Occupancy Modeling of Birds in the Amazon Rainforest

Executive Summary

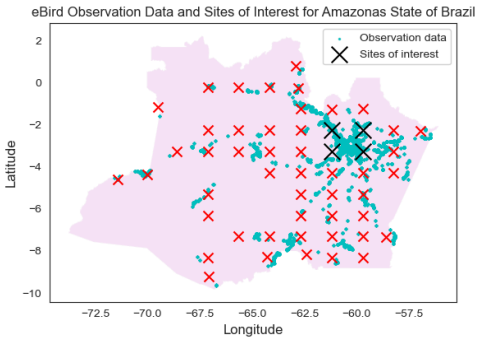
Jeremy Borden, Chelsey Hunts, Dawit Mengesha, Yusup Amat, Sriram Raghunath

Introduction

Our goals for this project were to test different occupancy modeling strategies to explore if and how climate change or forest loss has affected bird populations in the Amazonas region of Brazil over the time period of 2012 – 2021, and subsequently evaluate which models performed the best. We tested this for two species – a generalist species, Black vulture (*Coragyps atratus*) and a forest specialist, Screaming piha (*Lipaugus vociferans*). We used three different modeling approaches, two standard machine learning classification models – balanced random forest and binary logistic regression – and one modern occupancy modeling approach using the R package spOccupancy.

Data Collection

1. **Bird Detection Data**
   We obtained a dataset containing all eBird observations within Brazil directly from eBird.[1] The data are split into an observation dataset, containing observations of individual bird species, and a checklist dataset, containing records of 'checklists' - observation runs where several different species may have been observed. Generating a usable csv file for analysis and modeling requires some pre-preprocessing to combine the observation dataset and checklist dataset into a single dataset with detection/non-detection data for all species. Since most of the functionality to handle eBird data has been developed in R (like *auk*, the eBird Data Extraction and Processing in R package), we carried out this step in R rather than python. Details can be found in the R markdown file ebird_pre_preprocessing_example.rmd in the github.



   We subsequently narrowed our focus to the Amazonas state of Brazil, defined a grid of sites which could be used to group observations, and selected a set of four of these sites with data between the years of 2012-2021 on which to build and test models. We further restricted our analysis to a collection of two particular species - a generalist, Black vulture (*Coragyps atratus*) and a forest specialist, Screaming piha (*Lipaugus vociferans*).

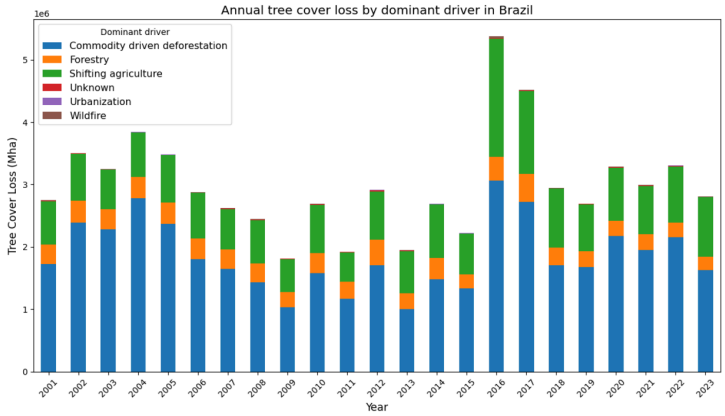
2. **Climate Data**
   Monthly weather data (temperature and precipitation) for the period 2012-2021 are obtained from WorldClim Historical Monthly Weather Data. These data are downscaled from CRU-TS-4.06 by the Climatic Research Unit, University of East Anglia, using WorldClim 2.1 for bias correction[2]. The features available are average minimum temperature (°C), average maximum temperature (°C) and total precipitation (mm). The spatial resolution we use is 10 minutes (~340 km^2). Each download is a "zip" file containing 120 GeoTiff (.tif) files, for each month of the year (January is 1; December is 12), for a 10 year period.

3. **Tree cover and tree cover loss Data**
   To analyze how climate change, especially increasing temperature and human-made deforestation, affects bird species occupancy we obtained the vegetation coverage data via the Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Vegetation Indices (MOD13Q1) Version 6.1, generated every 16 days at 250 meter (m) spatial resolution from the USGS.[3] For our analysis, we calculated the annual mean EVI for all sites in the

[1] Cornell Lab of Ornithology https://ebird.org/home
[2] CRU-TS 4.06 (Harris et al., 2020) downscaled with WorldClim 2.1 (Fick and Hijmans, 2017) https://www.worldclim.org/data/monthlywth.html
[3] MOD13Q1, USGS: https://lpdaac.usgs.gov/products/mod13q1v061/

Amazonas state. The tree cover loss data was obtained from Global Forest watch.[4] This dataset shows the dominant driver of tree cover loss from 2001-2023 for 27 states in Brazil using the following five categories:

- Commodity-driven deforestation: Large-scale deforestation linked primarily to commercial agricultural expansion.
- Shifting agriculture: Temporary loss or permanent deforestation due to small- and medium-scale agriculture.
- Forestry: Temporary loss from plantation and natural forest harvesting, with some deforestation of primary forests.
- Wildfire: Temporary loss, does not include fire clearing for agriculture.
- Urbanization: Deforestation for expansion of urban centers



With all our data collected and synthesized, we employ our occupancy covariates (the environmental features) and detection covariates (features from eBird related to human observation efforts, e.g. hours spent looking for birds) to model s**pecies occupancy**, a binary variable denoting whether or not a species occupies a particular site.

Standard Machine Learning Approaches

Since we are interested in modeling the effects of environmental change on species occupancy over time, we need to account for the time-ordered nature of our occupancy covariates (precipitation, temperature, EVI, and tree cover loss). Note we are not interested in the change of detection covariates (e.g. how many hours an observer spent looking for birds) over time. Since we still intend to use the standard techniques of supervised learning and classification, we must transform our time-series data into supervised learning features. We do this by creating new 'shifted' features for our occupancy covariates. These shifted features allow us to use the value of an occupancy covariate at time step t-1 to predict species occupancy at time step t. We use these shifted occupancy covariates along with the detection covariates in the following two models.

Target	Time Step	Feature (t)	Feature (t-1)
y4	4		
y3	3		
y2	2		
y1	1		
y0	0		

We take a single-species, single-site approach to model selection - selecting a model with data from one of our two species (Screaming piha) at one of our four sites. We later examine how the model generalizes to the other species and sites.

1. **Binary Logistic Regression with LASSO Regularization (Log Reg)**
   Our first model is a binary logistic regression (binary because we predict occupancy as True/False). We use LASSO regularization for feature selection and implement the regularization in the logistic regression model itself

[4] Curtis, P.G., C.M. Slay, N.L. Harris, A. Tyukavina, and M.C. Hansen. 2018. "Classifying Drivers of Global Forest Loss." Science. Accessed through Global Forest Watch on 02/12/2024. www.globalforestwatch.org.

as well, running a grid search to tune the regularization strength. We also implement a class weight inversely proportional to class frequency in an effort to accommodate class imbalance between detection and don-detection (eBird data tends to have many more non-detections than detections).

2. **Balanced Random Forest (BRF)**
   Our second model is a balanced random forest, very similar to the traditional random forest except that it draws a bootstrap sample from the minority class and samples the same number from the majority class in an effort to improve performance for a class-imbalanced dataset. We run a cross-validation grid search to tune the class-balancing hyperparameters class weight and sampling strategy (the ratio of the number of samples drawn from the minority class after resampling to the number of samples from the majority class), along with other random forest hyperparameters.
3. **Synthetic Minority Over-Sampling Technique (SMOTE)**
   In addition to the above efforts made to correct for class-imbalance, we also implement SMOTE to generate additional synthetic data from the minority class. We then re-run the above models, now passing in the (synthetically) oversampled data. Our SMOTE is conditional, however. If the data already has a ratio of minority to majority above the SMOTE threshold, we simply reduce to our original logistic regression and balanced random forest.

Evaluation

We compared our two models and their SMOTE-modified implementations with rolling cross-validation. To strike a balance between precision and recall for both classes (we are interested in performing well on both detections and non-detections in order to draw sound environmental and climate conclusions), we opt for the macro-averaged F1 score as our main performance metric. Macro-average is the simple average of the F1 scores for both classes. We also report the F1 score for just the detection class. This is a useful metric since, with strong class imbalance, performing well on the detection class is difficult while performing well on the non-detection class is fairly easy, even for dummy models. The performance on the detection class then offers another kind of insight into how our models have improved over baseline predictions.

Model	Average cross-validated performance	
	Macro F1	Detection F1
Uniform Dummy	0.407383	0.184598
Stratified Dummy	0.473880	0.061573
Log Reg	0.655341	0.409418
BRF	0.643427	0.413207
Log Reg w/ SMOTE	0.663651	0.423858
BRF w/ SMOTE	0.652909	0.422789

We compare model performance to two baseline dummy classifiers. Both dummies ignore the features, but the uniform dummy makes random guesses for each class with uniform frequency while the stratified dummy makes random guesses informed by the relative frequencies of the classes. The cross-validation scores for the ML classification models are shown in the accompanying table. Our ML classification models all perform fairly comparably, but the best, as measured by average cross-validated macro-averaged F1 score, is binary logistic regression with L1 regularization and SMOTE.

Occupancy Modeling with spOccupancy

$$
\begin{gathered}
y_i | z_i \sim \text{Bernoulli}\left(p \cdot z\right) \\
z_i \sim \text{Bernoulli}\left(\psi\right) \\
\text{logit}\left(p\right) = \alpha_0 + \sum_j \alpha_j \cdot A_j \\
\text{logit}\left(\psi\right) = \beta_0 + \sum_j \beta_j \cdot B_j \\
\text{with} \\
y_i = \text{data at site i} \\
p = \text{detection probability} \\
z_i = \text{true occupancy state at site i} \\
\psi = \text{occupancy probability} \\
\alpha_j = \text{model parameters relating detection probability and detection covariates } A_j \\
\beta_j = \text{model parameters relating occupancy probability and occupancy covariates } B_j
\end{gathered}
$$

We use spOccupancy[5] (an R library) to fit single–species and single–species integrated spatial occupancy models. We use these models to accommodate imperfect detection – when the species is present at (near) the site but it wasn't detected. The full (basic) model statement is shown to the left, but note that imperfection detection is accommodated

by the hierarchical structure and by the explicit modeling of the detection probability, *p*. In this way, occupancy models help distinguish occupancy probability from detection probability, something our more naive, traditional machine learning approach cannot do.

Our model is a layered logistic regression where the occupancy and detection random variables have Bernoulli distribution and conditional Bernoulli distribution respectively. These models are implemented in R using a Markov Chain Monte Carlo process. For the occupancy logit regression, our covariates are temperature, precipitation, forest cover loss, EVI (enhanced vegetation index), while for the detection logit regression, our covariates are day of the year, time of the day, and number of observations.

We first structured the detection non-detection data, the detection covariates, the occupancy covariates, and the coordinate matrix for the sites into the correct array formats that are needed for the spOccupancy models. We then fitted occupancy models with both temporal (tPGOcc) and spatial autocorrelation (stPGOcc). The temporal autocorrelation uses AR1 while the spatial autocorrelation uses the Nearest Neighbour Gaussian Process.

We also implemented posterior predictive checks on these models to evaluate the Bayesian p-value, which measures proportion of posterior samples of the fit statistic of the model generated data that are greater than the corresponding fit statistic of the true data, summed across all "grouped" data points. A Bayesian p-value around 0.5 indicates adequate model fit, while values less than 0.1 or greater than 0.9 suggest our model does not fit the data well.[6] We also evaluated the WAIC score [7] (Widely Applicable Information Criterion), where a lower value indicates better model fit.

Occupancy model	Bayesian p value grouped along sites (Freeman-Tukey statistic)	Bayesian p value grouped along replicates (Freeman-Tukey statistic)	WAIC (Widely Applicable Information Criterion)
tPGOcc (temporal autocorrelation)	0.0047	0.7594	3074.40
stPGOcc (spatial and temporal autocorrelation)	0.0015	0.0016	3074.03

Both the spOccupancy models performed comparably. The Bayesian p-value grouped along sites is not close to 0.5, which indicates that the model does not adequately represent variance in detection and occurrence probability across sites. See the notebook "Occupancy_modeling_with_SpOccupancy" for more details.

Results

The evaluation of our binary logistic regression model with L1 regularization and SMOTE is shown in the table to the right, compared with the baseline performance. We observe some decrease in performance relative to the cross validation score, but still reasonable improvement upon the baseline, particularly for the detection F1 score. The decrease in performance on the test set may be impacted by the class imbalance in the test set. The trend of detection frequency by year can be seen in the green line in the plot to the right. As our data is time-ordered, our test set is composed of the most recent 20% of the dataset. And because the data are more skewed toward recency (many more observations in more recent years than past years), the detection frequency in our test set is significantly lower than the

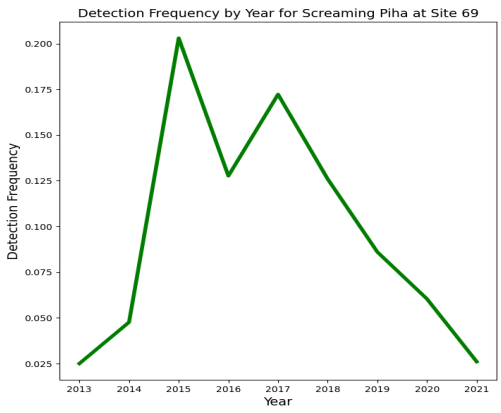
Model	Test set performance	
	Macro F1	Detection F1
Uniform Dummy	0.335315	0.027682
Stratified Dummy	0.467841	0.022727
Log Reg w/ SMOTE	0.566737	0.178571

[6] https://doserlab.com/files/spoccupancy-web/articles/modelfitting

[7] http://www.stat.columbia.edu/~gelman/research/published/waic_understand3.pdf

average detection frequency in our training set, which is skewed higher by the trend of greater detection frequency from, say, 2015-2018.

Relative to our logistic regression model's performance on the test set for the species and site with which it was selected, we find no significant decrease in performance when evaluating our model on the other three sites for Screaming piha, and in some cases moderate improvement. We find relatively consistent performance when evaluating on our second species, Black vulture. Compared to our model's performance on the test set for the species and site with which it was selected, only one site for Black vulture shows a markedly decreased performance. For full details, see the associated notebook 'ml_classification.'



Our L1 regularization and occupancy models in spOccupancy gave some insight into feature importance, with both types of models identifying as important the detection covariates of **effort hours, time of day, and number of observers**. Given that both models employ such different techniques, their agreement on the importance of these features seems significant. And the fact that these three common features are all detection covariates might suggest that the impact of environmental and climate features on species occupancy in our analysis is being washed out by the stronger effects of the features related to human observation efforts.

Conclusions, Shortcomings, and Future Directions

Broadly, we find binary logistic regression with L1 regularization and SMOTE to be our optimal traditional ML classification technique for site occupancy by Screaming pihas (a forest specialist) near Manaus, Brazil. We find this model generalizes reasonably well among nearby sites and for another species, the Black vulture (a generalist).

Based on both our L1 regularization and our SpOccupancy models, we find a common set of very important detection covariates: **effort hours, time of day, and number of observers**. The fact that both modelling approaches identify detection covariates as significant might suggest that climate change effects on occupancy and detection are overpowered by human effort effects in our analysis.

Some limitations and challenges of our modeling approach:

- Sites were clustered in the same region near Manaus, where there were the most eBird checklists over 2012 – 2021. Thus we have low variability in covariates between the sites. This might make it challenging for our models to generalize to more distant sites.
- A variable and sometimes very strong class imbalance between detections and non-detections.
- eBird data was skewed seasonally and over time, with the number of observations greatly increasing in more recent years.

Future directions to pursue:

- Test for more species over a broader area to get more variability in environmental covariates.
- Implement spatial correlation effects in our traditional ML models
- Find new ways to account for variable class imbalance