Quantifying Species-Specific Responses to Hunting Pressure

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

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

Abstract

We are experiencing an unprecedented loss of terrestrial biodiversity globally, 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, like hunting, to biodiversity are much more difficult to observe directly. Hunting is difficult to track and 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 collected studies of tropical mammal and bird abundance responses. However, only one predictive model has been considered to date (a mixed-effects generalised linear hurdle model) and a thorough inspection of model performance under realistic generalisation scenarios 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, thoroughly tested (nonlinear) predictive methods through application of automated machine learning, experimented with embeddings from pre-trained deep learning models as a supplement to existing spatial and species 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 when used on their own in conjunction 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 [1, 2], particularly in the species-rich tropics **(CITATION NEEDED!)**. Species extinctions are primarily driven by loss of habitat, which is relatively easy to monitor by, e.g., satellite remote sensing **(CITATION NEEDED!)**. Other anthropogenic threats to biodiversity (e.g., fragmentation, edge effects, hunting, invasive species) can be thought of as forms of habitat degradation, reducing habitat quality and thus suitability. However, most anthropogenic threats are very difficult to observe directly. One threat to which many terrestrial vertebrate species are subject is hunting **(CITATION NEEDED!)**. While hunting is a fundamental human practice, recent patterns of overexploitation have substantially affected animal populations **(CITATION NEEDED!)**. 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 **(CITATION MAYBE NEEDED!)**.

The recently proposed Land-cover change Impacts on Future Extinctions (LIFE) metric ties a species's probability of extinction to their current Area of Habitat (AoH) relative to their theoretical AoH in the absence of human interference [3]. Application of LIFE allows for the quantification of the biodiversity impact of land-use changes through the lens of expected extinctions. However, occupancy is not guaranteed simply because habitat matching a species's preferences falls within their geographic range: not all AoH is equally suitable because of, e.g., the presence of diverse anthropogenic threats, like hunting. Therefore, a better understanding of the extent and intensity of these anthropogenic threats is urgently needed to support the broader assessment of biodiversity loss.

Previous work towards quantifying hunting pressure has relied on accessibility maps (e.g., distance to nearest settlement) [4, 5], suites of indicators [6], or IUCN assessments [7]. 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. Motivations for hunting range from subsistence hunting to pet or wild meat trade and on to trophy hunting, each of which prioritises different species or species traits. Further, the local socioeconomic context plays an important role in determining the prevalence of the different forms of hunting and their intensity **(CITATION NEEDED!)**. Finally, a species's response to hunting can extend beyond direct offtake through ecological interactions (e.g., predator-prey relationships), life history parameters (e.g., time to maturity), and other unknown factors **(CITATION DEFINITELY NEEDED!)**. These complexities make it extremely difficult to model hunting pressure and species responses explicitly based on established ecological theory.

Several studies have instead taken a data-driven approach, relying on statistical methods to implicitly assess hunting pressure through collected studies of tropical mammal [8] and bird [9] abundance responses. [8] was the first to show the potential for estimating hunting pressure from satellite-observable variables (e.g., human population density, HPD), 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). [9] extended the methodology to birds using Bayesian GLMs, with a brief consideration of performance under different generalisation scenarios. Average abundance declines of 13% and 12% were estimated across tropical mammal and bird species, respectively, with pressure varying across body size, diet type, hunting motivation, and spatial region [8, 9]. However, only one predictive model was considered across the two studies (albeit with different fitting procedures), 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 **(CITATION MAYBE NEEDED!)**. 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 **(CITATION NEEDED!)**. 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) [10, 11]. However, 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 [12]. 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 (on some validation set) using the least computational resources [12, 13].

Deep Learning (DL) constitutes a subset of ML models called neural networks, which roughly mimic the structure of the human brain; DL tends to perform well on unstructured data like imagery or audio [14]. Recently, large DL models called foundation models, which are trained to extract generally useful features (i.e., *embeddings*) from vast unlabelled datasets, have been released with broad flexibility in use as predictors for downstream tasks **(CITATION NEEDED)**. Embedding models have been introduce for ecological ML (e.g., species [15] or coordinate [16] embeddings) and for spatial data more generally (e.g., coordinate [17] or satellite imagery [18] embeddings), but their utility for bespoke downstream ecological tasks has yet to be thoroughly tested.

It is generally the case that improvement in predictive performance 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 or to predict their behaviour in future scenarios **(CITATION NEEDED)**. To help address these concerns, ML practitioners have developed careful validation techniques to assess predictive performance; central to this area is the study of model behaviour under distribution shift (i.e., when testing data does not match training data) **(CITATION NEEDED)**, which most often arises in the ecological setting due to spatial or taxonomic model generalisation **(CITATION NEEDED)**. 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; existing datasets [19, 8, 9] are spatially and taxonomically biased, and therefore successful model generalisation is not guaranteed.

Following recent work towards quantifying mammal and bird defaunation in the tropics [8, 9], I present a comprehensive assessment of approaches for this predictive task. In particular, I (1) reproduced the previous state-of-the art, (2) thoroughly tested (nonlinear) ML methods through application of auto ML, (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 Data

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 $\text{RR} = {}^{X_H}/_{X_C}$, with $\text{RR} = 0$ indicating local extirpation, $0 \geq \text{RR} \geq 1$ a local abundance decline, $\text{RR} = 1$ no abundance change, and $\text{RR} > 1$ a local abundance increase. This RR is interpreted as the local abundance change that is directly attributable to hunting; however, varying quality of study site matching procedures to construct a pseudo-counterfactual can introduce noise and erode the validity of this interpretation.

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 (see [8, 9] for a detailed review). Spatial predictors included measures of accessibility like travel time to large cities **(CITATION NEEDED!)** and distance to hunter access points, human population density **(CITATION NEEDED!)**, protection status **(CITATION NEEDED!)**, prevalence of stunting **(CITATION NEEDED!)**, literacy rate **(CITATION NEEDED!)**, and livestock biomass **(CITATION NEEDED!)**. Species predictors included diet and body mass **(CITATION NEEDED!)**.

For the most part, I used the mammal meta-dataset of [8], which includes pre-extracted spatial and species predictors for a number of hunting studies across the tropics. Control sites were chosen to minimise differences in potential confounding variables that might affect local species abundance (e.g., other threats like logging) and matches were only included if the control site was further from hunter access points than the hunted area. I also used the tropical bird dataset of [9] and the reptile records from [20], both of which use similar inclusion criteria for studies.

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.

2.2 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 $f(\mathbf{x}_i) \approx y_i$. It is also helpful to inspect the model's capacity to sort RRs into defaunation intensity (DI) categories, where $\text{DI}_i = 1 - \text{RR}_i$; assuming $0 \leq \text{RR} \leq 1$, this is the predicted percent decline in abundance due to hunting. As such, in cases I discretised RRs into "high" ($\text{DI} \geq 0.7$), "medium" ($0.7 > \text{DI} > 0.1$), and "low" intensity ($0.1 \geq \text{DI}$) following [8].

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") [8]. 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 predicts the log-transform of the RR to ensure that model predictions are non-negative, since abundance ratios must be ≥ 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 τ 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-effect 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. I used the `Pymer4` Python package **(CITATION NEEDED!)**–which directly accesses the `lme4` R package **(CITATION NEEDED!)**–to implement the linear hurdle model.

To efficiently test a wide range of nonlinear tabular ML methods, I applied automated machine learning (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). I used the Fast and Lightweight AutoML (FLAML) Python package [13] to carry out model search; I chose FLAML because it is targeted towards minimising computational costs and allows for optimisation of an arbitrary validation metric. The models I included in the search space were: k-Nearest Neighbours (k-NN), regularised Logistic Regression, and various decision tree ensembling methods (either Gradient Boosted Decision Trees (GBDTs), like LGBM **(CITATION NEEDED!)** and XGBoost **(CITATION NEEDED!)**, or Bagged Ensembles, like Random Forests **(CITATION NEEDED!)** (RF) and Extra Tree **(CITATION NEEDED!)**). GBDTs are generally considered state-of-the-art for ML on tabular datasets **(CITATION NEEDED!)**.

Figure 2: 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 τ. 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 Deep learning embeddings

To elucidate the utility of recently-introduced general-purpose deep learning embeddings, I experimented with SatCLIP [17] and BioCLIP [15] embeddings as supplements to hand-chosen spatial and species predictors, respectively. Both were trained with a multimodal CLIP objective **(CITATION NEEDED!)**; SatCLIP was trained to correspond a location's geographic (latitude-longitude) coordinates to satellite imagery from that location and BioCLIP was trained to correspond a species's full taxonomic name to images of that species.

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. However, it is difficult to be sure of what precisely is represented in either model's embeddings due to the black-box nature of the deep neural networks applied. I tested both full embeddings and reduced-dimension embeddings using Principal Component Analysis (PCA), retaining only sufficient principle components to explain 90% of the embedding's variance. I focused on two settings: using embeddings as supplements to the predictors of [8] or as the only predictors.

2.4 Evaluation

To evaluate model performance, I assessed model balanced accuracy (BA) for classification into DI categories as well as mean absolute error (MAE) and root mean squared error (RMSE) on regressed abundance ratios. If there are 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 these metrics are defined as:

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

$$\mathrm{RMSE}(\mathbf{y}, \hat{\mathbf{y}}) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}, \tag{3}$$

$$\mathrm{BA}(\mathbf{c}, \hat{\mathbf{c}}, \mathbf{w}) = \left(\sum_{j=1}^{N} w_j \right)^{-1} \sum_{i=1}^{N} \mathbb{1}(c_i = \hat{c}_i) \cdot w_i, \tag{4}$$

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_α is all sample indices with abundance ratio in the 0 to α 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 assess model generalisability across several dimensions, I employed cross-validation. I constructed folds either randomly or via dedicated blocking methods: group blocking on species, family, or order and spatial blocking to test basic taxonomic and spatial generalisation, respectively. I used the Python package `scikit-learn` **(CITATION NEEDED!)** for group blocking and `verde` **(CITATION NEEDED!)** for spatial blocking. I also tested model generalisability in two extreme scenarios: cross-continent (South America ↔ Africa) and cross-class generalisation (birds ↔ mammals). Finally, I tested whether mixing datasets taxonomically (e.g., modelling birds and mammals jointly) confers performance benefits.

Figure 3: 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, and species identity is indicated by colour in the species blocking plot.

2.5 Implementation details

3 Results

3.1 Dataset description

The full dataset, which is composed of hunting effects from [8] (mammals) and [9] (birds), 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 (only 126 records have RR > 5) and usually reflect abundance declines ($0 <$ RR ≤ 1; 2,970 records) (Figure 4). 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). Low DI is most common (2,381 records) across the two datasets, with roughly the same number samples for medium and high DI (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 5). Brazil and Indonesia are the countries with the most records for the mammals and birds datasets, respectively.

Figure 4: The distribution of RRs in both datasets, across the full range of abundance ratios (top left, log-scale) and for only abundance declines (top right). Shown also is the distribution of DI categories across both datasets (bottom).

Figure 5: 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.

3.2 Reproducing the state-of-the-art

To establish a predictive baseline, I focused on reproducing the mixed-effects generalised linear hurdle model proposed in [8] for use with the mammals dataset. While I did not reproduce the precise approach of [9], the general structure of the model is very similar to [8] and therefore I believe application of the same linear hurdle model serves as an appropriate baseline. [8] used a Bayesian Information Criterion (BIC) for feature selection, which retained only a subset of the available predictor variables; I used the same regression equations as indicated in Table S5 of [8] (Figure 6). I did not fit on observations with RR$>$ 15, since I found that this generated a better model fit, but I performed model evaluation on *all* records.

When I fit the linear hurdle model as described, I found similar log-likelihood values (-947.0 and -3698.0 for the binomial and Gaussian models, respectively) as reported in Table S5 of [8]. I also obtained very similar coefficient estimates for the fixed-effects predictors, cross-referenced with Figure S5a-b of [8] (Figure 6). Finally, I performed random 5-fold cross-validation and inspected model sensitivity, specificity, and BA for the three DI categories (Figure 7). Again, while not precisely the same as reported in Figure S6b of [8], the obtained cross-validation metrics are similar and display many of the same patterns (e.g., high BA for high DIs, very high specificity for low DIs, and generally intermediate performance for medium DIs). While the results that I obtained do not perfectly align with [8], 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 RR. 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 corresponding figure in the original study [8] is Figure S5a-b.

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. The corresponding figure in the original study [8] is Figure S6b.

3.3 Model benchmarking

To broadly assess the effectiveness of different ML models for this predictive task, I first benchmarked model performance on the mammals and birds datasets independently using random, spatial, and species blocking for 5-fold cross-validation (Figures 8 and 9). I find that hurdle models are generally more suitable than direct regression or classification models with respect to the chosen evaluation metrics. Direct regression is, in almost all cases, 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 (Figures 8 and 9). Performance gains for nonlinear hurdles are most evident when the train and test distributions match well (i.e., random cross-validation), with a 21% and 17% relative improvement in MAE_1 and MAE_∞, 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

[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.

patterns hold for basic taxonomic and spatial generalisation (i.e., species and spatial cross-validation, respectively): the nonlinear hurdle tends to improve on MAE_∞ by $\approx 5-10\%$ across both datasets, but only improves on MAE_1 for the mammals dataset and not the birds dataset (Figure 9).

To gain further insight into hurdle model predictive behaviour, I compare the distribution of predicted RRs against actual RRs (Figure 10) and inspect the agreement between actual versus predicted RRs (Figure 11) over all folds for random 5-fold cross-validation. For both datasets, 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 10). While including random effects in the linear hurdle tends to rebalance the distribution of predicted RRs, 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 10). Similar patterns hold when inspecting actual versus predicted RRs (Figure 11): the vertical line at predicted RR $= 0$ for the linear hurdle (without random effects) demonstrates rampant false positive instances of local extirpation, 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.

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.

Figure 8: 5-fold cross-validation results across all blocking methods, all predictive models, and for three metrics (BA, MAE_∞, MAE_1). Results are shown for linear mixed-effects hurdle models with and without random effects at prediction time ("w/rfx" and "w/o rfx," respectively). All training and evaluation was performed only on the **mammals dataset** of [8].

Figure 9: 5-fold cross-validation results across all blocking methods, all predictive models, and for three metrics (BA, MAE_∞, MAE_1). Results are shown for linear mixed-effects hurdle models with and without random effects at prediction time ("w/rfx" and "w/o rfx," respectively). All training and evaluation was performed only on the **birds dataset** of [9].

3.4 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 12). 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 (i.e., it does better than random guessing) and roughly matched the performance of the linear hurdle model (without random effects), actually improving MAE_∞ by 2% and 7% for species and random blocking, respectively (Figure 12). 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 has 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 12). The same is true for the birds dataset, except for a slight improvement in MAE_1 for spatial blocking (Figure A15).

3.5 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 13 and 14). Both models perform no better or, in some cases, worse than random guessing when inspecting BA and MAE_∞ for cross-taxa generalisation; they do, however, improve on MAE_1 substantially when training on mammals and evaluating on birds (Figure 13). 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 14). 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 14).

Figure 10: Distribution of abundance ratios for random 5-fold cross-validation for the linear and nonlinear hurdle models. Bar colour indicates a data point's assigned fold. The red vertical line indicates the mean value for the presented range of abundance ratios (i.e., RR $\in [0, 2]$) and the blue vertical line the median. The distribution of actual abundance ratios is shown in grey to highlight major discrepancies.

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 perform approximately as well as the linear hurdle when random effects were included at prediction time, but outperformed 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 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 12). 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). Models handled extreme spatial generalisation (Figure 14) better than extreme taxonomic generalisation (Figure 13), in both cases displaying better generalisation when the mammals dataset was used for training.

While the nonlinear hurdle model semi-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

Figure 11: Actual vs. predicted abundance ratios for random cross-validation for the linear and nonlinear hurdle models. 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).

into the functional relationship learned by a GBDT; conversely, GLMs are relatively easy to interpret due to the simplifying assumption of global linearity. 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 4 and 11). Whether this is an issue depends on the response of local extirpation: 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 describe its predictive behaviour.

4.2 What is limiting predictive performance?

Even in the most optimistic evaluation setting–random cross-validation–model performance is relatively low. For example, the nonlinear hurdle obtained a MAE_1 of 0.27 and MAE_∞ 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.

Figure 12: Cross-validation results across all blocking methods and for three metrics (BA, MAE_∞, MAE_1), with a focus on the nonlinear hurdle models with DL embeddings as predictors (BioCLIP [15] for species and SatCLIP [17] for spatial). 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. One nonlinear hurdle model was trained on both datasets; otherwise, all training and evaluation was performed only on the **mammals dataset** of [8].

Figure 13: 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.

Figure 14: 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.

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 12), and therefore it is possible that we are missing key predictors. The measure of stunting prevalence is generally at first- or second-level administrative divisions; higher-resolution measures of poverty (e.g., the Gridded Relative Deprivation Index **(CITATION NEEDED)**) may help better characterise local socioeconomic conditions. Further, use of land-cover or habitat maps (e.g., **CITATIONS NEEDED**) may capture ease of hunting in a particular area **(CITATION NEEDED)** or ecosystem-level vulnerability to disturbance via hunting **(CITATION NEEDED)**.

(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. For example, the HPD data product from NASA is updated regularly using new census data (sometimes at lower administrative levels) and improved disaggregation methods **(CITATION NEEDED)**.

(2a) is perhaps the most difficult to address: the two datasets used are quite unique and already apply substantial vetting to ensure data quality. As alluded to earlier 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 the matching approach 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 [19]. Therefore, there will inevitably be substantial noise introduced due to other differences between sites that were not controlled for **(CITATION NEEDED)**. Statistics dictates that the higher the noise in the response variable, the more data that will be required to estimate a robust relationship between variables. Encouragingly, the two hunting RR datasets continue to grow (e.g., [8, 9] are expansions of the datasets collected in [19]); it is also possible that general abundance datasets like PREDICTS **(CITATION NEEDED)** or BioTIME **(CITATION NEEDED)** could be used somehow as noisy supplements or pre-training datasets for a DL model.

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 **(CITATION NEEDED)**; 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 NNs **(CITATION NEEDED)**; this exists as an avenue for further research, although I suspect that (1) and (2) are likely to confer larger performance benefits.

4.3 What are the applications of hunting pressure maps?

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 [8] and [9]. On the level of an individual species, these maps can be thought of as the general "contemporary" abundance effect of hunting; temporality is only injected into the models implicitly through predictor variables (e.g., a study from 2006 would use a HPD data product from, or as close as possible to, 2006), so it is difficult to pin predictions down to a particular year. 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 **(CITATION WOULD BE HELPFUL)**.

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 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 or locations distant from training points. Extreme generalisation experiments are somewhat unrealistic, but empirically underscore 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 motivated by application to the LIFE biodiversity metric [3]; an intuitive approach for integrating hunting effects into LIFE is by simply multiplying each pixel within a species's AoH by the estimated RR before estimating probability of persistence. In this case, rather than each unit of available habitat counting towards AoH equally, each unit is made more or less "valuable" to a species by a negative or positive abundance response to hunting, respectively. Other potential applications include conservation planning: [19] 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 establishing reserves **(CITATION WOULD BE HELPFUL)**.

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 factors relating to data. 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. DOI: 10.1126/sciadv.1400253. 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. DOI: 10.1073/pnas.1922686117. 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. DOI: 10.33774/coe-2023-gpn4p-v2. Available from: https://www.cambridge.org/engage/coe/article-details/65801ab4e9ebbb4db92dad33 [Accessed on: 2024 Mar 3]
4. 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. DOI: 10.1098/rspb.2019.2677. Available from: https://royalsocietypublishing.org/doi/full/10.1098/rspb.2019.2677 [Accessed on: 2024 Apr 18]
5. 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. DOI: 10.1111/j.1523-1739.2011.01660.x. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1523-1739.2011.01660.x [Accessed on: 2024 Feb 6]
6. 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. DOI: 10.1038/s41598-020-72010-w. Available from: https://www.nature.com/articles/s41598-020-72010-w [Accessed on: 2024 Feb 26]
7. 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. DOI: 10.1038/s41559-021-01542-9. Available from: https://www.nature.com/articles/s41559-021-01542-9 [Accessed on: 2023 Dec 18]
8. 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. DOI: 10.1371/journal.pbio.3000247. Available from: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000247 [Accessed on: 2023 Dec 18]
9. 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. DOI: 10.1111/ddi.13855. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/ddi.13855 [Accessed on: 2024 May 13]
10. 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. DOI: 10.1002/ecm.1422. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/ecm.1422 [Accessed on: 2024 Jun 17]

11. 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. DOI: 10.1038/s41467-022-27980-y. Available from: https://www.nature.com/articles/s41467-022-27980-y [Accessed on: 2024 Jun 23]

12. 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. DOI: 10.1109/SDS.2019.00-11. Available from: https://ieeexplore.ieee.org/document/8789865/ [Accessed on: 2024 Jun 23]

13. Wang C, Wu Q, Weimer M, and Zhu E. FLAML: A Fast and Lightweight AutoML Library. en

14. LeCun Y, Bengio Y, and Hinton G. Deep learning. en. Nature 2015 May; 521:436–44. DOI: 10.1038/nature14539. Available from: http://www.nature.com/articles/nature14539 [Accessed on: 2021 Nov 5]

15. 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]

16. 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. DOI: 10.48550/arXiv.2306.02564. Available from: http://arxiv.org/abs/2306.02564 [Accessed on: 2023 Oct 26]

17. 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]

18. 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]

19. 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. DOI: 10.1126/science.aaj1891. Available from: https://www.science.org/doi/full/10.1126/science.aaj1891 [Accessed on: 2024 Jan 30]

20. Morton O, Scheffers BR, Haugaasen T, and Edwards DP. Impacts of wildlife trade on terrestrial biodiversity. en. Nature Ecology & Evolution 2021 Apr; 5. Number: 4 Publisher: Nature Publishing Group:540–8. DOI: 10.1038/s41559-021-01399-y. Available from: https://www.nature.com/articles/s41559-021-01399-y [Accessed on: 2024 Feb 6]

Appendix

Figure 15: Cross-validation results across all blocking methods and for three metrics (BA, MAE_∞, 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 [8]. 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.

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

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