International Journal of English Literature and Social Sciences
Vol-11, Issue-6; November-December, 2026

Peer-Reviewed Journal
Journal Home Page Available: https://ijels.com/
Journal DOI: 10.22161/ijels

Spatial Stress Testing for Housing Valuation Model Risk Governance and Capacity-Constrained Review Allocation Evidence from California Census Block-Group Data and a Guangzhou Housing Sample

Xinyi Sun

Business School, University of Shanghai for Science and Technology, Shanghai 200093, China
Email: Xinyi520130@163.com

Received: 11 Oct 2026; Received in revised form: 03 Nov 2026; Accepted: 07 Nov 2026; Available online: 12 Nov 2026
©2026 The Author(s). Published by Infogain Publication. This is an open-access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

Abstract— *Housing valuation model selection and manual review require joint consideration of predictive accuracy, spatial applicability, and capacity constraints. Using 20,640 California census block-group observations and 1,824 Guangzhou housing records, this study conducts spatial stress tests and supplementary non-spatial review analyses. The results show that model rankings vary with spatial holdout design. Under a fixed feature specification, interquartile-range clipping reduces overall error but increases errors or underprediction bias for some sample groups. In the Guangzhou sample, under a nominal 10% review budget, the high-absolute-error capture rates of model-disagreement, training-only error-learning, and predicted-price-prioritization rules are 23.83%, 38.14%, and 41.31%, respectively, compared with 9.86% for random review. Rankings change under relative-error evaluation, partly because negative price predictions are identified. The findings highlight the importance of aligning model selection and review prioritization with deployment conditions, subgroup diagnostics, and loss objectives. The evidence is limited to offline prediction and screening; practical application requires further validation using target-market spatial and temporal information and actual review records..*

Keywords— ***housing valuation; model risk governance; regional risk; resource allocation; spatial cross-validation.***

I. INTRODUCTION

The managerial value of a housing valuation model cannot be determined solely by its overall prediction error. In tasks such as mass appraisal, mortgage underwriting, asset management, and risk screening, models face three practical constraints. First, a target property may be located in an area well covered by training observations or in an unseen area with sparse training coverage. Second, overall average error may conceal substantial errors in particular regions, value ranges, or sample groups. Third, personnel, time, and budgets for manual review are limited, making it impossible to subject every automated valuation to the same level of scrutiny. Evaluating housing valuation models is therefore not merely a technical question of which model has the lowest average error; it is also a managerial question of the conditions under which a model can be used and which cases should receive review priority.

Research on real estate valuation has long examined the relationships among housing attributes, spatial location, and prices. Hedonic price theory provides a classical basis for modeling associations between housing attributes and prices, while machine-learning methods improve the ability to fit nonlinear relationships and interactions. The validation design for spatial data also affects performance assessment. Random cross-validation may place geographically proximate observations in both training and test sets, approximating interpolation within the same market. Spatial grid grouping and regional holdout are closer to local spatial transfer or prediction in an entirely unseen area. Prior studies show that random splits alone

may overstate model performance when data have spatial structure [00000].

Model selection is also affected by hyperparameter tuning and data preprocessing. If the same test data are used both to select hyperparameters and to evaluate final performance, the resulting estimate may be subject to selection bias [00]. Likewise, using outer-test data to determine outlier thresholds, scaling parameters, or feature transformations may cause information leakage [0]. Model comparisons should therefore separate parameter selection, data processing, and outer-sample evaluation, while matching the validation design to the spatial scope of the intended prediction task.

The application of real estate appraisal in China also has specific institutional and operational characteristics. The Real Estate Appraisal Standard (GB/T 50291—2015) specifies requirements concerning the appraisal purpose, valuation date, appraisal object, procedures, methods, results, and appraisal report [0]. Consequently, predictive results obtained from historical foreign data cannot be transferred directly to the Chinese market without considering the appraisal object, valuation date, and business purpose. Applying machine-learning models to housing valuation in China requires redefining the observational unit—such as an individual dwelling, building, community, or administrative area—and the target variable and price unit, followed by temporal, regional, and target-market external validation.

This study uses two publicly available housing datasets that provide complementary information. The 20,640 observations from the 1990 U.S. Census for California census block groups are used to compare model performance under different spatial holdout designs and to examine training coverage and subgroup errors. A separate sample of 1,824 Guangzhou housing records is used to retrain models on Chinese data and compare risk-ranking rules under a fixed number of review slots [00000]. The two datasets are modeled and evaluated separately; error magnitudes expressed in different price units are not compared directly.

The geographic coordinates in the California data support spatial validation, whereas the housing attributes and prices in the Guangzhou data support non-spatial review screening. Because the publicly released Guangzhou file does not contain geographic identifiers or transaction-level dates, its supplementary analysis focuses on identifying large errors from common predictions and on the sensitivity of screening performance to the evaluation objective. The two datasets therefore provide evidence for spatial evaluation and review prioritization, respectively, rather than a cross-market transfer test of a single trained model.

The study addresses three questions concerning the connection between model evaluation and risk governance. First, how do model rankings and worst-region performance change across spatially mixed, local spatial-block, and complete regional holdout settings, and what do these results imply for determining model applicability? Second, do improvements in overall error resulting from training-fold feature processing coincide with changes in errors or underprediction bias for particular sample groups? Third, under a fixed number of review slots, how do predicted-price, model-disagreement, and training-only error-learning rules compare with random review in identifying severe prediction errors, and how sensitive are their rankings to absolute error, relative error, and abnormal predictions?

The study provides complementary evidence for three decisions: model selection, preprocessing assessment, and review prioritization. First, it compares model performance under spatially mixed, local spatial-block, and complete regional holdout conditions to identify differences between overall accuracy and adverse regional performance. Second, it compares overall errors, subgroup errors, and prediction bias before and after preprocessing on the same test observations. Third, using Guangzhou data with common out-of-fold predictions and a common review budget, it compares a simple price rule, model disagreement, and training-only error learning, and examines changes in rankings across error definitions and abnormal-output diagnostics. These analyses provide empirical information for determining model applicability, assessing preprocessing, and prioritizing review. The two datasets support spatial evaluation and non-spatial review analysis separately and do not constitute a cross-market transfer test of the same trained model.

II. LITERATURE REVIEW AND ANALYTICAL FRAMEWORK

2.1. Housing Valuation, Predictive Accuracy, and Model Risk: Conceptual Definitions

Housing valuation is the process of estimating the value of a specified appraisal object at a specified valuation date using observable information about the dwelling and its surrounding area. Traditional hedonic models estimate statistical associations between housing attributes—such as floor area, age, room count, location, and neighborhood environment—and prices [0]. Machine-learning models place greater emphasis on predictive performance and can fit complex relationships through regularization, tree-based methods, and other nonlinear approaches.

Research on housing valuation in China has also expanded from traditional hedonic analysis to data-driven mass appraisal. Zhang and Guo [0] review data-driven research on real estate mass appraisal, while Qu and Gu [0] survey applications of artificial intelligence in real estate appraisal. Wang et al. [0] examine geographically neural-network-weighted regression for housing valuation in Shenzhen, China. These studies provide a foundation for combining housing attributes, geographic information, and machine learning in valuation. Gyger et al. [0] further combine spatial Gaussian processes with tree boosting for hedonic real estate modeling. Building on this line of research, the present study shifts attention from the predictive performance of candidate models to the conditions under which valuation outputs are used, comparing the decision information provided by spatial validation, local errors, and review-capacity constraints.

Wang et al. [0] use second-hand housing records from Baiyun District, Guangzhou, covering 2023–2024, and make the associated data and code publicly available [0]. This study provides an accessible dataset for examining predictive evaluation using relatively recent Chinese housing observations. Based on the public workbook, the present study independently conducts model training, validation, and risk screening rather than adopting the prediction errors reported in the original study. Beyond average predictive accuracy, the supplementary analysis examines how review cases are ranked under a common automated valuation model and whether the rankings change with the definition of business loss.

Compared with studies that focus primarily on constructing valuation models and comparing their predictive performance, this study examines how model-evaluation results can inform risk governance and review prioritization. Specifically, it considers not only overall prediction errors under different spatial validation designs but also worst-region errors, training-sample coverage, subgroup bias, and the identification of high-error cases under a fixed review budget. The California and Guangzhou samples serve different purposes: spatial evaluation and non-spatial review testing, respectively. The Guangzhou analysis uses local model refitting and does not treat historical U.S. data as a substitute for the Chinese market or a single-region result as a general finding for China.

Predictive accuracy is defined as a model's error on outer-test observations and is evaluated using root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). Prediction and the identification of economic parameters address different objectives [0]. Accordingly, this study focuses on out-of-sample predictive performance and does not interpret regression coefficients, model predictions, or variable-interpretation measures as causal effects of housing attributes.

Model risk refers to potential problems arising under actual use conditions, including unstable predictions, increased regional errors, systematic bias in particular sample groups, insufficient training coverage, and inadequate review of model outputs. Model risk is not equivalent to overall RMSE. A model may have low average error across the full sample while producing substantial errors in an unseen region, a sparsely represented group, or the high-end target-value group. Model-risk assessment therefore considers average error, variation across folds, worst-region error, training coverage, and subgroup bias. In this study, "regional tail risk" refers operationally to adverse regional prediction performance, measured primarily by worst-region RMSE and minimum Skill; it is not treated as an unestimated financial loss quantile or conditional tail-risk measure.

2.2. Spatial Validation and Model Deployment Conditions

Spatial validation partitions training and test observations according to geographic locations or spatial groups to assess model performance under different geographic prediction conditions. Random five-fold cross-validation allows geographically proximate observations to appear in both training and test sets and mainly reflects prediction when the two sets are spatially mixed. It can approximate interpolation within the same market but should not be used alone to assess prediction in entirely unseen regions.

Grid-based spatial grouping first divides the study area into geographic grids and ensures that observations within the same grid do not appear simultaneously in the outer training and test sets. This reduces the likelihood that highly proximate observations occur in both sets and is primarily used to assess prediction under local spatial-block holdout. Because the ground distance represented by a degree of longitude varies with latitude, latitude–longitude grids are neither equidistant nor equal-area.

Leave-one-region-out validation partitions observations by geographic location and successively holds out one complete region as the outer test set. This more closely approximates offline prediction in an entirely unseen area. However, the boundaries are specified by the researcher or generated from the data and do not necessarily correspond to administrative divisions, real estate submarkets, or operational management boundaries. Results should therefore be interpreted under the specified spatial partition rather than as direct evidence of cross-city or cross-market deployment.

In this study, "deployment conditions" refer to the combination of spatial distance between prediction targets and training observations, sample coverage, covariate-distribution differences, target-variable distribution differences, and target-market boundaries. Spatial validation is intended not to identify one fixed split suitable for every business setting, but to make validation conditions as close as possible to the intended use of the model.

"Spatial stress testing" refers to assessing the sensitivity of model performance to changes in spatial application conditions by varying the spatial relationship between training and test observations under a given historical dataset, candidate-model set, and evaluation protocol. The designs include spatially mixed, local spatial-block, and complete regional holdout conditions, together with diagnostics of training coverage, covariate differences, overall prediction errors, and worst-region errors. This is an offline assessment of predictive stability and differs from stress testing in financial risk management, which estimates asset losses under prespecified macroeconomic shocks. This study does not construct exogenous scenarios involving housing-price shocks, interest-rate changes, or real estate policy changes, nor does it estimate actual economic losses in the housing market.

2.3. Subgroup Risk and Capacity-Constrained Review Allocation

Subgroup risk refers to differences in prediction errors or signed bias across groups defined in advance. The groups examined here are based on whether observations touch training-fold interquartile-range clipping boundaries, whether the target variable meets a prespecified high-end threshold, and spatial region. These are statistical groups defined by data attributes and model-processing rules; they are not demographic groups used in social-equity research, and the analysis does not assess discrimination against social groups.

The limited-review problem is methodologically related to selective prediction. Selective prediction allows a model to withhold automated predictions for some cases while evaluating both the proportion of cases retained for prediction and the associated prediction risk. Geifman and El-Yaniv [0] integrate a rejection mechanism into prediction and examine the trade-off between risk and coverage. Zaoui et al. [0] further study regression with a fixed rejection rate, highlighting the need to consider prediction and risk identification jointly. These studies indicate that reducing error among retained predictions cannot be evaluated independently of coverage.

Following the risk–coverage perspective in selective prediction, this study compares pre-prediction risk-ranking rules under a fixed number of review slots. The high-error capture rate measures the extent to which the selected list covers severe prediction errors, while the RMSE of unreviewed observations measures the error among cases retained for automated valuation. Together, these metrics describe offline screening performance. Cases selected for review have not necessarily received corrected valuations; the metrics therefore do not represent the final system error after human review.

"Capacity-constrained review allocation" is defined here specifically as prioritizing cases under a fixed number of review slots. Each case is treated as one review unit; slots are allocated to test batches according to a prespecified proportion, and cases are ranked within each batch using pre-prediction risk signals. The experiment does not reallocate slots across regions or estimate the actual review cost of individual cases. Model disagreement, distance to training observations, local sample sparsity, and covariate extrapolation can be calculated before the true target value is known and can therefore serve as pre-prediction ranking signals. The evaluation concerns risk identification per review slot, not optimal allocation under a monetary cost constraint.

prediction errors that are included in the proposed review list. It measures offline risk-ranking performance; it does not mean that human review has successfully corrected the predictions or generated cost savings or operating returns. The RMSE of unreviewed observations likewise describes the error composition of the cases retained for automated valuation and does not indicate that the overall valuation system has become more accurate.

2.4. Research Questions and Analytical Framework

Based on these concepts, the study develops an analytical framework linking validation conditions, training coverage, predictive performance, subgroup risk, and review screening. Random five-fold, latitude–longitude grid grouping, and regional holdout are used to compare overall accuracy and regional tail risk under different spatial prediction conditions. Nearest-training-observation distance, fifth-nearest-neighbor distance, the number of training observations within 10 km, and covariate-distribution differences are used to describe training coverage and the extent of extrapolation. Renewed nested tuning and training-fold preprocessing are used to examine the sensitivity of model rankings to parameter selection and data processing. Finally, pre-prediction risk rules are compared separately in the California and Guangzhou samples under fixed review budgets, with additional analyses of error definitions and abnormal predictions in the Guangzhou sample.

Three research questions follow.

Research Question 1: How do candidate-model rankings and worst-region performance vary across spatially mixed, local spatial-block, and complete regional holdout conditions? What information do overall error and adverse regional performance provide for determining model applicability?

Research Question 2: Do improvements in overall error resulting from training-fold feature construction and interquartile-range clipping coincide with deterioration in errors or signed prediction bias for particular sample groups?

Research Question 3: Under a fixed number of review slots, how do predicted-price, model-disagreement, and training-only error-learning rules compare with random review in identifying severe prediction errors? How do their relative rankings change with absolute-error, relative-error, and abnormal-output definitions?

The research questions are not framed as a universal ranking of algorithms. They are restricted to model-risk evaluation under the specified datasets, spatial partitions, and parameter-search spaces. The theoretical and managerial value lies in showing that housing valuation model selection should match the intended deployment setting, that average error cannot replace regional and subgroup diagnostics, and that limited review capacity should be prioritized using information available before the true target value is observed.

III. DATA AND RESEARCH DESIGN

3.1. Data Sources, Variables, and Spatial Dependence

The study uses the California Housing dataset, which contains 20,640 observations aggregated at the 1990 U.S. Census block-group level. Each observation represents a census block group rather than an individual dwelling or a contemporary real estate transaction. The dataset contains eight explanatory variables and one target variable. MedHouseVal represents the median house value for a block group and is measured in units of USD 100,000. An offline CSV file prepared by the authors is used for replication, and the target and explanatory-variable fields are checked against their definitions.

The observational unit in the California data is a census block group, and the target variable is the block-group median house value. The sample size, complete geographic coordinates, and public availability make the dataset suitable for comparing predictive stability under different spatial validation designs [000]. The resulting errors are interpreted at the block-group aggregate level and are not directly compared with the dwelling-level price errors in the Guangzhou sample.

The explanatory variables are median household income (MedInc), median house age (HouseAge), average rooms per household (AveRooms), average bedrooms per household (AveBedrms), block-group population (Population), average household occupancy (AveOccup), latitude, and longitude. Population, AveOccup, AveRooms, and AveBedrms exhibit substantial skewness or extreme values; their effects are examined through training-fold preprocessing. Latitude and longitude are used both to construct spatial groups and as model predictors. The use of unlabeled coordinates to define outer validation regions is therefore distinguished from the use of training observations to fit the prediction models.

The maximum value of MedHouseVal is 5.00001. Applying the threshold MedHouseVal ≥ 5.0 identifies 992 observations, or 4.81% of the sample, which are defined as the "high-end target-value group"; all remaining observations form the "other group." This group includes observations close to or at the coding ceiling, but not every observation satisfying the ≥ 5.0 threshold is assumed to equal the exact coded maximum of 5.00001. Error analyses for this group evaluate predictions of the recorded target values and cannot recover the true market values above the coding ceiling.

Table 1 reports variable definitions and descriptive statistics. The maximum values of AveRooms, AveBedrms, Population, and AveOccup are substantially higher than their means, suggesting large observations that may affect model fitting. The main specification retains the original variables, while robustness analyses compare logarithmic transformation, second-order spatial features, and training-fold interquartile-range clipping.

To describe spatial association in the target variable, Moran's I is calculated using an equal-weight eight-nearest-neighbor spatial weights matrix, with longitude adjusted by the sample mean latitude. The statistic summarizes whether neighboring observations tend to have similar target values. Under the specified weights, Moran's I is 0.799, indicating positive spatial clustering in housing values. This statistic is used descriptively to motivate spatial validation; no unreported significance test or causal interpretation is implied.

Table 1. Variable Definitions and Descriptive Statistics

Variable	Definition	Mean	Std. Dev.	Minimum	Maximum
MedInc	Median household income (USD 10,000)	3.871	1.900	0.500	15.000
HouseAge	Median house age (years)	28.639	12.586	1.000	52.000
AveRooms	Average rooms per household	5.429	2.474	0.846	141.909
AveBedrms	Average bedrooms per household	1.097	0.474	0.333	34.067
Population	Block-group population	1425.477	1132.462	3.000	35682.000
AveOccup	Average household occupancy	3.071	10.386	0.692	1243.333
Latitude	Latitude	35.632	2.136	32.540	41.950
Longitude	Longitude	-119.570	2.004	-124.350	-114.310
MedHouseVal	Median house value (USD 100,000)	2.069	1.154	0.150	5.000

Note: The sample size is 20,640 for all variables. Descriptive statistics are calculated from the CSV file used in the study.

3.2. Model Specifications

Ordinary least squares (OLS) serves as the linear benchmark and estimates the model by minimizing the sum of squared residuals. Ridge regression applies an L2 penalty to shrink regression coefficients [0], whereas Lasso regression applies an L1 penalty that can shrink some coefficients to zero [0]. These methods compare ordinary linear fitting, coefficient shrinkage, and sparse estimation. Let y_i denote the target value for observation i, x_i the explanatory-variable vector, β_0 the intercept, and β the coefficient vector; the linear prediction is ŷ_i = β_0 + x_i^{T}β. OLS, Ridge, and Lasso solve the following objective functions:

OLS: $\min \sum_i (y_i - \beta_0 - x_i^T\beta)^2$

Ridge: $\min \sum_i (y_i - \beta_0 - x_i^T\beta)^2 + \alpha \|\beta\|_2^2$

Lasso: $\min(\frac{1}{2n}) \sum_i (y_i - \beta_0 - x_i^T\beta)^2 + \alpha \|\beta\|_1$

where α ≥ 0 is the penalty strength and the intercept is not penalized. Because the Ridge and Lasso objective functions use different scaling conventions, their α values are not directly comparable. Lasso is estimated using coordinate descent [0]. All linear models are used for predictive association only; their coefficients are not interpreted causally.

To provide a nonlinear benchmark capable of capturing interactions, the study includes histogram-based gradient boosting regression (HGBR). Gradient boosting constructs an additive tree model by sequentially fitting the current residual direction [0]. HGBR is not treated as a spatial model; its predictive performance is evaluated under different spatial validation conditions. The four models represent an ordinary linear benchmark, coefficient shrinkage, sparse estimation, and nonlinear fitting, respectively.

3.3. Outer Validation Designs and Spatial Coverage Measures

Outer-test observations are used for final performance evaluation, while model selection and estimation of data-processing parameters are restricted to the outer training data. The main analysis includes random five-fold validation, five-fold grouping by 0.5° latitude–longitude grids, and K=5 leave-one-region-out validation. Supplementary analyses examine K=4 and K=6 regional holdouts and 50-km grid grouping based on the California Albers projection. The three main designs each produce five

outer test folds; K=4 and K=6 regional holdouts produce four and six outer folds, respectively. Errors are calculated using the number of folds specific to each design.

Random five-fold validation shuffles observations using a fixed random seed of 42, allowing geographically proximate observations to appear in both the outer training and test sets. It primarily evaluates prediction when training and test observations are spatially mixed and approximates interpolation within the same market or routine batch scoring.

For 0.5° grid grouping, latitude and longitude are each divided by 0.5° and rounded down to obtain grid identifiers. GroupKFold then ensures that observations from the same grid do not appear simultaneously in the outer training and test sets. The grids are defined in angular coordinates and are not equidistant or equal-area. Because the ground distance represented by one degree of longitude varies with latitude, actual grid widths differ across locations. No additional spatial buffer is imposed; this design evaluates local spatial-block holdout rather than strict extrapolation free of all geographic proximity.

For regional holdout, latitude and longitude are extracted for all observations, standardized using the full-sample coordinate means and standard deviations, and grouped into K geographic regions using K-means. Each region is successively held out as an outer test set. The main analysis uses K=5, with K=4 and K=6 in sensitivity analyses. The partition does not use target values or other housing attributes, but it does use unlabeled geographic coordinates from all observations. It is therefore an offline partitioning design in which target locations are known in advance; the resulting region identifiers do not correspond to administrative areas or real estate submarkets.

The outer spatial grouping and prediction-model fitting stages should be distinguished. Unlabeled coordinates from test observations are used to construct regional partitions, whereas housing-value labels from the outer test set are not used to estimate OLS, Ridge, Lasso, or HGBR parameters or to select hyperparameters. Because latitude and longitude are also model predictors, the design should not be described as a prospective validation that uses no information from test observations. More precisely, it is an offline spatial stress test that uses geographic locations known in advance without using test labels.

The three main validation designs approximate spatially mixed prediction, local spatial-block transfer, and complete regional holdout, respectively, but do not directly reproduce actual business deployment. Real applications may also be affected by transaction timing, policy conditions, market boundaries, covariate distributions, and training-sample density. The designs are therefore interpreted as different spatial evaluation conditions.

To describe geographic proximity between test and training observations, the Haversine formula is used to calculate the distance from each outer-test observation to its nearest training observation. The median, 90th percentile, and proportion of exactly coincident coordinates are reported. The distance to the fifth-nearest training observation and the number of training observations within a 10-km radius are also calculated. These measures describe training-sample coverage but do not, by themselves, capture the full difficulty of spatial extrapolation.

Even when two test observations have similar nearest-neighbor distances, the number and geographic configuration of surrounding training observations, as well as income and housing-attribute distributions, may differ substantially. The study therefore reports nearest-neighbor distance, fifth-nearest-neighbor distance, the proportion of test observations with no training observation within 10 km, and the two-sample Kolmogorov–Smirnov (KS) statistic for non-geographic explanatory variables. The KS statistic is the maximum absolute difference between the empirical cumulative distribution functions of the training and test samples for a given variable. These coverage and distribution measures jointly describe evaluation conditions; they are not used to isolate the independent effect of any one factor on prediction error.

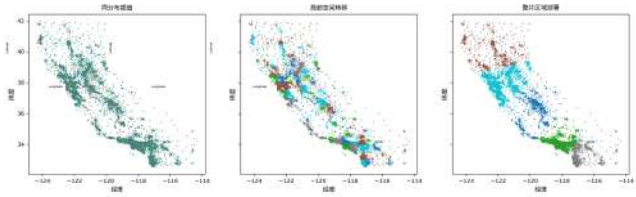


Fig.1. Geographic Partitions under the Three Outer Validation Designs

Table 2. Spatial Proximity Characteristics and Approximate Scenarios under Different Validation Designs

Validation design	Approximate application scenario	Median distance to nearest training observation (km)	90th-percentile distance (km)	Proportion of exactly coincident coordinates
Random five-fold	Interpolation within the same market / routine batch scoring	0.00	2.39	54.33%
0.5° grid grouping	Transfer to a neighboring spatial block / local expansion	10.13	23.11	0.00%
Leave-one-region-out	Deployment in an entirely unseen region / cross-region application	82.19	120.91	0.00%

Note: Distances are calculated using the Haversine formula. "Exactly coincident coordinates" means identical latitude and longitude and does not imply that all attributes are identical.

Table 3. Spatial Coverage and Covariate-Distribution Differences across Validation Designs

Validation design	Median nearest-neighbor distance (km)	Median fifth-nearest-neighbor distance (km)	Test observations with no training data within 10 km (%)	Mean KS across six variables
Random five-fold	0.00	1.11	1.73	0.0158
0.5° grid	10.13	12.08	51.95	0.0733
Regional holdout (K=5)	82.19	84.62	97.89	0.1395
Projected 50-km grid	10.19	12.27	52.24	0.0792

Note: Nearest-neighbor and fifth-nearest-neighbor distances are medians of geographic distances from test observations in each fold to training observations. The KS statistic is calculated by averaging the six non-geographic variables within each fold and then averaging across folds; it is descriptive only. The projected 50-km grid uses EPSG:3310

3.4. Nested Tuning, Preprocessing, and Robustness Design

Nested cross-validation separates model selection from model-performance evaluation. Outer cross-validation generates out-of-fold predictions that have not been used for model selection and is used to assess out-of-sample performance. Inner cross-validation is conducted only within the current outer training sample to select regularization strengths, tree-model parameters, and other learnable settings. This design prevents outer-test target values from being used to select parameters and then reused to evaluate the selected model, thereby reducing model-selection bias.

In the main analysis, three-fold inner cross-validation is conducted within each outer training fold. Random KFold is used for inner splits under random five-fold outer validation. GroupKFold is used for inner splits under grid grouping and regional holdout, with only the spatial-group labels corresponding to the current outer training observations passed to the inner procedure. Outer-test observations do not participate in inner parameter selection.

OLS is not subject to hyperparameter search. The Ridge regularization candidates are $\alpha \in \{0.1, 1, 10, 100, 1{,}000\}$; the Lasso candidates are $\alpha \in \{0.0003, 0.001, 0.003, 0.01, 0.03, 0.1\}$. For HGBR, the learning rate is 0.1, the maximum

number of iterations is capped at 120, the maximum leaf-node candidates are {15, 31}, and the L2 regularization candidates are {0, 1}, yielding four candidate configurations. The value 120 is an upper limit on the number of iterations. Each model selects the configuration with the lowest inner cross-validation RMSE, is refitted on the complete outer training set, and is evaluated only on the outer test set.

The parameter search is not exhaustive for every algorithm. The number of candidate parameters, search spaces, and computational budgets differ across models. The results should therefore be interpreted as relative performance under the specified experimental protocol, not as evidence that one algorithm must outperform another after exhaustive optimization. Spatial partitioning, training-sample structure, and parameter selection may jointly affect model rankings.

The main analysis uses the eight original explanatory variables. To examine feature representation and preprocessing, the robustness analysis adds second-order spatial features: Population is transformed using log(1 + Population), latitude and longitude are centered using means from the current training fold, and centered latitude-squared, longitude-squared, and latitude–longitude interaction terms are constructed.

With the feature representation held constant, an additional specification applies interquartile-range (IQR) clipping within each training fold. The variables clipped are MedInc, HouseAge, AveRooms, AveBedrms, Population_log, and AveOccup; latitude, longitude, and their derived spatial features are excluded. For each clipped variable, Q1, Q3, and the IQR are calculated using only the corresponding training data. The lower and upper bounds are Q1 − 1.5 × IQR and Q3 + 1.5 × IQR, respectively. Values below or above these bounds are replaced by the corresponding bound.

IQR clipping changes explanatory-variable values only; it does not delete observations or clip the target variable. Clipping thresholds are estimated separately in each outer training fold and then applied unchanged to the corresponding test fold. Within inner cross-validation, preprocessing parameters are likewise estimated only from the relevant inner training sample. Standardization parameters for OLS, Ridge, and Lasso are also calculated exclusively from their respective training data.

Accordingly, leakage prevention applies to learnable stages such as model fitting, standardization, feature transformation, outlier thresholds, and hyperparameter selection. The use of full-sample unlabeled coordinates to construct outer regional partitions is a prespecified offline spatial-partition design and should be distinguished from the prediction-model training pipeline.

3.5. Evaluation Metrics and Model-Selection Criteria

The study reports root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). For each outer validation design, the out-of-fold prediction obtained once for each observation is pooled to calculate overall metrics. Because outer test folds may differ in sample size and target distribution, the arithmetic mean of fold-specific R^2 values is not used to rank models overall. For the Guangzhou repeated validation, pooled out-of-fold metrics are first calculated within each repetition and then summarized across five repetitions.

$$RMSE = \sqrt{(1/n)\,\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAE = (1/n)\,\Sigma_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$

$$R^2 = 1 - \left[\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2\right] / \left[\Sigma_{i=1}^{n}(y_i - \bar{y}_i)^2\right]$$

To assess whether a model outperforms a simple deployable baseline, the target mean of the outer training set is used to predict the outer test set in each fold, and the RMSE skill score is defined as follows:

Skill = 1 − RMSE_model / RMSE_training-mean baseline

A Skill score greater than zero indicates performance better than the training-mean baseline; a negative value indicates performance worse than the simple mean prediction in that fold. The baseline used in Skill is the target mean of the corresponding outer training set, which differs from the evaluation-sample mean used in the denominator of R^2. In addition to overall prediction errors, the study reports fold-level RMSE standard deviations, worst-fold RMSE, minimum Skill, regional prediction bias, and errors for the high-end target-value group to examine risk distributions across regions and sample groups.

3.6. Risk Screening under a Fixed Manual-Review Budget

The California review experiment fixes the out-of-fold predictions from HGBR using the eight original explanatory variables as the object of diagnosis and compares risk-ranking rules on the same predictions and test observations. The design examines whether large errors from a given model can be identified using pre-prediction signals; it does not imply that HGBR should be selected in every spatial setting. Model selection is evaluated separately through the preceding candidate-model comparisons. Review budgets of 5%, 10%, and 20% are applied to each outer test batch. Screening uses only explanatory variables, training coverage, and model predictions available at the time, not

the true housing values in the test set. Because corrected values after actual appraisal review are not observed, the results describe offline screening performance.

The main-text results for California focus on three screening strategies: random review, prioritization by disagreement between HGBR and Ridge, and a prespecified composite risk score. Fifth-nearest-neighbor distance, local training-sample sparsity, and the number of covariates outside the training range are components of the composite score; their individual rankings are not reported as separate main-text findings. Covariate exceedances are counted using the 1st and 99th percentiles of six non-geographic explanatory variables in the outer training set. Model disagreement, fifth-nearest-neighbor distance, local sparsity, and covariate exceedances are converted to percentile ranks within the unlabeled test batch and assigned prespecified weights of 0.40, 0.30, 0.20, and 0.10, respectively. The weights are not fitted or optimized using outer-test errors, and review slots are allocated within each outer test fold according to the prespecified budget.

Screening performance is evaluated using the high-error capture rate and the RMSE of unreviewed observations. Within each validation design, severe prediction errors are defined ex post as the 10% of outer-test observations with the largest absolute prediction errors. True target values are used only to identify this evaluation set after prediction and to calculate evaluation metrics; they are not used to generate risk scores or determine the review list. The high-error capture rate is the proportion of ex post high-error observations included in the proposed review list.

Random review is repeated 100 times, and the average result is used as the benchmark. Because random review is subject to sampling variation, capture rates of 10.0% or 10.1% are descriptive benchmarks and should not be interpreted as statistically significant differences in the absence of repeated-sampling standard deviations or confidence intervals.

The RMSE of unreviewed observations describes the prediction error among cases that remain subject to automated valuation after screening. It does not mean that selected cases have been corrected by human reviewers or that the overall predictive ability of the model has improved. The high-error capture rate and unreviewed-sample RMSE are therefore offline risk-screening metrics and cannot be directly translated into review success rates, cost savings, or net operating benefits.

The experiment assumes that unlabeled explanatory variables, training-coverage measures, and model predictions for the review batch are available before review slots are allocated. It therefore more closely resembles offline batch valuation than a real-time system in which cases arrive sequentially. All regional partitions use the same historical California dataset. To examine some review rules in another market sample, a supplementary non-spatial analysis is conducted using publicly available Guangzhou housing data; the design is described in Section 3.7. Neither dataset contains actual human correction outcomes, so offline differences in screening performance cannot be interpreted as realized review benefits.

3.7. Supplementary Analysis Using the Guangzhou Housing Sample

3.7.1. Data Source and Scope

The Guangzhou supplementary analysis uses the second-version Figshare workbook associated with Wang et al. [00]. The original study describes the sample as second-hand housing records from Lianjia in Baiyun District, Guangzhou, during 2023–2024. The geographic, temporal, and transaction descriptions in this study follow the data provider's account and have not been independently verified against transaction-registration records. Sheet1 of "Second-hand Housing Price Data.xlsx" contains 1,824 observations and the following fields: record number, housing price, layout, floor, floor area, orientation, and decoration. Housing prices are measured in CNY 10,000 and floor area in square meters. The record number, other workbook sheets, and the original authors' prediction results are excluded from modeling.

None of the seven fields contains missing values, and no exact duplicate records remain after excluding the record number. The mean and median housing prices are CNY 2.3285 million and CNY 2.049 million, respectively, with a range of CNY 190,000 to CNY 20.069 million; mean floor area is 92.70 square meters. All observations are retained. The public file supports modeling with housing attributes and non-spatial review screening but does not provide latitude and longitude, community identifiers, or transaction-level dates. Accordingly, the supplementary analysis does not include geographic holdout or future-period prediction.

3.7.2. Feature Processing and Nested Validation

The housing-layout string is parsed into the numbers of bedrooms, living rooms, kitchens, and bathrooms; floor area is retained as a numeric variable. Four directional indicators are constructed according to whether the orientation field contains east, west, south, or north. Floor and decoration are one-hot encoded. All numerical standardization and categorical-encoding parameters are estimated within the corresponding training fold, with prespecified handling of unseen categories. The target variable is standardized within the training fold, and predictions are transformed back to CNY 10,000 for error calculation.

The Guangzhou experiment uses raw predictions converted back to CNY 10,000, without imposing a non-negativity constraint or replacing negative predictions with zero, so that all review rules share the same prediction baseline. Negative predictions that are inconsistent with the meaning of housing prices are reported separately in terms of their frequency and contribution to screening. Tables 13–15 evaluate offline rankings under this baseline; a combined procedure that first screens abnormal outputs and then ranks remaining cases was not included in the budget comparison.

To prevent identical encoded attribute combinations from appearing in both training and test sets, observations are grouped by their complete encoded attribute combinations, yielding 1,787 groups, and a shuffled grouped five-fold outer validation is used. These groups represent attribute combinations, not communities or geographic areas. Three-fold grouped parameter selection is conducted within each outer training set. The outer random seeds are 42, 43, 44, 45, and 46, yielding 25 outer training–test combinations. Each observation receives exactly one out-of-fold prediction in each repetition. The five repetitions use the same dataset; their means and standard deviations describe sensitivity to data partitioning and are not treated as five independent market samples.

The candidate valuation models are OLS, Ridge, Lasso, and HGBR. Ridge α candidates are 0.1, 1, 10, 100, and 1,000; Lasso α candidates are 0.0003, 0.001, 0.003, 0.01, 0.03, and 0.1. HGBR uses a learning rate of 0.1, a maximum of 120 iterations, a maximum leaf-node count of 15 or 31, L2 regularization of 0 or 1, and a minimum of 20 samples per leaf; early stopping is explicitly disabled. Model and parameter selection is based on the lowest mean RMSE in inner three-fold validation, followed by refitting on the complete outer training set. The Guangzhou and California experiments are conducted separately, and their parameter-selection results and prediction records are not pooled.

3.7.3. Review Rules under Common Predictions

Within each outer fold, an automated valuation model is selected using inner-validation results. Five review rules are compared on the same out-of-fold predictions and test observations: random review, predicted-price prioritization, covariate-exceedance prioritization, model-disagreement prioritization, and training-only error learning. Predicted-price prioritization ranks observations from highest to lowest predicted price and does not use true prices. Covariate exceedances count the number of five numerical attributes outside the 1st–99th percentile ranges of the outer training data and include unseen floor or decoration categories. Model disagreement is the absolute difference between HGBR and Ridge predictions for the same observation.

Training-only error learning uses absolute residuals from inner out-of-fold predictions of the selected valuation model as the risk target. Housing attributes, predictions from the selected model, and HGBR–Ridge disagreement are used as risk predictors. A separate risk regressor is fitted to estimate potential absolute errors for outer-test observations. Parameter selection and risk-model fitting are restricted to outer training data; outer-test prices are not used to optimize scores. The risk regressor is an HGBR with a learning rate of 0.05, 100 maximum iterations, a maximum of seven leaf nodes, a minimum of 20 samples per leaf, and L2 regularization of 1; early stopping is disabled, and the settings are not adjusted based on outer-test results.

3.7.4. Review Budgets and Evaluation Definitions

Nominal review budgets of 5%, 10%, and 20% are applied, with the number of slots in each outer test fold rounded down according to the budget proportion. Under the nominal 10% budget, 180 observations are selected in each repetition, corresponding to 9.87% of the full sample of 1,824. Ties in risk scores are resolved using a fixed random order independent of true prices. Random review is simulated 1,000 times and averaged. Each screening metric is first calculated within each repetition and then summarized using the mean and sample standard deviation across the five repetitions. For random review, the 1,000 simulations are averaged within each repetition before the across-repetition summary. The five repetitions are not independent market samples, and the 25 outer folds are not treated as 25 independent observations for ordinary significance testing.

In each repetition, all out-of-fold absolute errors are pooled, and the 183 largest are defined as the high-absolute-error set. The capture rate is the proportion of these 183 observations selected for review; the review-list hit rate is the proportion of selected cases that belong to the high-error set. Because all rules use exactly the same automated predictions, overall RMSE is unchanged by review ranking. The unreviewed RMSE evaluates only the subset retained for automated valuation.

Without changing the model, risk scores, or review lists, a high-relative-error set is separately defined using |predicted price − recorded price| / recorded price. Relative-error capture rates and the mean absolute percentage error (MAPE) of unreviewed observations are then calculated. This sensitivity analysis examines dependence on the loss definition; it does not retrain rules using test results or equate relative-error differences with actual business losses.

IV. EMPIRICAL RESULTS

The results first report pooled out-of-fold prediction metrics, followed by fold-level stability and worst-case performance. Pooled out-of-fold RMSE, MAE, and R^2 are calculated after combining the single out-of-fold prediction for each observation and reflect overall error across all test observations. By contrast, the "mean RMSE ± SD" in Table 5 is calculated by first computing RMSE within each outer test fold and then taking the arithmetic mean and standard deviation across folds. Because regional holdout folds may differ in sample size and target distribution, these two summaries are not equivalent. Pooled out-of-fold metrics are used for overall model ranking, while fold means, standard deviations, and worst-fold metrics are used to assess variation and adverse regional performance.

Before comparing model performance, the spatial coverage of training observations is examined across outer validation designs. According to Tables 2 and 3, the median distance from a test observation to its nearest training observation is 0.00 km under random five-fold validation, 10.13 km under 0.5° grid grouping, and 82.19 km under K=5 regional holdout. The corresponding median distances to the fifth-nearest training observation are 1.11, 12.08, and 84.62 km. The proportions of test observations with no training observation within 10 km are 1.73%, 51.95%, and 97.89%, respectively. These results indicate that spatial partitioning changes both nearest-observation distance and local training coverage.

4.1. Validation Design Changes Model Performance and Rankings

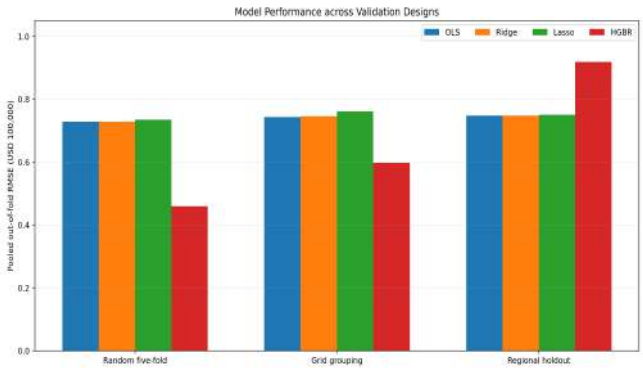


Fig.2. Pooled Out-of-Fold RMSE across Validation Designs

The empirical distributions of six non-geographic explanatory variables are also compared between training and test sets. For each fold, the two-sample KS statistic is calculated for each variable, averaged across variables, and then averaged across outer folds. The resulting mean KS statistics are 0.0158 for random five-fold validation, 0.0733 for 0.5° grid grouping, and 0.1395 for K=5 regional holdout. These results indicate differences in training coverage and covariate distributions across validation conditions. The regional prediction task therefore differs from the spatially mixed prediction condition approximated by random five-fold validation. The KS statistic is descriptive and does not identify the independent contribution of spatial distance, sample density, or covariate change to prediction error.

Figure 2 and Table 4 show that relative predictive performance depends on the spatial evaluation condition. Under random five-fold validation and 0.5° grid grouping, HGBR achieves pooled out-of-fold RMSE values of 0.4601 and 0.5977, respectively, approximately 36.8% and 19.6% lower than the corresponding OLS benchmarks. Under the spatially mixed and local spatial-block holdout conditions examined here, nonlinear fitting yields lower overall error.

Under K=5 complete regional holdout, HGBR has a pooled out-of-fold RMSE of 0.9192, compared with 0.7480 for both OLS and Ridge; the model ranking is reversed. Relative to random five-fold validation, HGBR's RMSE increases by approximately 99.8%, whereas Ridge's increases by about 2.7%. This difference shows that an accuracy advantage under spatially mixed observations cannot substitute for evaluation in an entirely unseen region. In this regional partition, the simpler linear models provide useful stability benchmarks.

The three designs simultaneously change training-sample composition, local coverage, and variable distributions. The results therefore represent combined performance under the specified validation designs and do not isolate the independent effect of any single spatial factor. Model selection should consider both overall error and adverse regional performance rather than generalizing a ranking from one scenario to all valuation tasks.

Table 4. Pooled Out-of-Fold Performance of the Main Models

Validation scenario	Model	RMSE	MAE	R^2
Random five-fold	OLS	0.7284	0.5317	0.6015
Random five-fold	Ridge	0.7284	0.5317	0.6015
Random five-fold	Lasso	0.7346	0.5370	0.5947

Validation scenario	Model	RMSE	MAE	R^2
Random five-fold	HGBR	0.4601	0.3060	0.8410
Grid grouping	OLS	0.7430	0.5448	0.5855
Grid grouping	Ridge	0.7462	0.5474	0.5819
Grid grouping	Lasso	0.7615	0.5631	0.5645
Grid grouping	HGBR	0.5977	0.4127	0.7317
Leave-one-region-out	OLS	0.7480	0.5361	0.5798
Leave-one-region-out	Ridge	0.7480	0.5365	0.5798
Leave-one-region-out	Lasso	0.7501	0.5393	0.5774
Leave-one-region-out	HGBR	0.9192	0.6647	0.3654

Note: Metrics are calculated by pooling one out-of-fold prediction per observation. RMSE and MAE are measured in units of USD 100,000. Bold indicates the model with the lowest RMSE in each scenario.

4.2. Average Advantage Does Not Guarantee Regional Stability

Table 5 further compares error variation and worst-fold performance across outer test folds. The mean RMSE in Table 5 is the arithmetic mean of fold-specific RMSE values, whereas the RMSE in Table 4 is calculated by pooling all out-of-fold predictions. Because regional holdout folds differ in sample size, the two summaries use different weighting and should not be treated as equivalent.

Under random five-fold validation, HGBR has a mean fold RMSE of 0.4600, a standard deviation of 0.0117, a worst-fold RMSE of 0.4731, and a minimum Skill of 0.589. Under grid grouping, the corresponding values are 0.5924, 0.0887, 0.7112, and 0.418. Under regional holdout, the mean fold RMSE rises to 0.8156, the standard deviation is 0.2579, the worst-fold RMSE reaches 1.2715, and the minimum Skill is −0.068. This indicates that HGBR performs worse than the training-mean baseline in at least one held-out region.

By comparison, Ridge has a mean fold RMSE of 0.7065, a standard deviation of 0.0625, a worst-fold RMSE of 0.7951, and a minimum Skill of 0.335 under regional holdout. The corresponding OLS values are 0.7070, 0.0611, 0.7951, and 0.335. Under the current regional partition, OLS and Ridge exhibit less fold-to-fold variation than HGBR. This finding describes stability in the current historical dataset and partition and should not be interpreted as a universal advantage in all real estate markets.

Model selection should therefore not rely solely on pooled RMSE rankings. For tasks involving prediction in entirely unseen regions, the worst-region error, minimum Skill, regional prediction bias, and training-sample coverage should also be considered.

Table 5. Fold-Level Stability and Worst-Case Performance of the Main Models

Validation scenario	Model	Mean RMSE ± SD	Worst-fold RMSE	Mean Skill	Minimum Skill
Random five-fold	OLS	0.7283±0.0167	0.7456	0.369	0.349
Random five-fold	Ridge	0.7283±0.0167	0.7456	0.369	0.349
Random five-fold	Lasso	0.7344±0.0158	0.7460	0.363	0.347
Random five-fold	HGBR	0.4600±0.0117	0.4731	0.601	0.589
Grid grouping	OLS	0.7375±0.1005	0.8998	0.366	0.255
Grid grouping	Ridge	0.7409±0.0987	0.8998	0.363	0.255

Validation scenario	Model	Mean RMSE ± SD	Worst-fold RMSE	Mean Skill	Minimum Skill
Grid grouping	Lasso	0.7570±0.0930	0.8999	0.349	0.255
Grid grouping	HGBR	0.5924±0.0887	0.7112	0.492	0.418
Leave-one-region-out	OLS	0.7070±0.0611	0.7951	0.403	0.335
Leave-one-region-out	Ridge	0.7065±0.0625	0.7951	0.404	0.335
Leave-one-region-out	Lasso	0.7113±0.0617	0.7956	0.400	0.335
Leave-one-region-out	HGBR	0.8156±0.2579	1.2715	0.312	-0.068

Note: Mean RMSE ± SD is the arithmetic mean and sample standard deviation of outer-fold RMSE values. The pooled out-of-fold metrics in Table 4 are calculated from all outer-test predictions and may therefore differ.

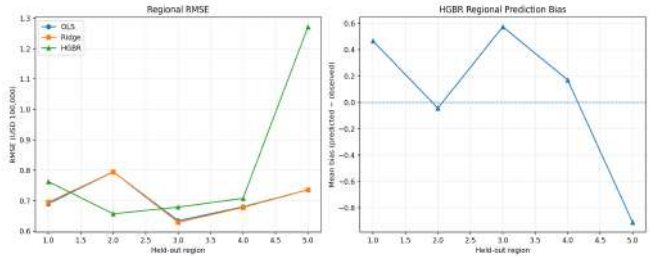


Fig.3. Regional Errors and Prediction Bias under Leave-One-Region-Out Validation

Figure 3 shows that model instability is geographically heterogeneous. HGBR's RMSE is only 0.6568 in Region 2 but rises to 1.2715 in Region 5. In Region 5, it underpredicts the mean housing value from 2.2459 to 1.3358, yielding a mean bias of −0.9101. OLS has an RMSE of 0.7356 and a mean bias of only −0.0930 in the same region. HGBR overpredicts in Regions 1 and 3 by 0.4661 and 0.5739, respectively. Its errors therefore do not reflect a uniform shift but rather asymmetric extrapolation across different geographic distributions.

Table 6. Sample Characteristics and Model Risk under Regional Holdout

Region	Sample size	Mean target value	OLS RMSE	Ridge RMSE	HGBR RMSE	HGBR bias
1	1794	1.219	0.690	0.695	0.763	0.466
2	8843	2.303	0.795	0.795	0.657	-0.043
3	969	0.926	0.634	0.629	0.679	0.574
4	2214	1.774	0.680	0.678	0.708	0.170
5	6820	2.246	0.736	0.736	1.272	-0.910

Note: Bias is defined as predicted value minus observed value. Regions are generated by K-means using standardized latitude and longitude and do not represent administrative areas.

4.3. Sensitivity to Spatial Partitioning Rules

The main regional holdout uses K-means to form five geographic regions; the number and boundaries of these regions do not correspond to actual administrative divisions or real estate submarkets. To examine sensitivity to spatial partitioning, the analysis compares K=4, K=5, and K=6 regional holdouts, as well as a projected 50-km grid validation, with inner parameter selection repeated under each design. The following analysis is based on the renewed nested-tuning results in Table 7.

To examine the combined effects of spatial partitioning and parameter selection, three-fold inner hyperparameter searches are repeated under K=4, K=5, and K=6 regional holdout. A 50 km × 50 km projected grid is also constructed using the California Albers projection (EPSG:3310). K=4, K=5, and K=6 regional holdout produce four, five, and six outer test folds, respectively. All K-based experiments use the same historical dataset and are not independent external samples.

After renewed nested tuning, HGBR's pooled out-of-fold RMSE values under K=4, K=5, and K=6 regional holdout are 0.8009, 0.9192, and 0.8089, respectively, all higher than the corresponding OLS values of 0.7775, 0.7480, and 0.7524. Under projected 50-km grid validation, HGBR's pooled RMSE is 0.5853, below the OLS value of 0.7383. These results show that model rankings vary with the prediction scenario under the specified dataset, spatial partitions, parameter candidates, and evaluation protocol. They do not establish the general superiority of one algorithm over another.

The comparisons indicate that local spatial-block holdout and complete regional holdout provide different

information for model selection. Because the designs also change training composition, spatial coverage, and inner parameter-selection conditions, they are used to assess scenario suitability rather than to isolate the independent effect of a single factor.

Table 7. Predictive Performance and Regional Error after Renewed Nested Tuning under Different Spatial Partitions

Partition design	OLS pooled	Ridge pooled	HGBR pooled	OLS worst fold	HGBR worst fold
K=4 regional holdout	0.7775	0.7775	0.8009	0.8016	0.9784
K=5 regional holdout	0.7480	0.7480	0.9192	0.7951	1.2715
K=6 regional holdout	0.7524	0.7824	0.8089	0.7981	1.2074
Projected 50-km grid	0.7383	0.7384	0.5853	0.8205	0.7092

Note: Three-fold parameter selection is repeated within the outer training data for K=4, K=5, and K=6 regional holdout and projected 50-km grid validation. "Pooled" RMSE is calculated from all outer-test predictions; "worst fold" is the maximum RMSE across outer test folds. Errors are measured in units of USD 100,000.

4.4. Preprocessing and Subgroup Risk

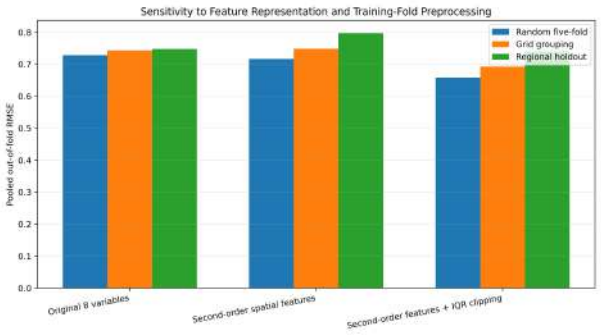


Fig.4. Sensitivity of OLS to Feature Representation and Training-Fold Preprocessing

The main specification uses the eight original explanatory variables. After adding centered latitude-squared, centered longitude-squared, centered latitude–longitude interaction, and a logarithmic transformation of Population, OLS RMSE decreases from 0.7284 to 0.7163 under random five-fold validation, but increases from 0.7430 to 0.7488 under grid grouping and from 0.7480 to 0.7973 under regional holdout. Feature expansion therefore has different effects across spatial validation conditions, and preprocessing should be matched to the intended prediction scenario.

Holding the second-order spatial feature representation constant, training-fold IQR clipping reduces OLS pooled out-of-fold RMSE under random five-fold, grid grouping, and regional holdout from 0.7163, 0.7488, and 0.7973 to 0.6579, 0.6926, and 0.7395, respectively. Under K=5 regional holdout, RMSE falls by approximately 7.25% relative to the unclipped second-order specification. Relative to the original eight-variable OLS benchmark of 0.7480, the final specification improves RMSE by approximately 1.14%. The former comparison isolates the clipping effect under a fixed feature representation, whereas the latter measures the net change relative to the original benchmark.

The reduction in overall error is not distributed uniformly across observations. Among the 3,677 observations for which at least one explanatory variable touches a clipping boundary, RMSE increases from 0.8614 to 0.8835, or by 2.57%. Among the remaining 16,963 observations, RMSE decreases from 0.7827 to 0.7045, or by 10.00%. For the high-end target-value group, RMSE falls from 1.8819 to 1.7191, but mean prediction bias changes from −1.4041 to −1.5717, indicating a larger degree of underprediction. Overall error, subgroup error, and signed bias therefore provide complementary evaluation information.

Table 9 classifies corresponding test observations using IQR bounds estimated from each outer training fold and then aggregates out-of-fold errors. The high-end target-value group overlaps with the group of observations touching explanatory-variable clipping boundaries; the two groups should not be treated as independent. These results support evaluating subgroup diagnostics alongside overall accuracy, within the current feature representation, clipping thresholds, and test sample.

Table 8. Sensitivity of OLS to Feature Representation and Preprocessing (Pooled Out-of-Fold RMSE)

Specification	**Random five-fold**	Grid grouping	**Leave-one-region-out**
Original eight variables	0.7284	0.7430	0.7480
Second-order spatial features	0.7163	0.7488	0.7973
Second-order spatial features + training-fold IQR clipping	0.6579	0.6926	0.7395

Note: All transformation parameters are estimated using only the current outer training fold. The three specifications use identical outer partitions.

Table 9. Subgroup Errors before and after IQR Clipping for OLS under Regional Holdout

Sample group	**Number of observations**	**RMSE before clipping**	**RMSE after IQR clipping**	**Change in RMSE**	**Original prediction bias**	**Prediction bias after IQR clipping**
All observations	20,640	0.7973	0.7395	-7.25%	-0.1664	-0.1596
At least one variable clipped	3,677	0.8614	0.8835	+2.57%	-0.1427	-0.3744
No variables clipped	16,963	0.7827	0.7045	-10.00%	-0.1715	-0.1131
High-end target-value group	992	1.8819	1.7191	-8.65%	-1.4041	-1.5717
Other observations	19,648	0.6993	0.6522	-6.74%	-0.1039	-0.0883

Note: Groups are determined in each test fold using IQR bounds estimated from the corresponding training set. Bias = predicted value − observed value; negative values indicate underprediction. The high-end target-value group overlaps with the group touching explanatory-variable clipping boundaries. RMSE is measured in units of USD 100,000.

4.5. Common Prediction Risks in the High-End Target-Value Group

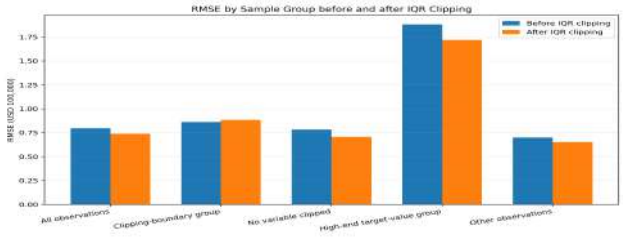


Figure 5. RMSE Differences between the High-End Target-Value Group and Other Observations

The 992 observations with MedHouseVal ≥ 5.0 account for 4.81% of the full sample. All models have larger prediction errors in this group than in the remaining observations. Under random five-fold validation, HGBR RMSE is 0.4256 for the remaining observations and 0.9032 for the high-end group. Under regional holdout, the corresponding values are 0.8364 and 1.9303, with a mean prediction bias of −1.6186 for the high-end group. Ridge has an RMSE of 1.7758 and a mean bias of −1.2203 for this group under regional holdout.

These results indicate that aggregate metrics may conceal larger errors and an underprediction tendency in the high-end target-value group. Because the target variable has a coding ceiling, the analysis evaluates prediction errors against recorded target values only; it cannot recover uncensored market values or interpret the group as necessarily representing genuinely high-value housing in the market.

The high-end target-value results from the main specification with eight original explanatory variables should be distinguished from the paired experiment using

second-order spatial features and training-fold IQR clipping. The two specifications should not be directly substituted or combined. In the paired experiment, IQR clipping reduces RMSE for the high-end group but changes mean prediction bias from −1.4041 to −1.5717, indicating greater underprediction. Model evaluation should therefore report both overall RMSE and errors and signed bias for the high-end group.

This section defines the group using the observed target value MedHouseVal ≥ 5.0 and is therefore an ex post error analysis. In actual valuation, the true target value is not yet available and cannot be used to identify cases for review. Pre-prediction screening must instead rely on available housing attributes, model predictions, model disagreement, and training-sample coverage. The review simulation in Section 4.6 and the subgroup analysis based on observed target values therefore represent distinct stages of evaluation.

4.6. Identification of Severe Errors under a Fixed Review Budget

Using the outer out-of-fold predictions from HGBR with the eight original explanatory variables, the study compares offline risk-screening performance under review budgets of 5%, 10%, and 20%. At a 10% budget, the capture rates of the ex post top 10% of absolute errors are approximately 10.1%, 10.1%, and 10.0% for random review under random five-fold, 0.5° grid grouping, and K=5 regional holdout, respectively. Model-disagreement prioritization achieves 26.4%, 16.6%, and 26.3%, while the prespecified composite risk score achieves 14.6%, 14.8%, and 27.0%.

Under the historical data and outer-evaluation conditions examined here, pre-prediction signals such as model disagreement identify a greater concentration of large prediction errors than random review. However, the performance of individual signals varies with spatial validation conditions, and the composite score does not dominate in every scenario. Model disagreement should therefore not be interpreted as a universal review rule for all housing valuation tasks.

Table 10. Capture of Severe Errors and Remaining Automated-Valuation Error under a 10% Review Budget

Validation design	Screening strategy	High-error capture rate	RMSE of unreviewed observations
Random five-fold	Random review	10.1%	0.4602
Random five-fold	Model disagreement	26.4%	0.4258
Random five-fold	Prespecified composite score	14.6%	0.4459
0.5° grid	Random review	10.1%	0.5977
0.5° grid	Model disagreement	16.6%	0.5842
0.5° grid	Prespecified composite score	14.8%	0.5882
Regional holdout (K=5)	Random review	10.0%	0.9191
Regional holdout (K=5)	Model disagreement	26.3%	0.8652
Regional holdout (K=5)	Prespecified composite score	27.0%	0.8635

Note: Based on HGBR out-of-fold predictions. Severe errors are defined as the 10% of observations with the largest absolute errors in each validation design. Review selection does not use test labels; true labels are used only for ex post evaluation. Random review is repeated 100 times and averaged. “Unreviewed RMSE” applies only to the subset retained for automated valuation and does not represent realized gains from human correction.

The “unreviewed RMSE” in Table 10 refers only to the error among observations that remain subject to automated valuation after risk ranking. It does not mean that selected observations have been corrected by human review or that the original model’s overall predictive performance has improved. Random-review results are averages over 100 repeated samples. Capture rates of approximately 10% are descriptive benchmarks for the specified sample and sampling rule, not confidence intervals for generalization to other datasets.

Table 11. Risk Identification under Different Manual-Review Budgets (K=5 Regions)

Review budget	Screening strategy	High-error capture rate	RMSE of unreviewed observations
5%	Random review	5.0%	0.9190
5%	Model disagreement	15.1%	0.8902
5%	Prespecified composite score	16.9%	0.8850
10%	Random review	10.0%	0.9191
10%	Model disagreement	26.3%	0.8652
10%	Prespecified composite score	27.0%	0.8635
20%	Random review	20.0%	0.9193
20%	Model disagreement	45.0%	0.8185
20%	Prespecified composite score	45.3%	0.8216

Note: Based on HGBR out-of-fold predictions under K=5 regional holdout. Severe errors are defined as the 10% of absolute errors among all outer-test observations in the validation design. Review budgets of 5%, 10%, and 20% are allocated separately within each outer test fold. Random review is repeated 100 times and averaged. Unreviewed RMSE is not the overall RMSE after human review.

Table 11 further compares the effect of different review budgets under K=5 regional holdout. Model-disagreement prioritization captures 15.1%, 26.3%, and 45.0% of ex post high-error observations at budgets of 5%, 10%, and 20%, respectively, exceeding the corresponding random-review benchmarks. This supports the feasibility of prioritizing cases using pre-prediction risk signals under a limited review budget, but does not establish improved human correction accuracy, cost savings, or operating returns.

Table 12. Robustness of Risk Identification under Different Numbers of Regions at a 10% Review Budget

Number of regions	Screening strategy	High-error capture rate	RMSE of unreviewed observations	Original overall RMSE
K=4	Random review	10.0%	0.8006	0.8009
K=4	Model disagreement	20.1%	0.7756	0.8009
K=4	Prespecified composite score	12.8%	0.7922	0.8009
K=5	Random review	10.1%	0.9189	0.9192
K=5	Model disagreement	26.3%	0.8652	0.9192
K=5	Prespecified composite score	27.0%	0.8635	0.9192
K=6	Random review	9.9%	0.8091	0.8089
K=6	Model disagreement	25.4%	0.7724	0.8089
K=6	Prespecified composite score	23.7%	0.7739	0.8089

Note: Based on HGBR and Ridge out-of-fold predictions after renewed nested tuning for K=4, K=5, and K=6. The review budget is 10% within each fold. The severe-error sets may differ across the numbers of regions. Random-review results are averages over repeated simulations; other methods use prespecified risk rankings.

Table 12 examines risk-screening results as the number of regions changes. Under the K=4, K=5, and K=6 renewed nested-tuning experiments, model-disagreement prioritization captures 20.1%, 26.3%, and 25.4% of high-error observations at a 10% budget, respectively, exceeding the corresponding random-review benchmarks. The composite score captures 12.8%, 27.0%, and 23.7%, indicating sensitivity to the regional partition. Because the high-error sets are not identical across regional designs and all experiments use the same historical dataset, the results support offline risk identification under the specified conditions rather than external validity in other housing markets.

4.7. Predictive Performance and Capacity-Constrained Review in the Guangzhou Housing Sample

4.7.1. Out-of-Sample Performance of Valuation Models

Table 13 shows that the pooled out-of-fold RMSE values of the three linear models are similar: 117.91, 117.95, and 117.94 CNY 10,000 for OLS, Ridge, and Lasso, respectively. HGBR has an RMSE of 121.64 CNY 10,000 but a lower MAE of 84.95 CNY 10,000 than the linear models. This indicates that RMSE, which is more sensitive to large errors, and MAE can produce different rankings; model comparisons should therefore be aligned with the relevant loss objective.

The inner-selection procedure selects Lasso in all 25 outer training–test combinations. The subsequent review-rule comparison therefore uses the common out-of-fold predictions from Lasso rather than selecting a model ex post based on final test errors. Its mean coefficient of determination is approximately 0.36, close to those of OLS and Ridge, reflecting the predictive performance supported by the available fields. The model is not changed after observing final test results.

Output diagnostics identify the same 10 housing records as having negative price predictions in all five repetitions. These records account for approximately 0.55% of the sample in each repetition, and the minimum prediction is −CNY 1.1708 million. Tables 13–15 retain these raw predictions as a common offline benchmark and further examine their contribution to relative-error screening. The results therefore cover both aggregate prediction error and the validity of individual outputs; negative predictions are not treated as usable valuations.

Table 13. Repeated Nested-Validation Performance in the Guangzhou Sample

Model	RMSE (CNY 10,000)	MAE (CNY 10,000)	R^2
OLS	117.91±0.52	86.98	0.3602
Ridge	117.95±0.55	86.97	0.3597
Lasso	117.94±0.49	86.87	0.3599
HGBR	121.64±1.14	84.95	0.3190

Note: RMSE is reported as the mean pooled out-of-fold RMSE across five repetitions ± the standard deviation across repetitions. MAE and R^2 are the means of the corresponding metrics across repetitions. Prices and errors are measured in CNY 10,000. The repetitions are not independent datasets, and the standard deviation is not a confidence interval for population-level generalization.

4.7.2. Comparison of Review Rules under a Common Budget

Under the same automated predictions and 180 available review slots, the rules differ in their coverage of high absolute errors (Table 14). The mean capture rate is 9.86% for random review, 23.83% for model disagreement, and 38.14% for training-only error learning. These represent increases of 13.97 and 28.28 percentage points over random review, or approximately 2.42 and 3.87 times the random-review rate. Predicted-price prioritization reaches 41.31%, an increase of 31.45 percentage points over random review and 3.17 percentage points above error learning. These comparisons indicate that pre-prediction information can improve the coverage of high absolute errors within a fixed number of review slots, while also showing why simple rules should remain necessary benchmarks for more complex risk-learning methods.

Before screening, all rules use the same predictions, with a pooled out-of-fold RMSE of 117.94 CNY 10,000. The unreviewed RMSE values for random review, model disagreement, training-only error learning, and predicted-price prioritization are 117.88, 106.24, 100.19, and 98.57 CNY 10,000, respectively. These values describe the error composition of the remaining automated-valuation subset; they do not indicate that the original model has improved or that reviewed observations have been corrected. The differences are descriptive across five repetitions and are not presented as statistically significant.

Table 14. Offline Screening Results under a Nominal 10% Review Budget in the Guangzhou Sample

Review rule	High-absolute-error capture rate (%)	Increase over random review (percentage points)	Review-list hit rate (%)	Unreviewed RMSE (CNY 10,000)
Random review	9.86±0.03	0.00	10.02±0.03	117.88±0.51
Predicted-price prioritization	41.31±1.32	31.45	42.00±1.34	98.57±0.73
Covariate-exceedance prioritization	16.94±2.01	7.08	17.22±2.04	108.94±0.83
Model-disagreement prioritization	23.83±2.21	13.97	24.22±2.24	106.24±0.76
Training-only error learning	38.14±1.95	28.28	38.78±1.98	100.19±0.98

Note: Capture rates, hit rates, and unreviewed RMSE are reported as means ± sample standard deviations across five repetitions. The percentage-point difference is the capture-rate mean minus the random-review mean; it is not a confidence interval or significance test. Each repetition selects 180 observations, while the high-absolute-error set contains 183 observations. Random review is averaged over 1,000 simulations within each repetition before aggregation across repetitions; its reported standard deviation therefore does not represent single-simulation variability. All rules use the same raw predictions, including negative price predictions. Unreviewed RMSE evaluates only the retained subset and does not represent human-correction benefits.

As the review budget increases, the proportion of high-absolute-error observations captured also rises. At nominal budgets of 5%, 10%, and 20%, model disagreement captures 13.55%, 23.83%, and 40.00%, respectively; training-only error learning captures 22.19%, 38.14%, and 61.31%; and predicted-price prioritization captures 23.28%, 41.31%, and 61.86%. At a 20% budget, the mean capture rates of error learning and price prioritization differ by only 0.55 percentage points, compared with 3.17 percentage points at a 10% budget. This indicates that budget size should also be considered when comparing rules. Similar means do not establish statistical equivalence or interchangeability.

4.7.3. Changes in Rankings under Alternative Error Measures

When the same model predictions and review lists under the nominal 10% budget are retained but high-risk observations are redefined using relative error, the mean capture rates change (Table 15). The high-relative-error capture rates are 12.46% for predicted-price prioritization, 17.60% for model disagreement, and 13.88% for training-only error learning. These values describe coverage under the original predictions, including negative price predictions; they do not establish that model disagreement retains the same incremental advantage after a basic output-validity screen. Training-only error learning is trained on absolute residuals, so its relative-error performance should not be interpreted as optimization for proportional error.

Table 15. Sensitivity to Relative-Error Evaluation with Fixed Review Lists in the Guangzhou Sample

Review rule	High-relative-error capture rate (%)	Unreviewed MAPE (%)
Random review	9.85±0.05	50.83±0.26
Predicted-price prioritization	12.46±1.52	50.62±0.18
Covariate-exceedance prioritization	12.90±2.98	49.62±0.64
Model-disagreement prioritization	17.60±1.79	48.78±0.28
Training-only error learning	13.88±2.14	50.30±0.27

Note: Values are means ± standard deviations across five repeated validations. The high-relative-error set comprises the 183 observations with the largest absolute percentage errors in each repetition and need not match the high-absolute-error set in

Table 14. Model predictions and review lists remain unchanged and include negative price predictions. MAPE denotes mean absolute percentage error; all recorded prices are positive. Random-review aggregation follows the same procedure as in Table 14. Across-repetition standard deviations describe sensitivity to data partitioning and are not confidence intervals for independent market validation.

To distinguish the identification of abnormal outputs from the identification of general relative-error cases, the composition of captured observations is examined while retaining the original set of 183 high-relative-error cases and the original review lists. Model disagreement captures an average of 32.2 observations, including 9.8 negative-price predictions and 22.4 nonnegative predictions. Predicted-price prioritization captures an average of 22.8 observations, all with nonnegative predictions. The number of nonnegative high-relative-error observations captured by the two rules is similar; thus, the overall capture difference between model disagreement and price prioritization mainly reflects the identification of negative predictions.

This diagnostic suggests that price-validity checks should be distinguished from subsequent error ranking. A negative price is an identifiable abnormal output, whereas a nonnegative prediction is not necessarily an accurate valuation. The composition analysis explains the source of the capture difference. Whether abnormal-output screening followed by risk ranking provides incremental identification under the same total budget requires a paired comparison using the same test observations and review capacity.

V. MODEL RISK GOVERNANCE AND CAPACITY-CONSTRAINED REVIEW ALLOCATION

The California and Guangzhou experiments provide evidence on scenario suitability, subgroup diagnostics, and review prioritization, respectively. Based on these findings, Table 16 organizes the use of housing valuation models into five stages: defining the application scenario, evaluating candidate models, identifying prediction risks, allocating review slots, and incorporating post-review feedback. The table specifies the information inputs and management outputs at each stage. The process connects existing evaluation results with deployment decisions; specific thresholds, review costs, and overall business outcomes should be examined using data from the target organization.

Table 16. A Risk-Diagnostic and Review Process for Housing Valuation Based on the Empirical Findings

Decision stage	Primary basis	Decision output	Management requirement
Application-scenario definition	Appraisal object, valuation date, target region, training-sample coverage	Select validation design	Validation conditions should match the intended application scenario
Candidate-model evaluation	Pooled out-of-fold RMSE, worst-region RMSE, minimum Skill, regional bias	Select candidate and benchmark models	Do not rely solely on average error from random cross-validation
Prediction-risk identification	Fifth-nearest-neighbor distance, local sample sparsity, covariate extrapolation, model disagreement	Generate a pre-prediction risk ranking	Do not use unavailable true target values
Allocation of limited review slots	Pre-prediction risk ranking and prespecified review budget	Determine the proposed review list	Prioritize potentially high-error cases within the fixed budget
Post-review feedback and model updating	Actual valuation feedback, changes in regional errors, model bias	Determine whether to continue use, retrain, or revert to a fallback model	Trigger conditions should be validated using target-market data and business losses

Note: The table organizes existing methods and selected findings from this study into a procedural framework. The California evidence concerns spatial evaluation, whereas the Guangzhou evidence concerns non-spatial review comparisons. Specific triggers, budget allocations, and the performance of the complete process require validation in the target business setting.

5.1. Model Selection According to the Application Scenario

Model selection should begin by specifying what is to be predicted, where the prediction will be made, and when the model will be used. If target observations are spatially mixed with training observations, the task may primarily involve interpolation. If targets are located in new areas with few training observations, the model faces greater spatial extrapolation risk. Because model rankings can change across prediction conditions, random cross-validation should not be used to establish a single model ranking for all applications.

The results show that HGBR has lower pooled out-of-fold RMSE under random five-fold and local spatial-grid grouping. However, after renewed tuning under K=4, K=5, and K=6 complete regional holdout, both its pooled RMSE and worst-region RMSE exceed those of OLS. This result is limited to the historical data, spatial partitions, and parameter search examined here and does not imply that linear models are universally superior to nonlinear models in cross-regional housing valuation.

For tasks with adequate training coverage and targets that are spatially mixed with training observations, HGBR may be considered as a candidate model, with OLS or Ridge serving as stability benchmarks. For local spatial-block transfer, overall RMSE should be accompanied by grid-level errors, training-sample density, and fold variation. For prediction in entirely new regions, priority should be given to worst-region RMSE, regional prediction bias, minimum Skill, and fifth-nearest-neighbor distance. Without external validation in the target market, historical average accuracy alone should not determine whether a model is deployed.

Before deployment, model reports should include pooled out-of-fold RMSE, MAE, and R^2; mean and standard deviation of fold-level RMSE; worst-fold RMSE; minimum Skill; regional bias; errors for the high-end target-value group; and training-sample coverage. Complex models may serve as candidate models, while simpler models provide stability benchmarks, abnormal-output checks, and fallback options. When a candidate model's error, bias, or data drift in a target region exceeds prespecified thresholds, local retraining, manual review, fallback to a benchmark model, or suspension of automated valuation should be considered.

5.2. Incorporating Validation Protocols into Model Governance

First, define the appraisal object and intended use before modeling. Specify whether the target is an individual dwelling, building, community, census block group, or administrative area; whether the target variable is a transaction price, appraised value, rent, collateral value, or another measure; the valuation date; and the information available at prediction time. Different appraisal purposes should not automatically share the same evaluation protocol.

Second, design validation according to the intended deployment scope. Random cross-validation may serve as a reference for interpolation within the same market. For prediction in new neighborhoods, across regions, or across cities, spatial grouping, complete regional holdout, and temporal holdout are more closely aligned with the intended use. Random five-fold results from historical U.S. data alone are insufficient to justify deployment in China.

Third, place all learnable data-processing steps within the training pipeline. Standardization parameters, outlier thresholds, logarithmic transformations, feature-selection rules, and hyperparameter searches should be estimated within the corresponding training fold. Fold-specific parameters, code versions, data checksums, and computing environments should be retained so that the outer test set remains representative of unseen data.

Fourth, extend evaluation from average accuracy to risk distribution and review triggers. Reports should include overall RMSE, fold variation, worst-region error, minimum Skill, regional prediction bias, ex post errors for the high-end target-value group, and training-sample coverage. Ex post error analysis must be distinguished from pre-prediction screening: whether a true target value exceeds the high-end threshold can be used only after the target is observed and cannot directly identify cases for review before valuation. For cases whose true values are unknown, screening should rely on information available before prediction review, including fifth-nearest-neighbor distance, local training-sample sparsity, covariates outside the training distribution, and disagreement between models. When such signals meet prespecified and validated trigger conditions, cases may be added to the manual-review list or automated valuation may be suspended. The thresholds in this study have not been validated using actual business data.

Fifth, establish challenger-model and fallback mechanisms. Complex models may offer potential accuracy gains, while simpler models provide stability benchmarks and fallback options. When a candidate model's errors or bias in a target region substantially exceed those of a challenger, training coverage, covariate distributions, target-variable coding, and parameter selection should be examined rather than attributing the difference solely to model complexity.

5.3. Conditions for Application to Housing Valuation in China

The Guangzhou supplementary analysis provides Chinese-sample evidence on review prioritization under a

fixed number of review slots and shows that rankings differ across monetary-error, percentage-error, and abnormal-output criteria. Applying these findings to a specific business setting requires an evaluation procedure aligned with the appraisal object and the information actually available.

First, align the observational unit and target variable. Individual dwellings, buildings, communities, and census block groups differ in attribute meanings and price definitions. The appraisal purpose should determine the target object, currency unit, and valuation date, with training and evaluation organized accordingly.

Second, construct spatial and temporal validation according to available information. The California data support spatial holdout because geographic coordinates are available, whereas the public Guangzhou file supports attribute modeling and non-spatial screening. Deployment evaluation in a specific Chinese city should use local data with reliable geographic identifiers and transaction dates to test performance across regions and periods.

Third, incorporate the loss objective into review-rule comparisons. Priorities based on monetary deviations may differ from those based on percentage deviations. Under the same number of review slots, random review, simple price-based rules, and model-risk signals should be compared, retaining simple benchmarks with clear operational interpretations.

Fourth, record output-validity checks separately from risk ranking. Outputs that are inconsistent with the meaning of prices should be flagged, while nonnegative predictions should still be evaluated for accuracy. If abnormal cases are included in the review list, their slots should count toward the same total budget to ensure comparability.

Fifth, use business feedback to refine resource allocation. Offline capture rates and errors among retained observations can provide an initial comparison of ranking rules. Selecting review proportions and trigger thresholds requires additional information on human correction outcomes, per-case review costs, delay costs, and business losses.

VI. DISCUSSION AND RESEARCH LIMITATIONS

6.1. Managerial Implications and Methodological Contributions

The study links housing valuation model performance to conditions of use and provides empirical evidence on three types of differences: model rankings vary with spatial validation conditions; improvements in overall error do not necessarily coincide with improvements for all sample groups; and review-rule comparisons under fixed capacity should consider both the loss objective and abnormal outputs. These findings add decision-relevant information beyond average-accuracy comparisons and inform the assessment of model applicability, preprocessing, and review priorities.

First, model applicability should be assessed in relation to spatial validation conditions. HGBR has lower average error under random five-fold and local spatial-grid grouping, but its pooled error and worst-region error exceed those of OLS under complete regional holdout. The evidence directly shows that model rankings change with validation conditions; it does not establish that overall error and worst-region error necessarily produce opposite rankings within the same condition. Regional metrics provide additional information about adverse-region error magnitude and performance relative to a simple benchmark, preventing model selection from relying solely on average rankings under spatially mixed samples.

Second, improvements in overall error may coincide with a redistribution of subgroup risk. Second-order spatial features and training-fold IQR clipping reduce overall OLS RMSE, but RMSE increases for observations touching clipping boundaries and underprediction bias expands for the high-end target-value group. Preprocessing should therefore be assessed using subgroup errors and signed bias in addition to overall RMSE. The subgroup risks examined here are statistical prediction risks and are not equivalent to social-equity assessments or evidence of realized losses among social groups.

Third, connecting prediction evaluation with capacity-constrained review requires consideration of both the loss objective and output validity. In the Guangzhou sample, training-only error learning has a higher mean high-absolute-error capture rate than random review and model disagreement, but it does not outperform predicted-price prioritization, a simple benchmark. Under relative-error evaluation, the higher capture rate of model disagreement is largely associated with the coverage of negative predictions and other abnormal outputs. It should not be interpreted as a stable advantage for general relative valuation risk. Practical review procedures should therefore define output-validity checks before comparing risk-ranking rules under the same budget.

From the perspective of housing valuation in China, the study adds local model refitting and non-spatial review screening using a publicly available Guangzhou sample. However, this supplementary analysis does not test cross-city or cross-period transfer and does not estimate actual human correction accuracy or review costs. The evidence should be interpreted as a conditional methodological test on a Chinese sample, not as a nationwide market regularity

or realized performance of a complete governance mechanism.

6.2. Scope and Directions for Further Research

The findings depend first on the information available in the datasets. The California data are 1990 census block-group aggregates, and target-value ceiling coding and aggregation limit interpretation at the individual-property level. The Guangzhou sample is described by its provider as covering a particular area and period, but the public file does not contain transaction-level dates, geographic identifiers, property age, or original transaction documents that would permit independent verification. The two datasets are therefore used separately for spatial-stability assessment and locally refitted non-spatial review screening. Future research could extend the analysis to actual communities, cities, and temporal boundaries using more complete property and transaction information.

The scope of model comparisons is also determined by the partitioning rules and parameter candidates. Latitude–longitude grids, projected grids, and K-means regions are experimental partitions; regional construction uses unlabeled coordinates from the full sample, and grid validation does not include a spatial buffer. Grouping Guangzhou observations by attribute combinations is not equivalent to grouping by property or community. The designs change training composition and sample coverage simultaneously, while repeated splits describe sensitivity within the same dataset. Future work could expand candidate models under a common computational budget and test ranking stability using prespecified real-world market boundaries.

The review results concern offline risk identification. High-error sets are defined ex post using true target values, and errors among unreviewed observations describe the risk composition after screening. Negative price predictions in the Guangzhou sample and the use of absolute residuals as the error-learning target also affect relative-error comparisons. Further research should evaluate an explicit output-validity screen under a fixed total budget and combine actual human corrections, review and delay costs, and business loss functions to assess the incremental value of risk ranking and the performance of the complete process.

VII. CONCLUSION

The study uses 20,640 California census block-group observations and 1,824 publicly available Guangzhou housing records to examine spatial validation, preprocessing-related subgroup risk, and capacity-constrained review ranking. The findings are summarized as follows.

First, spatial partitioning changes training coverage and covariate distributions. The proportions of test observations with no training observation within 10 km are 1.73%, 51.95%, and 97.89% under random five-fold, 0.5° grid grouping, and K=5 regional holdout, respectively. Evaluation conditions should therefore match the intended model use, and training distance should be interpreted together with coverage information.

Second, model advantages depend on the prediction scenario. HGBR has lower average error under local grid validation but higher pooled out-of-fold RMSE than OLS under K=4, K=5, and K=6 complete regional holdout. Model selection should consider both overall error and adverse regional performance rather than relying on the average ranking from a single validation design.

Third, preprocessing should be assessed using both overall and subgroup results. Under K=5 regional holdout, IQR clipping with fixed second-order spatial features reduces overall RMSE by approximately 7.25%, but increases RMSE by 2.57% for observations touching clipping boundaries and expands underprediction bias for the high-end target-value group. An improvement in overall error does not imply that all groups benefit.

Fourth, pre-prediction risk signals can help prioritize limited review slots. Across the examined California regional partitions, model-disagreement screening has a higher high-error capture rate than the corresponding random-review benchmark under a 10% budget, providing offline evidence for prioritization under the specified conditions.

Fifth, the Guangzhou comparison shows that simple benchmarks, loss objectives, and output validity should all be considered in review evaluation. Under a nominal 10% budget, the mean high-absolute-error capture rates of model disagreement, training-only error learning, and predicted-price prioritization are 23.83%, 38.14%, and 41.31%, respectively, compared with 9.86% for random review. Rankings change under relative-error evaluation, and the difference between model disagreement and price prioritization mainly involves negative price predictions.

Overall, the findings support incorporating scenario matching, subgroup diagnostics, and loss-oriented review ranking into housing valuation evaluation. The two datasets provide evidence for spatial assessment and non-spatial screening, respectively. The managerial implications require further validation using target-market spatial and temporal information and actual review feedback.

VIII. DATA AND CODE AVAILABILITY

The California analysis is based on the publicly available California Housing dataset, organized as California-Housing-Dataset.csv. The SHA-256 checksum of the data file is 7192af0bf33e94141194cfd91958c62d66262dc40c07cc2e6fbe0b2d02b0e1b3. The data are 1990 U.S. Census block-group aggregates rather than a modern individual-property transaction database.

To support replication, the study retains the original data, variable-validation records, out-of-fold predictions, outer-fold hyperparameters, inner-search results, spatial-coverage diagnostics, covariate-distribution differences, IQR group details, and risk scores, screening records, and evaluation metrics from the fixed-budget experiments.

The original California experiment environment and data-validation records are retained in the main-experiment archive. The Guangzhou supplementary analysis uses Sheet1 of the second-version Figshare workbook. The SHA-256 checksum of "Second-hand Housing Price Data.xlsx" is 8be62623ddbb0f3dd646d4ea444f93600ccb3a2fd51d47f034f74250a66e8c3e. The original authors' code is stored separately from the supplementary code used in this study; predictions displayed in the original authors' workbook are not substituted for independently recalculated results.

The Guangzhou supplementary experiment was run using Python 3.13.5, NumPy 2.3.5, SciPy 1.17.0, and scikit-learn 1.8.0. Data audits, feature processing, inner-search results for all candidate configurations, selected models in each fold, observation-level out-of-fold predictions, risk scores, review lists, budget-sensitivity results, and relative-error results are retained and provided as supplementary materials. The code restricts model selection and risk learning to outer training data and includes independent metric recalculation and checks that identical attribute groups do not cross folds. The Guangzhou experiment does not include spatial or temporal extrapolation validation and contains no corrected values from actual human review.

IX. DECLARATIONS

The author declares no known competing interests. No specific research funding is reported. The author takes responsibility for the research design, data analysis, manuscript content, and final version.

REFERENCES

[1] DEPPNER J, CAJIAS M. Accounting for spatial autocorrelation in algorithm-driven hedonic models: A spatial cross-validation approach[J]. The Journal of Real Estate Finance and Economics, 2024, 68(2): 235–273.

[2] ROBERTS D R, BAHN V, CIUTI S, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure[J]. Ecography, 2017, 40(8): 913–929.

[3] VALAVI R, ELITH J, LAHOZ-MONFORT J J, et al. blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models[J]. Methods in Ecology and Evolution, 2019, 10(2): 225–232.

[4] WANG Y, KHODADADZADEH M, ZURITA-MILLA R. Spatial+: A new cross-validation method to evaluate geospatial machine learning models[J]. International Journal of Applied Earth Observation and Geoinformation, 2023, 121: 103364.

[5] STOCK A. Choosing blocks for spatial cross-validation: Lessons from a marine remote sensing case study[J]. Frontiers in Remote Sensing, 2025, 6: 1531097.

[6] VARMA S, SIMON R. Bias in error estimation when using cross-validation for model selection[J]. BMC Bioinformatics, 2006, 7: 91.

[7] CAWLEY G C, TALBOT N L C. On over-fitting in model selection and subsequent selection bias in performance evaluation[J]. Journal of Machine Learning Research, 2010, 11: 2079–2107.

[8] KAPOOR S, NARAYANAN A. Leakage and the reproducibility crisis in machine-learning-based science[J]. Patterns, 2023, 4(9): 100804.

[9] ZHANG J J, GUO H Y. A review and prospect of data-driven research on real estate mass appraisal[J]. Journal of Xihua University (Philosophy & Social Sciences), 2024, 43(3): 13–27. (In Chinese)

[10] QU W D, GU H Y. Applications of artificial intelligence in real estate appraisal: A review and future prospects[J]. China Real Estate Finance, 2026(3): 57–66. (In Chinese)

[11] ROSEN S. Hedonic prices and implicit markets: Product differentiation in pure competition[J]. Journal of Political Economy, 1974, 82(1): 34–55.

[12] WANG Z, WANG Y, WU S, et al. House price valuation model based on geographically neural network weighted regression: The case study of Shenzhen, China[J]. ISPRS International Journal of Geo-Information, 2022, 11(8): 450.

[13] GYGER T, HAURI S, BÜHLMANN S, et al. Explainable spatial machine learning for hedonic real estate modeling[J]. Real Estate Economics, 2026, 54: 761–806.

[14] HOERL A E, KENNARD R W. Ridge regression: Biased estimation for nonorthogonal problems[J]. Technometrics, 1970, 12(1): 55–67.

[15] TIBSHIRANI R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society: Series B, 1996, 58(1): 267–288.

[16] FRIEDMAN J H. Greedy function approximation: A gradient boosting machine[J]. The Annals of Statistics, 2001, 29(5): 1189–1232.

[17] MULLAINATHAN S, SPIESS J. Machine learning: An applied econometric approach[J]. Journal of Economic Perspectives, 2017, 31(2): 87–106.

[18] PACE R K, BARRY R. Sparse spatial autoregressions[J]. Statistics & Probability Letters, 1997, 33(3): 291–297.
[19] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: Machine learning in Python[J]. Journal of Machine Learning Research, 2011, 12: 2825–2830.
[20] FRIEDMAN J, HASTIE T, TIBSHIRANI R. Regularization paths for generalized linear models via coordinate descent[J]. Journal of Statistical Software, 2010, 33(1): 1–22.
[21] SCIKIT-LEARN DEVELOPERS. California housing dataset[EB/OL]. Scikit-learn documentation, [2026-11-02]. https://scikit-learn.org/stable/datasets/real_world.html.
[22] MINISTRY OF HOUSING AND URBAN-RURAL DEVELOPMENT OF THE PEOPLE'S REPUBLIC OF CHINA. Real estate appraisal standard: GB/T 50291—2015[S]. Beijing: China Architecture & Building Press, 2015. (In Chinese)
[23] WANG J, JI H, WANG L. Forecasting second-hand house prices in China using the GA-PSO-BP neural network model[J]. PLOS ONE, 2025, 20(5): e0322821.
[24] JI H. Forecasting for second-hand house prices in China using the GA-PSO-BP neural-network model: Excel dataset and Python code[DS/OL]. Version 2. Figshare, 2025[2026-11-03]. DOI: 10.6084/m9.figshare.28171337.v2.
[25] GEIFMAN Y, EL-YANIV R. SelectiveNet: A deep neural network with an integrated reject option[C]//Proceedings of the 36th International Conference on Machine Learning. PMLR, 2019, 97: 2151–2159.
[26] ZAOUI A, DENIS C, HEBIRI M. Regression with reject option and application to kNN[C]//Advances in Neural Information Processing Systems. 2020, 33.