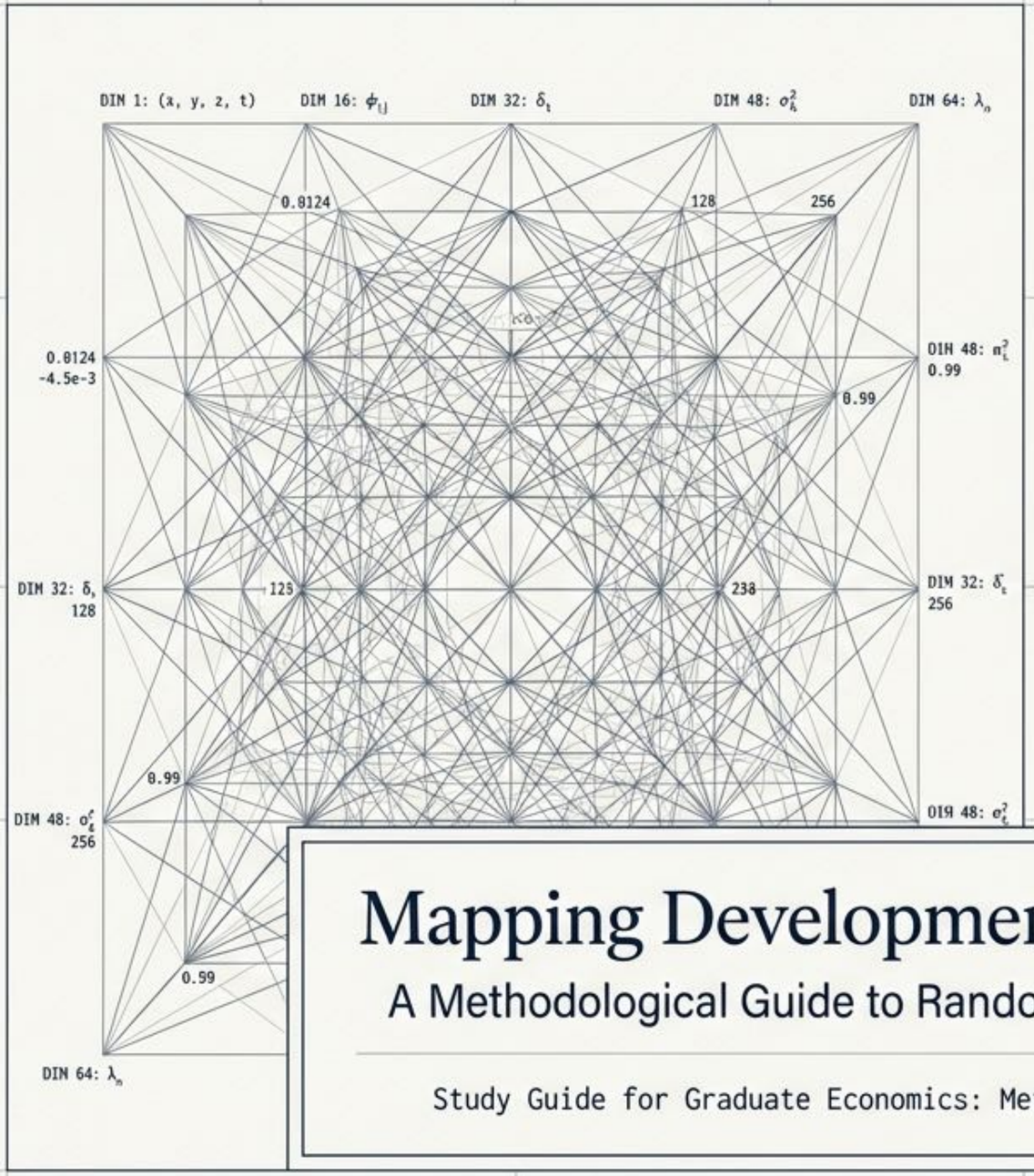
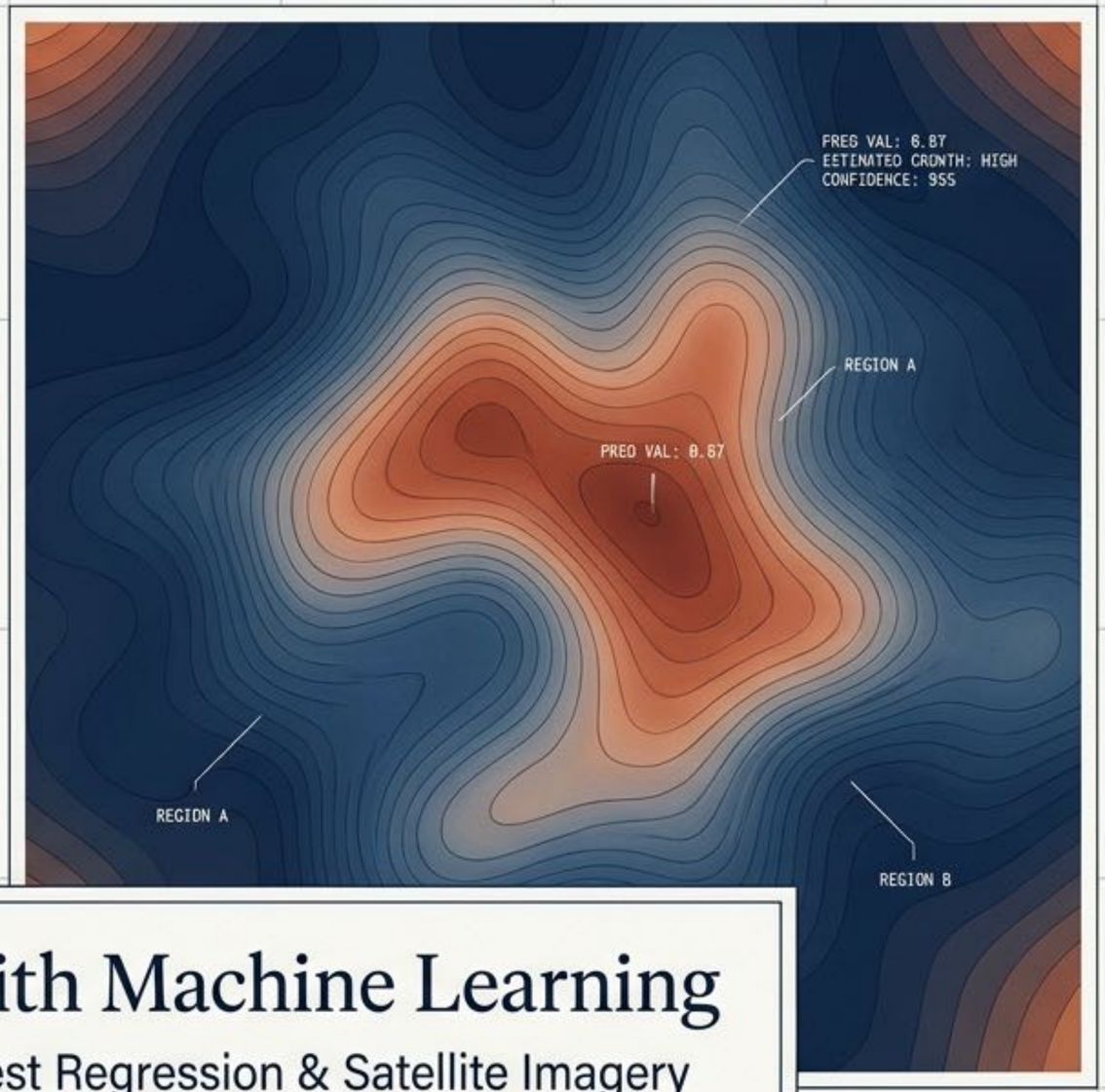




Mapping Development with Machine Learning

A Methodological Guide to Random Forest Regression & Satellite Imagery

Study Guide for Graduate Economics: Methodologies, Implementation, and Interpretation

The Problem Statement

Can high-dimensional satellite embeddings predict continuous socioeconomic outcomes where survey data is sparse?

Empirical Inputs

N = 339 Bolivian municipalities (complete coverage, no missing values).

Target: Municipal Sustainable Development Index (IMDS). 0-100 scale, μ=51.05,

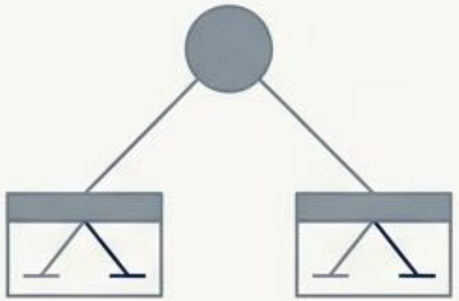
μ=51.05, σ=6.77

Features: 64-dimensional vectors (“A00” to “A63”) extracted from 2017 satellite imagery (land use, urbanization, terrain).

The Algorithmic Solution

Random Forest Regression.

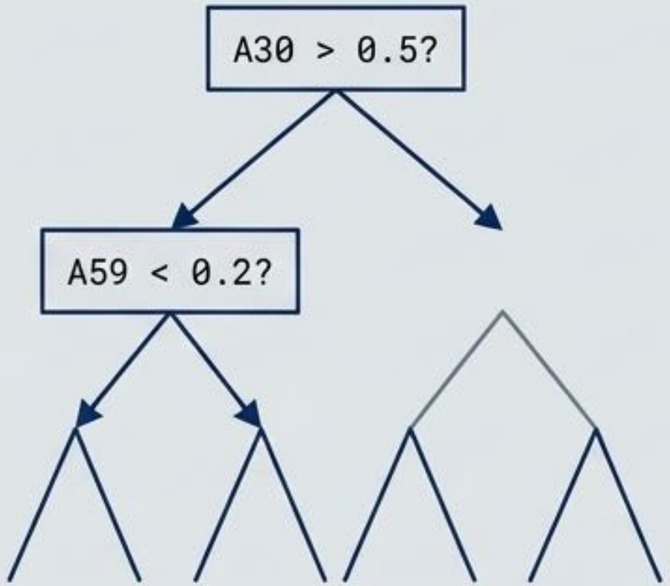
Selected for its innate capacity to handle non-linearities and spatial multicollinearity without requiring extensive manual preprocessing or interaction terms.



Term	Formal Definition	Intuitive Analogy	Bolivian Study Example
Decision Tree	`Recursive binary splits maximizing node purity.`	A sequence of yes/no questions ending in a verdict. 	Splitting first on A30 (built-up signal), then A59, outputting an IMDS prediction at the leaf.
Bagging (Bootstrap Aggregating)	$\hat{y} = \frac{1}{B}\sum T_b(\mathbf{x})$. Reduces variance via bootstrap resampling.	Polling many slightly different juries and averaging their verdicts. 	Growing 100 trees on 100 bootstrap samples of the 339 municipalities.
Random Forest	`Bagging + Random Feature Subsets` (decorrelates trees).	Blindfolding each juror to a random subset of evidence. 	Each split considers only $\sqrt{64} = 8$ embedding dimensions.
Out-of-Fold (OOF) Prediction	Predicting ŷ_i using a model trained on D \ D_fold_containing_i.	Every student sits the exam once, graded by a teacher who never tutored them. 	339 distinct IMDS predictions, preventing data leakage.

Model Architecture: Deconstructing the Random Forest

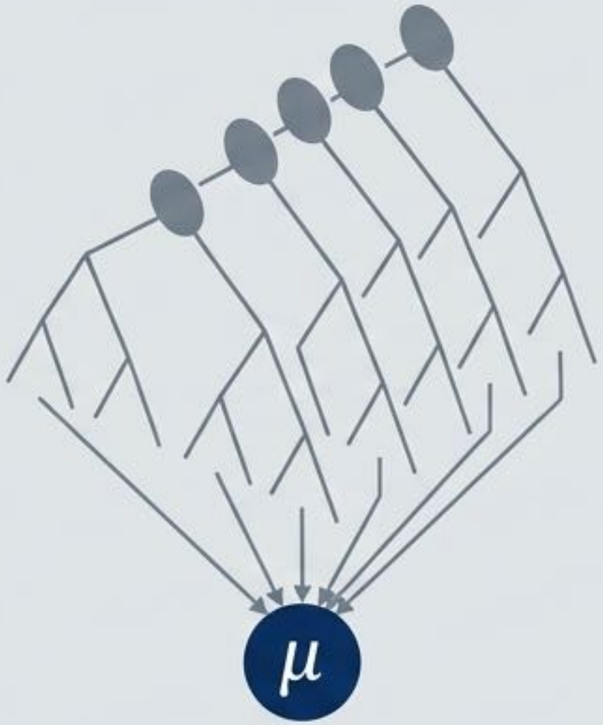
Step 1: Decision Tree



Note: Prone to high variance/overfitting the

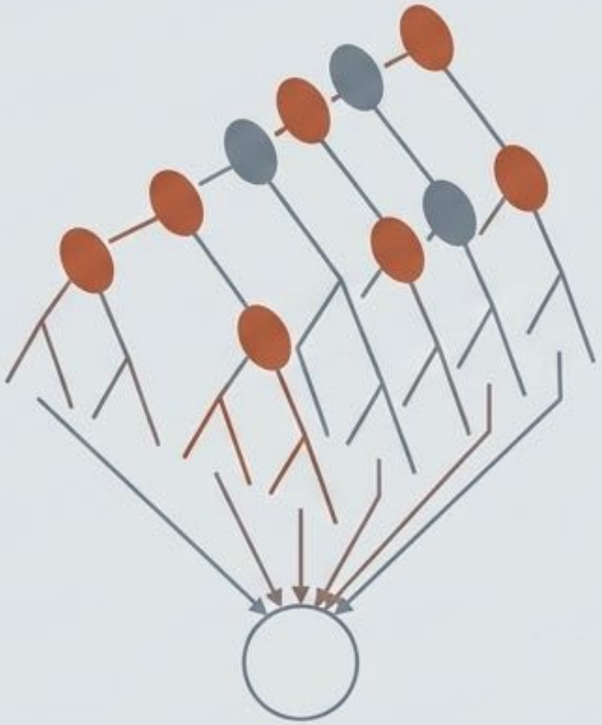
the specific training sample.

Step 2: Bagging



Variance Reduction. Bootstrapping samples the municipalities (N=339 with replacement), stabilizing the prediction.

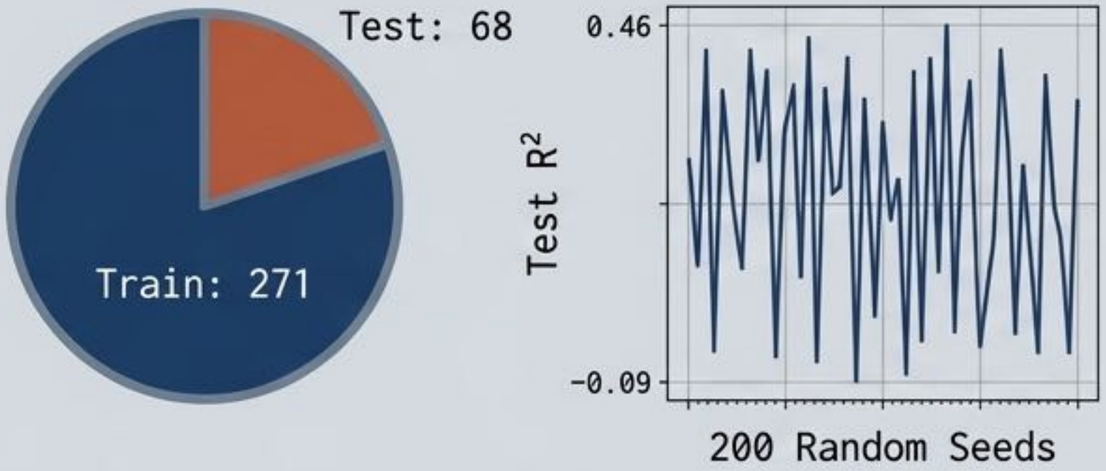
Step 3: Random Forest



Decorrelation. Feature Subsetting ($\sqrt{64} = 8$). Prevents dominant dimensions from masking weaker spatial signals. Handles satellite multicollinearity.

Evaluation Strategies Matrix: Small N Environments

Single Train/Test Split (80/20)



- **Data Efficiency:** Poor. 68 points are wasted, never helping the model learn.
- **Variance of Estimate:** Extreme. The score depends entirely on the 'luck of the draw.'
- **Risk of Cherry-Picking:** High. A researcher could easily publish R^2 = 0.40 or 0.05 based on the seed.

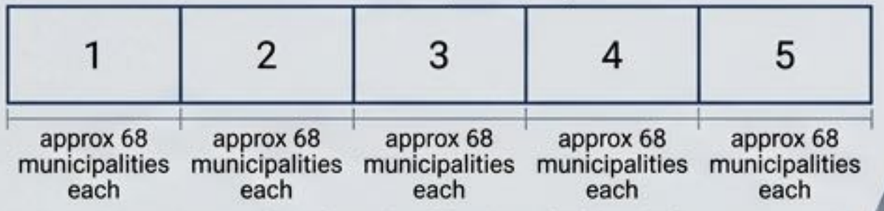
5-Fold Cross-Validation

- **Data Efficiency:** Optimal. 100% of the data is used for both training and testing.
- **Variance of Estimate:** Stable. Averages out the geographic anomalies of the Bolivian landscape.
- **Verdict for this Study:** Mandatory. With only N=339, cross-validation is required for an honest performance claim.

Implementation Pipeline: 5-Fold Cross-Validation

Step 1: Shuffle & Split

Dataset (N=339)



Data is randomly shuffled and divided into k=5 distinct, non-overlapping folds.

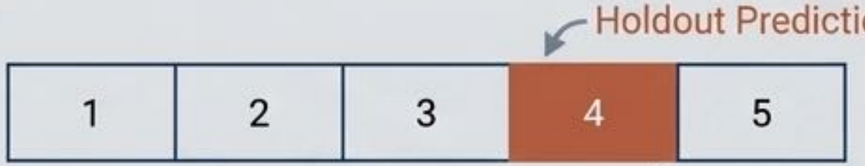
Step 2: Train on k-1


```
model.fit(X_train, y_train)
```

The model is trained using data from k-1 folds, excluding the current holdout set.

The Out-of-Fold (OOF) Prediction Engine

Step 3: Predict Holdout




```
model.predict(X_test)
```

Municipality X receives an honest prediction from a model that has never seen it.

Step 4: Rotate & Repeat

Repeats 5 times, changing the holdout fold each iteration. Each municipality is used as a holdout exactly once.

The Output Vector

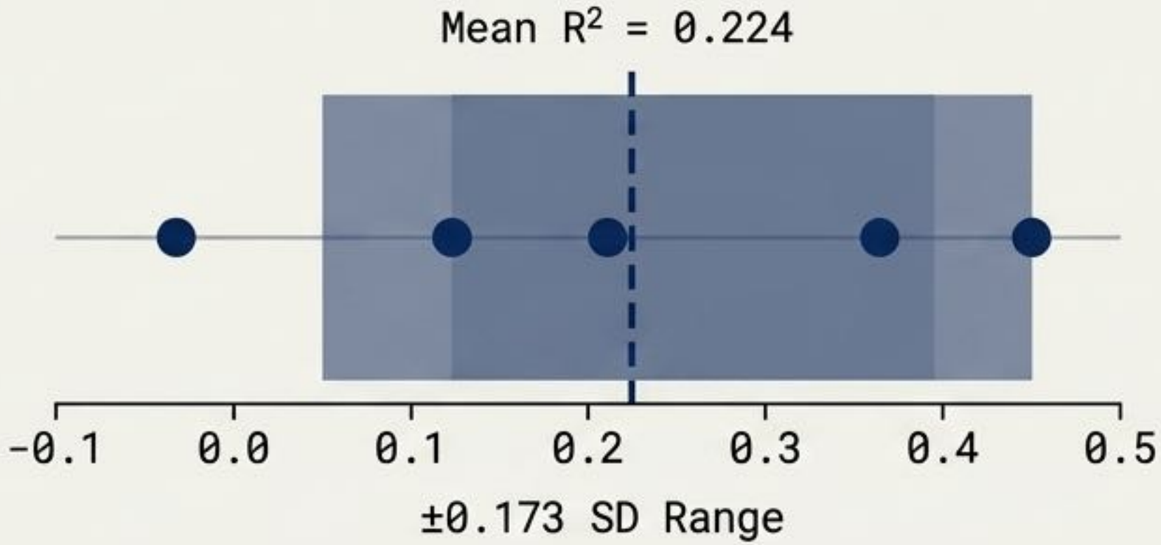
An N=339 vector of honest Out-of-Fold predictions, completely free of data leakage. This vector is used for final performance evaluation.

Interpreting Results I: The Importance of the Spread

The 5-Fold Evaluation Table

Fold	n	R^2	RMSE	MAE
1	68	0.209	6.61	4.73
2	68	0.121	7.34	5.05
3	68	-0.032	5.73	4.43
4	68	0.453	4.49	3.82
5	67	0.367	5.14	4.05
Mean		**0.224**	**5.86**	**4.42**
SD		**0.173**	**1.01**	**0.44**

The Variance Spread



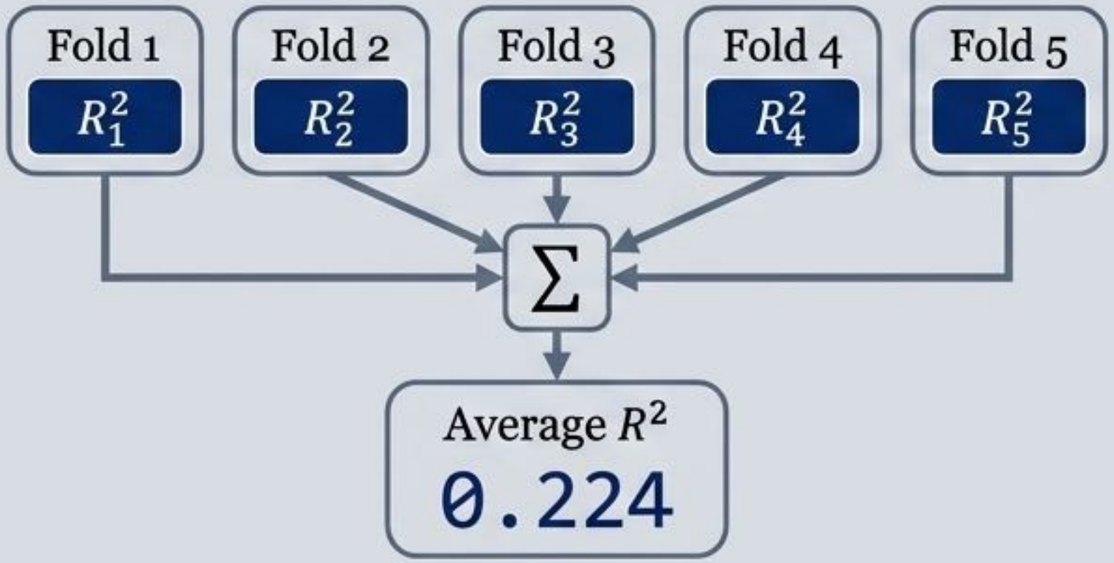
The Illusion of Certainty.

A mean R^2 of 0.224 is deceptive without its ±0.173 spread. The model's predictive power is highly unstable across different geographic slices. Fold 3 performs worse than guessing the national average.

Interpreting Results II: Pooled vs. Averaged Metrics

Averaged Per-Fold Metrics

$$\frac{1}{5}\sum R_k^2 = 0.224$$

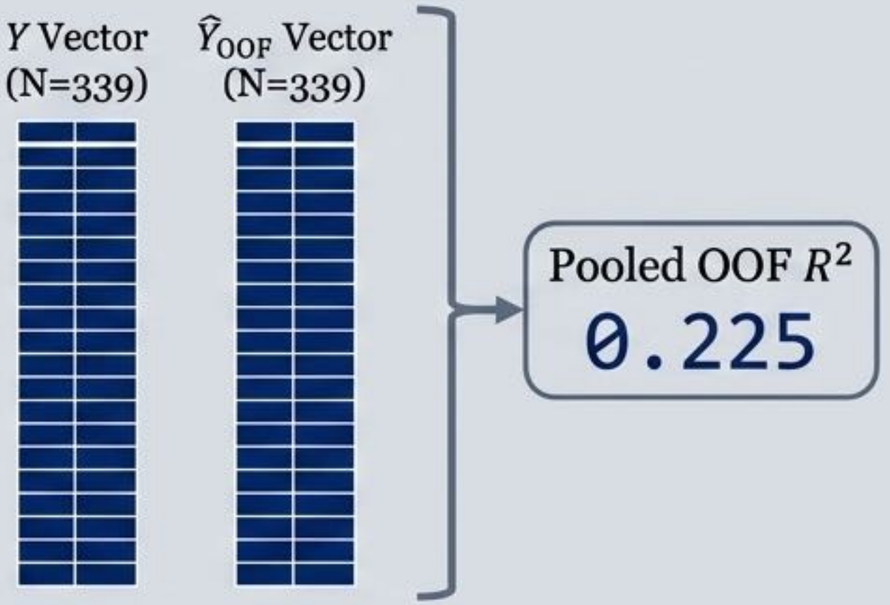


Econometric Use Case:

- Used to communicate overarching model performance alongside uncertainty (± Standard Deviation).
- It weights every *fold* equally.

Pooled Out-of-Fold (OOF) Metrics

$$R^2\left(Y, \hat{Y}_{\mathrm{OOF}}\right) = 0.225$$

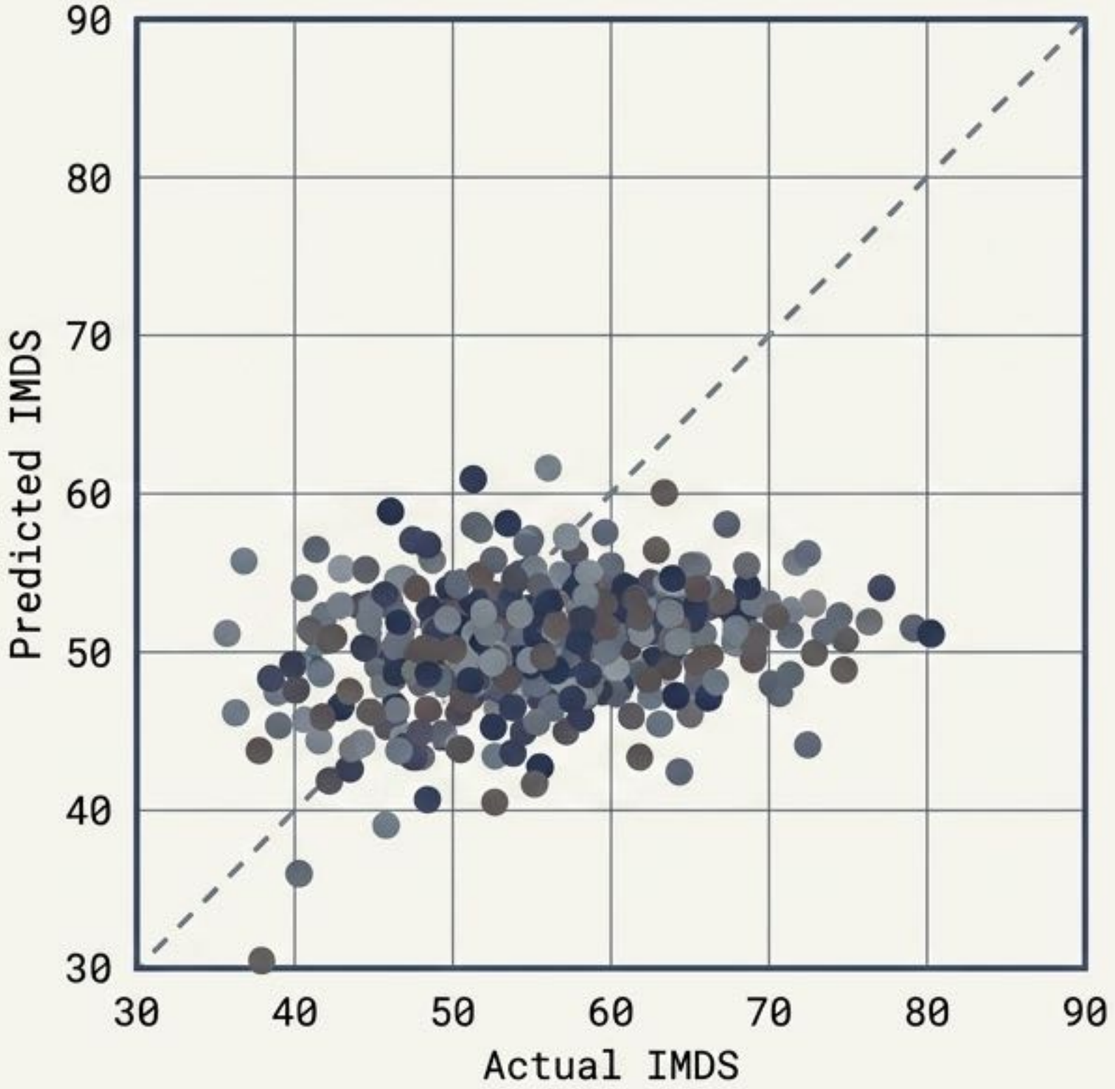


Econometric Use Case:

- Used for whole-dataset diagnostics, plotting, and residual analysis.
- It weights every *observation* equally.

While numerically similar here (0.224 vs 0.225), they mathematically diverge when fold sizes or difficulties vary. Distinguishing them is mandatory for replicable research.

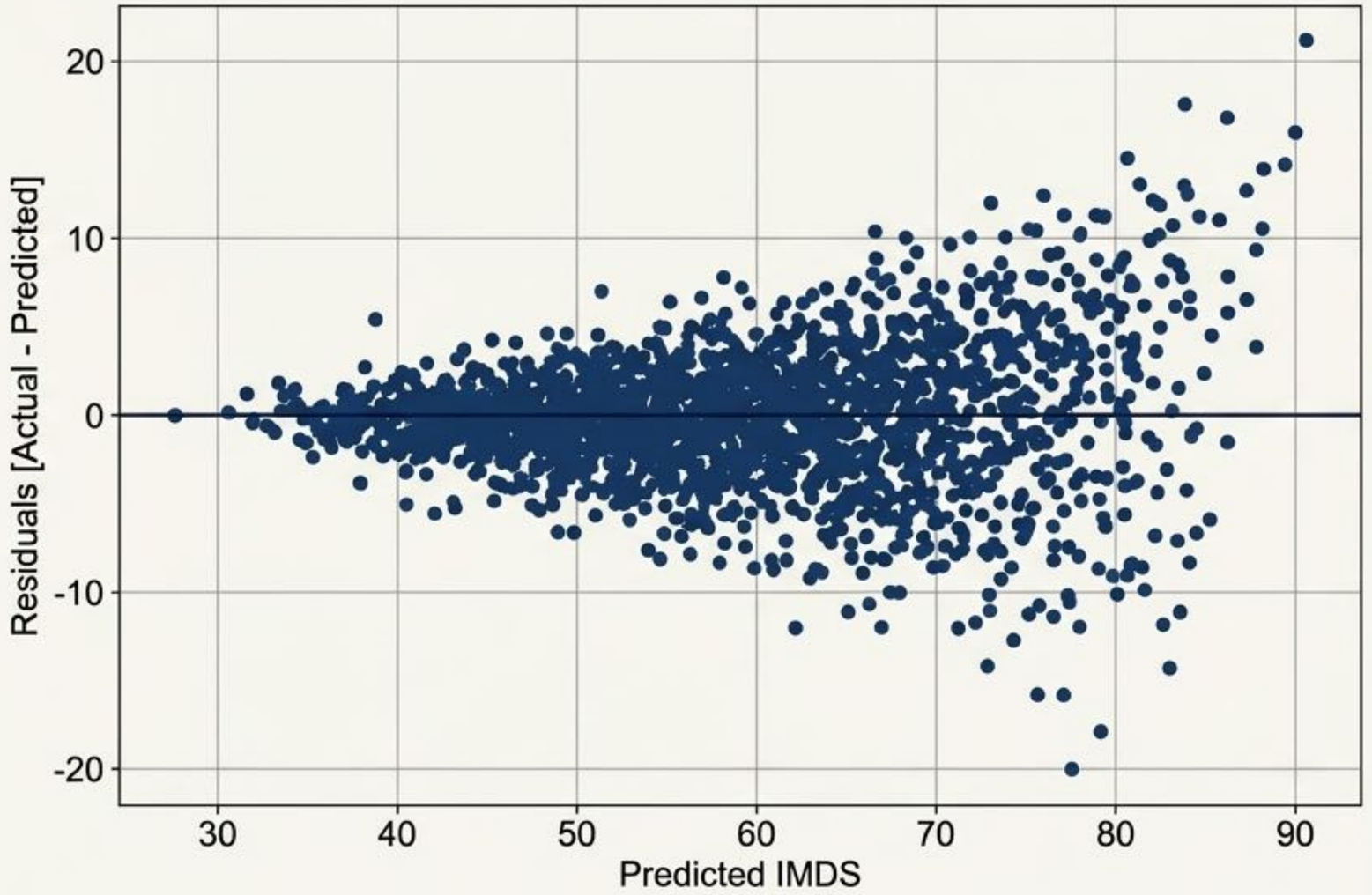
Diagnostic Analysis I: Regression to the Mean



Regression to the Mean

- **Observation:** The highest actual IMDS (80.2) is predicted at only ~51.
- **Economic Interpretation:** The model systematically over-predicts poor municipalities and under-predicts wealthy ones.
- Conclusion: This flattening is the statistical fingerprint of a model with limited predictive signal, causing it to hedge toward the safe national average.

Diagnostic Analysis II: Residual Heteroscedasticity

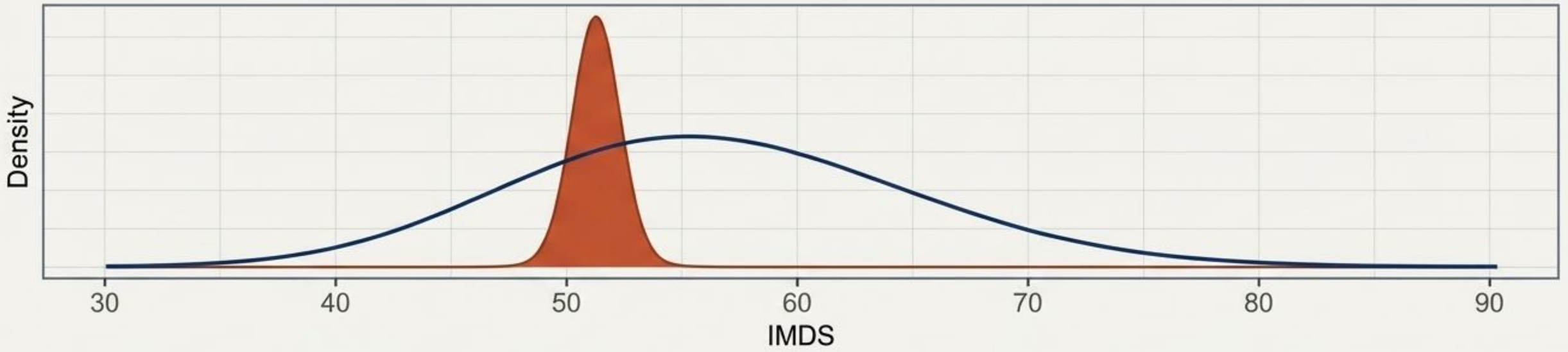


Systematic Residual Bias

- **Observation:** The largest positive residuals (errors) occur at the high-IMDS extremes.
- **Economic Interpretation:** The model structurally fails to capture the unique drivers of top-performing urban centers (e.g., La Paz, Santa Cruz).
- **Conclusion:** The drivers of their success are fundamentally invisible to 2017 satellite land-use embeddings alone.

Diagnostic Analysis III: Variance Compression

The Variance Compression Overlap



Summary Statistics

Statistic	Actual	Predicted (OOF)
Mean	51.05	51.02
Std. Dev	6.77	3.54
Min	35.70	40.66
Max	80.20	61.79

Diagnostic Callout

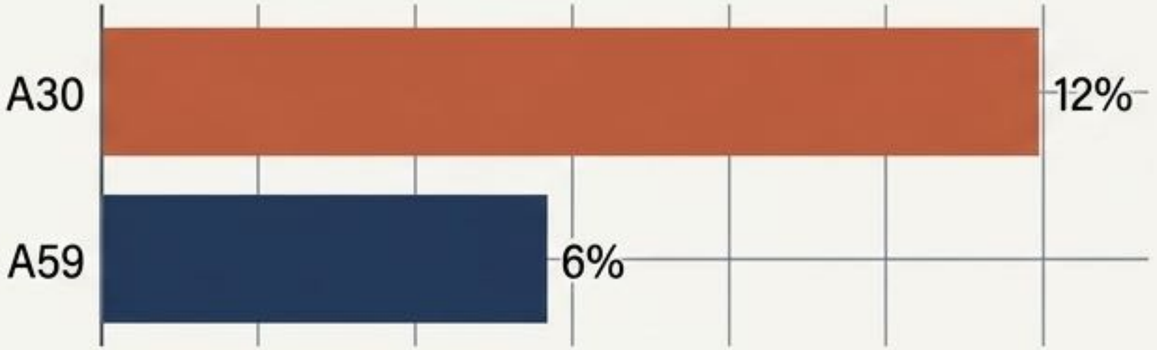
The Math: The means match perfectly (unbiased on average), but the predicted variance is compressed by 48%. A Kolmogorov-Smirnov test ($p < 0.001$) firmly rejects distribution equality.

The Econometric Reality: A model that explains only ~22% of variance *must* mathematically compress its predictions. A decent R^2 does not mean the predicted distribution mimics reality.

Feature Importance Diagnostics: Identifying the Signal

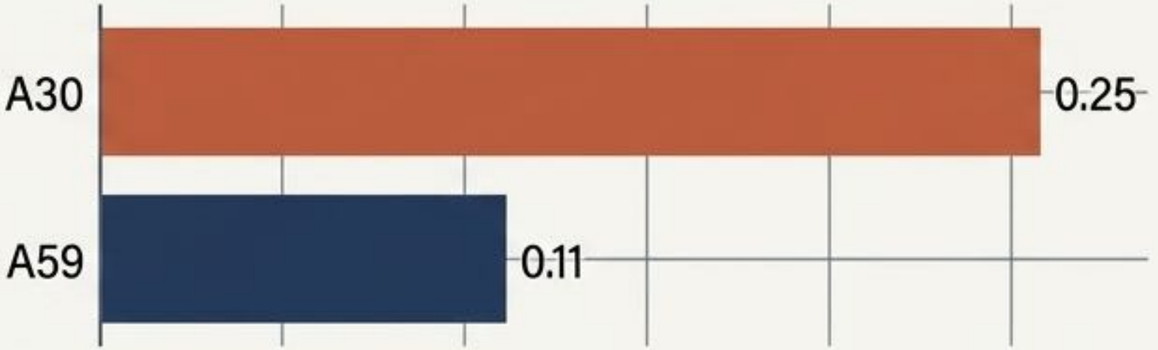
Mean Decrease in Impurity (MDI)

Mechanism: Measures impurity reduction across all splits. Built-in and fast.
Drawback: Inherently biased toward continuous/high-cardinality features.



Permutation Importance (The Gold Standard)

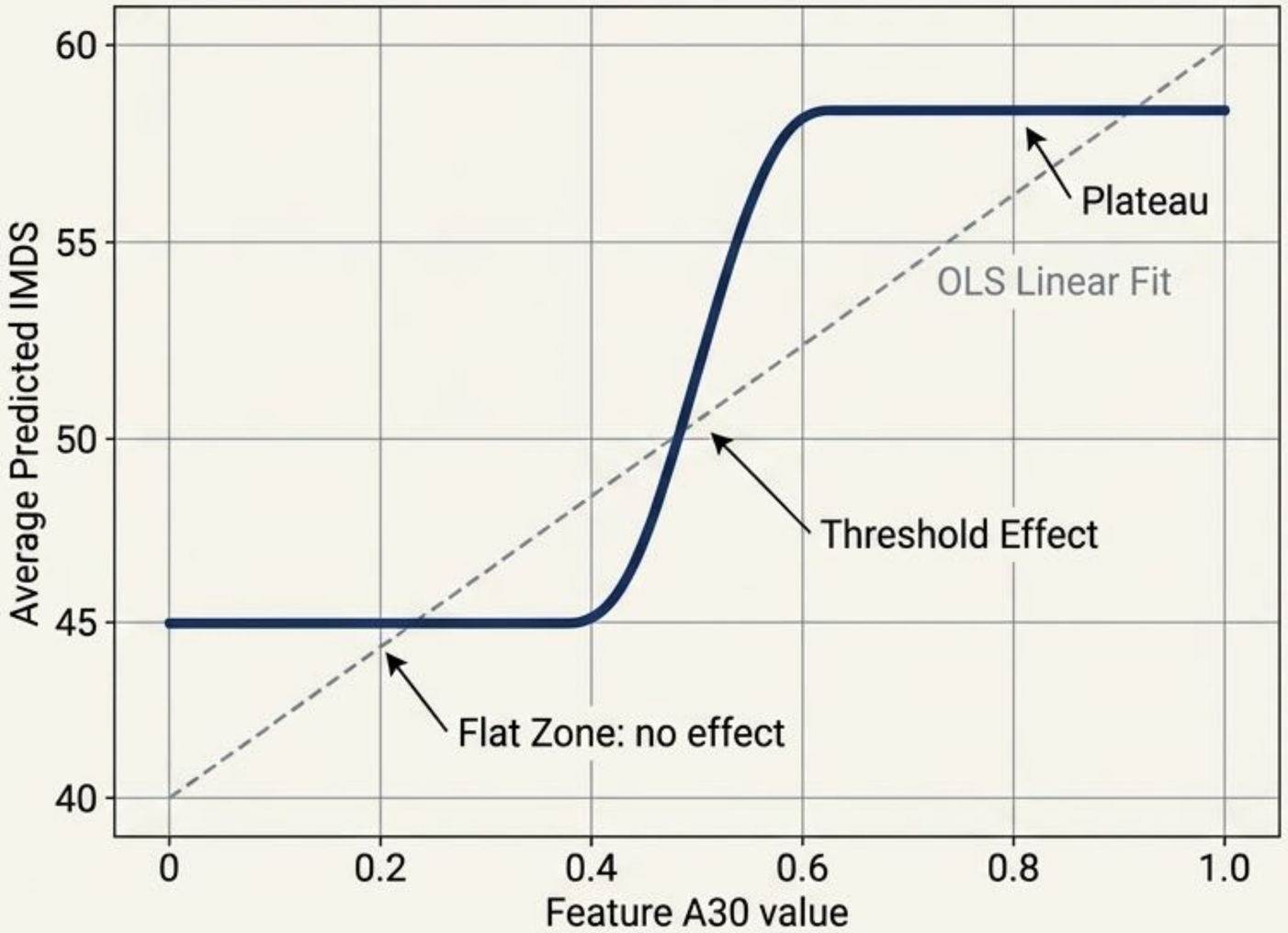
Mechanism: Scrambles a single column and measures the resulting drop in R^2.
Unbiased by scale or cardinality.



Verdict: Permutation confirms the MDI result. The predictive signal is highly concentrated. Shuffling just `A30` destroys the model entirely (dropping R^2 by more than the model's total 0.22 baseline).

Non-Linearity & Partial Dependence Plots (PDP)

Deconstructing the Partial Dependence Plot (PDP)



Theoretical Architecture

$$\bar{f}(x_k) = E_{X_{-k}}\left[\hat{f}(x_k, X_{-k})\right]$$

The average prediction as one feature varies, holding others at their empirical distribution.

Econometric Value:

This threshold effect perfectly justifies the use of Random Forest.

A standard linear regression (OLS) would force a straight line through this relationship, mis-specifying the spatial dynamics and averaging out the critical step-function behavior of the satellite embedding.

Methodological Synthesis: Hyperparameter Tuning

Hyperparameter Tuning Protocols

Grid Search	Random Search	Optuna (Bayesian/TPE)
Type: Exhaustive/Discrete **Mechanism**: Tries every listed combination. Blind to spaces between grid points.	**Type**: Sampled/Continuous **Mechanism**: Explores random combinations. Often more efficient than Grid.	**Type**: Smart/Probabilistic **Mechanism**: Builds a probabilistic model of promising regions to maximize yield.

The Reality Check (Yield in this Study)

`Untuned Baseline: R² = 0.224`

`Best Optuna Tuning: R² = 0.251`

The gain from the most advanced Bayesian tuning (+0.027) is drastically smaller than the cross-validation noise (± 0.173). When the signal is weak, hyperparameter tuning is rearranging deck chairs. The performance ceiling is dictated by the data (features), not the algorithm settings.

Methodological Synthesis & Future Research

Proven Limitations

- **Modest Signal:** An R^2 of ~22% proves satellite embeddings contain real, but insufficient, signal for mapping total development.
- **Variance Compression:** The collapse of the predicted distribution mathematically limits the model's utility for targeting extreme high/low policy interventions.
- **Abstract Interpretability:** Features like A30 and A59 are powerful mathematical drivers, but remain abstract geometric embeddings.
- **Temporal Mismatch:** Models rely on 2017 imagery predicting later SDG targets.

Next Steps for the Economist

Step 1: Data Fusion

Merge the 64-dimensional satellite embeddings with traditional administrative/survey data to break the 22% ceiling.

Step 2: Explainable AI

Employ SHAP (SHapley Additive exPlanations) values to bridge the gap between abstract embeddings and observable physical municipal traits.

Step 3: Panel Data Transition

Shift from cross-sectional outcome prediction to evaluating temporal changes (predicting Δ IMDS over time).