Transportation Noise and Urban Bird Richness: Baselines and Model Comparisons

Team member: Chris Ding, Zhuoxin Xia, Qinchuan Zhang, Chengwei Shen
AI for Conservation

Dec 2025

Abstract

Urban transportation noise is a pervasive stressor that can reduce bird occupancy, disrupt communication, and alter community composition. In this project we build a reproducible, spatially-explicit pipeline that combines citizen-science bird observations with a transportation noise raster, aggregates both to a common grid, and evaluates a set of baseline machine learning models for predicting mean bird species richness from noise. Our current feature set focuses on a single predictor (LAeq), which allows us to cleanly measure the strength and limitations of noise-only baselines. Across ten models, a degree-2 polynomial regression achieves the best test performance ($R^2 \approx 0.088$, RMSE ≈ 4.998), only slightly improving over a linear baseline ($R^2 \approx 0.077$, RMSE ≈ 5.029). More flexible models (e.g., random forests and extra trees) generalize poorly with the current feature set, yielding negative R^2 on held-out data. These results indicate that noise alone explains only a modest fraction of richness variation at our grid scale, motivating richer ecological covariates and spatial validation.

1 Conservation Problem and Technical Problem

1.1 Conservation context

Transportation noise from roads, rail, and aviation can reduce habitat quality by masking acoustic signals, increasing stress, and shifting behavior. In dense metropolitan regions, even small reductions in noise can potentially yield measurable biodiversity benefits, but agencies need actionable guidance: where are the noisiest places that also have high ecological value, and where might mitigation yield the highest marginal benefit?

1.2 Technical problem statement

We treat the task as supervised regression at a fixed spatial resolution. Given a grid cell with a measured transportation noise level (LAeq) and associated bird observation-derived richness, we predict mean species richness. The near-term technical goal is to establish strong, reproducible baselines and diagnose whether noise alone can explain meaningful variation. The longer-term goal is to extend the feature set (e.g., vegetation indices, land cover, distance to water, human density) and to produce maps and scenarios that can support screening and prioritization for conservation planning.

2 Data Summary and Key Findings

2.1 Data sources

- **Bird observations:** eBird Basic Dataset (EBD) sampling and observation files, restricted to a study region and time window.
- **Noise:** A transportation noise raster (LAeq) for the study area (roads/rail/aviation combined), sampled at grid-cell locations.

We cite the EBD release used in our pipeline as:

> eBird Basic Dataset. Version: EBD_relOct-2025. Cornell Lab of Ornithology, Ithaca, New York. Oct 2025.

2.2 Preprocessing and spatial aggregation

Our pipeline filters eBird sampling events to the NYC study bounding box (latitude $[40.45, 41.05]$ and longitude $[-74.30, -73.65]$) and to the years 2010–2024. We link each sampling event (checklist) to its set of reported species using the sampling event identifier, and compute per-checklist richness as the number of unique `COMMON NAME` values reported on that checklist. We then aggregate checklists into a regular **500 m** $\times$ **500 m** grid by projecting points to UTM 18N (EPSG:32618) and assigning each checklist to a cell using $\lfloor x/500 \rfloor$ and $\lfloor y/500 \rfloor$. For each grid cell we compute summary statistics including: (i) mean richness across checklists in the cell, (ii) median richness, and (iii) the number of contributing checklists ($n_{\text{checklists}}$). We also store a representative cell location as the **mean checklist latitude/longitude** within the cell. Next, we sample the transportation noise raster at this representative location (after transforming coordinates from EPSG:4326 to the raster CRS) using point sampling. We treat raster `nodata` values and extreme values (absolute value > 200) as missing, and then retain only grid cells with valid noise values and adequate sampling ($n_{\mathbf{checklists}} \geq 5$).

2.3 Exploratory findings

The noise-only relationship is negative in our fitted response curves (Figure 3), consistent with the ecological hypothesis that higher noise is associated with reduced richness. However, the scatter and the True-vs-Pred plots (Figure 2) show substantial unexplained variance and strong regression-to-the-mean: models tend to predict richness in a narrow band, underpredicting high-richness cells and overpredicting low-richness cells. This motivates (a) additional predictors and (b) validation strategies that account for spatial dependence.

3 Problem Formulation

Let each grid cell be indexed by $i = 1, \ldots, n$ with feature vector $x_i \in R^d$ and target $y_i \in R$. In the current baseline, $d = 1$ with $x_i = \text{LAeq}_i$ and $y_i = \overline{\text{richness}}_i$ (mean richness in the cell). We learn a function f_θ to minimize prediction error:

$$\hat{y}_i = f_\theta(x_i), \qquad \theta^\star = \arg\min_\theta \sum_{i \in \mathcal{T}} \ell\big(y_i, f_\theta(x_i)\big),$$

where $\mathcal{T}$ is the training set and ℓ is squared loss. We evaluate generalization using:

$$R^2 = 1 - \frac{\sum_{i \in \mathcal{S}} (y_i - \hat{y}_i)^2}{\sum_{i \in \mathcal{S}} (y_i - \bar{y}_{\mathcal{S}})^2}, \qquad \text{RMSE} = \sqrt{\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} (y_i - \hat{y}_i)^2},$$

on a held-out test set $\mathcal{S}$.

4 AI Techniques Developed and Used

We implement and compare a small model suite that spans linear, nonlinear parametric, local instance-based, kernel-based, and tree-ensemble methods:

- **Linear regression** (global linear baseline).
- **Polynomial regression** with degree 2 (global curved response in LAeq).
- **Ridge/Lasso/ElasticNet** (regularized linear variants).
- **KNN regression** (local averaging in feature space, with standardized input).

- **SVR with RBF kernel** (nonlinear kernel method, with standardized input).
- **Random Forest / Extra Trees / Gradient Boosted Trees** (tree ensembles).

All models are trained with standard scikit-learn pipelines; models that are sensitive to feature scale use standardization.

5 Training, Validation, and Testing

We use an 80/20 random train/test split with a fixed seed for reproducibility, and additionally report 5-fold cross-validation (CV) averages. CV is performed on the full dataset with shuffling enabled to estimate variability under random partitioning.

Limitations of random splits. Because grid cells are spatially structured, random splits can leak information through spatial autocorrelation. As a result, random-split metrics may be optimistic compared to spatial holdout. In future iterations we will implement spatial block cross-validation to better measure out-of-region generalization.

6 Results

6.1 Quantitative comparison

Table 1 summarizes test-set and CV performance for all evaluated models. The best-performing model is the degree-2 polynomial regression, which modestly improves over the linear baseline. Tree ensembles perform poorly with the current feature set, yielding negative R^2 on the test split, indicating worse performance than predicting the test mean.

Table 1: Model performance predicting mean cell richness from LAeq only.

Model	Type	R^2_{test}	$\text{RMSE}_{\text{test}}$	R^2_{CV}	RMSE_{CV}
Polynomial (deg=2)	Global curved	0.087737	4.998467	0.079037	4.848492
Linear	Global linear	0.076666	5.028706	0.075288	4.858485
Ridge	Linear regularized	0.076663	5.028716	0.075289	4.858484
Lasso	Linear regularized	0.076618	5.028837	0.075292	4.858487
ElasticNet	Linear regularized	0.076608	5.028866	0.075294	4.858486
Gradient Boosted Trees	Tree ensemble	0.031825	5.149368	0.014967	5.015113
KNN (k=15)	Local instance	0.017611	5.187030	-0.004382	5.063074
SVR (RBF)	Kernel method	-0.038128	5.332151	-0.026353	5.119255
Random Forest	Tree ensemble	-0.478099	6.362513	-0.401533	5.980846
Extra Trees	Tree ensemble	-0.665023	6.752849	-0.649511	6.485486

6.2 Visual diagnostics

Figure 1 shows performance bars for all models. Figure 2 shows that predictions are compressed relative to the true range, especially for high-richness cells. Figure 3 shows fitted response curves, suggesting a generally negative relationship between LAeq and richness, with curvature at high LAeq likely driven by sparsity and outliers.

7 Technical Discussion

7.1 Why do simple models win here?

With only one feature (LAeq), model capacity is not the limiting factor—signal-to-noise is. The polynomial model adds a small amount of flexibility and slightly improves fit, but the gains are marginal. Regularized linear models (ridge/lasso/elastic net) are nearly identical to ordinary linear regression, suggesting that the linear baseline is already stable.

7.2 Why do tree ensembles underperform?

In this setting, tree ensembles can overfit noise or leverage accidental structure in the training split without learning a stable monotonic relationship. Negative R^2 indicates that the model's prediction error exceeds the error of predicting the test mean. This is consistent with the True-vs-Pred plots, where predictions collapse into a narrow band and miss high-richness cells.

7.3 What is missing from the feature set?

Richness is influenced by habitat (vegetation, water, land cover), seasonality, observer effort, and sampling bias. Noise is only one stressor, and at our aggregation scale it may be correlated with other confounders (e.g., urbanization) in complex ways. Adding ecological and anthropogenic covariates, plus explicit season/time controls, is likely necessary to substantially improve predictive performance.

8 How the Solution Could Be Used for Conservation

8.1 Intended use: screening, not final decisions

We position predictions as a *screening tool* to help stakeholders prioritize where to look, not a replacement for field surveys. A practical workflow is: (1) generate a map of predicted richness and/or residuals, (2) identify noisy areas that are predicted to support relatively high richness (potentially high value but at risk), and (3) evaluate noise-reduction scenarios (e.g., barrier placement, rail lubrication programs, flight path adjustments) for potential benefits.

8.2 Is it accurate enough?

Not yet, for fine-grained decision making. With LAeq alone, R^2 is low, indicating large uncertainty in cell-level predictions. For operational use, we would add features, estimate uncertainty (e.g., bootstrap intervals), and validate using spatial holdouts.

8.3 Human guidance and safeguards

Human expertise is essential in three places:

- **Data QA/QC:** verify that high-richness predictions reflect plausible habitats rather than sampling artifacts.
- **Interpretation:** ensure stakeholders understand that results are associative, not causal.
- **Sensitive species:** aggregate or mask locations when rare or sensitive species could be exposed.

9 How This Challenges the State of the Art in AI

This problem stresses several common weaknesses in applied AI:

- **Observational bias:** citizen-science data is uneven in space/time and effort, which can dominate ecological signal.

- **Spatial generalization:** models must work across neighborhoods with different land cover and sampling intensity.
- **Causal ambiguity:** noise correlates with many urban factors; naive models can learn confounded associations.
- **Trust and transparency:** stakeholders need interpretable outputs, uncertainty estimates, and clear limitations.

Addressing these requires careful validation (especially spatial CV), uncertainty communication, and human-in-the-loop review.

References

[1] U.S. Department of Transportation, Bureau of Transportation Statistics. *DOT National Transportation Noise Map (2020)*. Available at: https://www.bts.gov/geospatial/national-transportation-noise-map.

[2] Cartwright, M., Mendez, A. E., Cramer, J., et al. *SONYC-UST v2: An Urban Sound Tagging Dataset with Spatiotemporal Context*. Zenodo (v2 annotations record). Available at: https://zenodo.org/records/3966543.

[3] Cornell Lab of Ornithology. *eBird Basic Dataset (EBD)*. Available at: https://science.ebird.org/en/use-ebird-data/download-ebird-data-products.

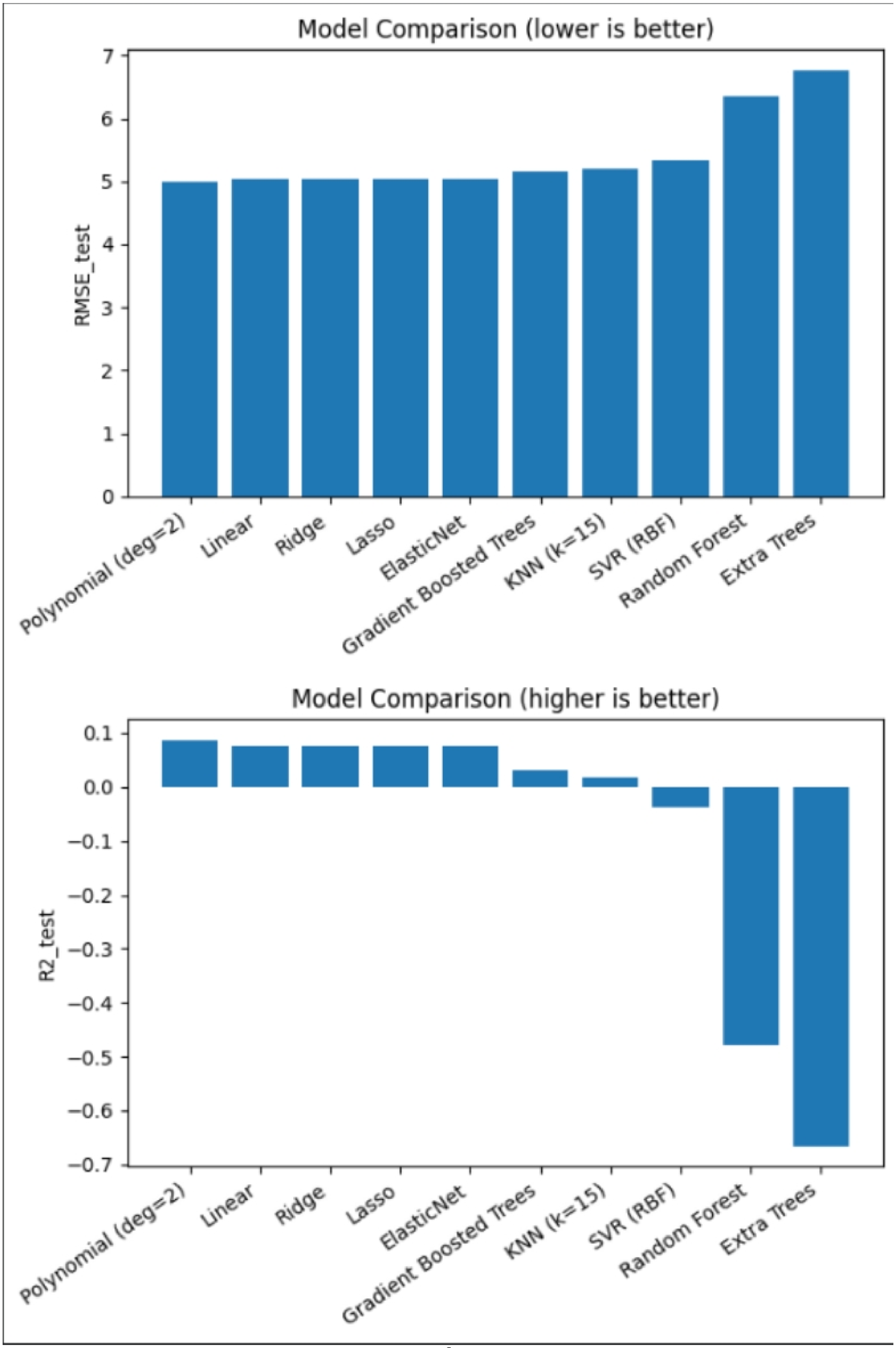
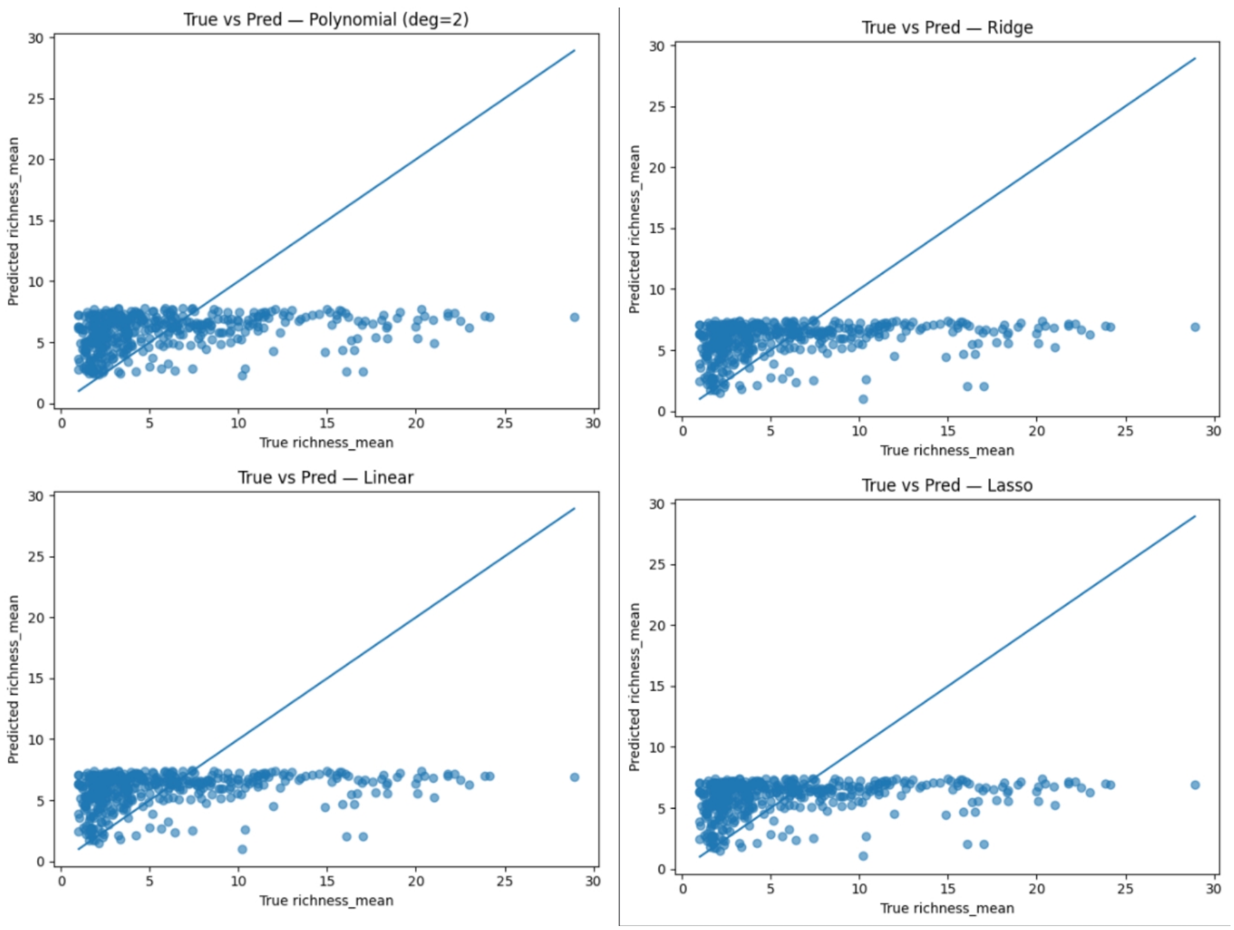


Figure 1: Model comparison using RMSE (lower is better) and R^2 (higher is better) on the test split.



(a) Polynomial vs. Linear

(b) Ridge vs. Lasso

Figure 2: True vs. predicted mean richness for representative baselines. The $y = x$ line indicates perfect prediction.

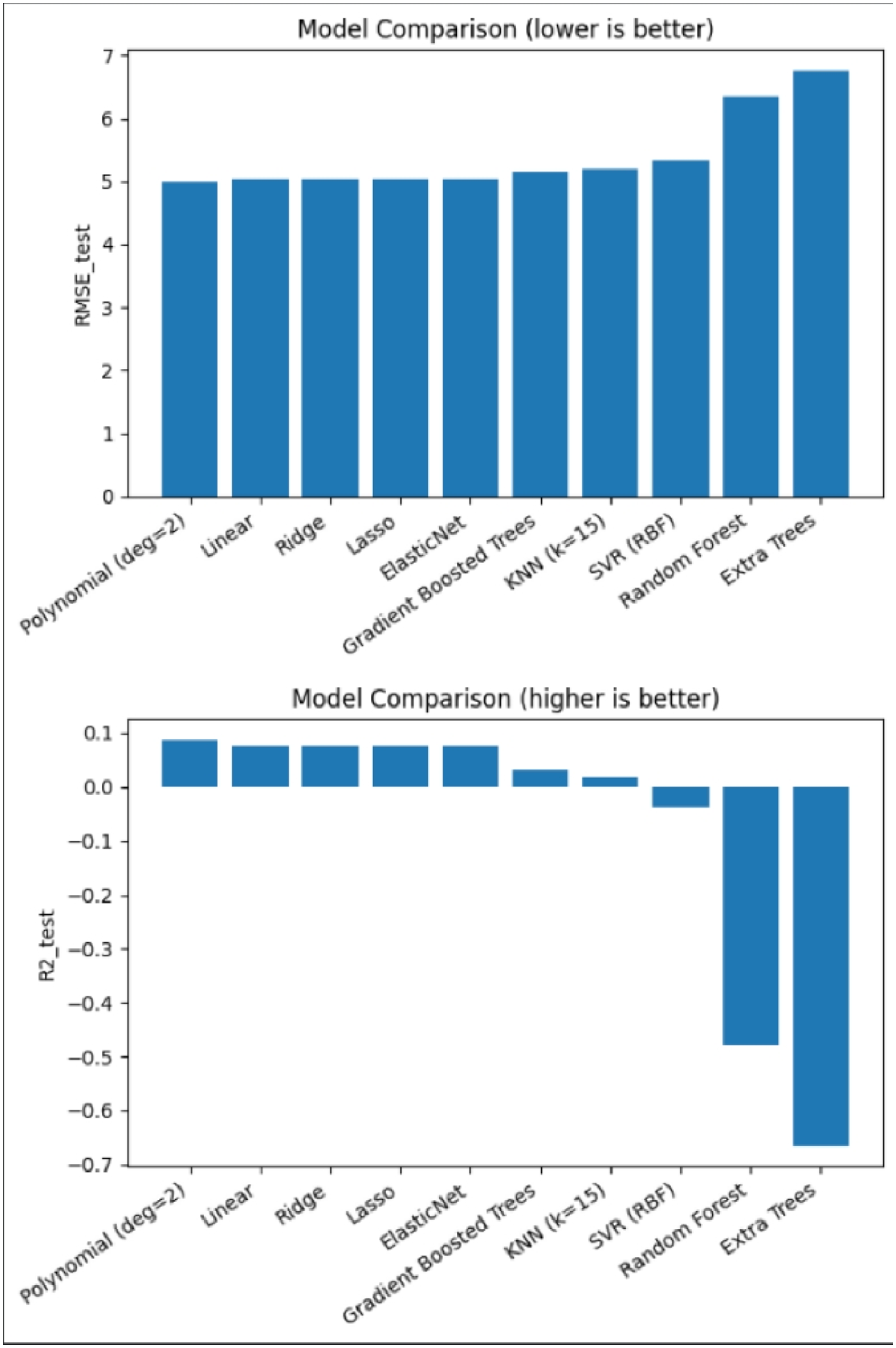


Figure 3: Noise-only response curves (LAeq → predicted richness) for several baselines.