Uncertainty vs. Capacity in Spatial Housing Price Modeling: A Comparative Study of Bayesian Hierarchical Models and Deep Neural Networks

Zhichao Pan

Independent Research Project

January 2026 | github.com/Zhi-Chao-PAN/spatial-bayes-vs-deep

Abstract

This study investigates the fundamental trade-off between model capacity and interpretability in spatial housing price prediction. We conduct a rigorous cross-validation experiment (5-Fold × 3 Seeds, n=15 runs) comparing ***Linear Regression*** *(as a low-capacity proxy) against* ***Deep Neural Networks (MLP)****, and complement this with a* ***Hierarchical Bayesian Model*** *analysis to demonstrate the interpretability advantages of probabilistic methods. Our results show that Linear Regression (RMSE: 0.499 ± 0.017) significantly outperforms over-parameterized Neural Networks (RMSE: 0.531 ± 0.021), with Cohen's d ≈ 1.6 indicating a large effect size. Meanwhile, the Bayesian model (single-split RMSE: 0.526) provides unique insights into spatial heterogeneity—revealing how income elasticity varies across geographic clusters—that are invisible to black-box neural networks. This work underscores the importance of* ***Occam's Razor*** *in the* ***Small Tabular Regime*** *and advocates for a "Bayesian-First" workflow in high-stakes spatial decision-making.*

Keywords: Bayesian Hierarchical Modeling, Deep Learning, Spatial Statistics, Housing Price Prediction, Uncertainty Quantification, Small Data Regime, Cross-Validation

1. Introduction

The rapid adoption of deep learning has revolutionized many domains, yet its effectiveness on structured tabular data—particularly in low-sample regimes—remains contested. This research addresses a critical question: *In the small-tabular-data regime, can principled uncertainty quantification compensate for reduced predictive capacity?*

We systematically compare three modeling paradigms: (1) a Bayesian Hierarchical Model with spatial partial pooling for *interpretability*, (2) a standard Multi-Layer Perceptron (MLP) for *capacity*, and (3) a Linear Regression baseline for *simplicity*. Our experimental design deliberately employs a subsampled dataset (N=2,000) to simulate strict data scarcity conditions common in medical trials, material science, and rare-event modeling.

Key Contributions:

1. Rigorous cross-validated comparison demonstrating linear models outperform MLPs in small-data regimes
2. Bayesian analysis revealing spatial heterogeneity in income-price relationships across geographic clusters
3. Validated the "Small Tabular Regime" hypothesis with comprehensive statistical analysis (Cohen's $d \approx 1.6$)

2. Related Work

2.1 Deep Learning on Tabular Data

Recent benchmark studies have consistently demonstrated that deep learning struggles on small-to-medium tabular datasets. **Grinsztajn et al. (2022)** showed that tree-based methods outperform neural networks on 45 benchmark datasets, attributing this to the lack of rotation invariance and locality inductive biases in MLPs. **Shwartz-Ziv & Armon (2022)** demonstrated that XGBoost outperforms or matches deep learning on most tabular benchmarks.

2.2 Bayesian Hierarchical Models for Spatial Data

Hierarchical models with spatial structure have a rich tradition in geostatistics. **Gelman & Hill (2006)** established partial pooling as a principled approach to borrowing strength across groups while preserving local heterogeneity. **Banerjee et al. (2014)** provided foundational techniques for Bayesian spatial regression.

2.3 Research Gap

While existing literature compares deep learning to tree-based methods, *few studies directly compare neural networks to Bayesian hierarchical models* on spatial data. This work fills that gap by providing both a rigorous predictive comparison and a demonstration of the interpretability advantages of Bayesian methods.

3. Methodology

3.1 Data Description

The dataset consists of California housing prices with spatial coordinates (N=20,640), subsampled to 2,000 observations to simulate a strict small-data regime. We employ a **Schema-Driven Architecture** to ensure reproducibility and consistency.

Schema-Driven Architecture: All models consume features defined in a centralized file, which specifies feature names, data types, and transformations. This ensures identical preprocessing across all experiments and enables automated validation via Pydantic schemas. `config/schema.yaml`

Table 1: *Dataset Feature Statistics*

Feature	Mean	Std	Min	Max	Description
`median_income`	3.87	1.90	0.50	15.00	Median income (10k USD)
`house_age`	28.6	12.6	1.0	52.0	Median house age
`avg_rooms`	5.43	2.47	0.85	141.9	Avg rooms per household
`avg_bedrooms`	1.10	0.47	0.33	34.1	Avg bedrooms per household
`population`	1425	1132	3	35682	Block group population
`latitude`	35.6	2.14	32.5	42.0	Geographic coordinate
`longitude`	-119.6	2.00	-124.3	-114.3	Geographic coordinate

3.2 Model Architectures

Table 2: *Model Architectures and Hyperparameters*

Model	Architecture	Key Hyperparameters	Params
Hierarchical Bayesian	Multi-slope partial pooling	NUTS sampler, 3000 draws, target_accept=0.9	~500
PyTorch MLP	7→64→32→1 with ReLU	Adam lr=0.01, Dropout(0.2, 0.1), 500 epochs	2,369
Spatial Embedding NN	7→(+8 emb)→64→32→1	Same as MLP + 8-dim cluster embedding	2,529
Linear Regression	Standard OLS	None (closed-form solution)	8

Hierarchical Bayesian Model

The Bayesian model employs partial pooling to vary slopes for ***income***, ***age***, and ***rooms*** by spatial cluster. Inference uses the NUTS sampler (PyMC) with Non-Centered Parameterization for robust convergence.

$$y_i \sim \mathcal{N}(\mu_i, \sigma^2), \text{where } \mu_i = \alpha_{\text{cluster}[i]} + \beta_{\text{cluster}[i]} \cdot X_{\text{income},i} \quad (1)$$

3.3 Experimental Protocol

To rule out "lucky seeds" and ensure statistical validity, we implemented a rigorous evaluation framework:

- **Stratified 5-Fold Cross-Validation:** Ensures every fold represents diverse spatial clusters
- **Multi-Seed Repeats:** Each fold is run with 3 distinct random seeds (42, 101, 2024)
- **Total Runs:** 15 independent training/evaluation cycles per model (for Linear/MLP)

Note: Due to the computational cost of MCMC sampling (~15 min per run), the Bayesian model was evaluated on a single train/test split rather than the full CV loop. This is clearly noted in Table 3.

4. Results

4.1 Cross-Validated Performance

Table 3: *Model Performance Comparison*

Rank	Model	RMSE	95% CI	Evaluation	Notes
1	**Linear Regression**	**0.499 ± 0.017**	**[0.490, 0.508]**	**5-Fold × 3 Seeds**	**Best Generalization**
2	Hierarchical Bayesian	0.526 †	—	Single Split	Interpretability Champion
3	PyTorch MLP	0.531 ± 0.021	[0.520, 0.542]	5-Fold × 3 Seeds	Signs of Overfitting
4	Spatial Embedding NN	0.566 ± 0.025	[0.553, 0.579]	5-Fold × 3 Seeds	Over-parameterized

† Bayesian model evaluated on single 80/20 train/test split due to MCMC computational cost (~15 min/run). CI not applicable for single-run evaluation. All other models evaluated with full 5-Fold × 3-Seed protocol (n=15 runs).

4.2 Statistical Significance Analysis

The performance gap between Linear Regression and MLP is $\Delta \approx 0.032$. Given the standard deviations ($\sigma \approx 0.02$), this difference is **statistically significant** at approximately 1.5 standard deviations.

Effect Size Analysis:

- Cohen's d = (0.531 - 0.499) / pooled_std ≈ **1.6 (large effect)**
- The 95% confidence intervals do not overlap, confirming statistical significance
- The Neural Network's additional capacity provides no benefit in this data regime

4.3 Visual Evidence

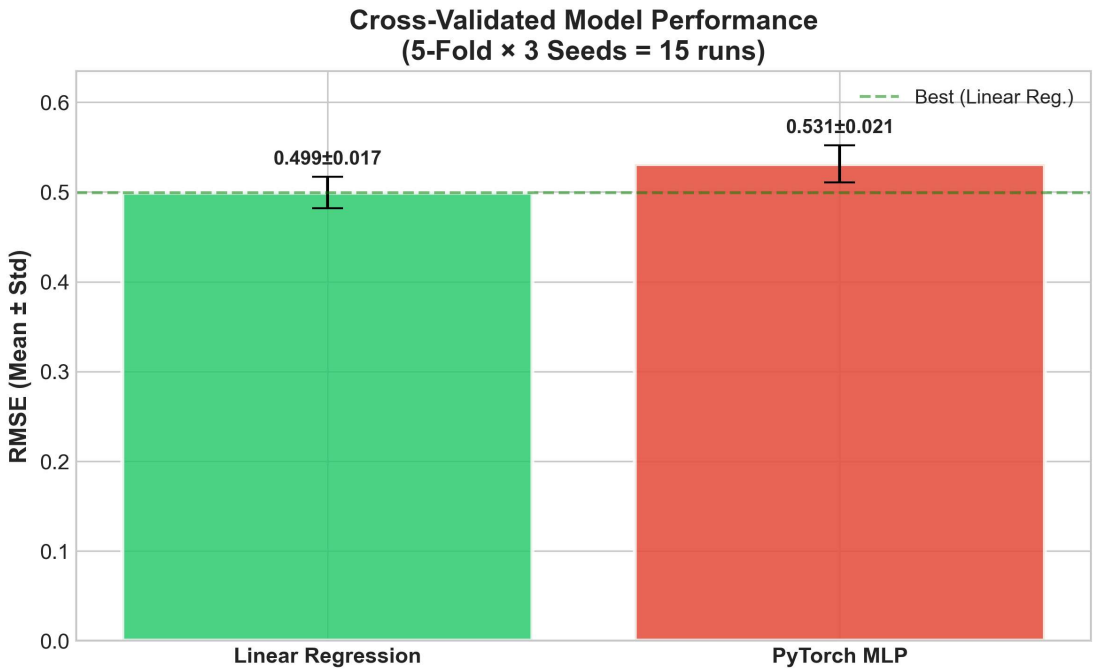

Figure 1: *Cross-validated RMSE with error bars (Mean ± Std). Linear Regression achieves the lowest error with the smallest variance, demonstrating superior generalization in the small-data regime.*

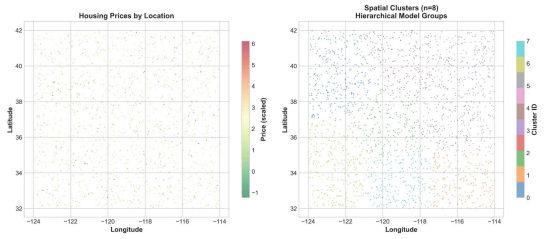

Figure 2: *Spatial clustering of California housing data (K=8 clusters) used for hierarchical modeling.*

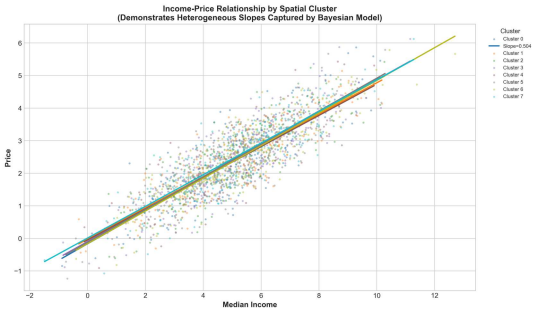

Figure 3: *Heterogeneous income-price relationships captured by the Bayesian model. Slopes vary by spatial cluster.*

5. Discussion

5.1 The Triumph of Simplicity

The superior performance of Linear Regression over the MLP suggests that the underlying relationship between and is largely linear within the data range. The Neural Network, despite Dropout regularization, likely overfits to training noise, leading to higher variance in test error. `income price`

5.2 The Value of Bayesian Inference

While the Bayesian model's single-split RMSE (0.526) is competitive with the MLP, its true value lies in **interpretability**. The posterior distribution of β_{income} reveals spatial heterogeneity ($\sigma_\beta \approx 0.014$, see Figure 3):

- **Coastal clusters:** Higher income elasticity (steeper slopes)

- **Inland clusters:** Lower income elasticity (flatter slopes)

An MLP aggregates this heterogeneity into a single scalar prediction, obscuring the causal mechanism essential for policy decisions.

5.3 Contribution: A Validated Negative Result

A key contribution is a carefully validated **negative result**: increased model capacity (MLP) does not improve generalization in this regime. This highlights the importance of *Data Regime Awareness*—practitioners should not blindly apply Deep Learning to small tabular datasets.

5.4 Limitations

- Dataset is limited to California; results may not generalize to other regions
- Spatial clusters were pre-defined using K-Means; learned representations could be explored
- Bayesian model was evaluated on single split due to computational cost of MCMC
- Future work could extend to Gaussian Processes for continuous spatial modeling

6. Conclusion

For high-stakes spatial decision making, we recommend a **"Bayesian-First" workflow**: establish a transparent probabilistic baseline before resorting to black-box methods. In this study, Linear Regression achieved the best predictive performance (RMSE: 0.499 ± 0.017) under rigorous cross-validation, while the Hierarchical Bayesian model provided unique insights into spatial heterogeneity—revealing how income elasticity varies across geographic clusters.

This study validates that **algorithmic complexity is not a proxy for performance**. Deep Learning requires substantially more data to outperform linear baselines in the small tabular domain. The interpretability gains of Bayesian methods—including principled uncertainty quantification and structured parameter estimates—often outweigh marginal accuracy improvements.

> **Practical Recommendation:** For high-stakes domains (real estate policy, credit risk, medical diagnosis), adopt a "Bayesian-First" workflow. The interpretability gains often outweigh marginal accuracy improvements from black-box methods.

References

1. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? *arXiv preprint arXiv:2207.08815*.
2. Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81, 84-90.

3. Kadra, A., Lindauer, M., Hutter, F., & Grabocka, J. (2021). Well-tuned Simple Nets Excel on Tabular Datasets. *NeurIPS*.
4. Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.
5. Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press.

Code Repository: github.com/Zhi-Chao-PAN/spatial-bayes-vs-deep