Innovation Economics and Management Research (IEMR), Volume 11, Issue 2, 2026. ISSN: 2949-1304
Proceedings of The 7th International Conference on Management, Economy and Law (ICMEL 2026)

Big-data-driven Commercial Real Estate Value Prediction and Spatial Layout Optimization in a Simulated U.S. Market

Junye Deng[1]

[1] *University of Washington-Seattle, Seattle 98195, USA*

ABSTRACT

Asset valuation and portfolio selection are both required for commercial real estate (CRE) investment, but they are often studied in isolation. This reproducible simulation links commercial value prediction to spatially constrained site selection in a stylized market of 120 sites across six submarkets, and the market reflecting stylized U.S. CRE conditions is not the same everywhere in the city. Four prediction models are evaluated by random five-fold and leave-one-submarket-out cross-validation, and their predictions are input into a binary optimisation model with capacity, budget and minimum-distance constraints. If the nonlinear value process is correctly specified, Hedonic II outperforms XGBoost in the simulated sample of 120-3000 sites (RMSE 0.052 vs. 0.083 at N=120). Compared with value-only selection, multi-criteria weighting reduces risk by 4.0% and competition by 2.5% at a 0.3% predicted-value cost, and the minimum-distance constraint achieves most of the geographic diversification. For the tested implementation and parameter budget, a genetic algorithm remains 11.2% below the certified exact solution. The results are methodological behaviours under known assumptions, not real-world performances.

Keywords: *Commercial real estate, Big data, Spatial layout optimization, Commercial value prediction, XGBoost, simulation.*

1. INTRODUCTION

Hedonic analysis and its machine-learning extensions have dominated property valuation, but their benefits are easily overestimated when ignoring spatial structure, thus motivating spatial cross-validation [1]. Evidence that tree ensembles outperform parametric models is strongest for large tabular datasets, and flexible learners are significantly more data-hungry; thus, they are less suitable for CRE submarkets with tens or a few hundred transactions [2]. On the allocation side, discrete site selection derives from covering formulations and p-dispersion, with recent multi-criteria suitability work. Scalarisation follows the weighted-attribute tradition, and evolutionary multi-objective methods are alternatives. Few studies have conducted prediction, spatial validation and constrained selection in a single reproducible environment. The linear binary formulation adopted here permits exact solution, enabling the comparison in Section 4.6.

2. RELATED WORK

Hedonic models and machine-learning methods have both been applied to property valuation so far. Flexible tree-based methods can handle non-linear relationships, but their performance may be relatively poor in CRE submarkets with a small number of transactions. Given the particular spatial structure of the real estate observations, random resampling may increase the estimated predictive power; thus, spatially structured validation offers a more reliable test of geographic transferability [3]. Location-allocation and p-dispersion models are typical ways to combine site attractiveness with geographic dispersion in portfolio construction [4]. This paper connects the above links by adding out-of-fold value predictions to a constrained binary portfolio model.

3. METHODOLOGY

3.1 Simulated Spatial Market and Value Process

The simulated market has N=120 candidate sites in six cluster submarkets. Each site *i* has normalised measures of accessibility, demand, competition, cost, risk and quality. Attributes are generated from a correlated logit-normal process with submarket-specific means, and realistic trade-offs are produced: highly accessible sites tend to be more costly and more contested. All parameters are in the following "Table 1".

Table 1. Simulation design parameters.

Component	Specification
Market	N = 120 sites; 6 submarkets of 20; coordinates bivariate normal truncated to the unit square, isotropic SD 0.055
Cluster centroids	(0.22, 0.78), (0.50, 0.85), (0.80, 0.72), (0.18, 0.30), (0.55, 0.22), (0.85, 0.32)
Attributes	Correlated logit-normal on the logit scale, SD 0.10, inverse-logit transformed
Submarket means (A/D/C/K/R/Q)	1 Core .82/.78/.72/.80/.55/.78; 2 Inner north .70/.68/.58/.66/.48/.68; 3 Technology edge .66/.74/.50/.62/.52/.72; 4 Legacy industrial .42/.40/.34/.38/.62/.44; 5 Logistics corridor .52/.62/.42/.46/.50/.56; 6 Suburban growth .46/.52/.30/.42/.44/.60. Correlations AD .55, AC .45, AK .60, AR −.25, AQ .35, DC .50, DK .40, DR −.20, DQ .30, CK .35, CR .15, CQ .10, KR −.10, KQ .45, RQ −.30
Value coefficients	β_0 = 0.20, β_1 = 0.30, β_2 = 0.35, β_3 = 0.15, β_4 = 0.20, β_5 = 0.18, β_6 = 0.22, β_7 = 0.80, β_8 = 0.60, β_9 = 2.00; τ = 0.40; $\varepsilon \sim N(0, 0.05^2)$
Constraints	Capacity 10 sites; baseline budget B = 5.50; distance threshold δ = 0.12
Decision weights	w_1 = 0.35 (predicted value), w_2 = 0.25 (demand), w_3 = 0.15 (cost), w_4 = 0.10 (risk), w_5 = 0.10 (competition); weights apply to normalized attributes and are not constrained to sum to one
Random Forest	100 trees, max depth 10, min samples per leaf 2, max features 0.8
XGBoost	100 rounds, learning rate 0.10, max depth 6, subsample 0.8, colsample 0.8, min child weight 3, L2 = 1.0
Genetic algorithm	Population 200, 300 generations, crossover 0.9, mutation 0.02, tournament size 3, elitism 4, penalty 50
Replications	30 independent seeds (1–30)
Software	Python 3.14.3, NumPy 2.5.2, scikit-learn 1.9.0, XGBoost 3.4.1, PuLP 3.3.2 with CBC, SciPy 1.18.1

The latent commercial value of site *i* is generated by

$$V_i^{\square}=\beta_0+\beta_1 A_i+\beta_2 D_i-\beta_3 C_i-\beta_4 K_i-\beta_5 R_i+\beta_6 Q_i+\beta_7 A_i D_i+\beta_8 D_i^2-\beta_9[max(0, C_i-\tau)]^2+\varepsilon_i \quad (1)$$

Equation (1) has three deviations from additive linearity: the interaction $\beta_7 A_i D_i$ shows a larger accessibility premium where demand is high; the quadratic term produces increasing returns to demand; and the threshold term makes competition harmful only beyond τ = 0.40, so it is a threshold nonlinearity rather than a constant slope. The disturbance is $\varepsilon_i \sim N(0, 0.05^2)$. Coefficients are design parameters, not estimates, and were selected so that the nonlinear components are relatively large compared to the noise; otherwise, a model-comparison experiment could not distinguish between functional forms. The realized standard deviation of latent value is 0.29.

3.2 Prediction Models and Validation

Four models are compared. Hedonic I is the traditional linear specification that includes the six main effects and the accessibility-demand interaction; it is misspecified relative to (1) because it omits the quadratic demand term and the competition threshold. Hedonic II adds exactly these two terms and is therefore correctly specified up to parameter linearity. Including both models separates functional-form misspecification from algorithmic flexibility. Random Forest and XGBoost, as shown in "Table 1", use the same features as Hedonic I and therefore need to recover the omitted nonlinearity from the data [5]. Hyperparameters ("Table 1") were set a priori. Accuracy is shown as mean absolute error (MAE), root mean squared error (RMSE) and R^2 of out-of-fold predictions.

Two validation schemes are used in all replications: random five-fold cross-validation, which is the most common design for applied valuation work, and leave-one-submarket-out

validation. The latter is a spatial block design that assesses transfer to a geographic area not seen in training, rather than interpolation among spatially close observations. In real-estate settings, spatial dependence can make random resampling estimates overly optimistic, so the two schemes are reported together to distinguish interpolation from geographic transfer performance [6]. The model with the lowest random five-fold RMSE in a replication is used to obtain the predicted values for the optimizer in that replication.

3.3 Spatial Portfolio Optimization Model

Let $x_i = 1$ if site i is selected and 0 otherwise, and let V_i^* be its predicted value. The portfolio is chosen by maximizing the composite utility

$$\max U(x) = \sum_{i=1}^{N}(w_1 V_i^* + w_2 D_i - w_3 K_i - w_4 R_i - w_5 C_i)x_i \quad (2)$$

subject to (3)–(4), with all x_i binary.

$$\sum_{i=1}^{N} x_i \leq 10, \sum_{i=1}^{N} K_i x_i \leq B \quad (3)$$

$$x_i + x_j \leq 1 \quad \text{for all } (i,j) \text{ with } d_{ij} < \delta \quad (4)$$

Constraint (3) sets a capacity limit of ten assets and a budget ceiling with baseline B = 5.50 normalized cost units. Pairwise exclusion in (4) prohibits selecting two sites with a Euclidean distance d_ij less than δ= 0.12; it is a linear minimum-separation rule in the spirit of p-dispersion literature that promotes geographic dispersion. Both δ and d_{ij} are in normalised coordinate units, not kilometres, and δ is varied in Section 4.5. Problems (2)–(4) are binary programs with a linear objective and constraints, so all reported portfolios were obtained exactly using the CBC branch-and-cut solver via PuLP [7], and every instance returned a certified optimal status. A genetic algorithm was also used for the same cases, and a penalty function for the three constraint sets was introduced to quantify the loss (Section 4.6).

3.4 Benchmarks, Scenarios and Protocol

Three strategies are compared under identical feasibility constraints: a value-only benchmark maximizing total predicted value, a demand-only benchmark maximizing total demand potential, and the integrated strategy maximizing (2). Set the same capacity, budget and distance constraints; otherwise, it would be easy to think that one strategy is better because it is allowed to spend more or cluster more tightly. Four market states are considered: S1 is the unshocked baseline; S2 increases demand in the two most accessible submarkets by 30%; S3 increases competition in the two highest-demand submarkets by 40%; and S4 increases risk in the two most vulnerable submarkets by 35%. Shocked attributes are re-clipped to the unit interval, the latent value is regenerated from (1), and then the prediction stage is re-run; thus, each scenario is a complete re-execution of the pipeline. Four sensitivity analyses vary the risk weight, budget, distance threshold and market size. Spatial dispersion is the mean pairwise distance of the selected sites, and each solution reports the chosen sites, budget utilisation, minimum realised distance and submarkets covered, so binding constraints can be identified.

The pipeline has been replicated 30 times independently, and all quantities are presented as means and standard deviations, with 95% confidence intervals included where applicable. As all seeds are common to all models and strategies, they are paired for comparison, and a two-sided Wilcoxon signed-rank test is used on the 30 paired differences to determine significance; only differences supported by this test are considered reliable. No external or proprietary data were used in this study, and the scripts and per-seed outputs are available from the corresponding author upon reasonable request; all design parameters, hyperparameters and random seeds required for replication are listed in “Table 1”.

4. RESULTS

4.1 Prediction Performance

As shown in “Table 2”, Hedonic II has the lowest out-of-fold error for both validation schemes. Random five-fold cross-validation shows that its RMSE is 0.0523, which is lower than that of XGBoost (0.0834) and Random Forest (0.0930); all pairs have a p-value less than 0.001. Leave-one-submarket-out validation increases the error for every model, but the increase is relatively small for Hedonic II (3.6%) compared with Hedonic I, Random Forest and XGBoost (18.1%-20.4%). In this known-process simulation, the correct functional specification is transferred more reliably across submarkets than flexible models trained on limited local data. The per-replication RMSE values for both cross-validation strategies are shown in “Figure 1”.

Table 2. Commercial-value prediction performance across 30 replications

Model	MAE	RMSE	RMSE 95% CI	R^2	p vs Hedonic II
Hedonic I	0.0566 ± 0.0045	0.0711 ± 0.0051	[0.0692, 0.0730]	0.941 ± 0.011	< 0.001
Hedonic II	0.0416 ± 0.0035	0.0523 ± 0.0042	[0.0507, 0.0539]	0.968 ± 0.006	Reference
Random Forest	0.0730 ± 0.0042	0.0930 ± 0.0056	[0.0909, 0.0951]	0.900 ± 0.016	< 0.001
XGBoost	0.0658 ± 0.0051	0.0834 ± 0.0064	[0.0810, 0.0858]	0.919 ± 0.016	< 0.001

Values are mean ± standard deviation; p-values are from two-sided Wilcoxon signed-rank tests on paired differences. Leave-one-submarket-out RMSE is 0.0856 ± 0.0095, 0.0542 ± 0.0058, 0.1113 ± 0.0092 and 0.0985 ± 0.0083 in row order; all differences from Hedonic II have $p < 0.001$.

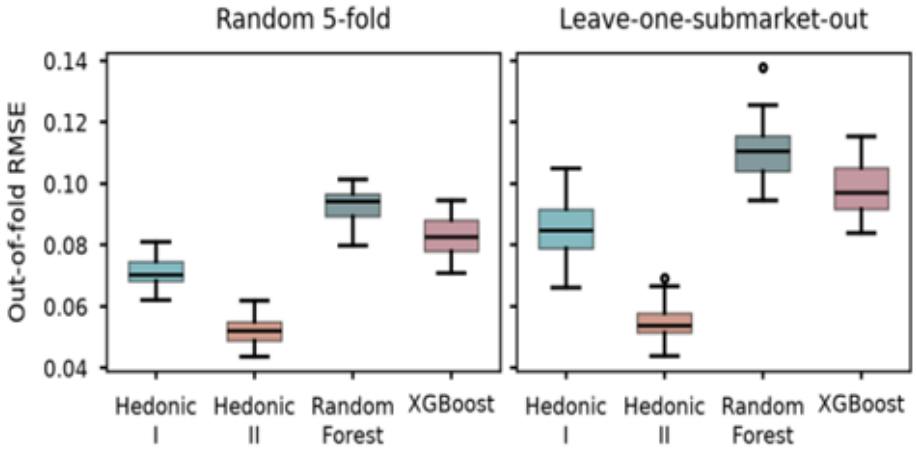


Figure 1 Out-of-fold RMSE across 30 replications under random five-fold (left) and leave-one-submarket-out (right) validation.

4.2 Sample-size Dependence of the Model Ranking

Section 4.1 is a small-sample result, and "Table 3" shows where the boundary is. Keep the process, hyperparameters and validation design fixed, and vary the market size from 120 to 3000 sites across ten seeds per size. Both tree ensembles improve monotonically with sample size, while the parametric models are flat beyond N = 300 because their error is dominated by specification rather than variance. XGBoost surpasses the misspecified Hedonic I benchmark between N = 300 and N = 600; Random Forest needs 600 to 1500 observations to achieve the same result, and neither ensemble reaches Hedonic II at any of the tested sizes. The reading is therefore conditional: gradient boosting is preferable when the functional form is unknown and the sample is large, whereas a structured hedonic model is preferable when plausible nonlinearities can be specified and the sample is small. The second system is more suitable for valuation in a lack of comparable transaction data. Hedonic II had the smallest error in all the replications at N = 120, so it provided the predicted values used in the optimisation stage.

Table 3. Out-of-fold RMSE by market size (ten seeds per size).

Sites N	Hedonic I	Hedonic II	Random Forest	XGBoost	Best model
120	0.0710 ± 0.0046	0.0540 ± 0.0050	0.0927 ± 0.0056	0.0844 ± 0.0053	Hedonic II
300	0.0702 ± 0.0023	0.0526 ± 0.0019	0.0814 ± 0.0039	0.0739 ± 0.0031	Hedonic II
600	0.0682 ± 0.0023	0.0506 ± 0.0011	0.0708 ± 0.0022	0.0653 ± 0.0019	Hedonic II
1500	0.0685 ± 0.0018	0.0502 ± 0.0013	0.0638 ± 0.0015	0.0601 ± 0.0017	Hedonic II
3000	0.0685 ± 0.0005	0.0497 ± 0.0007	0.0596 ± 0.0009	0.0562 ± 0.0008	Hedonic II

4.3 Baseline Portfolio Comparison

"Table 4" shows the three constraint-consistent strategies and reports the paired differences that identify the trade-off. All three select exactly ten sites in every replication and use 99.4%-99.6% of the budget, so capacity binds universally, and the realised minimum pairwise distance is at the δ boundary in every solution (0.126 vs. δ = 0.12), and the dispersion constraint is always active as well. The gains from multi-criteria weighting are modest. Compared with the value-only benchmark, the integrated strategy reduces risk by 4.0% and competition by 2.5% at a 0.3% value sacrifice; cost and dispersion do not differ reliably (p = 0.463 and 0.554). Against demand-only, differences are larger (risk −16.1%, competition −13.0%, value +4.6%; p < 0.001), but demand alone is not a good indicator. This small benefit is expected in the current design: competition, cost and risk have already reduced the latent value in Equation (1), so explicit weighting partially overlaps with information already contained in the predicted value. Weights may be necessary in practice if the valuation model omits decision-relevant risks.

Table 4. Baseline portfolio comparison and paired differences versus the value-only benchmark.

Strategy or metric	Predicted value	Demand capture	Mean risk	Mean competition	Spatial dispersion
Value-only	12.614 ± 0.375	7.495 ± 0.209	0.430 ± 0.036	0.488 ± 0.033	0.476 ± 0.037
Demand-only	12.026 ± 0.506	7.684 ± 0.158	0.492 ± 0.033	0.547 ± 0.035	0.489 ± 0.023
Integrated	12.576 ± 0.394	7.466 ± 0.205	0.413 ± 0.034	0.476 ± 0.033	0.480 ± 0.033
Mean difference	−0.039	−0.028	−0.017	−0.012	+0.003
95% CI	[−0.057, −0.020]	[−0.058, 0.001]	[−0.025, −0.009]	[−0.019, −0.005]	[−0.006, 0.012]
Relative change	−0.3%	−0.4%	−4.0%	−2.5%	+0.7%
p	< 0.001	0.068	< 0.001	0.002	0.554

Differences are integrated minus value-only, paired by seed (n = 30). All strategies select 10 sites, mean cost 0.546–0.548, budget use 99.4%–99.6%; cost does not differ reliably (p = 0.463).

4.4 Scenario Analysis

"Table 5" shows the integrated portfolio under the four market states. Under demand expansion (S2), the predicted value increases by 12.1% and demand capture by 11.4%; competition rises by 4.3%, and the portfolio response is in line with a shift to the shocked accessible submarkets and greater tolerance for crowding, but the distance constraint prevents co-location. Under competitive saturation (S3), the predicted value decreases by 6.2%, submarket coverage drops from 5.30 to 4.83, and dispersion declines by 4.5%; that is to say, the competition shock reduces the geographic spread rather than expanding it. S4 increases realised risk by 7.6% without raising the risk penalty and reduces predicted value by only 0.6%. The optimizer prioritises value over risk at the margin, but reallocation cannot fully compensate for a shock that reduces the opportunity set.

Table 5. Integrated portfolio response by market scenario

Scenario	Predicted value	Mean risk	Mean competition	Submarkets
S1 Baseline	12.576 ± 0.394	0.413 ± 0.034	0.476 ± 0.033	5.30 ± 0.47
S2 Demand expansion	14.103 ± 0.427 (+12.1%*)	0.422 ± 0.034 (+2.2%)	0.496 ± 0.034 (+4.3%*)	5.23 ± 0.50
S3 Competitive saturation	11.801 ± 0.321 (−6.2%*)	0.409 ± 0.037 (−0.8%)	0.498 ± 0.037 (+4.5%*)	4.83 ± 0.70
S4 High risk	12.501 ± 0.400 (−0.6%*)	0.444 ± 0.034 (+7.6%*)	0.477 ± 0.032 (+0.3%)	5.27 ± 0.45

Percentage change relative to S1; an asterisk denotes p < 0.05 on the paired Wilcoxon test (n = 30). Demand capture is 7.466 ± 0.205, 8.315 ± 0.247 (+11.4%*), 7.268 ± 0.197 (−2.7%*) and 7.457 ± 0.192 (−0.1%), and dispersion changes by +0.6%, −4.5%* and −0.1%, under S1–S4.

4.5 Sensitivity Analysis

"Table 6" shows the two decision parameters a practitioner would actually set; the budget results are presented in the text only. The risk-weight trade-off is strongly nonlinear: raising w_4 from 0.10

to 0.50 lowers mean risk by 10.2% at a value cost of 2.8%, an efficient exchange; pushing to 1.00 buys 19.2% risk reduction for 13.7% of value; at 1.50 the objective becomes risk-dominated, value collapses by 55.0% and coverage falls to 3.47. Any weight obtained from the investment policy should be compared with this curve before being used as a point value, and in this design, the efficient region ends near $w_4 = 0.50$. The budget experiment shows the binding constraint. At 80% of the baseline (B = 4.40), the budget is an effective constraint: utilisation reaches 99.6%, the portfolio drops below ten sites in a third of replications, and the predicted value falls by 12.7%. At 120% (B = 6.60) the budget develops 4.7% slack and capacity alone binds, so the extra capital raises predicted value by only 5.8%, because it can upgrade the quality of ten permitted assets but cannot acquire more.

The distance threshold is the main diversification instrument. Excluding it will concentrate the portfolio (dispersion 0.386, submarkets 4.00) and increase value by 4.8%; Increasing δ to 0.24 forces selection across all six submarkets, but reduces predicted value by 23.5% and makes a ten-site portfolio infeasible in most replications. The realized minimum separation tracks δ throughout, and diversification is thus achieved by constraints rather than objective weights (consistent with "Table 4").

Table 6. Sensitivity to the risk weight and to the minimum-distance threshold.

Parameter	Predicted value	Δ value	Mean risk	Δ risk	Spatial dispersion	Sub-markets
Risk weight w_4						
0.10	12.581 ± 0.394	—	0.415 ± 0.035	—	0.477 ± 0.037	5.30
0.25	12.528 ± 0.391	−0.4%	0.400 ± 0.029	−3.7%	0.481 ± 0.033	5.37
0.50	12.234 ± 0.535	−2.8%	0.373 ± 0.027	−10.2%	0.477 ± 0.035	5.43
1.00	10.861 ± 0.867	−13.7%	0.335 ± 0.030	−19.2%	0.483 ± 0.035	5.13
1.50	5.665 ± 1.784	−55.0%	0.290 ± 0.028	−30.2%	0.492 ± 0.088	3.47
Distance threshold δ						
0.00	13.181 ± 0.390	+4.8%	0.406 ± 0.041	−1.7%	0.386 ± 0.055	4.00
0.12	12.576 ± 0.394	—	0.413 ± 0.034	—	0.480 ± 0.033	5.30
0.18	11.767 ± 0.440	−6.4%	0.441 ± 0.036	+6.8%	0.517 ± 0.021	5.97
0.24	9.619 ± 0.719	−23.5%	0.439 ± 0.040	+6.3%	0.544 ± 0.026	6.00

Each block varies one parameter with others at the "Table 1" baseline; changes are relative to the baseline row of the block (w_4= 0.10, δ = 0.12). At w_4= 0.75 predicted value is 11.708 ± 0.493 (−6.9%) and risk 0.349 ± 0.030 (−16.0%); at δ = 0.06 they are 13.012 ± 0.386 (+3.5%) and 0.401 ± 0.039 (−2.9%). Realized minimum distances in the δ block, in normalized coordinate units, are 0.032, 0.071, 0.126, 0.185 and 0.249 at δ = 0.00, 0.06, 0.12, 0.18 and 0.24; 9.90 ± 0.31 and 8.17 ± 0.75 sites are selected at δ = 0.18 and 0.24.

4.6 Heuristic Versus Exact Optimization

The genetic algorithm returned feasible portfolios in all cases but never found the optimal set; the composite utility shortfall ranged from 6.9% to 15.7% and averaged 11.2%. This gap is several times larger than the 0.3-4.0% difference in allocation strategies (Section 4.3), so solver choice would dominate the decision criterion. With 120 variables and about 800 constraints, CBC certified optimality in all cases; for the tested implementation and parameter budget, the genetic algorithm did not get close enough to the certified optimum to justify its use as a substitute for exact optimization at this problem scale.

5. DISCUSSION AND LIMITATIONS

Five limitations qualify these results. First, the data are synthetic and not aligned with real-world prices; the application of Hedonic II requires a known specification, which is not available in practice. Second, the value process is smooth and has a single competition threshold; discontinuous or high-dimensional interactions would be more suitable for tree ensembles. Third, leave-one-submarket-out validation is a coarse block design; time-aware holdout and finer blocking are required. Fourth, portfolio comparisons are based on predicted values rather than latent values, so the decision quality relative to an oracle portfolio that knows the true value is not measured; reporting this regret is a natural extension. Fifth, weighted-sum scalarisation represents an ex ante preference structure. Where stakeholders cannot agree on fixed trade-offs among value, risk, competition and

dispersion, reporting a Pareto frontier or conducting multi-objective sensitivity analysis would make these trade-offs more explicit.

6. CONCLUSION

This paper has developed a reproducible simulation framework that connects machine-learning-based commercial-value prediction with constrained spatial portfolio optimisation, and has disclosed all design parameters, hyperparameters, solver settings and random seeds required for reproduction. Three results qualify common assumptions in applied data-driven real-estate work. First, when the nonlinear structure is correctly specified, a specification-aware hedonic model achieved the lowest prediction error at all market sizes from 120 to 3000 sites, and gradient boosting only surpassed a misspecified linear benchmark between 300 and 600 observations; thus, the preference for tree ensembles is sample-size dependent rather than general. Second, multi-criteria weighting reduced portfolio risk by 4.0% and competition exposure by 2.5% for a 0.3% value sacrifice; this was a reliable but modest gain, while geographic diversification was driven by the minimum-distance constraint and increased dispersion by 41% and submarket coverage from four to six. Third, under the tested implementation a genetic algorithm left an 11.2% mean optimality gap against the exact solution, exceeding the differences between the allocation strategies themselves. The above results are valid under the assumption of a known data-generating process and a synthetic market; future work should calibrate the framework with transaction-level data, incorporate rents, vacancy, financing terms and zoning, adopt spatially blocked and time-aware validation as the reporting standard, and extend the objective to a Pareto formulation where investor preferences are not fixed.

REFERENCES

[1] Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. Journal of Political Economy, 82(1), 34-55.

[2] Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why Do Tree-Based Models still Outperform Deep Learning on Typical Tabular Data?. Advances in Neural Information Processing Systems, 35, 507-520.

[3] Deppner, J., and Cajias, M. (2024). Accounting for Spatial Autocorrelation in Algorithm-Driven Hedonic Models: A Spatial Cross-Validation Approach. The Journal of Real Estate Finance and Economics, 68(2), 235-273.

[4] Church, R., and ReVelle, C. (1974). The Maximal Covering Location Problem. In Papers of the Regional Science Association, 32, 101-118.

[5] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. The Journal of Machine Learning research, 12, 2825-2830.

[6] Roberts, D. R., et al. (2017). Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. Ecography, 40(8), 913-929.

[7] Dunning, I., Mitchell, S., and O'Sullivan, M. (2011). PuLP: A Linear Programming Toolkit for Python. Department of Engineering Science, The University of Auckland.