

# Presence Only Prediction - Literature Review

Benton Tripp

## 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.

## Introduction

### Background and Significance

Species distribution modeling (SDM) has become increasingly crucial in the realm of ecological research. Driven by advancements in technology, remote sensing, and computational capabilities, there has been a proliferation of methods to predict species distributions (Renner & Warton, 2013; Thullier et al., 2008; O’Sullivan and Unwin, 2010). These models often center around unique datasets, such as presence-only data, which poses unique challenges for statistical modeling due to inherent uncertainty from data censoring and specific sampling procedures.

In this context, presence-only data is often defined as datasets “consisting only of observations of the organism but with no reliable data where the species was not found” (Pearce and Boyce, 2006). Such data forms the backbone of various sources, including atlases, museum and herbarium records, and incidental observation databases (Divino, 2013).

Two primary modeling approaches for this type of data have gained traction. The first leverages Poisson point processes in both likelihood and Bayesian frameworks, while the second adopts a modified case-control logistic model (Warton & Shepherd, 2010; Chakraborty et al., 2011; Ward et al., 2009). Another technique, MaxEnt, based on the maximum entropy principle, requires knowledge of species' prevalence for its estimator of occurrence to be consistent (Dorazio, 2012).

Recently, traditional machine learning models such as Random Forests, Support Vector Machines, and Gradient Boosting Machines have also been adopted for presence-only data modeling. These algorithms often function by identifying complex, non-linear relationships within the data, which can be particularly valuable for predicting species occurrences. They provide an avenue for species distribution modeling that is often interpretable, efficient, and robust to overfitting. By exploiting the power of ensemble methods or kernel-based strategies, these models can accommodate the intricate patterns and interactions seen in ecological datasets.

Bayesian models, especially in the context of presence-only data, have been highlighted due to their capability to encapsulate the complexities and uncertainties inherent in such datasets. Central to Bayesian modeling are techniques like Markov Chain Monte Carlo (MCMC) and the Integrated Nested Laplace Approximation with Stochastic Partial Differential Equation (INLA-SPDE). While MCMC offers a framework adept for data augmentation, simplifying the data distribution, INLA-SPDE presents a Bayesian approach that handles spatial and temporal autocorrelations. This method expands the predictive capacity of species distribution models, with the added advantage of utilizing Delaunay triangulation for more accurate predictions. The Bayesian foundation common to both allows for explicit quantification of uncertainties, crucial for management decisions (Divino, 2013; Lezama-Ochoa et al., 2020).

## **Objective of the Literature Review**

The objective of this review is to critically evaluate and synthesize the developments in presence-only prediction techniques. The exploration aims to shed light on the performance of traditional machine learning models against more specialized techniques, diving deep into the nuances of Bayesian methods, presence-only data processing mechanisms, and innovative sampling methodologies.

# **Sampling Techniques, Bias, and the Treatment of Pseudo-Absence Data**

## **Strategies to Address Bias**

### **Pseudo-Absence Data in Presence-Only Models**

## **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.

## **Machine Learning Models for Species Distribution Modeling**

### **Comparison with MaxEnt and Poisson Point Process models**

#### **Evaluation of Performance**

### **Bayesian Approaches in Species Distribution Modeling**

#### **Integrated Nested Laplace Approximation with Stochastic Partial Differential Equation (INLA-SPDE)**

The Integrated Nested Laplace Approximation with Stochastic Partial Differential Equation (INLA-SPDE) framework presents a Bayesian approach for handling the challenges of species distribution modeling (SDM). Bayesian models such as INLA-SPDE are adept at addressing complex datasets laden with spatial and temporal autocorrelations, thus providing an alternative to frequentist approaches, which yield fixed parameter estimates (Lezama-Ochoa et al., 2020; Martínez-Minaya et al., 2018; Blangiardo & Cameletti, 2015).

The INLA-SPDE method excels in capturing both well-known and more marginal areas where species are found, thereby expanding the predictive capacity of SDMs (Lezama-Ochoa et al., 2020). This framework incorporates multilevel structures with spatial random effects, which are stochastic processes indexed in space. This strategy adequately represents the various spatially explicit processes influencing species patterns (Lezama-Ochoa et al., 2020; Pennino et al., 2017; Redding et al., 2017).

One distinct advantage of the INLA-SPDE approach is its utilization of Delaunay triangulation over regular grids commonly used in SDMs. This triangulation congregates additional information in regions with higher density of observations, leading to more accurate predictions (Lezama-Ochoa et al., 2020).

Despite its capabilities, INLA-SPDE faces specific limitations such as difficulties in processing categorical variables and challenges in effective triangulation for analyzing spatial data (Lezama-Ochoa et al., 2020). Furthermore, although it offers a faster computational alternative to Markov Chain Monte Carlo (MCMC) methods, it should not replace MCMC entirely but rather serve as a complementary or alternative approach (Rue et al.; Lezama-Ochoa et al., 2020).

The Bayesian foundation of INLA-SPDE also allows for the explicit quantification of uncertainties, providing credible intervals and standard deviations in addition to point estimates (Lezama-Ochoa et al., 2020). This explicit quantification is particularly vital for management decisions, as it provides a fuller understanding of the model's predictions.

For a holistic understanding of species distribution, it is essential to contrast INLA-SPDE with other SDMs like Random Forests, MaxEnt, and Boosted Regression Trees. Each model

carries its unique set of strengths and limitations, warranting careful consideration for effective species management (Lezama-Ochoa et al., 2020).

## **Markov Chain Monte Carlo (MCMC)**

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

Presence-only data presents unique challenges in the realm of statistical modeling. These challenges primarily stem from the inherent uncertainty due to data censoring and the specific sampling procedures used in such studies. Divino (2013) aptly elucidates the Bayesian modeling approach to this problem, stating, “In this work we present a Bayesian approach to the problem of presence-only data based on a two levels scheme. A probability law and a case-control design are combined to handle the double source of uncertainty: one due to the censoring and one due to the sampling”.

As previously discussed, ecological surveys often record only species’ presence, leading to a unique dataset known as presence-only data (Araùjo & Williams, 2000). The situation where only attributes are noted without the corresponding response is not isolated to ecological studies, but has garnered attention in various realms of academic research. Lancaster and Imbens (1996) discuss this scenario, emphasizing its ubiquity in ecological contexts. Divino (2013) cites Pearce and Boyce (2006) to define such datasets: “Presence-only data [is] ‘consisting only of observations of the organism but with no reliable data where the species was not found’”. This dataset structure is the foundation for atlases, museum and herbarium records, species lists, incidental observation databases, and even radio-tracking studies.

There has been a surge in literature addressing modeling approaches for this type of data over the years. Some notable works include those by Keating and Cherry (2004), Pearce and Boyce (2006), Elith et al. (2006), Elith and Leathwick (2009), and Franklin (2010). Two primary approaches have been delineated for modeling presence-only data. The first approach considers Poisson point processes in both likelihood and Bayesian frameworks (Warton & Shepherd, 2010; Chakraborty et al., 2011). The second adopts a modified case-control logistic model, though, in works such as Ward et al. (2009) and Di Lorenzo et al. (2011), potential dependency structures within the observations are not taken into account. Another noteworthy modeling technique, MaxEnt, leans on the maximum entropy principle (Jaynes, 1957). As elucidated by Divino (2013), “In MaxEnt...the relative entropy between the distribution of covariates at locations where the species is present and the unconditional background distribution of covariates is maximized subject to some constraints concerning empirical statistics”. Dorazio (2012) notes a key limitation of the MaxEnt method: it requires knowledge of species’ prevalence for its estimator of occurrence to be consistent.

Divino (2013) then delves into the logistics of modeling presence-only data, articulating a two-tiered framework that encapsulates the linear logistic regression. The crux of this model

is to establish a relationship between observed stratum variable  $Z$  and covariates  $X$  in the midst of an absent binary response  $Y$ . This absence results in two layers of uncertainty, “the uncertainty due to the mechanism of censorship and the uncertainty due to the sampling procedure” (Divino, 2013). A crucial component in this Bayesian scheme is the case-control design which adjusts for the sampling mechanism.

Several pivotal equations are proposed by Divino (2013) in this context. The presence-only data approximation, defined as:

$$\phi_{\text{pod}}(x) \approx x\beta + \log\left(\frac{n_1u + np}{n_1u}\right)$$

modifies the standard linear regression function, tailoring it to suit presence-only datasets. Other equations that factor in the model include the approximation of the conditional probability of occurrence and the marginal probability  $P(Z|C = 1, x)$  after marginalization over  $Y$ .

Central to Bayesian modeling is the Markov Chain Monte Carlo (MCMC) computation. Divino (2013) underscores the data augmentation step as pivotal for easy implementation in the MCMC framework. This technique augments the set of observed data, making it follow a simpler distribution, thus facilitating the consistent derivation of the quantity  $n_1u$  crucial for adjusting the regression function to presence-only data.

The Bayesian modeling approach coupled with the MCMC computation, as presented by Divino (2013), provides an intricate yet potent method for analyzing presence-only data. Given the growing interest and reliance on such data, especially in ecological studies, this modeling technique offers a robust foundation for more reliable predictions and inferences in the future.

## **Discussion**

## **Synthesis of Findings**

## **Implications for Future Research**

## **Conclusion**

## **Summary of Major Findings**

## **Recommendations**

## **References**

Pearce and Boyce, 2006; Zarnetske, Edwards, and Moisen, 2007 Chefaoui and Lobo, 2008 Elith and Leathwick, 2007 Guisan et al., 2007 Zarnetske, Edwards, and Moisen, 2007 Phillips et al., 2009 Cressie, 1993 Baddeley et al., 2005 Elith et al., 2011; Thullier et al., 2008; O’Sullivan and Unwin, 2010;

Divino, F. (2013, May 6). Bayesian modeling and MCMC computation in linear logistic regression for presence-only data. arXiv.org. <https://arxiv.org/abs/1305.1232>

Lezama-Ochoa, N., Pennino, M. G., Hall, M., Lopez, J., & Murua, H. (2020). Using a Bayesian modelling approach (INLA-SPDE) to predict the occurrence of the Spinetail Devil Ray (Molecular mobular). *Scientific Reports*, 10(1). <https://doi.org/10.1038/s41598-020-73879-3>

Martínez-Minaya, J., Cameletti, M., Conesa, D. & Pennino, M. G. Species distribution modeling: a statistical review with focus in spatio-temporal issues. in *Stochastic Environmental Research and Risk Assessment* 1–18 (2018).

Pennino, M. G., Vilela, R., Bellido, J. M. & Mendoza, M. Comparing methodological approaches to model occurrence patterns of marine species. in *Research Advances in Marine Resources* (Eds: Norton, K.). (Nova Publisher, ISBN: 978-1-53612-177-3, 2017).

Redding, D. W., Lucas, T. C., Blackburn, T. M. & Jones, K. E. Evaluating Bayesian spatial methods for modelling species distributions with clumped and restricted occurrence data. *PLoS ONE* 12, e0187602 (2017).

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>

Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. (Stat. Method.)* 71, 319–392 (2009).



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-aos331>