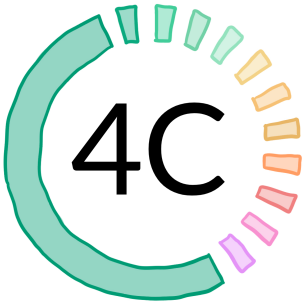
# Quantifying Species-Specific Responses to Hunting Pressure

**MRes Student:** Emilio Luz-Ricca
**Supervised by:** Andrew Balmford & Tom Swinfield

**Word count:** 6,089

---

**MRes Project Report**
**AI4ER CDT**
Department of Earth Sciences
University of Cambridge


# Abstract

We are currently experiencing an unprecedented loss of terrestrial biodiversity, particularly in the species-rich tropics. Species extinctions are primarily driven by loss of habitat, which is relatively easy to monitor by satellite remote sensing; other anthropogenic threats to biodiversity, like hunting, are much more difficult to observe directly. Further, little is known about how the local abundance of animal populations responds to hunting pressure. Recent studies have applied predictive modelling using statistical methods to implicitly assess hunting pressure through collections of tropical mammal and bird abundance responses. However, few predictive models have been considered to date and a thorough assessment of model generalisability is needed. Building on these studies, I present a comprehensive assessment of approaches for this predictive task. In particular, I reproduced the previous state-of-the art (a mixed-effects generalised linear hurdle model), thoroughly tested (nonlinear) predictive methods through application of automated machine learning, experimented with embeddings from pre-trained deep learning models as a supplement to hand-chosen predictors, and closely inspected spatial and taxonomic generalisability using cross-validation. I found that nonlinear hurdle models tend to outperform the existing mixed-effects linear hurdle model baseline, especially when random effects are excluded during prediction. Deep learning embeddings were largely unhelpful as supplemental predictors, but could be used to reliably predict hunting pressure on their own if used with the nonlinear hurdle model. Finally, spatial and taxonomic generalisation remained very difficult for all models tested, but improved in the presence of more training data. Through this work, I advance the state-of-the-art for predicting species-specific abundance responses to hunting pressure in the tropics and provide well-documented, reproducible code to support further predictive benchmarking for this task.


*This report is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text and/or bibliography.*

# 1 Introduction

We are experiencing an unprecedented loss of terrestrial biodiversity globally, particularly in the species-rich tropics [1, 2]. Species extinctions are primarily driven by loss of habitat, which is relatively easy to monitor by, e.g., satellite remote sensing. For instance, the recently proposed Land-cover change Impacts on Future Extinctions (LIFE) metric assesses a species's current Area of Habitat (AoH) relative to their theoretical AoH in the absence of human interference to estimate the impact of land-use changes on a species probability of extinction [3]. Other anthropogenic threats to biodiversity (e.g., fragmentation, edge effects, hunting, invasive species) can be thought of as forms of habitat degradation, thus reducing the effective AoH available to a species. However, most anthropogenic threats are very difficult to observe directly and are heterogeneously distributed. One threat to which many terrestrial vertebrate species are subject is hunting. While hunting is a fundamental human practice, recent patterns of overexploitation have substantially affected animal populations [4]. However, hunting is a cryptic activity and is difficult to track; unlike habitat loss, we cannot directly detect hunting using satellite remote sensing. Further, little is known about how the local abundance of animal populations responds to hunting pressure [5, 6].

Previous work towards quantifying hunting pressure has relied on accessibility maps (e.g., distance to nearest settlement) [7, 8], suites of indicators [9], or IUCN assessments [10]. While the threat maps produced have undoubtedly been valuable for guiding conservation action, the methods applied do not provide sufficient flexibility to capture the complex dynamics underlying the choice to hunt or the properties that govern population responses. Several studies have instead taken a data-driven approach, relying on statistical methods to implicitly assess hunting pressure through collected studies of tropical mammal [5] and bird [6] abundance responses. [5] was the first to show the potential for estimating hunting pressure from satellite-observable variables (e.g., human population density), global data products (e.g., travel time to major cities), and species traits (e.g., body mass); they accomplished this by gathering a comprehensive meta-dataset of georeferenced hunting studies for mammals and using generalised linear models (GLMs). [6] extended the methodology to birds using Bayesian GLMs. However, only one predictive model was considered across the two studies, providing little context for the reported predictive accuracy, and a thorough inspection of model performance under realistic generalisation scenarios was not provided.

Methods from Machine Learning (ML) have been adopted across the sciences as the state-of-the-art for predictive modelling [11, 12]. While classical statistical methods have proven useful for causal inference, many patterns in nature are not strictly linear and thus require more flexible predictive models capable of capturing dynamic, nonlinear relationships [13]. Use of ML has become popular in ecology for *purely predictive* tasks like species distribution modelling and wildlife monitoring (e.g., camera trapping or bioacoustics) [14, 15]. However, the improvement in predictive performance that ML provides comes at the cost of interpretability: the complicated, nonlinear structure of ML models makes it very difficult to understand the patterns they have picked up on [16]. To help address these concerns, ML practitioners have developed dedicated model validation techniques to assess predictive performance. Central to this area is the study of model behaviour under distribution shift, which most often arises in the ecological setting due to spatial or taxonomic model generalisation [17]. Predicting hunting pressure almost certainly displays these types of distribution shifts, since patterns of hunting and species responses are expected to vary substantially across species traits and local socioeconomic conditions, and are likely exacerbated by spatial and taxonomic biases in existing datasets [4, 5, 6].

Building performant ML models for bespoke tasks often requires substantial expert knowledge: it is difficult to predict which model or set of hyperparameters (i.e., parameters not fit using data) will work best for a given task [18]. Traditionally, a combination of intuition and manual trial-and-error experimentation has been used to select an appropriate model. Automated ML (AutoML) attempts to automate this process, performing both model selection and hyperparameter tuning with the goal of finding the best performing model using the least computational resources [18, 19]. Parallel to the development of AutoML for bespoke tasks is the development of large ML models called foundation models, which are trained to extract generally useful features (i.e., *embeddings*) from vast unlabelled datasets [12]. Foundation models use Deep Learning (DL), which constitutes a subset of ML models called neural networks; these models roughly mimic the structure of the human brain and tend to perform well on large, unstructured datasets [13]. Embeddings have been introduced in the ecological domain (e.g., species [20] or coordinate [21] embeddings) and for spatial data more generally (e.g., coordinate [22] or satellite imagery [23] embeddings), but their utility for bespoke downstream ecological tasks has yet to be thoroughly tested.

Following recent work towards quantifying mammal and bird defaunation in the tropics [5, 6], I

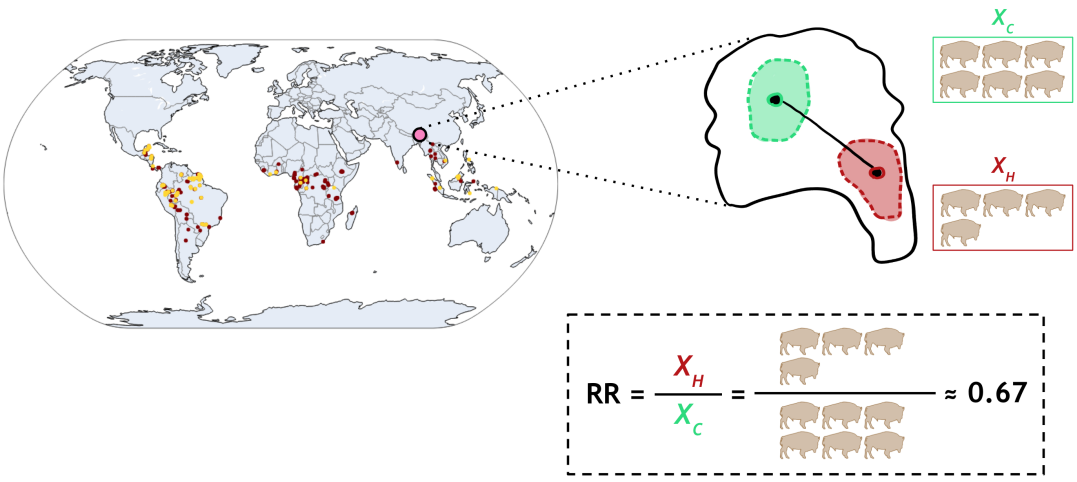



Figure 1: An example demonstrating how RRs are obtained. For a given study location (e.g., the pink point on the map), some measure of population abundance (or density) is obtained for an area with hunting ($X_H$, shown in red) and a similar area without hunting ($X_C$, shown in green). Then, the RR is calculated as the ratio $X_H/X_C$, interpreted as the loss of abundance in the hunted area that is directly attributable to hunting.

present a comprehensive assessment of approaches for this predictive task. In particular, I (1) reproduced the previous state-of-the art (a mixed-effects generalised linear hurdle model), (2) thoroughly tested (nonlinear) ML methods through application of AutoML, (3) experimented with embeddings from DL foundation models as a supplement to existing spatial and species predictors, and (4) closely inspected spatial and taxonomic generalisability. Finally, I provide well-documented, reproducible code to support further predictive benchmarking for this task.

# 2 Methods

## 2.1 Conceptual framework

To meaningfully predict species-specific responses to hunting pressure, we need both some measure of the effect of hunting on individual species (the *response*) and a suite of variables that determine the intensity of hunting pressure (the *predictors*).

I used a matched abundance ratio (or response ratio, RR) as the response variable; this is the ratio of some measure of abundance in a hunted study area ($X_H$) as compared to a similar unhunted (or lightly-hunted) control area ($X_C$). Mathematically, this is defined as RR $= X_H/X_C$, with RR $= 0$ indicating local extirpation, $0 \geq$ RR $\geq 1$ a local abundance decline, RR $= 1$ no abundance change, and RR $> 1$ a local abundance increase. This RR is interpreted as the local abundance change that is directly attributable to hunting, since confounding factors like habitat suitability and other anthropogenic threats are theoretically controlled for through site matching [4]. Predictor variables broadly fall into two categories: (1) *spatial* or (2) *species* predictors. These variables are expected to either indicate the level of hunting in a particular area, the extent to which a particular species is captured (either intentionally or through bycatch), or a species' (population) vulnerability to hunting.

## 2.2 Dataset description

I used the mammal meta-dataset of [5] (hereafter, the *mammals dataset*) and the bird meta-dataset of [6] (hereafter, the *birds dataset*), which each include pre-extracted predictors and corresponding RRs for a number of hunting studies across the tropics. Both datasets included the following spatial predictors: travel time to large cities [24] and distance to hunter access points, human population density [25], protection status [26], prevalence of stunting [27], and livestock biomass. The mammals dataset also included literacy rate (from the World Bank database; `https://data.worldbank.org/`), and the birds dataset included net primary productivity [28] and percent forest cover [29]. Species predictors included diet and body mass [30]. See [5] and [6] for a full description of each predictor variable.

The full dataset contains 4,765 records (i.e., RRs) in total. The mammals dataset contains 3,281 records from 163 studies and across 296 target groups, 271 of which are individual species and 25 are multiple species. The birds dataset contains 1,484 records from 55 studies and across 518 target groups. RRs are generally less than 5 (i.e., abundance in the hunted area is usually less than five times greater than abundance in the unhunted area), with only 126 records containing RR $> 5$; most records reflect abundance declines ($0 <$ RR $\leq 1$; 2,970 records) (Figure 2). Local extirpation was recorded in 12.7% of cases (RR$= 0$; 607 records) and, in the mammals dataset, 31.1% of records reflect no abundance change (RR $= 1$; 1,020 records). The low DI category is most common (2,381 records) across the two datasets, with roughly the same number samples for the medium and high DI categories (1,252 and 1,132 records, respectively). Studies cover all three tropical biogeographic realms (i.e., Neotropical, Afrotropical, and Indomalayan), but the Neotropical realm is the best represented (Figure 3). Brazil and Indonesia are the countries with the most records for the mammals and birds datasets, respectively.

In nearly all experiments, I model each dataset separately. However, I did test whether mixing datasets taxonomically (e.g., modelling birds and mammals jointly) confers performance benefits; results where the datasets were mixed will be clearly indicated.

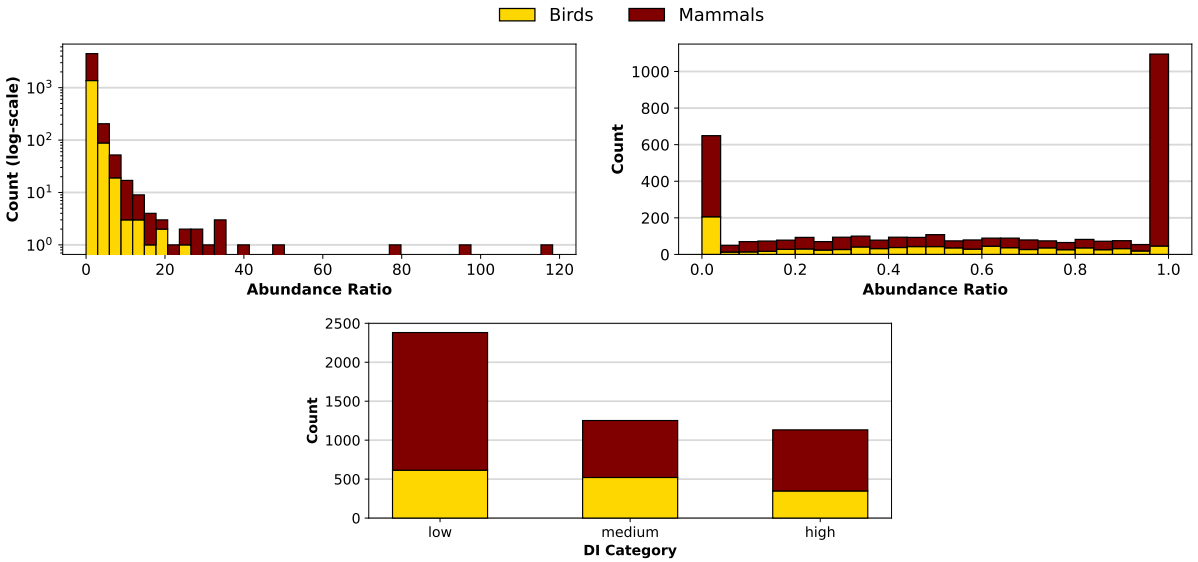



Figure 2: The distribution of abundance/response ratios in both datasets, across the full range (top left, log-scale) and for only abundance declines (top right). Shown also is the distribution of DI categories across both datasets (bottom).

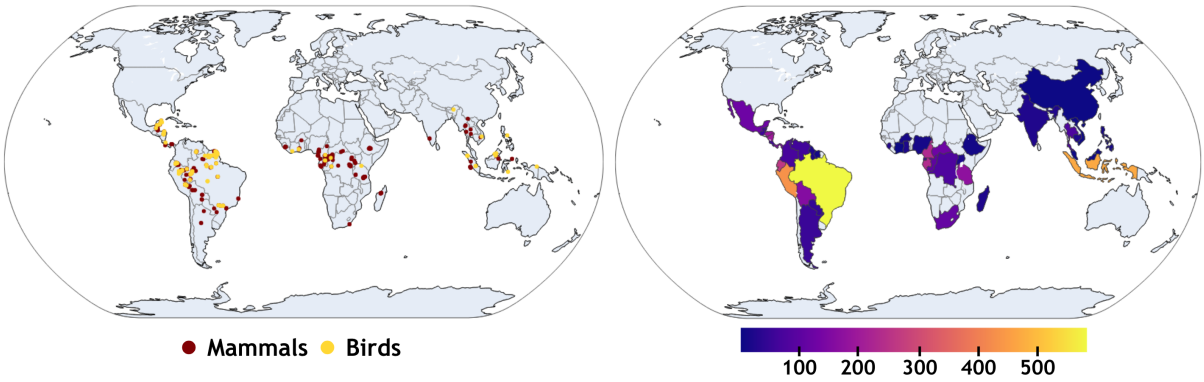



Figure 3: The location of study sites across the two datasets (left) and the number of records for each tropical country (right); countries with no records are indicated with grey shading.

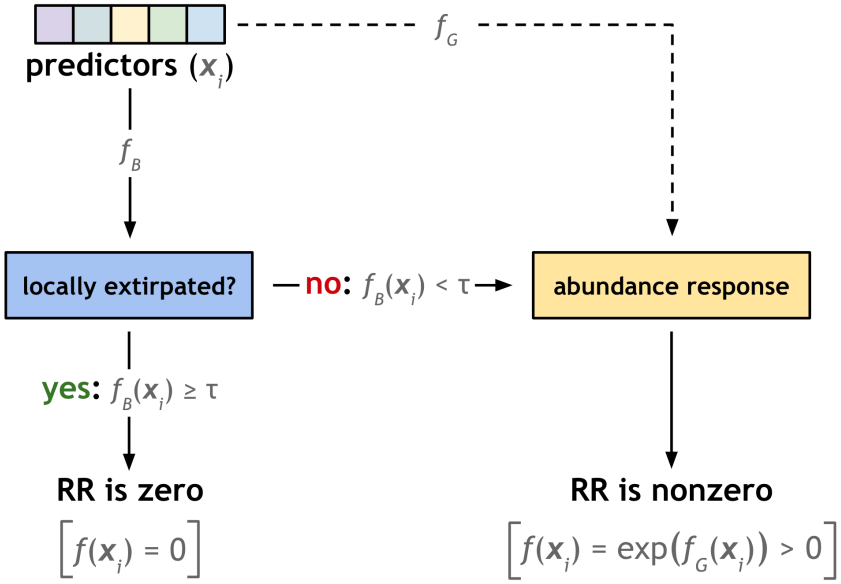



Figure 4: A visual representation of the functional structure of the hurdle model $f(\cdot)$. First, predictor variables are passed to a binary model $f_B(\cdot)$, which determines if the species will be locally extirpated at the query location; in practice, extirpation is predicted if the predicted probability is greater than some threshold $\tau$. If not, then the abundance response is regressed using the same predictors with a continuous model $f_G(\cdot)$, which predicts the log-transform of the RR.

### 2.3 Predictive models

To connect predictors to the response variable, I used methods from predictive modelling and ML. If we let the vector of $p$ predictors be $\mathbf{x}_i \in \mathcal{X} = \mathbb{R}^p$ and the corresponding RR is $y_i \in \mathcal{Y} = \mathbb{R}_{\geq 0}$, then the task is to learn a function mapping $f : \mathcal{X} \to \mathcal{Y}$ such that $f(\mathbf{x}_i) \approx y_i$.

As a predictive baseline, I replicated the previous state-of-the-art: a two-stage mixed-effects generalised linear hurdle model (henceforth, the *linear hurdle model*) [5]. This hurdle model is composed of two stages: first, we use a binomial generalised linear model (GLM) $f_B(\cdot)$ to determine the probability of extirpation at the chosen location and, if extirpation is sufficiently unlikely, we then use a Gaussian GLM $f_G(\cdot)$ to determine the abundance response (i.e., the RR). In practice, $f_G$ instead predicts the log-transform of the RR to ensure that model predictions are non-negative, since RR $\geq 0$. Therefore, the final function structure is

$$f(\mathbf{x}_i) = \begin{cases} 0 & f_B(\mathbf{x}_i) \geq \tau \\ \exp\left(f_G(\mathbf{x}_i)\right) & \text{otherwise} \end{cases}, \tag{1}$$

where $\tau$ is a chosen probability threshold (Figure 1). Continuous variables are included as fixed-effects predictors and the country, study, and species identity are used as grouping variables for random intercepts. While random intercepts can in theory be used during prediction, I report results both with and without the random-effects components to capture the change in performance when generalising to new groups and places. In general, I will refer to $f_B(\cdot)$ as the *zero model* and $f_G(\cdot)$ as the *nonzero model* to reflect that these can in principle be any arbitrary predictive model.

To efficiently test a wide range of nonlinear tabular ML methods, I applied AutoML. I experimented with direct classification into DI categories, direct regression of RRs, and regression via two-stage *nonlinear hurdle modelling* (i.e., both components of the model are nonlinear, but otherwise the model structure matches Equation 1). The models I included in the search space were: $k$-nearest neighbours, regularised logistic regression, and various decision tree ensembling methods. The latter class of models included Gradient Boosted Decision Trees (GBDTs [31]; LGBM [32] and XGBoost [33]), and bagged ensembles (random forests [34] and extra tree [35]). GBDTs are generally considered state-of-the-art for ML on tabular datasets [36]. To provide a lower bound on model performance, I also tested a method which I call the naïve regressor; this model ignores predictors and always predicts the mean RR from the training set, giving us a sense of how one would do when "guessing randomly."

## 2.4 Deep learning embeddings

To elucidate the utility of recently-introduced general-purpose DL embeddings, I experimented with SatCLIP [22] and BioCLIP [20] embeddings as supplements to hand-chosen spatial and species predictors, respectively. Both were trained with a contrastive multimodal CLIP objective [37]; SatCLIP distils information from satellite imagery into a representation of the latitude-longitude coordinates and BioCLIP distils images of a species into a representation of that species's taxonomic name. In downstream use, SatCLIP generates a (256-dimensional) vector representation for any given coordinate that is consistent with the local ground conditions (e.g., environmental, climatic, or socioeconomic) and BioCLIP generates a (512-dimensional) representation that captures species traits that are visually perceptible. Before using the embeddings as predictors, I reduced their dimensionality using principal component analysis, retaining only sufficient principle components to explain 90% of the embedding's variance. I tested two scenarios: using embeddings as supplements to the predictors of [5] or as the only predictors.

## 2.5 Evaluation metrics

For the most part, I assessed models through their ability to regress RRs. In cases, I look at the converse of the RR–called defaunation intensity (DI)–which is defined as $\text{DI}_i = 1 - \text{RR}_i$. When $0 \leq \text{RR} \leq 1$, the DI is interpreted as the percent decline in abundance due to hunting. Following [5], I discretised RRs into "high" ($\text{DI} \geq 0.7$), "medium" ($0.7 > \text{DI} > 0.1$), and "low" intensity ($0.1 \geq \text{DI}$). I inspected model balanced accuracy (BA) for classification into DI categories as well as two types of mean absolute error ($\text{MAE}_1$ and $\text{MAE}_\infty$) on regressed RRs. MAE is on the same scale as the response and gives us a sense of how "wrong," on average, a model's predicted RR is. On the other hand, BA is a percentage measure and demonstrates a model's ability to generally "sort" RR into correct DI categories; BA was used instead of accuracy because there is an imbalance in the representation of the three DI categories (Figure 2). See the Appendix for full details on the calculation of each metric.

## 2.6 Assessing spatial and taxonomic generalisability

To assess model generalisability across several dimensions, I employed $k$-fold cross-validation; $k = 5$ was used in all cases. I constructed folds either randomly or by using species or spatial blocking to test basic taxonomic and spatial generalisation, respectively (Figure 5). In the case of species blocking, all records for a particular species were included in a single fold so that a model would never be evaluated on species it was trained on. For spatial blocking, I used a $5 \times 5$ degree square grid and again placed all records in each grid cell into the same fold. I also tested model generalisability in two extreme scenarios: cross-continent and cross-class generalisation. For cross-continent generalisation, I held out all records from a single continent or biogeographic realm for evaluation (the former for the mammals dataset and the latter for the birds dataset), training on all other available records. For cross-class generalisation, I trained only on the mammals dataset and tested only on the birds dataset, and vice versa.

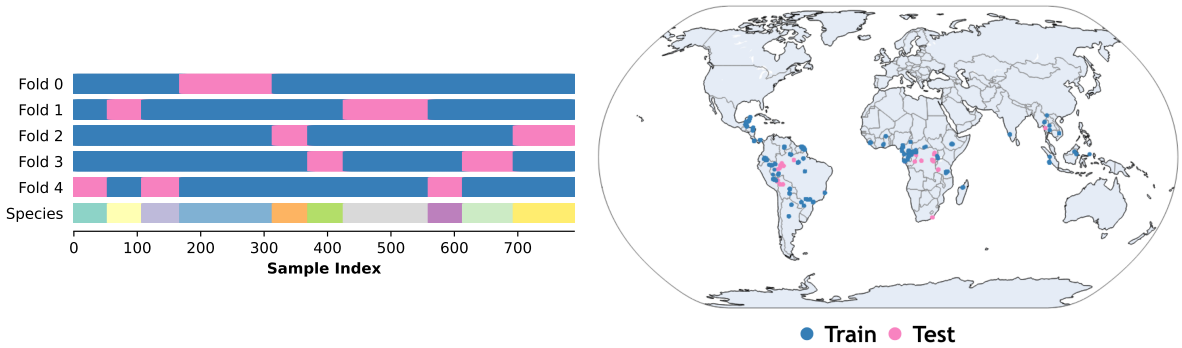



Figure 5: Example cross-validation sample splits for species (left) and spatial (right) blocking. Species blocking is shown for all splits for 5-fold cross-validation using only the ten most common species to improve readability. Spatial blocking was performed using a 5-degree regular grid; only one train/test split is shown. Train and test indices are indicated by blue and pink colouring, respectively, in both plots and species identity is indicated by the colour of the bottom bar in the species blocking plot.

## 2.7 Implementation details

Please refer to the Appendix for full implementation details for the reproduction of [5], the chosen AutoML settings, and the software packages used.

# 3 Results

## 3.1 Reproducing the state-of-the-art

When I fit the linear hurdle model as described, I found similar log-likelihood values as reported in [5]: $-947.0$ vs. $-928.6$ and $-3698.0$ vs. $-3701.5$ for mine vs. their zero and nonzero models, respectively. I also obtained very similar coefficient estimates for the fixed-effects predictors (Figure 6). The random 5-fold cross-validation achieved similar, but not precisely the same, metric values as reported in [5] (Figure 7); my reproduction displayed many of the same patterns, including high BA for the high DI category, very high specificity for the low DI category, and generally intermediate performance for the medium DI category. While the results that I obtained do not perfectly align with [5], I believe that they are similar enough to constitute a sufficient reproduction of their modelling approach.

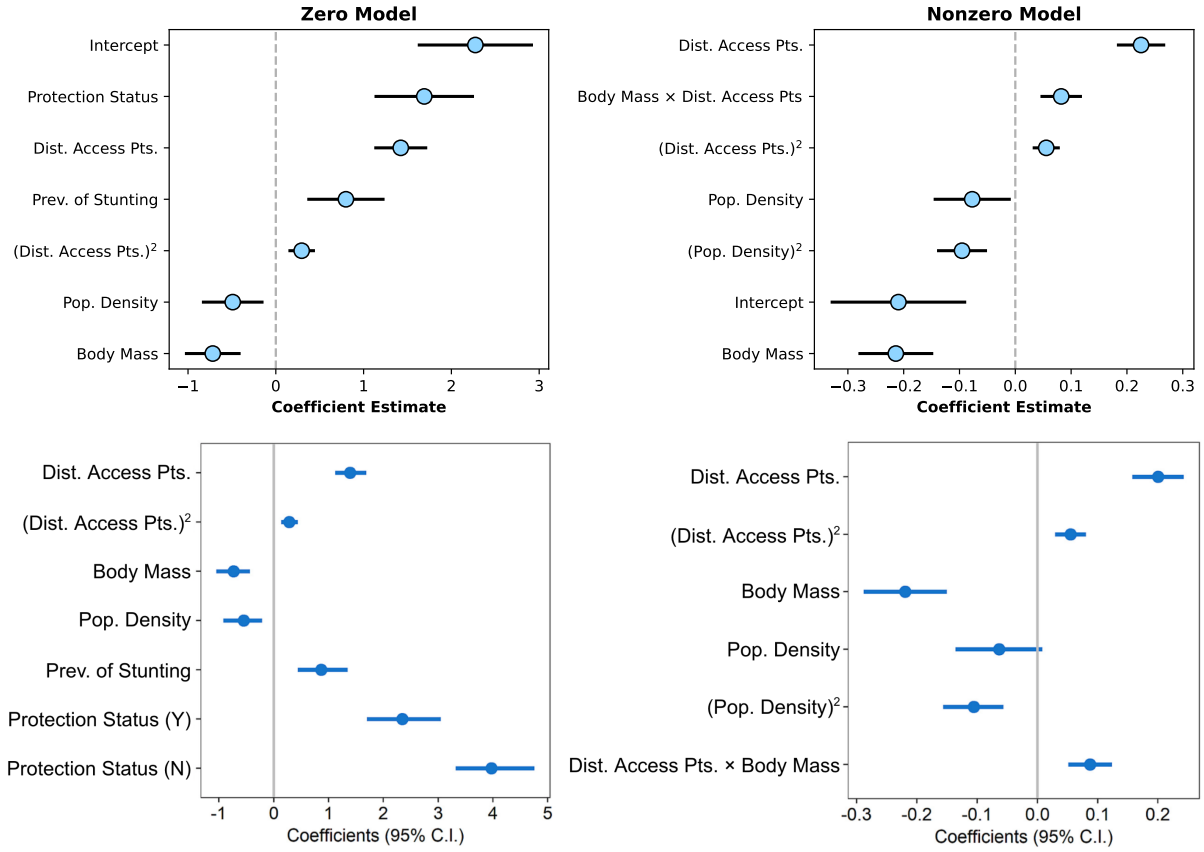



Figure 6: The coefficient estimates and 95% confidence intervals for the mixed-effects linear hurdle model. The response for the zero (binomial) model is binary (0 indicates locally extirpated, 1 indicates locally extant) and the response for the nonzero (Gaussian) model is the log-transform of the response ratio. All predictors were standardised using z-score scaling; "$(A)^2$" indicates a quadratic term for predictor $A$ and "$A \times B$" indicate a cross-term between predictors $A$ and $B$. The first row is my coefficient estimates and the second row is the coefficient estimates from [5]; I adapted the figure from the paper to align naming convention.

## 3.2 Model benchmarking

I find that hurdle models are generally more suitable than direct regression or classification models with respect to the chosen evaluation metrics (Figures 8 and 9). Direct regression is, in almost all cases,

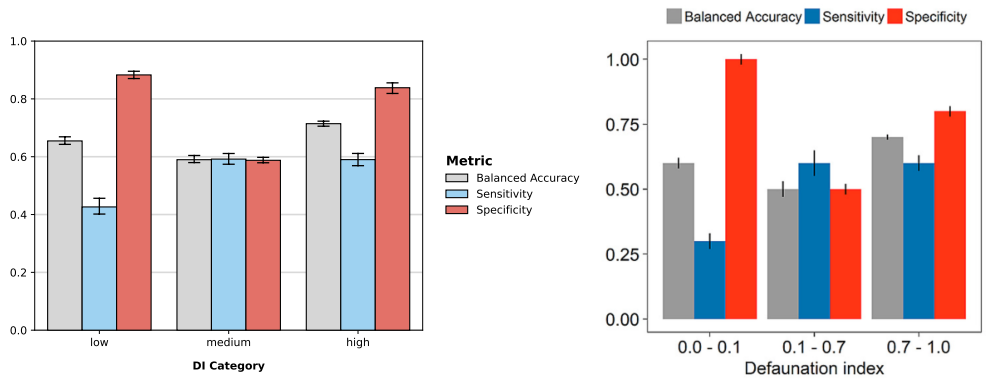



Figure 7: Random 5-fold cross-validation results for the mixed-effects linear hurdle model; metrics are presented for classification into the three DI categories. Average metric values are shown ± one standard deviation. To the left are my cross-validation results and to the right are the results from [5]; the figure is taken directly from the paper.

the worst performer for the metrics considered (often only slightly outperforming the naïve regressor) and, while direct classification into DI categories obtains the highest BA in almost all cases, RRs are a substantially more useful predictive target than DI categories. Nonlinear hurdle models tend to perform better than linear hurdle models, especially when random effects are not used during prediction. Nonlinear hurdles improve on linear hurdle most when train and test distributions match well (i.e., random cross-validation), with a 21% and 17% relative improvement in $\text{MAE}_1$ and $\text{MAE}_\infty$, respectively, over the linear hurdle (without random effects) for the mammals dataset and a $\approx 0\%$ and 13% improvement for the birds dataset.[1] Similar patterns hold for species and spatial blocking; in particular, the nonlinear hurdle tends to improve on $MAE_\infty$ by $\approx 5 - 10\%$ across both datasets (Figure 9). I inspect model predictive behaviour more closely in the Appendix (Figures A13 and A14).

Both the linear and nonlinear hurdle models substantially outperformed the naïve regressor (i.e., random guessing) for the mammals dataset; this is particularly evident when looking at random blocking or $\text{MAE}_1$ for any of the blocking methods (Figures 8 and 9). However, improvement over random guessing is less marked for the birds dataset and spatial blocking tends to provide the most difficult evaluation setting across the two datasets–average performance is only slightly better than random guessing, especially when considering the large amount of variation in results (i.e., large standard deviations between folds). Species blocking results are, for the most part, comparable to random blocking.

Retraining the nonlinear hurdle on the full datasets independently displayed good model search convergence (Figure A16); all final models were GBDTs (Tables A1 and A2). Feature importance scores for the zero and nonzero models of the nonlinear hurdle reveal a strong dependence on species body mass, especially for the nonzero model on the birds dataset (Figures A17 and A18). Otherwise, there are no clear "best predictors": the different models tend to prioritise different predictors and generally make use of all predictors to some degree.

## 3.3 Utility of DL embeddings & modelling multiple taxa simultaneously

I tested DL embeddings as both supplements to existing hand-chosen predictors, as well as on their own, for the mammals dataset (Figure 10). Overall, the nonlinear hurdle model with both hand-chosen predictors and DL embeddings performed no better than the nonlinear hurdle with just hand-chosen predictors. However, the nonlinear hurdle with *only* DL embeddings consistently out-performed the naïve regressor and roughly matched the performance of the linear hurdle model (without random effects) (Figure 10). Spatial blocking was particularly difficult for the nonlinear hurdle with only DL embeddings, though, indicating a failure to generalise spatially.

As with the DL embeddings, modelling multiple taxa simultaneously had little effect on nonlinear hurdle model performance; for the mammals dataset, performance is nearly identical whether training on mammals and birds or mammals alone (Figure 10). The same is true for the birds dataset, except for a slight improvement in $\text{MAE}_1$ in the case of spatial blocking (Figure A15).

[1] Relative improvement here is calculated as $100 \cdot \frac{M_{nl} - M_l}{M_l}$, where $M_{nl}$ and $M_l$ are the metric values obtained by the nonlinear and linear models, respectively.

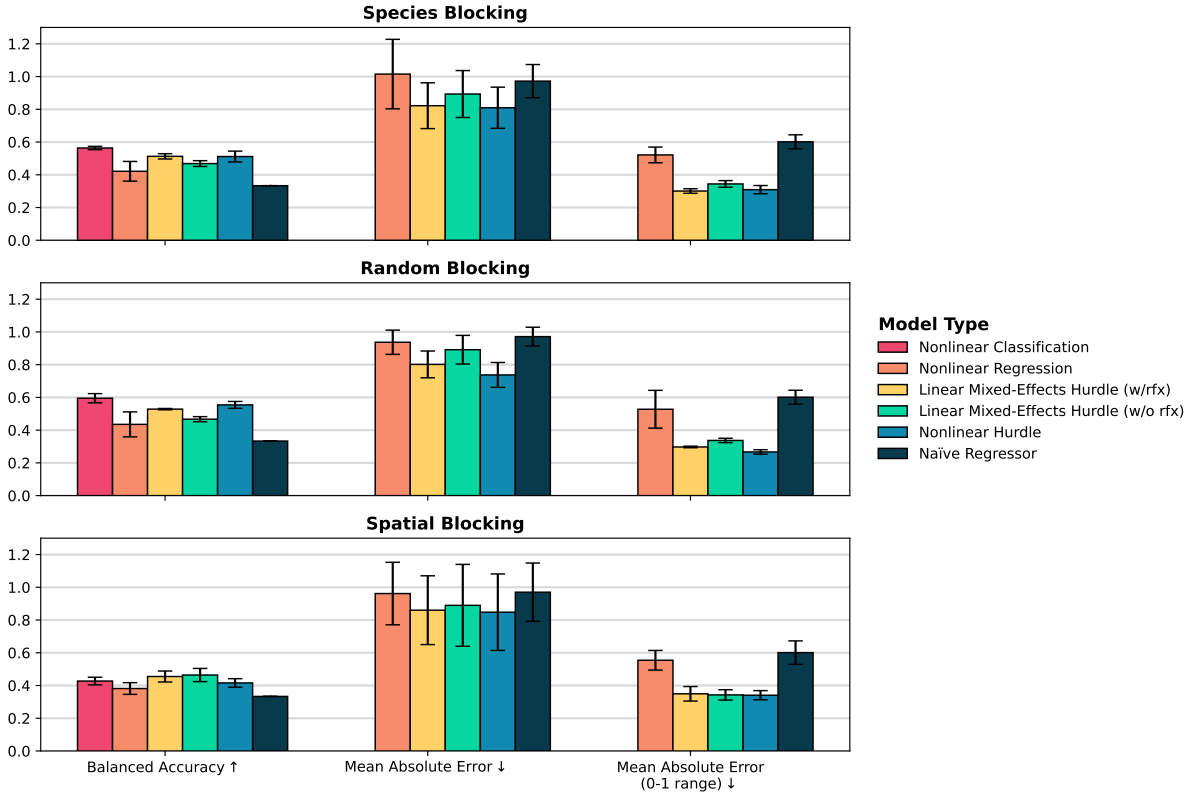



Figure 8: Performance metrics for the different predictive model and cross-validation approaches for predicting response ratios; all training and evaluation was performed only on the **mammals dataset**. 5-fold cross-validation was used and records were blocked either randomly, by species, or spatially. Three metrics were tested: balanced accuracy for classifying defaunation intensity categories, mean absolute error for regressing response ratios overall (i.e., $MAE_\infty$) and in the 0-1 range (i.e., $MAE_1$). Average metric values are shown ± one standard deviation. Results are included for linear mixed-effects hurdle models with and without random effects at prediction time ("w/rfx" and "w/o rfx," respectively).

### 3.4 Extreme spatial and taxonomic generalisability

As suggested by species and spatial cross-validation results (Figures 8 and 9), which test somewhat weaker forms of model generalisation, cross-taxa and cross-continent model generalisation are extremely difficult for both the linear and nonlinear hurdle models (Figures 11 and 12). Both models perform no better or, in some cases, worse than random guessing when inspecting BA and $MAE_\infty$ for cross-taxa generalisation; they do, however, improve on $MAE_1$ substantially when training on mammals and evaluating on birds (Figure 11). Similarly, when holding out a full biogeographic realm for the birds dataset, the nonlinear model improves $MAE_1$ consistently, but both models perform no better or worse than random guessing otherwise (Figure 12). When inspecting cross-continent generalisation for the mammals dataset, however, both models tend to outperform random guessing across all metrics, the nonlinear hurdle more so than the linear hurdle (Figure 12).

## 4 Discussion

### 4.1 Which predictive model should we use?

Overall, I found that a two-stage hurdle modelling approach performed best, likely due to differing dynamics governing local extirpation and nonzero RRs. Direct (nonlinear) classification on DI categories lacked flexibility and direct regression on RRs did not perform well (Figures 8 and 9). The nonlinear hurdle model tended to outperform the linear hurdle when random effects were not included (Figures 8 and 9). That being said, it is possible that the performance gain with the nonlinear hurdle can be primarily attributed to its ability to better take advantage of spatial autocorrelation in the data, since

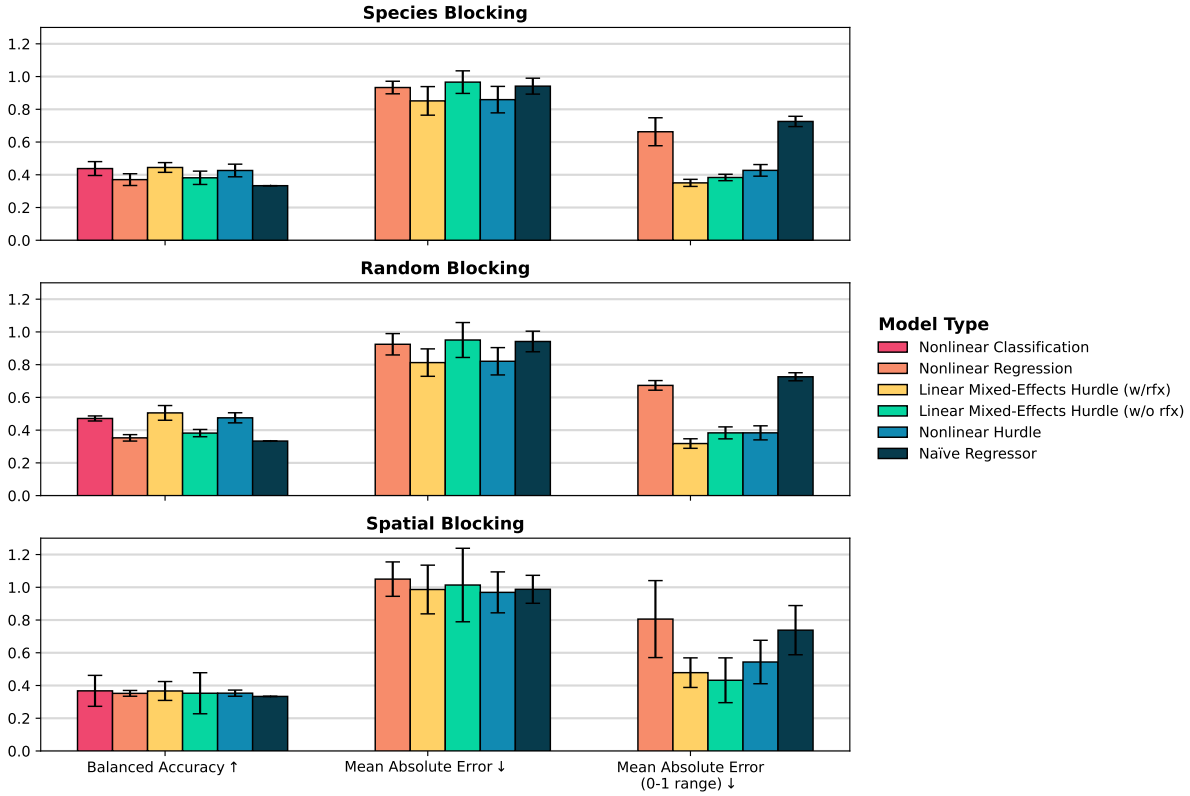



Figure 9: Performance metrics for the different predictive model and cross-validation approaches for predicting response ratios; all training and evaluation was performed only on the **birds dataset**. 5-fold cross-validation was used and records were blocked either randomly, by species, or spatially. Three metrics were tested: balanced accuracy for classifying defaunation intensity categories, mean absolute error for regressing response ratios overall (i.e., $MAE_\infty$) and in the 0-1 range (i.e., $MAE_1$). Average metric values are shown ± one standard deviation. Results are included for linear mixed-effects hurdle models with and without random effects at prediction time ("w/rfx" and "w/o rfx," respectively).

performance benefits largely disappeared under spatial blocking. Further, performance gains were more apparent for the mammals dataset than the birds dataset, potentially indicating that the nonlinear hurdle is sensitive to dataset size; for reference, the mammals dataset is about 120% larger than the birds dataset.

Spatial and species DL embeddings were largely unhelpful as supplemental predictors and modelling both birds and mammals simultaneously conferred no performance benefits (Figure 10). DL embeddings were useful on their own when used in conjunction with the nonlinear model, consistently outperforming random guessing and often matching the performance of the linear hurdle (without random effects). This suggests that the DL embeddings roughly capture the same ground conditions as the hand-chosen predictors. Models handled extreme spatial generalisation (Figure 12) better than extreme taxonomic generalisation (Figure 11), in both cases displaying better generalisation when the mammals dataset was used for training.

While the nonlinear hurdle model consistently performed better than the linear mixed-effects hurdle baseline, it is not clear whether the performance improvement is worth the loss of interpretability. While inspecting feature importances is illuminating (Figures A17 and A18), it gives us little insight into the functional relationship learned by a GBDT; conversely, GLMs are relatively easy to interpret. It is worth noting, however, that the linear hurdle was able to achieve low MAE in many cases because of its tendency to over-predict local extirpation events, particularly when random effects were excluded (Figures A13 and A14). Whether this is an issue depends on how we perceive local extirpation events: if, for instance, recovery from local extirpation requires substantially higher resource input on the part of conservation managers, false positives could be quite costly. In these cases, the nonlinear hurdle would likely be preferred, even though it is more difficult to summarise its predictive behaviour.

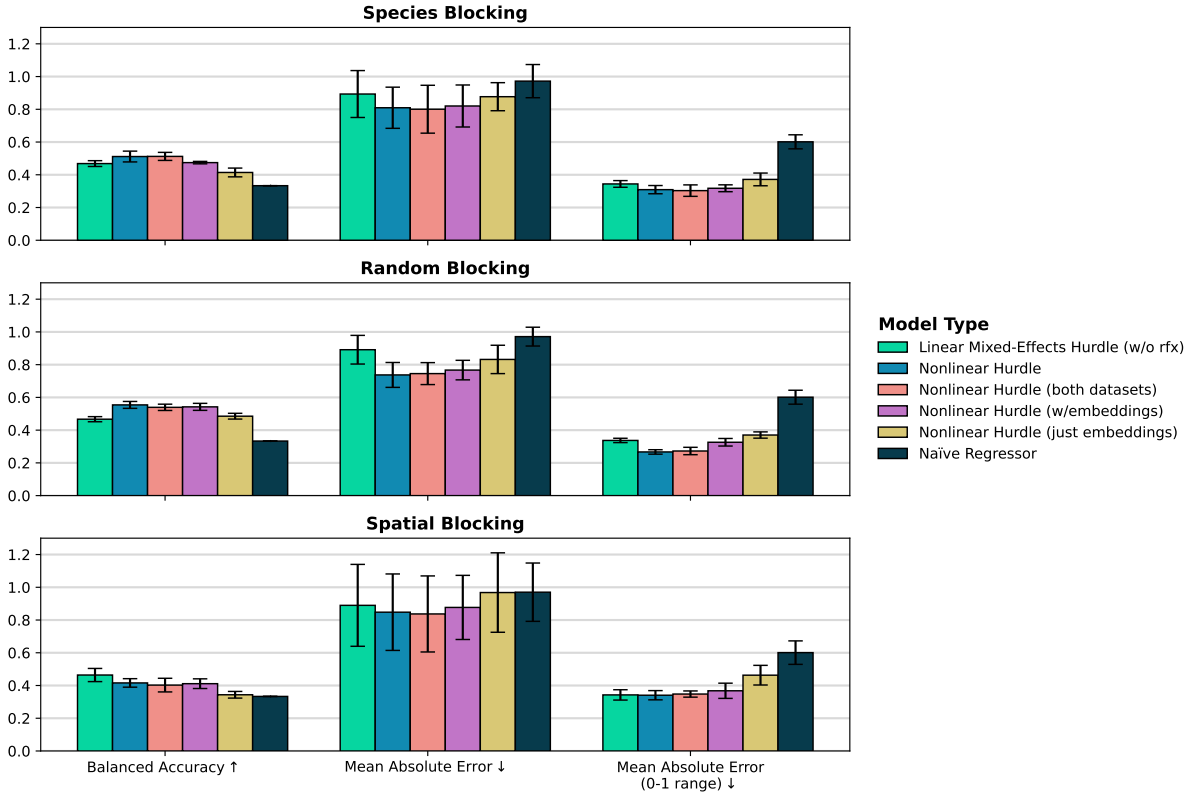



Figure 10: Performance metrics for the different predictive model and cross-validation approaches for predicting response ratios; one model was trained on both datasets, but otherwise all training and evaluation was performed only on the **mammals dataset**. 5-fold cross-validation was used and records were blocked either randomly, by species, or spatially. Three metrics were tested: balanced accuracy for classifying defaunation intensity categories, mean absolute error for regressing response ratios overall (i.e., $MAE_\infty$) and in the 0-1 range (i.e., $MAE_1$). Average metric values are shown ± one standard deviation. The focus is the nonlinear hurdle models with DL embeddings as predictors (BioCLIP [20] for species and SatCLIP [22] for spatial), but results from the nonlinear hurdle model and linear hurdle models (both using only hand-chosen features and the latter without random effects) are provided for context.

## 4.2 What is limiting predictive performance?

Even in the most optimistic evaluation setting–random cross-validation–model performance is relatively low. For reference, the nonlinear hurdle obtained a $MAE_1$ of 0.27 and $MAE_\infty$ of 0.74 on the mammals dataset (0.34 and 0.89, respectively, for the linear hurdle without random effects), but the mean RR in the $[0, 1]$ range is 0.43 and the mean RR overall is 1.16; under stronger generalisation scenarios (e.g., species- or spatial-blocking), performance is even less impressive. What then could be limiting predictive performance? The most compelling possibilities are:

1. The chosen predictors do not fully capture the motivations for or determinants of hunting and/or species susceptibility or likelihood of being captured by hunting.
2. There is too much noise in the predictor or response variables.
    (a) The response variable does not fully capture the phenomenon of interest, e.g., RRs do not actually capture the reduction in local abundance due to hunting.
3. The models cannot find the correct pattern because of limitations on model representational capacity.

Ideally, the DL embeddings would help address (1). However, DL embeddings conferred no performance benefits when used to supplement the hand-chosen predictors (Figure 10), and therefore it is possible that we are still missing key predictors. The measure of stunting prevalence is generally at

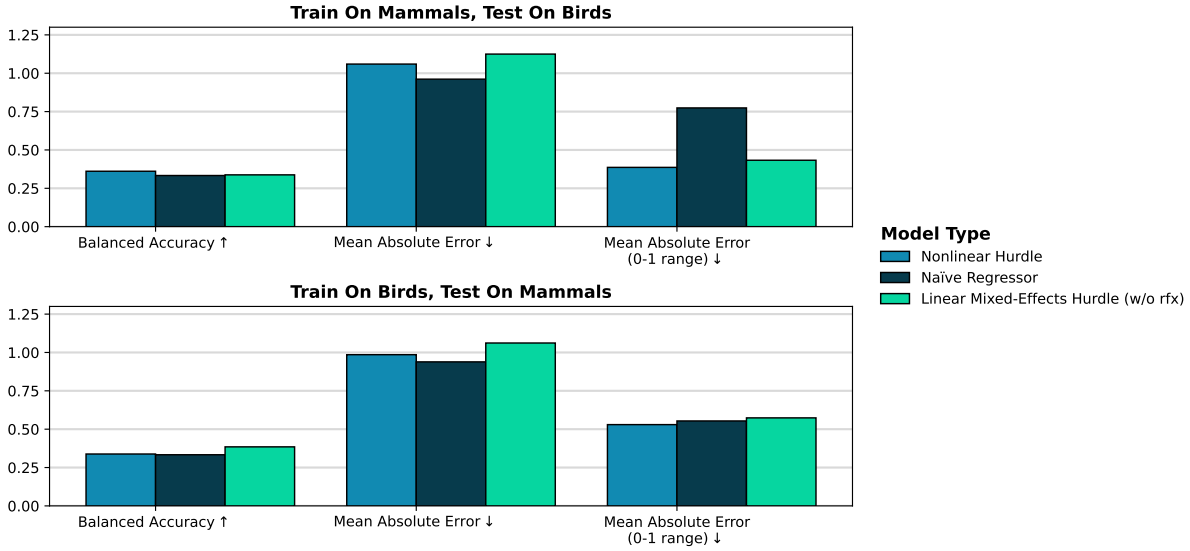



Figure 11: Model performance in the presence of extreme taxonomic generalisation, i.e., training on only the birds dataset and evaluating on only the mammals dataset, and vice versa. Results are shown only for the linear mixed-effects hurdle model without random effects at prediction time ("w/o rfx"), the nonlinear hurdle model, and the naïve regressor as a lower bound on performance.

first- or second-level administrative divisions; higher-resolution measures of poverty (e.g., the Gridded Relative Deprivation Index [38]) may help better characterise local socioeconomic conditions. Further, land-cover or habitat maps (e.g., [39]) may capture ease of hunting in a particular area or ecosystem-level vulnerability to disturbance via hunting.

(2) is fundamentally a data-side problem and is very difficult to address after-the-fact through improved modelling. Noise in predictor variables is inevitable given that many of them are not observed directly, but instead generated through some model with its own errors; in principle, however, noise should be reduced as newer versions using more advanced are released. (2a) is more concerning. The two dataset I used are the authoritative sources on abundance responses to hunting in the tropics and apply substantial vetting to ensure data quality. As alluded to earlier, though, spatial site matching is used to approximate a counterfactual scenario (i.e., $X_C$ should represent what the population abundance would have been if hunting were not present). But matching is structured around only two criteria: that the study "reports abundance at increasing distance from access points" and that no "potential confounding effects due to other disturbances" are present in the two sites [4]. Therefore, there will inevitably be substantial noise introduced due to other differences between sites that were not controlled for. Encouragingly, the two hunting RR datasets continue to grow (e.g., [5, 6] are themselves expansions of the datasets collected in [4]), but otherwise, noise in RR is difficult to address.

It is unlikely that (3) is the case, since GBDT methods like XGBoost and LGBM are generally considered state-of-the-art for tabular data [36]; they are extremely flexible in the nonlinear functions that they can learn. The only other class of model that is regularly competitive with GBDTs on tabular datasets are neural networks [36]; this exists as an avenue for further research, although I suspect that addressing (1) and (2) is likely to confer larger performance benefits.

## 4.3 How can hunting pressure maps be used?

Once a sufficiently performant predictive model has been obtained, the next logical step is to apply said model to all tropical bird and mammal species–indeed, this was the focus of the analysis in both [5] and [6]. On the level of an individual species, these maps can be thought of as the general "contemporary" abundance effect of hunting. Stacking species hunting pressure maps helps to highlight areas where many species are likely to be severely affected by hunting. In principle, any of the predictive models tested could be used to generate hunting pressure maps, with potential for ensembling model predictions to further improve accuracy.

We should, however, be cautious when interpreting any derived hunting pressure maps. I have shown the difficulties models face under realistic generalisation scenarios; for instance, species cross-validation

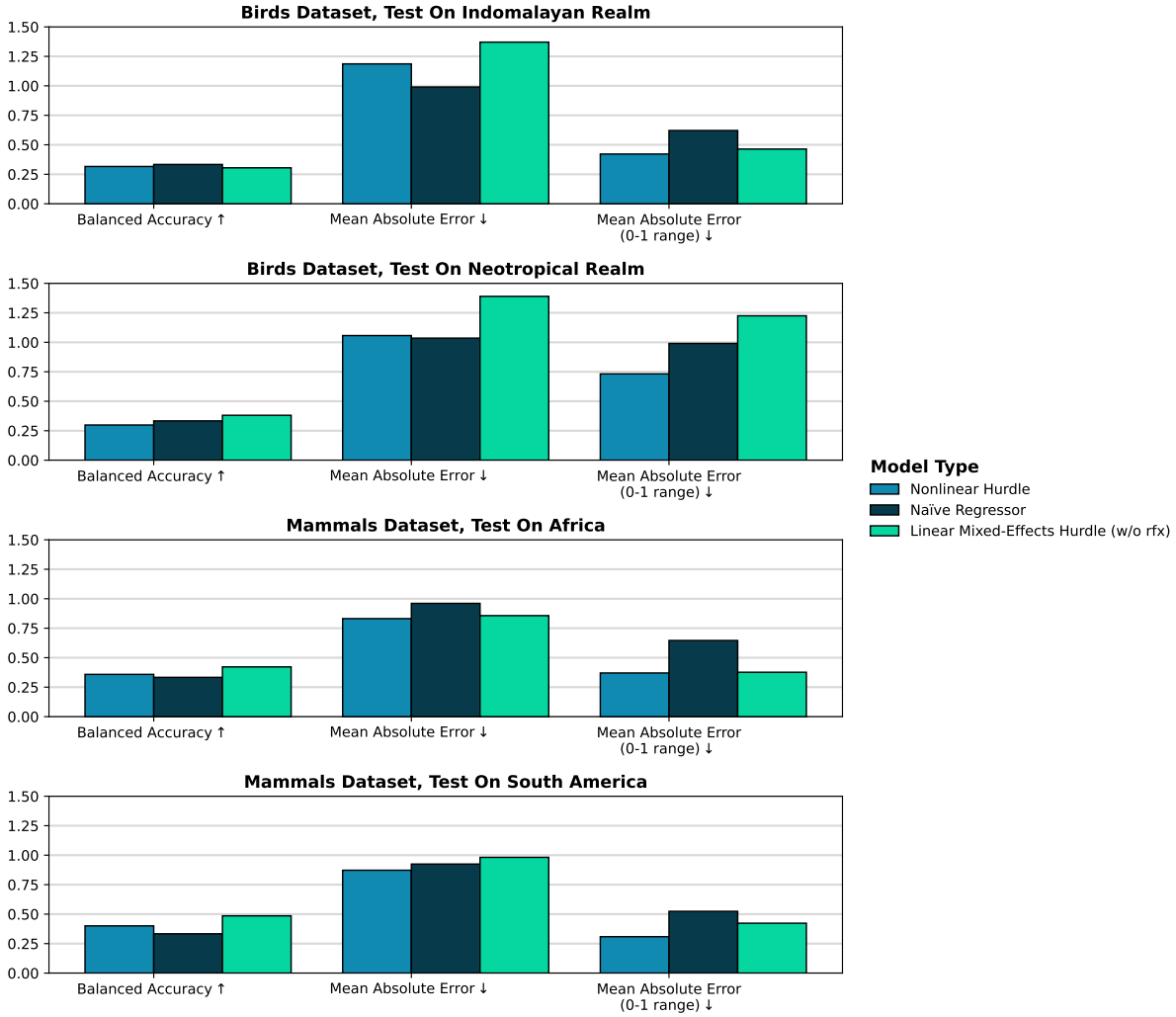



Figure 12: Model performance in the presence of extreme spatial generalisation, i.e., holding out all data from one continent or realm, training on all other data points, and evaluating on that continent or realm. Results are shown only for the linear mixed-effects hurdle model without random effects at prediction time ("w/o rfx"), the nonlinear hurdle model, and the naïve regressor as a lower bound on performance.

simulates generalisation to new species within the same class (e.g., other bird species for the a bird model). We have seen that model performance degrades when presented with new species (i.e., species blocking) and drops off substantially when spatial autocorrelation is accounted for (i.e., spatial blocking) (Figures 8 and 9). So, we should exercise particular caution when applying the model to unseen species (even of the same taxonomic class) or locations distant from training points. Extreme generalisation experiments are somewhat unrealistic, but emphasise an important point: we *should not* apply these models to completely unseen taxa (e.g., reptiles or amphibians) or locations for which there are no training points (e.g., North America or Europe). To do so would be inappropriate without at least some data to assess performance and, if possible, to train or fine-tune the predictive model.

This work was primarily undertaken to improve the limitations of the LIFE biodiversity metric [3]; I plan to use derived hunting pressure maps to adjust the value of units of AoH for each species according to their predicted responses to hunting. Other potential applications include conservation planning: [4] found that protected areas partially reduced the severity of hunting, so hunting pressure maps could in theory be used as an input to spatial conservation prioritisation systems for suggesting reserve sites (see, for example, [40]).

# 5 Conclusion

In this project, I pursued a comprehensive assessment of models for predicting species-specific abundance responses to hunting pressure. Focusing on two meta-datasets of RRs across the tropics for mammal and bird species, I assessed a broad array of possible modelling approaches, tested model performance under varying levels of taxonomic and spatial generalisation, and explored the utility of diverse predictors. While I found that a nonlinear hurdle model obtained through AutoML often outperformed the linear mixed-effects hurdle modelling baseline, spatial and taxonomic generalisation remained difficult for all models, and predictive performance seems to be fundamentally limited by data-side factors. Despite these limitations, the models I present will prove useful in improving pantropical hunting pressure maps and the prediction/evaluation framework I developed will lay the foundation for my PhD research in quantifying species-specific abundance effects of diverse anthropogenic threats.

# 6 Reproducibility statement

All code used to carry out analyses is available in the following GitHub repository: `https://github.com/emiliolr/life-hunting/tree/main`. The version from the MRes report deadline is archived in Zenodo: `https://zenodo.org/doi/10.5281/zenodo.12571509`. The hunting dataset for tropical birds is available at `https://github.com/IagoFerreiroArias/Bird_Defaunation/blob/main/Data/Bird_RR_data.csv` and the dataset for tropical mammals is available at `https://doi.org/10.6084/m9.figshare.6815288.v1`. Pre-trained BioCLIP weights were retrieved from `https://doi.org/10.57967/hf/1511` and pre-trained SatCLIP weights were retrieved from `https://huggingface.co/microsoft/SatCLIP-ResNet50-L40`.

# References


1. Ceballos G, Ehrlich PR, Barnosky AD, García A, Pringle RM, and Palmer TM. Accelerated modern human–induced species losses: Entering the sixth mass extinction. Science Advances 2015 Jun; 1. Publisher: American Association for the Advancement of Science:e1400253. Available from: https://www.science.org/doi/10.1126/sciadv.1400253 [Accessed on: 2024 Jun 15]
2. Ceballos G, Ehrlich PR, and Raven PH. Vertebrates on the brink as indicators of biological annihilation and the sixth mass extinction. Proceedings of the National Academy of Sciences 2020 Jun; 117. Publisher: Proceedings of the National Academy of Sciences:13596–602. Available from: https://www.pnas.org/doi/10.1073/pnas.1922686117 [Accessed on: 2024 Jun 15]
3. Eyres A, Ball T, Dales M, Swinfield T, Arnell A, Baisero D, Durán AP, Green J, Madhavapeddy A, and Balmford A. LIFE: A metric for quantitively mapping the impact of land-cover change on global extinctions. en. 2023 Dec. Available from: https://www.cambridge.org/engage/coe/article-details/65801ab4e9ebbb4db92dad33 [Accessed on: 2024 Mar 3]
4. Benítez-López A, Alkemade R, Schipper AM, Ingram DJ, Verweij PA, Eikelboom JAJ, and Huijbregts MAJ. The impact of hunting on tropical mammal and bird populations. Science 2017 Apr; 356. Publisher: American Association for the Advancement of Science:180–3. Available from: https://www.science.org/doi/full/10.1126/science.aaj1891 [Accessed on: 2024 Jan 30]
5. Benítez-López A, Santini L, Schipper AM, Busana M, and Huijbregts MAJ. Intact but empty forests? Patterns of hunting-induced mammal defaunation in the tropics. en. PLOS Biology 2019 May; 17. Publisher: Public Library of Science:e3000247. Available from: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000247 [Accessed on: 2023 Dec 18]
6. Ferreiro-Arias I, Santini L, Sagar HSSC, Richard-Hansen C, Guilbert E, Forget PM, Kuijk M van, Scabin AB, Peres CA, Revilla E, and Benítez-López A. Drivers and spatial patterns of avian defaunation in tropical forests. en. Diversity and Distributions 2024 May; n/a. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ddi.13855:e13855. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/ddi.13855 [Accessed on: 2024 May 13]
7. Deith MCM and Brodie JF. Predicting defaunation: accurately mapping bushmeat hunting pressure over large areas. Proceedings of the Royal Society B: Biological Sciences 2020 Mar; 287. Publisher: Royal Society:20192677. Available from: https://royalsocietypublishing.org/doi/full/10.1098/rspb.2019.2677 [Accessed on: 2024 Apr 18]
8. Mockrin MH, Rockwell RF, Redford KH, and Keuler NS. Effects of Landscape Features on the Distribution and Sustainability of Ungulate Hunting in Northern Congo. en. Conservation Biology 2011; 25. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1523-1739.2011.01660.x:514–25. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1523-1739.2011.01660.x [Accessed on: 2024 Feb 6]
9. Bogoni JA, Peres CA, and Ferraz KMPMB. Extent, intensity and drivers of mammal defaunation: a continental-scale analysis across the Neotropics. en. Scientific Reports 2020 Sep; 10. Number: 1 Publisher: Nature Publishing Group:14750. Available from: https://www.nature.com/articles/s41598-020-72010-w [Accessed on: 2024 Feb 26]
10. Harfoot MBJ, Johnston A, Balmford A, Burgess ND, Butchart SHM, Dias MP, Hazin C, Hilton-Taylor C, Hoffmann M, Isaac NJB, Iversen LL, Outhwaite CL, Visconti P, and Geldmann J. Using the IUCN Red List to map threats to terrestrial vertebrates at global scale. en. Nature Ecology & Evolution 2021 Nov; 5. Number: 11 Publisher: Nature Publishing Group:1510–9. Available from: https://www.nature.com/articles/s41559-021-01542-9 [Accessed on: 2023 Dec 18]
11. Broderick T, Gelman A, Meager R, Smith AL, and Zheng T. Toward a taxonomy of trust for probabilistic machine learning. Science Advances 2023 Feb; 9. Publisher: American Association for the Advancement of Science:eabn3999. Available from: https://www.science.org/doi/10.1126/sciadv.abn3999 [Accessed on: 2023 Dec 1]
12. Subramanian S, Harrington P, Keutzer K, Bhimji W, Morozov D, Mahoney MW, and Gholami A. Towards Foundation Models for Scientific Machine Learning: Characterizing Scaling and Transfer Behavior. en. Advances in Neural Information Processing Systems 2023 Dec; 36:71242–62. Available from: https://proceedings.neurips.cc/paper_files/paper/2023/hash/e15790966a4a9d85d688635c88ee6d8a-Abstract-Conference.html [Accessed on: 2024 Jun 28]

13. LeCun Y, Bengio Y, and Hinton G. Deep learning. en. Nature 2015 May; 521:436–44. Available from: http://www.nature.com/articles/nature14539 [Accessed on: 2021 Nov 5]

14. Lucas TCD. A translucent box: interpretable machine learning in ecology. en. Ecological Monographs 2020; 90. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ecm.1422:e01422. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/ecm.1422 [Accessed on: 2024 Jun 17]

15. Tuia D, Kellenberger B, Beery S, Costelloe BR, Zuffi S, Risse B, Mathis A, Mathis MW, Langevelde F van, Burghardt T, Kays R, Klinck H, Wikelski M, Couzin ID, Horn G van, Crofoot MC, Stewart CV, and Berger-Wolf T. Perspectives in machine learning for wildlife conservation. en. Nature Communications 2022 Feb; 13. Publisher: Nature Publishing Group:792. Available from: https://www.nature.com/articles/s41467-022-27980-y [Accessed on: 2024 Jun 23]

16. Carvalho DV, Pereira EM, and Cardoso JS. Machine Learning Interpretability: A Survey on Methods and Metrics. en. Electronics 2019 Aug; 8. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute:832. Available from: https://www.mdpi.com/2079-9292/8/8/832 [Accessed on: 2023 Dec 26]

17. Koh PW, Sagawa S, Marklund H, Xie SM, Zhang M, Balsubramani A, Hu W, Yasunaga M, Phillips RL, Gao I, Lee T, David E, Stavness I, Guo W, Earnshaw BA, Haque IS, Beery S, Leskovec J, Kundaje A, Pierson E, Levine S, Finn C, and Liang P. WILDS: A Benchmark of in-the-Wild Distribution Shifts. Number: arXiv:2012.07421 arXiv:2012.07421 [cs]. 2021 Jul. Available from: http://arxiv.org/abs/2012.07421 [Accessed on: 2022 Jun 25]

18. Tuggener L, Amirian M, Rombach K, Lorwald S, Varlet A, Westermann C, and Stadelmann T. Automated Machine Learning in Practice: State of the Art and Recent Results. en. *2019 6th Swiss Conference on Data Science (SDS)*. Bern, Switzerland: IEEE, 2019 Jun :31–6. Available from: https://ieeexplore.ieee.org/document/8789865/ [Accessed on: 2024 Jun 23]

19. Wang C, Wu Q, Weimer M, and Zhu E. FLAML: A Fast and Lightweight AutoML Library. en. *Proceedings of the 4th MLSys Conference*. San Jose, CA, USA, 2021

20. Stevens S, Wu J, Thompson MJ, Campolongo EG, Song CH, Carlyn DE, Dong L, Dahdul WM, Stewart C, Berger-Wolf T, Chao WL, and Su Y. BioCLIP: A Vision Foundation Model for the Tree of Life. arXiv:2311.18803 [cs]. 2023 Dec. Available from: http://arxiv.org/abs/2311.18803 [Accessed on: 2023 Dec 19]

21. Cole E, Van Horn G, Lange C, Shepard A, Leary P, Perona P, Loarie S, and Mac Aodha O. Spatial Implicit Neural Representations for Global-Scale Species Mapping. arXiv:2306.02564 [cs]. 2023 Jun. Available from: http://arxiv.org/abs/2306.02564 [Accessed on: 2023 Oct 26]

22. Klemmer K, Rolf E, Robinson C, Mackey L, and Rußwurm M. SatCLIP: Global, General-Purpose Location Embeddings with Satellite Imagery. arXiv:2311.17179 [cs]. 2023 Nov. Available from: http://arxiv.org/abs/2311.17179 [Accessed on: 2024 Mar 27]

23. Jakubik J, Roy S, Phillips CE, Fraccaro P, Godwin D, Zadrozny B, Szwarcman D, Gomes C, Nyirjesy G, Edwards B, Kimura D, Simumba N, Chu L, Mukkavilli SK, Lambhate D, Das K, Bangalore R, Oliveira D, Muszynski M, Ankur K, Ramasubramanian M, Gurung I, Khallaghi S, Hanxi, Li, Cecil M, Ahmadi M, Kordi F, Alemohammad H, Maskey M, Ganti R, Weldemariam K, and Ramachandran R. Foundation Models for Generalist Geospatial Artificial Intelligence. arXiv:2310.18660 [cs]. 2023 Nov. Available from: http://arxiv.org/abs/2310.18660 [Accessed on: 2024 Jan 8]

24. Nelson A, Weiss DJ, Etten J van, Cattaneo A, McMenomy TS, and Koo J. A suite of global accessibility indicators. en. Scientific Data 2019 Nov; 6. Publisher: Nature Publishing Group:266. Available from: https://www.nature.com/articles/s41597-019-0265-5 [Accessed on: 2024 Mar 27]

25. Center for International Earth Science Information Network - CIESIN - Columbia University. Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11. Palisades, New York, 2018. Available from: https://doi.org/10.7927/H49C6VHW

26. UNEP-WCMC and IUCN. The World Database on Protected Areas (WDPA). Cambridge, UK. 2024 Jun. Available from: www.protectedplanet.net

27. FAO. Chronic undernutrition among children: an indicator of poverty. 2003

28. Running S and Zhao M. MODIS/Terra Net Primary Production Gap-Filled Yearly L4 Global 500m SIN Grid V061. 2021 Mar. DOI: 10.5067/MODIS/MOD17A3HGF.06

29. Potapov P, Hansen MC, Laestadius L, Turubanova S, Yaroshenko A, Thies C, Smith W, Zhuravleva I, Komarova A, Minnemeyer S, and Esipova E. The last frontiers of wilderness: Tracking loss of intact forest landscapes from 2000 to 2013. Science Advances 2017 Jan; 3. Publisher: American Association for the Advancement of Science:e1600821. Available from: https://www.science.org/doi/10.1126/sciadv.1600821 [Accessed on: 2024 Jun 28]

30. Wilman H, Belmaker J, Simpson J, Rosa C de la, Rivadeneira MM, and Jetz W. EltonTraits 1.0: Species-level foraging attributes of the world's birds and mammals. en. Ecology 2014; 95. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1890/13-1917.1:2027–7. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1890/13-1917.1 [Accessed on: 2024 Jun 28]

31. Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics 2001; 29. Publisher: Institute of Mathematical Statistics:1189–232. Available from: https://www.jstor.org/stable/2699986 [Accessed on: 2024 Jun 28]

32. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, and Liu TY. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. Available from: https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html [Accessed on: 2024 Jun 3]

33. Chen T and Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. arXiv:1603.02754 [cs]. 2016 Aug :785–94. Available from: http://arxiv.org/abs/1603.02754 [Accessed on: 2024 Jun 28]

34. Breiman L. Random Forests. en. Machine Learning 2001 Oct; 45:5–32. Available from: https://doi.org/10.1023/A:1010933404324 [Accessed on: 2024 Jun 28]

35. Geurts P, Ernst D, and Wehenkel L. Extremely randomized trees. en. Machine Learning 2006 Apr; 63:3–42. Available from: http://link.springer.com/10.1007/s10994-006-6226-1 [Accessed on: 2024 Jun 28]

36. Grinsztajn L, Oyallon E, and Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? en. Advances in Neural Information Processing Systems 2022 Dec; 35:507–20. Available from: https://proceedings.neurips.cc/paper_files/paper/2022/hash/0378c7692da36807bdec87ab043cdadc-Abstract-Datasets_and_Benchmarks.html [Accessed on: 2024 May 10]

37. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, and Sutskever I. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs]. 2021 Feb. Available from: http://arxiv.org/abs/2103.00020 [Accessed on: 2024 Jun 28]

38. Center for International Earth Science Information Network - CIESIN - Columbia University. Global Gridded Relative Deprivation Index (GRDI), Version 1. Palisades, New York, 2022. Available from: https://doi.org/10.7927/3xxe-ap97

39. Jung M, Dahal PR, Butchart SHM, Donald PF, De Lamo X, Lesiv M, Kapos V, Rondinini C, and Visconti P. A global map of terrestrial habitat types. en. Scientific Data 2020 Aug; 7. Number: 1 Publisher: Nature Publishing Group:256. Available from: https://www.nature.com/articles/s41597-020-00599-8 [Accessed on: 2023 Dec 4]

40. Moilanen A, Franco AM, Early RI, Fox R, Wintle B, and Thomas CD. Prioritizing multiple-use landscapes for conservation: methods for large multi-species planning problems. Proceedings of the Royal Society B: Biological Sciences 2005 Aug; 272. Publisher: Royal Society:1885–91. Available from: https://royalsocietypublishing.org/doi/full/10.1098/rspb.2005.3164 [Accessed on: 2023 Oct 25]

41. Gallego-Zamorano J, Benítez-López A, Santini L, Hilbers JP, Huijbregts MAJ, and Schipper AM. Combined effects of land use and hunting on distributions of tropical mammals. en. Conservation Biology 2020; 34. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cobi.13459:1271–80. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/cobi.13459 [Accessed on: 2024 Feb 26]

42. Jolly E. Pymer4: Connecting R and Python for Linear Mixed Modeling. en. Journal of Open Source Software 2018 Nov; 3:862. Available from: https://joss.theoj.org/papers/10.21105/joss.00862 [Accessed on: 2024 Jun 28]

43. Bates D, Mächler M, Bolker B, and Walker S. Fitting Linear Mixed-Effects Models Using lme4. en. Journal of Statistical Software 2015 Oct; 67:1–48. Available from: https://doi.org/10.18637/jss.v067.i01 [Accessed on: 2024 Jun 28]

44. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, and Duchesnay É. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011; 12:2825–30. Available from: http://jmlr.org/papers/v12/pedregosa11a.html [Accessed on: 2024 Jun 28]

45. Uieda L. Verde: Processing and gridding spatial data using Green's functions. en. Journal of Open Source Software 2018 Oct; 3:957. Available from: https://joss.theoj.org/papers/10.21105/joss.00957 [Accessed on: 2024 Jun 28]

# Appendix

## Further details on evaluation metrics

If we have a dataset with $N$ data points in total with a single data point denoted $(\mathbf{x}_i, y_i)$ and the model's prediction as $f(\mathbf{x}_i) = \hat{y}_i$ (with true and predicted DI class as $c_i$ and $\hat{c}_i$, respectively), then the metrics used are defined as:

$$\mathrm{MAE}_\alpha(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N} \sum_{k \in K_\alpha} |y_k - \hat{y}_k|\,, \tag{2}$$

$$\mathrm{BA}(\mathbf{c}, \hat{\mathbf{c}}, \mathbf{w}) = \left( \sum_{j=1}^{N} w_j \right)^{-1} \sum_{i=1}^{N} \mathbb{1}(c_i = \hat{c}_i) \cdot w_i, \tag{3}$$

where $\mathbb{1}(\cdot)$ is the indicator function, $\mathbf{w}$ contains sample weights computed as

$$w_i = \left( \sum_{j=1}^{N} \mathbb{1}(c_j = c_i) \right)^{-1},$$

and $K_\alpha$ is all sample indices with abundance ratio in the 0 to $\alpha$ range, i.e., $K_\alpha = \{i : 0 \leq y_i \leq \alpha\}$. Of particular interest are $\alpha = 1$ and $\alpha = \infty$, the former focusing on abundance declines and the latter simply being MAE across all samples.

To validate my reproduction of [5], I also inspect per-class sensitivity and specificity for each the three DI categories, calculated as:

$$\mathrm{sensitivity}_C(\mathbf{c}, \hat{\mathbf{c}}) = \frac{\mathrm{TP}_C}{\mathrm{TP}_C + \mathrm{FN}_C}, \tag{4}$$

$$\mathrm{specificity}_C(\mathbf{c}, \hat{\mathbf{c}}) = \frac{\mathrm{TN}_C}{\mathrm{TN}_C + \mathrm{FP}_C}, \tag{5}$$

where $\mathrm{TN}_C$, $\mathrm{TP}_C$, $\mathrm{FN}_C$, and $\mathrm{FP}_C$ are the number of true negatives, true positives, false negatives, and false positives when DI category $C$ (e.g., "high" DI) is considered the positive class and the other two the negative class. Together, $\mathrm{sensitivity}_C$ and $\mathrm{specificity}_C$ inform us of the model's skill in picking out members of a particular DI category.

## Implementation details

### Reproducing the linear hurdle model

To establish a predictive baseline, I focused on reproducing the mixed-effects generalised linear hurdle model proposed in [5] for use with the mammals dataset. While I did not reproduce the precise approach of [6], the general structure of the model is very similar to [5] and therefore I believe application of the same linear hurdle model serves as an appropriate baseline. [5] used a Bayesian information criterion for feature selection, which retained only a subset of the available predictor variables; I used the same regression equations as reported. I did not fit on observations with RR$>$ 15, since I found that this generated a better model fit. I tuned the probability threshold $\tau$ for the zero model using the True Skill Statistic (TSS)

$$\mathrm{TSS} = \mathrm{sensitivity} + \mathrm{specificity} - 1, \tag{6}$$

following the approach of [41]. To validate the quality of the reproduction, I compared coefficient estimates and model log-likelihood with [5], and performed random 5-fold cross-validation to obtain model sensitivity, specificity, and BA for the three DI categories.

### AutoML settings

For direct regression, I used Mean Squared Error (MSE) as the validation metric, and for direct classification, I used BA. For the nonlinear hurdle model, I trained the zero and nonzero models separately, again using BA and MSE as the validation metrics, respectively. I limited model search to 2 minutes during cross-validation in all cases except when DL embeddings were used, where I instead limited to 3 minutes. All available spatial and species predictors were included, allowing models to select their own features via integrated regularisation strategies. Nonlinear hurdle models were re-trained with a 10-minute time budget on all records of each dataset to inspect model search convergence (Figure A16), feature importances (Figures A17 and A18), and final model and hyperparameter choices (Tables A1 and A2). The probability threshold $\tau$ was also chosen by tuning the TSS.

### Software packages

To implement the mixed-effects linear hurdle model, I used the `Pymer4` Python package [42], which directly accesses the `lme4` R package [43]. For AutoML model search, I used the Fast and Lightweight AutoML (`FLAML`) Python package [19]; I chose `FLAML` because it is targeted towards minimising computational costs and allows for optimisation of an arbitrary validation metric. For species blocking, I used the Python package `scikit-learn` [44] and for spatial blocking, I used the python package `verde` [45].

### Exploring hurdle model behaviour

To gain further insight into hurdle model predictive behaviour, I compared the distribution of predicted RRs against actual RRs (Figure A13) and inspected the agreement between actual versus predicted RRs (Figure A14), again only training and evaluating on a single dataset. The linear hurdle (without random effects) tends to substantially over-predict local extirpation events, with very few predicted RRs in the $(0, 0.5]$ interval (Figure A13). The nonlinear hurdle produces a more faithful overall distribution of predicted RRs: substantially fewer local extirpations are predicted, RRs $< 0.5$ are plentiful, and a larger cluster is present around RR $= 1$ for the mammals dataset (Figure A13). Similar patterns hold when inspecting actual versus predicted RRs (Figure A14): the vertical line at predicted RR $= 0$ for the linear hurdle (without random effects) demonstrates rampant false positive local extirpation predictions, whereas the nonlinear hurdle demonstrates generally better clustering around the 1:1 line, particularly for the mammals dataset. While not shown, results for distributional and actual vs. predicted under species and spatial blocking display similar trends.

Table 1: The models and hyperparameters chosen by FLAML during model search on the full mammals dataset.

| Sub-Model | Hyperparameter | Value |
|---|---|---|
| Zero Model (XGBoost) | `n_estimators` | 121 |
| | `max_leaves` | 377 |
| | `min_child_weight` | 0.131 |
| | `learning_rate` | 0.053 |
| | `subsample` | 0.980 |
| | `colsample_bylevel` | 0.742 |
| | `colsample_bytree` | 0.870 |
| | `reg_alpha` | 0.005 |
| | `reg_lambda` | 25.243 |
| Nonzero Model (LGBM, $\tau = 0.65$) | `n_estimators` | 2641 |
| | `num_leaves` | 4 |
| | `min_child_samples` | 3 |
| | `learning_rate` | 0.541 |
| | `log_max_bin` | 10 |
| | `colsample_bytree` | 0.385 |
| | `reg_alpha` | 0.007 |
| | `reg_lambda` | 1.325 |

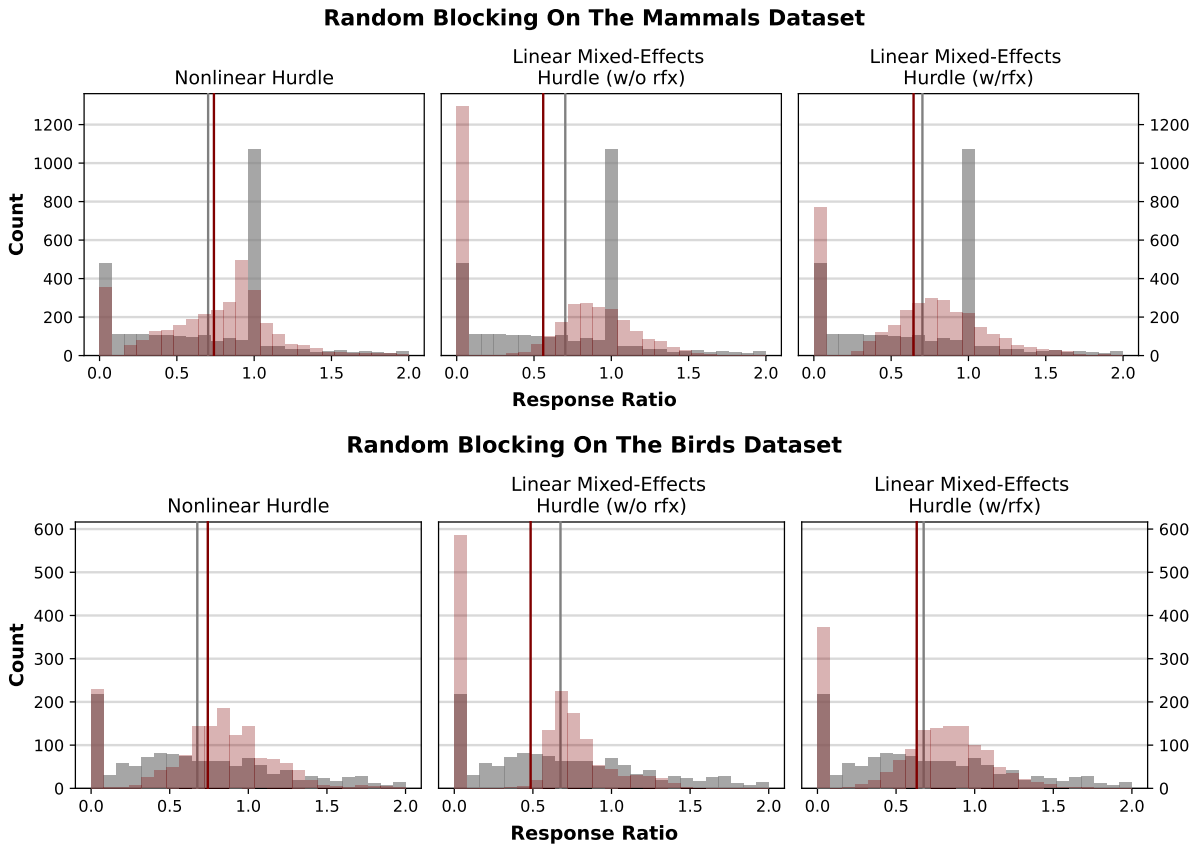



Figure 13: Distribution of response ratios for random 5-fold cross-validation for the linear and nonlinear hurdle models; the top row is for training and evaluating on only the mammals dataset and the bottom row the birds dataset. The red histogram shows the model prediction and the distribution of actual response ratios is shown in grey to highlight major discrepancies. The red vertical line indicates the mean predicted value for the presented range of response ratios (i.e., RR $\in [0, 2]$) and the grey vertical line the mean actual value.

Table 2: The models and hyperparameters chosen by FLAML during model search on the full birds dataset.

| Sub-Model | Hyperparameter | Value |
|---|---|---|
| Nonzero Model (LGBM) | `n_estimators` | 222 |
| | `num_leaves` | 16 |
| | `min_child_samples` | 7 |
| | `learning_rate` | 0.029 |
| | `log_max_bin` | 9 |
| | `colsample_bytree` | 0.967 |
| | `reg_alpha` | 0.001 |
| | `reg_lambda` | 6.386 |
| Zero Model (XGBoost, $\tau = 0.7$) | `n_estimators` | 473 |
| | `max_depth` | 6 |
| | `min_child_weight` | 1.821 |
| | `learning_rate` | 0.754 |
| | `subsample` | 0.893 |
| | `colsample_bylevel` | 0.642 |
| | `colsample_bytree` | 0.792 |
| | `reg_alpha` | 0.001 |
| | `reg_lambda` | 1.055 |

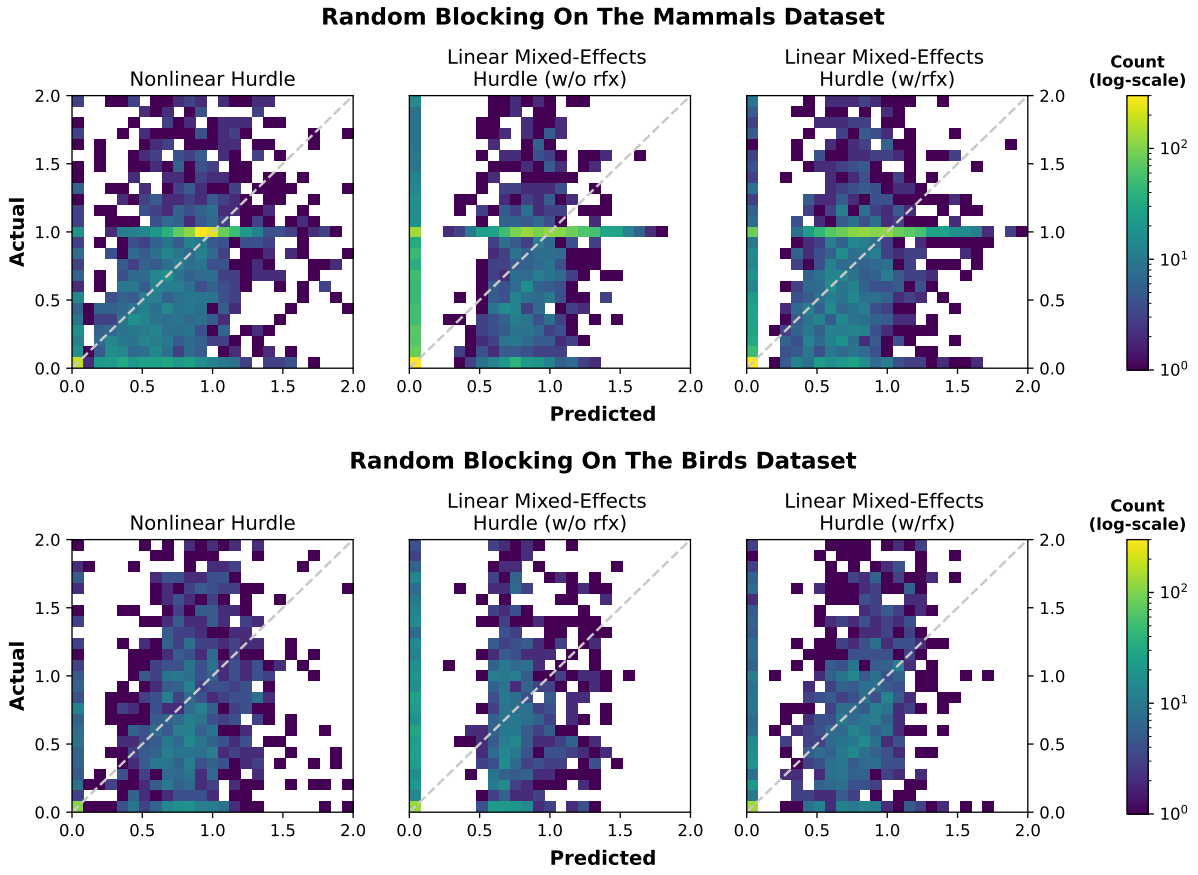



Figure 14: Actual vs. predicted response ratios for random cross-validation for the linear and nonlinear hurdle models; the top row is for training and evaluating on only the mammals dataset and the bottom row the birds dataset. The 1:1 line (i.e., where the predicted RR matches the actual RR) is shown in dashed grey. The space (i.e., $[0, 2]^2$) is discretised into equal-area bins, with the bin colour indicating the number of samples falling into that bin (on a log-scale).

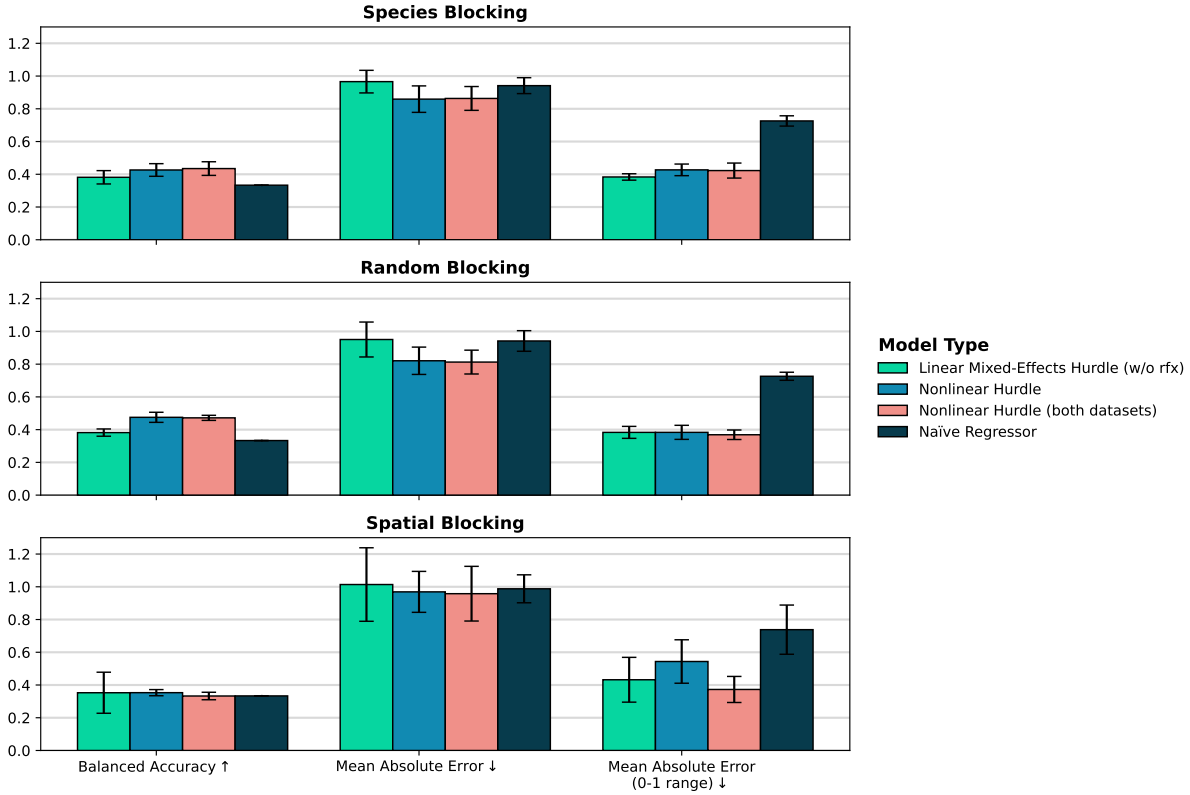



Figure 15: Cross-validation results across all blocking methods and for three metrics (BA, $MAE_\infty$, $MAE_1$), with a focus on the nonlinear hurdle model trained on both datasets. Otherwise, all training and evaluation was performed only on the **mammals dataset** of [5]. Results from the nonlinear hurdle model and linear hurdle models (both using only hand-chosen features and the latter without random effects) are provided for context.

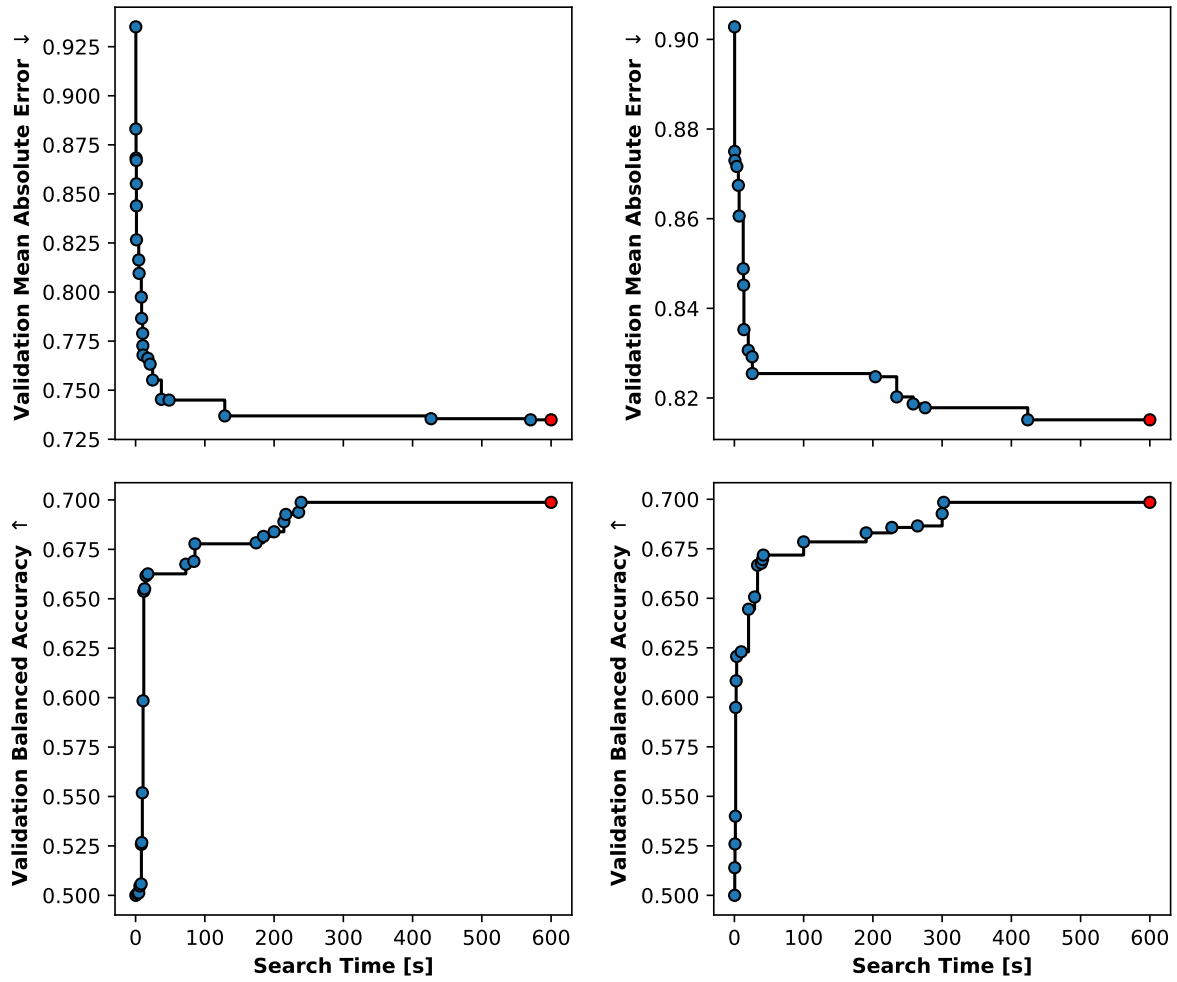



Figure 16: AutoML learning curves for the mammals (left) and birds dataset (right). Each point denotes a new best model found by FLAML (either a new model or new hyperparameters), which is accompanied by an improvement in validation performance (an increase for BA and a decrease for MAE).

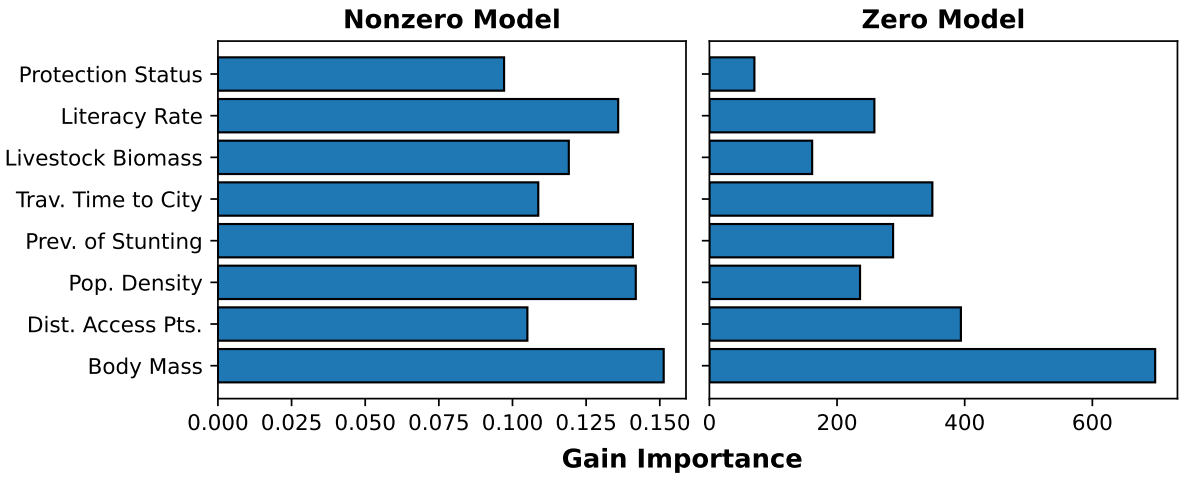



Figure 17: Model feature importance scores (expressed as gain importance) for the two components of the nonlinear hurdle model trained on the full mammals dataset. Importance scores are difficult to interpret directly and are on different scales different to different derivations between XGBoost and LGBM; relative differences between predictors are more insightful.

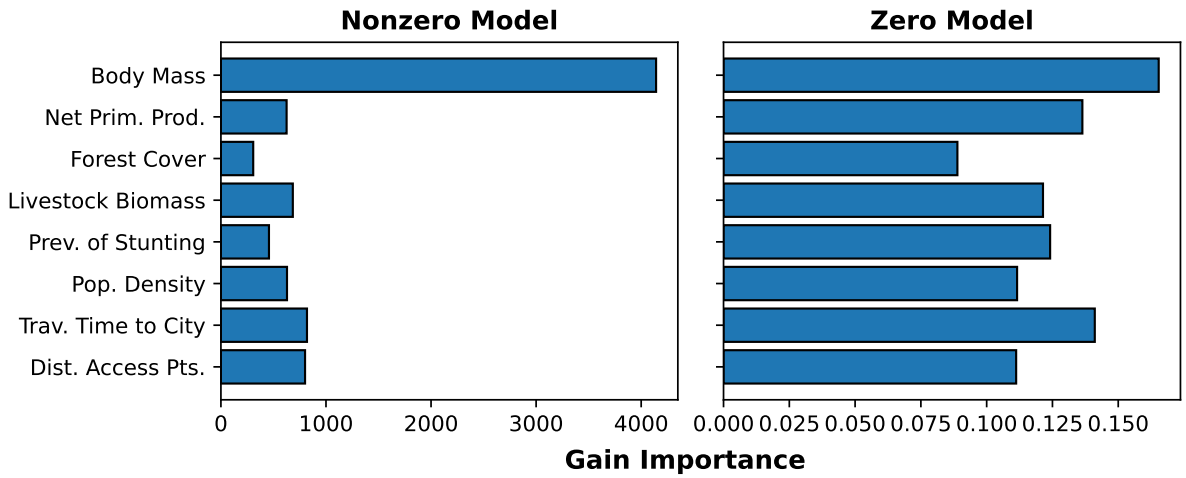



Figure 18: Model feature importance scores (expressed as gain importance) for the two components of the nonlinear hurdle model trained on the full birds dataset. Importance scores are difficult to interpret directly and are on different scales different to different derivations between XGBoost and LGBM; relative differences between predictors are more insightful.