 6. Simulation — MAE of 3-month index-change forecasts relative to the no-change forecast at the municipality level, by scenario (Monte Carlo s.e. over replications)

Model	rw	momentum	regimes	spillover	sparse
No change	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
Mean growth	0.78 (0.05)	1.13 (0.05)	0.82 (0.09)	0.98 (0.05)	1.13 (0.05)
Momentum	0.98 (0.01)	1.00 (0.01)	0.87 (0.03)	0.93 (0.01)	1.00 (0.01)
Random walk with drift	1.11 (0.06)	1.18 (0.06)	0.80 (0.06)	0.97 (0.03)	1.26 (0.07)
Seasonal naive	2.21 (0.16)	2.04 (0.12)	1.36 (0.13)	1.51 (0.06)	2.36 (0.18)
AR(2), cell	0.76 (0.05)	1.07 (0.05)	0.78 (0.08)	0.91 (0.03)	1.07 (0.05)
Local linear trend	2.29 (0.15)	1.90 (0.11)	1.13 (0.08)	1.33 (0.04)	2.26 (0.17)
Ridge (panel)	1.13 (0.13)	1.53 (0.08)	1.20 (0.20)	1.21 (0.10)	1.58 (0.12)
LightGBM (panel)	0.87 (0.06)	1.16 (0.05)	0.79 (0.07)	0.99 (0.04)	1.16 (0.06)
MLP (panel)	0.91 (0.08)	1.29 (0.04)	0.89 (0.11)	1.17 (0.13)	1.36 (0.14)
LSTM	0.90 (0.11)	1.18 (0.05)	0.91 (0.12)	1.08 (0.11)	1.33 (0.13)
TCN	0.88 (0.11)	1.21 (0.06)	0.91 (0.12)	1.07 (0.12)	1.31 (0.13)
TFT-lite	0.83 (0.09)	1.18 (0.05)	0.89 (0.16)	0.99 (0.09)	1.19 (0.10)
BHPF	1.16 (0.14)	1.61 (0.08)	1.22 (0.19)	1.22 (0.11)	1.65 (0.13)
RAREF	1.19 (0.14)	1.54 (0.08)	1.25 (0.19)	1.23 (0.14)	1.60 (0.09)
QANSS	2.04 (0.23)	1.51 (0.09)	1.30 (0.25)	1.15 (0.06)	1.57 (0.09)
DSGFN	1.08 (0.10)	1.32 (0.08)	0.95 (0.12)	1.29 (0.13)	1.37 (0.13)

rw: random walk with a constant drift; momentum: AR(1) drifts at every level; regimes: Markov-switching provincial drift; spillover: municipal drift follows the lagged regional drift; sparse: momentum with 35% of the sales. Values below 1 beat the no-change forecast. Classic and tree models: 10 replications; neural models: 6.

8. Empirical forecasting framework

Forecasts are produced in forward-chaining fashion. For Task A, twelve quarterly origins (December 2022 to September 2025) each yield a training set of all sales dated $\leq o$ (with horizon-dated features) and a test set of the sales of the following twelve months; for Task B, the models are re-estimated at every third origin on the rows whose targets were observed by the origin and used for the next three origins, which gives 34 to 45 evaluated origins per horizon. Point accuracy is reported as MAE, RMSE, median absolute error, bias and MASE relative to the naive reference; probabilistic accuracy as the mean pinball loss, the quantile-grid CRPS, 50% and 90% coverage and widths and the weighted interval score. Statistical comparisons use per-origin mean losses (which removes the cross-sectional dependence among forecasts issued at the same origin and leaves a short time series of n origins): the Diebold–Mariano test with a Newey–West long-run variance with $h-1$ lags and the Harvey–Leybourne–Newbold small-sample correction, the Giacomini–White conditional test with a constant and the lagged loss difference as instruments, stationary-bootstrap confidence intervals for mean loss differences, and the model confidence set at the 10% level with the $T_{\max}$ statistic and 500 bootstrap replications. With 34 to 45 origins these tests have limited power and are interpreted with the economic magnitudes alongside. Heterogeneity is reported by level, class, region, municipal size and price segment, and never for cells with fewer than 30 evaluated forecasts.

9. Main forecasting results

9.1. Task A: the price of a future sale

Table 7 reports the accuracy of the sale-price forecasts over 12 quarterly origins and 1,255,898 evaluated sales, by horizon band, with Diebold–Mariano tests against the naive municipal-median reference and the 90% model confidence set; Figure 3 plots the horizon profiles. Three results are robust across horizons. First, every model that uses the dwelling's characteristics beats the municipal median by a wide margin: the MASE of the hedonic OLS is 0.66 at one to three months, so a quarter of the error of the naive forecast is composition that any hedonic model removes. Second, gradient boosting is the most accurate model at every horizon (MAE 21.21 log points at one to three months, 22.04 at four to six and 22.70 at seven to twelve, against 23.24, 23.80 and 24.44 for the hedonic OLS), XGBoost and the random forest follow within one log point, and LightGBM is the only member of the 90% model confidence set at every horizon; the ranking is the one of the valuation literature (Mayer et al., 2019), now in a forecasting design. Third, the hierarchical hedonic network does not improve on the trees: its three pooling variants reach 22.36, 22.55 and 22.34 at one to three months, between the linear models and the trees, and the partial-pooling variant is indistinguishable from the unpooled one, so the learned municipality embeddings add nothing beyond the municipal market features that every model receives. The unembedded quantile MLP is the weakest learned model (26.14); its predictive intervals are far too wide (90% coverage 98%).

Table 7. Task A — Point accuracy of future sale-price forecasts by horizon band (months after the origin)

Model	1-3: MAE	1-3: MASE	4-6: MAE	4-6: MASE	7-12: MAE	7-12: MASE
Naive: municipal median	34.96	1.00	35.27	1.00	35.29	1.00
Hedonic OLS	23.24***	0.66	23.80***	0.67	24.44***	0.69
Ridge	23.24***	0.66	23.80***	0.67	24.44***	0.69
Elastic net	23.19***	0.66	23.73***	0.67	24.38***	0.69
SVR (subsample)	22.99***	0.66	23.42***	0.66	23.65***	0.67
Random forest	22.04***	0.63	22.59***	0.64	23.36***	0.66
XGBoost	21.36***	0.61	22.27***	0.63	23.02***	0.65
LightGBM	21.21***†	0.61	22.04***†	0.62	22.70***†	0.64
Quantile MLP	26.14***	0.75	25.97***	0.74	25.98	0.74
HSHN (no pooling)	22.34***	0.64	23.66***	0.67	24.82***	0.70
HSHN (full pooling)	22.55***	0.65	23.98***	0.68	25.19***	0.71
HSHN (partial pooling)	22.36***	0.64	23.68***	0.67	24.83***	0.70

Target: log sale price. MAE in log points ×100; MASE relative to the naive reference. Stars: Diebold–Mariano test of equal squared-error loss against the reference on per-origin mean losses (* 10%, ** 5%, *** 1%); † = in the 90% model confidence set of the horizon.

The horizon profile is informative about what the models learn. Accuracy deteriorates slowly with the horizon (LightGBM loses 1.5 log points of MAE between the first and the last band), but the bias grows: all models under-predict prices at seven to twelve months by 4 to 9 log points, which is the appreciation of the Quebec market over 2023–2026 that the models, trained on past levels,

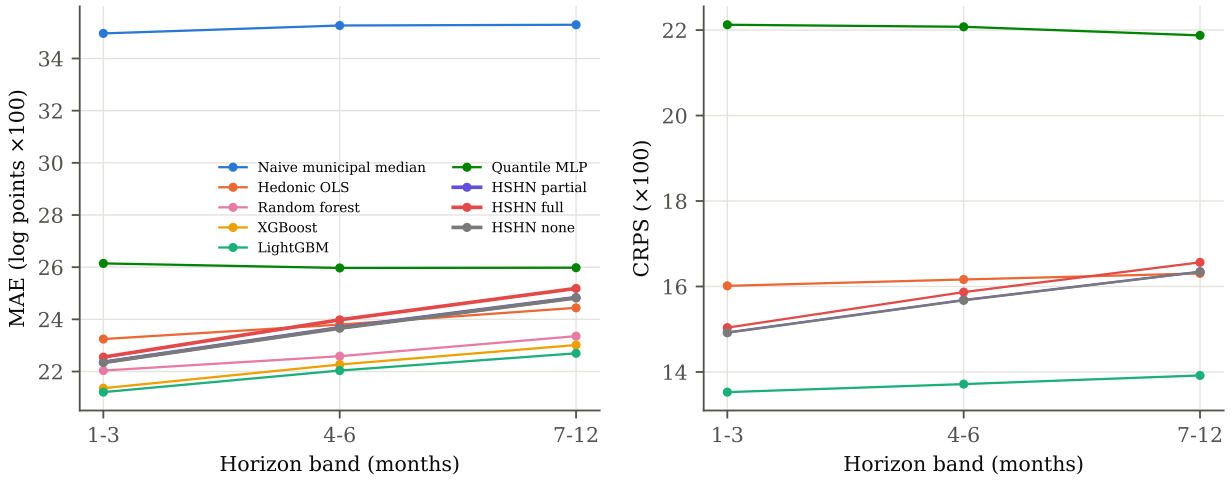

Figure 3. Task A: MAE and CRPS of sale-price forecasts by horizon band and model.

do not extrapolate. The naive median has the same bias, and the models that embed the horizon (HSHN) are not less biased than those that do not. Forecasting the level of a future sale is therefore two problems: valuing the dwelling at today's prices, which the trees do well, and forecasting the market, which none of the property-level models does better than assuming no change; Task B isolates the second problem. Figure 4 shows the MAE and bias of the main models at each of the twelve origins: the bias of the hedonic models is negative at every one of the twelve origins, between -3 and -8 log points for the origins from late 2023 to early 2025 and smallest (about -1 to -3) in mid-2023 and from mid-2025, when the market was decelerating; the unembedded quantile MLP swings from $+10$ to -19 and back, which is the instability of an under-regularised network on 50 to 150 thousand training sales rather than a market signal. In dollars (Table 8), the median absolute error of a one-to-three-month LightGBM forecast is \$61,102 for a house sold at the median price of \$420,000, \$32,005 for a condominium (median \$364,000) and \$80,115 for a plex (median \$510,000), against \$92,750, \$127,000 and \$135,000 for the municipal median: the characteristics reduce the dollar error by a third for houses and by three quarters for condominiums, whose prices vary most within a municipality, and the hedonic OLS is within 10% of LightGBM in every class.

9.2. Task B: the change of a quality-adjusted index

Table 9 reports the accuracy of the index-change forecasts for the 82 cells (province, 17 regions, 64 municipalities) by horizon, with the Diebold–Mariano test against the no-change forecast on per-origin mean losses, the Giacomini–White conditional test and membership of the 90% model confidence set; Figure 5 plots the horizon profiles and Figure 6 shows realised and forecast three-month changes for four cells. Five findings stand out.

Table 8. Task A, horizons 1–3 months — economic magnitude of forecast errors

Class	Model	MAE (log ×100)	Median price ($)	Median absolute error ($)
condo	LightGBM	13.3	364,000	32,005
condo	Hedonic OLS	15.6	364,000	37,900
condo	Naive: municipal median	34.9	364,000	127,000
house	LightGBM	23.0	420,000	61,102
house	Hedonic OLS	24.9	420,000	67,889
house	Naive: municipal median	34.5	420,000	92,750
plex	LightGBM	24.0	510,000	80,115
plex	Hedonic OLS	26.6	510,000	88,707
plex	Naive: municipal median	37.6	510,000	135,000

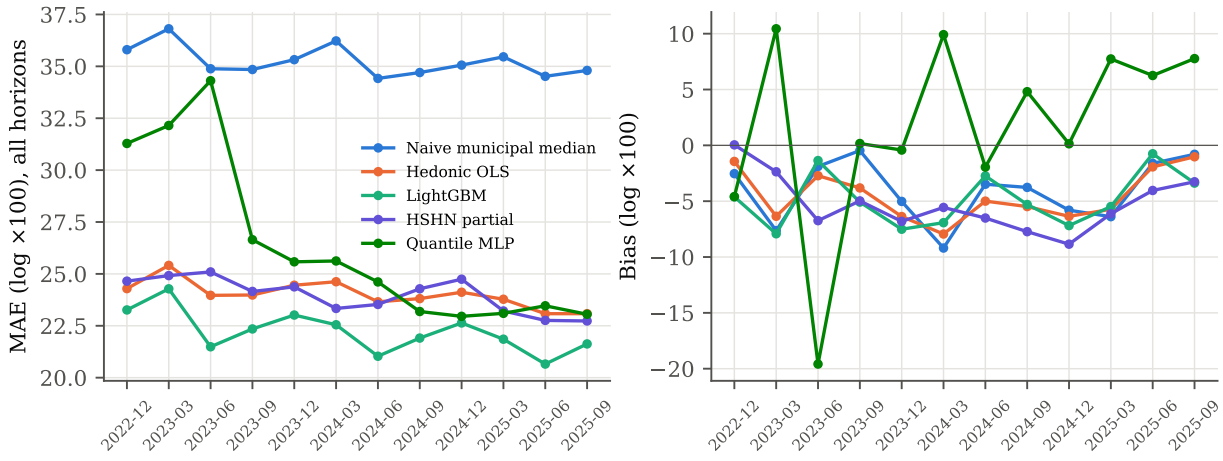

Figure 4. Task A: MAE (left) and bias (right) of sale-price forecasts by forecast origin, all horizons pooled.

First, quality-adjusted index changes are forecastable at short horizons, but modestly. At one month the best model, the pooled LightGBM, has a MASE of 0.87 relative to the no-change forecast (MAE 2.15 against 2.47 log points); at three months 0.75 and at six months 0.75. The Diebold–Mariano test rejects equal accuracy with the no-change forecast at the 1% level for LightGBM at three months and at the 5% level for QANSS and QREFA, and the Giacomini–White test agrees for LightGBM; with 31 to 45 origins per horizon the model confidence set is wide and separates only the clearly dominated models (local linear trend, seasonal naive, random walk with drift at some horizons).

Second, the source of the predictability is momentum and seasonality in the index, not spatial information: the AR(2) on the cell's own monthly changes, the mean-growth rule and the pooled ridge all beat the no-change forecast by similar margins at three months (3.76, 4.00 and 3.52 against 4.19), the simple momentum rule does not (its fitted slope is close to zero because three-month changes are dominated by a seasonal cycle that the rule extrapolates in the wrong direction), and the seasonal naive is the best forecast at the provincial level at three months, where the June–July cycle of the index is strongest. The ablations of Section 13 confirm that the cell's own momentum features carry the signal and that the regional and provincial momentum add little.

Third, among the proposed architectures, QANSS and the integrated QREFA are the only ones that improve on the no-change forecast at one, three and six months with statistical support (QANSS: 2.30, 3.48, 4.51; QREFA: 2.36, 3.42, 4.48), and neither beats the pooled LightGBM. BHPF and RAREF are indistinguishable from the pooled ridge they extend (3.57 and 3.58 against 3.52): the hierarchical intercepts shrink to nearly zero in a panel where the features already include the cell's own history, and the regime gating reproduces the ridge because the three experts converge to similar coefficients. DSGFN is worse than the no-change forecast at one month and no better than the ridge at three (3.69), consistent with the ablation that its graph carries no information.

Fourth, the deep sequence models (LSTM, GRU, TCN, TFT-lite) are the weakest learned models at every horizon and are worse than the no-change forecast at one month (3.29, 2.72 against 2.47): with 13-month histories and 34 to 45 origins they have too little data, and the robustness section shows that their errors move by half a log point across seeds. The Quebec panel reproduces the warning of Zeng et al. (2023) and the M-competitions (Makridakis et al., 2020): in short samples, simple and tree-based models dominate.

Fifth, at twelve months every pooled model fails. The AR(2) is the best forecast (4.74 against 7.52 for no change), the pooled models are worse than no change and carry positive biases of 6 to 11 log points: trained on origins from 2022 to mid-2024, they learned the appreciation of 2021–2023 and project it onto 2025–2026, when the market slowed. The twelve-month evaluation has only 34 origins concentrated in one macroeconomic episode, and the result says less about the models than about the impossibility of learning annual dynamics from a sample that contains one cycle; it is the

strongest argument in the paper for the protocol's insistence on short-sample humility.

Table 9. Task B — Point accuracy of quality-adjusted index-change forecasts by horizon (months)

Model	1: MAE	1: MASE	3: MAE	3: MASE	6: MAE	6: MASE	12: MAE	12: MASE
No change	2.47†	1.00	4.19†	1.00	5.37†	1.00	7.52	1.00
Mean growth	2.46†	1.00	4.00	0.95	4.59†	0.85	5.11	0.68
Momentum	2.48†	1.00	4.24†	1.01	5.73†	1.07	7.54	1.00
Random walk with drift	2.54**†	1.03	4.23	1.01	5.27	0.98	7.23	0.96
Seasonal naive	2.88***	1.16	4.29	1.02	5.21	0.97	7.23	0.96
AR(2), cell	2.43†	0.98	3.76*†	0.90	4.34†	0.81	4.74†	0.63
Local linear trend	3.04***	1.23	6.46***	1.54	10.27***	1.91	16.06***	2.14
Ridge (panel)	2.39*†	0.97	3.52*†	0.84	4.76†	0.89	9.11	1.21
LightGBM (panel)	2.15***†	0.87	3.16***†	0.75	4.02*†	0.75	5.69	0.76
MLP (panel)	2.71†	1.10	3.81†	0.91	5.00†	0.93	11.50	1.53
LSTM	3.29*†	1.33	3.88†	0.93	4.57†	0.85	7.56	1.00
GRU	3.07†	1.24	3.67†	0.88	4.60†	0.86	7.79	1.04
TCN	2.70†	1.09	3.81†	0.91	4.83†	0.90	8.31	1.10
TFT-lite	2.72†	1.10	4.13†	0.99	5.19	0.97	6.93	0.92
BHPF	2.40*†	0.97	3.57*†	0.85	4.74†	0.88	8.09	1.08
RAREF	2.39*†	0.97	3.58*†	0.85	4.69†	0.87	8.76	1.16
QANSS	2.30**†	0.93	3.48**†	0.83	4.51†	0.84	8.10	1.08
DSGFN	2.53†	1.02	3.69†	0.88	4.79†	0.89	8.33	1.11
QREFA	2.36*†	0.96	3.42**†	0.82	4.48†	0.83	7.61	1.01

Target: final SHARP-HPI log change. MAE in log points ×100; MASE relative to the naive reference. Stars: Diebold–Mariano test of equal squared-error loss against the reference on per-origin mean losses (* 10%, ** 5%, *** 1%); † = in the 90% model confidence set of the horizon.

9.3. Evolution of the errors over time, multiple testing and combinations

Figure 7 and Table 10 follow the three-month errors origin by origin. The gain of the pooled models over the no-change forecast is not uniform: it is concentrated in the origins of December to March, whose three-month windows cover the spring upswing, when the realised change is large and positive and the models' calendar-month and momentum features anticipate it (LightGBM reaches a MASE of 0.68 on the 2026 origins against 0.85 on those of late 2023), and it vanishes for the origins of summer and autumn, where no model does better than no change. Table 11 makes the same point with ex-post regimes of the provincial market: when the provincial house index rose by more than two log points over the target window (17 origins) LightGBM's MASE is 0.63 and every pooled model beats no change; when the window was flat (13 origins) LightGBM's MASE is 1.00, only the momentum rule edges out the no-change forecast (0.97), and DSGFN is the worst model (1.29); the single falling window, the three months from September 2025, is forecast better than no change by DSGFN, QREFA, TFT-lite, LightGBM and momentum, and worse by the AR(2), BHPF, RAREF, QANSS and the ridge, which extrapolate the positive drift. The forecastability measured in Section 9 is therefore the forecastability of seasonal upswings; a model that is good in rising windows and no worse than no change in flat ones, which is LightGBM's profile, is the best that this sample supports.

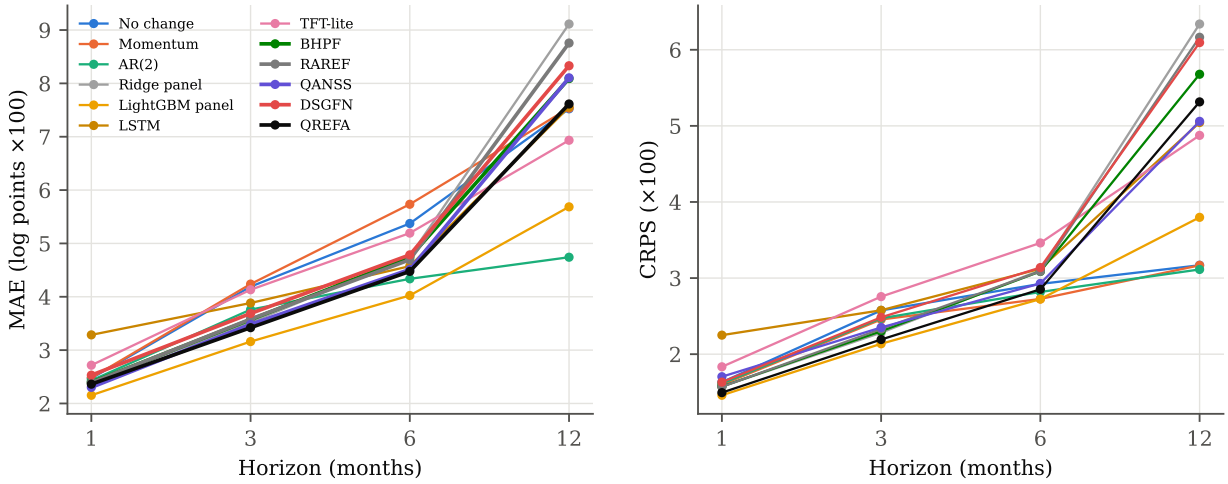

Figure 5. Task B: MAE and CRPS of index-change forecasts by horizon and model.

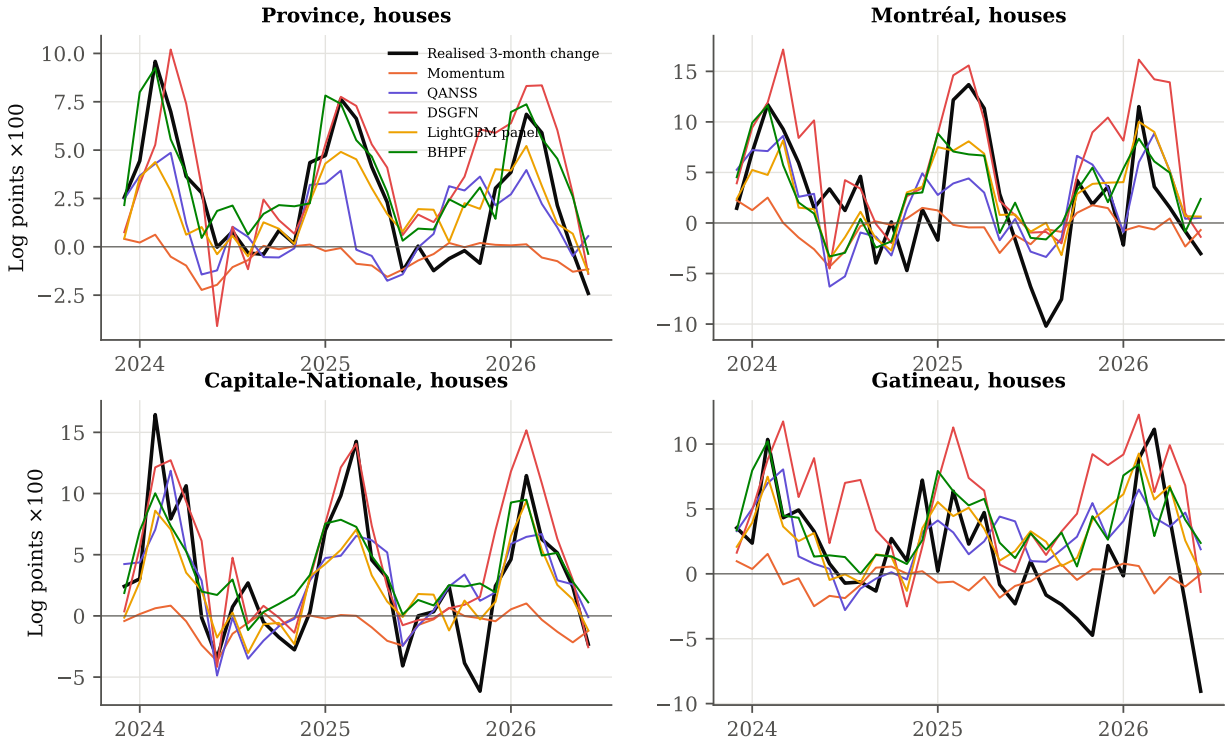

Figure 6. Task B: realised three-month change of the house index (black) and forecasts of five models, by origin, for the province, the Montréal and Capitale-Nationale regions and Gatineau.

Table 26 in Appendix D reports the Diebold–Mariano statistics with Holm-adjusted p-values across the 18 models of each horizon. The adjustment matters: at three months LightGBM's unadjusted p-value of 0.007 becomes 0.11 after Holm, QANSS's 0.037 becomes 0.59, and no improvement over the no-change forecast survives a 10% family-wise error rate at any horizon; only the deterioration of the local linear trend remains significant. With 31 to 45 origins the evidence that any model beats no change is suggestive at the 5% level for single comparisons and absent once the number of models tried is accounted for. The Giacomini–White conditional test, which asks whether the loss difference is predictable from its own lag, rejects for LightGBM ($p = 0.005$), whose advantage is itself predictable from the season, and for the random walk with drift, the seasonal naive and the mean-growth forecasts, whose relative performance alternates with the calendar.

Directional accuracy (Table 12) tells the same story in a form closer to use. At three months the pooled models predict the sign of the realised change in 75% of cell-months (LightGBM) to 77% (DSGFN), against 50% for a coin and 50% for the momentum rule, whose sign is that of the previous quarter; at turning points, origins where the realised change reverses sign, the accuracy falls to 65 to 66%, little above chance, and momentum, which by construction predicts the sign that has just reversed, does best (69%) for the wrong reason. At twelve months the directional accuracy of about 91% for every pooled model is uninformative: 91% of twelve-month changes in the sample were positive and the models predict a rise; at turning points none does better than a coin.

Table 13 evaluates equal-weight combinations of the strongest models on the same cells and origins. None improves on LightGBM alone (MASE 0.754 at three months against 0.760 for LightGBM + QANSS and 0.765 for the four-model combination), all improve on QREFA, the learned stack (0.816), and adding the no-change forecast as a fifth component (0.779) costs little. Combination protects against a bad component but, with one dominant component, does not beat it: the classical result that combination helps when the components have comparable accuracy and different information (Timmermann, 2006) holds here in its negative form, and the learned stack does worse than equal weights because its weights are estimated on the same short sample in which the components' relative accuracy changes from year to year (Table 10).

10. Spatial and regional heterogeneity

10.1. Task A by municipality size, class and price segment

Tables 28, 17 and 29 report the four-to-six-month MAE by municipality size (sales 2021–2026: large $\geq$ 5,000, medium 500–4,999, small $<$ 500), by class and by price segment. The ranking of models is the same in every group; what changes is the level of error and the gain over the naive reference. Errors are 40% larger in small municipalities than in large ones for every model,

Table 10. Task B, horizon 3 — MAE (×100) and MASE (in parentheses) by year of the forecast origin

Model	Origins in 2023	Origins in 2024	Origins in 2025	Origins in 2026
No change	3.75 (1.00)	4.40 (1.00)	3.94 (1.00)	4.35 (1.00)
Mean growth	3.63 (0.97)	4.15 (0.94)	3.81 (0.97)	4.13 (0.95)
Momentum	3.57 (0.95)	4.53 (1.03)	3.99 (1.01)	4.26 (0.98)
AR(2), cell	2.87 (0.77)	3.78 (0.86)	3.78 (0.96)	3.82 (0.88)
Ridge (panel)	3.28 (0.87)	3.23 (0.73)	3.63 (0.92)	3.89 (0.90)
LightGBM (panel)	3.17 (0.85)	3.35 (0.76)	3.07 (0.78)	2.95 (0.68)
TFT-lite	3.07 (0.82)	5.06 (1.15)	3.66 (0.93)	3.41 (0.78)
BHPF	3.59 (0.96)	3.32 (0.76)	3.66 (0.93)	3.89 (0.89)
RAREF	3.29 (0.88)	3.25 (0.74)	3.58 (0.91)	4.29 (0.99)
QANSS	3.16 (0.84)	3.48 (0.79)	3.40 (0.86)	3.69 (0.85)
DSGFN	2.77 (0.74)	3.91 (0.89)	3.76 (0.95)	3.25 (0.75)
QREFA	2.88 (0.77)	3.39 (0.77)	3.57 (0.91)	3.27 (0.75)

2023 covers origins from October to December only; 2026 covers January to June.

Table 11. Task B, horizon 3 — MAE (×100) and MASE (in parentheses) by ex-post provincial regime of the target window (rising: provincial house index change above +2 log points over the horizon; falling: below −2; flat otherwise)

Model	rising (17 origins)	flat (13 origins)	falling (1 origins)
No change	5.00 (1.00)	3.10 (1.00)	4.55 (1.00)
Mean growth	4.01 (0.80)	3.80 (1.23)	6.38 (1.40)
Momentum	5.20 (1.04)	3.00 (0.97)	3.87 (0.85)
AR(2), cell	3.82 (0.76)	3.54 (1.14)	5.66 (1.25)
Ridge (panel)	3.49 (0.70)	3.44 (1.11)	4.89 (1.08)
LightGBM (panel)	3.16 (0.63)	3.11 (1.00)	3.81 (0.84)
TFT-lite	4.68 (0.94)	3.46 (1.12)	3.50 (0.77)
BHPF	3.52 (0.70)	3.53 (1.14)	4.98 (1.10)
RAREF	3.52 (0.70)	3.55 (1.14)	5.03 (1.11)
QANSS	3.56 (0.71)	3.25 (1.05)	5.18 (1.14)
DSGFN	3.49 (0.70)	4.00 (1.29)	3.01 (0.66)
QREFA	3.36 (0.67)	3.51 (1.13)	3.39 (0.75)

Table 12. Task B — directional accuracy (% of forecasts with the sign of the realised change, changes above 0.5 log points) overall and at turning points (origins where the realised change reverses sign)

Model	h=1: all	h=1: turning pts	h=3: all	h=3: turning pts	h=6: all	h=6: turning pts	h=12: all	h=12: turning pts
Mean growth	60	51	70	50	78	51	90	47
Momentum	53	72	50	69	45	68	46	47
AR(2), cell	59	65	71	58	78	60	91	48
Ridge (panel)	65	68	73	58	76	59	91	48
LightGBM (panel)	70	77	75	65	78	59	91	47
TFT-lite	60	66	69	65	77	60	91	47
BHPF	65	67	73	57	77	59	91	49
RAREF	66	68	73	58	76	58	91	48
QANSS	67	75	73	61	76	64	75	50
DSGFN	63	66	77	65	79	60	91	47
QREFA	65	70	76	66	78	62	91	48

The no-change forecast has no sign; momentum has the sign of the last three-month change.

Table 13. Task B — MASE of equal-weight combinations versus their components (same cells and origins)

Forecast	h=1	h=3	h=6	h=12
LightGBM (panel)	0.871	0.754	0.749	0.756
QANSS	0.931	0.831	0.840	1.077
Ridge (panel)	0.969	0.839	0.886	1.211
AR(2), cell	0.983	0.897	0.807	0.630
QREFA	0.956	0.816	0.833	1.012
Equal weights: LightGBM + QANSS	0.870	0.760	0.735	0.845
Equal weights: LightGBM + AR(2)	0.892	0.787	0.751	0.673
Equal weights: LightGBM + QANSS + ridge + AR(2)	0.891	0.765	0.736	0.777
Equal weights: four models + no change	0.898	0.779	0.729	0.672

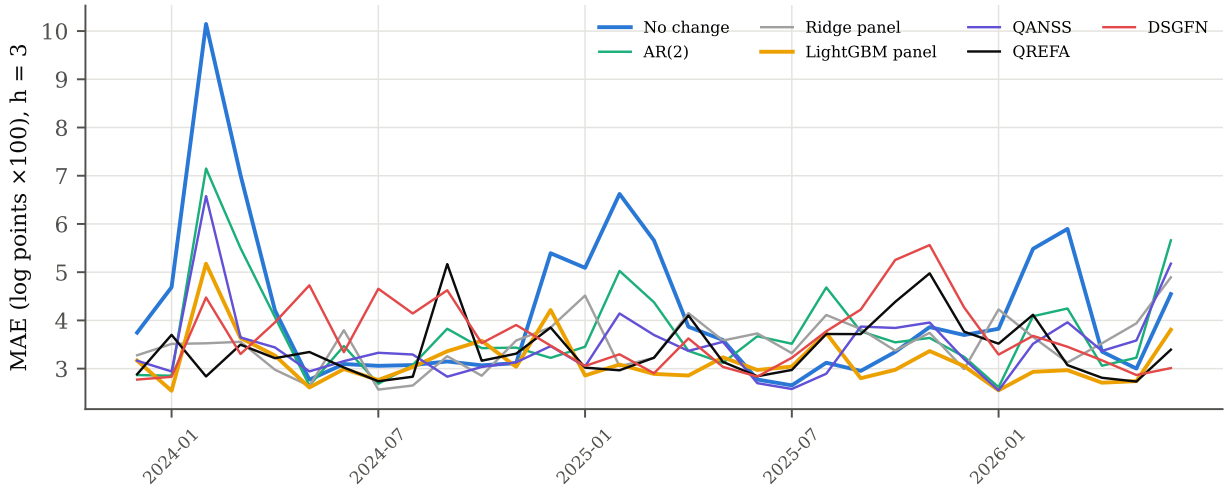

Figure 7. Task B, horizon 3: MAE of the index-change forecasts by forecast origin, selected models.

and the hierarchical network's partial pooling does not narrow the gap: in small municipalities HSHN-partial is no better than HSHN-none. The gain over the municipal median is largest for condominiums, whose prices are most heterogeneous within a municipality, and the errors are largest in the lowest price segment (below $250,000), where the models over-predict: the provider's $50,000 floor and the thinness of low-price markets make this segment the hardest and the most biased.

10.2. Task B by level, class and cell size

Tables 15, 16 and 30 report the three-month MAE by geographic level, class and municipal size (large: at least 50 sales in the origin month). Forecast errors scale with the noise of the target: municipal index changes are twice as volatile as regional ones, and the MASE of the best model relative to no change is similar across levels (0.69 for regions, 0.77 for municipalities with LightGBM), so the pooled models extract a similar share of the predictable variation at every level. Condominium indices are the most predictable in absolute terms and houses the least. Hierarchical pooling (BHPF, HSHN) does not improve on the pooled models in small municipalities: the gain over the no-change forecast is the same in large and small municipal cells, and the municipal intercepts shrink to zero. Figure 9 maps the three-month MAE of the best model relative to no change by region: the gain is largest in the regions with the strongest seasonal cycle and smallest in the sparse eastern regions. Table 14 and Figure 8 give the distribution of the three-month MASE across the 192 municipal cells: LightGBM beats the no-change forecast in 91% of the cells with a median MASE of 0.74, QANSS in 83%, the AR(2) in 80%, and the pooled ridge in 72% with a long right tail (90th percentile 1.44 against 0.97 for LightGBM): the linear pooled model is the one most often harmed by the heterogeneity of municipal dynamics, which is what the tree model's interactions absorb. Tables 30, 28 and 29 in Appendix E give the breakdowns by cell size and by price segment.

Table 14. Task B, horizon 3 — distribution of the MASE across the 192 municipal cells (64 municipalities × 3 classes)

Model	Median MASE	P10	P90	% of cells beating no change	Cells
LightGBM (panel)	0.74	0.62	0.97	91	192
Ridge (panel)	0.80	0.63	1.44	72	192
QANSS	0.82	0.68	1.08	83	192
AR(2), cell	0.91	0.76	1.08	80	192

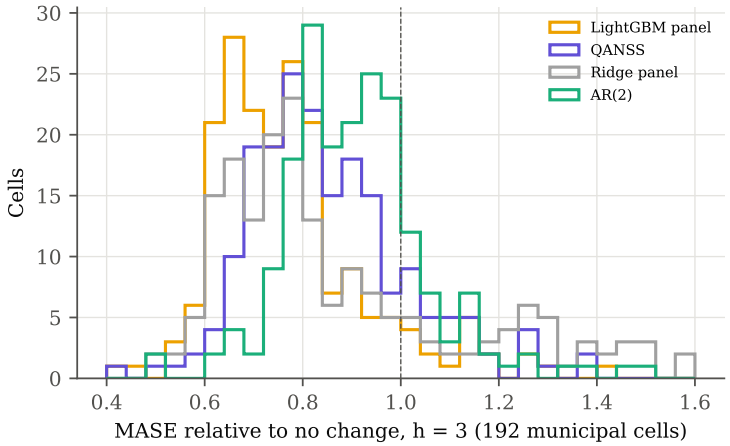

Figure 8. Task B, horizon 3: distribution of the MASE relative to no change across the 192 municipal cells.

Table 15. Task B — MAE (×100) by level, horizon 3

Model	municipality	province	region
No change	4.38	2.44	3.58
Mean growth	4.22	2.00	3.25
Momentum	4.39	2.73	3.73
Random walk with drift	4.52	1.90	3.27
Seasonal naive	4.71	1.37	2.85
AR(2), cell	3.95	2.06	3.14
Local linear trend	6.84	3.16	5.22
Ridge (panel)	3.73	2.06	2.78
LightGBM (panel)	3.36	1.52	2.47
MLP (panel)	4.04	2.08	3.04
LSTM	4.06	2.36	3.31
GRU	3.85	2.06	3.06
TCN	3.98	1.91	3.27
TFT-lite	4.34	2.09	3.46
BHPF	3.78	2.11	2.86
RAREF	3.80	2.12	2.82
QANSS	3.75	1.77	2.57
DSGFN	3.93	1.90	2.84
QREFA	3.67	1.66	2.57

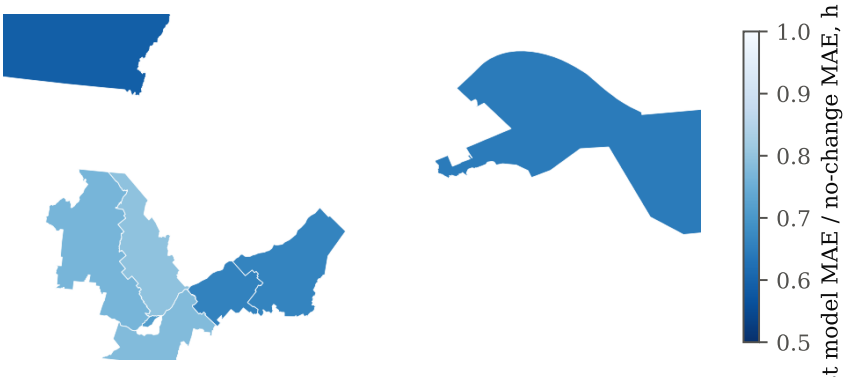

Figure 9. Task B: ratio of the best model's three-month MAE to the no-change MAE by administrative region.

Table 16. Task B — MAE (×100) by class, horizon 3

Model	condo	house	plex
No change	2.70	5.21	4.64
Mean growth	2.50	5.16	4.31
Momentum	2.81	5.20	4.68
Random walk with drift	2.76	5.31	4.62
Seasonal naive	3.01	5.09	4.77
AR(2), cell	2.58	4.79	3.90
Local linear trend	3.85	8.81	6.70
Ridge (panel)	3.11	3.69	3.75
LightGBM (panel)	2.24	3.56	3.67
MLP (panel)	2.82	4.24	4.35
LSTM	2.75	4.36	4.52
GRU	2.60	4.08	4.31
TCN	2.78	4.19	4.46
TFT-lite	2.80	4.59	4.98
BHPF	3.07	3.70	3.94
RAREF	3.13	3.84	3.76
QANSS	2.39	4.04	4.00
DSGFN	2.61	4.30	4.13
QREFA	2.58	3.86	3.81

Table 17. Task A — MAE (×100) by class, horizon 4-6

Model	condo	house	plex
Naive: municipal median	34.24	34.96	38.38
Hedonic OLS	15.86	25.49	27.20
Ridge	15.86	25.49	27.20
Elastic net	15.79	25.41	27.19
SVR (subsample)	15.08	25.15	27.19
Random forest	13.90	24.39	26.52
XGBoost	13.74	24.26	24.99
LightGBM	13.77	23.96	24.73
Quantile MLP	19.14	27.52	28.40
HSHN (no pooling)	17.31	25.05	26.17
HSHN (full pooling)	17.61	25.38	26.50
HSHN (partial pooling)	17.33	25.07	26.19

11. Probabilistic forecasting and calibration

Tables 18 and 20 report the CRPS, the coverage and the width of the 90% intervals; Figure 11 plots the share of outcomes below each nominal quantile. For Task A the LightGBM quantile models are the sharpest and are slightly under-covered (88% at one to three months for a nominal 90%), the linear models are over-covered because their residual quantiles are homoskedastic, the hierarchical network is close to nominal at short horizons and under-covers at seven to twelve months (82%) because its intervals do not widen with the horizon while the bias grows, and the unembedded MLP produces intervals two and a half times too wide. For Task B the ranking by CRPS follows the point ranking (LightGBM 2.14 at three months, QREFA 2.19, QANSS 2.35, no change 2.57), but calibration separates the models sharply. The neural quantile heads (LSTM, GRU, TCN, TFT-lite, DSGFN) under-cover at every horizon (73% to 73% at three months for a nominal 90%, falling below 50% at twelve months): trained with the pinball loss on 34 to 45 origins they fit the central quantiles and collapse the tails. The Gaussian posterior predictive of BHPF and the residual quantiles of the ridge are close to nominal at one and three months (90%, 90%) and under-cover at six and twelve months as the bias grows. QANSS is over-covered at every horizon (97% at three months, 93% at twelve) because its predictive variance adds the published filter variance of the origin level to the learned drift variance, and it is the only model whose twelve-month intervals contain the realised changes; the price is intervals twice as wide as LightGBM's. The decomposition of uncertainty the protocol asks for is only partly identified: the estimation component is the posterior variance of BHPF (small, because the panel is pooled), the transaction-noise component is the WP22 filter variance that QANSS carries, and the market-dynamics component is the residual drift variance; macroeconomic uncertainty is not modelled and is the reason every model fails at twelve months.

Table 18. Task A — Probabilistic accuracy (CRPS ×100, 90% interval coverage % and width ×100) of future sale-price forecasts by horizon band (months after the origin)

Model	1-3: CRPS	1-3: cov90	1-3: width90	4-6: CRPS	4-6: cov90	4-6: width90	7-12: CRPS	7-12: cov90	7-12: width90
Naive: municipal median	—	—	—	—	—	—	—	—	—
Hedonic OLS	16.02	92	116.5	16.16	92	116.6	16.31	92	116.4
Ridge	16.02	92	116.5	16.16	92	116.6	16.31	92	116.4
Elastic net	16.01	92	116.5	16.15	92	116.7	16.29	92	116.5
SVR (subsample)	—	—	—	—	—	—	—	—	—
Random forest	—	—	—	—	—	—	—	—	—
XGBoost	—	—	—	—	—	—	—	—	—
LightGBM	13.53	88	91.9	13.72	87	92.7	13.92	84	92.2
Quantile MLP	22.12	98	256.3	22.08	98	257.3	21.88	98	255.8
HSHN (no pooling)	14.92	89	107.8	15.68	86	108.3	16.34	82	106.5
HSHN (full pooling)	15.04	88	108.0	15.87	85	108.4	16.56	82	106.6
HSHN (partial pooling)	14.92	89	107.9	15.68	86	108.3	16.35	82	106.6

11.1. Probabilities of decline

The protocol asks for the probability of a price decline. Table 19 scores the probability of a decline of the quality-adjusted index over the horizon, read off the quantile forecasts, by the Brier score and its skill relative to the climatological frequency of declines (31% of cell-months at three months, 10% at twelve); Figure 10 plots reliability at three months. The integrated QREFA and LightGBM have the best skill at three months (+0.17 and +0.15), DSGFN, BHPF and the ridge follow, QANSS's over-wide distribution leaves it little skill (+0.07), and the no-change, TFT-lite and mean-growth probabilities are worse than the climatology. Reliability is good for QREFA, DSGFN, LightGBM and BHPF over the whole range of forecast probabilities, and QANSS's probabilities overstate the frequency of declines throughout (a forecast probability of 0.85 is followed by a decline 64% of the time), the mirror image of its over-wide intervals, which put too much mass below zero. At twelve months the skill of every model is negative, QANSS's severely so (-1.04): with fewer than 10% of twelve-month windows ending below their start, the climatology is hard to beat and the probabilistic forecasts inherit the bias of the point forecasts.

Table 19. Task C — probability of an index decline over the horizon: Brier score and Brier skill score relative to the climatological frequency of declines

Model	h=1: Brier	h=1: skill	h=3: Brier	h=3: skill	h=6: Brier	h=6: skill	h=12: Brier	h=12: skill
No change	0.244	-0.01	0.217	-0.02	0.179	-0.01	0.090	-0.01
Mean growth	0.249	-0.04	0.226	-0.06	0.187	-0.06	0.099	-0.12
Momentum	0.240	+0.00	0.205	+0.03	0.171	+0.03	0.089	-0.01
AR(2), cell	0.244	-0.01	0.208	+0.02	0.176	+0.00	0.089	-0.01
Ridge (panel)	0.231	+0.04	0.188	+0.11	0.186	-0.05	0.095	-0.07
LightGBM (panel)	0.219	+0.09	0.181	+0.15	0.174	+0.02	0.090	-0.02
TFT-lite	0.275	-0.14	0.223	-0.05	0.186	-0.05	0.095	-0.07
BHPF	0.226	+0.06	0.184	+0.13	0.179	-0.01	0.097	-0.10
RAREF	0.226	+0.06	0.194	+0.09	0.185	-0.05	0.094	-0.07
QANSS	0.226	+0.06	0.197	+0.07	0.181	-0.03	0.180	-1.04
DSGFN	0.252	-0.05	0.182	+0.14	0.170	+0.04	0.096	-0.08
QREFA	0.234	+0.03	0.176	+0.17	0.166	+0.06	0.097	-0.09

Probabilities are read off the five-quantile forecasts by linear interpolation of the quantile function. Skill = 1− Brier / Brier of the constant climatological probability.

12. Interpretation and economic mechanisms

12.1. What the property-level models use

Table 21 reports the mean absolute SHAP values (Lundberg and Lee, 2017; Lundberg et al., 2020) of the LightGBM sale-price model at three origins. The quality-stripped prices of the ten nearest past sales dominate (21.9 log points on average), followed by floor area (14.0), age (9.7) and lot area; the municipal and regional market features (levels, momentum, volumes) each contribute

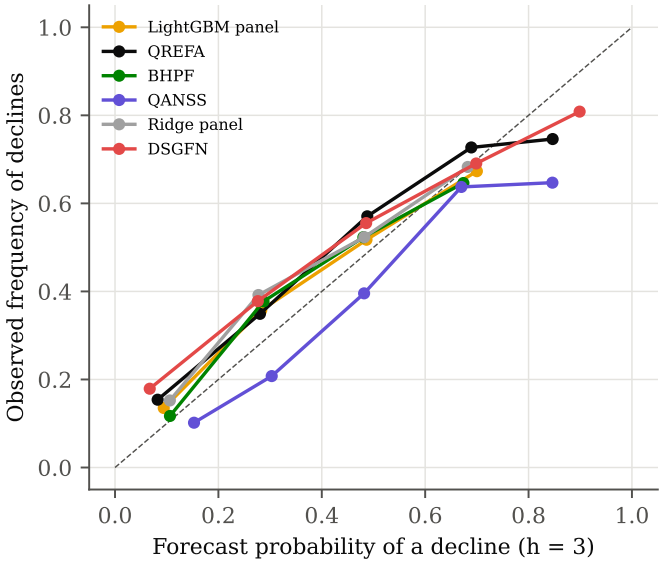

Figure 10. Task C: reliability of the forecast probability of an index decline over three months (bins of forecast probability).

Table 20. Task B — Probabilistic accuracy (CRPS ×100, 90% interval coverage % and width ×100) of quality-adjusted index-change forecasts by horizon (months)

Model	1: CRPS	1: cov90	1: width90	3: CRPS	3: cov90	3: width90	6: CRPS	6: cov90	6: width90	12: CRPS	12: cov90	12: width90
No change	1.63	91	11.7	2.57	92	18.4	2.92	94	22.3	3.17	88	18.2
Mean growth	1.65	91	11.7	2.65	91	18.1	3.05	92	21.7	3.57	80	14.6
Momentum	1.60	92	11.6	2.46	92	17.6	2.73	93	19.4	3.17	89	18.4
Random walk with drift	1.71	92	12.3	2.87	92	20.4	3.72	91	26.2	6.01	87	32.6
Seasonal naive	1.94	91	13.6	2.92	92	20.3	3.69	92	25.3	6.01	87	32.6
AR(2), cell	1.61	91	11.3	2.47	92	18.0	2.82	93	20.2	3.11	89	17.9
Local linear trend	2.04	91	14.8	4.38	92	32.3	6.93	95	58.5	11.21	95	94.2
Ridge (panel)	1.57	90	10.4	2.28	90	14.1	3.10	83	15.3	6.34	48	16.4
LightGBM (panel)	1.46	77	6.7	2.14	76	9.6	2.72	74	11.0	3.80	66	12.5
MLP (panel)	1.81	72	7.5	2.55	73	10.3	3.35	63	11.4	8.50	36	17.6
LSTM	2.25	69	8.0	2.58	72	10.6	3.12	72	12.8	5.04	64	14.6
GRU	2.05	70	8.0	2.33	81	11.5	2.93	81	14.2	4.88	76	17.9
TCN	1.82	69	7.2	2.51	71	10.0	3.07	75	13.1	5.18	72	17.5
TFT-lite	1.83	69	7.3	2.76	73	11.1	3.46	68	12.1	4.88	56	13.7
BHPF	1.58	90	10.7	2.30	90	14.6	3.09	83	14.9	5.68	50	14.2
RAREF	1.57	90	10.3	2.33	89	14.0	3.09	84	15.5	6.16	50	16.5
QANSS	1.70	98	17.2	2.35	97	20.7	2.93	95	23.3	5.06	93	35.0
DSGFN	1.63	78	7.6	2.48	73	10.2	3.14	72	12.9	6.09	45	12.4
QREFA	1.49	85	8.9	2.19	85	12.2	2.85	82	13.7	5.32	51	13.2

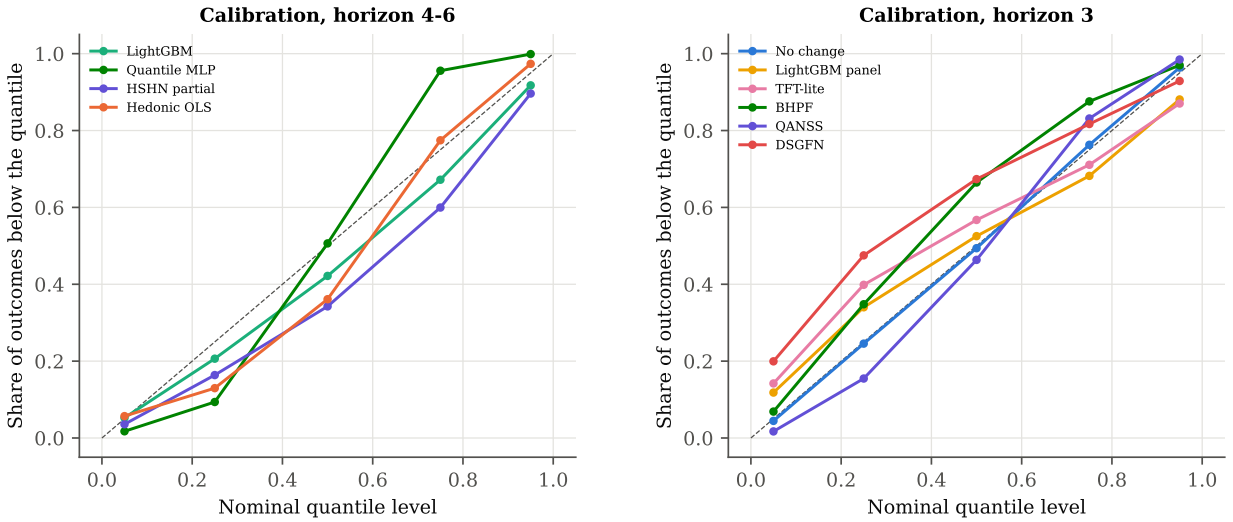

Figure 11. Calibration: share of outcomes below each nominal quantile, Task A (left, horizon band 4–6 months) and Task B (right, horizon 3).

less than 3.5 log points, and the time index, which carries the market trend during training, loses importance as the history lengthens (8.9 at the March 2023 origin against 1.0 at the March 2025 origin). The ranking is stable across origins. The accumulated local effect (Apley and Zhu, 2020) of the neighbour price is monotone and nearly linear over its range (85.6 log points from the lowest to the highest decile), which is the spatial counterpart of a location fixed effect; the effect of municipal momentum is small and, at the top decile, negative (-2.6 log points): dwellings in municipalities whose recent prices ran ahead of their twelve-month level sell for slightly less than the local level would imply, a mean-reversion pattern of the kind documented by Capozza et al. (2004), read here as a predictive association rather than a causal one.

12.2. What is forecastable at the market level

The protocol's eight questions receive the following answers from Sections 9, 10 and 13. (1) *Spatial spillovers* carry no predictive information beyond the parent (regional and provincial) momentum that every pooled model receives: the graph network does not beat the pooled ridge, its attention is diffuse, and removing its graph does not hurt. (2) *Local momentum* persists at one to six months: the cell's own recent changes are the dominant predictors, and AR(2) and mean-growth rules capture most of the gain; but it reverses at twelve months, where no pooled model beats the no-change forecast. (3) *Property characteristics* forecast the value of a dwelling, not the market: the trees halve the error of the municipal median in Task A, but their horizon profile shows no information about future market movements. (4) *Regimes*: Figure 12 shows the filtered state probabilities of the RAREF hidden Markov model, re-estimated at each origin from the provincial real-time

Table 21. Task A — mean absolute SHAP value (×100, log-price units) of the LightGBM model at three origins, 4,000 test sales each

Feature	2023-03	2024-03	2025-03	Mean
10 nearest past sales: quality-stripped log price	23.2	21.6	20.8	21.9
log floor area (centred)	13.9	14.0	14.1	14.0
age	10.0	9.8	9.3	9.7
origin month (time index)	8.9	3.1	1.0	4.3
log lot area (centred)	3.5	3.6	3.6	3.6
municipality: quality-stripped level, 12 m	3.4	2.7	4.0	3.4
northing	3.1	3.3	2.9	3.1
municipality: median log price, 3 m	2.7	3.1	2.3	2.7
distance to the 10 nearest sales	2.8	2.3	2.2	2.5
easting	2.5	2.5	2.3	2.4
dispersion of the 10 nearest sales	2.0	2.3	2.4	2.2
property_class=condo	2.1	2.1	2.4	2.2
subtype_c=detached	1.9	2.1	2.1	2.0
log floor area squared	1.4	1.4	1.5	1.4
municipality: sales, 3 m	1.6	1.4	1.1	1.4

Importance is predictive, not causal. Origin-month importance declines as the training history lengthens.

growth known then. The model identifies a high-growth state (mean monthly growth of about +2.6 log points at the last origin) that it assigns to the origins of March to June of 2024, 2025 and 2026 (and August 2024), a middle state for the rest of the year, and a low-growth state (mean −0.2) that it visits with near certainty only at the September 2025 origin, after the one negative provincial month of the evaluation period. The states are the seasonal cycle, not business-cycle regimes, and the gating therefore reproduces what the calendar-month feature already gives the ridge; separate regime experts cannot be estimated from a sample that contains one slowdown and one recovery, and Table 11 confirms that the forecast gains are a property of rising windows. (5) *Hierarchical pooling* in small municipalities does not help forecasting once the panel features are pooled: BHPF's municipal intercepts shrink to zero and HSHN's partial pooling matches no pooling. (6) *Uncertainty in low-volume periods*: QANSS, which carries the WP22 filter variance, widens its intervals in thin cells and is the only model with nominal or higher coverage at every horizon. (7) *Quality-adjusted versus median targets*: WP22 showed that the median index embeds composition; here the quality-adjusted index is forecastable at short horizons mainly through its seasonal and momentum structure. (8) *Turning points*: the 2025–2026 slowdown was not anticipated by any model; the twelve-month forecasts issued in 2024 carry the bias of the preceding boom. Predictive importance in these results is not evidence of causality.

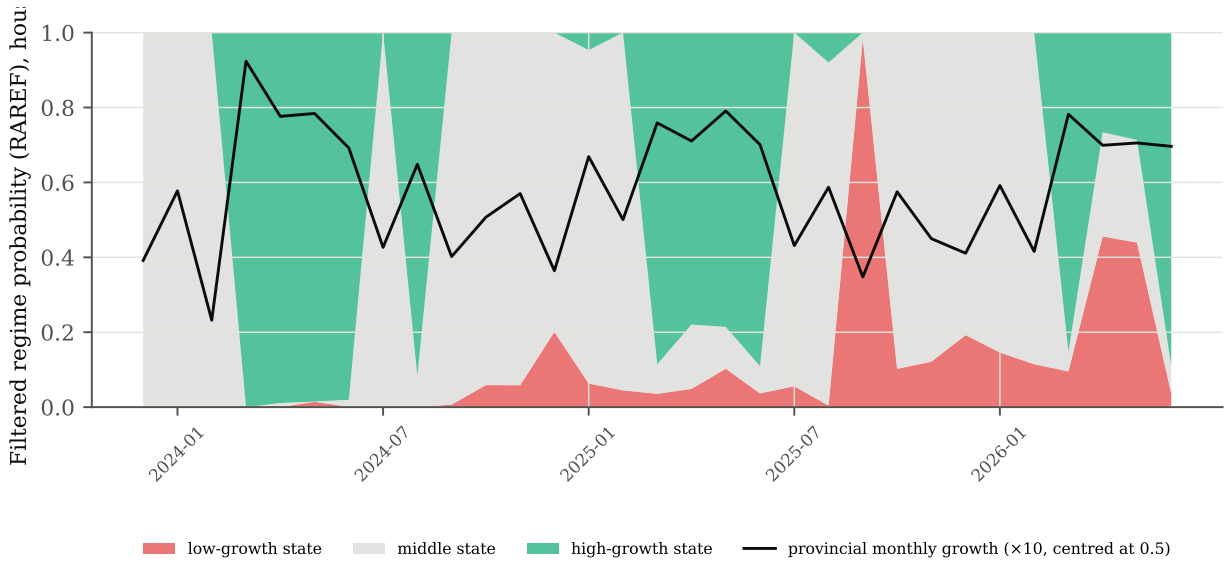

Figure 12. RAREF: filtered probabilities of the three hidden Markov states (ordered by mean growth) at each origin, estimated from the provincial real-time growth known at the origin (houses); the black line is the last observed provincial monthly growth, rescaled.

13. Ablations and robustness

13.1. Component ablations of the proposed architectures

Table 22 removes one component at a time from the Task B architectures at the municipal level and reports the change in MAE with a stationary-bootstrap interval over the per-origin loss differences. The results are the most informative of the paper. For DSGFN, removing the prior graph entirely (uniform prior, attention still learned) does *not* hurt: the variant is as accurate at one and three months and better at six months (-0.59 log points, interval excluding zero), and removing the distance kernel or the co-movement term changes little. The neighbourhood vector that DSGFN learns therefore carries no information beyond what the parent-momentum features already give every pooled model: on 64 municipal nodes and 45 origins, the attention learns a weighted average of other cells' histories that is at best a noisier version of the regional and provincial momentum. The attention diagnostics confirm the reading: the top neighbour receives a weight of 0.19 at three months, the attention entropy is close to that of a uniform distribution over the class's cells, the top neighbour is in the same region only 52% of the time and in the same MRC 12% of the time, with a median distance of 30 km. For QANSS, removing the parent drifts from the state changes the error by -0.02 log points at three months, within the bootstrap interval: the learned transition uses the cell's own recent drift. For the integrated architecture, the equal-weight combination of the four components beats the non-negative least-squares stack at every horizon (-0.30 log points at three months, -0.38 at six), the classical forecast-combination result (Bates and Granger, 1969;

Timmermann, 2006): with 45 origins, the stacking weights overfit. Removing any single component from the stack changes the error by less than the bootstrap interval, except BHPF at three and six months.

The feature-group ablations of the pooled ridge model locate the predictable signal: removing the cell's own momentum features raises the error by +0.30 log points at three months (interval excluding zero) and more at six months, removing the volume features changes it by +0.03, and removing the parent (regional and provincial) features by +0.05. Own momentum is the forecastable component; the parent momentum adds a little at short horizons and nothing at six months once the own momentum is present.

Table 22. Task B ablations at the municipality level: MAE (×100) and difference to the reference variant (per-origin mean, * = 95% stationary-bootstrap interval excludes zero)

Variant	h=1: MAE	vs ref	h=3: MAE	vs ref	h=6: MAE	vs ref
DSGFN (full)	2.66	ref	4.12	ref	5.05	ref
DSGFN without distance kernel	2.64	-0.02	3.95	-0.17	5.08	+0.02
DSGFN without co-movement	2.70	+0.04	4.01	-0.11	4.87	-0.19
DSGFN with uninformative prior graph	2.60	-0.06	3.92	-0.20	4.46	-0.59*
QANSS (full)	2.47	ref	3.75	ref	4.80	ref
QANSS without parent drifts	2.52	+0.05	3.73	-0.02	4.82	+0.02
QREFA (NNLS stack)	2.59	ref	3.75	ref	4.68	ref
Equal-weight combination of the four	2.38	-0.21*	3.45	-0.30	4.30	-0.38
QREFA without QANSS	2.54	-0.05	3.73	-0.02	4.95	+0.26*
QREFA without BHPF	2.53	-0.05	3.88	+0.13	5.04	+0.36
QREFA without DSGFN	2.43	-0.16	3.64	-0.11	4.94	+0.26
QREFA without RAREF	2.62	+0.03	3.65	-0.10	4.64	-0.04
Ridge panel (full features)	2.54	ref	3.73	ref	5.00	ref
Ridge without own momentum	2.65	+0.10*	4.03	+0.30*	5.54	+0.53
Ridge without parent (region/province) features	2.57	+0.02	3.78	+0.05	4.89	-0.11
Ridge without volume features	2.57	+0.03	3.76	+0.03	5.11	+0.10

Reference variants: DSGFN, QANSS, QREFA and the full ridge panel. Positive differences mean the ablated variant is worse.

13.2. Robustness

Table 23 re-runs a subset of Task B models at the three-month horizon under alternative designs. Four facts are stable. Rolling 24-origin windows help the pooled ridge (3.52 against 3.73 log points, as parameter drift in a short sample favours recent data) and hurt the neural models, which need the observations. Dropping the thinnest cell-months ($n_{\text{last}} < 10$) or winsorising the targets changes every model by less than 0.2 log points and leaves the ranking intact. The DSGFN bandwidth (30 or 120 km instead of 60) is irrelevant, consistent with the ablation that the graph carries no information. Most importantly, re-estimating the neural models with different seeds moves their error by as much as the differences between models: DSGFN ranges from 3.66 to 4.05 log points across three seeds and TFT-lite from 4.33 to 4.83, while the ridge and AR benchmarks have no seed. With 45 origins

and 64 cells the neural models are not reliably distinguishable from one another, and any ranking among them in Section 9 should be read with this seed variability in mind; the proposed QANSS is the most stable of the three (3.75, 3.96, 3.87).

Table 23. Task B robustness at the municipality level, horizon 3: MAE (×100) under alternative designs

Model	Expanding (baseline)	Rolling 24 origins	Cells with $n \geq 10$	Winsorised target	Seed 1	Seed 2	Bandwidth 30 / 120 km
No change	4.38	4.38	4.40	4.24	—	—	—
Momentum	4.39	4.38	4.40	4.25	—	—	—
AR(2), cell	3.95	3.92	4.06	3.82	—	—	—
Ridge (panel)	3.73	3.52	3.63	3.59	—	—	—
QANSS	3.75	3.88	3.69	3.72	3.96	3.87	—
TFT-lite	4.33	4.46	4.12	4.21	4.51	4.83	—
DSGFN	3.76	3.90	3.83	3.76	4.05	3.66	—
DSGFN, bandwidth 30 km	—	—	—	—	—	—	4.00
DSGFN, bandwidth 120 km	—	—	—	—	—	—	3.98

Rolling: models estimated on the 24 most recent origins only. Seeds: neural models re-estimated with different random seeds (the baseline is seed 0). Winsorised: targets clipped at the 1st and 99th percentiles in training and test.

14. Limitations and conclusion

14.1. Limitations

The sample is short: 69 months, of which 24 are history, 31 to 45 forecast origins per horizon and one macroeconomic cycle. Tests have low power, model confidence sets are wide, and the twelve-month results describe one episode. The property-level task values a dwelling whose characteristics are known; it does not forecast which dwellings sell, and selection into transaction is documented rather than corrected. The characteristics come from the assessment roll and omit bedrooms, bathrooms, condition and listing information. The neural architectures are trained with fixed hyper-parameters and few epochs, by design, and their errors vary across seeds by as much as the differences between them. The Task B target is itself an estimate (the WP22 final index), and the simulation study uses the WP22 simulator, whose assumptions are ours.

14.2. Conclusion

Do new forecasting architectures substantially improve residential price forecasts? On this data, no. For the price of a future sale, gradient boosting on hedonic and local market features is the most accurate model at every horizon and the hierarchical hedonic network does not improve on it; for the change of a quality-adjusted market index, the forecastable component at one to six months is the cell's own momentum and seasonality, which a pooled tree model or an autoregression captures, and the spatial graph, the regime gating and the hierarchical intercepts add nothing that survives an ablation; at twelve months nothing beats a simple autoregression and every pooled model inherits the bias of the preceding boom. Two of the proposed architectures earn a qualified place: the

quality-adjusted state-space forecaster, because it is stable across seeds and the only model with honest intervals at every horizon on the real data, although it is the worst model in the simulation when the real-time index is noisy; and the integrated architecture, because it is competitive and shows that an equal-weight combination of components is at least as good as a learned stack. The simulation study (Section 7) explains the empirical ranking: where latent dynamics are persistent the pooled models recover the drift, where they are a random walk nothing beats no change, and spatial spillovers are learnable only when they are strong and the sample is long. The value of the exercise is the measured answer and the design that produced it: strict chronology, leakage-audited information sets, inference that respects short samples, and ablations that attribute performance to mechanisms. The code, the panel and every result are released so that the next architecture can be tested the same way.

References

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

Apley, Daniel W. and Jingyu Zhu, "Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2020, *82* (4), 1059–1086.

Athanasopoulos, George, Rob J. Hyndman, Nikolaos Kourentzes, and Anastasios Panagiotelis, "Forecast reconciliation: A review," *International Journal of Forecasting*, 2024, *40* (2), 430–456.

Bai, Lei, Lina Yao, Can Li, Xianzhi Wang, and Can Wang, "Adaptive Graph Convolutional Recurrent Network for Traffic Forecasting," in "Advances in Neural Information Processing Systems 33 (NeurIPS 2020)" 2020.

Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," arXiv preprint arXiv:1803.01271 2018.

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.

Baldominos, Alejandro, Iván Blanco, Antonio José Moreno, Rubén Iturrarte, Óscar Bernárdez, and Carlos Afonso, "Identifying Real Estate Opportunities Using Machine Learning," *Applied Sciences*, 2018, *8* (11), 2321.

Bańbura, Marta, Domenico Giannone, and Lucrezia Reichlin, "Large Bayesian vector auto regressions," *Journal of Applied Econometrics*, 2010, *25* (1), 71–92.

Bates, J. M. and C. W. J. Granger, "The Combination of Forecasts," *Journal of the Operational Research Society*, 1969, *20* (4), 451–468.

Beracha, Eli and Hilla Skiba, "Momentum in Residential Real Estate," *The Journal of Real Estate Finance and Economics*, 2011, *43* (3), 299–320.

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe, "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, 2017, *112* (518), 859–877.

Bork, Lasse and Stig V. Møller, "Forecasting house prices in the 50 states using Dynamic Model Averaging and Dynamic Model Selection," *International Journal of Forecasting*, 2015, *31* (1), 63–78.

Boucher, Simon-Pierre, "SHARP-HPI: A Spatially Hierarchical, Adaptive and Robust Framework for Quality-Adjusted Housing Price Indices," Working Paper 22, Université du Québec en Outaouais, Département des sciences administratives 2026.

Bracher, Johannes, Evan L. Ray, Tilmann Gneiting, and Nicholas G. Reich, "Evaluating epidemic forecasts in an interval format," *PLOS Computational Biology*, 2021, *17* (2), e1008618.

Brady, Ryan R., "Measuring the diffusion of housing prices across space and over time," *Journal of Applied Econometrics*, 2011, *26* (2), 213–231.

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

Capozza, Dennis R., Patric H. Hendershott, and Charlotte Mack, "An Anatomy of Price Dynamics in Illiquid Markets: Analysis and Evidence from Local Housing Markets," *Real Estate Economics*, 2004, *32* (1), 1–32.

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

_ **and** _ , "Forecasting Prices and Excess Returns in the Housing Market," *Real Estate Economics*, 1990, *18* (3), 253–273.

Chen, Tianqi and Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System," in "Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining" ACM 2016, pp. 785–794.

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in "Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)" Association for Computational Linguistics 2014, pp. 1724–1734.

Clapp, John M., Carmelo Giaccotto, and Dogan Tirtiroglu, "Housing Price Indices Based on All Transactions Compared to Repeat Subsamples," *Real Estate Economics*, 1991, *19* (3), 270–285.

Clark, Todd and Michael McCracken, "Advances in Forecast Evaluation," in Graham Elliott and Allan Timmermann, eds., *Handbook of Economic Forecasting, Volume 2B*, Vol. 2, Elsevier, 2013, pp. 1107–1201.

Clark, Todd E and Michael W McCracken, "Tests of equal forecast accuracy and encompassing for nested models," *Journal of Econometrics*, 2001, *105* (1), 85–110.

Das, Sarkar Snigdha Sarathi, Mohammed Eunus Ali, Yuan-Fang Li, Yong-Bin Kang, and Timos Sellis, "Boosting house price predictions using geo-spatial network embedding," *Data Mining and Knowledge Discovery*, 2021, *35* (6), 2221–2250.

de Bézenac, Emmanuel, Syama Sundar Rangapuram, Konstantinos Benidis, Michael Bohlke-Schneider, Richard Kurle, Lorenzo Stella, Hilaf Hasson, Patrick Gallinari, and Tim Januschowski, "Normalizing Kalman Filters for Multivariate Time Series Analysis," in "Advances in Neural Information Processing Systems 33 (NeurIPS 2020)" 2020.

Deppner, Juergen and Marcelo Cajias, "Accounting for Spatial Autocorrelation in Algorithm-Driven Hedonic Models: A Spatial Cross-Validation Approach," *The Journal of Real Estate Finance and Economics*, 2024, *68* (2), 235–273.

Diebold, Francis X., "Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests," *Journal of Business & Economic Statistics*, 2015, *33* (1).

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

Elhorst, J. Paul, *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels* SpringerBriefs in Regional Science, Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

Eurostat, ILO, IMF, OECD, UNECE, and World Bank, *Handbook on Residential Property Price Indices (RPPIs)*, OECD Publishing / Eurostat, 2013.

Fay, Robert E. and Roger A. Herriot, "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," *Journal of the American Statistical Association*, 1979, *74* (366), 269–277.

Francke, Marc K., "Repeat Sales Index for Thin Markets," *The Journal of Real Estate Finance and Economics*, 2010, *41* (1), 24–52.

_ **and Alex van de Minne**, "The Hierarchical Repeat Sales Model for Granular Commercial Real Estate and Residential Price Indices," *The Journal of Real Estate Finance and Economics*, 2017, *55* (4), 511–532.

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

Gatzlaff, Dean H. and Donald R. Haurin, "Sample Selection Bias and Repeat-Sales Index Estimates," *The Journal of Real Estate Finance and Economics*, 1997, *14* (1-2), 33–50.

_ **and** _ , "Sample Selection and Biases in Local House Value Indices," *Journal of Urban Economics*, 1998, *43* (2), 199–222.

Gelman, Andrew, "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)," *Bayesian Analysis*, 2006, *1* (3), 515–534.

_ **and Jennifer Hill**, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, 2007.

Ghysels, Eric, Alberto Plazzi, Rossen Valkanov, and Walter Torous, "Forecasting Real Estate Prices," in Graham Elliott and Allan Timmermann, eds., *Handbook of Economic Forecasting, Volume 2A*, Vol. 2, Elsevier, 2013, pp. 509–580.

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

Glaeser, Edward L., Joseph Gyourko, Eduardo Morales, and Charles G. Nathanson, "Housing dynamics: An urban approach," *Journal of Urban Economics*, 2014, *81*, 45–56.

Gneiting, Tilmann and Adrian E Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 2007, *102* (477), 359–378.

_ **and Matthias Katzfuss**, "Probabilistic Forecasting," *Annual Review of Statistics and Its Application*, 2014, *1* (1), 125–151.

_ **, Fadoua Balabdaoui, and Adrian E. Raftery**, "Probabilistic Forecasts, Calibration and Sharpness," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2007, *69* (2), 243–268.

Guo, Yameng and Seppe vanden Broucke, "HeteroRoadNet: A Heterogenous Road-based Graph Neural Network Approach for Property Price Prediction," *ACM Transactions on Spatial Algorithms and Systems*, 2026, *12* (3), 1–21.

Gupta, Rangan, Alain Kabundi, and Stephen M. Miller, "Forecasting the US real house price index: Structural and non-structural models with and without fundamentals," *Economic Modelling*, 2011, *28* (4), 2013–2021.

_ **,** _ **, and Stephen Miller**, "Using Large Data Sets to Forecast House Prices: A Case Study of Twenty U.S. States," *Journal of Housing Research*, 2011, *20* (2), 161–190.

Guren, Adam M., "House Price Momentum and Strategic Complementarity," *Journal of Political Economy*, 2018, *126* (3), 1172–1218.

Hamilton, James D., "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, 1989, *57* (2), 357–384.

Hansen, Peter R., Asger Lunde, and James M. Nason, "The Model Confidence Set," *Econometrica*, 2011, *79* (2), 453–497.

Harvey, David, Stephen Leybourne, and Paul Newbold, "Testing the equality of prediction mean squared errors," *International Journal of Forecasting*, 1997, *13* (2), 281–291.

Head, Allen, Huw Lloyd-Ellis, and Hongfei Sun, "Search, Liquidity, and the Dynamics of House Prices and Construction," *American Economic Review*, 2014, *104* (4), 1172–1210.

Heckman, James J., "Sample Selection Bias as a Specification Error," *Econometrica*, 1979, *47* (1), 153–161.

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

_ , Michael Scholz, Chihiro Shimizu, and Miriam Steurer, "Rolling-Time-Dummy House Price Indexes: Window Length, Linking and Options for Dealing with Low Transaction Volume," *Journal of Official Statistics*, 2022, *38* (1), 127–151.

Hochreiter, Sepp and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, 1997, *9* (8), 1735–1780.

Hoffman, Matthew D. and Andrew Gelman, "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo," *Journal of Machine Learning Research*, 2014, *15* (47), 1593–1623.

Holly, Sean, M. Hashem Pesaran, and Takashi Yamagata, "A spatio-temporal model of house prices in the USA," *Journal of Econometrics*, 2010, *158* (1), 160–173.

_ , _ , and _ , "The spatial and temporal diffusion of house prices in the UK," *Journal of Urban Economics*, 2011, *69* (1), 2–23.

Hong, Jengei, Heeyoul Choi, and Woo sung Kim, "A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea," *International Journal of Strategic Property Management*, 2020, *24* (3), 140–152.

Hwang, Min and John M. Quigley, "Selectivity, Quality Adjustment and Mean Reversion in the Measurement of House Values," *The Journal of Real Estate Finance and Economics*, 2004, *28* (2-3), 161–178.

Hyndman, Rob J. and Anne B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, 2006, *22* (4), 679–688.

_ , Roman A. Ahmed, George Athanasopoulos, and Han Lin Shang, "Optimal combination forecasts for hierarchical time series," *Computational Statistics & Data Analysis*, 2011, *55* (9), 2579–2589.

Ishaak, Farley, Pim Ouwehand, and Hilde Remøy, "Constructing Limited-Revisable and Stable CPPIs for Small Domains," *Journal of Official Statistics*, 2024, *40* (3), 380–408.

Jacobs, Robert A., Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton, "Adaptive Mixtures of Local Experts," *Neural Computation*, 1991, *3* (1), 79–87.

Jin, Guangyin, Yuxuan Liang, Yuchen Fang, Zezhi Shao, Jincai Huang, Junbo Zhang, and Yu Zheng, "Spatio-Temporal Graph Neural Networks for Predictive Learning in Urban Computing: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, 2024, *36* (10), 5388–5408.

Jin, Ming, Huan Yee Koh, Qingsong Wen, Daniele Zambon, Cesare Alippi, Geoffrey I. Webb, Irwin King, and Shirui Pan, "A Survey on Graph Neural Networks for Time Series: Forecasting, Classification, Imputation, and Anomaly Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, *46* (12), 10466–10485.

Jordan, Michael I. and Robert A. Jacobs, "Hierarchical Mixtures of Experts and the EM Algorithm," *Neural Computation*, 1994, *6* (2), 181–214.

Jud, Donald and Terry Seaks, "Sample Selection Bias in Estimating Housing Sales Prices," *Journal of Real Estate Research*, 1994, *9* (3), 289–298.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in "Advances in Neural Information Processing Systems 30 (NIPS 2017)" 2017, pp. 3146–3154.

Kelejian, Harry H. and Ingmar R. Prucha, "Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances," *Journal of Econometrics*, 2010, *157* (1), 53–67.

Kim, Chang-Jin and Charles R. Nelson, *State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications*, Cambridge, MA: MIT Press, 1999.

Kipf, Thomas N. and Max Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in "Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)" 2017. arXiv:1609.02907.

Koenker, Roger and Gilbert Bassett, "Regression Quantiles," *Econometrica*, 1978, *46* (1), 33–50.

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

Kouwenberg, Roy and Remco Zwinkels, "Forecasting the US housing market," *International Journal of Forecasting*, 2014, *30* (3), 415–425.

Krishnan, Rahul, Uri Shalit, and David Sontag, "Structured Inference Networks for Nonlinear State Space Models," in "Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)" Association for the Advancement of Artificial Intelligence (AAAI) 2017. Vol. 31, No. 1.

Künsch, Hans R., "The Jackknife and the Bootstrap for General Stationary Observations," *The Annals of Statistics*, 1989, *17* (3), 1217–1241.

Lenaers, Ian and Lieven De Moor, "Exploring XAI techniques for enhancing model transparency and interpretability in real estate rent prediction: A comparative study," *Finance Research Letters*, 2023, *58*, 104306.

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

Li, Yaguang, Rose Yu, Cyrus Shahabi, and Yan Liu, "Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting," in "Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)" 2018. arXiv:1707.01926.

Lim, Bryan, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister, "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, 2021, *37* (4), 1748–1764.

Litterman, Robert B., "Forecasting With Bayesian Vector Autoregressions—Five Years of Experience," *Journal of Business & Economic Statistics*, 1986, *4* (1), 25–38.

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

Lundberg, Scott M. and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," in "Advances in Neural Information Processing Systems 30 (NIPS 2017)" 2017.

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

Makridakis, Spyros, Evangelos Spiliotis, and Vassilios Assimakopoulos, "The M4 Competition: 100,000 time series and 61 forecasting methods," *International Journal of Forecasting*, 2020, *36* (1), 54–74.

_ , _ , and _ , "M5 accuracy competition: Results, findings, and conclusions," *International Journal of Forecasting*, 2022, *38* (4), 1346–1364.

_ , _ , _ , Zhi Chen, Anil Gaba, Ilia Tsetlin, and Robert L. Winkler, "The M5 uncertainty competition: Results, findings and conclusions," *International Journal of Forecasting*, 2022, *38* (4), 1365–1385.

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.

Milunovich, George, "Forecasting Australia's real house price index: A comparison of time series and machine learning methods," *Journal of Forecasting*, 2020, *39* (7), 1098–1118.

Molnar, Christoph, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 3 ed., Self-published, 2025. ISBN 978-3-911578-03-5.

Munneke, Henry J. and Barrett A. Slade, "An Empirical Study of Sample-Selection Bias in Indices of Commercial Real Estate," *The Journal of Real Estate Finance and Economics*, 2000, *21* (1), 45–64.

Nie, Yuqi, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam, "A Time Series is Worth 64 Words: Long-term Forecasting with Transformers," in "Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)" 2023. arXiv:2211.14730.

Oreshkin, Boris N., Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio, "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting," in "Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)" 2020. arXiv:1905.10437.

Pace, R. Kelley and Darren Hayunga, "Examining the Information Content of Residuals from Hedonic and Spatial Models Using Trees and Forests," *The Journal of Real Estate Finance and Economics*, 2020, *60* (1-2), 170–180.

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

_ , _ , Otis W. Gilley, and C. F. Sirmans, "A method for spatial–temporal forecasting with an application to real estate prices," *International Journal of Forecasting*, 2000, *16* (2), 229–246.

Plakandaras, Vasilios, Rangan Gupta, Periklis Gogas, and Theophilos Papadimitriou, "Forecasting the U.S. real house price index," *Economic Modelling*, 2015, *45*, 259–267.

Politis, Dimitris N. and Joseph P. Romano, "The Stationary Bootstrap," *Journal of the American Statistical Association*, 1994, *89* (428), 1303–1313.

Prokhorenkova, Liudmila, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin, "CatBoost: unbiased boosting with categorical features," in "Advances in Neural Information Processing Systems 31 (NeurIPS 2018)" 2018.

Rangapuram, Syama Sundar, Matthias W. Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski, "Deep State Space Models for Time Series Forecasting," in "Advances in Neural Information Processing Systems 31 (NeurIPS 2018)" 2018.

Rapach, David E. and Jack K. Strauss, "Differences in housing price forecastability across US states," *International Journal of Forecasting*, 2009, *25* (2), 351–372.

Romano, Yaniv, Evan Patterson, and Emmanuel J. Candès, "Conformalized Quantile Regression," in "Advances in Neural Information Processing Systems 32 (NeurIPS 2019)" 2019.

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

Rue, Håvard, Sara Martino, and Nicolas Chopin, "Approximate Bayesian Inference for Latent Gaussian models by using Integrated Nested Laplace Approximations," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2009, *71* (2), 319–392.

Salinas, David, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski, "DeepAR: Probabilistic forecasting with autoregressive recurrent networks," *International Journal of Forecasting*, 2020, *36* (3), 1181–1191.

Shazeer, Noam, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean, "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer," arXiv preprint arXiv:1701.06538 2017.

Sijp, Willem P and Marc K. Francke, "Graph-Laplacian modeling of spatiotemporal effects for house price estimation," *Real Estate Economics*, 2026. Early view, article 1540-6229.70062.

Stock, James H and Mark W Watson, "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association*, 2002, *97* (460), 1167–1179.

_ and _ , "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business & Economic Statistics*, 2002, *20* (2), 147–162.

Taylor, James W., "A quantile regression neural network approach to estimating the conditional density of multiperiod returns," *Journal of Forecasting*, 2000, *19* (4), 299–311.

Timmermann, Allan, "Forecast Combinations," in Graham Elliott, Clive W. J. Granger, and Allan Timmermann, eds., *Handbook of Economic Forecasting, Volume 1*, Vol. 1, Elsevier, 2006, pp. 135–196.

Titman, Sheridan, Ko Wang, and Jing Yang, "The Dynamics of Housing Prices," *Journal of Real Estate Research*, 2014, *36* (3), 283–318.

Veličković, Petar, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio, "Graph Attention Networks," in "Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)" 2018. arXiv:1710.10903.

Viana, Darniton and Luciano Barbosa, "Attention-Based Spatial Interpolation for House Price Prediction," in "Proceedings of the 29th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '21)" ACM 2021, pp. 540–549.

Vovk, Vladimir, Alexander Gammerman, and Glenn Shafer, *Algorithmic Learning in a Random World*, New York: Springer, 2005.

West, Kenneth D., "Asymptotic Inference about Predictive Ability," *Econometrica*, 1996, *64* (5), 1067–1084.

Wickramasuriya, Shanika L., George Athanasopoulos, and Rob J. Hyndman, "Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization," *Journal of the American Statistical Association*, 2019, *114* (526), 804–819.

Wu, Zonghan, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu, "A Comprehensive Survey on Graph Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2021, *32* (1), 4–24.

_ , _ , Guodong Long, Jing Jiang, and Chengqi Zhang, "Graph WaveNet for Deep Spatial-Temporal Graph Modeling," in "Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)" International Joint Conferences on Artificial Intelligence Organization 2019, pp. 1907–1913.

Yeo, Jisu, Artyom Stitsyuk, and Jaesik Choi, "Deep-DFVAR: Dynamic factor vector autoregression with deep learning for regional house price index forecasting," *Neurocomputing*, 2025, *656*, 131103.

Yu, Bing, Haoteng Yin, and Zhanxing Zhu, "Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting," in "Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)" International Joint Conferences on Artificial Intelligence Organization 2018, pp. 3634–3640.

Zeng, Ailing, Muxi Chen, Lei Zhang, and Qiang Xu, "Are Transformers Effective for Time Series Forecasting?," in "Proceedings of the AAAI Conference on Artificial Intelligence" Association for the Advancement of Artificial Intelligence (AAAI) 2023, pp. 11121–11128. Vol. 37, No. 9.

Zhang, Weijia, Hao Liu, Lijun Zha, Hengshu Zhu, Ji Liu, Dejing Dou, and Hui Xiong, "MugRep: A Multi-Task Hierarchical Graph Representation Learning Framework for Real Estate Appraisal," in "Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21)" ACM 2021, pp. 3937–3947.

Appendices

A. Architectures: formal specification and diagrams

This appendix gives the equations and block diagrams of the six proposed architectures as implemented in `src/wp23/models_a_torch.py`, `models_b.py` and `models_b_torch.py`. Notation: x standardised feature vector, $\mathbf{h}$ the thirteen-month relative history of the real-time index ($h_j = y^{\text{rt}}_{o-12+j} - y^{\text{rt}}_{o}$), $\tau \in \{0.05, 0.25, 0.5, 0.75, 0.95\}$ the quantile levels, $\ell_\tau(u) = \max(\tau u, (\tau - 1)u)$ the pinball loss.

HSHN (Task A). With $g_i = e_{\mathrm{reg}(i)} + e_{\mathrm{mrc}(i)} + e_{\mathrm{mun}(i)} \in \mathbb{R}^8$ and $e_h \in \mathbb{R}^4$,

$$z_i = \mathrm{GELU}\big(W_2\,\mathrm{GELU}(W_1[x_i, g_i, e_{h_i}])\big), \quad v_i = \mathrm{GELU}\big(W_3[z_i, g_i, e_{h_i}]\big), \quad \hat{q}_i = W_4[z_i, v_i] \in \mathbb{R}^5, \quad (1)$$

trained by $\min \sum_i \sum_\tau \ell_\tau(\log P_i - \hat{q}_{i,\tau}) + \lambda_{\mathrm{mun}}\|e_{\mathrm{mun}}\|^2 + \lambda\,\|\theta\|^2$ with $\lambda_{\mathrm{mun}} = 5\times 10^{-3}$ (partial), $\lambda_{\mathrm{mun}} = 0$ (none) or $e_{\mathrm{mun}} \equiv 0$ (full).

DSGFN (Task B). Node encodings $z_u = \phi([x_u, \mathbf{h}_u])$; attention logits $a_{uv} = q_u^\top k_v/\sqrt{d} + \omega \log A_{uv}$ for $v \neq u$, $A_{uv} = \exp(-(d_{uv}/60\,\mathrm{km})^2)\,(\frac{1}{2} + \frac{1}{2}\rho_{uv})\mathbb{1}[\text{same level, class, origin}]$ with ρ_{uv} the correlation of the two nodes' monthly real-time changes over the history; $n_u = \sum_v \mathrm{softmax}_v(a_{uv})\,V z_v$; $\hat{q}_u = \psi([z_u, n_u])$.

RAREF (Task B). Provincial real-time growth $g_1, \ldots, g_T$ (one per past origin, each from its own vintage); three-state Gaussian HMM $(\pi, A, \mu_k, \sigma_k)$ by Baum–Welch; filtered probabilities $p_{k,o} = P(s_o = k \mid g_{1:o})$; experts $\hat{\beta}_k = \arg\min_\beta \sum_i p_{k,o(i)}(y_i - x_i^\top \beta)^2 + \alpha\|\beta\|^2$; forecast $\sum_k p_{k,o}\, x^\top \hat{\beta}_k$.

BHPF (Task B). $y = \beta_0 + x^\top\beta + u_r + u_c + \varepsilon,\ \beta \sim N(0, \tau_b^2 I),\ u_r \sim N(0, \tau_r^2),\ u_c \sim N(0, \tau_c^2)$, $\varepsilon \sim N(0, \sigma^2)$; posterior $N(\hat{\theta}, \Sigma)$ with $\Sigma = (D^\top D/\sigma^2 + \Lambda)^{-1}$, $\hat{\theta} = \Sigma D^\top y/\sigma^2$; variance components updated as $\tau_r^2 \leftarrow \mathrm{mean}(\hat{u}_r^2) + \mathrm{mean}(\mathrm{diag}\,\Sigma_{rr})$ (and likewise for τ_c^2, σ^2) for fifteen iterations; predictive $N(d^\top\hat{\theta}, d^\top\Sigma d + \sigma^2)$.

QANSS (Task B). State at the origin $s_o = (\bar{\Delta}_3, \mathrm{sd}(\Delta), n\text{-ratio} - 1, \Delta_3^{\mathrm{prov}}/3, \Delta_3^{\mathrm{reg}}/3, \Delta_{-1}, \Delta_{-2})$; network $(m(s), \log\varsigma(s)) = W_2 \tanh(W_1 s)$; forecast $h\,m(s_o)$ with predictive variance $\varsigma(s_o)^2 + \mathrm{se}_o^2$, estimated by maximising the Gaussian likelihood of the training targets.

QREFA. $\hat{y} = \sum_k w_k \hat{y}_k$ with $w = \arg\min_{w\geq 0} \|y - Fw\|^2$ on the training origins, normalised to sum to one; quantiles $\sum_k w_k \hat{q}_k$.

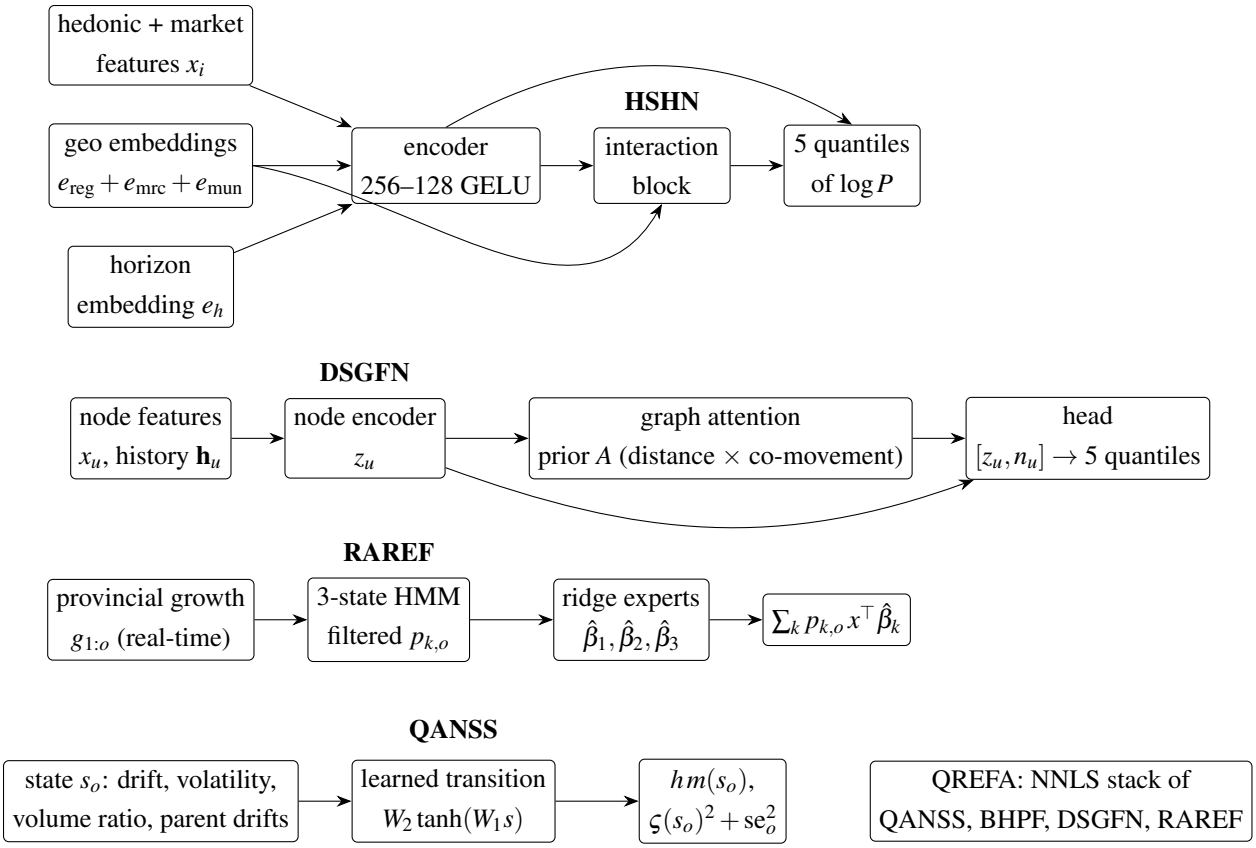

Figure 13. Block diagrams of the proposed architectures (BHPF is a conjugate linear-Gaussian hierarchy without a diagram).

B. Hyper-parameters

Table 24 lists the hyper-parameters of every model. They were fixed before any test origin was evaluated and were not tuned; the seed variability of Section 13 measures the uncertainty that tuning would have to overcome.

Table 24. Hyper-parameters

Model	Hyper-parameters (fixed in advance, never tuned on test origins)
HSHN	embeddings 8 (geo) / 4 (h); encoder 256-128 GELU, dropout 0.1; AdamW, one-cycle lr 2×10^{-3}, 8 epochs, batch 4,096; municipality weight decay 5×10^{-3}
MLP (A)	256-128 GELU, dropout 0.1; pinball loss on 5 quantiles; same optimiser
LightGBM (A)	800 trees, lr 0.04, 63 leaves, min child 30, subsample 0.8; quantile models 400 trees
XGBoost (A)	600 trees, lr 0.05, depth 8, subsample 0.8
Random forest (A)	150 trees, depth 18, min leaf 5, max features 0.5
SVR (A)	RBF, C=3, ε=0.05, 15,000-sale subsample
Panel MLP / LSTM / GRU / TCN / TFT-lite (B)	hidden 64 (MLP) / 32 (RNN, attention) / 16 channels (TCN); 60 epochs, batch 512, AdamW lr 2×10^{-3}, weight decay 10^{-4}; pinball loss
DSGFN (B)	node encoder 32; one attention layer; prior adjacency: Gaussian distance kernel (60 km) $\times$ co-movement; 60 epochs
RAREF (B)	3-state Gaussian HMM (60 EM iterations) on provincial real-time growth; 3 ridge experts (α=3)
BHPF (B)	conjugate Gaussian hierarchy; 15 variance-component iterations; priors τ_b^2=1 (slopes)
QANSS (B)	32 tanh units; 200 Adam steps, lr 5×10^{-3}; Gaussian likelihood
QREFA (B)	NNLS stacking of QANSS, BHPF, DSGFN (40 epochs), RAREF on training origins
LightGBM (B)	300 trees, lr 0.03, 15 leaves, min child 40; quantile models
Ridge (B)	α=3 on standardised features

C. Leakage audit

Table 25 records, for each item of the protocol's audit list, where chronological separation is enforced in the code; the test suite (`tests/test_wp23.py`) checks that the Task A features of test sales are computed from sales dated on or before the origin.

Table 25. Leakage audit

Item	Treatment
Hedonic coefficients	WP22 stage 1 re-estimated at every vintage; Task A features never use coefficients
Geographic clustering / location cells	WP22 nested grid rebuilt from data $\leq$ origin at every vintage; Task A uses raw coordinates
Normalisation	standardisation constants from training rows only; index levels enter as changes
Feature selection	fixed feature lists, no selection on test data
Imputation	WP22 conditional medians (no price information); indicators time-stable
Target encoding	none; one-hot with training categories
Spatial graph	DSGFN adjacency recomputed at each origin from centroids and the real-time history $\leq$ origin
Hyper-parameter tuning	fixed a priori (Table 24)
Index revisions	predictors = vintage published for the origin; target = final series (revision component reported)
Partial last month	October 2026 excluded; last target September 2026

D. Forecast comparison tests

Let $d_t = L_{A,t} - L_{B,t}$ be the difference of per-origin mean squared losses over n origins. The Diebold–Mariano statistic is $\bar{d}/\sqrt{\hat{\sigma}^2_{\mathrm{HAC}}/n}$ with a Bartlett long-run variance of $h-1$ lags, multiplied by the Harvey–Leybourne–Newbold factor $\sqrt{(n+1-2h+h(h-1)/n)/n}$ and compared with a t_{n-1} distribution. The Giacomini–White statistic regresses d_t on $z_{t-1} = (1, d_{t-1})$ and tests $\mathbb{E}[z_{t-1}d_t] = 0$ by a Wald statistic with the same HAC covariance, distributed χ^2_2. The model confidence set starts from all models, computes $t_i = \bar{d}_{i\cdot}/\hat{\sigma}_i$ for the deviation of each model's mean loss from the cross-model mean, eliminates the model with the largest t_i when the stationary-bootstrap p-value of $\max_i t_i$ (500 resamples, mean block length $\max(2,h)$) is below 10%, and stops otherwise; reported p-values are the monotone MCS p-values of Hansen et al. (2011). Holm step-down adjusted p-values of the Diebold–Mariano tests across the models of a horizon are given in Table 26; Table 27 gives the unadjusted p-values and the Giacomini–White p-values.

Table 26. Task B — Diebold–Mariano statistics against the no-change forecast (negative = better than no change) and Holm-adjusted p-values across the 18 models of each horizon

Model	h=1: DM	h=1: Holm p	h=3: DM	h=3: Holm p	h=6: DM	h=6: Holm p	h=12: DM	h=12: Holm p
Mean growth	+0.26	1.00	-0.59	1.00	-0.73	1.00	-0.26	1.00
Momentum	-1.00	1.00	-0.96	1.00	+0.57	1.00	+0.13	1.00
Random walk with drift	+2.40	0.33	+1.14	1.00	+0.90	1.00	+0.29	1.00
Seasonal naive	+5.00	0.00	+1.12	1.00	+1.00	1.00	+0.29	1.00
AR(2), cell	-1.10	1.00	-1.86	0.87	-1.67	1.00	-0.38	1.00
Local linear trend	+7.19	0.00	+8.08	0.00	+10.47	0.00	+4.88	0.01
Ridge (panel)	-1.90	0.80	-2.02	0.75	-0.92	1.00	+0.13	1.00
LightGBM (panel)	-3.85	0.01	-2.91	0.11	-1.79	1.00	-0.19	1.00
MLP (panel)	+0.73	1.00	-1.23	1.00	-0.52	1.00	+0.55	1.00
LSTM	+1.88	0.80	-1.02	1.00	-1.17	1.00	+0.07	1.00
GRU	+1.58	0.99	-1.67	1.00	-1.23	1.00	+0.14	1.00
TCN	+0.68	1.00	-1.29	1.00	-1.23	1.00	+0.15	1.00
TFT-lite	+1.24	1.00	+0.23	1.00	-0.16	1.00	-0.04	1.00
BHPF	-1.79	0.82	-1.92	0.83	-0.74	1.00	+0.11	1.00
RAREF	-1.99	0.72	-1.74	1.00	-0.93	1.00	+0.11	1.00
QANSS	-2.27	0.42	-2.18	0.59	-1.27	1.00	+0.27	1.00
DSGFN	-0.52	1.00	-1.24	1.00	-0.91	1.00	+0.10	1.00
QREFA	-1.74	0.82	-2.04	0.75	-1.27	1.00	+0.05	1.00

HAC variance with $h-1$ Bartlett lags and the Harvey–Leybourne–Newbold correction; t_{n-1} reference distribution; Holm step-down over the 18 two-sided p-values of each horizon.

Table 27. Task B — unadjusted Diebold–Mariano p-values and Giacomini–White conditional-predictive-ability p-values against the no-change forecast

Model	h=1: DM p	h=1: GW p	h=3: DM p	h=3: GW p	h=6: DM p	h=6: GW p	h=12: DM p	h=12: GW p
Mean growth	0.800	0.972	0.560	0.001	0.473	0.003	0.798	0.000
Momentum	0.324	0.075	0.344	0.063	0.574	0.069	0.896	0.376
Random walk with drift	0.022	0.044	0.261	0.000	0.375	0.000	0.773	0.001
Seasonal naive	0.000	0.000	0.270	0.001	0.327	0.000	0.773	0.001
AR(2), cell	0.278	0.496	0.072	0.099	0.109	0.035	0.712	0.000
Local linear trend	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Ridge (panel)	0.067	0.103	0.052	0.098	0.367	0.004	0.898	0.010
LightGBM (panel)	0.001	0.000	0.007	0.005	0.086	0.067	0.855	0.011
MLP (panel)	0.471	0.698	0.229	0.327	0.610	0.003	0.595	0.000
LSTM	0.069	0.150	0.314	0.184	0.254	0.002	0.945	0.107
GRU	0.124	0.243	0.106	0.190	0.229	0.007	0.894	0.007
TCN	0.501	0.649	0.208	0.220	0.229	0.045	0.880	0.046
TFT-lite	0.223	0.304	0.817	0.261	0.876	0.011	0.966	0.000
BHPF	0.082	0.118	0.064	0.120	0.469	0.001	0.918	0.000
RAREF	0.055	0.075	0.092	0.171	0.364	0.001	0.915	0.016
QANSS	0.030	0.027	0.037	0.066	0.217	0.001	0.792	0.004
DSGFN	0.605	0.320	0.223	0.154	0.372	0.015	0.919	0.000
QREFA	0.092	0.038	0.050	0.096	0.215	0.018	0.959	0.000

The Giacomini–White test regresses the loss differential on a constant and its own lag; a rejection means the relative accuracy of the two forecasts is predictable from the past.

E. Supplementary tables

Tables 28 to 33 supplement the main text: Task A errors by municipality size and by price segment (the breakdown by class is Table 17 in the main text); Task B errors by cell size and level; the regional and provincial simulation tables of Section 7; and the computational cost of every model, measured on one machine so that the accuracy gains of Section 9 can be weighed against their price in computing time (the tree models cost seconds per fit, the neural models minutes, and BHPF, which solves a dense linear system, is the most expensive Task B model).

Table 28. Task A — MAE (×100) by municipality size, horizon 4-6

Model	large	medium	small
Naive: municipal median	33.47	33.56	44.33
Hedonic OLS	20.70	23.88	33.29
Ridge	20.70	23.88	33.29
Elastic net	20.62	23.80	33.26
SVR (subsample)	19.67	23.41	35.09
Random forest	19.31	22.88	32.18
XGBoost	18.79	22.65	32.32
LightGBM	18.71	22.35	31.77
Quantile MLP	22.05	26.30	37.51
HSHN (no pooling)	20.61	23.78	32.90
HSHN (full pooling)	20.96	24.22	32.91
HSHN (partial pooling)	20.63	23.79	32.92

Table 29. Task A — MAE (×100) by price segment, horizon 4-6

Model	250-500k	<250k	>500k
Naive: municipal median	24.58	54.82	40.35
Hedonic OLS	18.81	41.38	22.23
Ridge	18.81	41.38	22.23
Elastic net	18.57	41.60	22.25
SVR (subsample)	17.85	44.09	21.16
Random forest	17.20	39.63	21.80
XGBoost	17.78	38.75	20.54
LightGBM	17.44	38.90	20.28
Quantile MLP	21.46	48.19	21.59
HSHN (no pooling)	18.01	41.11	23.02
HSHN (full pooling)	18.27	41.14	23.58
HSHN (partial pooling)	18.05	41.11	23.03

Table 30. Task B — MAE (×100) by cell size, horizon 3

Model	mun_large	mun_small	province	region
No change	3.79	4.47	2.44	3.58
Mean growth	3.99	4.26	2.00	3.25
Momentum	3.88	4.48	2.73	3.73
Random walk with drift	3.94	4.62	1.90	3.27
Seasonal naive	3.64	4.89	1.37	2.85
AR(2), cell	3.72	3.99	2.06	3.14
Local linear trend	6.74	6.85	3.16	5.22
Ridge (panel)	3.21	3.82	2.06	2.78
LightGBM (panel)	2.90	3.44	1.52	2.47
MLP (panel)	3.51	4.12	2.08	3.04
LSTM	3.65	4.12	2.36	3.31
GRU	3.37	3.93	2.06	3.06
TCN	3.56	4.05	1.91	3.27
TFT-lite	3.75	4.44	2.09	3.46
BHPF	3.20	3.87	2.11	2.86
RAREF	3.39	3.87	2.12	2.82
QANSS	3.06	3.86	1.77	2.57
DSGFN	3.44	4.02	1.90	2.84
QREFA	3.14	3.76	1.66	2.57

Table 31. Simulation — MAE of 3-month index-change forecasts relative to the no-change forecast at the region level, by scenario (Monte Carlo s.e. over replications)

Model	rw	momentum	regimes	spillover	sparse
No change	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
Mean growth	0.72 (0.06)	1.14 (0.06)	0.80 (0.10)	0.97 (0.05)	1.14 (0.06)
Momentum	0.96 (0.01)	0.98 (0.01)	0.83 (0.04)	0.91 (0.01)	1.00 (0.01)
Random walk with drift	0.80 (0.06)	1.05 (0.05)	0.70 (0.07)	0.92 (0.03)	1.08 (0.07)
Seasonal naive	1.43 (0.12)	1.66 (0.14)	1.14 (0.14)	1.33 (0.07)	2.00 (0.17)
AR(2), cell	0.71 (0.06)	1.06 (0.05)	0.75 (0.09)	0.90 (0.03)	1.08 (0.06)
Local linear trend	1.49 (0.10)	1.44 (0.11)	0.80 (0.06)	1.13 (0.04)	1.87 (0.18)
Ridge (panel)	1.13 (0.15)	1.71 (0.09)	1.20 (0.22)	1.21 (0.10)	1.65 (0.14)
LightGBM (panel)	0.84 (0.07)	1.23 (0.06)	0.76 (0.07)	0.99 (0.04)	1.20 (0.07)
MLP (panel)	0.87 (0.09)	1.36 (0.06)	0.88 (0.13)	1.15 (0.11)	1.42 (0.17)
LSTM	0.87 (0.13)	1.20 (0.04)	0.89 (0.14)	1.04 (0.10)	1.41 (0.18)
TCN	0.86 (0.13)	1.25 (0.07)	0.89 (0.14)	1.05 (0.11)	1.37 (0.17)
TFT-lite	0.76 (0.10)	1.16 (0.06)	0.85 (0.17)	0.96 (0.08)	1.20 (0.13)
BHPF	1.15 (0.16)	1.76 (0.11)	1.21 (0.21)	1.22 (0.12)	1.75 (0.15)
RAREF	1.20 (0.16)	1.71 (0.10)	1.25 (0.20)	1.23 (0.14)	1.68 (0.11)
QANSS	1.89 (0.25)	1.51 (0.14)	1.27 (0.28)	1.07 (0.07)	1.49 (0.07)
DSGFN	1.06 (0.12)	1.40 (0.09)	0.94 (0.12)	1.20 (0.13)	1.42 (0.15)

rw: random walk with a constant drift; momentum: AR(1) drifts at every level; regimes: Markov-switching provincial drift; spillover: municipal drift follows the lagged regional drift; sparse: momentum with 35% of the sales. Values below 1 beat the no-change forecast. Classic and tree models: 10 replications; neural models: 6.

Table 32. Simulation — MAE of 3-month index-change forecasts relative to the no-change forecast at the province level, by scenario (Monte Carlo s.e. over replications)

Model	rw	momentum	regimes	spillover	sparse
No change	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
Mean growth	0.68 (0.07)	1.12 (0.09)	0.79 (0.11)	0.94 (0.13)	1.12 (0.09)
Momentum	0.96 (0.01)	0.98 (0.01)	0.82 (0.04)	0.89 (0.02)	1.00 (0.01)
Random walk with drift	0.76 (0.06)	1.05 (0.08)	0.66 (0.07)	0.84 (0.09)	1.05 (0.08)
Seasonal naive	1.33 (0.12)	1.65 (0.19)	1.05 (0.14)	1.32 (0.16)	2.02 (0.23)
AR(2), cell	0.68 (0.08)	1.10 (0.08)	0.74 (0.10)	0.89 (0.11)	1.11 (0.09)
Local linear trend	1.37 (0.09)	1.35 (0.13)	0.70 (0.07)	1.11 (0.11)	1.89 (0.23)
Ridge (panel)	1.18 (0.16)	1.87 (0.15)	1.22 (0.22)	1.59 (0.26)	1.98 (0.20)
LightGBM (panel)	0.83 (0.10)	1.27 (0.13)	0.75 (0.08)	0.92 (0.11)	1.26 (0.12)
MLP (panel)	0.80 (0.11)	1.36 (0.13)	0.79 (0.11)	1.19 (0.27)	1.55 (0.28)
LSTM	0.87 (0.14)	1.29 (0.12)	0.87 (0.15)	1.10 (0.24)	1.48 (0.25)
TCN	0.82 (0.13)	1.31 (0.17)	0.85 (0.15)	1.08 (0.26)	1.38 (0.22)
TFT-lite	0.71 (0.11)	1.19 (0.13)	0.79 (0.15)	0.87 (0.12)	1.22 (0.17)
BHPF	1.19 (0.17)	2.02 (0.17)	1.24 (0.22)	1.63 (0.26)	2.04 (0.20)
RAREF	1.24 (0.18)	1.96 (0.20)	1.26 (0.21)	1.66 (0.29)	2.01 (0.17)
QANSS	1.90 (0.24)	1.58 (0.15)	1.24 (0.28)	1.38 (0.27)	1.68 (0.16)
DSGFN	0.99 (0.12)	1.46 (0.18)	0.90 (0.13)	1.33 (0.21)	1.47 (0.25)

rw: random walk with a constant drift; momentum: AR(1) drifts at every level; regimes: Markov-switching provincial drift; spillover: municipal drift follows the lagged regional drift; sparse: momentum with 35% of the sales. Values below 1 beat the no-change forecast. Classic and tree models: 10 replications; neural models: 6.

Table 33. Computational cost (wall-clock on a 32-core Mac Studio M3 Ultra; tree models 10 threads, neural models 8 threads)

Task	Model	Seconds per fit	Total seconds	Fits	Mean training rows
A	Ridge	0.1	1	12	314,363
A	Hedonic OLS	0.3	4	12	314,363
A	Elastic net	3.4	40	12	314,363
A	Quantile MLP	7.7	93	12	314,363
A	HSHN (full pooling)	9.0	108	12	314,363
A	HSHN (no pooling)	9.3	111	12	314,363
A	HSHN (partial pooling)	9.5	114	12	314,363
A	XGBoost	9.9	119	12	314,363
A	Random forest	29.9	359	12	314,363
A	SVR (subsample)	63.6	764	12	314,363
A	LightGBM	155.0	1860	12	314,363
B	No change	0.0	0	36	5,785
B	AR(2), cell	0.0	0	36	5,785
B	Seasonal naive	0.0	0	36	5,785
B	Mean growth	0.0	0	36	5,785
B	Ridge (panel)	0.0	0	36	5,785
B	Random walk with drift	0.0	0	36	5,785
B	Momentum	0.0	0	36	5,785
B	BHPF	0.2	7	36	5,785
B	QANSS	0.2	8	36	5,785
B	MLP (panel)	1.0	36	36	5,785
B	Local linear trend	1.6	58	36	5,785
B	RAREF	3.9	140	36	5,785
B	DSGFN	4.9	176	36	5,785
B	TFT-lite	5.0	181	36	5,785
B	LightGBM (panel)	5.8	210	36	5,785
B	GRU	7.1	254	36	5,785
B	LSTM	8.9	322	36	5,785
B	QREFA	10.4	373	36	5,785
B	TCN	186.0	6696	36	5,785