

How Local Should an AVM Be? Spatial Resolution, Market Regionalization in Automated Valuation Models

Dahoon Kim

zealot8100@naver.com

Research Article

Keywords: Automated valuation model, Commercial real estate, Spatial resolution, Property submarkets, Partial pooling, Modifiable areal unit problem

Posted Date: August 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10639860/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

How Local Should an AVM Be? Spatial Resolution, Market Regionalization in Automated Valuation Models

Abstract

Automated valuation models (AVMs) typically treat location as a question of inclusion: whether, and through which device, spatial information should enter the model. This paper reframes locality as a partial-pooling problem operating through three channels — the spatial resolution h at which location is represented inside a model, the number K of independent regional models into which the market is partitioned, and an external spatial anchor A supplied from outside the estimation sample. Using 22,063 arm's-length transactions of non-residential composite properties — commercial and office buildings sold together with their land — from Korea's seven metropolitan cities (2018–2025), with estimation on 2018–2020, joint selection of h , K , and the city partition on 2021–2022, and a single evaluation on 2023–2025, we estimate the full $h \times K$ risk surface across four resolutions and all 877 set partitions of seven cities. Four results follow. First, representation locality has a clear interior optimum: under complete pooling, future median absolute percentage error (MdAPE) moves $27.104 \rightarrow 22.599 \rightarrow 24.119 \rightarrow 26.118 \rightarrow 27.406$ as encoding refines from none to 500 m, and all four contrasts against the 5 km benchmark are significant after Holm adjustment under a city-stratified spatial block bootstrap. Second, market regionalization improves on complete pooling at every resolution, but returns diminish and the data do not identify an interior optimum in K . Third, the two endogenous channels are partial but incomplete substitutes: raising K to seven recovers 73.7% of the 2 km penalty, 53.7% of the 1 km penalty, and only 19.6% of the 500 m penalty — compensation falls as over-localization becomes more severe. Fourth, when h and K are selected jointly on validation data alone, both anchoring conditions select 5 km, establishing that the resolution optimum is not an artefact of imposed pooling; the anchor reduces the selected K from 5 to 3, cuts the realized regionalization gain from 1.483 to 0.444 percentage points, and eliminates the resolution frontier entirely — no anchored contrast survives multiplicity correction. Representation locality, model locality, and external anchors are alternative mechanisms for supplying local market information; they share a common partial-pooling logic but are asymmetric in strength and only partially interchangeable.

Keywords: Automated valuation model · Commercial real estate · Spatial resolution · Property submarkets · Partial pooling · Modifiable areal unit problem

JEL Classification: C21 · C52 · C53 · R33

1. Introduction

1.1 The question

Most AVM research treats finer locational information as unambiguously better. Coordinates outperform postcodes; postcodes outperform municipalities; submarket effects outperform none. The implicit design principle is that spatial detail is a good to be maximized subject to data availability.

This paper argues that locality is not one design decision but three, and that all three are governed by the same trade-off.

Representation locality (h). Within a single estimated model, at what spatial scale is location encoded? A 5 km grid and a 500 m grid both encode location; they differ in how finely space is partitioned before the learner sees it.

Model locality (K). Across models, into how many independent AVMs is the market divided? A pooled metropolitan-system AVM and seven city-specific AVMs may use identical covariates; they differ in how much of the predictive function is shared across markets.

External locality (A). What locational information is supplied from outside the estimation sample? The previous year's officially assessed parcel-level land price provides a local price level that the model does not have to infer from its own training transactions. We use the term external rather than exogenous throughout: the claim is about the origin of the information relative to the estimation sample, not about econometric exogeneity.

The organizing question is therefore not which grid is best, but:

How much locality should an AVM introduce, through which channel, and at what degree of pooling?

1.2 Research questions

- RQ1. How finely should spatial representation be resolved?
- RQ2. Does market regionalization improve on a pooled AVM?
- RQ3. Does model regionalization compensate for excessively fine spatial representation?
- RQ4. Does an external spatial anchor reduce the need for representation locality and model locality?
- RQ5. Does excessive locality increase spatially correlated error, beyond its effect on pointwise accuracy?

RQ3 binds the paper together. If the two endogenous channels were interchangeable, a sufficiently regionalized model would be indifferent to grid resolution. If they were unrelated, regionalization would leave the resolution frontier untouched. The answer determines whether “locality” is one design parameter or several.

1.3 Research gap

Spatial hedonic and AVM modelling. Following Rosen (1974), the hedonic framework decomposes property price into implicit prices of constituent attributes, and a large subsequent literature

establishes that spatial dependence and spatial heterogeneity are first-order features of that decomposition (Anselin 1988; Can 1992). The methodological effort has gone into devices that accommodate these features: spatial autoregressive specifications, geographically weighted regression (Brunsdon et al. 1996), geostatistical interpolation (Kuntz and Helbich 2014; Muto et al. 2023), and submarket indicators or separately estimated submarket equations (Bourassa et al. 2010). Over the past decade the estimator has changed. Tree ensembles and neural networks now dominate applied valuation work and consistently outperform parametric hedonic specifications (Bogin and Shui 2020; Lorenz et al. 2023; Deppner et al. 2025).

That shift has sharpened rather than resolved the spatial question. Pace and Hayunga (2020) show that tree-based learners leave systematically structured information in hedonic residuals, and Deppner and Cajias (2024) show that random k-fold cross-validation on spatially dependent transactions produces optimistic error estimates, proposing spatial cross-validation as a remedy — the real-estate counterpart of the block-resampling argument in Roberts et al. (2017). We adopt spatial resampling for inference and extend it by treating residual co-dependence as an outcome of interest rather than a nuisance to be corrected. What the machine-learning valuation literature has largely inherited unchanged from its parametric predecessor is the treatment of location as a fixed input — coordinates, distances to named points, or an administrative identifier chosen *ex ante*. How location should be represented has become an active question in GeoAI, where multi-scale encodings are learned rather than imposed (Mai et al. 2020, 2022). That question has not been carried back into valuation, where the granularity of the spatial identifier is rarely selected on predictive grounds.

Modifiable areal unit problem and spatial scale. First identified by Gehlke and Biehl (1934) and closely related to the ecological fallacy (Robinson 1950), the dependence of statistical results on the scale and zoning of areal aggregation was established as a systematic research problem by Openshaw and Taylor (1979) and Openshaw (1984), and extended to multivariate settings by Fotheringham and Wong (1991). The orientation of this classical literature is inferential: the concern is that estimated associations are artefacts of an arbitrary partition. A more recent strand in spatial machine learning does treat scale as a modelling decision. Zeng et al. (2021) document MAUP-induced instability in the inputs and predictions of deep traffic models; Jin et al. (2022) treat grid size explicitly as a tunable hyperparameter for spatiotemporal prediction; and Nikparvar and Thill (2021) survey how scale and spatial structure condition machine-learning performance more broadly. This work establishes that resolution is a prediction-relevant design choice. It has not been carried into property valuation, and it does not address the second question we pose — how resolution interacts with the partitioning of the estimation problem itself.

Property submarkets and regional modelling. Since Schnare and Struyk (1976), a substantial literature has established that urban housing markets are segmented and that submarket structure improves hedonic fit (Goodman and Thibodeau 1998; Bourassa et al. 1999; Watkins 2001). Part of this work is explicitly prediction-oriented: Bourassa et al. (2003) argue that submarkets should be defined so as to optimize out-of-sample valuation accuracy for mass appraisal, and Goodman and Thibodeau (2003) evaluate alternative delineations on prediction error. Later contributions derive segmentations algorithmically rather than from administrative geography (Chen et al. 2009; Helbich et al. 2013; Rey-Blanco et al. 2024), frequently imposing spatial contiguity through minimum-spanning-tree partitioning (Assunção et al. 2006), compare data-driven with expert-delineated

boundaries (Keskin and Watkins 2017), or evaluate the aggregation units used for machine-learning valuation directly (Gao et al. 2022). Multilevel formulations treat the nesting of properties within areas explicitly (Kang 2019). What this literature asks is whether submarkets exist, how they should be delineated, and whether they improve fit — with the number of independent predictive models usually fixed by the delineation procedure rather than chosen on out-of-sample grounds. Almost all of this evidence is residential. Commercial and office property, the setting of this study, has attracted far less segmentation research, even though its thinner and more heterogeneous transaction flow makes the support constraint at the centre of our argument bind harder; recent machine-learning work on commercial appraisal (Deppner et al. 2025) is a notable exception, but does not address market partitioning.

Each literature answers part of the question, and in the last five years each has moved closer to the others: valuation has adopted flexible learners and spatial resampling, spatial machine learning has begun tuning resolution, and submarket research has become algorithmic and prediction-oriented. What none provides is a joint treatment. The two endogenous choices — how finely to encode location within a model, and how finely to partition the market across models — have not, to our knowledge, been estimated on a common out-of-sample surface, so whether they substitute for one another is unknown. Nor has either been tested against the condition that most plausibly displaces both: the availability of an externally supplied local price level. This paper estimates that surface and conducts that test.

1.4 Contributions

Contribution 1 — An interior optimum in spatial representation, robust to inference and to market structure. Under complete pooling, future MdAPE is minimized at 5 km and all four alternative resolutions are significantly worse after Holm adjustment. When h and K are selected jointly on validation data, both anchoring conditions still select 5 km, and the best cell of the entire 4×7 future surface lies in the 5 km row under both. The optimum is a property of the spatial representation, not of the decision to pool.

Contribution 2 — Regionalization improves on pooling at every resolution, with diminishing returns. Complete pooling is dominated at all four resolutions under both anchoring conditions. But returns flatten quickly, and the data do not identify an interior optimum in K — a deliberate asymmetry with Contribution 1 that we explain rather than suppress.

Contribution 3 — The two endogenous channels are partial, not complete, substitutes. Raising K from 1 to 7 recovers 73.7% of the 2 km penalty, 53.7% of the 1 km penalty, and 19.6% of the 500 m penalty. Compensation declines as over-localization worsens: regionalization addresses market heterogeneity, not cell sparsity, and splitting the market makes sparsity worse.

Contribution 4 — External anchors substitute for both endogenous channels, on both outcome dimensions. The anchor cuts ex-post maximal regionalization gains from 2.0–3.7 to 0.5–1.0 percentage points, reduces the validation-selected K from 5 to 3, flattens the resolution frontier until no contrast survives multiplicity correction (interaction estimate -3.502 pp, 95% CI $[-4.804, -1.804]$), and significantly attenuates the co-error penalty at every resolution contrast.

2. Conceptual Framework

2.1 Representation locality

Let h index the spatial resolution at which location is encoded, with smaller h denoting finer encoding, and decompose predictive risk as

$$R_h(h) = B(h) + V(h).$$

Coarse encoding averages price levels over heterogeneous areas, absorbing genuine locational variation into the residual: approximation bias rises. Fine encoding assigns each observation to a cell containing few training transactions, so the estimated cell effect is driven by an idiosyncratic sample: variance rises. The terms move in opposite directions, implying an interior optimum.

H1 (Interior locality). Intermediate spatial resolution will outperform both excessively coarse and excessively fine representations.

2.2 Model locality

Let K denote the number of independent AVMs, with $K = 1$ complete pooling and $K = 7$ one model per city:

$$R_K(K) = B(K) + V(K).$$

Small K forces structurally different regional markets through a single predictive function, averaging away differences in the shape of the price function — not merely in its level, which city indicators absorb. Large K estimates each model on a fraction of the sample, so covariate configurations rare within one city cannot be learned from the others.

A terminological clarification is warranted. We use partial pooling to describe the continuum between $K = 1$ and $K = 7$, but the mechanism here is discrete sharing of the predictive function across market blocks rather than hierarchical shrinkage of parameters toward a common mean in the sense of Gelman and Hill (2007). The bias–variance logic is the same; the estimator is not.

Symmetry with Section 2.1 would suggest an interior optimum in K . We deliberately do not hypothesize one. The variance penalty from fragmentation depends on how much statistical support each block retains, and in a system of seven large metropolitan cities the per-model samples remain in the thousands even at $K = 7$. The hypothesis taken to the data is directional:

H2 (Partial regionalization). Market regionalization improves prediction relative to complete pooling, but its returns diminish as fragmentation increases.

2.3 Representation × model locality

This is the paper’s central extension. Both channels supply local information; the question is whether they supply the same information.

If the difficulty at fine h arose principally from forcing heterogeneous markets through one function, increasing K should largely restore fine- h performance. If instead fine h fails because individual cells are thinly populated, regionalization cannot help: splitting the market reduces per-model sample size and therefore worsens the support available to any given cell.

H3 (Partial substitution). Model regionalization partially offsets the penalty of excessively fine spatial representation, but does not eliminate it.

H3 makes $R(h, K)$ an empirical object rather than a metaphor. We estimate the surface directly.

2.4 External anchor

The lagged assessed land price supplies, from outside the estimation sample, precisely the quantity a fine grid or a regional model must otherwise estimate internally: the local price level. If the anchor carries that level, neither endogenous channel needs to manufacture it.

H4 (Anchor substitution). A stable external spatial anchor reduces the marginal value of both representation locality and model regionalization.

2.5 Spatial error dependence

Two models with identical mean error are not equivalent if one fails independently across properties and the other fails in the same direction across whole neighbourhoods. For mass appraisal and collateral portfolios, spatially clustered error does not diversify. Define

$$\Delta\rho_e(d) = \rho_e^{obs}(d) - E[\rho_e^{perm}(d)],$$

the observed correlation of prediction errors among pairs separated by distance d , net of its expectation under random permutation of errors across locations.

H5 (Co-error). Excessive locality increases the spatial persistence of correlated prediction errors.

2.6 The unifying statement

Across all three channels:

Localization creates information gains but simultaneously fragments statistical support.

The channels differ in who bears the fragmentation. Fine h fragments support within a model. Large K fragments support across models. An external anchor supplies local information without fragmenting anything — which is why it is predicted to dominate both where available.

3. Data and Empirical Design

3.1 Sample

The analysis dataset links three administrative sources. Transaction records come from the Ministry of Land, Infrastructure and Transport's real transaction price disclosure system; building characteristics from the building register; and land characteristics with officially assessed individual land prices from the land register. Records are linked on exact match of four fields: municipality-level address, gross floor area, site area, and construction year. Because the detailed addresses are the product of a de-masking procedure, the raw linked microdata cannot be redistributed; reproducibility is secured by

publishing the linkage procedure and variable-construction rules.

The merged output contains 22,076 transactions, of which 13 are dropped for missing centroid coordinates, yielding a final analysis sample of 22,063 non-residential composite-property transactions in Korea’s seven metropolitan cities over 2018–2025. The properties are commercial and office real estate transacted as a land-and-building bundle, so a sale price is jointly determined by parcel and structure characteristics; this is why the feature set carries land attributes and building composition side by side (Appendix Table A7). Mixed-use buildings are common in this segment, which is why residential use categories appear among the building-use shares even though purely residential properties lie outside the sampling frame. No other exclusion is applied: no target, address, or construction-year values are missing. City assignment follows the source file; 553 records exhibit a source–address discrepancy and are retained rather than dropped, a conservative choice that preserves the sample at the cost of some city-label noise. The dependent variable is the natural logarithm of transaction price per unit area.

Table 1A. Temporal split

Year	N	Role
2018	3,036	Inner training
2019	2,976	Inner training
2020	3,702	Inner training
2018–2020	9,714	Inner training (total)
2021	4,456	Inner validation
2022	2,314	Inner validation
2021–2022	6,770	Inner validation (total)
2018–2022	16,484	Full training (refit)
2023	1,651	Future test
2024	1,976	Future test
2025	1,952	Future test
2023–2025	5,579	Future test (total)
Total	22,063	

Table 1B. City composition

City	N	Share (%)
Seoul	6,413	29.07
Busan	4,047	18.34
Daegu	3,383	15.33
Incheon	2,913	13.20

Gwangju	2,220	10.06
Daejeon	1,868	8.47
Ulsan	1,219	5.52
Total	22,063	100.00

Table 1C. Common evaluation samples

Evaluation sample	N	Total	Coverage (%)
Common validation	4,815	6,770	71.12
Common future	4,331	5,579	77.63

The common evaluation samples consist of observations falling in training-observed cells at all four resolutions and all three grid origins simultaneously, so every specification is compared on an identical set of transactions.

All main results are reported on the common future sample; Appendix Table A3 reproduces the entire $h \times K$ surface on the full 5,579-transaction future sample and reaches the same conclusions. Table 1B is directly relevant to Section 2.2: the smallest city, Ulsan, still contributes 1,219 transactions, which is why per-model support remains substantial even at $K = 7$.

3.2 AVM specification

The learner is XGBoost gradient-boosted regression trees (Chen and Guestrin 2016), held fixed across every experimental condition — same hyperparameters, same random seed (42), same preprocessing order — so that differences in out-of-sample performance reflect changes in spatial representation, market pooling, or external anchoring rather than changes in the algorithm. Because the learner is tree-based, the compositional variables — building structure, use, and roof ratios, which sum to one within each property — do not create the singular-design-matrix problem they would generate in an unregularized linear specification.

The unanchored and anchored specifications at resolution h are

$$M0(h): y_i = f(X_i, T_i, City_i, Grid_i, h),$$

$$M1(h): y_i = f(X_i, T_i, City_i, Grid_i, h, A_{i,t-1}),$$

where X_i is a vector of 150 baseline predictors, T_i a continuous time index, and $A_{i,t-1}$ the logged previous-year assessed parcel land price. For $h = \text{None}$, the grid term is omitted, so that specification carries no sub-city spatial encoding beyond the metropolitan indicator.

The 150 baseline predictors are obtained from a candidate pool of 204 building and land characteristics (Appendix Table A7) plus the interest rate, after excluding all officially assessed land-price fields and removing variables entirely missing or constant in the pooled sample. No sparse-category merging, reference-category deletion, or near-zero-variance screening is applied. The anchor is held out of this set and enters only M1; its missingness rate is 3.96% and missing values are median-imputed from the training sample, as for all numeric variables. Pre-encoding predictor counts are 152, 153, 153, and 154 for $M0/\text{None}$, $M0/h>\text{None}$, $M1/\text{None}$, and $M1/h>\text{None}$; because city and grid identifiers are one-hot encoded, final design-matrix width varies with resolution and estimation sample.

Transaction-year indicators are used only to recover each record's year and do not enter the price

model; in their place every specification includes the continuous index $T_i = (\text{transaction year})_i - 2018$, which provides an ordered temporal control without introducing categories unobserved in training. Macro-financial conditions enter separately through the interest rate. Centroid coordinates are not arguments of the price function; they are used only for grid assignment, local-support computation, pairwise distances in the co-error diagnostics, and spatial block resampling.

Categories unobserved in training are ignored at prediction time. Hyperparameters are fixed throughout: 400 boosting iterations, maximum depth 6, learning rate 0.05, subsample and column-subsample 0.85, L1 regularization 0.05, L2 regularization 1.0.

3.3 Representation locality

$$h \in \{\text{None}, 5 \text{ km}, 2 \text{ km}, 1 \text{ km}, 500 \text{ m}\}$$

Cells are assigned as $\text{floor}((\text{coord} - s)/h)$ with the origin shift s applied jointly to both axes. Because a grid encoding depends on boundary placement as well as cell size, each resolution is estimated under three origins, $s \in \{0, 0.25h, 0.50h\}$. Appendix Table A2 reports coverage and performance by origin.

3.4 Model locality

$$K \in \{1, 2, \dots, 7\}$$

All set partitions of the seven cities are enumerated; the size of the partition search space by K is reported in Appendix Table A5. Estimating all 877 partitions requires fitting only the 127 non-empty city subsets, since a partition's performance is assembled from its constituent blocks. The city indicator is retained inside every regional AVM, so varying K changes the extent to which cities share a predictive function, not whether the model is informed of city identity; at $K = 7$ the indicator becomes constant within each model and carries no predictive variation.

3.5 Selection protocol

The protocol is strictly nested and applies to both locality dimensions simultaneously:

- Fit every $(h, K, \text{partition})$ configuration on 2018–2020.
- Select h, K , and the partition using the 2021–2022 common validation sample only.
- Refit the selected configuration on 2018–2022.
- Evaluate once on 2023–2025.

The evaluation window is never used for selection. Two quantities are reported and labelled separately. Descriptive frontiers (Tables 3, 4 and Appendix A3–A4) report future performance across all configurations and characterize the shape of $R(h, K)$; where a minimum across configurations is identified, it is identified ex post and is not a selection result. The jointly selected design (Table 5) reports the single configuration chosen at step 2 and is the basis for all deployment claims.

A consistency check confirms that the $K = 1$ column of the joint surface reproduces the standalone representation results to machine precision at every (h, anchor) cell.

3.6 Inference

Hypothesis tests use a city-stratified paired spatial block bootstrap: 5 km square blocks (201 blocks), 500 replications, percentile intervals, with the same resampled observations evaluating both candidate and baseline so differences are computed on a common sample. Because the paper documents spatial dependence among prediction errors, an independent row bootstrap would understate sampling variability (Lahiri 2003; Roberts et al. 2017). Within each family of contrasts sharing a benchmark, Holm-adjusted p-values (Holm 1979) accompany unadjusted intervals, and significance claims are made only on the adjusted values.

The co-error null distribution is generated by 499 permutations of residual labels within city, evaluated on 40,829 fixed within-city pairs in the 1–5 km band, capped at 1,500 pairs per city–distance bin to prevent the largest cities from dominating.

4. Results

4.1 Representation locality under complete pooling

We begin at $K = 1$, which isolates the representation channel.

Table 2. Spatial representation frontier under complete pooling ($K = 1$)

Anchor	h	Common MdAPE	Common RMSE	Common R^2	Common P30	Full MdAPE	Full RMSE
No	None	27.104	.4582	.8575	45.07	28.068	.4744
No	5 km	22.599	.3718	.9062	35.97	23.794	.3944
No	2 km	24.119	.3917	.8959	38.97	25.220	.4163
No	1 km	26.118	.4144	.8834	43.02	27.272	.4396
No	500 m	27.406	.4391	.8692	45.58	28.591	.4619
Anchor	None	19.496	.3286	.9267	32.63	20.853	.3473
Anchor	5 km	18.747	.3119	.9340	29.97	19.864	.3333
Anchor	2 km	18.814	.3132	.9334	30.13	19.977	.3352
Anchor	1 km	18.763	.3186	.9311	31.10	20.098	.3406
Anchor	500 m	19.346	.3238	.9288	31.84	20.527	.3445

Common $N = 4,331$; full $N = 5,579$. $P30 = P(APE > 30\%)$; lower is better. Row minima within each anchoring condition: 5 km on every metric except anchored MdAPE, where 5 km, 2 km, and 1 km are within 0.07 points.

The unanchored frontier is a clean inverted pattern:

$$27.104 \rightarrow 22.599 \rightarrow 24.119 \rightarrow 26.118 \rightarrow 27.406.$$

Moving from no sub-city encoding to 5 km improves MdAPE by 4.505 points; every subsequent refinement reverses part of that gain, and at 500 m the model is worse than one with no sub-city encoding at all. The ordering is identical in RMSE, R^2 , P30, and on the full 5,579-transaction future sample, which forecloses the objection that closed-set evaluation favours coarse grids by excluding the unseen cells fine grids generate.

Under the anchor the frontier is nearly flat across 5 km, 2 km, and 1 km (18.747, 18.814, 18.763) with only 500 m and no-encoding worse. The anchor does not move the optimum; it removes the consequences of missing it.

4.2 The joint representation × model locality surface

Table 3A. Future common-sample MdAPE, no anchor

h \ K	1	2	3	4	5	6	7
5 km	22.599	21.615	21.089	20.890	21.117	20.567	20.758
2 km	24.119	22.301	22.000	21.892	21.322	21.191	21.157
1 km	26.118	24.573	23.971	23.311	22.707	22.478	22.388
500 m	27.406	26.136	25.531	25.318	24.654	24.631	24.624

Table 3B. Future common-sample MdAPE, t-1 anchor

h \ K	1	2	3	4	5	6	7
5 km	18.747	18.504	18.302	18.070	17.881	17.958	17.849
2 km	18.814	18.343	18.284	17.888	17.889	18.034	17.980
1 km	18.763	19.058	18.361	18.274	18.312	18.238	18.221
500 m	19.346	19.182	19.006	18.579	18.524	18.399	18.374

Row minima: 20.567 (5 km, K = 6), 21.157 (2 km, K = 7), 22.388 (1 km, K = 7), 24.624 (500 m, K = 7) without the anchor; 17.849 (5 km, K = 7), 17.888 (2 km, K = 4), 18.221 (1 km, K = 7), 18.374 (500 m, K = 7) with it.

Three readings follow.

Regionalization helps at every resolution. Every row declines from K = 1, and in no row is K = 1 close to the minimum, under either anchoring condition. H2 is supported.

The 5 km row dominates the surface. In Panel A the global minimum is 20.567 at (5 km, K = 6), and every cell of the 5 km row except K = 1 beats every cell of the 1 km and 500 m rows at any K. Panel B's global minimum is 17.849 at (5 km, K = 7). This is the strongest available answer to the conjecture that the 5 km result reflects the pooled structure: it does not, because it survives at every level of market segmentation.

Regionalization does not rescue fine encoding.

Table 3C. Penalty relative to 5 km at the same K (no anchor, percentage points)

h	at K = 1	at K = 7	Recovered	Recovered (%)
2 km	1.519	0.399	1.120	73.7
1 km	3.519	1.630	1.889	53.7
500 m	4.807	3.866	0.941	19.6

Full fragmentation into seven city-specific models removes nearly three-quarters of the 2 km penalty, about half of the 1 km penalty, and a fifth of the 500 m penalty. Compensation declines monotonically in the severity of over-localization — exactly what Section 2.3 predicts, since regionalization addresses market heterogeneity, not cell sparsity, and splitting the market makes sparsity worse.

Regionalization compensates for, but does not eliminate, representation over-localization — and compensates least where over-localization is most severe.

H3 is supported.

4.3 How much does regionalization help?

Table 4. Maximum ex-post regionalization gain by representation scale

Anchor	h	Best K (common)	Gain vs. K=1 (pp)	Best K (full)	Gain vs. K=1 (pp)
No	5 km	6	2.033	6	1.961
No	2 km	7	2.961	7	2.989
No	1 km	7	3.730	7	3.766
No	500 m	7	2.782	6	2.948
Anchor	5 km	7	0.898	7	1.133
Anchor	2 km	4	0.926	5	0.981
Anchor	1 km	7	0.542	7	0.775
Anchor	500 m	7	0.972	7	0.953

Best K is identified on the future window: a descriptive frontier, not a selection result. Table 5 reports the honestly selected design.

Two patterns matter.

First, unanchored gains (2.0–3.7 points) are three to five times anchored gains (0.5–1.0), and the pattern is identical on the common and full samples. This is H4 measured on the model-locality channel.

Second, the unanchored gain is largest at 1 km (3.730), not at the optimal 5 km (2.033). Regionalization is most valuable precisely where the representation is most misspecified — the same U-shaped logic the anchor exhibits in Table 6, and further evidence that all three channels substitute for a common underlying quantity.

Best-K is 6 or 7 in seven of eight rows. Returns diminish sharply — most of the gain is captured by $K = 3$ or $K = 4$ — but the data do not identify an interior optimum in K . We report this asymmetry with H1 rather than forcing a parallel; Section 5.1 explains it.

4.4 Jointly selected design

Tables 3 and 4 describe the surface. The deployment question is what a modeller who never saw 2023–2025 would have chosen.

Table 5. Jointly selected AVM design using validation data only

Anchor	Sel. h	Sel. K	Validation MdAPE	Future common MdAPE	Full-future MdAPE	Common RMSE	Full RMSE	Gain vs. K=1
No	5 km	5	23.762	21.117	22.191	.3527	.3799	1.483
Anchor	5 km	3	18.768	18.302	19.305	.3094	.3331	0.444

Selected partitions

No anchor, $h = 5$ km, $K = 5$:

$$\{Seoul\} \mid \{Busan\} \mid \{Daegu\} \mid \{Incheon, Gwangju, Daejeon\} \mid \{Ulsan\}$$

Anchor, $h = 5$ km, $K = 3$:

{Seoul, Gwangju, Daejeon} | {Busan} | {Daegu, Incheon, Ulsan}

The headline result is the resolution:

Selecting representation and market structure jointly on validation data selects 5 km under both anchoring conditions. The 5 km optimum is not an artefact of imposing complete market pooling.

This closes the design's most exposed flank. A reviewer's natural conjecture — that 5 km looked best only because a single pooled function could not accommodate finer detail — is directly testable on the surface in Table 3 and is rejected.

Three further observations.

Selection is costly but not prohibitive. Without the anchor the selected $K = 5$ delivers 21.117 against the ex-post best of 20.567 at $K = 6$, so joint selection recovers 1.483 of 2.033 available points, or 73%. With the anchor it recovers 0.444 of 0.898, or 49%. The larger anchored shortfall reflects a genuinely flatter surface: with gains under one point, validation noise is a larger share of the signal being selected on. Appendix Table A4 shows the anchored validation column varying by only 0.61 points across K at 5 km.

The anchor reduces the selected degree of segmentation. K falls from 5 to 3. The anchor does not merely improve accuracy within a fixed architecture; it changes the architecture that is optimal.

The selected partitions are not administratively obvious. The unanchored solution isolates Seoul, Busan, Daegu, and Ulsan individually while grouping Incheon with Gwangju and Daejeon; the anchored solution groups Seoul with Gwangju and Daejeon, and Daegu with Incheon and Ulsan. Neither corresponds to conventional regional geography. We therefore describe these as predictive market blocks and make no claim that the procedure recovers economic regions — a caution reinforced by the flatness in Table 4, which implies many partitions are near-equivalent.

4.5 The anchor as a substitute for endogenous locality

Table 6A. Anchor improvement at $K = 1$, by resolution

h	No anchor	t-1 anchor	Improvement (pp)
None	27.104	19.496	7.608
5 km	22.599	18.747	3.853
2 km	24.119	18.814	5.304
1 km	26.118	18.763	7.355
500 m	27.406	19.346	8.060

Table 6B. Realized gain from jointly selected regionalization

Anchor	Sel. h	Sel. K	K = 1 at selected h	Selected design	Improvement (pp)
No anchor	5 km	5	22.599	21.117	1.483
t-1 anchor	5 km	3	18.747	18.302	0.444

Table 6A is U-shaped in resolution: the anchor contributes most where the endogenous representation is weakest (7.608 with no encoding, 8.060 at 500 m) and least where it is already best (3.853 at 5 km).

Table 6B shows the same substitution on the model-locality channel: the realized regionalization gain falls by 70% when the anchor is present.

External locational information substitutes for both endogenous locality channels. The better the external spatial signal, the less the model must manufacture locality internally — through fine encoding or through market segmentation.

H4 is supported on the pointwise dimension; Section 4.7 extends the test to the systemic dimension.

4.6 Inference on the resolution frontier

Table 7A. Spatial block bootstrap, future MdAPE relative to 5 km ($K = 1$)

Anchor	Contrast	Δ (pp)	95% CI	p	Holm p	Result
No	None – 5 km	+4.505	[2.107, 7.002]	.004	.016	5 km superior
No	2 km – 5 km	+1.519	[0.489, 2.411]	.008	.016	5 km superior
No	1 km – 5 km	+3.519	[2.031, 4.716]	.004	.016	5 km superior
No	500 m – 5 km	+4.807	[2.820, 6.899]	.004	.016	5 km superior
Anchor	2 km – 5 km	+0.067	[–0.805, 0.776]	.934	1.000	Not distinguishable
Anchor	1 km – 5 km	+0.016	[–0.465, 0.610]	.798	1.000	Not distinguishable
Anchor	500 m – 5 km	+0.600	[0.085, 1.307]	.036	.108	Not sig. after Holm

Table 7B. Anchor $\times$ resolution interaction

Contrast	Estimate (pp)	95% CI	p
(1 km – 5 km)_Anchor – (1 km – 5 km)_NoAnchor	–3.502	[–4.804, –1.804]	.004

201 spatial blocks of 5 km; 500 replications; Holm applied within anchoring condition.

Table 7A converts the Table 2 ranking into an inferential statement:

Every alternative resolution is significantly inferior to the 5 km specification in the unanchored model, with all four contrasts surviving Holm adjustment.

Under the anchor the frontier disappears entirely. The 2 km and 1 km contrasts are indistinguishable from zero at any level; the 500 m contrast is nominally significant but does not survive correction for the three-way family. The honest statement is that no anchored resolution contrast survives multiplicity correction — a stronger form of flattening than a simple comparison of point estimates would suggest. H1 is supported.

Table 7B tests H4 directly rather than by comparing two separately estimated differences:

The anchor reduces the 1 km-versus-5 km fine-resolution penalty by 3.502 percentage points (95% CI: –4.804 to –1.804).

4.7 Local support as the mechanism

Section 2.1 attributes the fine-resolution penalty to thin cell support. Appendix Table A1 confirms the sparsity: median local support falls from 628 training transactions within 5 km to 17 within 500 m, and the first quartile from 387 to 10.

Table 8. Fine-resolution penalty by local-support quartile (percentage points)

Anchor	Quartile	N	2 km – 5 km	1 km – 5 km	500 m – 5 km
No	Q1 (thinnest)	1,108	+4.047	+5.698	+7.322
No	Q2	1,088	+1.615	+3.126	+3.943
No	Q3	1,061	–0.238	+1.583	+2.455
No	Q4 (thickest)	1,074	+0.441	+3.656	+5.314
Anchor	Q1 (thinnest)	1,108	+0.818	+0.911	+1.529
Anchor	Q2	1,088	+0.881	+0.696	+0.869
Anchor	Q3	1,061	–0.539	–0.001	+0.355
Anchor	Q4 (thickest)	1,074	–0.828	–0.772	+0.303

Under the anchor the pattern is exactly as theory predicts and monotone across all three contrasts: the 500 m penalty falls from 1.529 points in the thinnest quartile to 0.303 in the thickest, and in the top two quartiles the 2 km and 1 km specifications are better than 5 km. Where support is abundant and an external level is supplied, finer encoding pays.

Without the anchor the penalty also falls sharply from Q1 to Q3 — the 500 m penalty from 7.322 to 2.455 — but rebounds at Q4 (5.314). Support is therefore a substantial part of the mechanism but not the whole of it: the thickest-support quartile is concentrated in the densest urban cores, where price variation at fine scales is itself greatest, and without an external anchor a fine grid must estimate that variation from its own thin cells.

Local support governs the fine-resolution penalty, and where an external anchor supplies the local level, abundant support reverses the penalty’s sign.

4.8 Spatial error dependence

Table 9A. Excess co-error dependence, 1–5 km band

Anchor	5 km	2 km	1 km	500 m
No anchor	.0235	.0502	.0924	.1202
t–1 anchor	.0237	.0281	.0399	.0505

Table 9B. Fine resolution minus 5 km (spatial block bootstrap, Holm-adjusted)

Anchor	Contrast	Δ excess	95% CI	p	Holm p	Result
No	2 km – 5 km	+.0268	[.0026, .0600]	.032	.032	Significant
No	1 km – 5 km	+.0689	[.0250, .1337]	.004	.012	Significant
No	500 m – 5 km	+.0967	[.0339, .1815]	.004	.012	Significant
Anchor	2 km – 5 km	+.0045	[–.0157, .0296]	.599	.599	Not significant
Anchor	1 km – 5 km	+.0162	[–.0093, .0495]	.224	.447	Not significant
Anchor	500 m – 5 km	+.0268	[–.0032, .0688]	.096	.287	Not significant

Table 9C. Anchor × resolution interaction on co-error dependence

Contrast	Estimate	95% CI	p	Holm p
----------	----------	--------	---	--------

(2 km – 5 km)_A – (2 km – 5 km)_NA	-.0223	[-.0399, -.0069]	.024	.024
(1 km – 5 km)_A – (1 km – 5 km)_NA	-.0527	[-.0855, -.0239]	.004	.012
(500 m – 5 km)_A – (500 m – 5 km)_NA	-.0699	[-.1172, -.0255]	.004	.012

Excess dependence increases monotonically with refinement under both conditions: .0235 → .0502 → .0924 → .1202 without the anchor, and .0237 → .0281 → .0399 → .0505 with it.

The inferential statement respects the multiplicity correction:

H5 is strongly supported in the unanchored specification: relative to the 5 km benchmark, excess 1–5 km co-error persistence increases significantly at 2 km, 1 km, and 500 m, with all three contrasts surviving Holm adjustment. The anchored specification exhibits the same monotonic ordering, but the differences are smaller and none is significant after correction.

The anchored 1 km and 500 m unadjusted intervals do not exclude zero, and we do not describe them as significant.

Table 9C establishes that the anchored attenuation is not merely a failure to detect. Every interaction contrast is significantly negative after Holm adjustment, and the magnitudes are large: at 500 m the anchor removes .0699 of a .0967 unanchored penalty, or 72%. This places the systemic claim on the same inferential footing as the pointwise claim in Table 7B and unifies H4 across both outcome dimensions:

The anchor significantly attenuates both the pointwise accuracy cost and the systemic spatial-error cost of over-localization.

Note finally the two benchmark values in Table 9A: .0235 without the anchor and .0237 with it. At 5 km the anchor delivers essentially no co-error benefit — its entire systemic contribution consists of preventing the deterioration that refinement would otherwise cause.

One qualification of scope: Tables 9A–9C are estimated under complete pooling, so they characterize the representation channel. Whether regionalization also affects co-error persistence is unexamined.

5. Discussion

5.1 Common principle, asymmetric manifestations

The paper’s organizing claim is that three apparently unrelated design choices — grid size, submarket structure, and whether to import an external price signal — are instances of one problem:

Localization creates information gains but simultaneously fragments statistical support.

The empirical manifestations are not symmetric, and stating this precisely is more defensible than forcing a parallel:

Representation locality shows a clear interior optimum; model locality shows a decisive rejection of pooling followed by a plateau; the external anchor dominates

both where it is available.

The asymmetry has a direct explanation in the framework. Both endogenous channels trade specificity against support, but at very different scales of support. A 500 m cell holds a median of 17 training transactions (Appendix Table A1); the smallest city block at $K = 7$ holds 1,219 (Table 1B). The variance term that makes a 500 m grid fail has not begun to bind by $K = 7$ in a system of seven large metropolitan cities. The theory does not merely accommodate the asymmetry; it predicts where an interior optimum in K would appear — in systems partitioned into many small markets — and thereby generates a testable extension.

The finding that would have been unexplained by the framework is the one we do not observe: had regionalization fully rescued 500 m performance, the two channels would have been supplying the same information, and there would be one locality parameter rather than three.

5.2 Partial, not complete, substitution

Table 3C is the paper’s most theoretically informative result. Raising K from 1 to 7 removes 73.7% of the 2 km penalty, 53.7% of the 1 km penalty, and 19.6% of the 500 m penalty — compensation declining monotonically as over-localization worsens.

The two channels fragment different supports. Regionalization addresses heterogeneity in the shape of the price function across markets, and a model estimated on Seoul alone can fit Seoul’s function exactly. It cannot put more transactions into a 500 m cell; if anything it removes the borrowed strength that neighbouring cities’ observations previously supplied. When over-localization is mild, most of the residual penalty is market heterogeneity and regionalization removes it. When over-localization is severe, most of the penalty is cell sparsity and regionalization is powerless.

Section 4.7 supplies the corroborating micro-evidence: the fine-resolution penalty is largest exactly where support is thinnest, and under the anchor it reverses sign in the two highest support quartiles.

Representation locality, model locality, and external anchors are alternative mechanisms for supplying local market information. They share a common partial-pooling logic, but their empirical functions are asymmetric and only partially substitutable.

Stated this way, the result speaks to a question beyond valuation. Grid encoding is the simplest member of a family of location representations that now includes learned multi-scale embeddings (Mai et al. 2020, 2022). Our finding is that the useful scale of such a representation is bounded below by the local density of training observations, and that this bound cannot be relaxed by subdividing the estimation problem — subdivision tightens it. Any location encoder estimated from transaction data, learned or imposed, should face the same constraint.

5.3 External anchors and the architecture of the model

The literature would predict that adding a parcel-level assessed price improves valuation accuracy, and it does. That is not the finding of interest. The finding of interest is what the anchor does to the rest of the design.

It eliminates the resolution frontier — no anchored contrast survives Holm correction, with a directly

estimated interaction of -3.502 pp (CI $[-4.804, -1.804]$). It cuts realized regionalization gains from 1.483 to 0.444 points and ex-post maximal gains from 2.0–3.7 to 0.5–1.0. It significantly attenuates co-error dependence at all three resolution contrasts, removing 72% of the 500 m penalty. And it reduces the validation-selected K from 5 to 3.

The last is the most consequential and least anticipated. The anchor does not merely raise accuracy within a fixed architecture — it changes the architecture that is optimal. A modeller with access to a reliable assessed-value series should build a simpler system: fewer regional models, less sensitivity to grid choice, less exposure to spatially correlated failure.

The practical implication is conditional on institutional setting. Where comprehensive public assessment exists, much of the return to both endogenous channels is already captured, and the operational simplicity of a coarser, more pooled system is nearly free. Where it does not — common in markets without universal assessment — locality must be manufactured internally, and both design levels matter considerably more. The value of investing in fine spatial encoding or in maintaining multiple regional AVMs is therefore not a property of the modelling technology but of the surrounding data infrastructure.

5.4 Open-set behaviour

Supplementary Table S1 documents a structure invisible to closed-set analysis. Refinement degrades coverage — from 99.6% at 5 km to 85.7% at 500 m — while reducing conditional error on the observations that fall outside coverage. Under the anchor the severity ordering is clean and monotone: unseen-cell MdAPE falls from 35.31 at 5 km to 25.76 at 500 m as unseen counts rise from 22 to 797.

The two components require different operational responses. A coverage problem is addressed by fallback logic — hierarchical backoff to a coarser cell, or nearest-neighbour imputation of the cell effect. A severity problem is addressed by flagging and manual review. A coarse-grid system needs a small, high-quality exception process; a fine-grid system needs a large, lower-stakes one. We treat this as a secondary finding: the unanchored severity column is non-monotone and the 5 km unseen cell count is only 22 observations, too few to support strong claims.

5.5 Limitations

Learner dependence. A single tree ensemble with fixed hyperparameters and a fixed seed is used throughout. This isolates the locality effects cleanly — no result can be attributed to differential tuning — but leaves open whether the 5 km optimum is a property of spatial representation or of this learner’s capacity to represent interactions and high-cardinality one-hot encodings. Comparative evidence suggests gradient-boosted trees are a reasonable default for this task (Bogin and Shui 2020; Lorenz et al. 2023), and holding the learner fixed is the same identification strategy used by Deppner and Cajias (2024) to isolate a resampling effect; neither substitutes for direct evidence. Reproducing the resolution ordering under one alternative learner would settle the question at modest cost, and is the single most valuable remaining robustness check.

Inference on H2 and H3. The resolution frontier, the anchor interaction, and the co-error results are all supported by spatial block bootstrap intervals. The regionalization gains in Tables 4–6 and the compensation ratios in Table 3C are reported as point estimates. Paired block-bootstrap intervals for

the $K = 1$ versus selected- K contrast at each resolution, and for the differences among compensation ratios, would place H2 and H3 on the same inferential footing as H1, H4, and H5.

Setting. Non-residential property in seven metropolitan cities of one country over eight years. Commercial and office transactions are sparser and more heterogeneous than residential ones, so the support constraint that drives our central result binds at a coarser scale here than it plausibly would in a residential market of comparable extent; the resolution frontier should be re-estimated, not transferred, when moving to housing. More generally, the 5 km optimum is an instance of an interior optimum, not a transferable constant; the optimal scale should depend on transaction density, market extent, and the granularity of price variation. The seven-city partition space is exhaustively enumerable but too coarse to characterize the K frontier finely.

Zoning robustness. Three grid origins shift both axes jointly, tracing one diagonal through the space of admissible zonings. A full 3×3 design over independent x and y offsets would speak more completely to the zoning component of the MAUP (Openshaw 1984). Appendix Table A2 shows origin-induced variation is an order of magnitude below resolution-induced variation, which bounds but does not eliminate the concern.

Co-error scope. Tables 9A–9C are estimated under complete pooling; the interaction of regionalization with co-error persistence is unexamined.

Linkage. The four-field exact-match key is not a unique parcel identifier, and the pre-merge stages of the linkage funnel are not recoverable from the retained logs (Appendix Table A6). Residual matching error cannot be quantified. In addition, 553 records show a source-file/address city discrepancy; these are retained rather than dropped, so city labels carry some measurement error.

Anchor generality. The anchor is a specific institutional artefact with particular construction and revision properties, and 3.96% of observations lack it and receive imputed values. Its substitutability with endogenous localization may not extend to other external price signals.

6. Conclusion

First, spatial representation has a clear interior optimum. Under complete pooling, future MdAPE is minimized at 5 km, and all four alternative resolutions are significantly worse after Holm adjustment under a spatial block bootstrap. The ordering holds in RMSE, R^2 , P30, and on the full deployment sample.

Second, market regionalization is a second partial-pooling channel. Complete pooling is dominated at every resolution, but returns diminish sharply and the data do not identify an interior optimum in the number of models.

Third, the two endogenous channels are partial but incomplete substitutes. Full fragmentation recovers 73.7% of the 2 km penalty, 53.7% of the 1 km penalty, and 19.6% of the 500 m penalty. Regionalization addresses market heterogeneity; it cannot repair cell sparsity, and compensation declines precisely as over-localization becomes severe.

Fourth, selecting representation and market structure jointly on validation data selects 5 km under both anchoring conditions, establishing that the resolution optimum is not an artefact of imposed pooling. The anchor reduces the selected number of models from five to three and cuts the realized

regionalization gain by seventy percent, changing the architecture that is optimal rather than merely improving performance within one.

The relevant question for AVM design is not whether location should be modeled, but how much locality the available information can support, through which channel, and at what degree of market pooling.

References

- Anselin L (1988) Spatial econometrics: methods and models. Kluwer Academic Publishers, Dordrecht
- Assunção RM, Neves MC, Câmara G, da Costa Freitas C (2006) Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *Int J Geogr Inf Sci* 20(7):797–811
- Bogin AN, Shui J (2020) Appraisal accuracy and automated valuation models in rural areas. *J Real Estate Financ Econ* 60(1):40–52
- Bourassa SC, Cantoni E, Hoesli M (2010) Predicting house prices with spatial dependence: a comparison of alternative methods. *J Real Estate Res* 32(2):139–160
- Bourassa SC, Hamelink F, Hoesli M, MacGregor BD (1999) Defining housing submarkets. *J Hous Econ* 8(2):160–183
- Bourassa SC, Hoesli M, Peng VS (2003) Do housing submarkets really matter? *J Hous Econ* 12(1):12–28
- Brunsdon C, Fotheringham AS, Charlton ME (1996) Geographically weighted regression: a method for exploring spatial nonstationarity. *Geogr Anal* 28(4):281–298
- Can A (1992) Specification and estimation of hedonic housing price models. *Reg Sci Urban Econ* 22(3):453–474
- Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 785–794
- Chen Z, Cho SH, Poudyal N, Roberts RK (2009) Forecasting housing prices under different market segmentation assumptions. *Urban Stud* 46(1):167–187
- Deppner J, Ahlefeldt-Dehn B, Beracha E, Schaefers W (2025) Boosting the accuracy of commercial real estate appraisals: an interpretable machine learning approach. *J Real Estate Financ Econ* 71(2):314–351
- Deppner J, Cajias M (2024) Accounting for spatial autocorrelation in algorithm-driven hedonic models: a spatial cross-validation approach. *J Real Estate Financ Econ* 68(2):235–273
- Fotheringham AS, Wong DWS (1991) The modifiable areal unit problem in multivariate statistical analysis. *Environ Plan A* 23(7):1025–1044
- Gao Q, Shi V, Pettit C, Han H (2022) Property valuation using machine learning algorithms on statistical areas in Greater Sydney, Australia. *Land Use Policy* 123:106409
- Gehlke CE, Biehl K (1934) Certain effects of grouping upon the size of the correlation coefficient in census tract material. *J Am Stat Assoc* 29(185A):169–170
- Gelman A, Hill J (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge
- Goodman AC, Thibodeau TG (1998) Housing market segmentation. *J Hous Econ* 7(2):121–143
- Goodman AC, Thibodeau TG (2003) Housing market segmentation and hedonic prediction accuracy. *J Hous Econ* 12(3):181–201
- Helbich M, Brunauer W, Hagenauer J, Leitner M (2013) Data-driven regionalization of housing markets. *Ann Assoc Am Geogr* 103(4):871–889

- Holm S (1979) A simple sequentially rejective multiple test procedure. *Scand J Stat* 6(2):65–70
- Jin J, Cheng P, Chen L, Lin X, Zhang W (2022) GridTuner: reinvestigate grid size selection for spatiotemporal prediction models. In: *Proceedings of the 2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE
- Kang CD (2019) Effects of spatial access to neighborhood land-use density on housing prices: evidence from a multilevel hedonic analysis in Seoul, South Korea. *Environ Plan B Urban Anal City Sci* 46(4):603–625
- Keskin B, Watkins C (2017) Defining spatial housing submarkets: exploring the case for expert delineated boundaries. *Urban Stud* 54(6):1446–1462
- Kuntz M, Helbich M (2014) Geostatistical mapping of real estate prices: an empirical comparison of kriging and cokriging. *Int J Geogr Inf Sci* 28(9):1904–1921
- Lahiri SN (2003) *Resampling methods for dependent data*. Springer, New York
- Lorenz F, Willwersch J, Cajias M, Fuerst F (2023) Interpretable machine learning for real estate market analysis. *Real Estate Econ* 51(5):1178–1208
- Mai G, Janowicz K, Yan B, Zhu R, Cai L, Lao N (2020) Multi-scale representation learning for spatial feature distributions using grid cells. In: *Proceedings of the 8th international conference on learning representations (ICLR)*
- Mai G, Janowicz K, Hu Y, Gao S, Yan B, Zhu R, Cai L, Lao N (2022) A review of location encoding for GeoAI: methods and applications. *Int J Geogr Inf Sci* 36(4):639–673
- Muto S, Sugawara S, Suzuki M (2023) Hedonic real estate price estimation with the spatiotemporal geostatistical model. *J Spat Econom*. <https://doi.org/10.1007/s43071-023-00039-w>
- Nikparvar B, Thill JC (2021) Machine learning of spatial data. *ISPRS Int J Geo-Inf* 10(9):600
- Openshaw S (1984) *The modifiable areal unit problem. Concepts and techniques in modern geography*, no 38. Geo Books, Norwich
- Openshaw S, Taylor PJ (1979) A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In: Wrigley N (ed) *Statistical applications in the spatial sciences*. Pion, London, pp 127–144
- Pace RK, Hayunga D (2020) Examining the information content of residuals from hedonic and spatial models using trees and forests. *J Real Estate Financ Econ* 60(1):170–180
- Rey-Blanco D, Arbués P, López FA, Páez A (2024) Using machine learning to identify spatial market segments: a reproducible study of major Spanish markets. *Environ Plan B Urban Anal City Sci*. <https://doi.org/10.1177/23998083231166952>
- Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, Hauenstein S, Lahoz-Monfort JJ, Schröder B, Thuiller W, Warton DI, Wintle BA, Hartig F, Dormann CF (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40(8):913–929
- Robinson WS (1950) Ecological correlations and the behavior of individuals. *Am Sociol Rev* 15(3):351–357
- Rosen S (1974) Hedonic prices and implicit markets: product differentiation in pure competition. *J Polit Econ* 82(1):34–55
- Schnare AB, Struyk RJ (1976) Segmentation in urban housing markets. *J Urban Econ* 3(2):146–166

- Watkins CA (2001) The definition and identification of housing submarkets. *Environ Plan A* 33(12):2235–2253
- Zeng W, Lin C, Lin J, Jiang J, Xia J, Turkay C, Chen W (2021) Revisiting the modifiable areal unit problem in deep traffic prediction with visual analytics. *IEEE Trans Vis Comput Graph* 27(2):839–848

Appendix

Table A1. Local training support by radius (common future sample, N = 4,331)

Radius	Mean	SD	Min	Q25	Median	Q75	Max
500 m	24.73	24.97	1	10	17	30	219
1 km	68.72	58.10	1	30	52	90	314
2 km	188.55	151.72	1	83	148	243	798
5 km	680.59	393.56	1	387	628	956.5	1,521

Support falls by a factor of 27 from 5 km to 500 m. A quarter of future transactions sit in 500 m neighbourhoods with ten or fewer training comparables, against 387 at 5 km.

Table A2. Grid coverage and performance by origin shift

h	Shift	Valid. closed N	Future closed N	Future cov. (%)	MdAPE (No anchor)	MdAPE (Anchor)
5 km	0.00	6,719	5,557	99.61	22.599	18.747
5 km	0.25	6,733	5,558	99.62	22.870	18.245
5 km	0.50	6,726	5,560	99.66	23.497	19.001
2 km	0.00	6,558	5,440	97.51	24.119	18.814
2 km	0.25	6,596	5,481	98.24	23.799	18.788
2 km	0.50	6,558	5,457	97.81	24.219	18.814
1 km	0.00	6,212	5,248	94.07	26.118	18.763
1 km	0.25	6,221	5,234	93.82	26.082	19.183
1 km	0.50	6,253	5,260	94.28	25.753	18.949
500 m	0.00	5,547	4,782	85.71	27.406	19.346
500 m	0.25	5,528	4,795	85.95	27.320	19.913
500 m	0.50	5,553	4,810	86.22	27.426	19.698
Intersection	—	4,815	4,331	77.63	—	—

Summary across origins (future common MdAPE): without the anchor, 5 km 22.989 (SD .460), 2 km 24.045 (SD .219), 1 km 25.984 (SD .201), 500 m 27.384 (SD .056); with the anchor, 5 km 18.664 (SD .384), 2 km 18.805 (SD .015), 1 km 18.965 (SD .210), 500 m 19.652 (SD .286). Without the anchor the ranges of adjacent resolutions do not overlap: the worst 5 km origin (23.497) still beats the best 2 km origin (23.799), and the same holds at every step down the scale. The resolution effect is therefore an order of magnitude larger than the origin effect.

Table A3. Full-deployment $h \times K$ surface, all 5,579 future transactions

Panel A. No anchor

h \ K	1	2	3	4	5	6	7
-------	---	---	---	---	---	---	---

5 km	23.794	22.788	22.290	22.180	22.191	21.833	22.015
2 km	25.220	23.697	23.128	22.912	22.312	22.312	22.231
1 km	27.272	25.598	24.940	24.506	24.084	23.891	23.506
500 m	28.591	27.041	26.618	26.391	25.888	25.643	25.730

Panel B. $t-1$ anchor

h \ K	1	2	3	4	5	6	7
5 km	19.864	19.616	19.305	19.094	18.913	19.034	18.731
2 km	19.977	19.523	19.337	19.132	18.996	19.301	19.013
1 km	20.098	19.994	19.593	19.559	19.375	19.396	19.323
500 m	20.527	19.935	19.840	19.632	19.607	19.617	19.574

Every substantive conclusion drawn from Table 3 holds here: the 5 km row dominates, complete pooling is worst in every row, and the global minimum is at 5 km under both anchoring conditions (21.833 and 18.731).

Table A4. Validation and future performance by $h \times K$

No anchor

h	K	Validation	Future common	Future full	Selected partition
----------	----------	-------------------	----------------------	--------------------	---------------------------

5 km	1	25.390	22.599	23.794	all seven pooled
5 km	2	24.285	21.615	22.788	{SE,DG,DJ,US} {BS,IC,GJ}
5 km	3	23.851	21.089	22.290	{SE,IC,US} {BS} {DG,GJ,DJ}
5 km	4	23.767	20.890	22.180	{SE,IC,US} {BS} {DG} {GJ,DJ}
5 km	5	23.762	21.117	22.191	{SE} {BS} {DG} {IC,GJ,DJ} {US}
5 km	6	23.801	20.567	21.833	{SE} {BS} {DG} {IC} {GJ,DJ} {US}
5 km	7	24.103	20.758	22.015	city-specific
2 km	1	26.791	24.118	25.220	all seven pooled
2 km	2	25.724	22.301	23.697	{SE,DG,DJ} {BS,IC,GJ,US}
2 km	3	25.196	22.000	23.128	{SE,GJ} {BS,DJ,US} {DG,IC}
2 km	4	24.955	21.892	22.912	{SE,DG} {BS} {IC,GJ,DJ} {US}
2 km	5	24.898	21.322	22.312	{SE,DJ} {BS} {DG} {IC,GJ} {US}
2 km	6	25.110	21.191	22.312	{SE,DJ} {BS} {DG} {IC} {GJ} {US}
2 km	7	25.377	21.157	22.231	city-specific
1 km	1	27.673	26.118	27.272	all seven pooled
1 km	2	26.404	24.573	25.598	{SE,DG,DJ} {BS,IC,GJ,US}
1 km	3	25.965	23.971	24.940	{SE,GJ,US} {BS} {DG,IC,DJ}
1 km	4	25.747	23.311	24.506	{SE,GJ,US} {BS} {DG,DJ} {IC}
1 km	5	25.954	22.707	24.084	{SE,US} {BS} {DG,DJ} {IC} {GJ}
1 km	6	26.012	22.478	23.891	{SE} {BS} {DG,DJ} {IC} {GJ} {US}
1 km	7	26.211	22.388	23.506	city-specific
500 m	1	28.439	27.406	28.591	all seven pooled
500 m	2	26.991	26.136	27.041	{SE,DG,DJ} {BS,IC,GJ,US}
500 m	3	26.883	25.531	26.618	{SE,DG} {BS,US} {IC,GJ,DJ}
500 m	4	26.806	25.318	26.391	{SE,DG} {BS,US} {IC} {GJ,DJ}
500 m	5	26.806	24.654	25.888	{SE} {BS,US} {DG} {IC} {GJ,DJ}
500 m	6	26.809	24.631	25.643	{SE} {BS} {DG} {IC} {GJ,DJ} {US}
500 m	7	27.079	24.624	25.730	city-specific

t-1 anchor

h	K	Validation	Future common	Future full	Selected partition
5 km	1	19.324	18.747	19.864	all seven pooled
5 km	2	18.858	18.504	19.616	{SE,DG,IC,GJ,DJ,US} {BS}
5 km	3	18.768	18.302	19.305	{SE,GJ,DJ} {BS} {DG,IC,US}
5 km	4	18.808	18.070	19.094	{SE,IC} {BS} {DG} {GJ,DJ,US}
5 km	5	18.881	17.881	18.913	{SE,IC} {BS} {DG} {GJ} {DJ,US}

5 km	6	19.031	17.958	19.034	{SE,IC} {BS} {DG} {GJ} {DJ} {US}
5 km	7	19.378	17.849	18.731	city-specific
2 km	1	19.716	18.814	19.977	all seven pooled
2 km	2	19.270	18.343	19.523	{SE,DG,US} {BS,IC,GJ,DJ}
2 km	3	19.054	18.284	19.337	{SE,DG,US} {BS} {IC,GJ,DJ}
2 km	4	19.293	17.888	19.132	{SE,GJ} {BS} {DG,IC} {DJ,US}
2 km	5	19.360	17.889	18.996	{SE} {BS} {DG,IC} {GJ} {DJ,US}
2 km	6	19.623	18.034	19.301	{SE} {BS} {DG,IC} {GJ} {DJ} {US}
2 km	7	19.948	17.980	19.013	city-specific
1 km	1	20.096	18.763	20.098	all seven pooled
1 km	2	19.574	19.058	19.994	{SE,IC,GJ,DJ,US} {BS,DG}
1 km	3	19.446	18.361	19.593	{SE} {BS,DG,IC,DJ,US} {GJ}
1 km	4	19.476	18.274	19.559	{SE} {BS} {DG,IC,DJ,US} {GJ}
1 km	5	19.615	18.312	19.375	{SE} {BS,DG,US} {IC} {GJ} {DJ}
1 km	6	19.747	18.238	19.396	{SE} {BS} {DG} {IC,DJ} {GJ} {US}
1 km	7	19.892	18.221	19.323	city-specific
500 m	1	20.011	19.346	20.527	all seven pooled
500 m	2	19.596	19.182	19.935	{SE,DG,DJ} {BS,IC,GJ,US}
500 m	3	19.462	19.006	19.840	{SE,DJ} {BS,DG,IC} {GJ,US}
500 m	4	19.510	18.579	19.632	{SE,DG,DJ} {BS} {IC} {GJ,US}
500 m	5	19.611	18.524	19.607	{SE,DG,DJ} {BS} {IC} {GJ} {US}
500 m	6	19.739	18.399	19.617	{SE} {BS} {DG} {IC} {GJ,US} {DJ}
500 m	7	19.911	18.374	19.574	city-specific

City abbreviations: SE Seoul, BS Busan, DG Daegu, IC Incheon, GJ Gwangju, DJ Daejeon, US Ulsan. The validation column is the only one used for selection. Its global minima — 23.762 at (5 km, $K = 5$) without the anchor and 18.768 at (5 km, $K = 3$) with it — determine the configurations in Table 5.

Table A5. Partition search space

K	1	2	3	4	5	6	7	Total
Candidate partitions	1	63	301	350	140	21	1	877

Estimating all 877 partitions requires fitting only the 127 non-empty city subsets, since a partition's performance is assembled from its constituent blocks.

Table A6. Record linkage funnel

Step	N	Source
Source transaction records	—	Requires original linkage log
4-field exact 1:1 matches	—	Requires original linkage log
Records with multiple candidates	—	Requires original linkage log
Merged output after resolution rule	22,076	Seven merged city files

Dropped: missing centroid coordinates	13	Current analysis
Final analysis sample	22,063	Current analysis

The three pre-merge stages are not reconstructible from the retained processing logs. Their absence bounds what can be claimed about matching error: 99.94% of merged records survive to analysis, but the share of source transactions entering the merge and the treatment of multiple-candidate cases cannot be stated.

Table A7. Feature dictionary

The administrative records yield a high-dimensional feature set whose components fall into seven blocks: the log unit-price outcome; land- and building-scale measures; building composition (structure, use, roof) entered as within-building proportions; building vintage; categorical land attributes (zoning, land-use situation, terrain, lot shape, road frontage); the institutional price signal; and time, location, and macro controls. The same schema applies identically across all seven cities. Building composition variables enter as within-building proportions that sum to one within each set, because a single building commonly combines several structural systems, uses, or roof types; land attributes enter as binary indicators. The single lagged assessed land value enters only in the anchored specification; eleven year-specific published columns are excluded from both specifications to prevent leakage. Parcel centroids are used only to form the spatial grid identifier and are not arguments of either price function.

Block / variable	Type / encoding	k	Description (source field)
A. Unit price (log)	Continuous	1	Log transaction unit price; regression outcome (단가).
B. Parcel area, GFA, height, facilities	Continuous / count	10	Land- and building-scale measures: parcel area, gross floor area, storey counts, weighted-average and maximum height, elevators, mechanical parking (토지·건물 규모).
C. Structure type	Proportion	37	Within-building share by structural system (건물 구조).
C. Use type	Proportion	32	Within-building share by use (건물 용도).
C. Roof structure	Proportion	11	Within-building share by roof composition (건물 지붕구조).
D. Construction-year band	Indicator	6	Decade/period bands 1980s–2020s (건물 건축년도).
E. Zoning district	Indicator	24	Use-zone classification (토지 용도지역명).
E. Land-use situation	Indicator	58	Actual land-use category (토지 토지이용상황).
E. Terrain / lot shape / road frontage	Indicator	26	Topographic position, parcel geometry, road-access configuration (토지 지형·형상·도로접면).
F. Assessed land value (log)	Continuous	1	Log lagged official land value (gongsijiga); anchored specification only (토지 공시지가).
F. Year-specific assessed value	Continuous	11	Per-year published columns; excluded from both specifications to prevent leakage.
G. Transaction year	Indicator	8	One-hot transaction-year dummies 2018–2025 (거래시점); not entered in either price function, used only to recover the year for the continuous time index.
G. Interest rate	Continuous	1	Benchmark interest rate at the transaction date (금리).
G. Spatial grid identifier	Indicator	varies	One-hot grid cell at resolution h, constructed from parcel centroids; entered only when h > None. Centroids are dropped after assignment.

k is the number of columns in each block in the pooled feature matrix; for categorical blocks it is the count of distinct levels observed across all seven cities, before the pooled all-missing and constant-column screen. Blocks B–E sum to 204 candidate columns, which with the interest rate reduce to the 150 baseline predictors described in Section 3.2. The grid block expands to a number of columns that varies with *h* and with the estimation sample, which is why the final design-matrix width is not fixed across specifications. Korean source-field names are given in parentheses to aid replication.

Table A8. Variable construction and model usage

Block	Content	Count	Use in model
A	Log transaction price per unit area	1	Dependent variable
B–E	Building and land characteristics	204	Screened with interest rate to 150 baseline predictors
F	Log t–1 assessed individual land price	1	Anchor specification only
G1	Interest rate	1	Within the 150 baseline predictors
G2	Transaction-year indicators	8	Not entered; recovers transaction year
G3	Continuous time index	1	All models
G4	City identifier	1	All models; one-hot encoded
G5	Spatial grid identifier	1	Entered when h > None; one-hot encoded
H	Centroid coordinates (x, y)	2	Not entered directly; grid, support, co-error, resampling
I	Address string	1	Not entered; linkage and audit

Supplementary Material

Table S1. Open-set coverage and severity (full future sample)

Anchor	h	Coverage (%)	Full MdAPE	Seen MdAPE	Unseen N	Unseen MdAPE
No	None	100.00	28.068	28.068	0	—
No	5 km	99.61	23.794	23.744	22	50.59
No	2 km	97.51	25.220	24.844	139	37.07
No	1 km	94.07	27.272	26.663	331	38.65
No	500 m	85.71	28.591	27.778	797	35.22
Anchor	None	100.00	20.853	20.853	0	—
Anchor	5 km	99.61	19.864	19.801	22	35.31
Anchor	2 km	97.51	19.977	19.738	139	33.08
Anchor	1 km	94.07	20.098	19.679	331	28.26
Anchor	500 m	85.71	20.527	19.774	797	25.76

Table S2. Future MdAPE by local-support quartile

Panel A. No anchor

Quartile	N	None	5 km	2 km	1 km	500 m
Q1 (thinnest)	1,108	28.166	22.285	26.332	27.982	29.606
Q2	1,088	28.104	24.241	25.856	27.367	28.184
Q3	1,061	24.084	22.180	21.942	23.763	24.635
Q4 (thickest)	1,074	28.522	21.975	22.417	25.631	27.289

Panel B. t-1 anchor

Quartile	N	None	5 km	2 km	1 km	500 m
Q1 (thinnest)	1,108	21.063	19.493	20.311	20.403	21.022
Q2	1,088	21.954	20.304	21.185	20.999	21.173
Q3	1,061	18.148	18.241	17.702	18.239	18.595
Q4 (thickest)	1,074	17.551	17.399	16.571	16.627	17.701

Table S3. Co-error decomposition, 1–5 km band

Anchor	h	Observed p	Permuted E[p]	Excess	Permutation p
--------	---	------------	---------------	--------	---------------

No	5 km	.0424	.0189	.0235	.002
No	2 km	.0703	.0200	.0502	.002
No	1 km	.1099	.0175	.0924	.002
No	500 m	.1371	.0170	.1202	.002
Anchor	5 km	.0415	.0178	.0237	.002
Anchor	2 km	.0418	.0137	.0281	.002
Anchor	1 km	.0503	.0104	.0399	.002
Anchor	500 m	.0600	.0095	.0505	.002

40,829 within-city pairs; 499 permutations of residual labels within city. The permuted expectation is itself informative: it declines with refinement under the anchor (.0178 $\rightarrow$.0095), so the rise in excess dependence is driven entirely by the observed correlation, not by a shifting null.