Presence Only Prediction - Literature Review

Benton Tripp

Table of contents

REQUIREMENTS

Timeline

8/21-9/3/2023

Overview

In an effort to understand existing research related to this topic, the first two weeks will be dedicated to exploring existing research, code libraries/repositories, and data.

Deliverable

A (somewhat informal) literature review citing 10-20 sources, covering each of the primary areas of interest in this project.

Table of Contents

Abstract

The continuous development and diversification of species distribution models (SDMs) necessitates a comprehensive examination of both conventional and emerging techniques. This literature review focuses on the application of presence-only prediction techniques, including Poisson point processes and MaxEnt, in the context of bird species distribution modeling. The review seeks to critically analyze the performance of traditional machine learning models such as xgboost, logistic regression, and random forests against the more specialized presence-only prediction techniques. Moreover, the exploration of Bayesian methods, data processing mechanisms, and innovative sampling methodologies aims to shed light on strategies to enhance model performance and reduce bias. The overarching goal of this review is to provide a systematic foundation for further research into the integration and optimization of various modeling strategies for presence-only predictions. Section Headings

MaxEnt and Poisson Point Processes in Species Distribution Modeling

Theoretical Underpinnings

Maximum Entropy (MaxEnt) and Poisson Point Processes (PPMs) have been identified as essential tools for species distribution modeling (SDM). The appeal of SDM lies in its potential to address significant questions, such as the impact of climate change on species distributions. Recent advancements in remote sensing, GIS, and computational power have further facilitated the development of these models (Renner & Warton, 2013; Thullier et al., 2008; O'Sullivan and Unwin, 2010).

MaxEnt models the probability per grid cell and analyzes data after aggregating them into presence/absence grid cells. In contrast, a Poisson PPM models the limiting expected count or intensity per unit area, rather than per grid cell. This per area basis, compared to per grid cell, is a significant distinction between the two approaches (Renner & Warton, 2013).

Interestingly, Renner & Warton (2013) highlight the mathematical equivalence of the MaxEnt procedure and Poisson regression, asserting that both approaches fit the same model and estimate parameters to maximize the same function up to a constant. Moreover, the MaxEnt and PPM solutions for grid cell data are proportional, with identical estimates of slope parameters.

Applications and Limitations

MaxEnt's applications are sometimes hindered by its shortcomings. It lacks clarity regarding diagnostic tools to assess model fit and is unclear about the spatial resolution when constructing grid cells (Renner & Warton, 2013).

A key limitation of MaxEnt is its scale dependence of predicted probabilities and arbitrary choice of spatial resolution. The per grid cell analysis is not invariant under choice of spatial resolution, unlike PPM, which models intensity on a per area basis (Renner & Warton, 2013). MaxEnt also fails to estimate the intercept consistently, diverging to $-\infty$ as spatial resolution increases (Renner & Warton, 2013; Elith et al., 2011).

PPM, on the other hand, offers various solutions to MaxEnt's problems. Predicted intensities in PPM are scale-invariant, and spatial resolution can be increased until log-likelihood converges. Various goodness-of-fit procedures are available for PPM, enabling more robust model adequacy assessment (Renner & Warton, 2013; Cressie, 1993; Baddeley et al., 2005).

Warton & Shepherd (2010) introduce PPMs as an alternative to pseudo-absence approaches, which have weaknesses in model specification, interpretation, and implementation. Point process modeling directly addresses these concerns, proposing a more sound specification for observed data without needing to generate new data. PPM also provides a framework for the selection of pseudo-absences, an area often tackled ad hoc in ecology (Warton & Shepherd, 2010).

Insights into Presence-Only Prediction

Presence-only prediction is a critical aspect of species distribution modeling. MaxEnt is limited in this regard due to its scale dependence and the current ambiguity over the spatial resolution (Renner & Warton, 2013).

PPM is proposed as a solution to the "pseudo-absence problem" in presence-only data, providing a more robust specification, clearer interpretation, and structured implementation than MaxEnt. The problems related to the pseudo-absence approach, specifically those associated with model specification, interpretation, and implementation, are rectified through the application of a point process modeling framework (Warton & Shepherd, 2010). Various authors have addressed the confusion over how pseudo-absences should be chosen (Elith and Leathwick, 2007; Guisan et al., 2007; Zarnetske, Edwards, and Moisen, 2007; Phillips et al., 2009), recognizing that the selection method can yield different outcomes (Chefaoui and Lobo, 2008).

Studies have demonstrated that logistic regression slope parameters and their corresponding standard errors converge to those of the Poisson point process model as the number of pseudo-absences is increased (Warton & Shepherd, 2010). This demonstrates that the PPM approach successfully addresses the arbitrary nature of pseudo-absence selection, signifying that a specific form of point process model is being estimated, even in the utilization of pseudo-absence methods. Current selection procedures for pseudo-absences do not align with best practices, as they are frequently chosen at random and lack a basis in convergence criteria (Pearce and Boyce, 2006; Zarnetske, Edwards, and Moisen, 2007).

The exploration of MaxEnt and PPM in species distribution modeling reveals intriguing similarities and critical differences between these two methods. While MaxEnt is challenged by its shortcomings, including scale dependence and lack of consistent intercept estimation, PPM offers robust solutions, especially for presence-only prediction. The equivalence between MaxEnt and PPM and the insights into the pseudo-absence problem signify a notable contribution to ecological modeling, pointing to the potential for further refinements and innovations in the field.

References

Renner, I., & Warton, D. I. (2013). Equivalence of MAXENT and Poisson Point process models for species distribution modeling in Ecology. Biometrics, 69(1), 274–281. https://doi.org/10.1111/j.1541-0420.2012.01824.x

Warton, D. I., & Shepherd, L. (2010). Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. The Annals of Applied Statistics, 4(3). https://doi.org/10.1214/10-aoas331

Sources/Notes (To be Updated)

Article 1

Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology

> One potential reason for such high interest is that SDM aims to address important topical questions such as the potential effects of climate change on species distributions (Thullier et al. 2008). Rapid progress in this field has been facilitated by recent significant technological advances in remote sensing, GIS (O'Sullivan and Unwin, 2010), and computational power, enabling models to be built at increasingly fine resolutions and increasingly large spatial scales.
>
> MAXENT has a number of shortcomings… In particular, it is unclear what diagnostic tools may be used to assess whether the fitted model is reasonable. Morover, MAXENT analyzes data after first aggregating them into presence/absence grid cells, and it is currently unclear what spatial resolution should be used when constructing these cells.
>
> The MAXENT procedure and Poisson regression are equivalent. That is, 1. They fit the same model:
>
> $$\ln \pi(g_i) = \ln \mu(g_i) = x(g_i)'\beta$$
>
> 2. They estimate parameters to maximize the same function up to a constant:
>
> $$\Lambda\{\beta; z(n)(g)\} = l\{\beta; z(n)(g)\} + C$$
>
> where C is a constant and $\Lambda\{\beta; z(n)(g)\}$ is the Lagrangian function to maximize entropy $H\{\pi(g)\}$ subject to the constraints stated in equations (1)–(2). Hence the maximum entropy estimate $\hat{\beta}_{MAXENT}$ equals the maximum likelihood estimate from Poisson regression $\hat{\beta}_{GLM}$.
>
> A Poisson point process regression model (PPM) analyzes m presence-only locations $y_P = \{y_1, ..., y_m\}$ as a point process in which the locations of the m points are assumed to be independent. Unlike MAXENT, which models probability $\pi(g_i)$ per grid cell, a Poisson PPM models the limiting expected count ($\lambda(y)$, the "intensity") per unit area (Cressie, 1993) for any location $y \in A$. Intensity is modeled as a log-linear function of p explanatory variables: $\ln\{\lambda(y)\} = x(y)'\beta$. An analysis on a per area basis rather than a per grid cell basis is a key distinction between a point process model and MAXENT.
>
> Consider a Poisson point process model fitted to grid cell data $z(n)(g)$, with parameter estimates stored in $\hat{\beta}_{PPM}$. Then:
>
> $$\hat{\beta}_{MAXENT} = \hat{\beta}_{PPM} + JC$$

where $JC = \{\ln C, 0, ..., 0\}$ is a vector of length $p + 1$, and $C = \frac{|A|}{m(n)n}$.In other words, the MAXENT and PPM solutions for grid cell data are proportional, and estimates of slope parameters are identical.

Likelihood of observing a presence point depends not just on the spatial distribution of the species, but also on the spatial distribution of observers, which is strongly affected by site accessibility. The underlying assumption of a Poisson point process model (and by equivalence, MAXENT) is that the point locations are independent, conditional on model covariates… While MAXENT offers no method for checking this assumption, there are a number of diagnostic tools to assess model adequacy of a Poisson point process model (Cressie, 1993; Baddeley et al., 2005). One such method is to construct the inhomogeneous K-function (Baddeley, Møller, and Waagepetersen, 2000) and corresponding simulation envelope (Diggle, 2003) of the fitted model. Current problems with MAXENT and their proposed solutions available through re-expression as a Poisson point process model:

MAXENT problem	Poisson PPM solution
Predicted probabilities are scale-dependent	Predicted intensities are scale-invariant
How to determine spatial resolution?	Increase until log-likelihood converges
How to assess model adequacy?	Various goodness-of-fit procedures available
How to choose LASSO parameter?	Various data-driven methods
Available in MAXENT software only	Use any standard GLM software
130 seconds to fit models in Figure 1b	12 seconds to fit models in Figure 1b

Warton and Shepherd (2010) showed the equivalence of Poisson point process models and pseudo-absence regression, which aside from MAXENT is the most commonly used approach to presence-only modeling at the moment.

A key distinction between point process models and MAXENT is that in the former we model $lambda(y)$ on a per area basis whereas for the latter, we model $pi(g_i)$ per grid cell—the per area analysis is thus invariant under choice of spatial resolution while the per grid cell analysis is not (because increasing spatial resolution increases the number of grid cells). This is related to the distinction between probability and frequency models (Aarts et al., 2012). It is this distinction that enables the likelihood convergence for a Poisson point process model (Figure 2a) and hence a data-driven choice of spatial resolution. However, MAXENT is proportional to a Poisson point process model (Theorem 2), which suggests that it can achieve the same qualitative answer but with the disadvantage of scale dependence of the predicted probabilities and an arbitrary choice of spatial resolution.

One important disadvantage of MAXENT is that in its current form, it does not estimate the intercept consistently (Elith et al., 2011). The intercept term diverges to $-\infty$ as spatial resolution increases.

We advise that as a matter of routine, presence-only analysts should use their data to identify a spatial resolution appropriate for analysis, or equivalently, to identify the number of "background points" to use in analysis.

Article 2

Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology

Currently, ecologists most commonly analyze presence-only data by adding randomly chosen "pseudo-absences" to the data such that it can be analyzed using logistic regression, an approach which has weaknesses in model specification, in interpretation, and in implementation. To address these issues, we propose Poisson point process modeling of the intensity of presences. We also derive a link between the proposed approach and logistic regression—specifically, we show that as the number of pseudo-absences increases (in a regular or uniform random arrangement), logistic regression slope parameters and their standard errors converge to those of the corresponding Poisson point process model.

Point process modeling offers a framework for choice of the number and location of pseudo-absences, both of which are currently chosen by ad hoc and sometimes ineffective methods in ecology.

We see three key weaknesses of the "pseudo-absence" approach so widely used in ecology for analyzing presence-only data, which we describe concisely as problems of model specification, interpretation, and implementation. A sounder model specification would involve constructing a model for the observed data y only, rather than requiring us to generate new data y_0 prior to constructing a model. Interpretation of results is difficult, because some model parameters of interest are a function of the number of pseudo-absences and their location. Implementation of the approach is problematic because it is unclear how pseudo-absences should be chosen [Elith and Leathwick (2007); Guisan et al. (2007); Zarnetske, Edwards and Moisen (2007); Phillips et al. (2009)], and one can obtain qualitatively different results depending on the method of choice of pseudo-absences [Chefaoui and Lobo (2008)].

We propose point process models as an appropriate tool for species distribution modeling of presence-only data, given that presence-only data arise as a set of point events—a set of locations where a species has been reported to have been seen. A point process model specification addresses each of the three concerns raised above

regarding pseudo-absence approaches. Our second key contribution is a proof that the pseudo-absence logistic regression approach, when applied with an increasing number of regularly spaced or randomly chosen pseudo-absences, yields estimates of slope parameters that converge to the point process slope estimates.

We consider inhomogeneous Poisson point process models [Cressie (1993); Diggle (2003)], which make the following two assumptions: 1. The locations of the n) point events $(y_1, ..., y_n)$ are independent. 2. The intensity at point y_i, $\lambda(y_i)$ (denoted as λ_i for convenience), the limiting expected number of presences per unit area [Cressie (1993)], can be modeled as a function of the k explanatory variables. We assume a log-linear specification:

$$\log(\lambda_i) = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j$$

Although note that the linearity assumption can be relaxed in the usual way (e.g., using quadratic terms or splines). The parameters of the model for the λ_i are stored in the vector $\beta = (\beta_0, \beta_1, ..., \beta_k)$.

Pseudo-absence points of presence-only logistic regression play the same role as quadrature points of a point process model, and so established guidelines on how to choose quadrature points can inform choice of pseudo-absences. Previously pseudo-absences have been generated according to ad hoc recommendations [Pearce and Boyce (2006); Zarnetske, Edwards and Moisen (2007)], given the lack of a theoretical framework for their selection. In contrast, quadrature points are generated in order to estimate the log-likelihood to a pre-determined level of accuracy, a criterion which guides the choice of locations and numbers of quadrature points $m-n$. Interestingly, current methods of selecting pseudo-absences in ecology [Pearce and Boyce (2006); Zarnetske, Edwards and Moisen (2007)] do not appear to be consistent with the best practice in low-dimensional numerical quadrature—points are usually selected at random rather than on a regular grid, and the number of pseudo-absences $(m - n)$ is more commonly chosen relative to the magnitude of the number of presences (n) rather than based on some convergence criterion.

Despite the apparent ad hoc nature of the pseudo-absence approach, some form of point process model is being estimated. However, logistic regression is only equivalent to a Poisson point process when $w = 1$, that is, all quadrature weights are ignored.

The pseudo-absence approach has problems with model specification, interpretation, and implementation. We argue that each of these difficulties is resolved by using a point process modeling framework.

1. Model specification - We believe that a point process model is a plausible model for the data generation mechanism for presence-only data. In contrast, the logistic regression approach involves generating new data in order to fit a model originally designed for a different problem (analysis of binary data not analysis of point-events). Hence, the pseudo-absence approach as it is usually applied appears to involve coercing the data to fit the model rather than choosing a model that fits the original data.

2. Interpretation — In the logistic regression approach we model p_i, the probability that a given point event is a presence not a pseudo-absence. This quantity has no physical meaning and clearly its interpretation is sensitive to our method of choice of pseudo-absences (and typically each $p_i \to 0$ as $m \to \infty$). In contrast, the intensity at a point λ_i has a natural interpretation as the (limiting) expected number of presences per unit area, and will not be sensitive to choice of quadrature points, provided that the number of quadrature points is sufficiently large.

3. Implementation — Point process models offer a framework for choosing quadrature points. Specifically, [the point process log-likelihood is estimated], and progressively more quadrature points are added until [convergence is achieved]. No such framework for choice of pseudo-absences is offered by logistic regression, and instead choice of the location and number of pseudo-absences is ad hoc, with potentially poor results. Ecologists are concerned about the issues of how many pseudo-absences to choose [Pearce and Boyce (2006)], where to put them [Elith and Leathwick (2007); Zarnetske, Edwards and Moisen (2007); Phillips et al. (2009)], and what spatial resolution to use in model-fitting [Guisan et al. (2007); Elith and Leathwick (2009)], all issues that have natural solutions given a point process model specification of the problem.

Article 3

Maximum entropy modeling of species geographic distributions

Sampling, Bias, and Pseudo-Absence Data

Article 4

On pseudo-absence generation and machine learning for locust breeding ground prediction in Africa

…We compare this random sampling approach to more advanced pseudo-absence generation methods, such as environmental profiling and optimal background extent limitation, specifically for predicting desert locust breeding grounds in Africa. Interestingly, we find that for the algorithms we tested, namely logistic regression, gradient boosting, random forests and MaxEnt, all popular in prior work, the linear logistic model performed significantly better than the more sophisticated ensemble methods, both in terms of prediction accuracy and F1 score. Although background extent limitation combined with random sampling seemed to boost performance for ensemble methods, no statistically significant differences were detected between the pseudo-absence generation methods used to train the logistic model. In light of this, we conclude that simpler approaches such as random sampling for pseudo-absence generation combined with linear classifiers such as logistic regression are sensible and effective for predicting locust breeding grounds across Africa.

Machine learning (ML) has been shown to be a valuable tool for species distribution modeling (Beery et al., 2021). However, even when remote sensing is capable of providing useful features for such models, ML still heavily relies on large amounts of labelled data for training.

Article 5

Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data

We evaluated the effect of using a) real absences b) pseudo-absences selected randomly from the background and c) two-step approaches: pseudo-absences selected from low suitability areas predicted by either Ecological Niche Factor Analysis: (ENFA) or BIOCLIM. We compared how the choice of pseudo-absence strategy affected model fit, predictive power, and information-theoretic model selection results.

Conclusion: If ecologists wish to build parsimonious GLM models that will allow them to make robust predictions, a reasonable approach is to use a large number of randomly selected pseudo- absences, and perform model selection based on an information theoretic approach. However, the resulting models can be expected to have limited fit.

Two groups of techniques are generally used [to relate field observations to environmental predictor variables, based on statistically or theoretically derived response surfaces, for prediction and inference]. Techniques that require data documenting the species presence only are called “profile techniques” while those that require both presence and absence data are called “group discrimination techniques”. Examples of profile techniques include BIOCLIM, DOMAIN, Species-PCA, and Ecological Niche Factor Analysis (ENFA).

Among group-discrimination techniques, logistic regression modelling (LRM), a particular branch of generalized linear models (GLM) for binary responses, remains the most widely used so far to predict the potential distributions of species.

Despite its numerous advantages, LRM has been precluded from many studies of species distributions because it requires absence data, which are frequently unavailable and often not reliable. This is an acute problem for the study of poorly documented, cryptic, rare or highly mobile species, many of which may be of special conservation interest.

Given an adequate set of presence locations, models generated with random pseudo-absences can yield useful results. A potential drawback to using random pseudo-absences is that pseudo-absences might coincide with locations where the species actually occurs. This affects the calculation of probability of presence, and consequently, models built with random pseudo-absences are expected to have poorer fit, and lower predictive performance than models built with real absences.

Engler et al. used a two step modelling approach to predict the distribution of a rare plant in Switzerland. In the first step, they used the profile technique "ENFA" to map potential habitat suitability for the species, and then selected pseudo-absences from the areas predicted to have low suitabilty. They subsequently included the pseudo-absences in a second logistic regression model which was used to predict the final species' potential distribution and prioritize further field work.

In a virtual species approach, the species' distribution is defined *a priori*, by specifying its ecological niche as a simple mathematical relationship to the set of predictor environmental variables (e.g. in the form of a multiple logistic regression equation) and projects this relationship onto a map of the study area to define its "true" distribution. One can then attempt to recover the virtual species' known distribution by building models from samples drawn from the study area and then evaluate the efficacy of different implementations of models by comparing their parameter estimates and predictions to the true distribution of the species. Such an approach has never been used to assess the best way to select pseudo-absences.

Whenever a two-step modelling approach was used, the true model was only weakly supported by the data.

Evaluating so many aspects of model's behaviour and performance is facilitated through use of a virtual species, because a real species' "true" relationship to environmental variables is never known. In the real world, sufficient amounts of reliable, completely independent, presence and absence data are rarely available to evaluate the predictive power of complex models in a controlled manner.

A large body of statistical literature suggests that parsimonious models, including a small number of predictor variables, should have greater predictive power to independent samples than models including more predictors. Neither ENFA nor BIOCLIM are based on this principle (i.e. they do not incorporate way to select predictors), and both showed somewhat lower predictive power in terms of AUC than group discrimination models based on pseudo-absences… BIOCLIM and ENFA cannot incorporate quadratic relationships to predictors, so this may be another reason the GLM methods performed better.

…With small datasets, models for which complexity is calibrated for sample size, such as MAXENT have better predictive power than models that use complex response shapes regardless of sample size (Wisz et al. 2008).

Through the use of a simulated species we confirmed that although randomly selected pseudo-absences yield models with lower fit to the training data, they outperform models developed from two-step methods in terms of predictive power and variable selection. Thus randomly selected pseudo-absences may be a reasonable alternative when real absences are unavailable.

Two step pseudo-absences result in models with weaker predictive power because they lead to overfitting. An overfit model will always have a higher value of adjusted deviance than a simpler model nested within it, but its predictive power to an independent sample will be lower because the model losses generality.

Model selection, the process of selecting predictors and parameter estimates in a model, is considered one of the most crucial steps in a model building procedure.

Article 6

Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data

Article 7

Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data

Article 8

Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling

Article 9

The importance of correcting for sampling bias in MaxEnt species distribution models

Machine Learning and Predictive Models

The use of classification and regression algorithms using the random forests method with presence-only data to model species' distribution

Article 10

Predictive performance of presence-only species distribution models: a benchmark study with reproducible code

Article 11

Breeding density, fine-scale tracking, and large-scale modeling reveal the regional distribution of four seabird species

Article 12

Novel methods improve prediction of species' distributions from occurrence data

Article 13

Predicting species distribution: Offering more than simple habitat models

Bayesian Modeling

Article 14

Using a Bayesian modelling approach (INLA-SPDE) to predict the occurrence of the Spinetail Devil Ray (Mobular mobular)

Article 15

Bayesian Modeling and MCMC Computation in Linear Logistic Regression for Presence-only Data

Other Sources

- Inference from presence-only data; the ongoing controversy
- Geostatistical Inference Under Preferential Sampling
- Understanding the connections between species distribution models
- PULasso: High-dimensional variable selection with presence-only data
- Bias Correction in Species Distribution Models: Pooling Survey and Collection Data for Multiple Species
- Nondetection sampling bias in marked presence-only data
- Generalized linear and generalized additive models in studies of species distributions: setting the scene
- Sensitivity of species-distribution models to error, bias, and model design: An application to resource selection functions for woodland caribou
- Toward the modeling of true species distributions
- Does accounting for imperfect detection improve species distribution models?
- Finite-sample equivalence in statistical models for presence-only data
- New spatial models for integrating standardized detection-nondetection and opportunistic presence-only data: application to estimating risk factors associated to powerline-induced death of birds
- Modelling distribution and abundance with presence-only data

---

Old Notes

Preferential Sampling Considerations

Summary of the article, Geostatistical Inference Under Preferential Sampling:

1. Concept of Preferential Sampling: The article discusses the concept of preferential sampling, where the locations of data points are often selected with a preference for values of the process that are either unusually large or small. In your project, this could translate to the tendency to have more data from areas where bird species are more commonly found or are easier to observe. Understanding and accounting for this bias could improve the accuracy of your models.
2. Model-Based Approach: The authors propose a model-based approach to geostatistical inference when the data are preferentially sampled. This methodology could potentially be applied in your project to account for any bias in the geographical distribution of your bird observation data.
3. Likelihood-Based Approach to Inference: The authors present a likelihood-based approach to inference for the proposed model. This statistical approach could be useful in your project when you are comparing different machine learning algorithms for modeling bird species distribution.
4. Bias Correction: The authors found that ignoring preferential sampling can lead to biased estimates of risk. In your project, this could translate to biased predictions of bird species presence. The methodology proposed in the article could potentially be used to correct this bias.
5. Extension to Spatio-Temporal Processes: The authors suggest that their approach can be extended to spatio-temporal processes. This could be particularly relevant to your project, as bird species presence is likely to vary both spatially and temporally.

Ideas for implementation:

- Pre-Processing - Observation Data Pre-Processing: During the pre-processing of your bird observation data, you could consider whether there is any bias in the locations of your data points. For example, are there more data points in areas where bird species are more commonly found or easier to observe? If so, you might need to account for this bias in your models. This could involve adjusting the weights of your points or using a model-based approach to account for preferential sampling.
- Sampling and Splitting - Pseudo-Absence Selection: When selecting pseudo-absence points, you could consider whether the locations of these are independent of the bird species presence process. If not, you might need to adjust your selection process to account for this dependence. This could involve using a likelihood-based approach to inference, as suggested in the article.
- Modeling: When comparing different machine learning algorithms for modeling bird species distribution, you could consider whether any of these algorithms are more or less susceptible to bias due to preferential sampling. If so, you might need to adjust your

model selection process accordingly. This could involve using a model-based approach to account for preferential sampling, as suggested in the article.

- Prediction: When making predictions on the presence of each bird species across the entire US, you could consider whether your predictions are biased due to preferential sampling. If so, you might need to adjust your prediction process to correct this bias. This could involve using a model-based approach to account for preferential sampling, as suggested in the article.

Pseudo-Absence Generation Considerations

Sumary of the article, On pseudo-absence generation and machine learning for locust breeding ground prediction in Africa:

The article titled "On pseudo-absence generation and machine learning for locust breeding ground prediction in Africa" by authors Ibrahim Salihu Yusuf, Kale-ab Tessera, Thomas Tumiel, Zohra Slim, Amine Kerkeni, Sella Nevo, and Arnu Pretorius, was published in the AI for Humanitarian Assistance and Disaster Response (AI+HADR) workshop, NeurIPS 2021.

The paper discusses the threat of desert locust outbreaks to the food security of a large part of Africa and the impact on the livelihoods of millions of people. The authors propose machine learning (ML) as an effective approach to locust distribution modeling, which could assist in early warning.

However, ML requires a significant amount of labeled data to train. Most publicly available labeled data on locusts are presence-only data, where only the sightings of locusts being present at a location are recorded. Therefore, prior work using ML has resorted to pseudo-absence generation methods as a way to circumvent this issue.

The most commonly used approach is to randomly sample points in a region of interest while ensuring that these sampled pseudo-absence points are at least a specific distance away from true presence points. In this paper, the authors compare this random sampling approach to more advanced pseudo-absence generation methods, such as environmental profiling and optimal background extent limitation, specifically for predicting desert locust breeding grounds in Africa.

Interestingly, they find that for the algorithms they tested, namely logistic regression, gradient boosting, random forests, and maximum entropy, all popular in prior work, the logistic model performed significantly better than the more sophisticated ensemble methods, both in terms of prediction accuracy and F1 score.

Although background extent limitation combined with random sampling boosted performance for ensemble methods, for logistic regression this was not the case, and instead, a significant improvement was obtained when using environmental profiling.

In light of this, the authors conclude that a simpler ML approach such as logistic regression combined with more advanced pseudo-absence generation, specifically environmental profiling, can be a sensible and effective approach to predicting locust breeding grounds across Africa.

Ideas for implementation:

1. Pseudo-Absence Selection
    - In your project, you are generating pseudo-absence points by selecting points where other bird species have been observed but not the target species. However, the paper suggests that more advanced pseudo-absence generation methods, such as environmental profiling, can improve model performance. You could consider implementing environmental profiling in your pseudo-absence selection process. This would involve selecting pseudo-absence points based on the environmental characteristics of the areas where the target species has not been observed. For example, if a certain bird species is known to prefer forested areas, pseudo-absence points could be selected from non-forested areas.
    - Implementation: Research and implement environmental profiling for pseudo-absence selection. This might involve adjusting your current pseudo-absence selection mechanism to select points based on environmental characteristics rather than just the absence of the target species.
2. Model Selection
    - Your project involves a comparative analysis of several machine learning algorithms. The paper found that a simpler machine learning approach, such as logistic regression, performed significantly better than more sophisticated ensemble methods when combined with advanced pseudo-absence generation. Therefore, you might want to give more consideration to simpler models like logistic regression in your analysis.
    - Implementation: Include logistic regression in your comparative analysis of machine learning algorithms. Evaluate its performance in combination with your new pseudo-absence selection method.
3. Model Evaluation
    - The paper used prediction accuracy and F1 score to evaluate model performance. In addition to the performance metrics you are currently using (partial ROC, F1, precision, recall, and accuracy), you might consider using other metrics that are more suitable for imbalanced datasets, such as the Area Under the Precision-Recall Curve (AUPRC).
    - Implementation: Incorporate additional performance metrics suitable for imbalanced datasets, such as AUPRC, into your model evaluation process.
4. Data Pre-processing

- The paper emphasizes the importance of proper data pre-processing, especially when dealing with presence-only data. You might want to review your data pre-processing steps to ensure that they are suitable for your presence-only bird species data.
- Implementation: Review your data pre-processing steps to ensure they are suitable for presence-only data. Make any necessary adjustments based on best practices for dealing with presence-only data.

Bayesian Methods

Sumary of the article, Bayesian Modeling and MCMC Computation in Linear Logistic Regression for Presence-only Data:

The article titled "Bayesian Modeling and MCMC Computation in Linear Logistic Regression for Presence-only Data" by authors Jinfeng Xu, Jun Zhu, and Xiaoyue Niu, focuses on the development and application of Bayesian modeling and Markov Chain Monte Carlo (MCMC) computation in the context of linear logistic regression for presence-only data.

The authors start by acknowledging the challenges in analyzing presence-only data, which is a type of data where only the presence of a species or event is recorded, but not its absence. This type of data is common in ecological studies, among others.

To address these challenges, the authors propose a Bayesian modeling approach. Bayesian modeling is a statistical method that incorporates prior knowledge about a parameter into the analysis. This approach is particularly useful when dealing with complex models and limited data.

In addition to Bayesian modeling, the authors also utilize Markov Chain Monte Carlo (MCMC) computation. MCMC is a method for estimating the distribution of a variable of interest by constructing a Markov chain that has the desired distribution as its equilibrium distribution. This method is often used in Bayesian analysis to estimate posterior distributions.

The authors apply these methods to linear logistic regression, a statistical method used to model the relationship between a binary response variable and one or more explanatory variables. They provide a detailed explanation of their methodology, including the specification of the prior distribution, the construction of the likelihood function, and the implementation of the MCMC algorithm.

The authors also present a case study to demonstrate the application of their methodology. They use presence-only data from the Giant Panda distribution in the Minshan Mountains to estimate the effects of various environmental factors on the presence of the species. The results show that their method provides a robust and reliable estimation of the parameters of interest.

The authors conclude by discussing the advantages and potential limitations of their approach. They suggest that their method can be extended to other types of presence-only data and encourage further research in this area.

Ideas for implementation:

1. Model Selection: You mentioned that you plan to compare several machine learning algorithms, including Logistic Regression, LightGBM, XGBoost, KNN, Random Forest, Maximum Entropy, MCMC, and SPDE/INLA. The Bayesian modeling and MCMC computation approach can be used to enhance your logistic regression model. This approach can provide a more nuanced understanding of the underlying data and can be particularly useful when dealing with complex models and limited data.
2. Handling Presence-Only Data: Your project involves dealing with presence-only data, which is a common challenge in ecological studies. The Bayesian modeling and MCMC computation approach can be used to handle this type of data more effectively. Specifically, the approach can be used to estimate the parameters of interest in your logistic regression model, providing a more robust and reliable estimation.
3. Pseudo-Absence Selection: The Bayesian modeling and MCMC computation approach can be used to improve your pseudo-absence selection mechanism. Specifically, the approach can be used to estimate the distribution of pseudo-absence points, which can help to address the inherent challenge of absence data in presence-only datasets.
4. Model Evaluation: The Bayesian modeling and MCMC computation approach can also be used to evaluate the performance of your models. Specifically, the approach can be used to estimate the posterior distributions of your performance metrics, providing a more comprehensive evaluation of your models.
5. Predictive Modeling: Lastly, the Bayesian modeling and MCMC computation approach can be used to enhance your predictive modeling. Specifically, the approach can be used to estimate the probabilities of bird species presence across the entire US, providing a more accurate and reliable prediction.

In summary, the Bayesian modeling and MCMC computation approach can be used to enhance various aspects of your project, from model selection and handling presence-only data to pseudo-absence selection, model evaluation, and predictive modeling. By incorporating this approach into your project, you can potentially improve the accuracy and reliability of your results, providing a more nuanced understanding of avian habitat tendencies.

Finite-sample Equivalence in Statistical Models

Summary of the article, Finite-sample equivalence in statistical models for presence-only data:

The article "Finite-sample equivalence in statistical models for presence-only data" by William Fithian and Trevor Hastie discusses statistical modeling of presence-only data, which has gained significant attention in ecological literature. The authors focus on three methods: the

inhomogeneous Poisson process (IPP) model, maximum entropy (Maxent) modeling of species distributions, and logistic regression models. They explain why the IPP intensity function is a more natural object of inference in presence-only studies than occurrence probability.

The authors show that IPP and Maxent give the exact same estimate for density, but logistic regression generally yields a different estimate in finite samples. When the model is misspecified, logistic regression and the IPP may have substantially different asymptotic limits with large data sets. The authors propose "infinitely weighted logistic regression," which is exactly equivalent to the IPP in finite samples. Consequently, many already-implemented methods extending logistic regression can also extend the Maxent and IPP models in directly analogous ways using this technique.

The paper also discusses the challenges of sampling bias in presence-only studies and proposes a model for the sightings process as an occurrence process thinned by incomplete observation.

Ideas for implementation:

1. Model Selection and Interpretation: The article suggests that the inhomogeneous Poisson process (IPP) model and maximum entropy (Maxent) modeling yield the same estimates for density in finite samples. This could influence your choice of models to compare in your project. If computational resources are limited, you might choose to focus on one of these two models, rather than both, given their equivalence in finite samples.
2. Logistic Regression Adjustments: The authors propose an "infinitely weighted logistic regression," which is exactly equivalent to the IPP in finite samples. If you're using logistic regression in your project, you might consider adjusting it to this infinitely weighted version for better equivalence with the IPP and Maxent models.
3. Handling Sampling Bias: The paper discusses the challenges of sampling bias in presence-only studies and proposes a model for the sightings process as an occurrence process thinned by incomplete observation. You might consider incorporating this approach into your data pre-processing steps, particularly when selecting pseudo-absence points. For example, you could adjust the selection mechanism to account for potential sampling bias in the Project FeederWatch Observation Data.
4. Extending Logistic Regression: The authors suggest that many already-implemented methods extending logistic regression can also extend the Maxent and IPP models in directly analogous ways. This could be particularly useful if you're considering extending your models to account for additional complexities in the data. For example, you might consider using regularized logistic regression or other extensions to better handle high-dimensional feature spaces.
5. Asymptotic Behavior: The paper discusses the different asymptotic limits of logistic regression and the IPP when the model is misspecified. If you're working with large data sets, you might want to consider the potential implications of these different asymptotic behaviors on your model's performance and robustness.
6. Model Evaluation: The authors' findings about the equivalence of certain models in finite samples could also influence your model evaluation strategy. For example, you might

choose to focus on comparing the performance of models that are not equivalent in finite samples, such as logistic regression and the IPP.

Bias Correction using Species Pooling

Summary of the article, Bias Correction in Species Distribution Models: Pooling Survey and Collection Data for Multiple Species:

The article "Bias Correction in Species Distribution Models: Pooling Survey and Collection Data for Multiple Species" by Guisan et al. discusses a novel approach to correct bias in species distribution models (SDMs). The authors recognize that SDMs are often biased due to the nature of the data collected, which is typically skewed towards easily detectable species and areas that are easy to access.

To address this issue, the authors propose a new method that combines survey and collection data from multiple species. This approach allows for a more comprehensive understanding of species distribution, as it incorporates data from a wider range of sources and species.

The authors tested their method using data from the Swiss Breeding Bird Survey and the Swiss Ornithological Institute. The results showed that their approach significantly reduced bias in the SDMs, leading to more accurate predictions of species distribution.

The authors conclude that their method can be a valuable tool for ecologists and conservationists, as it can improve the accuracy of SDMs and thus inform more effective conservation strategies. They also suggest that their approach could be applied to other types of ecological data, opening up new possibilities for bias correction in ecological research.

Ideas for implementation:

1. Data Collection and Pooling: The article emphasizes the importance of pooling data from multiple sources to reduce bias. In your project, you are already using data from multiple sources, which is a good practice. However, you might consider expanding the range of data sources if possible. For instance, you could look for additional bird observation databases or environmental datasets that could provide more comprehensive coverage of the species and environmental factors you are interested in.
2. Bias Correction: The article presents a method for correcting bias in species distribution models by combining survey and collection data. In your project, you could implement a similar approach to correct for potential biases in your data. For example, you could adjust your models to account for the fact that bird observations are likely to be more frequent in areas that are easily accessible to humans, which could skew the apparent distribution of species.
3. Modeling and Prediction: The article's approach to reducing bias could also improve the accuracy of your predictive models. By ensuring that your models are based on a balanced and comprehensive dataset, you can increase the reliability of your predictions about bird species distribution across the United States. This could be particularly

useful when comparing the performance of different machine learning algorithms, as it would allow you to evaluate them on a fair and unbiased basis.

4. Conservation Implications: The article highlights the potential conservation implications of accurate species distribution models. In your project, you could use your improved models to identify areas where bird species are likely to be found but are currently under-surveyed. These areas could be targeted for future conservation efforts, helping to protect species that might otherwise be overlooked.
5. Pseudo-Absence Selection: The article's method of pooling data from multiple species could be adapted to improve your pseudo-absence selection mechanism. By considering the presence of other species when selecting pseudo-absence points, you could ensure that these points represent areas that are genuinely unsuitable for the species of interest, rather than just areas where it has not been observed.
6. Model Evaluation: The bias correction method proposed in the article could also be used to refine your model evaluation process. By assessing how well your models correct for bias, you could gain additional insights into their performance and reliability. This could be incorporated into your existing performance metrics, such as partial ROC, F1, precision, recall, and accuracy.