Proceedings of CONF-APMM 2026 Symposium: Applied Mathematics and Quantitative Modeling
DOI: 10.54254/2753-8818/2026.PU37987

Frequentist Versus Bayesian Hedonic Regression for Miami House Prices: Prediction Intervals, Coverage Calibration, and Prior Sensitivity

Hongyi Huang

School of Statistics, University of Minnesota, Minneapolis, USA
huan2748@umn.edu

Abstract. Housing price prediction requires both accurate point estimates and a useful description of prediction uncertainty. This study compares frequentist, regularized, Bayesian, and machine-learning approaches using the Miami Housing 2016 dataset. After removing duplicate properties, 13,776 observations were analyzed, with log sale price used as the response variable. Six models were compared: ordinary least squares (OLS), Ridge, Lasso, Bayesian Ridge, Automatic Relevance Determination (ARD), and Random Forest. Random Forest achieved the strongest point-prediction performance, with a test R^2 of 0.9237 and a root mean squared error (RMSE) of 0.1589, while the five linear models produced R^2 values near 0.766. Prediction intervals from OLS, Bayesian Ridge, and ARD showed nearly identical performance with the full training sample, with 95% empirical coverage of approximately 0.939 and similar average widths. Small-sample experiments showed wider and more variable intervals as training size decreased, although Bayesian Ridge produced slightly narrower intervals than OLS with similar coverage. Additional experiments demonstrated the connection between Ridge regularization and Bayesian shrinkage and showed that an extreme prior-strength setting can reduce Bayesian Ridge performance when training data are limited. Overall, the results show that strong point-prediction performance does not necessarily imply better uncertainty estimation, and that prior specification can become more influential when less training data are available.

Keywords: housing valuation, Bayesian regression, prediction intervals, random forest, regularization

1. Introduction

Housing is a differentiated good whose observed price reflects a bundle of structural, locational, and environmental characteristics. The hedonic pricing framework provides a natural basis for modeling this relationship by expressing price as a function of individual property attributes [1]. For prediction, however, an interpretable linear model is only one possible approach. Recent property-valuation research has increasingly applied machine-learning methods to capture complex relationships in detailed housing data [2].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.PU37987

This study compares frequentist, regularized, Bayesian, and machine-learning approaches using the Miami Housing 2016 dataset. The analysis evaluates six models: ordinary least squares (OLS), Ridge, Lasso, Bayesian Ridge, Automatic Relevance Determination (ARD) Regression, and Random Forest. It considers both out-of-sample point-prediction accuracy and prediction uncertainty. Point predictions are evaluated using test-set R^2, root mean squared error (RMSE), and mean absolute error (MAE), while uncertainty is evaluated using empirical interval coverage, average interval width, and calibration across nominal levels.

The study also examines three questions beyond the main model comparison. First, it tests how OLS and Bayesian Ridge prediction intervals change when the training sample is reduced. Second, it evaluates the connection between Ridge regularization and Bayesian shrinkage by deriving a Bayesian-equivalent Ridge penalty and comparing fitted coefficients. Third, it investigates the sensitivity of Bayesian Ridge to prior-strength settings under full and reduced training samples. Together, these analyses distinguish point-prediction performance from uncertainty estimation and clarify when regularization, sample size, and prior specification matter.

Using one fixed test set makes the comparison directly reproducible and separates point accuracy from interval calibration. The reduced-sample and prior experiments then identify when conclusions from the full sample cease to hold.

2. Literature review

Housing valuation is commonly grounded in Rosen's hedonic framework, which treats a differentiated property's price as a function of its characteristics [1]. In housing, this framework links value to size, quality, age, and location and makes linear regression a useful interpretable baseline. More recent valuation research has expanded beyond linear specifications and demonstrated the growing role of machine learning for land and structure valuation [2]. The resulting choice is not simply between an old and a new technique, but between models that emphasize interpretable average relationships and models that can represent nonlinearities and interactions.

Regularization addresses instability when many correlated housing attributes enter a linear model. Ridge shrinks coefficients with an L2 penalty [3], whereas Lasso uses an L1 penalty and may set coefficients exactly to zero [4]. Recent automated-valuation evidence shows that shrinkage can support data-driven model specification [5]. Bayesian regression imposes related control through parameter distributions; under corresponding precision and penalty settings, its fitted coefficients can align closely with Ridge without making the two inferential frameworks identical.

Flexible algorithms can capture relationships that fixed linear terms miss. Automated-valuation research consequently stresses transparent out-of-sample metrics [6]. Random Forest combines randomized decision trees [7]. A recent housing study incorporates spatial dependence [8], and another compares machine-learning algorithms for Dubai property prices [9]. However, a lower point error does not describe uncertainty for an individual property. Prediction intervals must also be judged by empirical coverage and width: useful intervals approach their stated coverage while remaining reasonably sharp [10]. This study therefore compares point accuracy, interval calibration, sample-size effects, and shrinkage sensitivity within the same held-out test design.

DOI: 10.54254/2753-8818/2026.PU37987

3. Methodology

3.1. Data source and preprocessing

The Miami Housing 2016 dataset was obtained from OpenML (data_id=43093) [11, 12]. It contains 13,932 public single-family transactions and 17 numerical variables covering price, location, property size and quality, environmental distances, age, noise exposure, and sale month.

Duplicate PARCELNO records were removed with the first transaction retained, reducing the sample to 13,776 properties and preventing the same parcel from leaking across training and test sets. The right-skewed SALE_PRC response was transformed to ln_price. Fifteen predictors remained after excluding the parcel identifier and price variables.

Data were split 80/20 with random_state=42, yielding 11,020 training and 2,756 test observations. The test set was fixed for every comparison. StandardScaler was fitted only to training data; Ridge, Lasso, Bayesian Ridge, and ARD used standardized predictors, whereas OLS and Random Forest used original scales.

Equation (1) defines the response.

$$y_i = \ln(\mathrm{SALE_PRC}_i) \tag{1}$$

Here y_i is the natural logarithm of $\mathrm{SALE_PRC}_i$, the observed price of property i.

Table 1 defines every variable in the original OpenML dataset, including its interpretation and unit of measurement.

Table 1. Data dictionary for the Miami Housing 2016 dataset

Variable	Meaning	Unit
LATITUDE	Property latitude	degree
LONGITUDE	Property longitude	degree
PARCELNO	Parcel identifier	ID
SALE_PRC	Sale price	USD
LND_SQFOOT	Land area	sq ft
TOT_LVG_AREA	Total living area	sq ft
SPEC_FEAT_VAL	Value of special features	USD
RAIL_DIST	Distance to nearest railway	ft
OCEAN_DIST	Distance to ocean	ft
WATER_DIST	Distance to water	ft
CNTR_DIST	Distance to central business district	ft
SUBCNTR_DI	Distance to subcenter	ft
HWY_DIST	Distance to highway	ft
age	Age of house	years
avno60plus	Excess aircraft noise indicator	0/1
month_sold	Month sold (1–12)	month

DOI: 10.54254/2753-8818/2026.PU37987

Table 1. (continued)

structure_quality	Structure quality score (1–5)	score

Data from: Miami Housing 2016 dataset

3.2. Frequentist models

OLS was the frequentist baseline, using a constant and all 15 predictors on their original scales. Heteroskedasticity-consistent (HC3) standard errors were computed for inference; they alter estimated uncertainty but not coefficients.

Equation (2) gives the hedonic model.

$$y_i = \beta_0 + x_i^{\mathrm{T}}\beta + \varepsilon_i \tag{2}$$

Here β_0 is the intercept, x_i and β are the predictor and coefficient vectors, T denotes the transpose, and ε_i is the error.

Ridge and Lasso supplied L2 and L1 regularization [3, 4]. Five-fold cross-validation (CV) in RidgeCV searched 13 penalties from 10^{-3} to 10^{3} and selected a_{CV}=31.6228; five-fold LassoCV selected 0.000405.

Estimation used the training sample only, and each fitted model generated predictions for the unchanged test observations. OLS estimates the coefficient vector by minimizing the sum of squared residuals, making it an interpretable benchmark for the average conditional relationship between housing attributes and log price. Ridge minimizes the same loss plus the sum of squared coefficients, whereas Lasso adds the sum of absolute coefficient magnitudes. Standardization before penalized estimation was necessary because a common penalty should not depend on whether a predictor is measured in feet, dollars, years, or degrees. Cross-validation selected penalty strength within the training data, preventing the held-out test set from influencing model tuning. The comparison therefore isolates whether coefficient shrinkage improves generalization relative to the unpenalized baseline while retaining the same response, predictors, split, and evaluation sample.

3.3. Bayesian models

Bayesian Ridge and ARD used standardized predictors and returned predictive means and standard deviations (SD). Bayesian Ridge applies common Gaussian shrinkage, while ARD permits predictor-specific shrinkage. Their inclusion extends recent use of shrinkage in automated valuation [5].

Both Bayesian estimators use Gaussian likelihoods for the log-price response, but they differ in how shrinkage is distributed across predictors. Bayesian Ridge estimates a shared coefficient-precision structure, so all standardized coefficients are regularized through the same general mechanism. ARD estimates separate precisions and can strongly shrink weakly supported coefficients while preserving predictors with clearer evidence. Hyperparameters were learned from the training observations using the estimators' iterative evidence-maximization procedures; the test outcomes were never supplied during fitting. For uncertainty analysis, return_std=True produced a predictive standard deviation for each test property rather than a single global error measure. This design permits direct comparison of posterior predictive intervals with OLS observation-level intervals on identical houses and nominal coverage levels.

DOI: 10.54254/2753-8818/2026.PU37987

3.4. Random forest and point-prediction evaluation

The nonlinear benchmark was a 300-tree Random Forest fitted on original predictors with random_state=42 [7]. It supplied point predictions and feature importance, not intervals. Held-out R^2, RMSE, and MAE followed automated-valuation evaluation guidance [6], while spatial Random Forest research motivated retaining location variables [8].

Each tree was trained on a bootstrap sample with all predictors considered at every split, so aggregation reduces the instability of an individual decision tree. Three hundred trees were used to obtain a stable ensemble without tuning against the test set. Unlike the linear models, the forest can partition the predictor space and represent thresholds and interactions, such as a living-area effect that changes with structure quality or proximity to the coast. R^2 measures explained test variation, whereas RMSE gives greater weight to large log-price errors and MAE summarizes a more robust typical absolute error. Reporting all three prevents a model ranking from depending on one loss function. Impurity-based feature importance was used descriptively to identify influential predictors, not as a causal estimate of their effects.

3.5. Prediction intervals and calibration

Prediction uncertainty was compared for OLS, Bayesian Ridge, and ARD. OLS used observation-level intervals for future prices; the Bayesian models used their predictive means and standard deviations.

Coverage was the test proportion inside each interval and width was the mean upper-minus-lower difference on the log-price scale. Calibration was evaluated at nominal levels from 0.50 to 0.99; good intervals balance calibration and sharpness [10].

For each nominal level, the OLS limits were obtained from the fitted observation-prediction distribution, which includes both uncertainty in the conditional mean and irreducible residual variation. Bayesian Ridge and ARD limits were formed symmetrically around their predictive means using the appropriate standard-normal quantile and observation-specific predictive standard deviation. All interval calculations remained on the log-price scale, ensuring that widths were comparable across methods. Coverage below the nominal level indicates undercoverage and overconfident intervals; coverage above it indicates conservative intervals. Width must be interpreted jointly with coverage because an arbitrarily wide interval can cover nearly every observation without being informative. The seven-level calibration curve was therefore used to evaluate behavior across the distribution rather than relying only on the conventional 95% level.

3.6. Small-sample experiment

Twenty subsets (seeds 0–19) were drawn for each training size, n=500 and n=200, while the test set remained fixed. The scaler, OLS, and Bayesian Ridge were refitted for each subset; RMSE, 95% coverage, and width are reported as means and standard deviations.

Subsampling was performed only within the original training partition, so every replicate was evaluated against the same 2,756 unseen properties. For each seed and sample size, preprocessing was repeated from the selected observations: the scaler was refitted, the two models were re-estimated, and fresh prediction intervals were produced. This repetition avoids carrying full-sample information into the reduced-data conditions. The n=500 and n=200 settings approximate increasingly data-scarce valuation problems while preserving the original predictor dimension. Means across seeds describe average performance, whereas standard deviations show sensitivity to

DOI: 10.54254/2753-8818/2026.PU37987

which transactions happened to be available. The experiment was designed to determine whether Bayesian shrinkage becomes practically distinguishable from OLS as estimation uncertainty rises, rather than to search for a sample size that favors either method.

3.7. Ridge-bayesian ridge connection

The Ridge-Bayesian connection was drawn from the two precision parameters returned by the Bayesian Ridge fit; their ratio defines the equivalent Ridge penalty in equation (3).

$$a_{\text{eq}} = \lambda/\alpha \tag{3}$$

Here a_{eq} is the Bayesian-equivalent Ridge penalty, λ is the coefficient precision, and α is the noise precision.

With λ=92.0847 and α=13.6019, equation (3) gives a_{eq}=6.7700. Ridge at a_{eq} was compared with all 15 Bayesian posterior-mean coefficients. A path from 10^{-3} to 10^{3} marked both a_{eq} and a_{CV}=31.6228.

The equivalence check used the same standardized design matrix and intercept convention for both estimators. After deriving a_{eq}, Ridge was refitted rather than comparing coefficients from the independently tuned RidgeCV model. Coefficient agreement was assessed across all predictors against the 45-degree identity line, so matching predictions could not conceal offsetting coefficient differences. The regularization path then showed how every coefficient changed as the penalty increased, while vertical markers separated the Bayesian-equivalent value from the cross-validated value. This distinction matters because a_{eq} translates the Bayesian posterior-mean optimization problem into Ridge form, whereas a_{CV} is selected solely for predictive performance. The experiment tests a computational correspondence under the fitted assumptions and does not claim that frequentist and Bayesian interpretations are identical.

3.8. Prior-sensitivity experiment

Prior sensitivity was tested at L=0.01, 1, 100, and 10,000. Equation (4) defines the coefficient-precision hyperparameters.

$$\lambda_1 = 10^6 L, \quad \lambda_2 = 10^6 \tag{4}$$

Here L is the prior precision multiplier, λ_1 is the shape of the Gamma prior on the coefficient precision λ, and λ_2 is its fixed rate.

The experiment used the full training set and seed-0 subsets at n=500 and n=200. RMSE and 95% coverage were recorded; reduced-sample results describe these tested subsets rather than all possible samples.

Changing L modifies the prior placed on coefficient precision while leaving the response data, predictors, test set, and remaining model settings unchanged. The logarithmically separated values span weak, moderate, strong, and deliberately extreme conditions, making it possible to identify when the likelihood ceases to dominate the prior. The full-sample fit establishes whether abundant observations protect predictions from prior changes. The two reduced subsets test the same question when information is limited, where prior assumptions should have more opportunity to influence shrinkage. RMSE records the consequence for point prediction and 95% coverage records the consequence for uncertainty calibration. Because only seed 0 was used in the reduced prior

DOI: 10.54254/2753-8818/2026.PU37987

experiment, the analysis is a controlled sensitivity demonstration rather than a repeated-sampling estimate of average prior effects.

4. Results

4.1. Descriptive statistics and exploratory analysis

After deduplication, 13,776 houses remained. Figure 1 shows strong right skew in price and a more symmetric log response. Figure 2 identifies living area as the strongest correlate of ln_price (r=0.713), followed by special-feature value (r=0.506) and structure quality (r=0.463), and also shows correlation among distance variables. Figure 3 contrasts the clear positive living-area relationship with the weaker negative ocean-distance relationship. Table 2 summarizes all analysis variables.

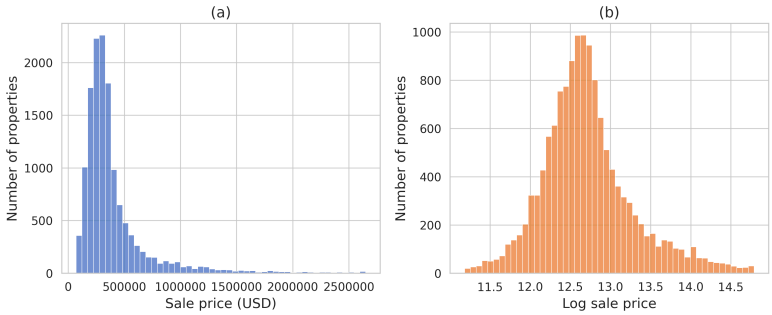

Figure 1. Original and log-transformed house-price distributions: (a) sale price; (b) log sale price

Data from: Miami Housing 2016 dataset

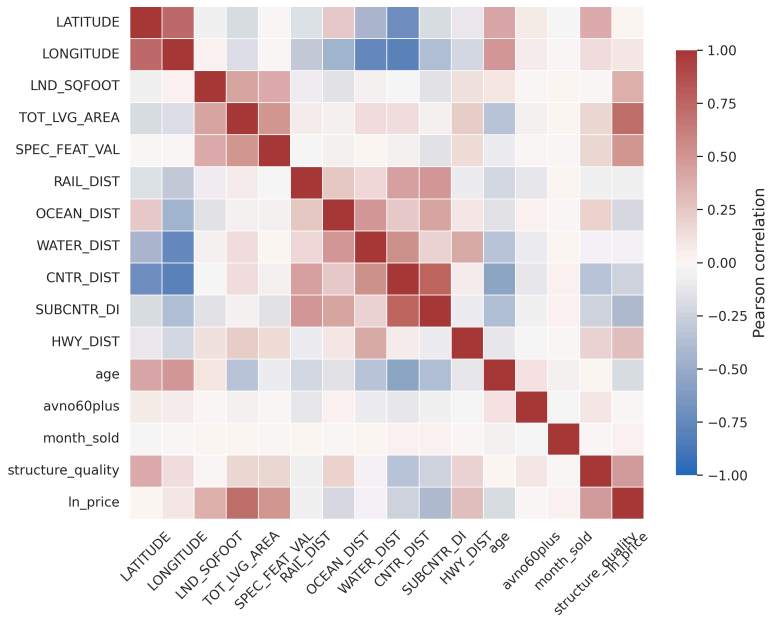

Figure 2. Correlation heatmap for the 15 predictors and log house price

Data from: Miami Housing 2016 dataset

DOI: 10.54254/2753-8818/2026.PU37987

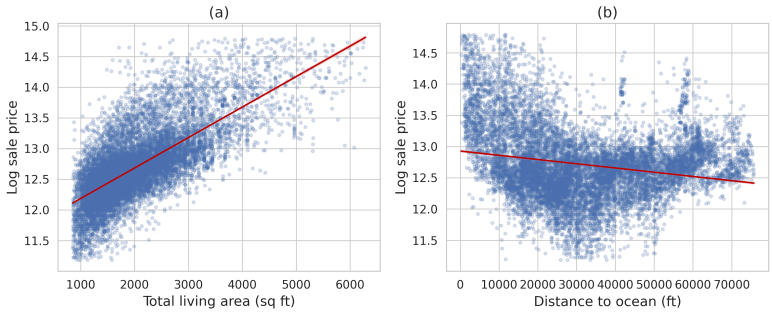

Figure 3. Selected predictor relationships with log house price: (a) living area; (b) distance to ocean

Data from: Miami Housing 2016 dataset

Table 2. Descriptive statistics after deduplication

Variable	Mean	SD	Median	Min	Max
LATITUDE	25.7283	0.1406	25.7310	25.4343	25.9744
LONGITUDE	−80.3279	0.0891	−80.3394	−80.5422	−80.1197
LND_SQFOOT	8623.96	6063.47	7500.00	1248.00	57064.00
TOT_LVG_AREA	2061.50	814.36	1881.00	854.00	6287.00
SPEC_FEAT_VAL	9604.75	13923.55	2784.00	0.00	175020.00
RAIL_DIST	8354.85	6175.57	7124.95	10.50	29621.50
OCEAN_DIST	31710.96	17609.83	28595.55	236.10	75744.90
WATER_DIST	11994.25	11941.12	6979.95	0.00	50399.80
CNTR_DIST	68625.84	31990.84	66039.80	3825.60	159976.50
SUBCNTR_DI	41149.98	22167.83	41156.40	1462.80	110553.80
HWY_DIST	7731.19	6074.76	6158.85	90.20	48167.30
age	30.5873	21.1168	26.00	0.00	96.00
avno60plus	0.0150	0.1214	0.00	0.00	1.00
month_sold	6.6352	3.3024	7.00	1.00	12.00
structure_quality	3.5173	1.0963	4.00	1.00	5.00
ln_price	12.7128	0.5682	12.6479	11.1844	14.7901

Data from: Miami Housing 2016 dataset

4.2. Point-prediction comparison

Table 3 shows near-identical performance for the five linear models: OLS obtained R^2=0.7658, RMSE=0.2784, and MAE=0.2084, with only negligible changes from regularization or Bayesian shrinkage.

Random Forest improved R^2 to 0.9237 and reduced RMSE to 0.1589 and MAE to 0.1048. Figure 4 ranks living area first in importance (0.440), followed by distance to the central business district,

DOI: 10.54254/2753-8818/2026.PU37987

structure quality, ocean distance, and subcenter distance.

Table 3. Test-set point-prediction performance

Model	Test R^2	RMSE	MAE
OLS	0.765823	0.278380	0.208448
RidgeCV	0.765883	0.278344	0.208400
LassoCV	0.765888	0.278341	0.208493
Bayesian Ridge	0.765841	0.278369	0.208437
ARD Regression	0.765535	0.278551	0.208680
Random Forest	0.923732	0.158869	0.104809

Data from: Miami Housing 2016 dataset

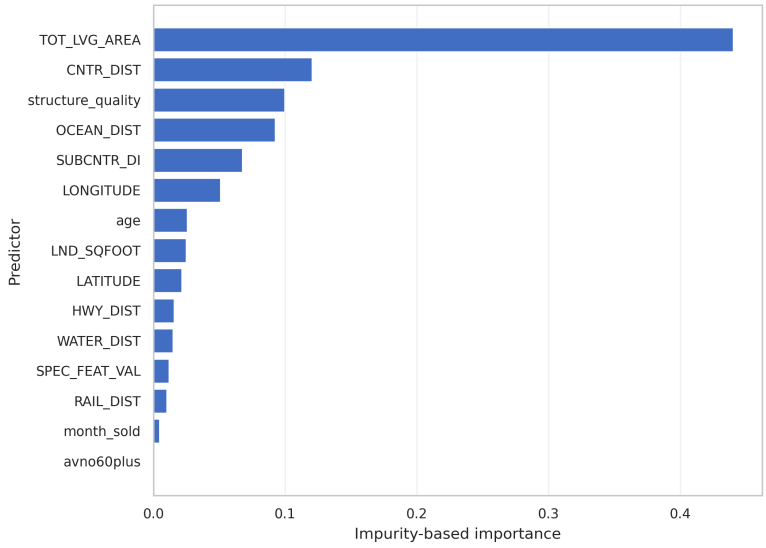

Figure 4. Random Forest feature importance

Data from: Miami Housing 2016 dataset

The roughly 43% lower RMSE confirms a substantive rather than marginal Random Forest advantage. Because all five linear estimators performed similarly, their shared linear form—not the choice of shrinkage rule—was the main predictive constraint. Forest feature importance nevertheless remains descriptive: it ranks variables used for splitting but does not estimate causal price effects. Table 3 and Figure 4 therefore support a focused interpretation: nonlinearities and interactions improve prediction, while the linear models provide transparent benchmarks for the subsequent uncertainty comparison.

4.3. Full-sample prediction intervals and calibration

Table 4 shows almost identical 95% intervals: OLS and Bayesian Ridge coverage was 0.9387, ARD coverage was 0.9394, and widths were approximately 1.064. All three slightly undercovered relative to 0.95.

DOI: 10.54254/2753-8818/2026.PU37987

The nearly overlapping curves in Figure 5 were above target at lower levels but below target at 0.90 (about 0.895), 0.95, and 0.99.

Table 4. Prediction-interval coverage and average width

Method	Nominal	Empirical	Average width
OLS	80%	0.822932	0.695557
Bayesian Ridge	80%	0.823295	0.695447
ARD Regression	80%	0.822932	0.695588
OLS	90%	0.895501	0.892760
Bayesian Ridge	90%	0.895501	0.892597
ARD Regression	90%	0.895138	0.892777
OLS	95%	0.938679	1.063816
Bayesian Ridge	95%	0.938679	1.063594
ARD Regression	95%	0.939405	1.063809

Data from: Miami Housing 2016 dataset

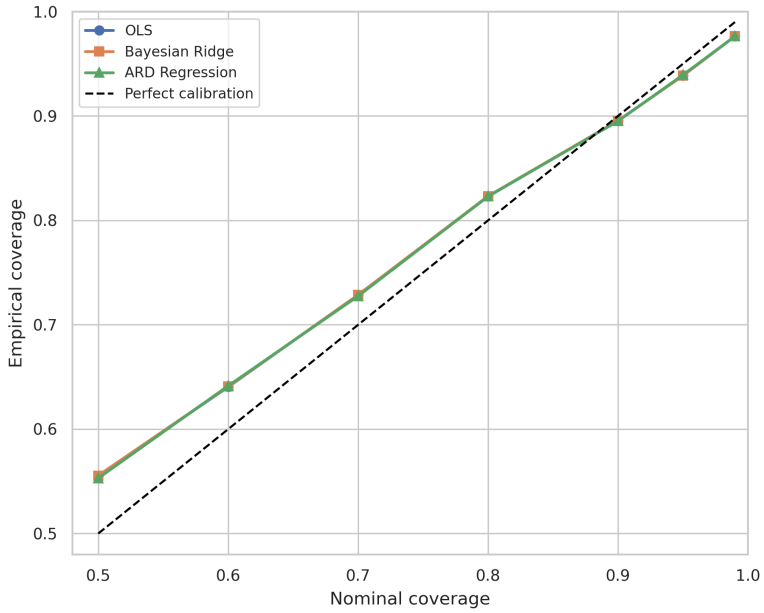

Figure 5. Nominal versus empirical prediction-interval coverage

Data from: Miami Housing 2016 dataset

Width differences at the 95% level were below three ten-thousandths, so no method had a meaningful sharpness advantage. Coverage was about 1.1 percentage points below target, indicating mild undercoverage. Figure 5 also shows conservative behavior at several lower levels and slight overconfidence at the highest levels. The almost coincident curves imply that this pattern is shared across the fitted linear predictive distributions rather than being specific to one estimator. Thus, small differences in point accuracy did not translate into distinguishable full-sample interval performance.

DOI: 10.54254/2753-8818/2026.PU37987

4.4. Small-sample results

Table 5 shows wider intervals as training data decreased. OLS/Bayesian Ridge widths rose from 1.0638/1.0636 at full size to 1.0746/1.0681 at n=500 and 1.1164/1.0976 at n=200.

Coverage remained similar across methods, while variation across the 20 subsets increased at smaller n. At n=200, coverage was 0.9378 for OLS and 0.9354 for Bayesian Ridge. Figure 6 visualizes the stable mean coverage and increasing interval width and variability.

Table 5. Small-sample prediction performance and uncertainty

Training	Method	RMSE mean (SD)	Coverage mean (SD)	Width mean (SD)
Full	OLS	0.278380 (-)	0.938679 (-)	1.063816 (-)
Full	Bayesian Ridge	0.278369 (-)	0.938679 (-)	1.063594 (-)
500	OLS	0.284375 (0.001984)	0.934452 (0.014607)	1.074580 (0.070995)
500	Bayesian Ridge	0.283685 (0.002066)	0.933599 (0.014821)	1.068116 (0.070255)
200	OLS	0.292684 (0.007301)	0.937827 (0.016931)	1.116365 (0.105203)
200	Bayesian Ridge	0.289804 (0.006098)	0.935396 (0.016695)	1.097568 (0.102504)

Data from: Miami Housing 2016 dataset

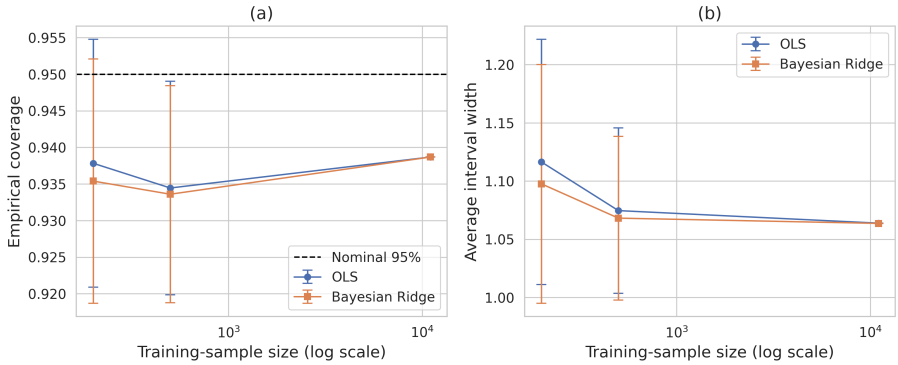

Figure 6. Prediction intervals across training-sample sizes: (a) 95% empirical coverage; (b) average 95% interval width

Data from: Miami Housing 2016 dataset

From the full sample to n=200, average width increased by about 0.053 for OLS and 0.034 for Bayesian Ridge, consistent with greater parameter uncertainty. Bayesian Ridge was modestly narrower but also had slightly lower coverage, so narrowness alone does not establish better performance. The overlapping error bars in Figure 6 show that between-method differences were small relative to variation among subsamples. The dominant result is therefore declining stability as training data become scarce, which cautions against treating intervals from one small sample as generally representative.

4.5. Ridge-bayesian ridge relationship

RidgeCV selected a larger a_{CV}=31.6228. Figure 7 locates both penalties on the coefficient path: cross-validation optimizes predictive performance, whereas a_{eq} follows from Bayesian precision

DOI: 10.54254/2753-8818/2026.PU37987

estimates.

Bayesian Ridge produced λ=92.0847, α=13.6019, and a_{eq}=6.7700. Figure 8 shows that Ridge at a_{eq} matches all 15 Bayesian posterior-mean coefficients to numerical precision.

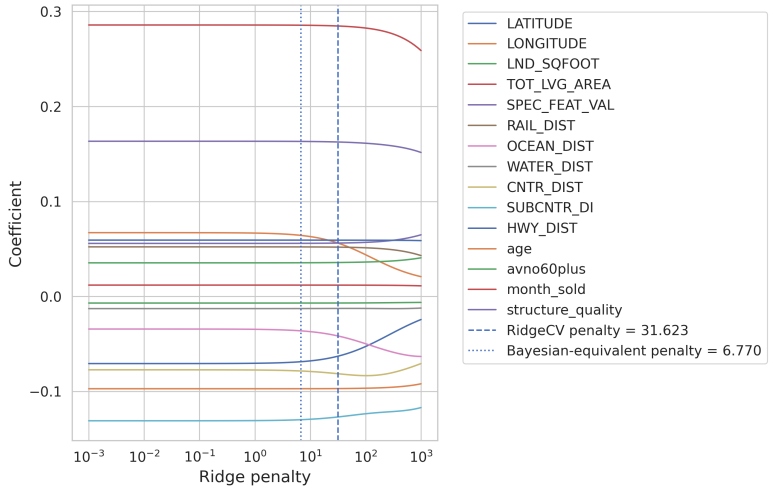

Figure 7. Ridge regularization path with cross-validated and Bayesian-equivalent penalties

Data from: Miami Housing 2016 dataset

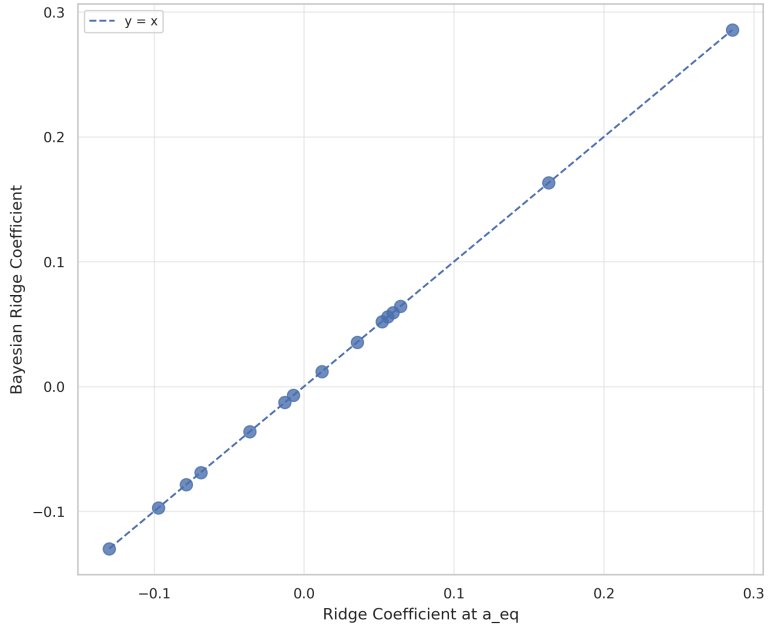

Figure 8. Ridge and Bayesian Ridge coefficient comparison

Data from: Miami Housing 2016 dataset

The two penalties differ by a factor of about 4.67, with the cross-validated value farther along the shrinkage path. Gradual path changes around both values help explain their similar test errors. More importantly, Figure 8 places every standardized coefficient on the identity line after the penalties are matched, establishing parameter-level agreement rather than only similar predictions. This confirms the expected optimization correspondence between Gaussian coefficient shrinkage and an L2

DOI: 10.54254/2753-8818/2026.PU37987

penalty for this design. The distinct penalty values still show that cross-validation and Bayesian evidence estimation use different selection criteria.

4.6. Prior-sensitivity results

Table 6 shows stable Bayesian Ridge performance for L=0.01, 1, and 100. At full size, RMSE stayed near 0.2784 and coverage near 0.939; even L=10,000 changed RMSE only to 0.2793.

The extreme prior mattered in reduced samples: at n=500 RMSE rose to 0.4354 and coverage fell to 0.9075; at n=200 RMSE rose to 0.5219 and coverage fell to 0.8973. Figure 9 makes the RMSE divergence visible.

Table 6. Bayesian Ridge prior-sensitivity results

Training	L	RMSE	Coverage
Full	0.01	0.278380	0.938679
Full	1	0.278380	0.938679
Full	100	0.278368	0.938679
Full	10,000	0.279311	0.934325
500	0.01	0.285340	0.930697
500	1	0.285297	0.930697
500	100	0.284421	0.930334
500	10,000	0.435373	0.907475
200	0.01	0.287063	0.924165
200	1	0.286730	0.924528
200	100	0.282155	0.926705
200	10,000	0.521932	0.897315

Data from: Miami Housing 2016 dataset

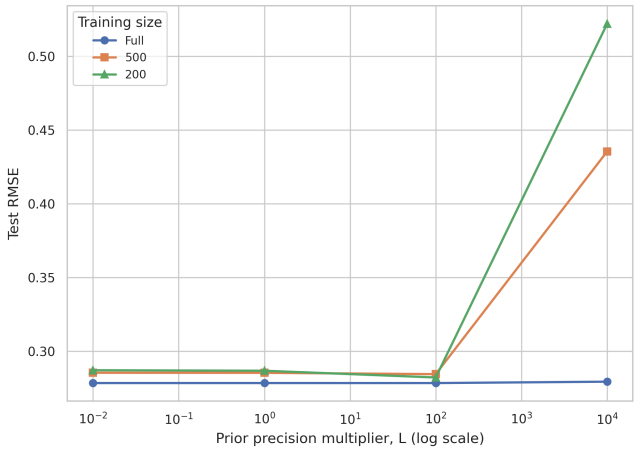

Figure 9. Prior-strength sensitivity across training sample sizes

Data from: Miami Housing 2016 dataset

DOI: 10.54254/2753-8818/2026.PU37987

The full-sample curve is nearly flat across the tested values, indicating that abundant data dominated the prior. Results for n=500 and n=200 also remained stable through L=100, but both deteriorated sharply at L=10,000. The smaller subset showed the largest RMSE increase and the greatest coverage shortfall. Because accuracy and calibration worsened together, the extreme setting offered no compensating benefit. The evidence supports a conditional conclusion: ordinary Bayesian Ridge settings were robust, whereas excessive prior precision could dominate limited data and materially harm forecasts.

5. Discussion

Random Forest, with an R^2 of 0.9237 and an RMSE of 0.1589, substantially outperformed the linear models (R^2 near 0.766; RMSE near 0.278). Regularization changes coefficient estimation but preserves linearity, whereas forests capture nonlinearities and interactions. The result therefore points to functional form, rather than insufficient shrinkage, as the main limitation of the linear specifications.

Interval results tell a different story. With 11,020 training observations, OLS, Bayesian Ridge, and ARD all achieved coverage near 0.939 and width near 1.064, with nearly overlapping calibration curves. Bayesian modeling therefore did not automatically deliver narrower or better-calibrated intervals when the likelihood was supported by abundant data.

Smaller training sets widened intervals and increased variability across repeated subsets. At n=200, Bayesian Ridge was modestly narrower than OLS (1.0976 versus 1.1164) with similar coverage. Sample size had a larger practical effect on uncertainty than the choice between these two linear estimators.

The coefficient experiment explains the similar Ridge and Bayesian Ridge predictions: at a_{eq}=6.7700, their standardized coefficients coincide numerically. RidgeCV selected 31.6228 because it optimized cross-validated prediction, while a_{eq} was derived from estimated precisions. The computational equivalence of Ridge and Bayesian Ridge at a corresponding penalty does not imply identical procedures for choosing shrinkage.

Prior sensitivity was negligible for L up to 100 but severe at L=10,000 in reduced samples. The n=200 RMSE rose to about 0.522 and coverage fell to 0.897, while the full-sample fit changed little. Priors can therefore dominate when data are scarce and should be stress-tested rather than assumed beneficial.

The evidence comes from one city-year and one fixed outer split, so generalizability to other markets and periods remains uncertain. Random Forest intervals were not evaluated, and the prior experiment used one reduced subset per size. Future work should add repeated outer splits, other markets, and nonlinear uncertainty methods such as conformal prediction or quantile regression forests.

6. Conclusion

This study compared frequentist, regularized, Bayesian, and machine-learning approaches for predicting log house prices in the Miami Housing 2016 dataset. The five linear models produced similar point-prediction performance, with test R^2 values close to 0.766. Random Forest performed substantially better, achieving a test R^2 of 0.9237 and an RMSE of 0.1589. The uncertainty comparison produced a different pattern: with the full training sample, OLS, Bayesian Ridge, and ARD had nearly identical prediction-interval coverage and width. The similarity of the full-sample linear intervals suggests that methodological complexity alone does not ensure a material calibration

DOI: 10.54254/2753-8818/2026.PU37987

advantage when data are abundant. With reduced training samples, intervals became wider and more variable, while Bayesian Ridge produced only slightly narrower intervals than OLS with similar coverage.

The additional experiments clarified the relationship between regularization and Bayesian modeling. Ridge fitted with the Bayesian-equivalent penalty reproduced the Bayesian Ridge coefficients to numerical precision, illustrating the connection between L2 regularization and Gaussian Bayesian shrinkage under corresponding settings. Bayesian Ridge remained stable under weak to strong prior-strength settings, while the extreme setting had a much larger negative effect in small samples. Overall, point-prediction accuracy and uncertainty estimation should be treated as related but separate goals.

For practitioners, model comparisons should report uncertainty alongside point error, and Bayesian shrinkage assumptions should be stress-tested when samples are small. Future studies should repeat the analysis across years and markets and compare nonlinear interval methods, including quantile regression forests and conformal prediction. Such extensions would test whether the forest's point-prediction advantage can be paired with reliable intervals and whether the present conclusions remain stable under changes in housing-market conditions.

References

[1] Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82*(1), 34-55. https://doi.org/10.1086/260169

[2] Mayer, M., Bourassa, S. C., Hoesli, M., & Scognamiglio, D. (2022). Machine learning applications to land and structure valuation. *Journal of Risk and Financial Management, 15*(5), 193. https://doi.org/10.3390/jrfm15050193

[3] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67. https://doi.org/10.1080/00401706.1970.10488634

[4] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[5] Hinrichs, N., Kolbe, J., & Werwatz, A. (2021). Using shrinkage for data-driven automated valuation model specification: A case study from Berlin. *Journal of Property Research, 38*(2), 130-153. https://doi.org/10.1080/09599916.2021.1905690

[6] Steurer, M., Hill, R. J., & Pfeifer, N. (2021). Metrics for evaluating the performance of machine learning based automated valuation models. *Journal of Property Research, 38*(2), 99-129. https://doi.org/10.1080/09599916.2020.1858937

[7] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[8] Hu, L., Chun, Y., & Griffith, D. A. (2022). Incorporating spatial autocorrelation into house sale price prediction using random forest model. *Transactions in GIS, 26*(5), 2123–2144. https://doi.org/10.1111/tgis.12931

[9] Elnaeem Balila, A., & Shabri, A. B. (2024). Comparative analysis of machine learning algorithms for predicting Dubai property prices. *Frontiers in Applied Mathematics and Statistics, 10*, 1327376. https://doi.org/10.3389/fams.2024.1327376

[10] Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102*(477), 359–378. https://doi.org/10.1198/016214506000001437

[11] Bourassa, S. C. (2021, September 24). MiamiHousing2016 [Data set]. *OpenML*. https://www.openml.org/d/43093

[12] Feurer, M., van Rijn, J. N., Kadra, A., Gijsbers, P., Mallik, N., Ravi, S., Müller, A., Vanschoren, J., & Hutter, F. (2021). OpenML-Python: An extensible Python API for OpenML. *Journal of Machine Learning Research, 22*(100), 1–5. https://www.jmlr.org/papers/v22/19-920.html