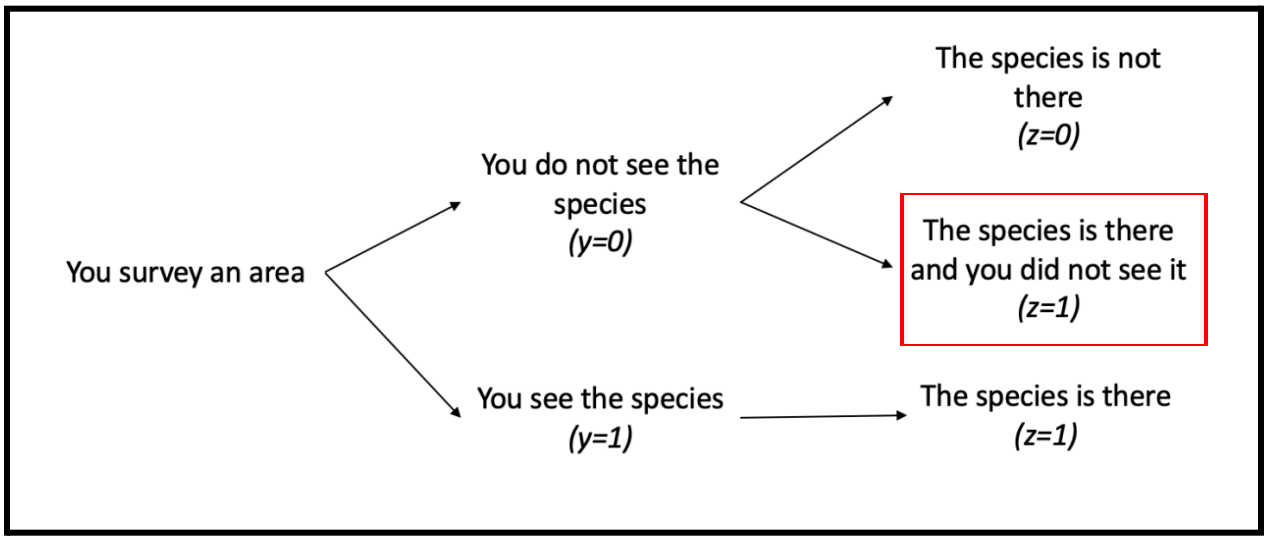
Occupancy Modeling Approach

Broadly, our aim is to use the ecological notion of the occupancy model to improve predictive capabilities regarding bird species occupancy in the Amazonas state. More traditional machine learning models do not account for imperfect detection - that is, if a species is not detected at a particular site then, as far as the model is concerned, the species truly is not there. On the other hand, modern occupancy modeling techniques attempt to account for imperfect detection - that is, the possibility that the species is present at the site but was not detected by the observer.



Observational flow chart relevant for occupancy modeling. The red box contains the possibility of imperfect detection. Variables y and z correspond to detection and occupancy, respectively. Taken from kevintshoemaker.github.io, produced by Morgan Byrne, James Golden.

Occupancy models accommodate imperfect detection by implementing a hierarchical, linked model. By this we mean the occupancy of a site appears in both the data modeling and the latent process modeling. The basic model statement can be written with two stochastic processes (Bernoulli) and link functions. In this way it is a sort of generalized linear model. [1]

$$y_i|z_i \sim Bernoulli(p \cdot z_i)$$
$$z_i \sim Bernoulli(\psi)$$
$$logit(p) = \alpha_0 + \alpha_1 \cdot covariate_1$$
$$logit(\psi) = \beta_0 + \beta_1 \cdot covariate_1$$

Where:
y_i = data at site i
p = detection probability
z_i = true occupancy state at site i
ψ = occupancy probability
α = parameters to estimate for the detection probability p
β = parameters to estimate for the occupancy probability ψ

[1] *Occupancy Modeling*, Morgan Bryne & James Golden, Nov 29, 2021. Course: Advanced Analysis Methods in Natural Resources and Environmental Sciences, Prof. Kevin Shoemaker, kevintshoemaker.github.io

To investigate the predictive power of the occupancy model and its accommodation of imperfect detection, we will implement several baseline models which do NOT accommodate imperfect detection.

- Baseline models NOT accounting for imperfect detection, however we could include some effort covariates. The naive approach would be to run a model for each species in our dataset.
    - Binary logistic regression, possibly with LASSO (L1) regularization
        - Good link for binary logistic regression example: https://library.virginia.edu/data/articles/logistic-regression-four-ways-with-python
    - Random Forest
        - https://cornelllabofornithology.github.io/ebird-best-practices/encounter.html#encounter-sss
            - This is a great site showing how to implement a random forest for strongly class-imbalanced dBird data (because non-detections are much more common than detections). It's done in R but there's solid random forest infrastructure in python, so this shouldn't be a problem. They're predicting encounter rate, rather than occupancy, but I think we could modify things appropriately to work with occupancy instead.

- Occupancy Models - will use extensively the R package spOccupancy.[2] I've included the relevant model names. I've selected three different types of models, each with a single-species version and a multi-species version (for a total of 6 models). Single-species models would be run individually for each species of interest. Multi-species models will be run for all species simultaneously (accounting for correlations among different species).
    - Single-Species, Multi-Season Models (run for each species)
        - tPGOcc() for single-species multi-season occupancy model - good basic occupancy model
        - stPGOcc() for single-species multi-season spatio-temporal occupancy model - more advanced model with spatial and temporal correlations
        - svcTPGOcc() for single-species spatially-varying coefficient multi-season occupancy model - even more advanced - SVC (spatially-varying-coefficient) occupancy models are a flexible extension of spatial occupancy models that allow for not only the intercept to vary across space, but also allows the effects of the covariates themselves to vary spatially, resulting in each spatial location having a unique effect of the covariate (Pease, Pacifici, and Kays 2022)

    - Multi-Species, Multi-Season Models (will allow for correlations between different species' occupancies)
        - tMsPGOcc() for multi-species, multi-season occupancy model
        - stMsPGOcc() for multi-species, multi-season spatial (temporal?) occupancy model
        - svcTMsPGOcc() for multi-species, multi-season, spatially-varying coefficient occupancy model

See the table below for an organized layout of the (five) types of models along with each type's lead
MODEL LEADS TO BE ASSIGNED AT NOV 8 MEETING

[2]Doser, J. W., Finley A. O., Kéry, M., & Zipkin E. F. (2022). spOccupancy: An R package for single-species, multi-species, and integrated spatial occupancy models Methods in Ecology and Evolution, 13, 1670-1678. https://doi.org/10.1111/2041-210X.13897

Model Class	Model Type	Single-Species Implementation	Multi-Species Implementation	Model Lead
ML	Binary logistic regression, possibly with LASSO regularization	Yes sklearn - includes built in L1 regularization	Possibly, might take some creativity	
	Random Forest	Yes sklearn	Possibly, might take some creativity	
		Single-Species Implementation	Multi-Species Implementation	
Occupancy	Multi-season (basic)	tPGOcc()	tMsPGOcc()	
	Multi-season, spatio-temporal	stPGOcc()	stMsPGOcc()	
	Muli-season, spatially-varying coefficient	svcTPGOcc()	svcTMsPGOcc()	