University of Cape Town

Enhancing Species Distribution Models through Integrated Modeling and Bias Mitigation in African Bird Studies

Author:
Wayne Jiang

Student Number:
JNGWEN002

December 13, 2025

Abstract

The increasing availability of biodiversity data provides significant opportunities to improve species distribution modeling, particularly through the integration of multiple datasets. The overarching aim of this dissertation is to construct an integrated species distribution model (ISDM) for African bird species. A central challenge in developing ISDMs is that different datasets follow distinct sampling protocols and embody different assumptions. In particular, widely available presence-only (PO) data are prone to severe sampling bias, which can substantially distort model inference if not properly addressed.

In this dissertation, we evaluate how sample size, degree of spatial bias, and species prevalence influence the accuracy and stability of ISDMs. This is achieved through simulation experiments using the virtual ecologist approach, which allows controlled manipulation of ecological and sampling processes. We also examine methods for mitigating sampling bias in PO datasets, including modeling the bias using covariates and incorporating an additional spatial random field specifically designed to account for the bias component.

Our simulation results show that when the volume of presence-only data greatly exceeds that of presence-absence (PA) data, the PO dataset dominates model behaviour, resulting in decreased precision and reduced predictive performance. Consequently, when applying ISDMs to real-world data, PO data must be thinned to reduce the influence of sampling bias. Guided by the insights gained from the simulation study, ISDMs were then constructed for three African bird species with differing ecological and data-related characteristics. These models were developed using eBird (PO) data and SABAP2 (PA) data. As informed by the simulations, the eBird dataset was thinned prior to model construction, with thinning intensity determined using the inhomogeneous pair correlation function to minimise residual sampling bias.

Contents

1 Introduction **4**
1.1 Research Aim . . . 4
1.2 Background . . . 4
1.3 Methodology . . . 5
1.4 Outline of the Thesis . . . 5

2 Literature review **7**
2.1 Why use data integration methods to estimate species' distributions . . . 7
2.2 Characteristics of abundance data types . . . 8
2.2.1 Presence-absence data . . . 8
2.2.2 Count data . . . 8
2.2.3 Presence-only data . . . 8
2.3 General approaches for integrating data to model species distributions . . . 9
2.3.1 Data pooling . . . 9
2.3.2 Ensemble modeling . . . 9
2.3.3 Using one dataset as covariate data (Auxiliary data) . . . 10
2.3.4 Informed priors . . . 10
2.3.5 Correlation model . . . 11
2.3.6 Integrated modeling via joint likelihood . . . 11
2.3.7 Evaluation and comparison of approaches . . . 12
2.4 Handling bias in presence-only data . . . 12
2.4.1 Down-weighting presence-only data . . . 13
2.4.2 Accounting for bias via intensity function . . . 13
2.4.3 Penalizing the joint likelihood . . . 14
2.4.4 Spatial thinning . . . 14
2.5 Validation and Model evaluation . . . 16
2.6 Related work . . . 17

3 Data **20**
3.1 Real world data . . . 20
3.2 Virtual data via Virtual Ecologist approach . . . 21
3.3 Predictor variables . . . 22
3.3.1 Bioclimatic variables . . . 23
3.3.2 World Population Density . . . 23
3.3.3 Topographical and Landcover variables . . . 24

4 Methods **26**
4.1 Simulating Virtual Species . . . 26
4.2 Point process models . . . 28
4.3 Spatial thinning . . . 33
4.4 Performance measures . . . 34
4.4.1 Area under the ROC curve (AUC) . . . 34
4.4.2 Mean absolute error (MAE) . . . 35
4.4.3 Pearson correlation . . . 35
4.4.4 Tjur R^2 . . . 35
4.4.5 Watanabe-Akaike Information Criterion (WAIC) . . . 36
4.4.6 The accuracy and precision of the estimates . . . 36

5 Simulation experiment **37**
5.1 Simulation of virtual species . . . 37
5.2 Sampling . . . 38
5.3 Integrated distribution models . . . 40
5.4 Data Simulations . . . 41
5.5 Results and discussion . . . 41
5.5.1 Mean absolute error (MAE) . . . 41
5.5.2 Pearson correlation and Area under the curve (AUC) . . . 42
5.5.3 Tjur R^2 . . . 44
5.5.4 Discussion . . . 44

6 Real world application and result **46**
6.1 Martial Eagle . . . 46
6.2 Long-crested Eagle . . . 50
6.3 Peregrine Falcon . . . 53

7 Conclusion **57**
7.1 Limitations and future research opportunities . . . 57

8 References **59**

9 Appendix **63**
9.1 R code: Simulation . . . 63
9.2 R code: Real world application . . . 66

1 Introduction

1.1 Research Aim

The general aim of this study is to construct an integrated species distribution model for African birds. In the process, we aim to determine what the best modeling choices are in the construction of integrated species distribution models when using presence-only (PO) and presence-absence (PA) data.

1.2 Background

Species distribution modeling is important for many problems in ecology and conservation. Nowadays, there has been an increasing availability of data sources for modeling species distributions. Reliably combining data sources has great potential, but it can be challenging due to variations in their design, the environmental gradients they capture, and potential sampling biases. Species distribution models describe the relationship between species presence and environmental variables, helping to estimate the geographic areas where a species is likely to occur. It is also useful in determining the environmental conditions most suitable for a species. Gathering structured data to build accurate large-scale species distribution models can be expensive and time-consuming (Isaac et al., 2020). As technology advances, the availability of information for species distributions is radically growing. Therefore building a species distribution model using multiple data sources becomes a great opportunity, but researchers face the challenge of using different sources that have their own sampling protocols and assumptions. In the past, researchers integrated multiple data sources through the point process model (Miller et al., 2019). However, the simple integrated model does not always outperform the model that uses only structured data (Simmonds et al., 2020). The reason is unstructured datasets like presence-only data have the advantage of availability but they suffer from different kinds of bias. Simply combining structured datasets and unstructured datasets can lead to a biased result. Therefore, an integrated species distribution model that only keeps the advantage of each dataset can be an important need.

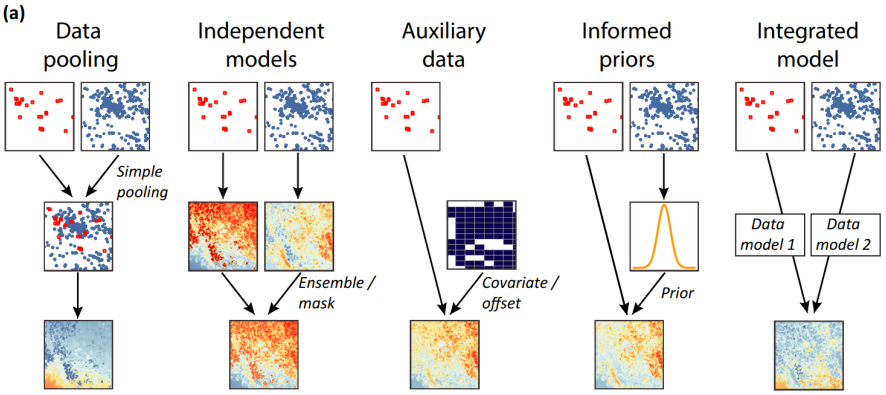


Figure 1: An illustration of different approach, directly taken from Fletcher et al. (2019)

Fletcher et al. (2019) have provided a comprehensive insight into the methodologies and considerations involved in integrating diverse data sources to model species distributions. They reviewed

recent developments in combining multiple datasets for species distribution modeling. They identified five methods by which multiple sources of data are typically combined to model the distributions of species: Data pooling, independent models, Auxiliary data, Informed priors, and Integrated models. An illustration of each approach is shown in Figure 1.

The capacity of these methods to account for bias, uncertainty, and sampling design when quantifying environmental relationships in species distribution models varies (Fletcher et al., 2019). Figure 1 shows the basic manner in which the approaches are implemented. Fletcher et al. (2019) believed that integrated species distribution models can solve the challenge of combining data because they account for most available information, allow the inclusion of different predictors, and operate within a simultaneous modeling framework.

Building on the need for integrated species distribution models, Sherman's fox squirrels distribution in USA (Fletcher et al., 2015) and fishery data (Paradinas et al., 2023) have been studied using integrated species distribution modeling to effectively combine data from multiple sources while addressing biases inherent in unstructured datasets. This approach has demonstrated its potential in improving species distribution and population dynamics analysis. However, integrated species modeling has yet to be applied to African bird species, presenting an exciting opportunity to explore their ecological patterns and inform conservation efforts with this robust and flexible framework.

1.3 Methodology

In this project, we aim to combine one structured presence-absence (PA) dataset and one unstructured presence-only (PO) dataset to leverage the accuracy of PA data while minimizing bias from the PO data to construct species distribution models for some bird species. To achieve the aim of the project, we applied a virtual ecologist approach that separated this project into three parts. Firstly, we simulated two virtual species using various environmental covariates and two levels of species prevalence and mimicked a sampling regime that would occur in the real world to generate the two types of data. This was all done via the R package Virtualspecies (Leroy et al., 2015). Secondly, we fitted and evaluated different integrated models on the generated virtual data via the R package PointedSDMs (Mostert et al., 2023), and performance was compared using metrics such as AUC, TSS, Tjur R^2 and WAIC. Finally, insights from the simulations were applied to real data, fitting integrated models to eBird and SABAP2 datasets for three bird species: Martial Eagle (*Polemaetus bellicosus*), Long-crested Eagle (*Lophaetus occipitalis*), and Peregrine Falcon (*Falco peregrinus*), chosen for their differing prevalence, habitat preference, and detectability.

1.4 Outline of the Thesis

This Dissertation is organised as follows.

section 2 reviews existing approaches for integrating presence–absence and presence-only data, and summarises common sources of sampling bias and mitigation strategies.

section 3 describes the datasets used in this study, including both virtual species data and real-world observations from SABAP2 and eBird, together with the environmental covariates.

section 4 presents the methodological framework, covering the generation of virtual species, the construction of integrated species distribution models, the spatial thinning procedure, and the performance metrics used for evaluation.

section 5 reports results from simulation experiments designed to examine how sample size, spatial

bias and prevalence influence model performance, and to assess the effectiveness of different bias-mitigation approaches.

section 6 applies the recommended workflow to three African bird species to demonstrate real-world performance of the proposed methods.

Finally, section 7 provides concluding remarks and discusses limitations and potential directions for future research.

2 Literature review

2.1 Why use data integration methods to estimate species' distributions

Data integration, the process of combining data, is known by various terms, such as data fusion, assimilation, combination, or integration (Isaac et al., 2020). With recent technological advancements, the increased availability of ecological data presents both opportunities and challenges. Ideally, utilizing data from multiple projects could lead to more reliable and comprehensive insights into species distributions and ecological patterns, as it allows researchers to leverage the complementary strengths of different datasets. However, maximizing the benefits of the data revolution is complex because datasets are often designed for specific objectives, meaning that different data types have distinct strengths and limitations, which must be carefully accounted for during integration (Isaac et al., 2020). Research is conducted by different countries and institutions, resulting in variations in data collection protocols, data quality, and data types. Common ecological data types include counts, presence-absence data, capture-recapture data, spatial capture-recapture, mark-recapture data, and presence-only data. These data types can provide valuable insights into species abundance, occurrence, and information about their dynamics over time and space.

Traditional modeling methods are typically designed to suit a specific data type; for instance, widely used MaxEnt method for species distribution modeling is optimized for presence-only data (Isaac et al., 2020). So usually when modelers have multiple datasets available to them, they are forced to either pool them and ignore any differences in how the datasets were collected or to choose between them (Isaac et al., 2020). A common choice is between small amounts of structured data and large amounts of unstructured data. Structured data come from surveys with formal protocols, which are expensive to gather and often geographically limited. In contrast, unstructured data make up the majority of available information but are influenced by various biases. Examples of structured data and unstructured data are presence-absence data and presence-only data respectively. This decision can be challenging, and modelers may lose sight of the advantages of having multiple datasets available to them.

Also, another problem raised for species distribution studies is that single data might not be a well-selected sample for the area that the researcher is interested in. For example, one single dataset may not cover the whole distribution of a species. Species distribution of models built on that single dataset have limited predictive power in the area of interest and fail to describe the species and environment relationship well. When data is hampered by small sample sizes, low detectability, or limited spatial coverage, researchers enhance predictive performance and reduce uncertainty by employing integrated models that incorporate multiple datasets (Nephin et al., 2023).

Consequently, ecologists are faced with the crucial question of how to optimize the combination of datasets to enhance the accuracy of species distribution estimates. This allows the advancement of data integration techniques that simultaneously leverage information from multiple datasets while addressing the unique strengths and deficiencies of each dataset (Miller et al., 2019). More recently, efforts have concentrated on the integration of data from specifically designed surveys and presence-only data. Most applications have used the point-process framework, a flexible and likelihood-based approach. The more recent advancements that implement data integration for multiple data types are supported by this flexible likelihood-based framework (Miller et al., 2019). By leveraging this methodology, researchers can enhance the precision and robustness of species distribution models, even in data-limited or unevenly sampled regions. In this research, we focus on building SDMs using presence-absence data, and presence-only data. We note, however, using the point-process framework, it is possible to combine data of various other types (Miller et al., 2019).

2.2 Characteristics of abundance data types

2.2.1 Presence-absence data

Presence-absence data are records of whether a species is present or not at each sampling location (Isaac et al., 2020). Presence-absence data are typically collected from pre-selected sites in carefully designed surveys; therefore, the data are derived from a well-defined sampling protocol, ensuring that observations are comparable across time and space. Since presence-absence data follow formal protocols or sampling design, it requires data collectors to be experienced and/or well-trained. Hence, it is expensive to collect, and usually, fewer samples are generated as compared to ad-hoc surveys. For each site where the presence-absence data were sampled, presence is modeled as a single Bernoulli trial with a probability of presence p (Simmonds et al., 2020).

Presence-absence data may be biased due to imperfect detection, but it can be avoided by repeated surveys of sites within a relatively short timeframe (MacKenzie, 2005). Commonly, presence-absence data are assumed to be free from spatial bias, but this is not guaranteed (Isaac et al., 2020). Presence-absence data are often called structured data, therefore they are known to be of good quality. A good example of Presence-absence data in Africa is the Southern African Bird Atlas Project 2 (SABAP2). It is a citizen science project but it is reliable as only actual scientists with advanced training can contribute and it has a well-outlined survey protocol.

2.2.2 Count data

Similar to presence-absence data, count data is also collected in pre-selected study sites. Count data are the total number of individuals from a given species within each study site. Unlike presence-absence data, which only records whether the researcher found species or not, count data requires the researcher to count the number of observations within study sites. These data may be presented in the form of direct counts or an index of abundance that is derived from the original counts (Isaac et al., 2020). Similar to presence-absence data, Count data requires a formal protocol or sampling design, and it usually takes a substantial amount of time to collect. It is also costly to collect, and hence records are generally few. For each site where the count data are collected, the number of observations is modeled as a single Binomial trial with a probability of observing the species p and the total number of observations n.

Similar to presence-absence data, count data is also called structured data, and it has good quality. An example of count data is the International Waterbird Census (IWC), a monitoring program that operates in 143 countries to collect data on the number of waterbirds at wetland sites (IWC, 2024).

2.2.3 Presence-only data

Presence-only data only records the locations where a species was observed. These data do not have information about where a species was not observed, in contrast to presence-absence data (Isaac et al., 2020). Presence-only data are often called unstructured data. This is because they are usually collected without formal protocol or sampling design causing bias in the data. People without experience can easily participate in presence-only data collection. Therefore, these data are cheap to collect and massive datasets can be generated. It is commonly assumed that presence-only data locations arise from a Poisson point process. Under this assumption the total number of presences in a sub-region A are Poisson distributed (Simmonds et al., 2020).

Presence-only data suffer from many types of biases, sampling bias is one of them as people may not collect samples equally over the study region or area of interest. eBird is one of the most

famous presence-only data in the world (Sullivan et al., 2009). It is a citizen science dataset of bird observations collected by scientists, researchers, and amateurs.

2.3 General approaches for integrating data to model species distributions

Each data type has its own strengths and limitations, and the purpose of combining multiple datasets is to maximise the benefits of each while reducing their individual weaknesses. General approaches for integrating data in species distribution modeling are Data pooling, Ensemble modeling, Auxiliary data, Informed priors, Correlation model, and Integrated modeling via joint likelihood. Fletcher et al. (2019) analysed the literature to determine how frequently each integration approach has been used. Their review showed that data pooling is by far the most commonly applied method, followed by ensemble modeling, integrated modeling, and the auxiliary-data approach. The informed-prior approach was used least often, appearing in only about 1% of the studies they examined. These patterns indicate the prevailing preferences of researchers when combining multiple datasets. Although data pooling has historically been the dominant approach, integrated modeling has gained increasing popularity in recent years.

2.3.1 Data pooling

Data pooling is the combination of different data sources to model species distribution without explicit acknowledgment of the different data characteristics in the modeling process (Fletcher et al., 2019). It is commonly used for opportunistic presence-only data that are collected from different sources. Data pooling ignores differences between each data source such that simple pooling may be of limited value (Fletcher et al., 2019). However, it is easy to use and can be helpful to increase the number of point occurrences used for modeling. This might be the reason for the high frequency of data pooling literature. Pooling data is generally straightforward when the data are of the same type. However, when combining different data types, the appropriate statistical distribution is often unclear, which can lead to unwarranted inferences and conclusions (Fletcher et al., 2019). Take for instance the indiscriminate pooling of presence-only data and count data. We can either choose the Poisson distribution or the Inhomogeneous Poisson Process (IPP) which requires either count data to be assigned to a spatial location or presence-only data to be assigned to grid cells. This extra process generates location errors and misleading inferences. Data pooling may be of some use in instances where the pooled data are of a similar type. For instance, in Palaoro et al. (2013) 6 presence-only datasets are pooled together to build species distribution models of an invasive species. The similarity of the data types allows them to be pooled together although each dataset was obtained via a different protocol.

2.3.2 Ensemble modeling

Ensemble modeling usually builds independent models for each data source first and then ensembles them. This two-step process can be useful for understanding how different data sources can vary in terms of statistical inference and predictions (Fletcher et al., 2019). However, this approach does not allow each model to share information to enhance the estimate's accuracy. Also, it is difficult to combine parameter estimates formally, and usually, the coefficients are not directly comparable between models (Fletcher et al., 2019). For example, count data and presence-only data may be modeled using Poisson regression and IPP respectively. Both Poisson regression and the IPP model can be used to estimate coefficients, but since count data and presence-only data have different spatial support (grid cells and locations) the coefficients are not directly comparable.

To address these limitations, ensemble modeling provides an alternative strategy that focuses on combining predictions rather than parameters. By integrating outputs from multiple SDMs—often based on different algorithms—ensemble approaches reduce model-specific biases and yield more stable and robust predictions. Common ensemble procedures include simple averaging, weighted averaging based on model performance, and consensus voting.

While the discussion above highlights ensemble modeling in the context of integrating different data sources, ensemble techniques are also widely applied within a single data type to improve predictive accuracy. Grenouillet et al. (2011) illustrated the effectiveness of this approach using presence–absence data for 35 common species in France. They fitted eight different SDMs, each trained with 70% of the data and evaluated with the remaining 30%, and repeated this split-sampling procedure 100 times. The 800 resulting predictions were converted into binary maps using a threshold that maximized sensitivity and specificity. The authors then averaged the 100 predictions from each modeling technique and again applied a binary threshold to obtain a single prediction per SDM. Finally, the eight SDMs were combined into an ensemble average model. Across all species, the ensemble model consistently outperformed the individual SDMs, demonstrating the benefits of integrating multiple modeling approaches into a single consensus prediction.

2.3.3 Using one dataset as covariate data (Auxiliary data)

Covariate models are defined as the inclusion of information from one dataset, typically the lower quality dataset, in the model of a second dataset through a fixed effect covariate, enabling the sharing of information between datasets (Ahmad Suhaimi et al., 2021). This relationship provides a straightforward means to incorporate auxiliary data. However, similar to combining independent models, it is often unclear how one can account for variation in the spatial or temporal support of the data when modeling the primary data source (Fletcher et al., 2019). When using this approach, decisions have to be made on which dataset to include as a covariate when similar quality datasets are available and the spatial scale to be aggregated to produce a suitable covariate. Trainor and Schmitz (2014) applied the auxiliary data approach with presence-only data of the Canada lynx and its primary prey, the snowshoe hare. They connected the two species by providing the means to spatially characterize variation in consumer trophic dependencies on resources, a method they termed the trophic interaction distribution model (TIDM). The TIDM was developed using data from a multiyear telemetry and ground-tracking study that monitored lynx movement and lynx–snowshoe hare encounters following the lynx reintroduction to Colorado, USA. It is designed to account for the spatial locations of observed lynx–snowshoe hare interactions. Finally, the species distribution model is improved by incorporating the lynx TIDM as a covariate in the lynx SDM.

2.3.4 Informed priors

Combining data via informed priors is a different approach whereby the first dataset influences the second via informative priors rather than a fixed effect (Ahmad Suhaimi et al., 2021). Although informed priors in ecological models were first suggested by Ellison (1996), it has been rarely implemented in the context of species distribution modeling when data is combined. Informed priors are applied sequentially, with one data source providing a prior distribution for one or more parameters when modeling the second data source. The requirement to use informed priors is that the model for the first set of data has the same parameters as the model for the second source of data (Fletcher et al., 2019). For instance, a prior that is derived from Poisson regression on count data should not be employed as a prior in IPP regression with presence-only data.

2.3.5 Correlation model

Ahmad Suhaimi et al. (2021) suggest a correlation model approach. In correlation modeling, a shared covariance matrix indirectly connects the datasets, capturing common patterns in both data sources. Ahmad's findings indicate that the correlation model works well even when there were unknown biases and when high-quality PA data were spatially restricted. Ahmad Suhaimi et al. (2021) suggested that correlation models offer a reliable alternative to joint likelihood models if covariates associated with effort or detection in presence-only data are unavailable. The correlation model is recently proposed by Ahmad Suhaimi et al. (2021) and has not yet been tested in the literature.

2.3.6 Integrated modeling via joint likelihood

A joint likelihood is the most intuitive approach for dealing with two datasets that we wish to integrate into a singular model. This likelihood is obtained by multiplying the likelihoods of each dataset (Ahmad Suhaimi et al., 2021). Because the datasets we are dealing with are not the same data types, point processes are used to connect them formally as all data types can be assumed to be generated by a latent point process. This is discussed in detail in the subsection 4.2. This approach has several desirable characteristics and has been shown to improve estimates of environmental relationships and predictions of species distributions (Fletcher et al., 2019). A joint likelihood for the integrated model is the product of the component likelihoods as in Equation 1.

$$L(\boldsymbol{\beta}; \boldsymbol{\theta}) = \prod_{m=1}^{M} L_m(\boldsymbol{\beta}; \boldsymbol{\theta}_m) \tag{1}$$

Where $\boldsymbol{\beta}$ are shared parameters and $\boldsymbol{\theta}_m$ are parameters associated with the observation process of the m^{th} dataset and are typically not shared. The joint likelihood method was found to be more sensitive to the quality of the unstructured data source than covariate and correlation modeling, although it performed relatively well overall. Joint likelihood models have been widely recommended in the literature—such as Ahmad Suhaimi et al. (2021), Fletcher et al. (2015)—as the best-performing models when spatial bias in PO data is low or bias can be modeled. However, it gave poor estimates when there were unknown biases in the data (Ahmad Suhaimi et al., 2021). Therefore, when using Integrated modeling, bias handling methods seem to be important.

Fletcher et al. (2015) applied an integrated modeling framework to assess the distribution of Sherman's fox squirrel, a species primarily found in central and northern Florida with extensions into southern Georgia. They compared the integrated model with two alternatives: a generalized linear model (GLM) fitted solely to presence–absence data, and an ensemble model that averaged predictions from separate presence–absence and presence-only models. Prior to analysis, the authors verified all suspicious occurrence records and removed any that could not be confirmed, and they conducted intensive camera-trap surveys across 40 landscapes to obtain high-quality presence–absence data. In total, 2,785 presence-only records and 252 presence–absence observations were incorporated into the models.

A key strength of the integrated model was its ability to explicitly account for sampling bias by including a covariate representing distance to roads. This allowed the model to address spatial bias in the presence-only data more effectively than the GLM or the ensemble approach. As a result, the integrated model produced coefficient estimates that were smaller in magnitude but more precise, reflecting reduced uncertainty and improved parameter identifiability. In terms of predictive performance, the integrated model consistently outperformed the single-source and ensemble

models, as demonstrated through block validation—a spatially structured cross-validation method that assesses a model's ability to generalize to unsampled areas. This suggests that integrating multiple data types within a joint likelihood framework not only mitigates bias but also enhances spatial prediction accuracy relative to models that treat each dataset independently.

2.3.7 Evaluation and comparison of approaches

Table 1 which is adopted from Fletcher et al. (2019) shows the different approaches for combining data and their characteristics. It can be seen that informed priors and correlation models account for all variability but they perform sequentially. Out of the 6 approaches, only simple pooling and integrated models are simultaneously performed. In practice, simultaneous models are easier to use than sequential models therefore they are preferred by researchers. Integrated models appear to have the most advantageous characteristics among all approaches when combining multiple datasets. Also, Fletcher et al. (2019) believe integrated species distribution models are the best approach to tackling the challenge of combining data.

Characteristic	Simple pooling	Ensemble modeling	Auxiliary data	Informed priors	Correlation model	Integrated models
Account for different sampling issues	No	Yes	Yes	Yes	Yes	Yes
Account for variation in spatial or temporal support among data	No	Yes	No	Yes	Yes	Yes
Account for uncertainty from both data sources	No	No	No	Yes	Yes	Yes
Allow for different predictors for each data source	No	Yes	Yes	Yes	Yes	Yes
Sequential vs. simultaneous modeling of data sources	Simultaneous	Sequential	Sequential	Sequential	Sequential	Simultaneous

Table 1: Some characteristics of different approaches for combining data, taken from Fletcher et al. (2019)

2.4 Handling bias in presence-only data

Bias seems to be the most important problem when combining with other data types. Studies show that model performance can vary substantially depending on whether or not the model accounts for bias. In particular, joint likelihood approaches are highly sensitive to bias in the data. These models tend to perform well when spatial bias in PO data is minimal, but their estimates become unreliable when biases are unknown or unaccounted for (Ahmad Suhaimi et al., 2021). Also, when combining PO data with PA data or count data, the amount of PO data usually is much more than others. This can result in the larger dataset having a dominant influence on the outcome, as the integrated joint log-likelihood function is additive, causing the larger dataset to contribute disproportionately to the joint likelihood (Fletcher et al., 2019). The bias can therefore dominate the integrated model and affect the estimates hugely.

There are 3 prescribed strategies for handling sampling bias when we use the joint log-likelihood model:

1. Down-weighting the presence-only data
2. Accounting for sampling bias via the intensity function
3. Penalizing the joint log-likelihood

2.4.1 Down-weighting presence-only data

Down-weighting the presence-only data and accounting for bias via the intensity function is often used when handling bias. For example in Equation 2, the log-likelihood of the presence-only data has less weight and higher weight on the log-likelihood of the presence-absence data.

$$\log(L(\boldsymbol{\beta};\boldsymbol{\theta})) = w * \log(L_{PA}(\boldsymbol{\beta};\boldsymbol{\theta}_{PA})) + (1-w) * \log(L_{PO}(\boldsymbol{\beta};\boldsymbol{\theta}_{PO})) \tag{2}$$

In Fletcher et al. (2019) down-weighting the presence-only data was found to improve model performance for all species tested. Intuitively, the more weight is given to structured data the more improvement in model accuracy, but overly weighting on PA data leads to a model that is similar to PA only model and lose the benefit of combining multiple datasets. As a result, a decision must be made regarding which dataset should be assigned the highest weight and the exact value of that weight. It remains unclear what the optimal weighting strategy should be, and developing objective criteria for weighting datasets in IDMs remains a key research priority and an open question (Simmonds et al., 2020).

2.4.2 Accounting for bias via intensity function

Accounting for bias via the intensity function is also a commonly used method. This is achieved by assuming that the locations of individuals follow a log-Gaussian Cox process, which represents a doubly stochastic Poisson process, as the intensity itself is stochastic (Simmonds et al., 2020). The intensity function for this process is defined by Equation 3:

$$\log(\lambda(s)) = \alpha_0 + \beta_x x(s) + \xi(s) + \varepsilon(s) \tag{3}$$

where the log of the intensity $\lambda(s)$ is defined by an intercept α_0, a linear relationship β_x with an environmental covariate $x(s)$, a Gaussian random field $\xi(s)$ which simulates spatial variation not explained by the environmental covariate, and some random error $\varepsilon(s)$ (Simmonds et al., 2020).

Based on this intensity function, the bias can be accounted for in two ways. Firstly, the bias can be accounted by adding a covariate $z(s)$ as shown in Equation 4. Adding a covariate $z(s)$ improves joint likelihood model performance well if the bias covariates can well explain the bias contained in the data (Ahmad Suhaimi et al., 2021). However, it requires additional information from experts and this strategy can be useless when the spatial bias in the data cannot be explained properly.

$$\log(\lambda_{PO}(s)) = \alpha_{PO} + \beta_x x(s) + \beta_z z(s) + \xi(s) + \varepsilon(s) \tag{4}$$

Another way to account for bias is via an additional spatial field $\zeta(s)$ as shown in Equation 5, which also improved joint likelihood model performance (Simmonds et al., 2020). The main difference between the approach given in Equation 4 and that of Equation 5 is that in the latter no knowledge is on the nature of the bias is needed.

$$\log(\lambda_{PO}(s)) = \alpha_{PO} + \beta_x x(s) + \xi(s) + \zeta(s) + \varepsilon(s) \tag{5}$$

2.4.3 Penalizing the joint likelihood

Penalizing the joint log-likelihood is another possible solution to handling bias. However, it is not widely used and no bench-mark studies for its performance have been reported. Therefore, researchers have to determine the form of penalty. Hutchinson et al. (2015) used penalized likelihood methods to improve parameter estimates in occupancy models. They compared the model by maximum likelihood estimation to the Ridge penalty, the Bayes penalty, and the logistic regression penalty. An objective function of the ridge penalty is defined in Equation 6.

$$\log(L(\theta)) - k\frac{1}{2}(\alpha_1^2 + \beta_1^2) \tag{6}$$

where k is the tuning parameter, α_1^2 and β_1^2 are two non-intercept coefficients of the model and the $\frac{1}{2}$ is introduced to simplify derivatives. The ridge penalty shrinks estimates toward zero by penalizing large coefficients. Hutchinson et al. (2015) applied this to occupancy models by adding a penalty to the non-intercept coefficients, encouraging a model with constant occupancy and detection probabilities.

The bias, variance, and mean squared error (MSE) of parameter estimates obtained from each method on synthetic data have been examined. The penalized approach exhibited a lower MSE compared to the standard approach across all of the synthetic datasets (Hutchinson et al., 2015). Penalized likelihood methods showed improved performance across all of the tests conducted in their study. Bayes penalty performed well for small sample size and Ridge penalty performed well when sample size is large. But both methods require tuning the value of k via cross-validation. On the other hand, the logistic regression penalty's performance varies in cases. Penalizing the joint log-likelihood shows potential, given that penalty terms have been a longstanding practice in statistics. However, a benchmark study would be necessary.

2.4.4 Spatial thinning

Another way to handle bias is to perform spatial thinning before modeling. Although this method was applied in single data species distribution models to the best of our knowledge it has not been used in SDMs combining multiple datasets. Spatial thinning of species occurrence records can address issues related to geographical sampling biases. This procedure keeps the most valuable information while deleting the fewest records required to significantly lessen the effects of sampling bias (Aiello-Lammens et al., 2015).

Current spatial thinning methods generally fall into one of two categories, either employing stratified random sampling or thinning based on nearest neighbor distance (Aiello-Lammens et al., 2015). One method in the first category entails overlaying a grid on the study region and randomly sampling a set number of occurrence records from each grid cell, where grid cells should have equal area (Aiello-Lammens et al., 2015). Other methods involve stratifying based on the density of occurrence records and randomly selecting records for inclusion based on the geographic sampling density. The second category involves removing occurrence records so that no two are closer than a linear distance x, resulting in a minimum nearest neighbor distance (NND) greater than or equal to x (Aiello-Lammens et al., 2015). To retain the greatest amount of useful information, records should be thinned such that the largest possible number of records is retained. This method presents several challenges. Like all thinning approaches, the optimal degree of thinning remains subject to empirical determination. Specifically, the optimal NND (x) likely varies across species and study regions. Additionally, this category presents serious computational challenges. Determining the optimal number of occurrence records that meet the NND constraint can

be viewed as the classic set-packing problem in computational complexity theory (Aiello-Lammens et al., 2015), which is considered nondeterministic polynomial-time (NP) hard. While solutions to such problems can be checked quickly, it remains unclear whether a solution can be found quickly (Aiello-Lammens et al., 2015).

Aiello-Lammens et al. (2015) applied spatial thinning to a set of occurrence records for the Caribbean spiny pocket mouse, *Heteromys anomalus*. The spatial thinning algorithm reduces spatial sampling bias in ecological datasets by enforcing a user-defined minimum nearest neighbor distance (NND) between occurrence records. The authors used a thinning distance of 10 km to the dataset. The steps of the algorithm are as follows:

1. Set Minimum Distance: User specifies the minimum NND (x).
2. Calculate Distances: Pairwise distances between all records are computed.
3. Identify Violations: Records with neighbors closer than x are flagged.
4. Remove Records: Records with the most violations are randomly removed.
5. Repeat: Steps 3–4 continue until no violations remain.
6. Output: One or more thinned datasets are generated, retaining as many records as possible.

The occurrences are located along the coastal mainland (hereafter, mainland) of northern South America (174) and three nearby Caribbean islands: Trinidad (21), Tobago (4), and Margarita (2). Aiello-Lammens et al. (2015) ran the thinning algorithm 10 and 100 times. Most repetitions resulted in an optimal number of 110 occurrence records for the mainland dataset and 124 occurrence records for all data. Figure 2 illustrates how the spatial thinning algorithm works. There are two squares in A with dense occurrence points, which is a pattern commonly observed in regions where biased sampling is expected. Purple records were removed by the thinning algorithm and red records were retained by the function on B and C. Aiello-Lammens et al. (2015) constructed Ecological Niche Models (ENMs) using an unthinned dataset, a manually thinned dataset, and two spatially thinned datasets generated with 10 and 100 repetitions of the thinning algorithm. The ENMs from the algorithm-thinned datasets differed significantly from those of the unthinned dataset but closely matched the manually thinned results, indicating the algorithm effectively reduces spatial bias.

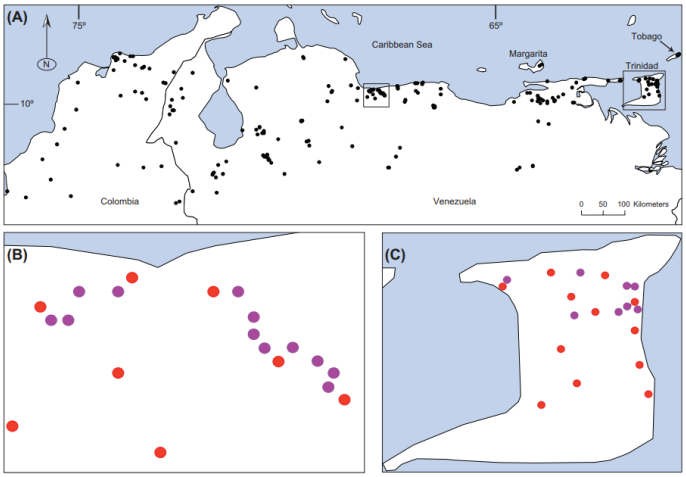


Figure 2: An illustration of spatial thinning algorithm, directly taken from Aiello-Lammens et al. (2015). Purple points represent locations removed during thinning, while red points represent locations retained, as shown in B and C.

2.5 Validation and Model evaluation

Researchers in the past compared data combining approaches using mostly real-world data (e.g., Aiello-Lammens et al. (2015); Ahmad Suhaimi et al. (2021)). Fletcher et al. (2019) compared the integrated model to independent models using planned survey data and eBird data on 24 species of birds. The utility of the models was evaluated with 3-fold block validation. They built models using data from two geographic regions and then projected the predictions onto a third region that was not included in model building. The predictive accuracy was measured using the area under the curve (AUC) of the receiver operating characteristic (ROC) curve and block-validated log-likelihoods. AUC is a widely used metric for assessing model discrimination, while block-validated log-likelihoods provide a more accurate measure of model calibration (Fletcher et al., 2019).

Nephin et al. (2023) tested whether the integration of different datasets could improve distribution predictions across an area not properly sampled by a single dataset using Dungeness crab data from British Columbia as a case study. They evaluated model performance with spatially buffered leave-one-out cross-validation to prove that integrated modeling is better than the independent model. The model's performance was evaluated using Tjur's R^2 and AUC, which favor models with good discrimination abilities that predict a very low probability of detection in the absence of a species and a very high probability of detection in the presence of a species. Tjur R^2 is in contrast to AUC, which only considers the rank order of probabilities and not their magnitude (Nephin et al., 2023).

On the other hand, some literature used simulated studies to compare model performance. Simmonds et al. (2020) first generated true background data, and then obtained samples of both PA and PO data. They built models based on the two datasets. They did not estimate observation probability from the model. Instead, they estimated a relative pattern of occurrence in relation to the mean prediction, as they assumed that using PA and PO data for a single species without additional information could not provide an accurate estimate of the intercept of the original true intensity surface (Simmonds et al., 2020). The study evaluated the models using three measures. First, to evaluate the quality of the model parameter estimation, the accuracy and precision of

the estimate $\beta_x \hat{x}(s)$ were evaluated. Second, the correlation between the true intensity and the expected intensity was calculated. This measure evaluated each model's ability to represent the geographic pattern in species distributions. Finally, the unsigned difference between the true and forecasted intensities was used to compute the mean absolute error (MAE). In the study by Simmonds et al. (2020), the MAE does not assess the models' ability to return absolute intensity values, as all validation was performed relative to the mean prediction. Instead, the MAE shows how well the models can represent the intensity variation.

In a study by Ahmad Suhaimi et al. (2021), the authors performed a simulated study to compare model performance. Similarly, Simmonds et al. (2020) also used mean absolute error (MAE) to compare the goodness of fit of each model by measuring the difference between the predicted and true log intensities. Additionally, the Pearson correlation between predicted and actual log intensities was calculated to assess the similarity in the predicted spatial distributions. A correlation coefficient closer to one indicates a better spatial match between the predicted and actual values. Furthermore, the authors assessed the bias in the estimate of the environmental covariate coefficient, as the true parameter value was known.

When we want to know which model has the best model performance, it is important to compare it to independent models or other multiple-data model approaches. It appears logical to validate the model using the data source with better quality to compare methods. When the quality of all data sources is unknown then model selection is not clear (Miller et al., 2019). Miller suggests that multi-objective optimization may provide a path forward that combines measures of fit across the different sources of data using a weighted average. Although weight selection remains a subjective decision, researchers can validate the integrated model by selecting a suitable weight. Since the quality of each dataset is known in this study, the challenge of subjectively weighting sources during model validation is not necessary.

Zurell et al. (2010) have provided another approach by evaluating the model using the virtual ecologist (VE) approach. The main difference between the virtual ecologist approach and the empirical ecologist approach is that in the VE approach, a virtual species is simulated, then a sample is taken, and models are built based on the sample. The information about virtual species is all known. Therefore it overcomes the main challenge of not having a true distribution model. This method's strength lies in its capacity to conduct an in-depth assessment of method performance in comparison to a known truth (Zurell et al., 2010). By using this approach one is able is test models on virtual data to determine their performance characteristics and then apply the best model and modeling strategies on real data.

2.6 Related work

Ahmad Suhaimi et al. (2021) tested several modeling techniques proposed for implementing integrated species distribution models (IDMs), including joint likelihood models, informative priors, and integration via correlation method. The goal was to evaluate the performance of various integrated models in realistic ecological data scenarios. Additionally, they used a virtual ecologist approach to explore which integrated model performs best under different levels of spatial bias in PO data, varying sample sizes of PA data, and different degrees of spatial overlap between datasets.

They created an intensity surface over a 100 by 100 grid to represent the true spatial patterns of species distributions. This intensity surface was modeled using a log-Gaussian Cox process. The intensity function in this model accounts for both environmental effects and random spatial terms in determining the species distribution. Two types of observation processes were simulated: a presence-absence (PA) dataset, which simulated a structured survey, and a presence-only (PO)

dataset, representing unstructured citizen science data. It was assumed that each location in the PA samples was visited only once, with perfect detection. For the PO samples, the study area was divided into five regions, with the probability of presence in each region depending on the specific detection probabilities defined in each scenario. Presence-only data, generated from a realization of the log-Gaussian Cox process, were then thinned according to these detection probabilities.

Three performance metrics were used for evaluation. First, accuracy was determined by calculating the mean absolute error (MAE) between the predicted and true log intensity values. Smaller errors indicate that the predictions are closer to the actual values, suggesting a good model fit. Second, the Pearson correlation between predicted and actual log intensities was computed to assess the similarity of the predicted spatial distributions. This correlation provides insight into how well the model captures spatial patterns. Finally, bias in the environmental covariate coefficient estimates was examined, as the true parameter value was known. Fitting and model performance can be measured by the difference of mean estimates and credible intervals of the posterior distribution.

There are 3 approaches tested in this literature, simple joint likelihood model, informed prior model, and correlation method. The authors consider the covariates method to be similar to the informed prior model, and in their study, only the informed prior model was tested. PA-only and PO-only models were also tested for comparison. Each approach was tested as one with a covariate explaining bias and one without. Ahmad tested each approach in 4 different scenarios: Default, Large sample, spatially biased, and spatially unbiased. Each scenario with different probabilities of observing PO data that determines spatial bias in the data.

The results in Ahmad Suhaimi et al. (2021) demonstrate that the extent of performance differences in response to changes in spatial bias in PO data depended on the availability of a covariate to account for the bias. When a bias covariate was included, all integrated model types performed well, showing similar or lower MAE compared to the PA-only model. Among these, the informed prior model showed the smallest improvement over the PA-only model. Conversely, when no covariate was available to explain the bias, the spatial bias in PO data had a significantly greater impact on the joint likelihood model than on the informed prior or correlation models.

Ahmad Suhaimi et al. (2021) also conducted simulations examining different levels of overlap between PA and PO data. They found that the results varied based on how effectively the bias in PO data could be explained. When a covariate was available to account for the bias, incorporating a small amount of PA data anywhere within the domain enhanced performance relative to the PA-only and PO-only models in terms of MAE. Additionally, adding a small amount of PA data improved the correlation with the true intensity for both joint likelihood and correlation IDMs. However, when no covariate was available to explain the bias, the performance of the joint likelihood model in terms of MAE was influenced by the placement of the PA data. In contrast, the informed prior and correlation models outperformed the PA-only model and were less affected by the placement of PA data.

The conclusion was then drawn based on the result from testing. The joint likelihood model had the best performance when spatial bias in PO data was low or could be modeled. However, it gave poor estimates when there were unknown biases in the data. Correlation models consistently provided good model performance. The informative prior methods had little improvement over modeling PA data alone and were inferior compared to either the joint likelihood or correlation approach. According to their findings, correlation models offer a reliable substitute for joint likelihood models in situations when covariates related to effort or detection in PO data are not available. Therefore, we should be aware of the limitations of each approach and consider how well biases in the data can be modeled. Ahmad's result suggests bias from PO data has a big impact on the performance of

IDMs. Integration modeling has the best performance when bias handling methods have properly worked. On the other hand, the correlation approach can perform well even when the bias is unknown.

3 Data

3.1 Real world data

The distributions of three bird species over Africa were studied. These are the Martial Eagle (*Polemaetus bellicosus*), Long-crested Eagle (*Lophaetus occipitalis*), and Peregrine Falcon (*Falco peregrinus*). Each species had a unique distribution which allows us to evaluate the model properly. The Martial Eagle is widespread in Southern Africa but uncommon, with a patchy distribution. The Long-crested Eagle has a smaller range but is commonly seen where it lives, while the Peregrine Falcon is widespread and rare all over Africa (BirdLife, 2012).

The Martial Eagle is a large bird of prey primarily found in sub-Saharan Africa. It is known for its powerful build and broad wings with a wingspan reaching up to 2.6 meters and it is one of the largest eagles in Africa. Martial Eagles are exceptional hunters, preying mainly on medium-sized mammals, birds, and reptiles. They typically inhabit open grasslands, savannas, and semi-arid regions, nesting in tall trees (Kemp et al., 2020). Despite being apex predators, their numbers are declining due to habitat loss and human activities. The primary cause of habitat loss is the deforestation of large trees used for nesting which have been transformed into agricultural fields. Therefore, knowledge of the Martial Eagle's distribution is crucial for shaping conservation policies in those areas. Additionally, two other bird species are fitted to the distribution model in this study: the Long-crested Eagle and the Peregrine Falcon. Both birds share some distribution similarities with the Martial Eagle, though there are also distinct differences (BirdLife, 2012). The Long-crested Eagle has a relatively limited range in Africa but maintains a considerable population within its habitats. In contrast, the Peregrine Falcon is widely distributed across Africa, although it is found rarely.

The datasets used in this project are the eBird dataset (Sullivan et al., 2009) and the Second Southern African Bird Atlas Project (SABAP2) dataset (Brooks et al., 2022). Table 2 is a comparison of two datasets.

Dataset	Data type	Sampling unit
eBird	Presence only (PO)	checklist
SABAP2	Presence-absence (PA)	Pentad (5min by 5min area)

Table 2: Comparison of two datasets

eBird is the world's largest citizen science project related to biodiversity, with more than 100 million bird sightings contributed annually by eBirders worldwide. Managed by the Cornell Lab of Ornithology, it provides scientists, researchers, and amateur naturalists with real-time data on bird distribution and abundance. eBird employs various observation protocols to standardize data collection and enhance its utility for research and conservation (Sullivan et al., 2009). The primary protocols include:

- Stationary Count: Observers remain at a fixed location, recording all bird species detected within a specified time frame.
- Traveling Count: Observers move over a defined distance, documenting all bird species encountered along the route.
- Incidental Observation: Bird detections occur during activities where birding is not the primary focus; these observations are recorded without systematic effort.

Each protocol requires specific effort details, such as duration, distance traveled, and start time, to provide context for the observations. Selecting the appropriate protocol ensures the data's accuracy and relevance for analysis. For this dissertation, only eBird data collected under the first two primary standardized protocols: Stationary Counts and Traveling Counts were used.

On the other hand, SABAP2 is a citizen science project that is driven by the energy of several hundred volunteers who are mapping the distribution of birds across southern Africa (Brooks et al., 2022). South Africa, Lesotho, Botswana, Namibia, Mozambique, Eswatini, Zimbabwe, and Zambia are all included in the project, which attempts to map the distribution and relative abundance of birds in southern Africa. To gather data, volunteers follow a strict protocol, where they select a location, named a pentad and record all the bird species seen within a set time frame. After that, the information is added to the SABAP2 database for examination and study. It is supported by the South African National Biodiversity Institute and the FitzPatrick Institute of African Ornithology and has its headquarters at the University of Cape Town (Brooks et al., 2022). Since only individuals with extensive ornithological knowledge and experience contribute to SABAP2, the data are highly reliable. It employs a single standardized data collection protocol to map bird distributions and relative abundances across southern Africa(Brooks et al., 2022). Participants, known as citizen scientists, conduct surveys within defined 5-minute latitude by 5-minute longitude grid cells, termed 'pentads'. For a 'full-protocol' survey (Brooks et al., 2022), observers spend a minimum of two hours and up to five days recording all bird species detected within the pentad, ensuring coverage of as many different habitats as possible. Species are documented in the order of detection to help assess their relative commonness.

The data used in this dissertation were the SABAP2 data collected in South Africa between 2018 and 2020, and eBird data collected across Africa up to 2024. A longer time span of eBird data was used to increase the amount of presence-only data, whereas a shorter time interval was chosen for the SABAP2 data to ensure that the presence-absence data are fewer than the presence-only data.

3.2 Virtual data via Virtual Ecologist approach

The Virtual Ecologist (VE) approach is a simulation-based method that enables researchers to generate artificial species distribution data while incorporating observer behavior (Zurell et al., 2010). This approach allows us to evaluate statistical models against known simulated truths, addressing challenges posed by the lack of true species distributions in empirical research.

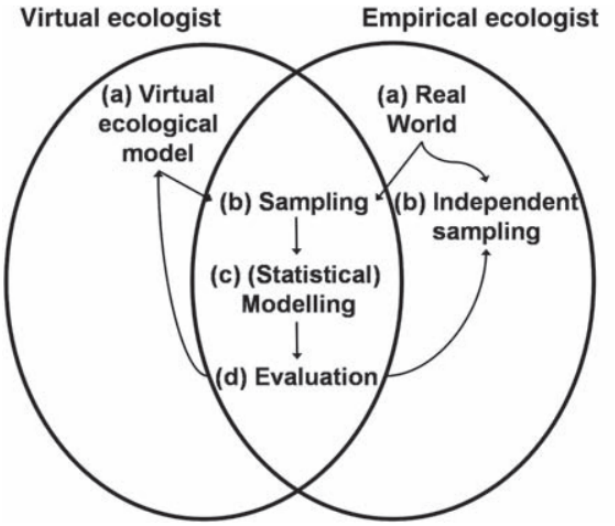


Figure 3: The elements of the virtual ecologist approach, directly taken from Zurell et al. (2010)

As shown in Figure 3, the VE approach consists of four key components:

1. A virtual ecological model that defines species distributions based on ecological factors
2. A virtual sampling model that mimics real-world observation biases
3. Statistical modeling
4. Evaluation

This framework provides a controlled environment for testing sampling protocols and modeling approaches.

In this study, we employed the VE approach to generate virtual species data and assess model performance under different sampling scenarios. We created two virtual species with distinct prevalence levels: one with low prevalence (30%) and another with moderate prevalence (60%). Suitable habitats were defined using four environmental variables: annual mean temperature, maximum temperature of the warmest month, minimum temperature of the coldest month, and annual precipitation. PO data were then sampled with varying levels of bias to simulate real-world citizen science data, while PA data were assumed to be unbiased and were simply sampled randomly from the map.

By leveraging the VE approach, we systematically analyze the impact of sampling bias on species distribution models and compare different methods for mitigating these biases. This controlled experimental setup allows us to explore hypothetical scenarios and assess model robustness under varying conditions.

3.3 Predictor variables

In this study, a range of environmental and anthropogenic variables were used as predictors. These variables capture key ecological factors that influence species occurrence and habitat suitability.

Bioclimatic variables from the WorldClim database (Fick and Hijmans, 2017) provide temperature and precipitation data, while additional covariates such as land cover, soil, elevation, and population density account for habitat characteristics and potential sampling biases. By incorporating these predictors, the models aim to improve the accuracy of species distribution estimates and mitigate biases associated with PO data.

3.3.1 Bioclimatic variables

The WorldClim database provided a set of bioclimatic characteristics that were utilized as predictors (Fick and Hijmans, 2017). While some covariates, such as cloud cover, were taken from the MODIS satellite platform (Hufkens, 2023), others were interpolated using data that was normally collected at weather stations. The gridded data for these covariates have a geographic resolution of one kilometer. A detailed description of the variables used is provided below:

BIO1	Annual Mean Temperature
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3	Isothermality (BIO2/BIO7) ($\times$100)
BIO4	Temperature Seasonality (standard deviation $\times$100)
BIO5	Max Temperature of Warmest Month
BIO6	Min Temperature of Coldest Month
BIO7	Temperature Annual Range (BIO5-BIO6)
BIO8	Mean Temperature of Wettest Quarter
BIO9	Mean Temperature of Driest Quarter
BIO10	Mean Temperature of Warmest Quarter
BIO11	Mean Temperature of Coldest Quarter
BIO12	Annual Precipitation
BIO13	Precipitation of Wettest Month
BIO14	Precipitation of Driest Month
BIO15	Precipitation Seasonality (Coefficient of Variation)
BIO16	Precipitation of Wettest Quarter
BIO17	Precipitation of Driest Quarter
BIO18	Precipitation of Warmest Quarter
BIO19	Precipitation of Coldest Quarter

Table 3: Table of Bio climatic variables, directly taken from Fick and Hijmans (2017)

3.3.2 World Population Density

The world population density is a measurement of the human population per unit of land area. World population density is a key geographical term. As we had PO data from the citizen science, world population density measurement has been included as a bias covariate to represent possible sampling bias and reachability of people. Also, due to the special characteristics of Africa, the land is vast and the population is sparse, we believe world population density could be one of the major factors driving sampling bias. Figure 4 is a map of world population density in Africa.

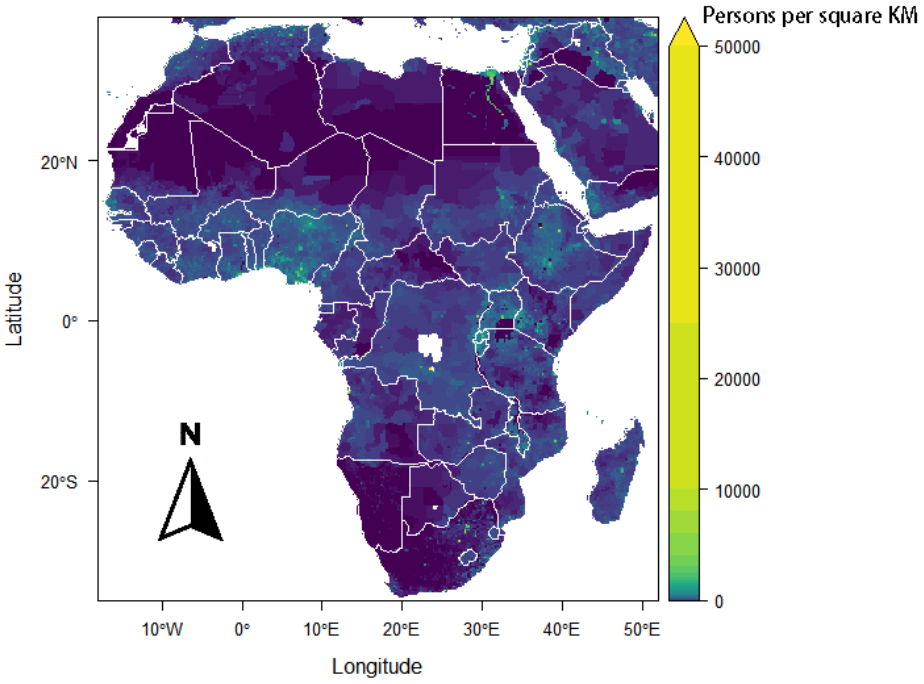


Figure 4: Population density across Africa, used as a covariate to account for sampling bias in citizen-science data

3.3.3 Topographical and Landcover variables

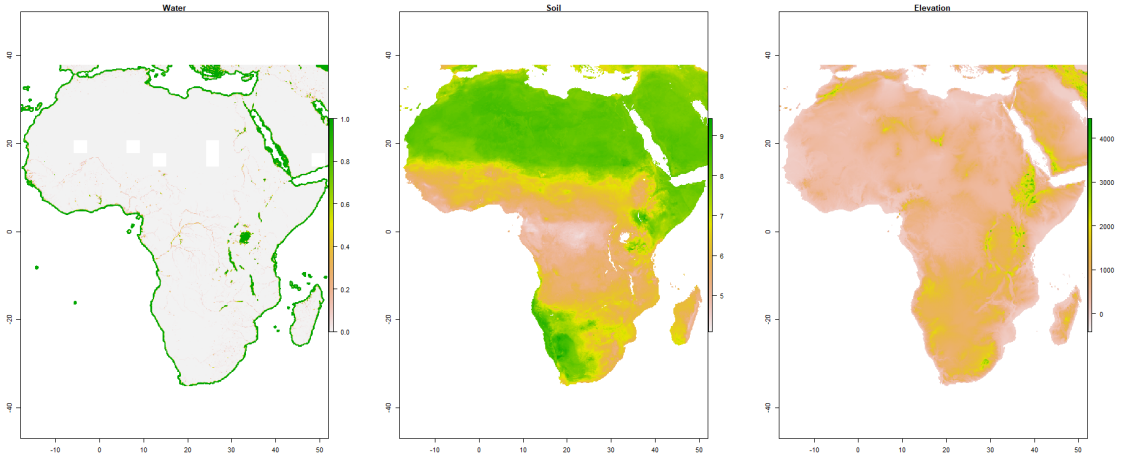


Figure 5: The maps of water landcover, Soil and elevation in Africa (From left to right: Water Land Cover, Soil, and Elevation)

Figure 5 is the map of water land cover, Soil, and elevation in Africa. Water landcover, soil, elevation, and some more land cover variables were used as predictor variables:

- Water Landcover: Refers to areas covered by water, such as lakes, rivers, and wetlands. It influences species that depend on water availability for survival and can predict biodiversity patterns in aquatic or riparian ecosystems.
- Soil: Soil properties affect plant growth, influencing vegetation types and indirectly impacting animal species reliant on certain habitats.
- Elevation: Elevation impacts climate (temperature, humidity) and it helps model variations in ecosystems across altitudinal gradients.
- Trees: The tree land cover represents the proportion or density of tree cover in a specific area, often used to assess vegetation patterns, habitat quality, and ecosystem health. It is quite useful as birds generally prefer to rest in trees.
- grassland: The grassland land cover covariate represents the proportion or density of grassland land cover in a specific area to support ecological analysis and species habitat studies.
- shrubs: The shrubs land cover variable indicates the presence and density of shrub vegetation in a given area, commonly used in studies of habitat composition and ecosystem analysis.
- cropland: The cropland land cover variable represents the extent and distribution of cultivated agricultural land within an area, often used to analyze land use patterns and agricultural impact on ecosystems.
- built: The built land cover variable indicates the extent and distribution of developed areas, such as urban infrastructure and built environments, used to assess human impact on land use and ecological dynamics.
- wetland: The wetland land cover variable represents the area and characteristics of wetland ecosystems, used to evaluate hydrological functions and biodiversity within these habitats.

These variables are generally expected to influence species distributions, although their effects are not always clear. In this dissertation, they are included to provide a broad exploratory analysis. The data come from Hijmans et al. (2023).

4 Methods

4.1 Simulating Virtual Species

Environmental suitability is defined as the probability of occurrence of a species in a given environment, based on ecological and climatic conditions. This probability, denoted as $E(z)$, represents the conditional probability of species occurrence given the environment, $P(y = 1|z)$, and can be expressed using Bayes' rule in Equation 7.

$$E(z) = P(y = 1|z) = \frac{f_1(z)P(y = 1)}{f(z)} \tag{7}$$

where z is a vector of environment measurements at location i, $P(y = 1)$ represents species prevalence, $f(z)$ is the environmental distribution (the joint density of environments in nature), $f_1(z)$ is the species occurrence distribution (the joint density of environments where the specie is found) (Drake and Richards, 2018).

Equation 7 is the basis of generating a virtual species distribution and represents a species' response to different environmental gradients. The entire simulation process was carried out using three functions from the virtualSpecies R package (Leroy et al., 2015):

1. `generateSpFromFun`: simulates environmental suitability
2. `convertToPA`: converted the environmental suitability into presence-absence data (1 for presence, 0 for absence)
3. `sampleOccurrences`: Samples observed occurrences from the presence-absence map

To simulate the environmental suitability, we defined a response function for a virtual species to chosen environmental variables and combine these responses to define the environmental suitability. Response functions for different predictor variables were combined using a weighted average to obtain a habitat suitability value. That is the most commonly used method to create virtual species distributions and is the method implemented in the function `generateSpFromFun`. It has also been applied in numerous studies, such as Hirzel et al. (2001), Zurell et al. (2010), and Leroy et al. (2015), which used similar frameworks to simulate species with predefined ecological response functions.

Essentially, simulating a species' environmental suitability map involves capturing the species' response to various environmental gradients. Higher values of environmental suitability on the map suggest a higher likelihood of species presence. To determine realistic preferences and species distributions, and to ensure that the species-environment relationships are consistent across the landscape, mechanistic response curves are developed sequentially. (Inman et al., 2021).

Initially, a single bioclimatic covariate was chosen, and an appropriate response curve was selected from linear, Gaussian, logistic, or quadratic functions to create a habitat map based solely on that covariate. The resulting habitat values were then linearly rescaled to fall between 0 and 1, with values exceeding 0.5 indicating areas of high environmental suitability. For the remaining covariates, the user defined a response curve, which is used to predict new habitat values by calculating the weighted sum of the individual response functions. The final habitat values were then rescaled linearly to a range between 0 and 1. This procedure is repeated until we have the distribution of habitat values for all selected bioclimatic covariates. A schematic framework adopted from Inman et al. (2021), illustrating the creation of virtual species as described above,

provides a visual summary, as shown in Figure 6. Afterward, all selected bio-climatic covariates and their habitat functions are combined using the `generateSpFromFun` function.

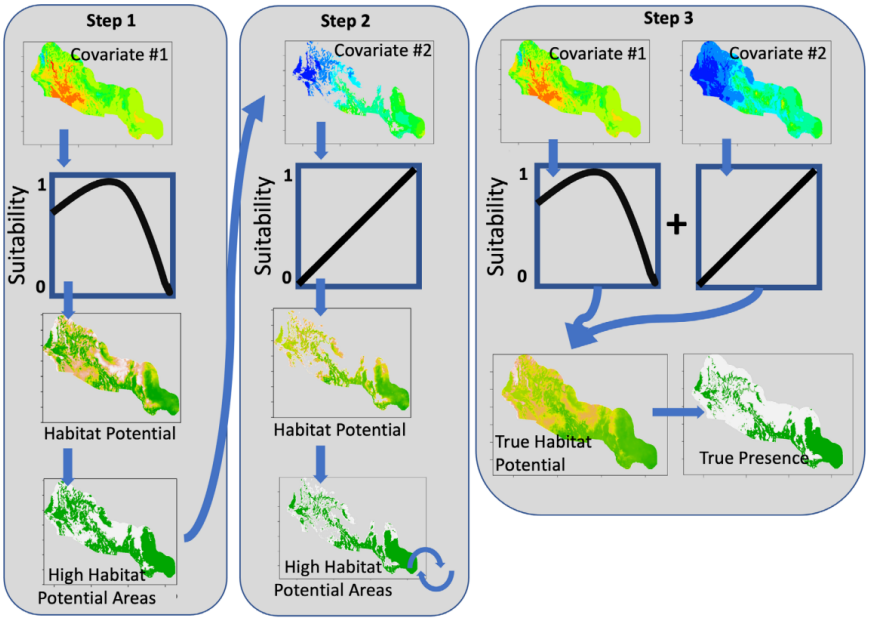


Figure 6: A visualization of the framework for virtual species creation, directly taken from Inman et al. (2021)

The environmental suitability was then converted to presence-absence data. The traditional approach typically involves setting a threshold to transform environmental suitability into a binary presence-absence map (Inman et al., 2021). This method will always result in threshold responses and does not replicate the random processes acting on species occupancies. Such unrealistic virtual species can produce deceptive findings regarding the ability of modeling techniques to predict species distributions. An appropriate alternative applies a probabilistic approach to convert environmental suitability into a probability of occurrence first and then subsequently random draws using the probability of occurrence to determine whether a particular cell is turned into a presence or an absence. As a result, a binomial experiment is conducted in each cell, where the probability of occurrence represents the success probability. For example, a cell with a 0.4 probability of occurrence will have a presence assigned in 4 out of 10 cases according to this probabilistic method. This probabilistic conversion to presence-absence implies that repetitions of the conversion process differed, each providing a valid realization of the true species distribution map (Leroy et al., 2015). This improved approach is implemented in the virtualSpecies package using the function `convertToPA`.

Lastly, we sampled observed occurrences for the virtual species using the function `sampleOccurrences`. This function can be used to generate occurrence data in different ways. It can produce samples of presence-absence or presence-only data. The function, `sampleOccurrences` allows us to assign a probability of detection to the virtual species which is important given the impact of imperfect detection on SDM performance. This probability of detection can be weighted by environmental suitability to simulate smaller populations in less suitable areas (Leroy et al., 2015). Additionally, a

sampling intensity bias can be applied to simulate over-sampled or under-sampled regions, further enhancing the realism of the species distribution modeling process.

4.2 Point process models

Point process models, particularly inhomogeneous Poisson point process (IPP) models, are powerful tools in species distribution modeling. Several commonly used algorithms—such as logistic regression, MaxEnt, and boosted regression trees—can be interpreted as approximations to the IPP model, each estimating its intensity function with varying degrees of accuracy (Fletcher et al., 2019). Also, though the general framework for integrated modeling allows for the combination of datasets of the same resolutions, when datasets have different resolutions or different data protocols, the general framework cannot work, and the IPP framework can be a solution.

The distribution of points in space is described statistically by the point process. One way to conceptualize the points in an ecological context is as the instantaneous positions of each individual of the species (Isaac et al., 2020). The number of points inside a particular region is the site abundance, and the presence or absence of points inside a particular region is site occupancy (Isaac et al., 2020). This framework covers a range of data types and model structures including count data, presence-absence (PA) data, and presence-only (PO) data. By using an IPP framework wherein all the data are assumed to be just different realizations of the same Poisson process, we can combine datasets with different resolutions.

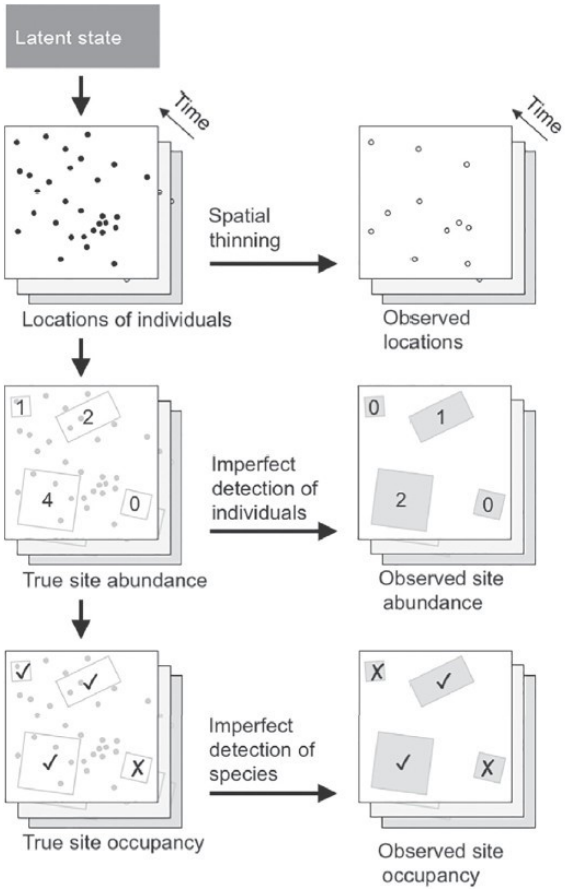


Figure 7: An illustration of the interrelationships between different types of ecological data and models, directly taken from Isaac et al. (2020)

Figure 7 illustrates how the point process links each data type. Pictures in the first column show the same distribution of true locations of individuals in the region. In collecting abundance data, the detection of individuals is nearly always imperfect, and the resulting data will be from a subset of all present individuals in the region. The first row illustrates how PO data arise, where the data consist of point locations at which individuals were observed. Due to imperfect detection, the PO data (top-right panel) are thinned locations of where all the species individuals are located in the region (top-left panel). There may even exist some spatial sampling bias if observation effort is not uniform in the region. The second row illustrates how count data arises. In the collection of count data areas are demarcated (middle left panel) and the number of species individuals observed in each sub-region are recorded (middle right panel). Again some individuals may go unobserved due to imperfect detection. Lastly, the third row, illustrates how PA data arise. PA data like count data is also recorded by sites. In this case, four sites have been designed and the location within four sites has been counted. Due to imperfect detection, only some of the locations have been counted in the data. Lastly, structured presence-absence is also recorded by sites. Observed site occupancy is then generated from whether each site has detection or not. Under the point process framework, each data type follows a specific distribution and it allows us to combine them via the joint likelihood function.

IPP framework provides a foundation for combining multiple datasets. As a data-generating mechanism, a realization from an IPP generates random points in geographic space. The IPP offers a solid basis for extensions to less-ideal data and is a logical choice for modeling a species distribution using idealized data. The probability density function of a realization from an IPP is

$$f(n, s_1, \cdots, s_n) = \frac{e^{-\int_S \lambda(s)ds} \left(\int_s \lambda(s)ds\right)^n}{n!} n! \prod_{i=1}^{n} \frac{\lambda(s_i)}{\lambda(s)ds}, \tag{8}$$

which simplifies to $e^{-\int_S \lambda(s)ds} \prod_{i=1}^{n} \lambda(S_i)$, where s_i is a vector that contains the coordinates of the i^{th} individual location and n is the total number of locations (Fletcher et al., 2019). The spatially varying intensity function $\lambda(s)$ controls the expected number and location of the random points within the study area S. An important quantity in Equation 8 is the integrated intensity function, $\int_S \lambda(s)ds$, which can be used to show the connection between exact point locations (presence-only data) and grid-based locations (presence-absence and count data) (Fletcher et al., 2019).

Regression models can be formulated by specifying a probability distribution for the response variable and defining a deterministic relationship between predictor variables and the expected value of the response variable. The IPP can be formulated as a regression model by assuming the intensity function depends on location-specific covariates $x(s)$ as Equation 9.

$$\log(\lambda(s)) = \beta_0 + x(s)'\boldsymbol{\beta}, \tag{9}$$

where β_0 is the intercept and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ is a vector of regression coefficients associated with covariates.

The connection between presence-only data tends to be the exact coordinates of where individuals occurred (Fletcher et al., 2019). Presence–absence and count data, which tend to be assigned to grid cells, can be interpreted through a statistical method called the change of support (Fletcher et al., 2019). The support of a probability distribution is the set of all random values that can be generated. The support of the IPP in Equation 8 is the infinite set of all possible locations within the study area S. In contrast, the support of count data is limited to a finite number of grid cells where the random variable is the number of points within each cell (Fletcher et al., 2019). Thus, the IPP can be formally connected to count data by transitioning from continuous spatial support to discrete spatial support. By implementing this change of support, the number of points y_j was contained within the j^{th} grid cell A_j becomes a Poisson random variable:

$$y_j \sim Poisson\left(\int_{A_j} \lambda(s)ds\right), \tag{10}$$

where the expected number of points is $\int_{A_j} \lambda(s)ds$. Presence–absence data are related to count data by recording the grid cells 0 or 1 depending on whether $y_j = 0$ or $y_j > 0$. Since the Poisson process is just a continuous time version of the Bernoulli trials process, we can use that relationship to connect them:

$$I(y_j > 0) \sim Bernoulli\left(1 - e^{-\int_{A_j} \lambda(s)ds}\right), \tag{11}$$

where $I(y_j > 0)$ is an indicator function that takes on a value of 1 if the species is present and 0 if the species is absent. IPP can thus effectively connect different data types and describe several distribution modeling frameworks. They therefore allow for the fitting of integrated models with different data types.

In our modeling, a log-Gaussian Cox process (LGCP) was assumed for the locations of individuals for all data types as in Simmonds et al. (2020). LGCPs represent a doubly stochastic Poisson process as the intensity itself is stochastic. It is a commonly used model for the analysis of spatial point pattern data (Møller et al., 1998). The intensity function is defined by Equation 12 and Equation 13. Let $N(A)$ denote the number of individuals observed in the region A.

$$N(A) \sim Poisson\left(\int_A \lambda(s)ds\right), \tag{12}$$

$$\log(\lambda(s)) = \alpha_0 + \hat{\beta}_x x(s) + \hat{\xi}(s) + \varepsilon(s), \tag{13}$$

where the log of the intensity $\lambda(s)$ was defined by an intercept α_0, a linear relationship β_x with an environmental covariate $x(s)$, a Gaussian random field $\xi(s)$ which simulates spatial variation not explained by the environmental covariates, and some random error $\varepsilon(s)$ (Simmonds et al., 2020). As a result, all three data types can be modeled as originating from the same underlying state as Equation 13, but with different observation processes and therefore different intercepts.

Fitting of all models was done via an R package PointedSDMs (Mostert et al., 2023). PointedSDMs provide tools to integrate different types of occurrence data through point process species distribution models (SDMs) conveniently. It does so using the now well-established integrated nested Laplace approximation (INLA) methodology and by constructing wrapper functions around the R package inlabru (Mostert et al., 2023).

The PointedSDMs package uses a hierarchical modeling structure with an underlying process model that provides a statistical description of how points are distributed in space and the role of such is a reflection of how multiple data types emerge from the same system (Mostert et al., 2023). This process has a spatially varying intensity function $\lambda(s)$ which is a response function of environmental covariates X and bias parameters such that a higher intensity indicates that the species is more abundant in a given area. An illustration of the hierarchical setup of this model can be seen in Figure 8.

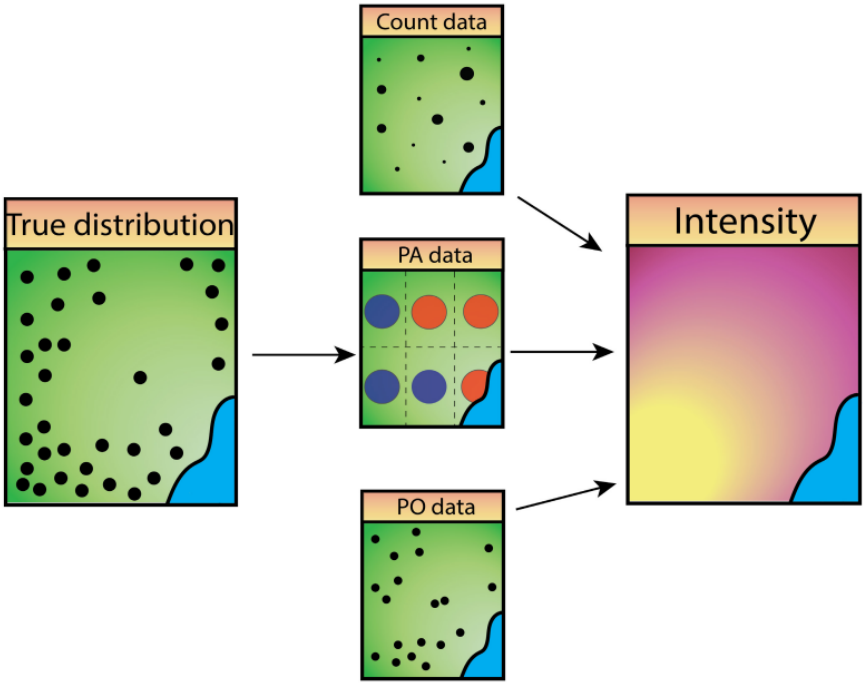


Figure 8: An illustration of the integrated species distribution model's structure, in which every dataset represents a distinct realization of the true species distribution, directly taken from Mostert et al. (2023)

For the hierarchical model, it is assumed that the underlying process model is a log-Gaussian Cox process (LGCP) with an intensity function given $\lambda(s) = exp\{\eta(s)\}$ which describes the expected number of individuals at some location s. The log of this intensity function is thus given as Equation 14, which is a rewritten form of Equation 13.

$$\log(\lambda(s)) = \eta(s) = \alpha + \sum_{u=1}^{k} \beta_i X_i(s) + \zeta(s), \tag{14}$$

where α is a dataset-specific intercept term, β_i is the coefficient associated with the ith environmental covariate, and $\zeta(s)$ is a zero-mean spatially continuous Gaussian random field (GRF), included in the model to account for potential spatial autocorrelation and the effects of all the environmental covariates not included in the model (Mostert et al., 2023). Therefore, the expected number of species' presences within a region S is given by the integral of the intensity function across the entire region as Equation 15.

$$\mu(s) = \int_S \lambda(s) ds. \tag{15}$$

Next, it is assumed that each dataset process has its own observation model, which provides a statistical description of the data-collection process as described earlier. These models link the

intensity function to the dataset's assumed likelihood, given by $L(Y_i|\lambda(s), \theta_i)$, where θ_i are the parameters for the ith observation model (Mostert et al., 2023). Then, by combining the process model with the observation models, the full likelihood for the data processes $\boldsymbol{Y} = Y_1, Y_2, \ldots, Y_m$ is given by Equation 16 (Mostert et al., 2023).

$$L(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\phi}) \propto p(\lambda(s), \boldsymbol{X}, \phi) \cdot \prod_{i=1}^{n} L(Y_i|\lambda(s), \theta_i), \tag{16}$$

which is the model component for the latent state of the model multiplied by the product of the individual likelihoods for the data processes where $p(\lambda(s), \boldsymbol{X}, \phi)$ is the latent state of the model follows a log-Gaussian Cox process.

4.3 Spatial thinning

As discussed in Section 2.4.4, spatial thinning can be applied to spatially biased presence-only data prior to modeling to reduce sampling bias and improve inference. Several thinning algorithms are available, as described in Section 2.4.4. A key practical challenge is selecting an appropriate thinning parameter, typically a minimum inter-point distance. To guide this choice, we proposed using the pair correlation function (PCF), a summary statistic that characterizes the spatial structure of a point pattern at different scales.

The PCF describes how the density of points changes as a function of distance from an arbitrary point, providing a non-cumulative measure of local clustering or inhibition. It is defined in terms of the derivative of Ripley's K function:

$$G(r) = \frac{K'(r)}{2\pi r}, \tag{17}$$

where $K'(r)$ is the derivative of $K(r)$, which summarizes the accumulated number of neighboring points within distance r. Because the PCF evaluates spatial structure at specific distances rather than aggregating information across scales, it offers greater sensitivity to local patterns than cumulative statistics such as the K or L functions (Bevan and Lake, 2022). This property makes it particularly useful for identifying characteristic spacing in the data, and therefore well suited for selecting thinning parameters.

When the underlying point pattern exhibits spatially varying intensity, the standard PCF may confound true interactions with large-scale intensity gradients. In such cases, the inhomogeneous pair correlation function $G_{inhom}(r)$ provides a more appropriate summary by adjusting for spatial variation in point density (Baddeley et al., 2015). Conceptually, the probability $p(r)$ of observing two points at locations x and y separated by distance r can be expressed as

$$p(r) = \lambda(x)\lambda(y)g(r)dxdy, \tag{18}$$

where $\lambda(\cdot)$ denotes the spatially varying intensity of the point process. For a homogeneous Poisson process, $g(r) = 1$, indicating no interaction among points. Deviations from this baseline reveal clustering ($g(r) > 1$) or inhibition ($g(r) < 1$) at scale r.

In practice, estimation of the inhomogeneous PCF requires smoothing to obtain a stable derivative of the K function. This is commonly achieved via kernel density estimation (KDE), resulting in a smoothed estimator:

$$G_\sigma(r) = \frac{1}{2\pi r\lambda} \sum_{i=1}^{n} \sum_{j \neq i} K_\sigma(r - d_{ij}), \tag{19}$$

where $K_\sigma(\cdot)$ is a kernel function, typically Gaussian:

$$K_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{x^2}{2\sigma^2}). \tag{20}$$

The bandwidth parameter σ controls the degree of smoothing, with larger values producing smoother curves that emphasize broad-scale trends, and smaller values allowing finer-scale structure to be detected (Wiegand and Moloney, 2013; Diggle, 2013). Choosing an appropriate level of smoothing is therefore essential when using the PCF to identify clustering scales relevant for thinning.

4.4 Performance measures

Performance measures are used to evaluate the performance of statistical and machine learning models. The Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) assesses classification accuracy by measuring the model's ability to distinguish between classes. Mean Absolute Error (MAE) quantifies the average magnitude of prediction errors, providing an intuitive measure of model accuracy. Pearson correlation evaluates the linear relationship between predicted and observed values. Tjur's R^2 is a goodness-of-fit metric for logistic regression, measuring the difference in predicted probabilities between groups. Watanabe-Akaike Information Criterion (WAIC) estimates model fit while accounting for complexity, helping to compare and select models.

Both real-world and virtual datasets are used in this project. The first four metrics (AUC, MAE, Pearson correlation, and Tjur R^2) are used to assess the accuracy of models on virtual data, while WAIC is used to measure model performance on real-world data, ensuring a balance between model fit and complexity. Additionally, for virtual datasets, the accuracy and precision of the estimates were also used to assess model performance, providing further insight into the reliability of the predictions.

4.4.1 Area under the ROC curve (AUC)

One often-used statistic for model discrimination is the area under the curve (AUC). It indicates the proportion of the area under the Receiver Operating Characteristic (ROC) curve, a graph that shows the varying false positive rates and true positive rates that a model can generate for different threshold values.

This performance metric is very helpful for classification problems since it shows how well the model works in different scenarios, giving a better idea of how accurate it is. Essentially, the classification of an observation into one class or another is determined by a threshold k. Therefore, if the predicted value for an observation, based on its covariates, is y^*, the model will classify the observation into class 1 if $y^* < k$ and into class 2 if $y^* > k$, for a given value of the threshold k. The threshold k is determined by selecting the value that maximizes the Area Under the Curve (AUC), ensuring optimal model performance.

AUC is used to assess the discriminatory power of models in distinguishing presence from absence locations. We compare different models' AUC values to evaluate their relative performance.

4.4.2 Mean absolute error (MAE)

The accuracy was assessed by computing the mean absolute error between the predicted log intensity and the actual log intensity. MAE sums up the absolute difference between two intensities. It is less influenced by outliers compared to the mean squared error. Smaller errors suggest that the predicted values are closer to the validation dataset, indicating a better model fit. By using the virtual ecologist approach, we had true intensity; therefore, we are able to use MAE to inform on the ability of the models to return the absolute intensity values. MAE is used to compare model performance in estimating species intensity and to determine which model best captures the true underlying distribution.

4.4.3 Pearson correlation

The Pearson correlation between the predicted and actual log intensity was computed to assess the similarity in the predicted spatial distributions (Ahmad Suhaimi et al., 2021). The Pearson correlation is calculated as in Equation 21:

$$\rho_{Y,Y^*} = \frac{cov(Y, Y^*)}{\sigma_Y \sigma_{Y^*}}, \tag{21}$$

where $cov(Y, Y^*)$ is the covariance between the predicted log intensity Y^* and the actual log intensity Y, σ_Y is the standard deviation of actual log intensity and σ_{Y^*} is the standard deviation of predicted log intensity. A positive correlation indicates that the predicted log intensity increases alongside the true value, while a negative correlation suggests that the prediction pattern moves in the opposite direction to the actual values. The closer the correlation coefficient is to one, the more closely the predicted and actual spatial distributions align. Therefore Pearson correlation can indicate how well spatial patterns have been captured by the model.

Pearson correlation is used to assess the models' ability to capture spatial patterns of species intensity and to compare different models based on their spatial predictive accuracy

4.4.4 Tjur R^2

Tjur's R-squared is defined as the absolute difference between the mean predicted value on the absence locations and the mean predicted value on the presence locations.

$$T_{\text{jur}} = \left| \bar{\hat{p}}_{\text{absence}} - \bar{\hat{p}}_{\text{presence}} \right|, \tag{22}$$

where $p_{absence}$ is the average predicted probability for observations where the outcome was absent (0) and $p_{presence}$ is the average predicted probability where the outcome was present (1). This metric ranges between 0 and 1, with higher values indicating greater discriminatory power of the model. Tjur's R squared favors models with good discrimination ability that predict a very high probability of detection where a species was detected and a very low probability of detection where a species was not detected (Nephin et al., 2023).

It is used to evaluate the discrimination ability of the models, specifically how well they distinguish between presence and absence locations, as well as to assess the overall fit of the models in separating detected and non-detected species.

4.4.5 Watanabe-Akaike Information Criterion (WAIC)

Watanabe-Akaike Information Criterion (WAIC) is a Bayesian model selection criterion that evaluates a model's predictive performance while accounting for the complexity and uncertainty inherent in Bayesian models. It is particularly suitable for comparing complex hierarchical or structured models.

WAIC is designed to estimate the expected log pointwise predictive density, combining elements of model fit and model complexity. The calculation steps are as follows:

- Step 1: Calculate log-Likelihood for each data point: For each data point y_i and each posterior sample θ_s from the posterior distribution of parameters, compute the log-likelihood $\log p(y_i|\theta_s)$
- Step 2: Average over the posterior: For each data point, calculate the mean of the log-likelihood across all posterior samples, which approximates the posterior predictive likelihood.
- Step 3: Calculate the effective number of parameters: The effective number of parameters is estimated by computing the variance of the log-likelihoods across posterior samples. This variance reflects the model's complexity, indicating how sensitive the model is to the data points.
- Step 4: Compute WAIC:

$$\text{WAIC} = -2\sum_{i=1}^{n} \log \mathbb{E}_\theta \left[p(y_i \mid \boldsymbol{\theta})\right] + 2 \cdot \text{Effective Parameters}$$

  This formula combines model fit ($\sum_{i=1}^{n} \log \mathbb{E}_\theta \left[p(y_i \mid \boldsymbol{\theta})\right]$) and complexity (Effective Parameters).

WAIC provides a way to compare Bayesian models, with lower WAIC values indicating better models in terms of predictive accuracy and complexity balance (Spiegelhalter et al., 2002). By using all posterior samples, WAIC better accounts for model uncertainty compared to traditional criteria, making it especially useful for complex models with hierarchical structures or varying levels of certainty in parameter estimates. WAIC penalizes overly complex models by accounting for the effective number of parameters, helping to avoid overfitting while promoting models with better generalization (Vehtari et al., 2016).

Overall, WAIC is valuable in Bayesian analysis for selecting models that provide a balance between accuracy and complexity, particularly in settings with structured data or uncertain parameters.

4.4.6 The accuracy and precision of the estimates

For models tested in the virtual ecologist approach, the accuracy and precision of the estimates were assessed to evaluate how well the model parameters were estimated (Simmonds et al., 2020). Accuracy is measured by the difference between the parameter estimates and the true parameter values, while precision is measured by the variation of estimates across simulations. Through the virtual ecologist approach, we could directly measure model performance using true estimates. We intended to use accuracy and precision to evaluate the models' performance based on how close the parameter estimates are to the true values, which provided insights into the models' reliability. However, when applying the model to real data, it is not possible to assess accuracy and precision directly, as we do not have access to the true species distribution parameters.

5 Simulation experiment

5.1 Simulation of virtual species

In our study, we created two virtual species distributions with low and moderate prevalence using the R package virtualSpecies (Leroy et al., 2015). The habitat of each species is defined using four bio-climatic covariates from the WorldClim database (Fick and Hijmans, 2017): Annual Mean Temperature (bio1), Max Temperature of Warmest Month (bio5), Min Temperature of Coldest Month (bio6), and Annual Precipitation (bio12). These choices reflect our aim to simulate species with distributions across Africa, where temperature and precipitation gradients play a significant role in determining species' environmental suitability.

To generate the species' distribution, the habitat values derived from all selected covariates are processed to ensure they are consistent and cannot contradict each other, as described in section 5.1. The species response functions for bio1, bio5, and bio6 are modeled using linear response functions, while bio12 is modeled using a Gaussian response function. The linear response functions for temperature variables (bio1, bio5, bio6) were chosen because they provide a straightforward representation of how species might respond to changes in these climatic factors. For bio12, we chose a Gaussian response function because precipitation often has a non-linear relationship with species' distributions, making this function more appropriate for capturing the complex response to varying rainfall patterns across the continent. The resultant species response functions for bio1, bio5, bio6 and bio12 are as follows, respectively:

$$f(bio1) = 2 * bio1 + 0.05,$$

$$f(bio5) = 1 * bio5 + 0.1,$$

$$f(bio6) = -1.2 * bio6 + 1.5$$

and

$$f(bio12) = \frac{1}{3000\sqrt{2\pi}} e^{-\frac{(bio12-700)^2}{2*3000^2}}.$$

The virtual species were presumed to be in equilibrium with their environment for the purpose of simplicity. We converted the environmental suitability into a probability of occurrence for species occupying 30% (low prevalence) and 60% (moderate prevalence) of the study region as shown in Figure 9. Our interest in focusing on low and moderate prevalence was driven by two key considerations: first, highly prevalent species are generally more accurately estimated by statistical models, even under conditions of poor data quality. However, such species are also more likely to be influenced by sampling bias. Second, our interest lies in endangered and rare species which typically exhibit low prevalence. We wanted to investigate how the methods and models being investigated perform in these scenarios.

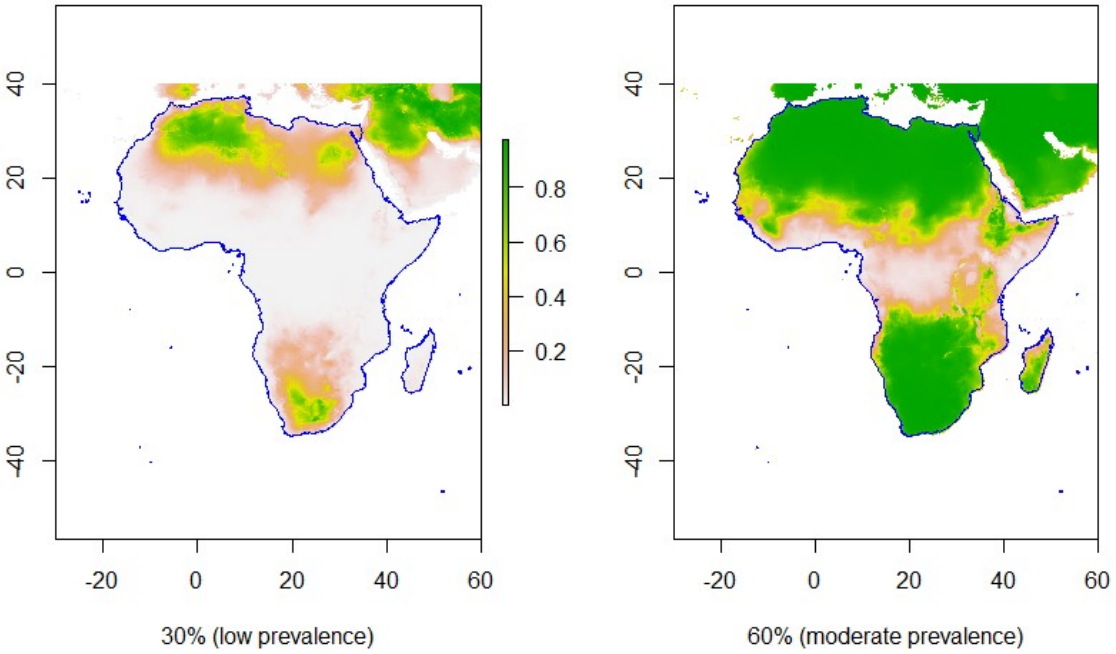


Figure 9: Conversion of environmental suitability in Africa into a probability of occurrence for species with low and moderate prevalence

5.2 Sampling

In this study, both presence-absence (PA) and presence-only (PO) data were sampled, considering a total of 12 scenarios with varying PO data sampling settings. The presence-absence data were assumed to be unbiased, meaning that they were sampled randomly from the map. For each scenario, 200 PA points are sampled, with occurrences determined directly by species suitability.

On the other hand, to simulate the sampling biases encountered in real-world citizen science data, our PO data were intentionally generated as biased samples. This approach aims to reflect biases often present in ecological data collected by volunteers or through other citizen science initiatives. We drew polygons all over Africa in which the sampling is biased. Sampling in each polygon was biased by an arbitrarily selected strength of 1, 3, and 10. A bias strength level of 10 means that there are 10 times more sample points in the chosen region compared to the rest of the sampled area. Conversely, a bias strength level of 1 represents an even sampling distribution across the entire space. A bias strength level of 1 was chosen as a control group with no effect from sampling bias. In this study, polygons were drawn according to the population density in Africa to mimic sampling bias that actually might occur due to reachability and population density as shown in Figure 10. The different colors of the polygons have no specific meaning; they are used only to indicate that there are distinct polygons.

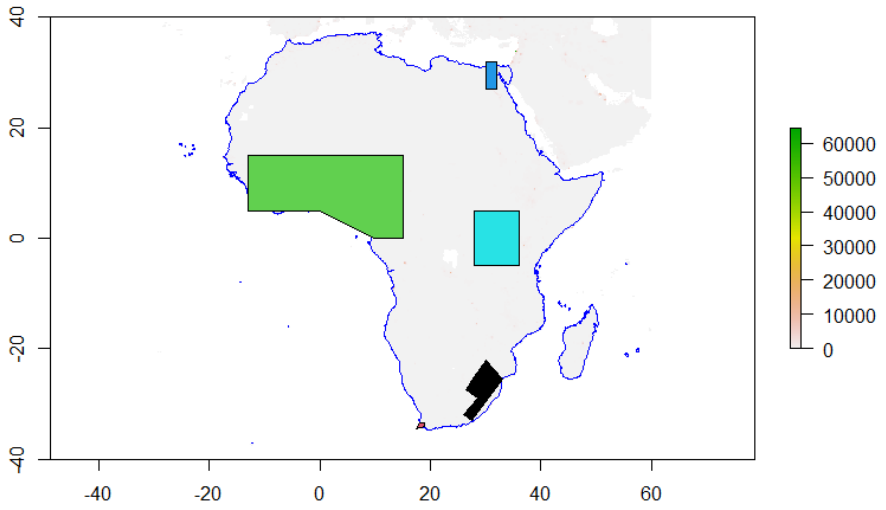


Figure 10: sample are biased in 5 polygons that are drawn around high population density areas. Different colors indicate distinct polygons and have no specific meaning.

Another consideration was to examine the impact of sample size on the ability of the methods being studied to accurately predict the true intensities and to assess whether increasing PO data has a positive effect. In order to do this, we biasedly sampled 500 and 5000 species occurrences in virtual ecosystems with low and moderate prevalence. A summary of the species location datasets we ended up with is given in Table 4. For both low (30%) and moderate (60%) prevalence virtual ecosystems, we sampled 6 times using 2 sample sizes (500 and 5000) across three different bias levels, resulting in a total of 12 datasets for each species prevalence.

Dataset	Prevalence	Bias level	Sample size
1	Low	1	500
2	Low	1	5000
3	Low	3	500
4	Low	3	5000
5	Low	10	500
6	Low	10	5000
7	Moderate	1	500
8	Moderate	1	5000
9	Moderate	3	500
10	Moderate	3	5000
11	Moderate	10	500
12	Moderate	10	5000

Table 4: Summary of the 12 datasets generated from both 30% and 60% prevalence virtual ecosystems for presence-only (PO) data

5.3 Integrated distribution models

In our study, we aimed to combine presence-absence data with presence-only data. As outlined in the discussion in section 5.2 through the inhomogeneous point process (IPP), we are able to multiply the likelihoods of each dataset as a joint likelihood to connect them formally. After formulating models for each data source, the joint likelihood for the integrated model is obtained by multiplying the component likelihoods:

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{m=1}^{M} L_m(\boldsymbol{\beta}, \boldsymbol{\theta}_m)$$

where $\boldsymbol{\beta}$ are shared parameters (among component likelihoods) explaining the observed density $\lambda(s)$ of the IPP and $\boldsymbol{\theta}_m$ are parameters associated with the observation process of the m^{th} dataset such as the probability of detection for presence-absence data or the bias covariates for presence-only data. The parameter $\boldsymbol{\theta}_m$ is typically not shared. The joint-likelihood approach for combining datasets typically assumes independence between the datasets, although this assumption can be relaxed (Fletcher et al., 2019).

Spatial bias from presence-only data is one of our main concerns when we build the integrated model. The flexibility of the framework allows us to handle the bias directly. Table 5 are standard models and two modifications to the model fitted to the data.

Model	Model description	Type	Response	Predictor
A	IDM	Integrated	PA	$\alpha_{PA} + \hat{\beta}_x x(s) + \hat{\xi}(s)$
			PO	$\alpha_{PO} + \hat{\beta}_x x(s) + \hat{\xi}(s)$
B	IDM with bias covariate	Integrated	PA	$\alpha_{PA} + \hat{\beta}_x x(s) + \hat{\xi}(s)$
			PO	$\alpha_{PO} + \hat{\beta}_x x(s) + \hat{\beta}_z z(s) + \hat{\xi}(s)$
C	IDM with second spatial field	Integrated	PA	$\alpha_{PA} + \hat{\beta}_x x(s) + \hat{\xi}(s)$
			PO	$\alpha_{PO} + \hat{\beta}_x x(s) + \hat{\xi}(s) + \hat{\zeta}(s)$

Table 5: Model types fit in this study

In the given models, $x(s)$ represents the shared environmental covariates between the PO and PA data, while $\hat{\xi}(s)$ represents the shared spatial field between the two data types, $z(s)$ and $\hat{\zeta}(s)$ are two additional spatial components. Model A is the standard integrated model that jointly combines presence-only and presence–absence data to estimate species–environment relationships, but does so without incorporating an explicit bias component for the PO data. Spatial bias present in the PO data is accounted for by extending the models in two ways. Model B includes the covariate $z(s)$ to model sampling bias and hence accounts for the spatial bias in the observation probability of the PO data. The main purpose of including bias covariates in the model is to correct for relative biases inherent in presence-only (PO) datasets. These covariates primarily correct relative sampling biases, but do not fully address systematic biases arising from non-detection. Model C includes a second spatial field, $\hat{\zeta}(s)$, which is informed solely by the PO data and is intended to reflect spatial variation not explained by either the shared spatial field or the environmental covariates. According to Simmonds et al. (2020), the addition of a second spatial field is useful for accounting for spatial bias in PO data. Therefore, we investigate whether the second spatial field $\hat{\zeta}(s)$ can effectively mitigate this bias.

The models specified in this experiment enable a comparative evaluation of incorporating a covariate for bias and adding a second spatial field to account for spatial bias. This will help to determine whether using the complex models is beneficial over the simple integrated model framework specified by model set A.

5.4 Data Simulations

In our study, each model has been run with every dataset a total of 100 times to ensure robust results. During each time, the presence-absence and presence-only datasets were resampled from the generated virtual species. This resampling process involves creating multiple datasets to mimic real-world scenarios, which allows us to assess how well the models perform under various conditions. Once the models have been built, they are stored for the purpose of calculating performance measures. All simulations were performed using RStudio (R Core Team, 2021).

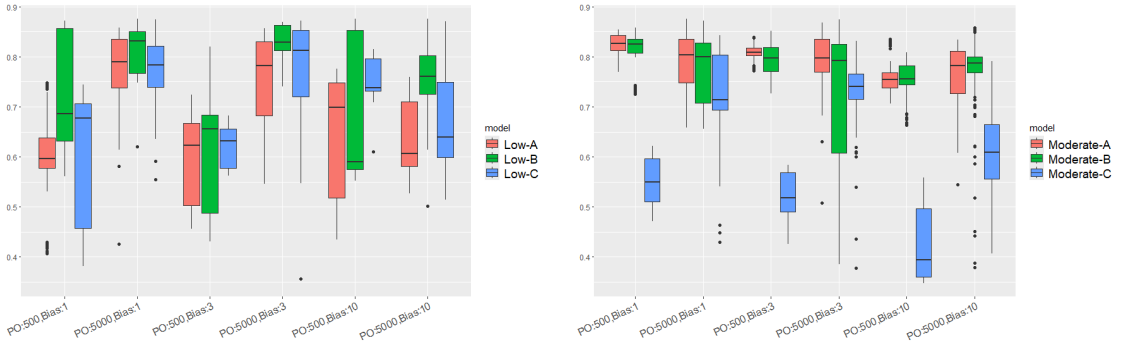
5.5 Results and discussion

5.5.1 Mean absolute error (MAE)

Figure 11a and Figure 11b illustrate the mean absolute error (MAE) for each dataset and model. A smaller MAE indicates lower prediction error, reflecting better model performance. For low-prevalence species (30% probabilities of occurrence), when the dataset contained 500 presence-only data points, the median MAE values of all models are very similar at bias levels 1 and 3. However, at bias level 10, the MAE increased significantly for all models except Model B, with the entire boxplot shifting upward noticeably. In contrast, with 5000 presence-only data points, the MAE was relatively high at bias levels 1 and 3, but at bias level 10, it decreases significantly for all models except Model B.

For Model A, when the dataset includes a small number of presence-only data points, the MAE increases as the bias level rises. However, with a large number of presence-only data points, the MAE instead decreases as bias increases. For Model B, under the same presence-only data quantity, the MAE showed a slight decrease with increasing bias level. Model C follows a pattern similar to Model A, with similar MAE values to Model A across different scenarios.

Overall, low-prevalence species tend to exhibit lower MAE when fewer presence-only data points are included in the dataset.



(a) MAE for Low-prevalence species (b) MAE for Moderate-prevalence species

Figure 11: Mean absolute error (MAE)

In contrast, moderate-prevalence species showed relatively stable MAE across different datasets, regardless of the amount of presence-only data except Model C. Model C shows variation in MAE across different scenarios, with overall lower MAE compared to the other two models, particularly when the dataset contains only 500 presence-only data points. Additionally, the MAE of Model C decreases as the bias level increases.

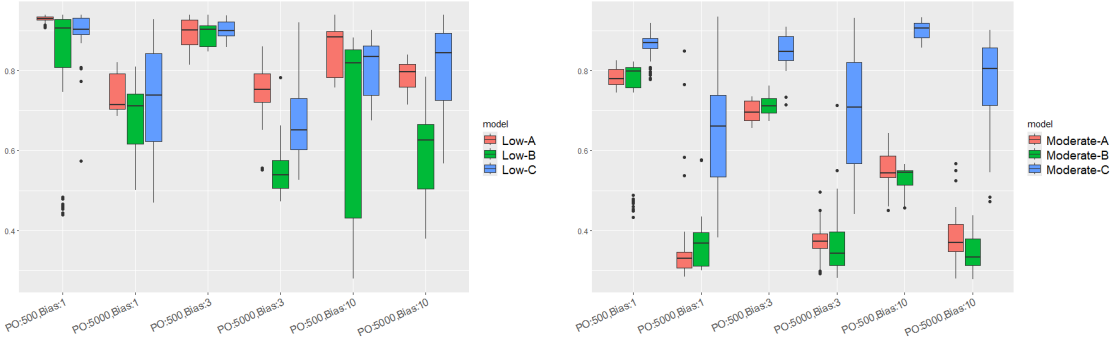
When comparing datasets that contain different amount of presence-only data:

- Lower MAE is observed when the dataset contains less presence-only data. However, this comes with a trade-off in the form of reduced precision, indicating that fewer presence-only records may improve the model's fitting but because data is limited variation may be large.
- As the amount of presence-only data increases, the difference between models diminishes, and performance becomes more uniform, indicating that the added data does not always provide a clear advantage for model performance.

5.5.2 Pearson correlation and Area under the curve (AUC)

Figure 12 shows the Pearson correlation for each dataset across different models, while Figure 13 illustrates the area under the curve (AUC) values. Both figures reveal similar patterns in model performance, allowing for a more comprehensive comparison. A higher correlation indicates a stronger relationship between predicted and observed values, suggesting that the model effectively captures the underlying patterns in the data. Similarly, a higher AUC reflects better model performance, as it demonstrates the model's ability to accurately distinguish between different classes. Together, these metrics highlight the model's overall effectiveness in prediction and classification tasks.

For low-prevalence species shown in Figure 12a, the AUC is generally lower with 5000 PO data points than with 500 PO data points across different bias levels. Model B consistently exhibits the lowest correlation compared to Model A and Model C across all scenarios. Model A shows the highest correlation when there are 500 PO data points and at bias level 1. However, as the number of PO data increases, the correlation for Model A decreases, and it continues to decline as the bias level rises. Model C exhibits similar behavior and correlation values to Model A.



(a) Pearson correlation for Low-prevalence species (b) Pearson correlation for Moderate-prevalence species

Figure 12: Pearson correlation

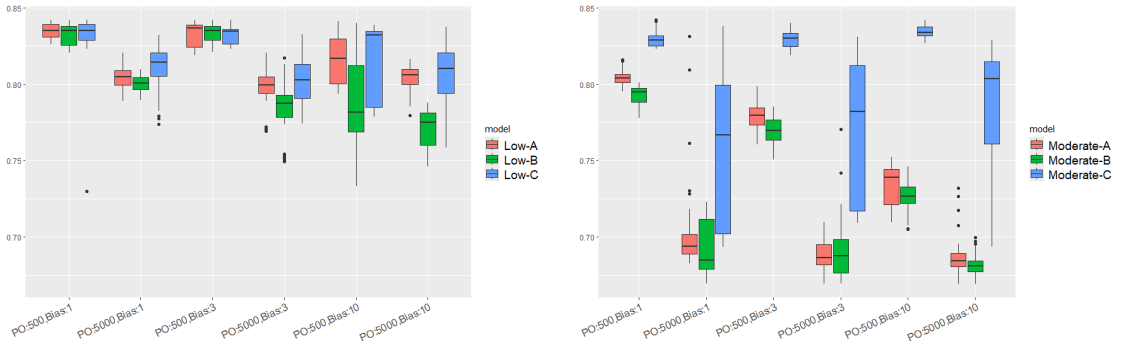
As shown on Figure 12b, when the species have higher prevalence, the differences in correlation

between the models become more pronounced. Model C consistently shows the highest correlation across all scenarios. In contrast, Models A and B have relatively similar correlations in all cases. With 500 PO data points, the correlation for both Model A and Model B decreases as the bias level increases. However, when the PO data increases to 5000, the correlation for both models remains low and relatively stable. Model C, on the other hand, maintains a high and stable correlation with fewer PO data points, and as the number of PO data increases, the correlation rises with higher bias levels. The gap in correlation between Model C and the other models widens as the bias level increases.

When we compare datasets with varying amounts of presence-only data:

- A higher correlation was observed when the dataset contains less presence-only data with low precision except for Model B in the low-prevalence, bias level 10 scenario.
- As the amount of presence-only data increases, the difference between models widens, indicating that the added data provides a clear advantage for model performance for correlation.
- Model C consistently delivered the best performance when the dataset contains a higher number of presence-only records, suggesting that this model handles dense presence-only data more effectively than the others.

Figure 13 and Figure 12 are very similar to each other. For low-prevalence species, the AUC patterns of Models A, B, and C closely mirror their correlation patterns. However, the smaller AUC quantiles suggest that the precision of AUC is higher. For moderate-prevalence species, the patterns are similar to those of low-prevalence species, with the exception that Model C shows a wider AUC quantile range when there are 5000 PO data points. When comparing datasets with varying amounts of presence-only data for AUC, the conclusions are almost identical to those for correlation.

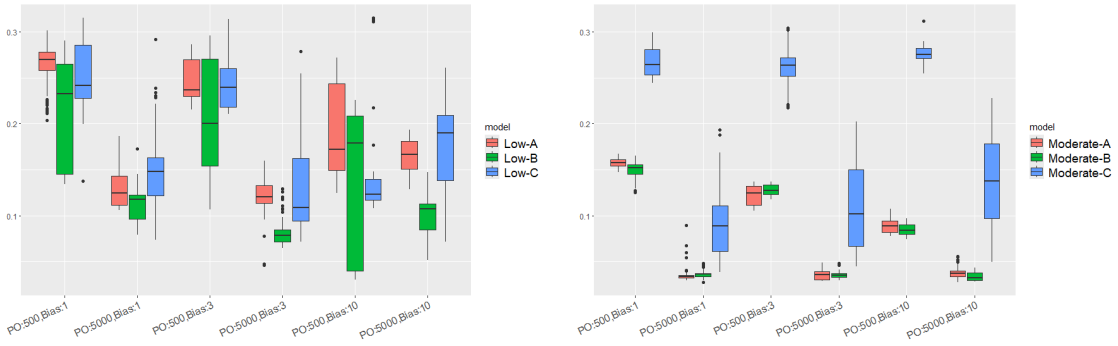


(a) AUC for Low-prevalence species

(b) AUC for Moderate-prevalence species

Figure 13: Area under the curve (AUC)

5.5.3 Tjur R^2



(a) Tjur R^2 for Low-prevalence species

(b) Tjur R^2 for Moderate-prevalence species

Figure 14: Tjur R^2

Figure 14 shows the Tjur R^2 for each dataset across different models. A higher Tjur R^2 favors models with better discrimination ability, indicating stronger predictive performance. For low-prevalence species, the Tjur R^2 values of all three models remain fairly consistent across different scenarios. Unlike AUC on Figure 13a, where Model C generally performs better than the other two, Model C's Tjur R^2 only surpasses the others when the PO data is 5000 and the bias levels are 1 or 10. In scenarios with 500 PO data points, the Tjur R^2 for all models decreases as the bias level increases, whereas with 5000 PO data points, the Tjur R^2 increases with higher bias levels. On the other hand, for moderate-prevalence species, the Tjur R^2 values closely mirror Figure 12b and Figure 13b, with Model C consistently outperforming the other models across all scenarios.

If we compare datasets with varying amounts of presence-only data:

- A higher Tjur R^2 is observed when the dataset contains less presence-only data. However, this comes with a trade-off in the form of reduced precision, indicating that fewer presence-only records may improve the model's fitting but because data is limited variation may be large.
- Model C consistently delivers the best performance when the dataset contains a higher number of presence-only records, confirming our finding from correlation and AUC.

The Tjur R^2 favors models with good discrimination ability. Model C has outstanding Tjur R^2 on moderate-prevalence species indicating its effectiveness in distinguishing between the presence and absence.

5.5.4 Discussion

Model C consistently outperforms models A and B across Pearson correlation, AUC, and Tjur R^2 particularly when dealing with increased bias or increased presence-only data. It is the most adaptable model for moderate-prevalence species, showcasing lower error rates, stronger correlations, and better classification accuracy overall. For low-prevalence species, Model C also performs well, showing comparable MAE, correlation, and Tjur R^2 values while achieving a slightly higher AUC than the other two models. This highlights Model C's consistent performance across different species prevalence levels, often matching or surpassing the other two models.

Presence-only data introduces bias, particularly in moderate-prevalence datasets, negatively impacting models A and B more than Model C. This suggests that including more presence-only data in the integrated modeling setting does not always improve model performance and in fact leads to degraded performance. This could be because a large volume of presence-only data can dilute and overwhelm the presence-absence data, reducing its contribution, which is crucial for the integrated model. As a result, this imbalance can negatively affect the overall model performance.

Low-prevalence species are easier for models to handle, as shown by lower MAE, higher Pearson correlations, and higher AUC values. However, as species prevalence increases, the models, particularly A and B, struggle with increased bias, reflected in higher errors, lower correlations, and decreased AUC.

Interestingly, increasing the amount of unbiased presence-only data (PO:5000 vs. PO:500) did not improve model performance for low-prevalence species. This likely reflects the strong imbalance between presences and absences: adding more data increases the total sample size but not the proportion of presences, causing models to be dominated by absence patterns and perform worse on most metrics.

As bias increases, all models experience some degree of degradation in performance, particularly for moderate-prevalence species. Model C, however, is the least affected by increasing bias, highlighting its robustness. Model C consistently outperforms or matches Model A in all scenarios. Although not quantitatively assessed in the results section, the addition of a second spatial field generally increases computational time. Nevertheless, this trade-off is often worthwhile given the substantial gains in predictive accuracy and bias correction. This advantage is particularly evident as bias increases, clearly showing that the extra computational cost is justified by the accuracy and stability Model C provides.

Model C demonstrates strong performance with an additional random spatial field. In contrast, Model B requires a bias covariate and relies on the user's prior knowledge of the spatial bias mechanism. This highlights Model C's advantage in handling presence-only data, as it effectively accounts for bias through the second spatial field, achieving comparable or even superior performance without requiring predefined bias information. Meanwhile, Model B's performance depends on the accuracy and availability of bias covariates, making it less flexible when such information is limited or uncertain.

In conclusion, Model C appears as the most effective model across all species prevalence levels and bias conditions. Models A and B, while effective for low-prevalence species, show greater sensitivity to bias which limits their broader applicability. The use of large amounts of presence-only data does not always bring advantages to the integrated model. Therefore, implementing a spatial thinning process could be valuable, especially before applying Model C to reduce sampling bias in presence-only data. Reducing the amount of presence-only data may improve Model C's predictive power and solidify its role as the most reliable model across varying prevalence conditions.

6 Real world application and result

According to the results from simulations, we believe thinning data and then applying integrated distribution models (IDM) with a second spatial field have the best performance. We apply this sequential modeling procedure to three different species and compare performance between standard IDM and IDM with a second spatial field. Each species has a unique distribution which allows us to evaluate the model properly. The Martial Eagle is widespread in Southern Africa but uncommon, with a patchy distribution. The Long-crested Eagle has a smaller range but is commonly seen where it lives, while the Peregrine Falcon is widespread over all of Africa and rare.

6.1 Martial Eagle

The Martial Eagle is a large bird of prey primarily found in sub-Saharan Africa. It is known for its powerful build and broad wings with a wingspan reaching up to 2.6 meters and it is one of the largest eagles in Africa. Martial Eagles are exceptional hunters, preying mainly on medium-sized mammals, birds, and reptiles. They typically inhabit open grasslands, savannas, and semi-arid regions, nesting in tall trees (Kemp et al., 2020). Despite being apex predators, their numbers are declining due to habitat loss and human activities. The primary cause of habitat loss is deforestation, particularly the clearing of large trees used for nesting, which are being converted into agricultural fields. Therefore, having knowledge of the Martial Eagle's distribution is essential for guiding conservation policies in those areas.

The data applied in the model are the second Southern African bird atlas project (SABAP2) data collected in South Africa between 2018 and 2020, and eBird data for Martial Eagle collected across Africa continuously up to 2024. The number of observations for each dataset is 7,545 and 22,149 respectively. To find the correct thinning parameter, eBird data is first checked with the inhomogeneous pair correlation function (PCF). Since the observations of the Martial Eagle are primarily concentrated in the southeastern part of Africa, a moderate value of sigma 0.75 was used.

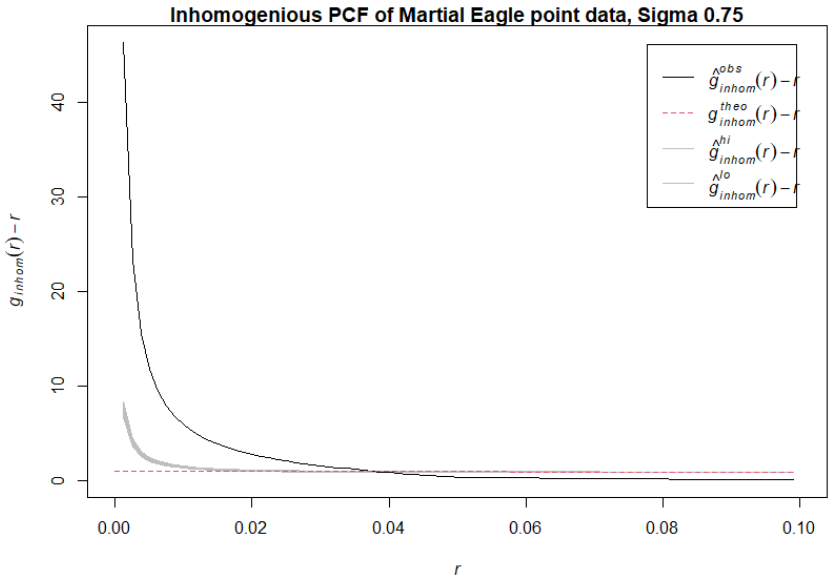


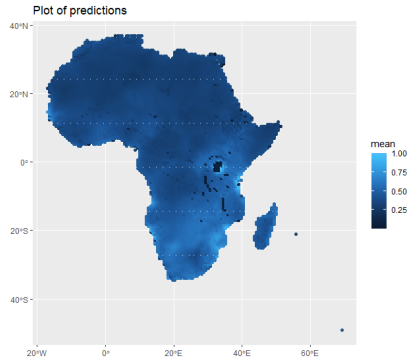
Figure 15: PCF for presence-only data of Martial Eagle. The black curve shows the observed inhomogeneous PCF; the pink dashed curve indicates the theoretical value, and the gray shaded area represents the confidence envelope.

On the Figure 15, the black line represents PCF for the observed data and the red line represents PCF for an inhomogeneous Poisson point process in theory. We observe non-random clustering at short distances and non-random dispersion at larger distances, with the transition occurring at the crossover point, where the spatial patterns shift from clustering to dispersion. PCF suggests a distance of 0.04 units, which is what we would expect if the pattern represented complete spatial randomness.

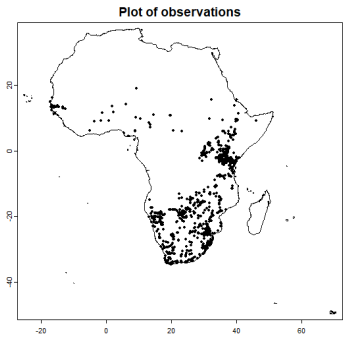
To thin the presence-only data, we used the thinning parameter 0.04 units from PCF, which corresponds to a spatial distance of 4.4 km. This approach was implemented using the `spThin` function from the package spThin (Aiello-Lammens et al., 2015). This function is a commonly used tool in ecological studies for spatial thinning to reduce autocorrelation in presence-only data. The original eBird dataset for the Martial Eagle consisted of 22,149 observations. After applying the thinning process the dataset was reduced to 3,324 observations. On the other hand, SABAP2 has a total of 7,541 observations, of which 601 represent presences and 6,940 represent absences. In this case, presence-absence data is more abundant than presence-only data after thinning. However, this is likely to improve over time, as the availability of citizen science data increases, presence-only data is expected to become more prevalent in the future. Also, One advantage of citizen science data is its ability to cover larger areas compared to structured datasets. In our case, SABAP2 data is limited to southern Africa, whereas eBird provides coverage across the entire African continent and even worldwide.

Secondly, we construct an integrated species distribution model (IDM) with a second spatial random field as described in section 6.2 using the thinned eBird data along with the SABAP2 data. The covariates used in modeling include 19 world climate covariates, water land covers, soil properties, elevation, tree cover, grassland extent, shrubs cover, cropland cover, built cover, and wetland. These variables provide essential context for predicting Martial Eagle's distribution (Kemp et al.,

2020). Multicollinearity between covariates was tested before modeling by calculating the variance inflation factor. High correlation covariates were removed to avoid instability in parameter estimation in model fitting.



(a) Model predictions for the Martial Eagle using the IDM with a second spatial random field



(b) Actual observations for the Martial Eagle

Figure 16: A comparison between prediction and actual observations

Figure 16 illustrated the model predictions for the Martial Eagle, with predictions shown in Figure 16a and actual field observations in Figure 16b. Only presence data from eBird and SABAP2 were shown in Figure 16b, while absences are omitted for clarity of visualization. In the prediction map, lighter shades indicate areas with a higher probability of observing Martial Eagles, while darker shades indicate lower probabilities. This visual representation highlights the model's capability to capture the distribution of this species, particularly in Eastern and Southern Africa, consistent with the observed data in Figure 16b. Notably, the bright areas are surrounded by very dark regions, which appear to result from gaps or missing values in the underlying geospatial data. Despite these local absences, the bright areas themselves remain highly consistent with observed distributions, indicating that the model reliably identifies key hotspots even in regions with incomplete environmental coverage.

The spatial distribution of both predicted and observed Martial Eagle populations suggests that the model effectively reflects real-world patterns. Observations were concentrated in Eastern and Southern Africa and so are areas of high predicted densities hence the predicted surface aligns well with known habitats for the species.

Regarding environmental variables in Table 6, the analysis reveals that "Precipitation of Driest Month", "Precipitation Seasonality", water, and cropland have negative coefficients, indicating these land cover types may be less conducive to the presence of Martial Eagles. Conversely, the variables "Mean Diurnal Range", shrubs, grassland, built environments, and wetlands exhibited positive coefficients, suggesting they are favorable for the species.

	mean	sd
bio2	1.0855	0.0490
bio14	-0.2935	0.0109
bio15	-0.1938	0.0059
water	-7.4064	0.2912
grassland	10.9801	0.3265
shrubs	0.5531	0.3434
cropland	-6.0608	0.8117
built	19.2904	0.6489
wetland	97.6421	1.6649
$PO\_intercept$	-0.064	0.182
$PA\_intercept$	-4.6678	0.2906

Table 6: Estimated coefficients of environmental covariates for Martial Eagle using the IDM with a second spatial random field

The positive association with grassland and wetlands aligns with the ecological preferences of Martial Eagles, as these habitats typically provide essential resources such as prey and suitable nesting sites. The positive coefficient for the "Mean Diurnal Range" implies that temperature fluctuations and weather patterns could also influence habitat suitability. This understanding of variable influence can inform future research and conservation strategies aimed at preserving critical habitats for the Martial Eagle.

The comparison of Watanabe-Akaike Information Criterion (WAIC) values between the standard Integrated Distribution Model (IDM) and the IDM with a second spatial field was conducted using cross-validation over 10 iterations as is shown in Figure 17. WAIC measures model fit, A lower value indicates that the model provides a better fit to the data while maintaining good generalization ability. This evaluation allows us to assess the models' predictive performance systematically.

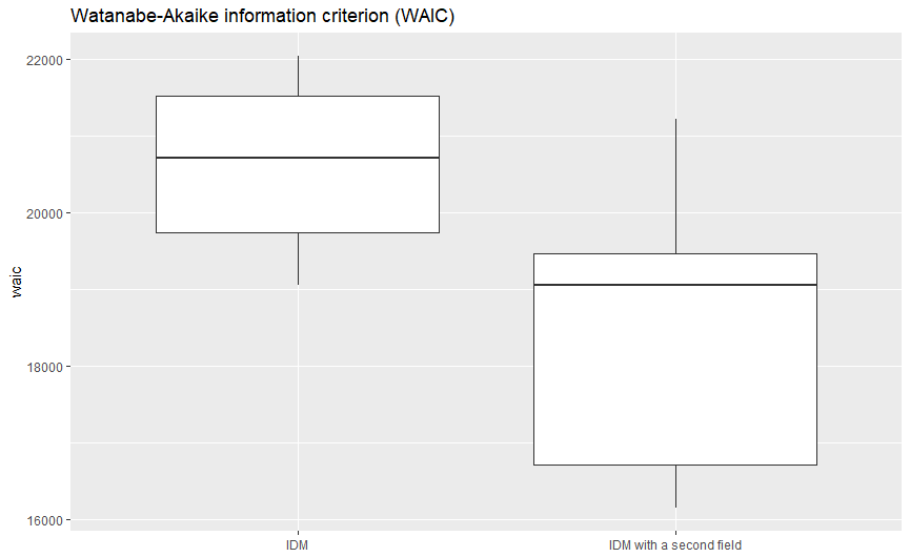


Figure 17: WAIC for standard IDM and IDM with a second spatial field

On average, the results indicate that the IDM incorporating a second spatial field yielded lower WAIC values compared to the standard IDM. This suggests that the inclusion of an additional spatial dimension enhances the model's ability to capture the underlying data structure, leading to better predictive accuracy. The reduced WAIC reflects improved model fit, as it penalizes complexity while rewarding better predictive performance. However, it is essential to note that the IDM with a second spatial field also exhibited lower precision in its estimates. This decrease in precision may stem from the increased complexity of the model, which, while improving overall fit, introduces additional variability in predictions. Overall, the IDM with a second spatial field is advantageous in terms of predictive capability as indicated by lower WAIC.

6.2 Long-crested Eagle

The Long-crested Eagle is native to sub-Saharan Africa, where it is commonly found in open woodlands, forest edges, and savannahs. It primarily inhabits areas with tall trees, using these trees for nesting. The eagle is a skilled hunter, preying on small mammals, birds, and reptiles. The Long-crested Eagle is known for its distinctive long crest on its head and its ability to soar at great heights. Unlike the Martial Eagle, it tends to be more adaptable to different environments, though its range is still somewhat restricted compared to other eagle species. While it shares some distribution similarities with the Martial Eagle, there are notable differences (BirdLife, 2012). The Long-crested Eagle has a relatively limited range in Africa but maintains a considerable population within its habitats.

The observations of the Long-crested Eagle were mainly concentrated in eastern Africa, with sparse distribution in the west. Therefore, a larger sigma was used to account for this broader spread in the observation data. Sigma 1 was used for the PCF.

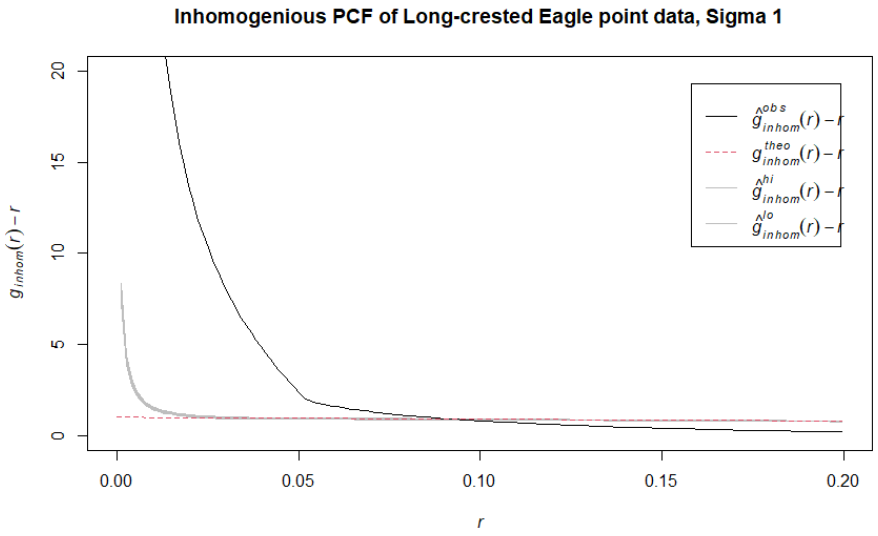
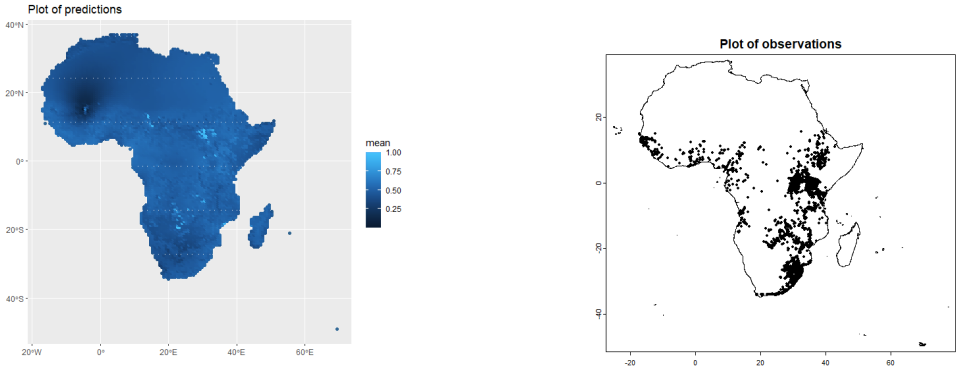


Figure 18: PCF for presence-only data of Long-crested Eagle. The black curve shows the observed inhomogeneous PCF; the pink dashed curve indicates the theoretical value, and the gray shaded area represents the confidence envelope.

On the Figure 18, the black line represents PCF for the observed data and the red line represents PCF for an inhomogeneous Poisson point process in theory. Similar to what we see for Martial Eagle, non-random clustering at short distances and non-random dispersion at greater distances. PCF suggests a distance of 0.09 units, which is what we would expect if the pattern represented complete spatial randomness.

The thinning parameter of 0.09 units from PCF is converted to a spatial distance of 10 km and thinning was also done via `spThin`. The original eBird dataset for the Long-crested Eagle consisted of 29,901 observations. After applying the thinning process the dataset was reduced to 2,582 observations. On the other hand, SABAP2 has a total of 7,541 observations, of which 646 represent presences and 6,895 represent absences.

After that, we performed IDM with the second spatial field on the thinned eBird dataset and SABAP2 dataset.



(a) Model predictions for the Long-crested Eagle using the IDM with a second spatial random field

(b) Actual observations for the Long-crested Eagle

Figure 19: A comparison between prediction and actual observations for the Long-crested Eagle

Figure 19 illustrates the model predictions for the Long-crested Eagle, with predictions shown in Figure 19a and actual field observations in Figure 19b. Only presence data from eBird and SABAP2 were shown in Figure 19b, while absences are omitted for clarity of visualization. In the prediction map, as before, lighter shades represent areas with a higher probability of observing the species, while darker shades indicate areas with lower probabilities. This visual representation highlights the model's ability to capture the species' distribution, particularly in northwestern Africa, aligning well with the gap in observed data shown in Figure 22b. High probabilities on eastern Africa as shown in the plot of real observations.

In Table 7, we have coefficients for environmental variables. Water, shrubs, grassland, and cropland have negative coefficients, indicating these land cover types may be less conducive to the presence of Long-crested Eagles. On the other hand, the variables "Mean Diurnal Range", "Precipitation of Driest Month", built, and wetlands exhibited positive coefficients, suggesting they are favorable for the species. The Long-crested Eagle shares similar environmental preferences with the Martial Eagle, as their observed distributions across Africa are quite similar.

	mean	sd
bio2	0.7949	0.0363
bio14	0.1992	0.0078
water	-16.4476	0.7098
grassland	-1.0200	0.2907
shrubs	-4.6117	0.4280
cropland	-0.5309	0.4400
built	12.2551	0.6654
wetland	25.8694	1.5230
$PO\_intercept$	-6.5179	0.4612
$PA\_intercept$	-10.3477	0.6081

Table 7: Estimated coefficients of environmental covariates for Long-crested Eagle using the IDM with a second spatial random field

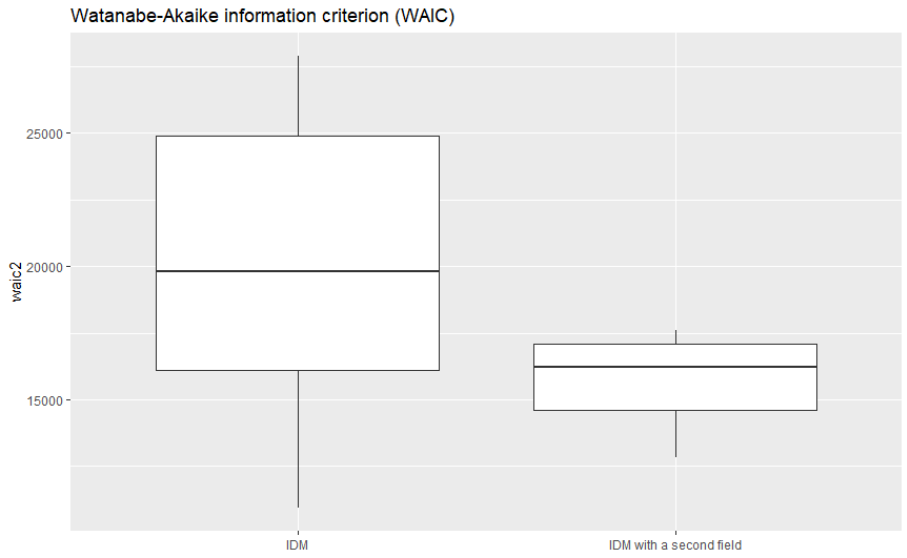


Figure 20: WAIC for standard IDM and IDM with a second spatial field

The comparison of WAIC values between the standard Integrated Distribution Model IDM and the IDM with a second spatial field was conducted using cross-validation over 10 iterations shown on Figure 20. On average, the results indicate that the IDM incorporating a second spatial field yielded slightly lower WAIC values compared to the standard IDM. The IDM with a second spatial field had smaller precision in its estimates compared to the standard IDM indicating it has better performance while having better stability in this case.

6.3 Peregrine Falcon

The Peregrine Falcon is a widely distributed species across Africa, though it is found less frequently in certain regions. This bird of prey inhabits a variety of environments, including coastal cliffs, mountain ranges, and increasingly urban areas. It is renowned for its exceptional speed, capable of reaching over 300 km/h during its characteristic hunting stoop. The Peregrine Falcon primarily preys on medium-sized birds, capturing them mid-air with remarkable precision. Its adaptability to diverse habitats, from remote wilderness areas to bustling cities, underscores its resilience and ecological versatility. The Peregrine Falcon is less frequently observed across the continent and shares some distribution similarities with the Martial Eagle and Long-crested Eagle (BirdLife, 2012).

As with previous analyses, we first examined the eBird data using the inhomogeneous PCF. Due to the more dispersed distribution of Peregrine Falcon observations, a larger sigma value of 1.25 was used in the PCF analysis.

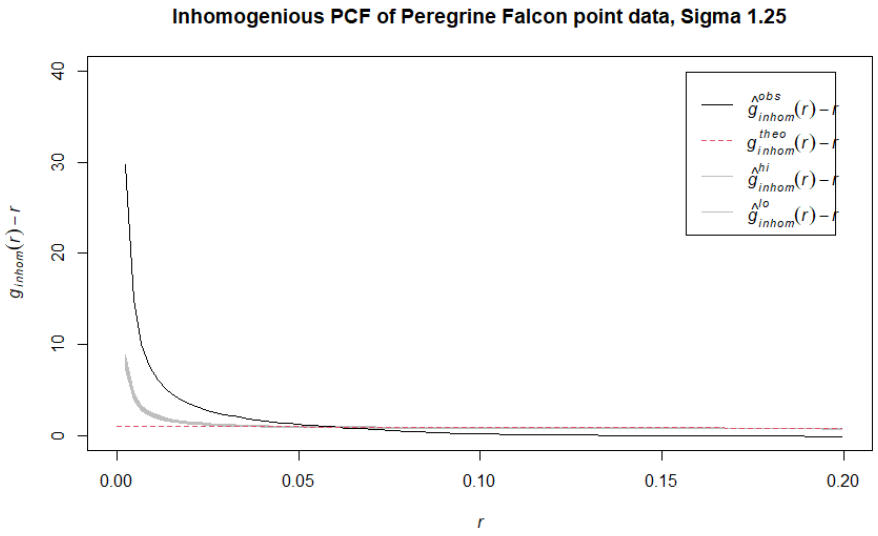
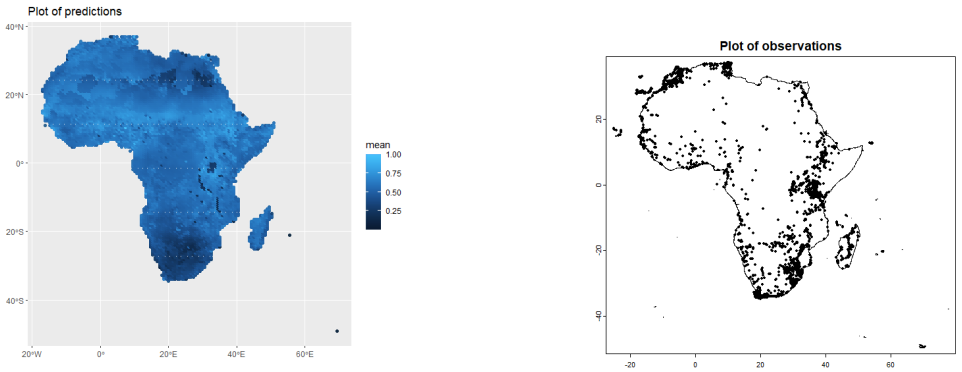


Figure 21: PCF for presence-only data of Peregrine Falcon. The black curve shows the observed inhomogeneous PCF; the pink dashed curve indicates the theoretical value, and the gray shaded area represents the confidence envelope.

In Figure 21, PCF suggests a distance of 0.06 units, which is what we would expect if the pattern represented complete spatial randomness. The thinning parameter 0.06 units from PCF is then converted to a spatial distance of 6.6 km. The original eBird dataset for the Peregrine Falcon consisted of 13,497 observations. After applying the thinning process the dataset was reduced to 2,346 observations. On the other hand, SABAP2 has a total of 7,541 observations, of which 444 represent presences and 7,097 represent absences.



(a) Model predictions for the Peregrine Falcon using the IDM with a second spatial random field

(b) Actual observations for the Peregrine Falcon

Figure 22: A comparison between prediction and actual observations for the Peregrine Falcon

Then, we performed IDM with the second spatial field on the thinned eBird dataset and SABAP2

dataset. Figure 22 illustrates the model predictions for the Peregrine Falcon, with predictions shown in Figure 22a and actual field observations in Figure 22b. Only presence data from eBird and SABAP2 were shown in Figure 22b, while absences are omitted for clarity of visualization. In the prediction map, lighter shades indicate areas with a higher probability of observing Peregrine Falcon, while darker shades indicate lower probabilities. This visual representation highlights the model's capability to capture the distribution of this species, particularly on the edge of Africa, consistent with the observed data in Figure 22b. Low probabilities on northeastern Africa as shown in the plot of real observations. Apart from the southeastern part of South Africa, discrepancies existed between the actual and predicted data with predictions indicating low probability while real observations show a high abundance in this area. Similar to the Martial Eagle map, there are dark regions around eastern Africa, which appear to result from gaps or missing values in the underlying geospatial data. The area surrounding these dark patches is relatively bright and corresponds well to the clusters of presences in the observation map, indicating that the model still captures the main concentration of occurrences despite the missing geodata.

In Table 8, the variables trees, grassland, cropland, built environments, and wetlands exhibited positive coefficients, suggesting they are favorable for Peregrine Falcon. Water and shrubs have negative coefficients, indicating these land cover types and climate variables may be less conducive to the presence of the species. The positive coefficient on built environments confirms its tendency to inhabit urban areas. Unlike the Martial Eagle and Long-crested Eagle, the Peregrine Falcon's distribution is not affected by world climate covariates. These habitat differences can be used to distinguish their distributions and design different conservation measures.

	mean	sd
water	-6.6079	0.2960
trees	6.6495	0.3163
grassland	3.2512	0.2726
shrubs	-2.9819	0.3497
cropland	5.3070	0.3101
built	13.8353	0.4817
wetland	20.1287	1.6472
$PO\_intercept$	-3.4937	0.1784
$PA\_intercept$	-8.5426	0.2732

Table 8: Estimated coefficients of environmental covariates for Peregrine Falcon using the IDM with a second spatial random field

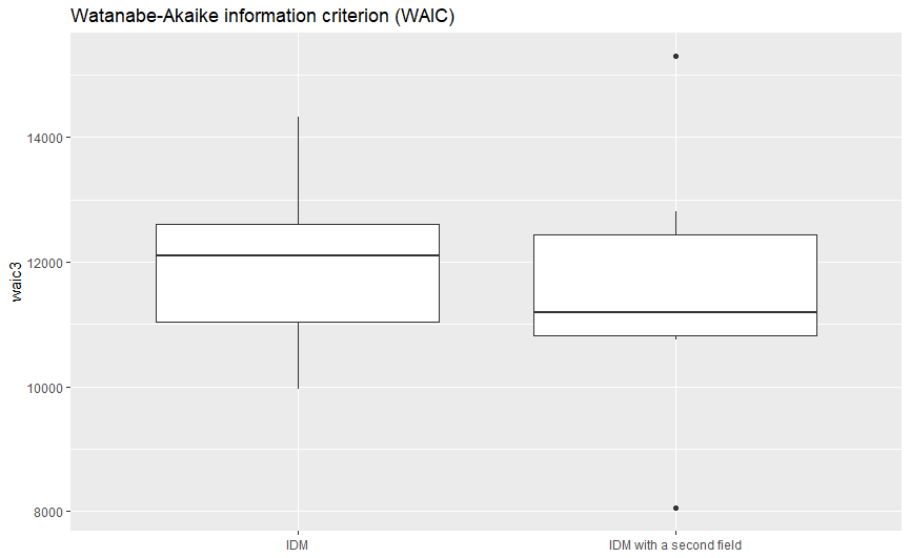


Figure 23: WAIC for standard IDM and IDM with a second spatial field

In Figure 23, the results indicate that the IDM incorporating a second spatial field yields similar but slightly lower WAIC values compared to the standard IDM. The IDM with a second spatial field has similar precision in its estimates compared to the standard IDM. In this case, the WAIC for the IDM with a second spatial field is slightly lower than that of the standard IDM, though the difference is not significant. However, for the other two species, the WAIC results clearly demonstrate the advantages of the IDM with a second spatial field.

7 Conclusion

In this dissertation, we set out to identify the best modeling approach for constructing an integrated species distribution model using multiple datasets, specifically presence-absence (PA) and presence-only (PO) data. We found that the integrated distribution model (IDM) with joint likelihoods has various advantages over other data-combining methods. We used the approach to evaluate variations of the IDM model with joint-likelihoods via a simulation study. Virtual species for the simulation were generated via the Virtualspecies package (Leroy et al., 2015). The package was also employed to generate samples of PA and PO datasets. The virtual data were then used to investigate the predictive ability of integrated distribution models (IDM) under a range of scenarios. The scenarios were created by varying the species prevalence rates between 30% and 60%, varying the PO data size to be either 500 or 5000, and also varying the bias levels in samples of PO data to be either 1, 3, or 10. Three IDMs were tested: the standard IDM, the IDM with bias covariates, and the IDM with a second spatial field.

Among the models investigated, the model incorporating a second spatial field consistently outperformed its alternatives across the scenarios. The standard IDM and IDM with bias covariate, while effective for low-prevalence species, show greater sensitivity to bias, which limits their broader applicability and leads to a decline in performance as bias levels increase. In contrast, the model with a second spatial field remained unaffected by bias changes. Furthermore, when combining multiple datasets, we discovered that an increase in unstructured data does not always bring advantages to the model, and in some cases, it can even degrade performance. Therefore we applied spatial thinning as a tool to remove bias from data to improve model performance more. Also, we can recommend a pair correlation function to determine the thinning parameter. In the Martial Eagle case, the effective parameter for spatial thinning was 4.4km, but this depended fully on the clustering and bias present in the dataset. PCF will help the user find a thinning parameter without random clustering and random dispersion. When performing the PCF, different sigma values should be selected based on the general degree of dispersion or clustering in the data.

In this project, three bird species' distributions were modeled using the integrated distribution model. Each species exhibited distinct distribution patterns and varying levels of rarity, reflecting differences in habitat preferences and population density. On average, the model with a second spatial field performed better than the standard model for all three species and most of them had better precision.

The virtual ecologist approach was used to measure model performance before applying it to real-world data. The virtual ecologist approach overcomes the challenge of no actual species distribution. It allows us to make a direct comparison between model prediction and true species distribution, enabling a more rigorous evaluation of our modeling techniques. Mean absolute error is the main performance metric used for the virtual ecologist approach. It provides a clear indication of how closely our predictions align with the true distribution. On the other hand, Watanabe-Akaike Information Criterion (WAIC) was used to measure the performance of real-world data models. These performance measures effectively evaluated the quality of the model and provided valuable guidance for model selection.

7.1 Limitations and future research opportunities

In this dissertation, our final recommendation is to first apply spatial thinning and then proceed with integration modeling with a second spatial field. However, this sequential approach may not be the most efficient or convenient for complex modeling tasks. A key limitation lies in the

lack of a simultaneous optimization method that combines both spatial thinning and spatial field integration.

Moreover, determining an optimal thinning parameter and using Integrated Nested Laplace Approximation (INLA) for large datasets can significantly increase computational time. Future research could explore ways to streamline this process, developing a more simultaneous approach to improve model efficiency and potentially enhance accuracy.

Also, we focused on integrated modeling of presence-only data and presence-absence data. Future research could expand on this by integrating additional data types, such as count data, and other species data. Incorporating these diverse data sources would enhance the robustness of species distribution models and provide a more comprehensive understanding of species patterns and drivers across various ecological contexts.

8 References

Ahmad Suhaimi, S. S., Blair, G. S., and Jarvis, S. G. (2021). Integrated species distribution models: A comparison of approaches under different data quality scenarios. *Diversity and Distributions*, 27(6):1066–1075.

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., and Anderson, R. P. (2015). Spthin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, 38(5):541–545.

Baddeley, A., Rubak, E., and Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R*. Chapman and Hall/CRC Press, London.

Bakka, H., Rue, H., Fuglstad, G.-A., Riebler, A. I., Bolin, D., Illian, J., Krainski, E., Simpson, D. P., and Lindgren, F. K. (2018). Spatial modelling with INLA: A review. *WIRES (Invited extended review)*.

Bevan, A. and Lake, M. (2022). *Computational approaches to archaeological spaces*. Routledge.

BirdLife (2012). BirdLife South Africa.

Brooks, M., Rose, S., Altwegg, R., Lee, A. T., Nel, H., Ottosson, U., Retief, E., Reynolds, C., Ryan, P. G., Shema, S., and et al. (2022). The african bird atlas project: A description of the project and birdmap data-collection protocol. *Ostrich*, 93(4):223–232.

Cervantes, F., Altwegg, R., Strobbe, F., Skowno, A., Visser, V., Brooks, M., Stojanov, Y., Harebottle, D. M., and Job, N. (2023). Birdie: A data pipeline to inform wetland and waterbird conservation at multiple scales. *Frontiers in Ecology and Evolution*, 11.

Diggle, P. J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns*.

Drake, J. M. and Richards, R. L. (2018). Estimating environmental suitability. *Ecosphere*, 9(9).

Ellison, A. M. (1996). An introduction to bayesian inference for ecological research and environmental decision-making. *Ecological Applications*, 6(4):1036–1046.

Fick, S. E. and Hijmans, R. J. (2017). Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12):4302–4315.

Fletcher, R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., and Dorazio, R. M. (2019). A practical guide for combining data to model species distributions. *Ecology*, 100(6).

Fletcher, R. J., McCleery, R. A., Greene, D. U., and Tye, C. A. (2015). Integrated models that unite local and regional data reveal larger-scale environmental relationships and improve predictions of species distributions. *Landscape Ecology*, 31(6):1369–1382.

Grenouillet, G., Buisson, L., Casajus, N., and Lek, S. (2011). Ensemble modelling of species distribution: The effects of geographical and environmental ranges. *Ecography*, 34(1):9–17.

Hamner, B. and Frasco, M. (2018). *Metrics: Evaluation Metrics for Machine Learning*. R package version 0.1.4.

Hijmans, R. J. (2023). *raster: Geographic Data Analysis and Modeling*. R package version 3.6-26.

Hijmans, R. J. (2024). *terra: Spatial Data Analysis*. R package version 1.7-71.

Hijmans, R. J., Barbosa, M., Ghosh, A., and Mandel, A. (2023). *geodata: Download Geographic Data*. R package version 0.5-9.

Hirzel, A., Helfer, V., and Metral, F. (2001). Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, 145(2):111–121.

Hufkens, K. (2023). The modistools package: an interface to the modis land products subsets web services.

Hutchinson, R. A., Valente, J. J., Emerson, S. C., Betts, M. G., and Dietterich, T. G. (2015). Penalized likelihood methods improve parameter estimates in occupancy models. *Methods in Ecology and Evolution*, 6(8):949–959.

Inman, R., Franklin, J., Esque, T., and Nussear, K. (2021). Comparing sample bias correction methods for species distribution modeling using virtual species. *Ecosphere*, 12(3).

Isaac, N. J., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman, S. N., Golding, N., Guillera-Arroita, G., Henrys, P. A., and et al. (2020). Data integration for large-scale models of species distributions. *Trends in Ecology amp; Evolution*, 35(1):56–67.

IWC (2024). International waterbird census.

Jung, M. (2023). An integrated species distribution modelling framework for heterogeneous biodiversity data. *Ecological Informatics*, 76:102127.

Kemp, A. C., Boesman, P. F., and Marks, J. S. (2020). Martial eagle (polemaetus bellicosus). *Birds of the World*.

Komori, O., Eguchi, S., Saigusa, Y., Kusumoto, B., and Kubota, Y. (2020). Sampling bias correction in species distribution models by quasi-linear poisson point process. *Ecological Informatics*, 55:101015.

Lawson, C. R., Hodgson, J. A., Wilson, R. J., and Richards, S. A. (2013). Prevalence, thresholds and the performance of presence–absence models. *Methods in Ecology and Evolution*, 5(1):54–64.

Leroy, B., Meynard, C. N., Bellard, C., and Courchamp, F. (2015). Virtualspecies, an R package to generate virtual species distributions. *Ecography*, 39(6):599–607.

MacKenzie, D. I. (2005). What are the issues with presence–absence data for wildlife managers? *Journal of Wildlife Management*, 69(3):849–860.

Miller, D. A., Pacifici, K., Sanderlin, J. S., and Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, 10(1):22–37.

Mostert, S, P., O'Hara, and B, R. (2023). PointedSDMs: An R package to help facilitate the construction of integrated species distribution models. *Methods in Ecology and Evolution*, 14(5):1200–1207.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log gaussian cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482.

Naimi, B., a.s. Hamm, N., Groen, T. A., Skidmore, A. K., and Toxopeus, A. G. (2014). Where is positional uncertainty a problem for species distribution modelling. *Ecography*, 37:191–203.

Nephin, J., Thompson, P. L., Anderson, S. C., Park, A. E., Rooper, C. N., Aulthouse, B., and Watson, J. (2023). Integrating disparate survey data in species distribution models demonstrate the need for robust model evaluation. *Canadian Journal of Fisheries and Aquatic Sciences*, 80(12):1869–1889.

Niemelä, J., Breuste, J., Elmqvist, T., Guntenspergen, G., James, P., and McIntyre, N. (2011). Introduction. *Urban Ecology*, page 1–4.

Pacifici, K., Reich, B. J., Miller, D. A., Gardner, B., Stauffer, G., Singh, S., McKerrow, A., and Collazo, J. A. (2017). Integrating multiple data sources in species distribution modeling: A framework for data fusion*. *Ecology*, 98(3):840–850.

Palaoro, A. V., Dalosto, M. M., Costa, G. C., and Santos, S. (2013). Niche conservatism and the potential for the crayfishprocambarus clarkiito invadesouthamerica. *Freshwater Biology*, 58(7):1379–1391.

Paradinas, I., Illian, J. B., Alonso-Fernändez, A., Pennino, M. G., and Smout, S. (2023). Combining fishery data through integrated species distribution models. *ICES Journal of Marine Science*, 80(10):2579–2590.

Pebesma, E. and Bivand, R. (2023). *Spatial Data Science: With applications in R*. Chapman and Hall/CRC, London.

Pebesma, E. J. and Bivand, R. (2005). Classes and methods for spatial data in R. *R News*, 5(2):9–13.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rowlingson, B. and Diggle, P. (2023). *splancs: Spatial and Space-Time Point Pattern Analysis*. R package version 2.01-44.

Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J., and O'Hara, R. B. (2020). Is more data always better? a simulation study of benefits and limitations of integrated distribution models. *Ecography*, 43(10):1413–1422.

Smith, D. A. (2020). World population density map.

South, A. (2011). rworldmap: A new R package for mapping global data. *The R Journal*, 3(1):35–43.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(4):583–639.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., and Kelling, S. (2009). Ebird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10):2282–2292.

Trainor, A. M. and Schmitz, O. J. (2014). Infusing considerations of trophic dependencies into species distribution modelling. *Ecology Letters*, 17(12):1507–1517.

Vehtari, A., Gelman, A., and Gabry, J. (2016). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D. (2023a). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.4.

Wickham, H., Pedersen, T. L., and Seidel, D. (2023b). *scales: Scale Functions for Visualization*. R package version 1.3.0.

Wiegand, T. and Moloney, K. A. (2013). *Handbook of Spatial Point-pattern analysis in ecology*.

Wieland, T. (2019). REAT: A Regional Economic Analysis Toolbox for R. *REGION*, 6(3):R1–R57.

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., Farooq, H., Herdean, A., Ariza, M., Scharn, R., Svanteson, S., Wengstrom, N., Zizka, V., and Antonelli, A. (2019). Coordinatecleaner: standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, (10):–7. R package version 3.0.1.

Zurell, D., Berger, U., Cabral, J. S., Jeltsch, F., Meynard, C. N., Münkemüller, T., Nehrbass, N., Pagel, J., Reineking, B., Schröder, B., and et al. (2010). The virtual ecologist approach: Simulating data and observers. *Oikos*, 119(4):622–635.

9 Appendix

9.1 R code: Simulation

```
library(raster)
library(geodata)
library(sp)
library(rworldmap)
library(geodata)
library(PointedSDMs)
library(ggplot2)
library(terra)
library(splancs)
library(INLA)
library(virtualspecies)
library(scales)
library(Metrics)
library(stars)
library(CoordinateCleaner)
library(spatstat)
library(REAT)

#download and prepare world climate variables
worldclim <- worldclim_global(var =  bio , res = 10,path=tempdir())
worldclim = raster::stack(worldclim)
names(worldclim)=c( bio1 ,  bio2 , bio3 , bio4 , bio5 , bio6 , bio7 , bio8 , bio9 ,
                    bio10 , bio11 , bio12 , bio13 , bio14 , bio15 , bio16 , bio17 ,
                    bio18 , bio19 )
pop <- population(2020, 10, path=tempdir())
pop = raster(pop)

#create species habitat
my.parameters <- formatFunctions ( bio1 = c( fun = 'linearFun', a = 2 , b =
                                               0.05) ,
                                   bio5 = c( fun = 'linearFun', a = 1 , b =
                                               0.1) ,
                                   bio6 = c( fun = 'linearFun', a = -1.2 , b =
                                               1.5) ,
                                   bio12 = c( fun = 'dnorm', mean = 700 , sd =
                                                3000) )

#generate virtual species
species <- generateSpFromFun(raster.stack = worldclim[[c( bio1 ,  bio5 ,  bio6 ,  bio12 )]],
                                     parameters = my.parameters,
                                     plot = TRUE)
#create species with 0.3 prevalence
species_0.3 <- convertToPA ( species,
                             PA.method =  probability ,
                             species.prevalence = 0.3 ,
                             plot = TRUE )
#create species with 0.6 prevalence
species_0.6 <- convertToPA ( species,
                             PA.method =  probability ,
                             species.prevalence = 0.6 ,
                              plot = TRUE )

bru_options_set(inla.mode =  experimental )

#download African map
countries <- geodata::world(resolution = 5, path =  maps )
cntry_codes <- country_codes()
countries <- merge(countries, cntry_codes, by.x =  GID_0 , by.y =  ISO3 , all.x = TRUE)
continents <- aggregate(countries, by =  continent )
africa.bdry <-  as(continents[1],  Spatial ) %>% inla.sp2segment()


#Create polygons to introduce sampling bias in presence only data
x_coords <- c(28,26.5,28.5,26,27.5,33,30)
y_coords <- c(-25,-27.5,-29,-32,-33,-25.5,-22)
poly1 <- sp::Polygon(cbind(x_coords,y_coords))

x_coords <- c(18,19,19,17.5)
```

```
y_coords <- c(-33.5,-33.5,-34,-34.5)
poly2 <- sp::Polygon(cbind(x_coords,y_coords))

x_coords <- c(-13,15,15,10,0,-13)
y_coords <- c(15,15,0,0,5,5)
poly3 <- sp::Polygon(cbind(x_coords,y_coords))

x_coords <- c(30,32,32,30)
y_coords <- c(32,32,27,27)
poly4 <- sp::Polygon(cbind(x_coords,y_coords))

x_coords <- c(28,36,36,28)
y_coords <- c(5,5,-5,-5)
poly5 <- sp::Polygon(cbind(x_coords,y_coords))

Srs1 = Polygons(list(poly1),  s1 )
Srs2 = Polygons(list(poly2),  s2 )
Srs3 = Polygons(list(poly3),  s3 )
Srs4 = Polygons(list(poly4),  s4 )
Srs5 = Polygons(list(poly5),  s5 )

SpP = SpatialPolygons(list(Srs1,Srs2,Srs3,Srs4,Srs5), 1:5)


#function to do integration
intgration = function(species,bias,sample_size,addbias,addcov)
{

#Sampling of presence only data from the species
presence.points <- sampleOccurrences(species,
                                     n = sample_size,
                                     type =  presence only ,
                                     bias =  polygon ,
                                     bias.strength = bias,
                                     bias.area = SpP,
                                     replacement = T ,
                                     sampling.area =  Africa ,
                                     plot=F)


# Sampling of presence-absence data from the species
PA.points <- sampleOccurrences(species,
                               n = 200,
                               type =  presence-absence ,
                               replacement = T ,
                               sampling.area =  Africa ,
                               plot=F)

#prepare the data
PA <- PA.points$sample.points
PO <- presence.points$sample.points
PA=PA[,c(1,2,4)]
rownames(PA)=1:nrow(PA)
colnames(PA)=c( X   ,        Y    ,      Present )
PO=PO[!is.na(PO[,4]),c(1,2,4)]
rownames(PO)=1:nrow(PO)
colnames(PO)=c( X   ,        Y    ,      Observed )
speciesdata = list(PO=PO,PA=PA)

max.edge= diff(range(speciesdata$PO[,1]))/(3*5)
bound.outer = diff(range(speciesdata$PO[,1]))/3

#add additional bias covariate
if(addcov==T)
{
  NPP <- raster::stack(worldclim[[c( bio1 ,  bio5 ,  bio6 ,  bio12 )]],pop)
}
else
{
  NPP <- raster::stack(worldclim[[c( bio1 ,  bio5 ,  bio6 ,  bio12 )]])
}

#create a 2d mesh for the model
```

```
max.edge= diff(range(speciesdata$PO[,1]))/(3*5)
bound.outer = diff(range(speciesdata$PO[,1]))/3
mesh <- inla.mesh.2d(boundary = africa.bdry,
                     max.edge = c(1,2)*max.edge,
                     offset=c(max.edge, bound.outer),
                     cutoff = max.edge/5,
                     crs=st_crs(NPP))
projection <-  +proj=longlat +datum=WGS84 +no_defs

#set up the model
model <- intModel(speciesdata, spatialCovariates = as(NPP,  SpatRaster ), Coordinates = c('X', 'Y'),
                  Projection = projection, Mesh = mesh, responsePA = 'Present')

#add a second spatial field
if(addbias== TRUE)
{
  model$addBias( PO )
}

#perform the integrated model
modelRun <- fitISDM(model, options = list(control.inla = list(int.strategy = 'eb'),
                                          safe = TRUE))
return(modelRun)
}

#function to calculate all performance measures
pom=function(result)
{
  #prepare predictor variables
  NPP <- raster::stack(worldclim[[c( bio1 ,  bio5 ,  bio6 ,  bio12 )]],pop)
  #create 2d mesh
  mesh <- inla.mesh.2d(boundary = africa.bdry,
                       max.edge = 2,
                       offset=c(2,2),
                       cutoff = 0.1,
                       crs=st_crs(NPP))
  #make prediction based on the model
  predictions <- predict(result, mesh = mesh,
                         fun = 'linear',
                       predictor=T)
  #sampling validation data point
  PA.points <- sampleOccurrences(species_0.6,
                                 n = 10000,
                                 type =  presence-absence ,
                                 replacement = T ,
                                 sampling.area =  Africa ,
                                 plot=F)

#extract and prepare data from the prediction and validation data point
rp<-st_rasterize(predictions$predictions %>% dplyr::select(mean, geometry))
rp$mean=scales::rescale(rp$mean,to=c(0.0000001,1))

aaaa= as.matrix(PA.points$sample.points[,1:2])
aaaaaa=as(species_0.6$suitab.raster, SpatRaster )
sr=as(rp, SpatRaster )
train <- terra::extract(sr, vect(aaaa))
vali <- terra::extract(aaaaaa, vect(aaaa))

train2=train[which(!is.na(train[,2])),2]
vali=vali[which(!is.na(train[,2])),2]
PA.points=PA.points$sample.points[which(!is.na(train[,2])),]

#AUC
auc1=NA
aucs=seq(0,1,length=100)
for(i in 1:100)
{
  ccc=train2>aucs[i]
  auc1[i]=Metrics::auc(PA.points$Real,ccc)
}
auc=max(auc1)

```

```
#TSS
ccc=train2>(aucs[which(auc1==max(auc1))][1])
a=table(ccc,PA.points$Real)[4]
b=table(ccc,PA.points$Real)[3]
c=table(ccc,PA.points$Real)[2]
d=table(ccc,PA.points$Real)[1]
TSS=(a*d-b*c)/((a+c)*(b+d))

#MAE
mae= mean(abs(log(train2)-log(vali)))


#correlation
cor= cor(log(train2),log(vali))

#Tjur R2
tr=abs(mean(train2[which(PA.points$Real==0)])-mean(train2[which(PA.points$Real==1)]))
return(c(auc,TSS,mae,cor,tr))
}
```

9.2 R code: Real world application

```
library(raster, quietly = T)
library(sp)
library(rworldmap, quiet = TRUE)
library(PointedSDMs)
library(ggplot2)
library(terra)
library(splancs)
library(INLA)
library(virtualspecies)
library(scales)
library(Metrics)
library(stars)
library(CoordinateCleaner)
library(REAT)
library(spatstat)
library(spThin)
library(dplyr)
library(geodata)
library(ABAP)
library(usdm)

#download world climate variables
worldclim <- worldclim_global(var =  bio , res = 5,path= D:/2024/trees/ )

names(worldclim)=c( bio1 ,  bio2 , bio3 , bio4 , bio5 , bio6 , bio7 , bio8 , bio9 ,
                    bio10 , bio11 , bio12 , bio13 , bio14 , bio15 , bio16 , bio17 ,
                    bio18 , bio19 )

climebird = raster::stack(worldclim)
ext <- extent( c(-18,52,-35,38))
xxx <- crop(climebird,ext)
#download water land cover variables
water=geodata::landcover(var='water',res=5,path= D:/2024/trees/ )
water= water %>% raster()
water <- crop(water,ext2)
water <- aggregate(water, fact=10)
xxx<-addLayer(xxx,water)
#download soli land cover variables
soil=geodata::soil_world(var= phh2o ,depth=5,path= D:/2024/trees/ )
soil=soil %>% raster()
soil <- crop(soil,ext2)
soil <- aggregate(soil, 10)
xxx<-addLayer(xxx,soil)
#download elevation variables
elev=geodata::elevation_global(res=5,path= D:/2024/trees/ )
elev <- crop(elev,ext2)
elev= elev %>% raster()
xxx<-addLayer(xxx,elev)
#download trees land cover variables
```

```
tree=geodata::landcover(var='trees',res=5,path= D:/2024/trees/ )
tree= tree %>% raster()
tree <- crop(tree,ext2)
tree <- aggregate(tree, fact=10)
xxx<-addLayer(xxx,tree)
#download grass land cover variables
grass=geodata::landcover(var='grassland',res=5,path= D:/2024/trees/ )
grass= grass %>% raster()
grass <- crop(grass,ext2)
grass <- aggregate(grass, fact=10)
xxx<-addLayer(xxx,grass)
#download shrub land cover variables
shrub=geodata::landcover(var='shrubs',res=5, D:/2024/trees/ )
shrub= shrub %>% raster()
shrub <- crop(shrub,ext2)
shrub <- aggregate(shrub, fact=10)
xxx<-addLayer(xxx,shrub)
#download cropland cover variables
crops=geodata::landcover(var='cropland',res=5,path= D:/2024/trees/ )
crops= crops %>% raster()
crops <- crop(crops,ext2)
crops <- aggregate(crops, fact=10)
xxx<-addLayer(xxx,crops)
#download built land cover variables
built=geodata::landcover(var='built',res=5,path= D:/2024/trees/ )
built= built %>% raster()
built <- crop(built,ext2)
built <- aggregate(built, fact=10)
xxx<-addLayer(xxx,built)
#download wetland cover variables
wetland=geodata::landcover(var='wetland',res=5,path= D:/2024/trees/ )
wetland= wetland %>% raster()
wetland <- crop(wetland,ext2)
wetland <- aggregate(wetland, fact=10)
xxx<-addLayer(xxx,wetland)
#check multicollinearity problems between variables
cortest <-vifcor(as(xxx,  SpatRaster ),0.6)
xxx <- exclude(xxx,cortest)

#download african map
countries <- geodata::world(resolution = 1, path= maps )
cntry_codes <- country_codes()
countries <- merge(countries, cntry_codes, by.x =  GID_0 , by.y =  ISO3 , all.x = TRUE)
continents <- aggregate(countries, by =  continent )
africa.bdry <-  as(continents[1],  Spatial ) %>% inla.sp2segment()

#read ebird raw data
me_raw=read.delim( D:/2024/martial eagle/occurrence.txt )
#clean and prepare data
me_raw$collectionCode[which(me_raw$collectionCode==  )]= GRIN
me_raw$collectionCode[which(me_raw$collectionCode== NBDB RECORDS: 1900s-2000s )]= NBDB
recs_raw <- me_raw[,c(  decimalLongitude , decimalLatitude , occurrenceStatus , species ,
     collectionCode , continent )]
recs_raw=subset(recs_raw, !is.na(decimalLongitude))
recs_raw=subset(recs_raw, !is.na(decimalLatitude))
cl_recsebird <- clean_coordinates(recs_raw, lon= decimalLongitude ,
                                  lat= decimalLatitude ,tests=c( centroids , outliers ))
recs_raw <- recs_raw[cl_recsebird$.summary,]
alldata<-recs_raw
alldata=cbind(1:nrow(alldata),alldata)
colnames(alldata)[1]= location
ebird_data=alldata[alldata$collectionCode!= GRIN &alldata$collectionCode!= NBDB ,]

#Define ppp object on observed location
p=as.ppp(ebird_data[,2:3],W=c(-18,52,-35,40))
#calculate and plot inhomogeneous PCF
ep=envelope(p, pcfinhom, sigma = 1.25, correction =  Ripley , verbose = F)
plot(ep,xlim=c(0,0.2),ylim=c(0,40),. - r ~ r, main =  Inhomogenious PCF of Peregrine Falcon
     point data, Sigma 1.25 )
#find the correct thinning parameter
ep$r[which((ep$obs-ep$theo)==min(abs(ep$obs-ep$theo)))]

#thin data by 6.6km
```

```
thin_data<-thin(ebird_data,lat.col= decimalLatitude ,long.col= decimalLongitude ,
                spec.col= species ,thin.par =0.06*111, reps=1,out.dir= D:/2024/thin/Peregrine falcon )
#read thinned data
thin_data = read.csv( D:/2024/thin/thinned_data_thin1_new_new.csv )

#check Martial Eagle code
eagles <- searchAbapSpecies( Eagle )
eagles[eagles$Common_species== Martial , Spp ]
#download SABAP2 data for Martial eagle in south africa
martial.eagle.compl <- getAbapData(142, .region_type =  country , .region = c( South Africa ),
                                   .years = c(2017,2018,2019))
#clean and prepare data
decimalLatitude <-((as.numeric(stringr::str_sub(martial.eagle.compl$CardNo, start =1, end =4)))+2.5)*(-1/100)
decimalLongitude <-((as.numeric(stringr::str_sub(martial.eagle.compl$CardNo, start =6, end =9)))+2.5)*(1/100)

martial.eagle.compl <- cbind.data.frame(martial.eagle.compl,decimalLatitude,decimalLongitude)
martialEagle_SABAP_compl <-martial.eagle.compl[,c(  CardNo ,  decimalLatitude , decimalLongitude , TotalHours ,  Spp )]

martialEagle_SABAP_compl[ Spp ][martialEagle_SABAP_compl[ Spp ] ==  142 ] <- 1
martialEagle_SABAP_compl[ Spp ][martialEagle_SABAP_compl[ Spp ] ==  - ] <- 0
martialEagle_SABAP_compl <- na.omit(martialEagle_SABAP_compl)

PA_data=martialEagle_SABAP_compl[,c(2,3,5)]
PA_data$Spp=as.numeric(PA_data$Spp)
PA_data=PA_data %>% group_by(decimalLongitude,decimalLatitude) %>%
  summarise(Spp=sum(Spp),.groups = 'drop')
PA_data$Spp =ifelse(PA_data$Spp>0,1,0)

#function to generate a random sample of 70% from both data set to do cross validation
valid_data=NA
cvme=function()
{
valid_data=NA
sample_num=sample(1:nrow(thin_data),round(0.7*nrow(thin_data)),replace=F)
thin_data= cbind(thin_data, Observed =1)
train_data=thin_data[sample_num,]
vali_data=thin_data[-sample_num,]

sample_num2=sample(1:nrow(PA_data),round(0.7*nrow(PA_data)),replace=F)

PA_data2=cbind( Polemaetus bellicosus ,PA_data)
colnames(PA_data2)[4]= Present


train_data2=PA_data2[sample_num2,]
vali_data2=PA_data2[-sample_num2,]

valid_data=as.matrix(rbind(vali_data,vali_data2))

speciesdata = list(PO=train_data,PA=train_data2)
return(speciesdata)
}

#function to do integrated distribution modeling
intgration2 = function(data,addbias)
{
  #create 2d mesh
  max.edge= diff(range(data$PO[,2]))/(3*5)
  bound.outer = diff(range(data$PO[,2]))/3
  mesh <- inla.mesh.2d(boundary = africa.bdry,
                       max.edge = c(1,2)*max.edge,
                       offset=c(max.edge, bound.outer),
                       cutoff = max.edge/5,
                       crs=st_crs(xxx))
  projection <-  +proj=longlat +datum=WGS84 +no_defs

  #prepare the model using mesh and covariates prepared
  model <- intModel(data, spatialCovariates = as(xxx,  SpatRaster ), Coordinates = c('
      decimalLongitude', 'decimalLatitude'),
```

```
                    Projection = projection, Mesh = mesh, responsePA = 'Present')
  #add a second spatial field
  if(addbias== TRUE)
  {
    model$addBias( PO )
  }
  #run the model
  modelRun <- fitISDM(model, options = list(control.inla = list(int.strategy = 'eb'),
                                            safe = TRUE))
  return(modelRun)
}

#create a 2d mesh
mesh <- inla.mesh.2d(boundary = africa.bdry,
                     max.edge = 1,
                     offset=c(2,2),
                     cutoff = 0.1,
                     crs=st_crs(xxx))

#calculate performance measure
pom2=function(result)
{
  predictions <- predict(result, mesh = mesh,
                         fun = 'linear',spatial=T,
                         mask = continents[1],
                         predictor=T)

  rp<-st_rasterize(predictions$predictions %>% dplyr::select(mean, geometry))
  rp$mean=scales::rescale(rp$mean,to=c(0.0000001,1))
  sr=as(rp, SpatRaster )
  train <- terra::extract(sr, vect(valid_data[,2:3]))

  train2=train[which(!is.na(train[,2])),2]
  vali= valid_data[which(!is.na(train[,2])),4]

  #AUC
  auc1=NA
  aucs=seq(0,1,length=100)
  for(i in 1:100)
  {
    ccc=train2>aucs[i]
    auc1[i]=Metrics::auc(vali,ccc)
  }
  auc=max(auc1)

  #TSS
  ccc=train2>aucs[which(auc1==max(auc1))]
  a=table(ccc,vali)[4]
  b=table(ccc,vali)[3]
  c=table(ccc,vali)[2]
  d=table(ccc,vali)[1]
  TSS=(a*d-b*c)/((a+c)*(b+d))

  #Tjur R2
  tr=abs(mean(train2[which(vali==0)])-mean(train2[which(vali==1)]))
  return(c(aucs[which(auc1==max(auc1))],auc,TSS,tr))
}

#generate predication based on the model and plot the prediction
predictions1 <- predict(model, mesh = mesh,
                        fun = 'linear',
                        mask = continents[1],
                        spatial=T,
                        predictor=T)
plot(predictions1)
```