Estimating Wildlife Spatial Distributions from Presence-Only Observations Using Deep Learning

Rory Bell

4th Year Project Report
Computer Science and Mathematics
School of Informatics
University of Edinburgh

2024

Abstract

Species Distribution Modelling (SDM) is crucial for conservation efforts, but traditional methods rely on expert knowledge or presence-absence data, which can be expensive and limited. This study investigates the effectiveness of deep learning architectures for building range maps using citizen science data with presence-only information. We explore two deep learning models: a single-species model and a multi-species model that considers interactions between coexisting species. Our analysis focuses on the impact of different pseudo-absence generation techniques, a method for estimating non-observations, on model performance. We demonstrate that deep learning models can achieve accurate range maps using presence-only data, with the choice of pseudo-absence generation technique playing a significant role. Our findings suggest that random sampling is most effective for single-species models, while a hybrid approach combining random and target-group sampling benefits multi-species models. This research contributes to the development of cost-effective and data-driven approaches for SDM using readily available citizen science data.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Rory Bell*)

Acknowledgements

To my supervisor, who was both my compass and my clock,

To my flatmates, with whom I wasted many valuable hours,

To my parents and my family, for getting me here,

And to others, who dug me up when I was buried underground,

Thank you.

Table of Contents

Chapter 1

Introduction

> "We are at a unique stage in our history. Never before have we had such an awareness of what we are doing to the planet, and never before have we had the power to do something about that."
>
> *David Attenborough*

1.1 Motivation

Species Distribution Modelling (SDM) is an important facet of environmental science. Range maps—the geographical spread of a species—often underpin policy decisions [47, 100], and without informed models, expert knowledge is required to produce them. There are problems with an expert-based approach for map generation: it is expensive, as experts must measure presence and absence of a given species over an area of interest; it is hard to maintain, as species can migrate over time or are displaced by effects of climate change [34, 14]; and it is not complete, as maps do not exist for many species which may be of interest. In addition, data is often species-specific and much is closed-source; when data is made available for other purposes it is frequently outdated and of limited utility [16].

Computational models are a good alternative to manually producing range maps; they can be crafted to produce maps at low cost in a potentially generalised manner [105]. Some models consider temporal aspects, such as migratory seasonal patterns or long term trends across time [37, 10], while others focus on abundance [56]—the number of a species present in a location—or ecological niches—the role and interaction of a species with its environment [76]. Many are single species only, and do not consider interactions of different species cohabiting the same environmental space [62, 45]. Where ecosystems are diverse, this singular approach can result in poor range maps [93]— particularly when high precision is important, such as modelling presence within one habitat (e.g., a forest).

Citizen-science sites such as iNaturalist [5], which records "encounter[s] with an individual organism at a particular time and location," have contributed to a global increase of available *presence-only* data—data documenting observation of a species

[33]. This lacks explicit non-observation information, which *presence-absence* data includes, but avoids the need for comprehensive, expensive studies across a region—which presence-absence requires [38]. However, presence-only data often relies on *pseudo-absence* sampling, a method of data generation, to estimate absences. These methods vary in implementation and can significantly affect model performance [13]. Presence-absence data does not require pseudo-absence sampling, but the increased data-collection requirements render it more scarce.

1.2 Project Goals and Contributions

In this study we seek to determine effective practices and architectures to produce range maps using data-driven techniques. Specifically, we investigate whether there are underlying trends in presence-only distributional data which can be well generalised by deep learning, and further by joint-species learning, and the effect of pseudo-absence techniques on model performance.

We first investigate a "shallow" statistical learning technique using random forests on presence-only data from iNaturalist, which provides us with a valuable baseline to compare to our deep models.

We then create a deep neural-network architecture which models single-species at train and test time. We conduct analysis into the best performing architecture, considering standard machine learning implementations with the inclusion of specific domain knowledge (e.g., spatial bias alignment, global land distribution).

We further extend our architecture to handle a multi-species scenario, where we consider the impact of considering many species at once. This multi-label approach considers species interactions not present many single-species models and in many cases outperforms such models. The adaptation requires few new assumptions but is computationally more expensive to train and requires greater architecture engineering to avoid common machine learning pitfalls such as vanishing gradient or overfitting.

Finally, to assess the impact of pseudo-absence selection strategies, we evaluate the performance of two common techniques—random sampling and target-group sampling—and a combined approach, on both model architectures.

1.3 Dissertation Structure

This dissertation comprises of six main chapters.

Chapter 2 provides the relevant background for understanding the context of this project by exploring historical data-driven approaches and the migration to current state-of-the-art, deep models. It also provides an overview of the presence-absence problem arising from the use of such models.

Chapter 3 introduces both our training and evaluation datasets, detailing dataset statistics, advantages, and disadvantages of each.

Chapter 4 describes the design and the implementation of the application, introducing two main models and methods of evaluation.

Chapter 5 describes the design of the experimental methods together with an analysis of the results from each experiment. Additionally, both model architectures are compared and suggestions for use are described.

Chapter 6 concludes the report, explaining how the aims of the project were achieved and providing recommendations for future work.

Chapter 2

Background

2.1 Species Distribution Modelling

Species Distribution Modelling (SDM) is a field of statistics which uses computer algorithms to analyze data on known plant or wildlife information [40]. By considering species' relationship with the environment, SDM can estimate their distribution across space, which can be used to predict habitat suitability [43] and identify areas of potential conservation concern [37]. We focus specifically on the task of spatial estimation, not including periodic (seasonal) or long-term variation (migratory patterns) [19]. We will first discuss the rise and success of machine learning in this task, followed by the challenge of generating absence data from datasets containing only presence records.

2.2 Machine Learning in SDM

2.2.1 History

Traditional SDM relies on large-scale surveys to determine presence or absence of a species. These surveys are time-consuming, expensive, and result in data often kept local to the institution who has carried these surveys out [17]. Furthermore, these closed-source practices can lead to outdated information. Surveys often focus on one species at a time [62, 45] which limits understanding of ecological communities species interactions. Additionally, the lack of open-source culture contributes hinders reproducability of studies, which can undermine confidence in their results due to lack of peer-reviewing opportunities [61].

More recently, there has been a shift in data-sharing attitudes within the ecological community [17]. However, this growing access has introduced a scalability challenge: analysis of such datasets is difficult, as it is typically rich but unclean. Data from environmental systems, for example bat observation systems, can be complex and require significant processing. It must be categorised, isolated from background noise, and distinguished from other wildlife sounds, which is intractable by manual processes [63]. These limitations highlight the need for more efficient advanced data processing

methods to effectively analyse the large amounts of data collected by modern SDM approaches.

2.2.2 Rise of Machine Learning

It is partially due to these factors that Machine Learning (ML) models have become increasingly utilised in SDM. ML is effective at handling large-scale data and can automatically analyse unclean data at large scales [67]. This is particularly appropriate for SDM, where inherent complexities can be contained in animal competition or in the dynamics of depletable resources [42]. In addition, ML methods can effectively learn complex distributions from data itself obviating the need for bottom-up statistical approaches mentioned above [74].

The relative portability of ML models makes studies easier to reproduce than typical methods, which can increase confidence in their results [9]. In addition, ML models more often use publicly-available data, such as iNaturalist [5], eBird [91], and Pl@ntNet [72], which are large-scale data repositories of reported sightings of wildlife species. Use of similar datasets makes studies externally comparable as benchmarks exist on the data [96, 35], though certain limitations exist when using these datasets which are later discussed in Section 2.3.

The rising use of ML in ecology has resulted in some friction. ML approaches, which are data-driven, can emphasise results with limited grounding in little ecological theory, potentially causing less consideration to be given to conclusions [98]. The non-transparent [31] and non-invertible [78] nature of ML models contribute to this interpretability issue; many ML studies warn against using results without complimentary ecological information [27]. While research is actively ongoing—in the field and for ML in general—to understand how model predictions can be explained and defended [101], ML models are currently less suitable for isolated use compared to more traditional (and often more interpretable) ecolocal aproaches.

While ML offers significant capabilities, it also has limitations, such as challenges in interpreting model outputs. Nevertheless, the ability to analyze complex, large-scale SDM data makes ML a valuable tool for ecological research. In order for ML models to be relied upon in the ecology field, they must be easy to evaluate, show a clear predictive path that can be easily verified, and be widely applicable across different species.

2.2.3 Machine Learning Approaches

As ML has grown for use in SDM, different approaches have attempted to effectively model the distribution of species using data-driven, statistical, or heuristic techniques. We discuss some common approaches in this section.

Modelling the distribution of species is a highly complex system due to the intricate relationships between species and their environment [42]. Traditional statistical techniques often struggle with inherent variability, non-linearity, and interacting factors involved in these systems [17, 36].

Modern approaches have moved towards data-driven machine learning models which

instead assume a complex distribution is present and tries to learn it from the data itself [96]. The models typically apply supervised learning—where models are trained on data with known species presence and absence locations [28]. Through iterative learning, these models learn to to predict the spatial or spatial-temporal distribution of a species. They differ in several key aspects and we briefly summarise movement from assumption-driven statistical techniques to data-driven deep learning techniques (current state-of-the-art) below.

We introduce the terminology of *ensemble* learning, when multiple models' predictions are combined to improve overall accuracy (*bagging*) [22] or when models are sequentially training to address errors from previous models (*boosting*) [41]. We also discuss *overfitting*, which is decreased performance on novel data due to memorisation of training data [50].

2.2.3.1 Models

Gradient Boosted Regression Trees (GBT) were one of the first migrations from statistical techniques, building upon earlier statistical techniques [32]. Tree-based methods such as GBT recursively split data using binary decisions at each level, allowing them to fit complex, non-linear relationships in the data. GBTs additionally do not require manual feature extraction techniques—when information is extracted from the data—making them less reliant on expert input [36]. While single trees are prone to biases of the training data, GBTs combine trees through *boosting*, which minimises training error (the difference between predicted and actual values) However, boosting can contribute to overfitting, especially when the training set has distinct outliers or unrepresentative samples as boosting focuses on the worst performing parts of the data—those which may be unrepresentative [99].

Due to the sequential nature of boosting, GBT cannot be parallelised—when computations are performed simulatneously on multiple processors—and therefore can be computationally expensive when data is complex. As such, GBTs are less suitable for large, multi-species models which are more complex than single-species models due to increased data volume [16]. One approach to address this scalability issue is *Extreme Gradient Boosting (XGBoost)* [26] which introduced more tunable parameters to allow for more efficient model training and improved scalability on complex datasets. While GBTs have been surpassed by other methods, GBTs and XGBoost remain relevant techniques for tackling some problems in SDM.

Random Forest (RF) models address limitations encountered by GBTs by introducing a different ensemble learning tecnique, *bagging*. Bagging—sampling from the training set with replacement—produces decorrelated trees [23] and allow parallelisation. This makes RFs significantly more scalable for complex datasets, especially multi-species models [95]. However, this increased scalability brings increased complexity, meaning RF models can be harder to interpret compared to simpler models like GBTs [11]. This trend of increased complexity for improved performance is a recurring theme in machine learning [44].

Following the success of RFs, other data-driven models have emerged such as Support

Vector Machines, which excel at handling high-dimensional data [52], and Generalised Additive Models, which can accurately capture non-linear relationships in ecological data [48, 17]. However, in recent years, deep learning models such as the *Feedforwad Multi-layer Perceptron* (MLP) have emerged as state-of-the-art for SDM tasks [60]. This genre of models, inspired by the connection of synapses in the brain, can learn complex patterns from data through iterative learning processes. We discuss deep learning models in the following section.

2.2.4 Deep Learning

MLPs are a type of deep learning model consisting of interconnected nodes arranged in multiple layers. They are *universal function approximators*, meaning that, given appropriate complexity (number of layers and nodes), are able to represent any complex relationship between inputs and outputs [55]. This makes them well suited for SDM, where the relationship between species and environmental factors can be intricate. In this section and hence, we will categorize the models described earlier (Section 2.2.3) as 'shallow' models, and MLP as 'deep' models. We note there are many powerful deep learning architectures beyond MLPs such as Transformers [97] and Recurrent Neural Networks [80]. However, in this study we limit our foucs to fully-connected MLPs due to their suitability to the task.

Deep models in SDM can be categorised into single-species or multi-species approaches. Single-species models are computationally less intensive to train and may be suitable for lower-end systems, such as when models are required to run on environmental measurement systems and perform inference at measurement time [21]. However, multi-species models have an intrinsic advantage over multi-species models by considering species interaction [73]. While these interactions may not be explicitly measured, presence of a species may consider hidden (latent) information about the distribution of other species. For example, the presence of a competitor (or predator) may inform of the presence or absence of a rival (or prey). Deep models are particularly adept at capturing these underlying trends in location data, allowing them to infer species interations even when not directly measured.

Deep learning models can incorporate various data sources to improve performance. Beyond location data, some studies consider environmental factors such as elevation or 'meta factors' like photographer bias [64]. Additoinaly, some models incorporate temporal data (information on the data over time), while others weight presence data differently (e.g., giving more weight to recent sightings).

It's important to consider limitations when evaluating existing literature on SDM. First, some studies might suffer from reporting bias, where researchers are more likely to publish results that show improvement over previous work. Second, studies can be difficult to compare directly due to their focus on different geographic scales (e.g., local vs global [103, 81]) or specific taxa (groups of organisms). In addition, training data often contains bias. Teaching models to learn this data includes teaching it the bias, whereas non data driven or shallow models may be more robust [58]. We note that many methods perform point-wise evaluation based on single input and output points. Unexplored in literature is an investigation into performance of models with multiple

points as input (e.g. a potential range), which could benefit from not only joint learning between species, but also between environmental factors.

2.3 The Presence-Absence Problem

ML models for SDM require data on both presence (positive observations) and absence (negative observations) of a species in a particular area. Supplying a model with only positive data will result in a model which predicts true everywhere, as it will have no notion of negative space. **Presence-only** data typically takes the form of reported sightings of a species, but provides no evidence for what species were not observed in the area. As people may only report sightings for areas they have visited, presence-only data is typically spatially biased towards well trodden or accessible areas such as national parks or tourist destinations [39]. This can lead to underrepresentation of areas with a high biodiversity, such as the Amazon rainforest, or low biodiversity, such as the Arctic [46, 30]. In contrast, **Presence-absence** data contains both positive observations (where a species is present) and negative non-observations (where it is not). This approach is less inclined to this bias, as it is typically generated for areas of interest through scientific study, which may include these areas [86, 90].

While presence-absence data offers advantages over presence-only data by including reliable information on species absence, it is typically less extensive due to the effort required to collect negative observations [38]. Unlike presence-only data, which can be crowd-sourced from reports of sightings, collecting absence data requires expert-knowledge and significant time investment, and may be specific to a particular species or study area. This limitation, when put in context of data-hungry machine learning approaches, often makes presence-absence data less generalisable compared to presence-only data.

The wide accessibility of presence-only data makes it attractive for many SDM studies. However, as discussed earlier, lack of absence data requires generation of **pseudo-absences**, where artificial points representing species absence are created. We investigate techniques of pseudo-absence sampling in Section 2.3.1.

2.3.1 Absence Generation Techniques: How many, where, and by what method?

The method of absence generation has a significant impact on the final performance of presence-only deep learning SDM models [13]. Despite this, there is currently no standardized method of producing pseudo-absences, and research for the best approaches for multi-species models is limited [103]. Two methods are popular in literature, which we will refer to as **random sampling** and **target-group sampling**. Random sampling consists of generating a number of points across a study area uniformly without bias; target-group sampling is when the presences of species are used as absences for others at training time. We discuss advantages and disadvantages of both in the following section.

How many?

The number of negatives generated is an important consideration in SDM. Too few negatives result in poor learned negative space—absences of species—and worse overall performance [57]. Generating too many pseudo-absences will result in an imbalanced dataset, where the number of negative points far outweigh the number of positive presence points. ML models perform better when the dataset is balanced, as imbalance will tend to overfit towards the majority class [77]. To address this imbalance, *oversampling* is sometimes used, where the positive samples are duplicated during training to increase total positive observation numbers and balance with the number of negatives [25]. However, while oversampling might create a numerical balance between positive and negative observations, it doesn't necessarily address the underlying issue of potentially unrepresentative positive data.

By what method?

Pseudo-absences can be attributed with having some form of 'information content' [65]. A point which is far away, in already negatively learned space has 'low' information content— the model will not learn anything new given this point during training. Similarly, a point near a boundary of the predicted presence area, will be highly informative: an absence will constrict the zone to increase its specificity.

Random sampling, when pseudo-absences are generated uniformly across a target zone, is a simple approach utilised by many studies, with good results [85]. However, by uniformly sampling across the space, the most common environmental conditions in that space are overrepresented [29], meaning less common environments may be underrepresented in resulting range maps. Furthermore, random sampling—particularly when considering the global scale—may result in sampling many uninformative negatives [103]. That is, points far away from a species' zone of interest, or points over water, terrestrial species will not inhabit, may form the bulk of negative observations. Consequently, resulting SDM models might be less specific, predicting the species' presence in wider areas than it actually occupies.

Target-group pseudo-absences, absences taken from the presence observations of other species, mitigate some of these issues by providing points with a bias aligned to that of the observation points [12]. However, using target group-only points, range maps can fail to learn important negative space and ranges can become 'ballooned', for example over water where no presences for any species are recorded. In addition, the bias contained in target-groups will overrepresent well travelled areas. However, this method has also been shown to produce accurate and precise range maps, and is a popular method used in many studies [71].

In summary, target-group sampling is efficient but not exhaustive, but random point generation is exhaustive, but not necessarily efficient. It is possible that a combination of these points may result in complimentary bias alignment.

Chapter 3

Data

The quality and relevance of the data used to train machine learning models significantly impact their final performance [84]. For our purpose, presences (positives) and absences (negatives) were required in order to train a supervised machine learning model to predict global species range, and ground truth data was required to evaluate our model.

Our positive data comes from open source citizen-scientist website iNaturalist [5]. Section 3.1 provides a brief overview of the dataset and highlights particular limitations and challenges arising from its use. Ground-truth—gold-standard positive and negative data used for model evaluation—comes from IUCN Redlist [6], discussed in Section 3.2.

3.1 iNaturalist

iNaturalist focuses on collecting documented observations of organisms. An observation, as defined by iNaturalist, is "an encounter with an individual organism at a particular time and location." [5]. Encounters may include direct sightings—with evidence—or indirect observations such as stool or tracks. To ensure data quality and confidence, the dataset excludes unverifiable observations which are not considered 'research grade' (Section 3.1.2). Additional exclusionary criteria are non-recent observations such as fossilisation or historical sightings and observations with unverifiable or inaccurate locations.

The iNaturalist dataset is a rich source of information used in many research applications [54, 79], but is also applicable to other purposes. iNaturalist has been claimed to improve the outdoor experience by providing identification suggestions [20], increase public knowledge of conservation efforts [87] and decrease biodiversity naivety—limited understanding of nature's rich variety—in young people [68].

3.1.1 What is the dataset?

Each observation in the iNaturalist dataset is associated with a taxon ID, which refers to a group of organisms categorised hierarchically, ranging from broad classifications (kingdoms) to specific ones (species). However, the non-expert user is not required

to have full knowledge of the taxa they observe: iNaturalist allows submission into a "Needs ID" category, where taxa will be labeled by species experts. 'Research grade' observations are downgraded to casual if the data cannot be verified or there is low agreement between identifiers [5].

iNaturalist utilises computer vision (CV) to aid casual observer classification. The CV system is trained on species with greater than 100 observations, and is seen to be more accurate in high activity areas such as North America or New Zealand. While CV is useful in the context of identifying common species, it is limited when differentiating between visually similar animals. CV assisted identifications are labeled as such in the iNaturalist dataset, but considering this distinction is beyond the scope of our study.

Due to endangerment or particular sensitivity, some observations may be privatised—in which case they are only available to government or research institution—or obscured, where observations coordinates are replaced within a 0.2x0.2 degree cell (equivalent to 500 km^2 area at the equator). Examples of original, private, and obscured data is given in Figure 3.1. Our dataset does not include such observations, which may limit its effectiveness for affected species.

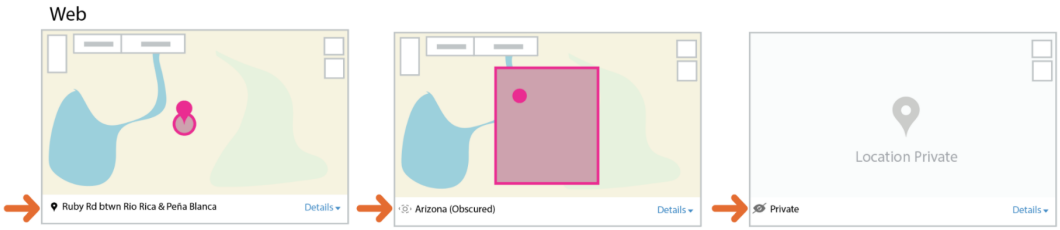


Figure 3.1: Normal, obscured, and private locations from left to right. Figure adapted from iNaturalist. Available from https://www.inaturalist.org/pages/help. Accessed 23/03/24

Not all species are equally represented on the iNaturalist website. Birds are well represented, representing 24.8 million of the 36.8 million total vertebrate observations on the site [1, 4], wheras mammals and reptiles represent only 3.8 million observations each [2, 3]. This over representation of some taxa over others continues in our dataset.

3.1.2 How is the data collected?

The process for recording an observation is outlined as follows:

1. A photograph or auditory capture is taken and submitted by an observer, complete with GPS coordinates, the time documentation was recorded, and other meta-data.
2. By expert knowledge or with aid of built-in species classification, species are labeled and added to the 'casual observations' map.
3. If an observation is verifiable, contains photo or audio documentation, and more than two-thirds of human annotators agree on an identification, the observation is granted 'research grade' status.
4. Research grade observations are exported and used for large scale surveys of single and multiple species.

iNaturalist data is presence-only. As described in Section 2.3, no absence data is supplied and so in order to be utilised for statistical modelling, absences must be generated. We refer the reader to Section 2.3 for an overview of these techniques and we discuss our method later in Section 4.5.

3.1.3 iNaturalist Exploration

Despite the large number of potential species available, some species in the iNaturalist set have very few observations. We limit our study to species with more than 50 observations, and limit observations per species at 2000. Our dataset consists of 272,037 observations across 500 randomly selected taxa which meet this criteria. Figure 3.2 shows that the observations follow a positively skewed power-law distribution—except for at our observation limit. For each species in this dataset, there is a corresponding expert range map from our ground truth dataset discussed in Section 3.2.

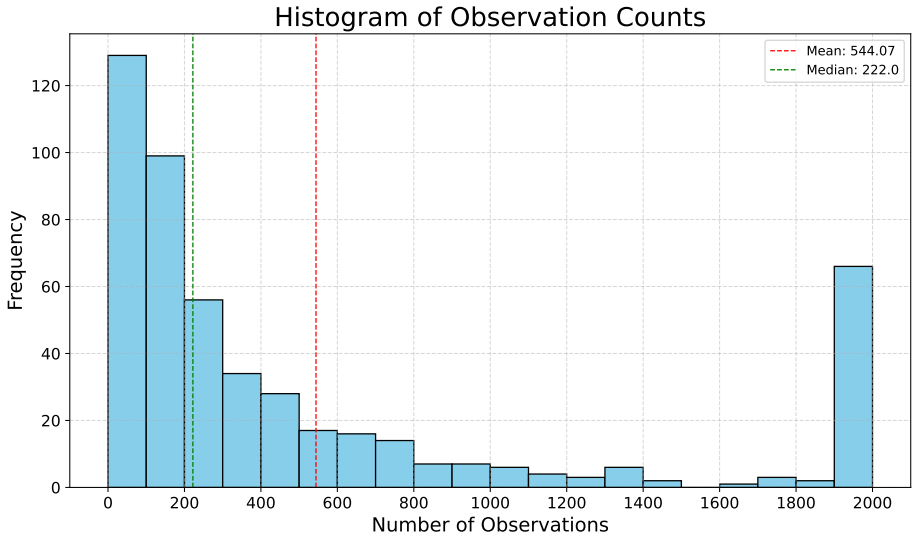


Figure 3.2: Number of observations for species in our iNaturalist dataset.

Figure 3.2 shows a skewed distribution of observations across species.The median number of observations is 222, which is lower than the mean, 544, meaning the majority of species have fewer than 250 observations and many species fall into this low observation category. 63 species in the dataset have the maximum number of observations while only 3 have the minimum, though a significant portion fall into the 50-100 observation bin. This observation distribution may pose challenges when training models for species with this limited data, as fewer data points can lead to overfitting, where the model performs well on training data but generalises poorly to unseen data.

Figure 3.3 illustrates the geographic distribution of observations. As expected, observations are concentrated in more developed areas such as North America, Europe, and New Zealand. Conversely, there are fewer reported observations in sparsely populated regions areas such as Russia (though accentuated by map projection in the Figure) and less developed regions of central Africa.

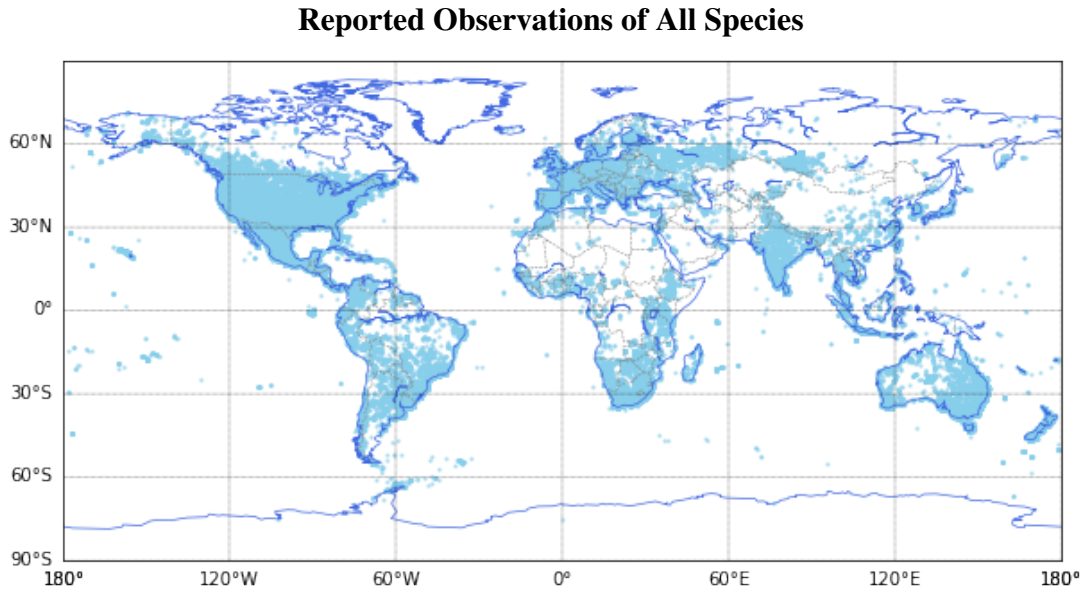


Figure 3.3: Locations of all species observations in our dataset. Showcases higher concentration of observations in developed regions like North America, Europe, and New Zealand and fewer in sparsely populated regions like Russia and less developed regions of central Africa.

This study does not consider temporal (time-dimension) data. As such, each entry in our dataset consists of one verifiable observation—location and species id—with a high confidence of accuracy, providing invaluable presence-only data.

3.1.4 The Validation Problem

Many machine learning methods use a random subset of the training set for validation to measure overfitting—poor performance on novel data despite good training performance—or predict generalisation performance—performance on unseen data [49]. However, in SDM, partitioning the training set in this way would place validation points interspersed within training locations, and therefore is a poor estimation of novel performance or indicator of overfitting.

Splitting data geographically for training and validation (as seen in some SDM studies [65, 15]) may not fully address these issues, as each set will retain spatial biases present in the original presence-only data (Section 2.3). Additionally, presence reports are often unevenly distributed within a species' true range. This means some areas with confirmed presence might have no recorded observations, leading to inconsistencies or poor model performance during evaluation.

To address the limitations of validation with presence-only data, we adopt a different approach. We do not use validation during training and instead train the model until it converges on the training data (details in Section 4). To assess model performance, we then evaluate its ability to generalise to different scenarios. This evaluation involves testing models trained on distinct species using the same model architecture (single-species) or applying the same model to predict the distribution of multiple species

(multi-species). For this evaluation, we use the unbiased IUCN test set discussed next in Section 3.2.

3.2 IUCN Redlist: Our Ground Truth

Though iNaturalist provides extensive presence-only observations for many species, it lacks ground truth information. Through historical studies, human expert-generated maps are available for many species; many are collated in the International Union for Conservation of Nature (IUCN) Redlist [6], an online species record collection with a focus on endangered species. The IUCN Redlist categorizes species into one of nine categories, ranging from 'least concern' to 'extinct,' based on information gathered by trained individuals [6]. Given its extensive data and reliable information, IUCN Redlist provides an invaluable tool for evaluating our generated range maps.

3.2.1 Comparison to iNaturalist Data

We match IUCN data with iNaturalist at the species level through a common taxon identifier. Provided by Mac Aodha et al. [64], test location data is in the form of a dense output grid for which every species has a presence marking for the locations within the range map determined by the IUCN. Observations not marked as present are assumed as negative for evaluation purposes. The observations consist of a dense grid covering the globe which allows point-wise evaluation for model predictions for each species at every location.

Given that IUCN is based on expert-derived conclusions and not observations, the distribution of presences in the iNaturalist dataset is different from data in IUCN: iNaturalist suffers from spatial and other reporting biases discussed in Section 2.3, but IUCN is less affected by these trends. IUCN is not completely unbiased: species present in less active areas are underrepresented in IUCN as well as iNaturalist (people tend to study local species of interest), but we expect species ranges present in the IUCN set to be accurate. exhaustive.

3.2.2 Limitations of IUCN

Though IUCN has excellent presence-absence information for each species it has in its set, many species are underrepresented. Data is spatially biased towards land-based (in particular, forest) ecosystems, and towards animal observations over plants and fungi [6]. This does not affect our application, as we consider wildlife distributions only, but does affect the type of studies done and accentuates the gap as species with more resources are likely to be the subject of more studies.

However, our study does not consider species with less than 50 observations in the iNaturalist dataset, and rarer species are more likely to fall below this threshold. This is not unusual in data-driven SDM [75], and it is therefore important as future work to consider under-represented species and investigate effective ways of producing accurate range maps for such species. In particular, 'endangered' and 'critically endangered'

species fall into this low-observation category, and so finding ways to improve low-observation performance by including more latent information is an important next step to increase confidence in using data-driven SDM for conservation efforts.

Despite these limitations, the IUCN set contains ground-truth data for 157,190 species and so forms invaluable evaluation for our models on the species we have available. We display three example expert range maps for species of varying concern in Figure 3.4.

Expert Range Maps of Least Concern, Vulnerable, and Endangered Species from IUCN Dataset

(a) Expert range map of least concern American Robin (taxa ID 103889499) from IUCN. Map cropped to region for clarity.

(b) Expert range map of vulnerable Giant Panda (taxa ID 712) from IUCN. Map cropped to region for clarity.

(c) Expert range map of critically endangered Eastern Gorilla (taxa ID 39994) from IUCN. Map cropped to region for clarity.

Figure 3.4: Ground truth maps of American Robin, Giant Panda and Eastern Gorilla from IUCN dataset [6]. Accessed 14/3/24.

Chapter 4

Methodology

This chapter focuses on model implementation, first outlining the problem definition and defining notation used throughout the chapter. We introduce two models: a single-species model that predicts the probability of presence for a single species at a given location, and a multi-species model that predicts the probability of presence for all species simultaneously. Both are fully-connected feedforward multi-layer perceptron (MLP) models.

We later investigate the impact of pseudo-absence generation techniques on model performance, and next introduce two methods: random and target-group sampling. The chapter is concluded with a brief discussion of SDM evaluation, introducing metrics later used in Chapter 5.

4.1 Problem Definition

Let $\mathbf{x} = [x_1, x_2]$ be a geographical location and $\mathbf{y} \in \{0,1\}^{|\mathcal{S}|}$ denote the ground truth presence (1) or absence (0) for each species $s \in \mathcal{S}$ at location $\mathbf{x}$. Our goal is to predict $\hat{\mathbf{y}} \in [0,1]$ for any $\mathbf{x} \in \mathbb{X}$ such that $\hat{\mathbf{y}}$ estimates $\mathbf{y}$, given observed data $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ for some N ground truth locations and some $\mathbb{X}$, a set of points across the globe. As our dataset is presence-only, we only have entries corresponding to $y_i = 1$, meaning absence (0) is lacking for most locations $\mathbf{x} \in \mathbb{X}$. To generate absences, we utilise pseudo-sampling to generate $\hat{\mathbf{y}}' \in \mathbb{X}$, a pseudo-absence in our study area.

For a fixed location encoder $g : \mathbb{R}^2 \to \mathbb{R}^4$ such that $\tilde{\mathbf{x}} = g(\mathbf{x})$ and some parameters θ, each model is a mapping $f_\theta : \mathbb{R}^4 \to [0,1]^{\bar{S}}$ for $\bar{S} = |\mathcal{S}|$. The model is parameterised by $\hat{\mathbf{y}} = f_\theta(\tilde{\mathbf{x}})$, and model output $\hat{\mathbf{y}} \in [0,1]^{\bar{S}}$ is a measure of likelihood for all species $s \in \mathcal{S}$ to appear at location $\mathbf{x}$. We look to find

$$\theta^* = \text{argmin}_\theta \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\hat{\mathbf{y}}_i, \mathbf{y}_i)$$

for a suitable loss function $\mathcal{L}$. Once θ^* has been estimated, we may use f_θ to predict species presence for any location on the globe.

4.2 Single-Species Model

As described above, our model is a mapping $f_{\theta_{ss}} : \mathbb{R}^4 \mapsto [0,1]^{\bar{S}} : f(\tilde{\mathbf{x}})$. In the single-species case, $\bar{S} = |\mathcal{S}| = 1$, which corresponds to the species of interest. As such, output dimension for each input location $\mathbf{x}$ is one, corresponding to $\hat{y} \in [0,1]$.

The single-species architecture first selects input coordinates, $\mathbf{x} \in \mathbb{X}$, using a balanced sampler. This sampler randomly selects presences for the target species with replacement, ensuring an equal number of absences are included in each batch. Each $x_i \in \mathbf{x}$ is then scaled to [-1,1] before being mapped to $\mathbb{R}^4$ via the location encoder (described in Section 4.4) to get $\tilde{\mathbf{x}} = g(\mathbf{x})$.

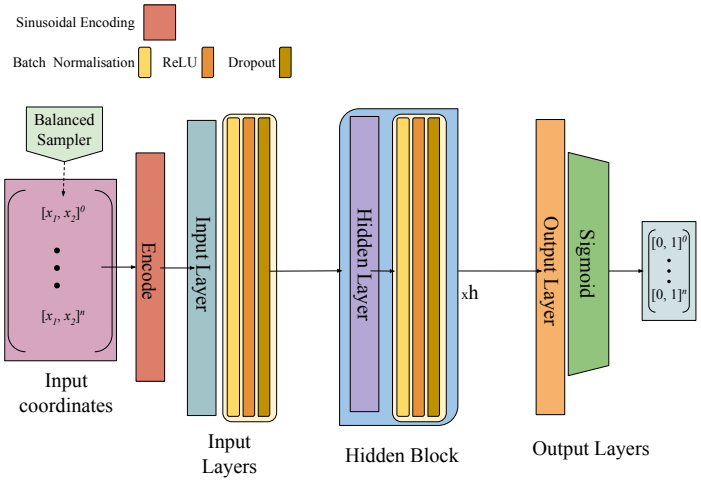


Figure 4.1: One batch example of single-species model architecture. 'h' is number of hidden blocks, which varies in experimentation; 'n' is size of batch. Number of input points equals number of output predictions. Input layer projects from input to hidden dimensions, output layer projects from hidden to binary classification.

A prediction, $\hat{y} = f_{\theta_{ss}}(\tilde{\mathbf{x}})$, is then generated through the architecture seen in Figure 4.1. First, $\tilde{\mathbf{x}}$ is passed to an input layer which projected each input into the hidden dimension. Data was then passed though a *regularisation* block consisting of batch normalisation—which helps mitigate internal covariate shift, where the input distribution varies during training [82]—a ReLU activation function [66], and dropout, which randomly drops input weights during training to avoid becoming overly reliant on any one feature [88]. Data was then passed through h hidden blocks, where h was varied during experimentation. Finally, data was passed to an output layer which projected to a binary classification and then a sigmoid activation to normalise to $\hat{y} \in [0,1]$. We used no residual connections—where input data is combined with model output after perceptron layers to facilitate the training of deeper networks—as we experienced no vanishing gradient during training, though other work has used these in their models [103, 104, 27].

For training and initial results, we trained on a batch size of 128 with 128 hidden units and 2 hidden layers. During our experiments, we increased our batch size to 2048 and hidden units to 256 to align with prior work [64]. We used Adam optimiser [59]

with 0.001 learning rate, no learning rate scheduler, and dropout probability 0.1. As the single-species task is binary classification, we used PyTorch's built in binary cross entropy (BCE) loss function [69], which aims to minimise difference between predicted and ground truth labels. It computes the negative log likelihood of observed data under the predicted probability distribution by penalizing incorrect predictions in proportion to how they are from being correct. Mathematically, for $y \in \{0,1\}$ ground truth label and predicted outcome $\hat{y}$, the BCE loss for each location is given as:

$$\mathcal{L}_{\text{BCE}}(\hat{y}, y) = -((y\log(\hat{y}) + (1-y)\log(1-\hat{y})). \tag{4.1}$$

Minimizing the BCE loss trains the model to produce predicted probabilities that are close to the true labels, resulting in improved classification performance as training progresses.

4.3 Multi-Species Model

Our multi-species model is a mapping $f_{\theta_{ms}} : \mathbb{R}^4 \mapsto [0,1]^{\bar{S}} : f(\tilde{\mathbf{x}})$ for $\bar{S} = |\mathcal{S}|$ and $s \in \mathcal{S}$ species, which, for each input location $\mathbf{x}$, predicts the probability of presence for all species $\hat{\mathbf{y}} \in [0,1]^{\bar{S}}$. In contrast to the single-species case, this is a multi-class problem.

To account for multiple species and with inspiration from Mac Aodha et al. [64], the balanced sampler was adapted to ensure all species were equally represented over a training epoch (iteration through the training set). We adapt the custom loss function which pairs a randomly generated negative observation for a species with each presence: when combined with the balance of species across an epoch, we note that our dataset is fully balanced. For ground truth $\mathbf{y}$ and model output predictions $\hat{\mathbf{y}}$, we have

$$\mathcal{L}_{\text{random}}(\hat{\mathbf{y}}, \mathbf{y}) = -\bar{S}^{-1} \sum_{j=1}^{\bar{S}} \log(\hat{y}_j) + \log(1-\hat{y}'_j)$$

for randomly generated point $\hat{y}'_j = f_{\theta_{ms}}(g(\mathbf{a}))$ where $\mathbf{a} \sim \text{Uniform}(\mathbb{X})$ is a randomly sampled point on the globe.

We also created a new model pipeline, incorporating residual connections [51] and a learning rate scheduler [102] to account for greater model complexity introduced by training on all species simultaneously. Residual connections increase information flow across layers which facilitate the training of deeper networks [51]. A learning rate scheduler decreases the learning rate at set intervals; ours was a polynomial scheduler updated each epoch by the formula given in Equation 4.2.

$$\text{learning\_rate} = \text{initial\_learning\_rate} \times \text{epoch}^{0.98} \tag{4.2}$$

This architecture—potentially due to the greater batch size and therefore less variation among training batches—did not require batch normalisation during initial testing,

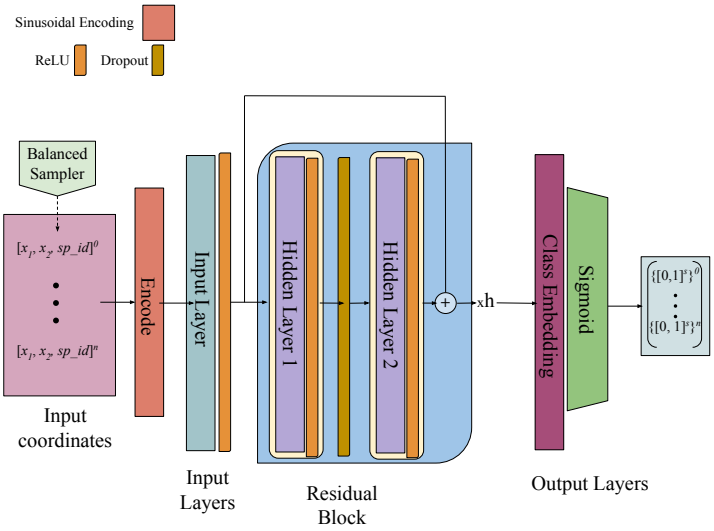


Figure 4.2: One batch example of multi-species model architecture. 'h' is number of residual blocks, which is set to 4 in our study. Output dimension is $[N, \bar{S}]$ for N input points and $\bar{S}$ species. Input layer projects from input to hidden dimensions, output layer projects from hidden to output. Note lack of BN, as not required during training. '+' indicates a residual connection, which allows information to flow between deeper layers in complex models.

though we did still incorporate dropout to decrease model reliance on specific features (e.g. presence of one very informative species).

To generate output predictions $\hat{\mathbf{x}}$, input locations $\mathbf{x} \in \mathbb{X}$ were still sampled, scaled, and encoded to get $\tilde{\mathbf{x}} \in \mathbb{R}^4$ as in Section 4.2. Data was then passed to a residual block consisting of two hidden layers, each followed by a ReLU activation function with dropout applied between. The skip connection was applied, and the block was repeated h times, where here h was fixed at 4. After going through all residual blocks, model outputs were projected back to number of species through a learned *class embedding*, which allows models to relate hidden dimensional representations to discrete output (our species IDs) [8]. The outputs are passed through a sigmoid activation as in the single-species case and a prediction is made for each species for each location, $\hat{\mathbf{y}} \in [0,1]^{\bar{S}}$. During training, this output was passed to the loss function as described above. This pipeline is shown in Figure 4.2. We train using Adam (Adaptive Moment Estimation) Optimiser [59] with initial learning rate set to 0.001 and batch size 2040 for 10 epochs.

4.4 Sinusoidal Encoding

By artefact of projecting a sphere (the globe) onto a plane (map), two surfaces with different Gaussian curvature, any coordinate base will have a line of discontinuity at $(0, 2\pi)$, meaning points nearby in spherical space may be far apart in coordinate space—points at either edge of a map. Motivated by this problem, we map input coordinates $\mathbf{x} \in \mathbb{R}^2$ to $\tilde{\mathbf{x}} \in \mathbb{R}^4$ by rescaling input coordinates to the interval $[-1, 1]$ and a sinusodial mapping $g : \mathbb{R}^2 \mapsto \mathbb{R}^4$. This encoding, previously utilised in prior studies [27], ensures

continuity in the input space.

To formally analyze this behavior we introduce the concept of *continuity around a boundary*, a generalization of continuity relating to functions on closed domains. In this context, we prove continuity around the interval $[-1,1]$ for the function $h(x) = (\sin \pi x, \cos \pi x)$.

Proposition 1. *Consider* $h : \mathbb{R} \to \mathbb{R}^2 : h(x) = (\sin \pi x, \cos \pi x)$. *Then* h *is continuous on* $[-1,1]$ *and further is continuous around the boundary* $x = -1$ *to* $x = 1$.

Proof. We first observe that $h_1(x) = \sin \pi x$ and $h_2(x) = \cos \pi x$ are both globally continuous functions since sin and cos are continuous on $\mathbb{R}$. Therefore, h is continuous on the closed interval $[-1,1]$ as each component function is continuous on this interval.

Furthermore, h is continuous around $x = -1$ to $x = 1$ if $\lim_{x \to -1} h(x) = \lim_{x \to 1} h(x)$. We calculate these limits as follows:

$$\begin{aligned}
\lim_{x \to -1} h(x) &= \lim_{x \to -1} (\sin \pi x, \cos \pi x) \\
&= (\sin(-\pi), \cos(-\pi)) \\
&= (0, -1) \\
&= (\sin \pi, \cos \pi) \\
&= \lim_{x \to 1} h(x).
\end{aligned}$$

Thus, h is continuous around $[-1,1]$.

□

For each input coordinate $\mathbf{x} = [x_1, x_2] \in \mathbb{R}^2$ we then have, for $i \in \{1,2\}$, $x_i \in \mathbb{R}$. We then define $\tilde{x_i} = h(x_i)$ and thus

$$\tilde{\mathbf{x}} = [\sin \pi x_1, \cos \pi x_1, \sin \pi x_2, \cos \pi x_2] \tag{4.3}$$

Through this encoding, our models $f_{\theta_{ss}}$ and $f_{\theta_{ms}}$ can better learn species distributions without being negatively affected by map boundary effects.

4.5 Pseudo-Absence Sampling

As motivated in Section 2.3, presence-only datasets require pseudo-absence generation to represent areas where a species is likely to be absent. This allows machine learning models to learn "negative space" and prevent them from predicting true across the entire region. We initially used **random sampling** to generate pseudo-absences—where points are generated from an entire zone without bias—using which we generated initial results. However, to explore the impact of pseudo-absence generation techniques on model performance, we outline an alternative method, **target-group sampling**—where absences are sampled from the presences of other species. A visual representation of target-group and random sampling is shown in Figure 4.3

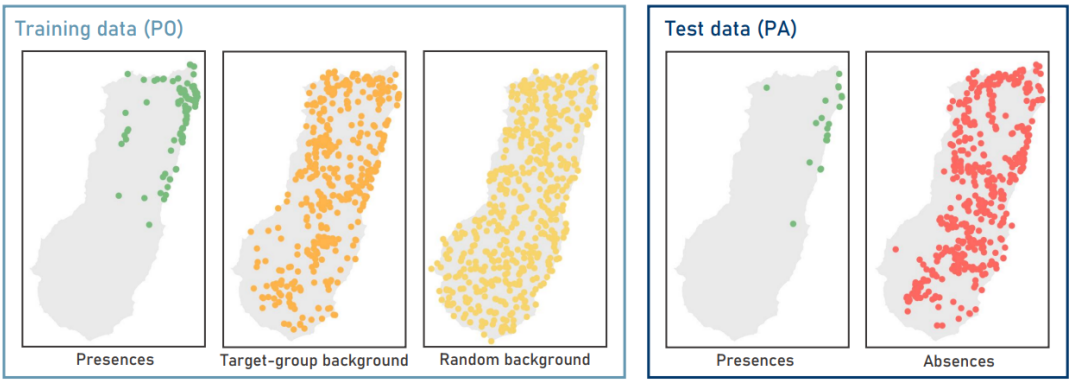


Figure 4.3: Machine learning requires presence and absence data to train models, and sometimes absence data is not available and must be generated. This may be done through *pseudo-absences*, and different generation techniques exist to generate these absences. Target-group points use presences of other species, and random sampling selects random locations from within a region. Figure from Zbinden et al. [103]. Accessed 23/03/2024.

4.5.1 Pseudo Sampling Methods

Random Sampling

Our initial random sampling approach was a random point generation which generated random points on a map between limits of latitude and longitude. However, this method suffers from a bias towards the poles due to the convergence of lines of longitude at those locations—one degree toward the pole represents a progressively smaller distance as the line is traversed. To address this bias, we employed a different method, *random-sphere sampling*. This method samples two uniformly distributed random variables $(\alpha_1, \alpha_2 \sim \text{Unif}[0,1))$ to generate points uniform across the sphere through Equations 4.4-4.7. This adaptation facilitates greater uniformity of samples across the surface of the Earth, which assists models in learning negative space.

$$\theta_1 = 2\pi\alpha_1, \tag{4.4}$$

$$\theta_2 = \arccos(2\alpha_2 - 1), \tag{4.5}$$

$$x_1 = \frac{\theta_1}{\pi} - 1, \tag{4.6}$$

$$x_2 = 1 - \frac{2\theta_2}{\pi}. \tag{4.7}$$

Target-group Sampling

Target-group sampling required no additional point generation, as negative locations are drawn from the presences of other species in the dataset. For $f_{\theta_{ss}}$, the balanced sampler was adapted by replacing random sampling by the presence of another random species

presence. For $f_{\theta_{ms}}$, the loss function was modified to select target-group points:

$$\mathcal{L}_{\text{target-group}}(\hat{\mathbf{y}}, \mathbf{y}) = -\bar{S}^{-1} \sum_{j=1}^{S} \log(\hat{y}_j) + \log(1 - \hat{y}_{j'})$$

where $j' \sim \text{Uniform}(\{j_p \in \{1, ..., \bar{S}+1\} : j_p \neq j, j \in \mathbb{N}\})$ is a species index distinct from j and $\bar{S} = |\mathcal{S}|$.

Hybrid Approach

To determine optimal pseudo-absence sampling techniques, we also investigated the effect of combining both target-group and background sampling approaches. For $f_{\theta_{ss}}$, we merged the balanced samplers from target-group and random sampling, taking half of each method in each batch. For $f_{\theta_{ms}}$, a final loss function was utilised, adapted from Cole et al. [27], which combined both approaches: the input point $\mathbf{x} \in \mathbb{X}$ was applied as negative to all but the input species (target-group) and a random point was sampled for the input species. This loss function is given as

$$\mathcal{L}_{\text{both}}(\hat{y}, z) = -\bar{S}^{-1} \sum_{j=1}^{S} \left(\log(1 - \hat{y}_j) + \log(1 - \hat{y}'_j)\right) + \mathbb{1}_{[z_j=1]} \log(1 - \hat{y}_j),$$

where $\hat{y}'_j = f_{\theta_{ms}}(g(\mathbf{a}))$ for $\mathbf{a} \sim \text{Uniform}(\mathbb{X})$ as in $\mathcal{L}_{\text{random}}$ and $z \in \{0, 1\}$ represents species presence (1) or absence (0) for indicator variable $\mathbb{1}$, which takes the value '1' if the condition $(z_j = 1)$ is true, and '0' otherwise.

Additional Techniques

We considered additional negative sampling techniques such as sampling in spatially distinct locations from training data [92, 53], though studies have shown that this does not outperform random sampling when enough points are sampled [13] possibly due to many presences overshadowing the same zone as the absences (especially when oversampling). In fact, bias may be amplified due to reinforcing known zones and poor generalisation to unknown ones. Thus we limit our investigation to target-group and random samples.

4.6 Evaluation

Species may only be *present* or *absent* at a location, so single-species models are binary classifiers, and multi-species models are multi-label classifiers. Locations where the model and ground truth—true labels—align are *true positives* and *true negatives*; where they do not are *false positives* (model predicts true when it should not) and *false negatives* (vice versa).

A species will be absent at more locations where it is present, and our test dataset will therefore be unbalanced. As such, accuracy (number correct over all points) is a poor metric, as a trivial classifier which always predicts negative will achieve over 99% accuracy for a species occupying less than 1% of its potential range.

$$\text{Precision} = \frac{tp}{tp + fp} \tag{4.8}$$

$$\text{Recall} = \frac{tp}{tp + fn} \tag{4.9}$$

$$\text{Average Precision} = \sum_{t} (R_t - R_{t-1}) P_t \tag{4.10}$$

To address this challenge, we evaluate single-species models on Average Precision (AP; Equation 4.10) and multi-species models on Mean Average Precision (MAP), the average precision normalised over all species. In Equation 4.10, R_t is the recall at threshold t, P_t the precision, where for Equations 4.8 and 4.9, 'tp' is true positives and 'fn' is false negatives. AP is widely used in SDM [96, 27, 64, 24] as it emphasises true positive predictions by considering both precision (ability to identify true positives; Equation 4.8) and recall (proportion of actual positives identified; Equation 4.9). In our implementation we used Scikit Learn's AP method provided in the Python package [70].

For comprehensive evaluation and to measure overfitting, we must combine metric performance with visual map inspection in order to evaluate range maps which are precise and generalised. Model may perform well on generalised metrics such as MAP, but fail to meet key criteria required for ecological purposes. Models must be *specific*– they must learn negative space well–and they must be able to interpolate between groups of locations seen in training.

Chapter 5

Experiments

The accuracy of range maps, fundamental to SDM, is dependant on many factors. This chapter investigates best practices for creating informative range maps through focusing on two key questions:

1. **Optimising Pseudo-Absence Generation**: Given a presence-only dataset, how can we generate the most informative pseudo-absences to train deep models to effectively distinguish from presence and absence areas?
2. **Fine-Tuning Single- and Multi-Species Models**: How can we optimise model performance for single- and multi-species models while avoiding overfitting?

We begin with a focus on our single-species model. We first introduce a representative case study species for model tuning. Next, we investigate pseudo-absence generation methods while maintaining a balanced dataset on single-species model. This is followed by an optimal model depth investigation and an ablation study in order to understand the impact of individual model components on performance.

Following this investigation, we shift to our multi-species model. Here, we examine the effect of increasing the number of randomly generated pseudo-absences and additionally explore the effectiveness of using target-group pseudo-absences instead of—and as well as—random background points.

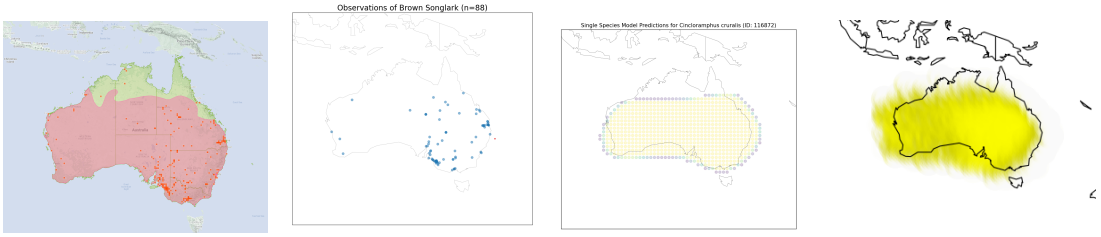
Finally, we analyse why some species perform better than others in our multi-species model. By correlating average precision against several factors, we aim to identify limitations in order to motivate future work to address these limitations.

5.1 Case Study: *Cincloramphus Cruralis (Brown Songlark)*

For initial model investigation, we selected the brown songlark (*Cincloramphus Cruralis*) as a case study species. This medium-sized bird species offers an interesting study due to its average-case distributional characteristics; with only 88 observations, it is in the 21st percentile of all observations in our dataset (Figure 3.2) and is classified in the 'least concern' IUCN classification [7].

The spatial distribution of the songlark's observations poses a common challenge. While ground truth places presence of the songlark over the entire continent, observations are concentrated in two spatially distinct areas on either coastline. This allows us to investigate a model's ability to interpolate between observed locations, acting as a proxy for gaps in observation data in rare locations where it may be valuable e.g., Amazon Rainforest.

Ground Truth, Single-Species, and Multi-Species Predictions for Brown Songlark



(a) Ground Truth. (b) Observations. (c) Single-species model predictions. (d) Multi-species model predictions.

Figure 5.1: IUCN Ground Truth, best single-species model, and best multi-species model predictions for brown songlark.

Building on the ability to assess negative space learning through coastal observations, the east coast's proximity to the high-activity area of New Zealand (Figure 3.3) can help investigate how target-group pseudo-absences influence model performance. If this results in more specific range maps on the east coast compared to the west coast (where no observations for any species are reported beyond Australia, see Figure 5.1b), we may conclude that target-group pseudo-absences assist in learning negative space. In addition to AP, visual inspection of maps to assess these factors will result in a thorough investigation on the drawbacks and benefits of both techniques. Figure 5.1 shows songlark ground truth, observations, and both single- and multi-species model performance on the species.

However, we earlier noted (Section 3.1) that birds are over-represented in our dataset. It is possible that by tuning our models to this species, we bias our models towards better performance on this, and similar, species. This is a common problem in SDM, and emphasises the need to consider low-resource species when developing models, as 'at risk' species will often be the subject of ecological studies. We attempt to avoid this pitfall by carefully measuring overfitting: we observe output range maps and validate on other species at each stage of the process.

5.2 Single-Species Model

This section investigates the performance of our single-species model architecture, introduced in Section 4.2. We train the model on one species at a time using a randomly selected subset of species from our dataset (shown in Table 5.1). This approach allows us to compare the effect of different pseudo-absence generation techniques (addressed first) and later, to compare the single-species model with our multi-species model.

5.2.1 Optimising Single-Species Pseudo-Absences

We focus on the performance of model SS_BN4, which will be introduced in detail in Section 5.2.3. The model, denoted $f_{\theta_{ss}}$ is trained on data pairs consisting of coordinates ($x \in \mathbb{X}$) and presence labels ($\mathbf{y}$) for some N species. To assess the effectiveness of this model, we consider several criteria: learning highly negative space (e.g., accurately predicting absence over oceans), maintaining zone specificity (predicted zones should not extend beyond known presence areas), and overall Average Precision (AP). For baseline comparison, we also train a Random Forest (RF) model and present the results in Table 5.1.

Species	Random	Target-group	Both	RF	Num. Obs.
Brown Songlark†	**0.87**	0.40	0.86	0.45	88
Rufous Woodpecker	0.29	0.16	0.26	**0.42**	1933
Mexican Free–Tailed Bat	0.41	0.10	0.26	**0.54**	1371
White-tailed Mongoose	0.68	0.36	**0.69**	0.51	230
Rainforest Rocket Frog	**0.39**	0.22	0.37	0.25	230
Hong Kong Warty Newt*	**0.54**	0.36	0.45	0.01	508
Subset MAP	**0.46**	0.24	0.41	0.35	-

*Threatened Globally. † Not included in subset MAP.

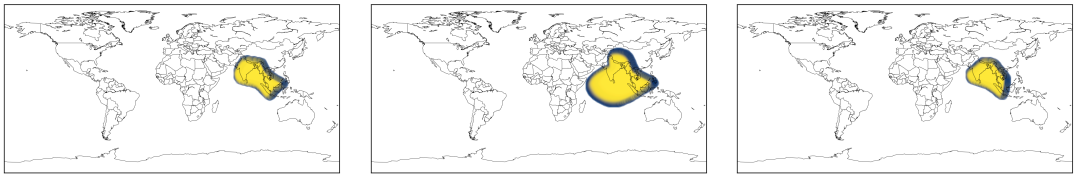
Table 5.1: Average precision for single species models on randomly selected species with varying pseudo-absence generation techniques. Random sampling samples uniformly across the sphere, target-group samples from the presence of other species. Both is a hybrid approach with half random and half target-group pseudo-absences. Brown songlark (case study species) not included in MAP value. RF trained with random background sampling. Bolded best results per species.

Table 5.1 shows that pseudo-absence point selection technique has a significant impact on model performance. In our random subset, target-group sampling significantly underperformed random sampling, with an MAP of 0.24 vs 0.46 for random sampling alone. Furthermore, our hybrid method, which utilised both random and target-group sampling, resulted in worse MAP compared to random sampling. The worse performance of target-group samples implies that information was not gained from the inclusion of target-group pseudo-sampling, even when combined with random sampling. The hybrid approach resulted in similar MAP in most cases, and so we undertake a visual inspection of range maps to determine if there were additional benefits to including target-group points. The range map produced though each sampling technique is seen in Figure 5.2.

In this figure, we see that target-group sampling results in a ‘ballooned’ range map,

where negative space is poorly learned over the ocean. As no presences are reported there, it is not surprising that target-group sampling failed to learn this area. Random and hybrid sampling resulted in similar range maps, though the hybrid approach did generalise better to Indonesian regions; this is represented in MAP of 0.42 vs 0.29 for random sampling.

Rufous Woodpecker SS_BN4 Random, Target-Group, and Hybrid Pseudo-Absence Generated Range Maps



(a) Random sampling (b) Target-group sampling (c) Hybrid sampling

Figure 5.2: Estimated range maps by SS_BN4 for Rufous Woodpecker and each random sampling technique. We note 'ballooned' range map for target species, and in general a lack of specificity for all. This improved slightly in the multi-species case.

The RF model served as a strong baseline results, outperforming all deep models on two occasions, and target-group in all but one occasion. This implies that the deep approach is not always the best one, and suggests perhaps an ensembling modelling approach—where model predictions are combined—can offer more consistent results.

We note that the Hong Kong Warty Newt is a 'near threatened' species [6], but still had an acceptable AP value of 0.54 for SS_BN4 with random sampling. This implies that global rarity is not necessarily a barrier to accurate range map estimation using deep learning. Other factors, such as habitat requirements of a species or quality of the data available may influence model performance. Interestingly, the RF model performed very poorly for this species (AP: 0.01), highlighting the potential advantage of deep learning approaches for low-resource species.

5.2.2 How Good is Random Sampling?

Our best pseudo-sampling method from Section 5.2.1 was random sampling, and we here investigate limitations and range maps produced for each species using this approach.

Figure 5.3 provides visual analysis of models on each species when random sampling was applied, with column (c) showcasing single-species model predictions. We see that the model predicts poor negative space for the Mexican free-tailed bat, which exists over the Indonesian region, and the white-tailed mongoose, which is present across regions in both American continents. Additionally, the model failed to learn connection across the American continents for the bat species, instead discretising predictions to a zone on each continent. This is reflected in poor AP values in Table 5.1, highlighting limitations of the single-species model.

The mongoose model produced an acceptable range map, with an MAP value of 0.69 in the best case (both). Figure 5.3-2c shows that a negative spatial indentation was learned,

and further the west coast of Africa was seen to be accurately bounded. The frog and newt existed in smaller geographical locations compared to the other species, nearby the ocean in both cases. While the model located these areas in both cases, the predicted presence did over encompass the ground truth, an observation which was quantified in the relatively poor performance for the rocket frog model.

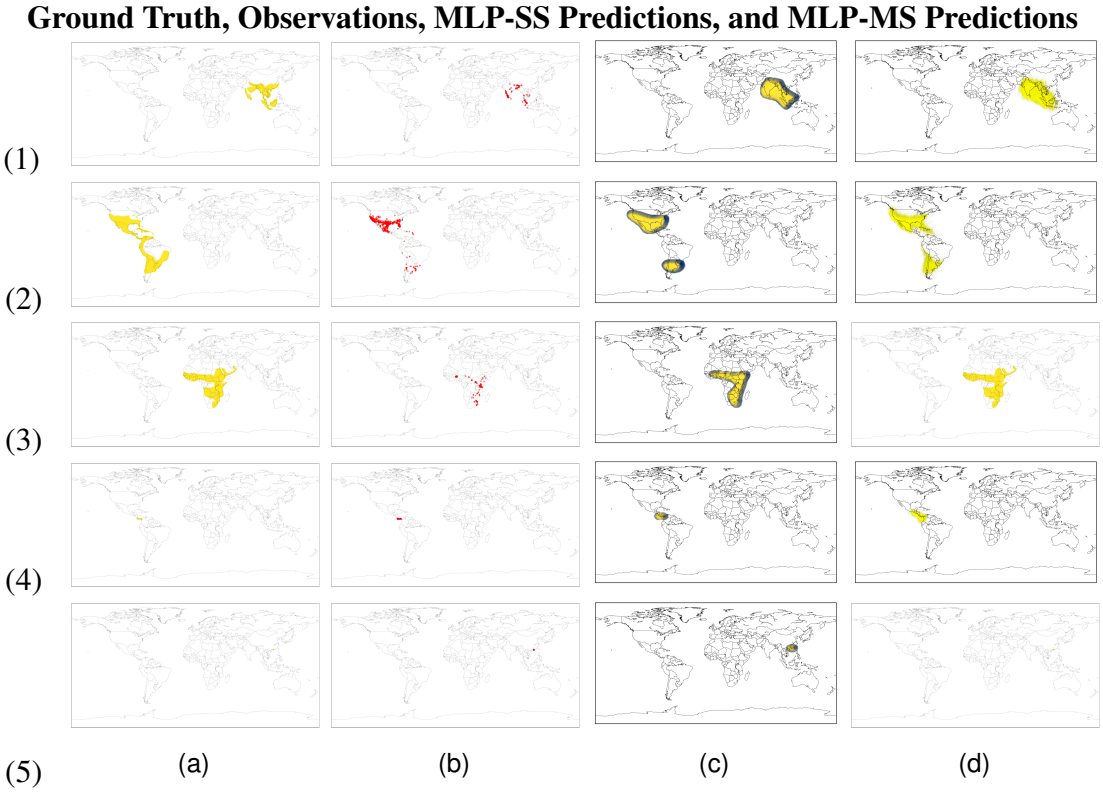


Figure 5.3: (a) Ground truth, (b) species reported observations, (c) SS_4BN range map predictions, (d) and MS_1.75k range map predictions for subset of species studied in Section 5.2, (1) Rufous Woodpecker, (2) Mexican Free-tailed Bat, (3) White-tailed Mongoose, (4) Rainforest Rocket Frog, and (5) Hong Kong Warty Newt.

5.2.3 Hidden Layers: Deep, but not too Deep

To investigate optimal range map production methods, we we investigated the effect of model complexity on single-species model predictions. More complex models are more intensive to train and infer from, so in order to make sure additional complexity results in a performance boost, we here investigate the effect of increasing number of hidden layers. Here and hence we notate a single-species with batch norm as SS_BN and with dropout SS_DO,

Model complexity can be increased by increasing the *depth* of the model—more hidden layers— or increasing the *width*—number of hidden units; both will increase the number of parameters available to a model. We here elect to focus on the result of changing model depth over width as we wish our model to capture underlying features, and by focusing on depth, we can allow our model to abstractly capture trends in stages [89].

Number of Hidden Layers	Average Precision
2	0.909
3	0.919
4	0.**938**
5	0.930
6	0.924

Table 5.2: Single-Species Average Precision for Species Cincloramphus cruralis (ID: 116872) with Dropout and Batch Normalisation.

Model Configuration	Average Precision
BN+Dropout	0.938
BN	**0.949**
Dropout	0.913
None	0.898

Table 5.3: Ablation Study for BN+4HL Model (Species of Interest: Cincloramphus cruralis). BN = batch normalisation; dropout p = 0.1.

Table 5.2 reveals that our single-species model benefits from increased depth up to a maximum of 4 hidden layers. Performance starts to decrease after this point, indicating overfitting. This suggests a limit to the complexity that can be learned from the training data: despite the challenges mentioned in our case study (Section 5.1), the songlark's 88 observations contain only so much extractable information. Attempting to learn beyond this capacity results in overfitting—decreased generalisation performance due to learning particularities specific to training data. An average precision of 0.938, our highest, is on the upper end of all model performance, and so it is likely that once data has been learned well, additional layers—which can learn more complex trends—hinder performance.

Table 5.3 investigates the performance of our model with and without batch normalisation and dropout. We see that implementing dropout and batch normalisation both help improve generalisation performance over the baseline when applied independently and together. Interestingly, BN improves performance more when applied without dropout, potentially indicating that this model does not learn to rely too heavily on individual features and therefore randomly dropping weights is not necessary.

Based on the above configuration, our best model configuration was 4 hidden layers with batch normalisation, which we use in Section 5.1 below.

In general, with optimal tuning there is still only so much that we can draw from a single-species result. Tuning for one species may result in worse performance on the rest, so we note the choices made in this section could have been different had we selected a different case study. Though, we note that we attempted to avoid overfitting and looked to produce a generalisable model.

5.3 Multi-Species Model

In this section, motivated by the importance of background points on model performance, we first investigate varying the number of observations per species on model performance. We then investigate the importance of the type of pseudo-absence methods, testing random sampling, where points are selected uniformly from the globe, target-group, where presences of other species are used, and a combination of both. We here denote MS_*obs* as a multi-species model $f_{\theta_{ms}}$ as described in Chapter 4, where *obs* denotes the number of samples taken per species with replacement.

5.3.1 Optimising Number of Multi-Species Pseudo-Absences

We expect varying the number of background points to have a twofold effect on our model.

First, the number of presences considered for each species are increased. This has the result of learning the positive space very well, but we anticipate overfitting when the number of samples greatly exceeds the number of observations. This may result in models failing to generalise and instead forming localised groups.

Second, our model balances positives and negative observations and so the number of background points are increased: more negatives are considered. In the random case, this has the effect of more comprehensively covering the surface of the globe, but also the byproduct of more negatives impacting positive space. In the target group case, negative locations are reinforced more strongly.

We experiment on the results of using different observations of 100 (baseline), 1k (average case), 1.75k (below max observation oversampling threshold), 2.5k (required oversampling), and 5k (high oversampling).

		Average Precision					
Obs. per species	MAP	Songlark	Woodpecker	Bat	Mongoose	Frog	Newt
100	0.54	0.79	0.39	0.35	0.69	0.17	0.06
1k	0.59	0.90	0.41	0.40	0.85	0.25	0.05
1.75k	0.62	*0.92*	0.44	0.50	0.81	0.24	0.15
2.5k	0.63	0.89	0.49	0.54	0.78	0.15	0.19
5k	0.65	0.89	0.47	0.50	0.70	0.18	0.12
SS_RAND	N/A	0.87	0.29	0.41	0.68	0.39	0.54

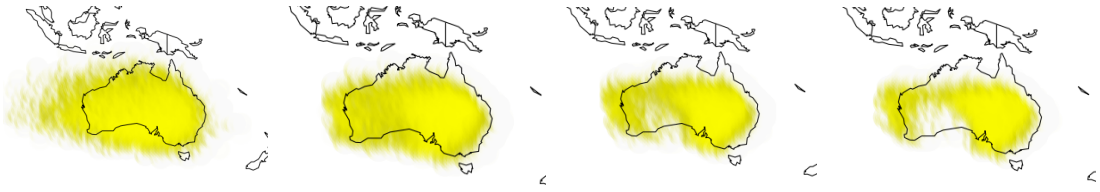
Table 5.4: Multi-species MAP over all species (in contrast to 5 species subset in Section 5.2.1) and AP for select species with varying numbers of observations per species.

Table 5.4 shows that MAP increases with the number of background points up to a maximum of 0.65 MAP for MS_5k observations per species. This is greater than any MAP on the single-species subset, and in addition this value is indicative of the entire training set, making it a more comprehensive measure than AP alone–which is per species—or the subset MAP used in Section 5.2.1—which was an average over only 5 species.

5.3.1.1 Qualitative Analysis

Despite MS_5k having a higher mean average precision (MAP) than others, visual inspection of range maps produced by the model seen in Figure 5.4, suggests that our 5k model overfits on the training data and fails to learn negative space. MS_5k begins to discretise observational groups on each coastline, while ground truth places the species over the entire continent. It is due to this observed overfitting that we use MS_1.75k for future experiments, which Figure 5.4 shows interpolates well over space unseen in training, learns negative space well, and has high MAP.

Ground Truth, Single-Species, and Multi-Species Predictions for Brown Songlark



(a) 100 Observations per Species.

(b) 1.75k Observations per Species.

(c) 2.5k Observations per Species.

(d) 5k Observations per Species.

Figure 5.4: Range maps produced by MS_*obs* for 100, 1.75k. 2.5k, and 5k sampled observations per species.

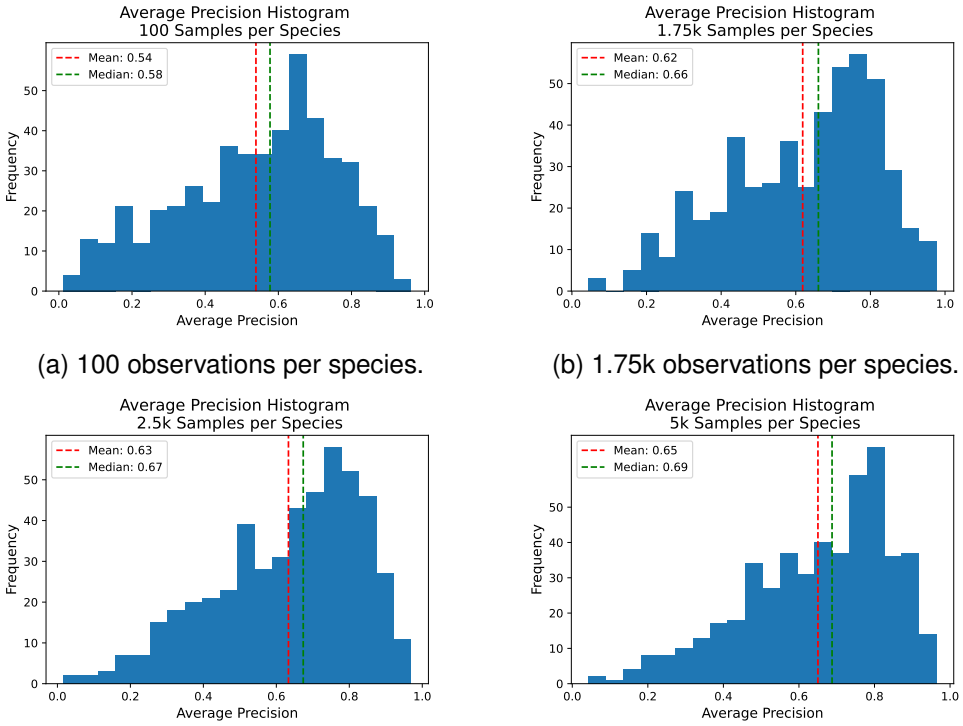
Figure 5.5 exemplifies this observation, as we see AP values become more balanced from Figures 5.4a-5.4c, pooling towards the right hand side of the histogram as number of observations increases. This is another indication of overfitting between 1.75k-2k sampled observations, as the MS model learns the training distribution very well, resulting in outlier-specific gaps in range maps seen in Figure 5.4d.

5.3.1.2 Quantitative Analysis

We conducted an ANOVA [83] test on the distributions of average precision and found differences to be statistically significant ($p = 1.12 \times 10^{-18}$). Investigating pairwise significance using Tukey's HSD pairwise comparison test [94]—a post-hoc test investigating which groups are statistically significant—we found all changes in number of points to have statistically significant differences except between our 2.5k and 5k number of observations models. For these, we hypothesise that the model begins to overfit around 2.5k observations per species, and this is further reinforced for 5k observations. Model-by-model comparisons can be seen in Table 5.5.

Table 5.5 shows that we reject the null hypothesis—that there is no difference between the means of the groups— for all comparisons except between our 2.5k and 5k number of observations models. This means there is evidence to suggest that the means of all other groups are different—that changing the number of points of observation for our groups results in better or worse performing models. This is reinforced in Table 5.4, where the MAP increases as number of observations is increased.

Failing to reject the null hypothesis between MS_2.5k and MS_5k may be the result of many factors, including the greater sample sizes when compared to other groups. The



(a) 100 observations per species.

(b) 1.75k observations per species.

(c) 2.5k observations per species.

(d) 5k observations per species.

Figure 5.5: Average precision of all species for multi-species models trained with both random and target-group pseudo-absence sampling, with varying number of samples per species (includes oversampling when relevant). 1k observations figure omitted (Appendix 7.1).

maximium number of presence points in the training set was 2000, meaning MS_2.5k and MS_5k are guaranteed to oversample (use the same observation more than once) and as such may have lower variance between groups. This is not true of any other of our models, and may be a factor in the absence of statistical difference of the two results. In addition, as our number of observations increases, the true difference of the mean of the AP values produced by models may decrease. This will render them harder to detect statistically, which may result in failing to reject the null.

It is interesting that MS_1k is statistically similar to MS_1.75k but dissimilar to both MS_2.5k and MS_5k, while MS_1.75k shows statistical significance with MS_2.5k and MS_5k. However, statistical significance is not transitive[1]; this result could be due to adjusted p-values (which account for multiple comparisons [18]), sample variability (leading to overlapping confidence intervals), or simply a greater magnitude of differences between MS_1k and MS_1.75k vs MS_1.75k and MS_2.5k.

Despite MS_5k having greater MAP than other models, visual inspection of range maps produced by the model, seen in Figure 5.4d, imply that our 5k model overfits on training data and fails to learn negative space. For our case study example, the brown songlark, MS_5k discretises observational groups on either Australian coastline, while ground

[1] $(A \Rightarrow B) \wedge (B \Rightarrow C) \not\Rightarrow A \Rightarrow C$, or, A implies B and B implies C does not imply A implies C

Group 1	Group 2	Mean Diff.	P-adj	lower	upper	reject
100	1k	0.05	0.00	0.02	0.09	True
100	1.75k	0.08	0.00	0.12	0.04	True
100	2.5k	0.09	0.00	0.06	0.13	True
100	5k	0.11	0.00	0.07	0.15	True
1k	1.75k	0.03	0.23	0.06	0.01	False
1k	2.5k	0.04	0.01	0.01	0.08	True
1k	5k	0.06	0.00	0.02	0.10	True
1.75k	2.5k	0.01	0.86	-0.02	0.05	False
1.75k	5k	0.03	0.12	0.00	0.07	False
2.5k	5k	0.02	0.76	-0.02	0.05	False

Table 5.5: Tukey's HSD pairwise comparisons. p = 1.12×10^{-18}. P-adj = 0.00 if $< 1 \times 10^{-3}$. All values rounded to 3 s.f.

truth places the species over the entire continent.

It is due to factors in our qualitative and quantitative assessments regarding measured and observed overfitting that we use MS_1.75k for future experiments, which interpolates well over space unseen in training, learns negative space well, and has high MAP.

5.3.2 Type of Pseudo-Absences: Multi-Species

In this section we train MS_1.75k, our best generalisation model from Section 5.3.1, using both background and target group point generation independently and together as discussed in Section 4.5. We attempt to investigate the balance of maximising the 'information' content of a background point by investigating the effect of model performance when considering target-group pseudo-absences, whose bias is aligned with that of the species of interest, random pseudo-absences, which are drawn randomly, and a balanced combination of both..

Pseudo-Absence Generation Technique	MAP
Random	0.61
Target-group	0.39
Both	0.62

Table 5.6: MAP scores for different pseudo-absence generation techniques for MS_1.75k.

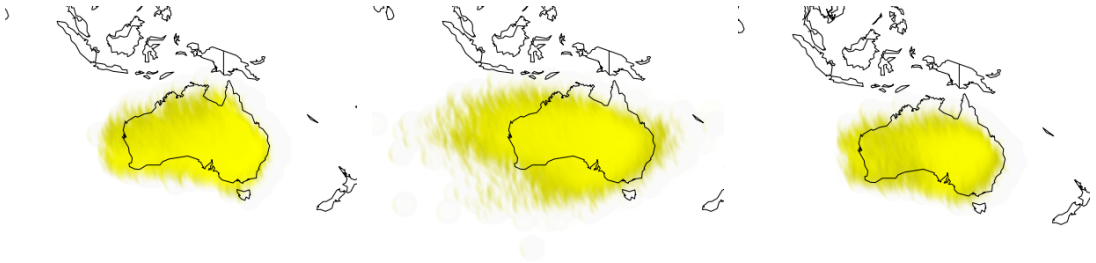
We find that, as in to Section 5.2.1, despite target group points sharing bias of our presence input points, when training on target group points alone the model performs worse than when training with random only or a combination of random and target. In the multi-species case, MAP improves for the hybrid approach by 0.01 MAP, and we perform another Tukey's HSD pairwise comparison test to determine statistical significance, the results of which are seen in Table 5.7.

Group 1	Group 2	Mean Diff.	p-adj	lower	upper	reject
Both	Random	-0.01	0.73	-0.04	0.02	False
Both	Target	-0.23	0.00	-0.26	-0.20	True
Random	Target	-0.22	0.00	-0.25	-0.19	True

Table 5.7: Tukey's HSD pairwise comparisons. p=5.5×10^{-70}. P-adj = 0.00 if $(< 1 \times 10^{-3})$. All values rounded to 3 s.f.

Table 5.7 states that we reject the null hypothesis between both combiations of random-only pseudo-absences and our hybrid approach. Given that the differnces in the distribution is highly likely to appear by chance, we may conclude that the inclusion of target-group points had negligible effect on model performance. The three methods are plotted in Figure 5.6, which reinforces the weaknesses of target-group sampling in its poor negative learned space over the ocean. Hybrid sampling does appear to increase zone specificity on the west coastline, though this may be a stochastic effect and no concrete conclusions may be drawn from this image alone. Thus, our best generalisable multi-species model was MS_1.75k with a hybrid sampling method.

Brown Songlark MS_1.75k Random, Target-Group, and Hybrid Pseudo-Absence Generated Range Maps



(a) Random sampling (b) Target-group sampling (c) Hybrid sampling

Figure 5.6: Estimated range maps by SS_BN4 for Rufous Woodpecker and each random sampling technique. We note 'ballooned' range map for target species, and in general a lack of specificity for all. This improved slightly in the multi-species case.

5.4 Why are some species better than others?

Figure 5.7 shows that there was almost no correlation between number of observations and average precision for MS_1k. This is surprising, as we typically expect a larger number of observations to lead to better model performance. This finding suggests that for multi-species model MS_1k, the quality and distribution of the observations might be more important than the overall quantity.

Species in our dataset with fewer than 2000 observations imply that these are all available observations of that species, whereas greater than 2000—given that this is our observation cap—imply that these are a subset of all observations. As such, species with 2000 observations in our training set may have a more representative distribution across our training set compared to those with fewer. Indeed, the correlation coefficient—a

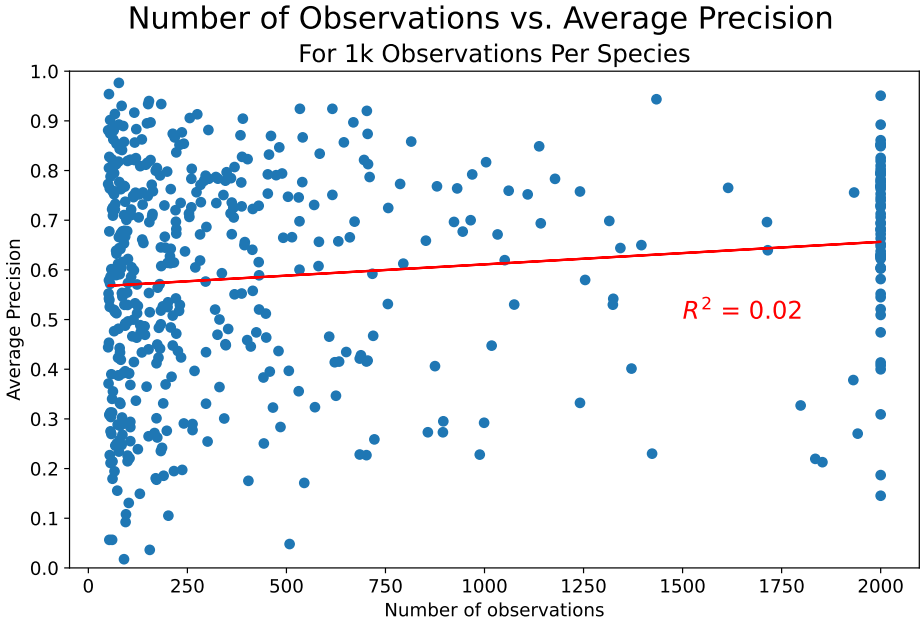


Figure 5.7: Number of Observations plotted against AP with line of correlation drawn (red). $R^2 = 0.02$, implying number of observations do not explain variation of AP about the mean.

measurement of how much variation in the dependent variable (average precision) can be explained by the independent variable (number of observations)— decreases from 0.2 to 0 when species with 2000 observations are removed, implying that number of observations explain no variation in the average precision without this subset.

Our motivation behind producing a multi-species model was the inclusion of joint-modelling, which provides latent information to a model which may increase its performance. An implication of this assumption is that species with fewer observational counts may be paired with and informed by those with greater, reducing potential effects of low observation on model average precision. We see this effect here, showcasing that many observations are not required in order to produce accurate range maps.

However, some species AP values were very low, and as such AP variance must be attributable to other factors, such as geographic location.

5.5 Single- vs Multi-Species Models

We have shown that neither multi-species models and single-species each outperform each other in all scenarios, and so we are left with the question "which one, and when?" The answer is situation-dependent, and we discuss advantages and disadvantages of both in this section.

Which is computationally easier?

Single species models are computationally more efficient to run. Running one inference

at test time for SS_B4 takes 0.82s on a standard computer[2], while in the same situation an MS iteration takes 2.29s.

Training a single-species model is also less resource-intensive. The time taken to train SS_B4 on the same computer was roughly eight times less (198s) compared to MS_1.75k (1531s). This efficiency gap stems from the inherent complexity of multi-species models, which may be a barrier on resource-constrained systems, especially if training with other environmental covariates. This highlights a trade-off within SDM: computationally efficient single-species models versus potentially more comprehensive multi-species models which can capture inter-species interactions.

Which is more consistent?

We showed in Section 5.4 that multi-species models offer consistency and defence against low observer count bias—when predictions for a species are overly influencedby a limited number of observations. As they consider many species simultaneously, multi-species models can utilise the collective strength of the data, comparable to an ensembling approach which combines predictions from multiple sources. In addition, our experiments trained single- and multi-species models with the same number of hidden dimensions, which may have limited the multi-species models from fully exploiting their potential. Doubling the number of hidden units to 512 raises the number of model parameters from 656k to 2.4 million, and with this extra capacity, it is likely that more complex interactions can be learned—if overfitting is avoided.

On the other hand, we observe for certain species, some multi-species performance is very poor. Most of our multi-species models, except MS_1.75k, produced a range map with AP of less than 0.05 for Gran Canaria giant lizard. Despite our balanced sampler ensuring equal representation of all species across a training epoch, the nature of a multi-species model focusing on all species at once may cause it to lose focus on individual species, more so if the loss function takes no measure of individual species into account, only total deviation from the ground truth distribution.

Given the architecture of single-species models, this situation would be less likely to occur. We see in Figure 5.2 that, although range maps produced by single species models are less specific—additional negative space is considered where it is guaranteed to be incorrect (over water)—in no observed cases was performance as poor as the worst-case scenario for the multi-species model. It is possible that a species not located by our experiments would have this result in the single-species case, but it is also feasible that—by focusing on one species only—the lower bound of performance is raised, at least for species with minimum 50 observations.

Interestingly, not all multi-species models had poor performance on *Gallotia stehlini*. MS_1.75k had an AP value of 0.58 MAP for this species, indicating good knowledge of the species' spatial distribution. This effect could be contributed to stochasticicity during training time, where the number of examples seen (particularly in early epochs) influences final performance. Our exclusion of batch normalisation supports this hypothesis, as despite not seeing the requirement for it during training, by not including it we allowed possible internal covariate shift, meaning a specific species showing up

[2]Intel(R) Core(TM) i7-10510U CPU @ 1.80GHz, 2304 Mhz, 4 Cores, 8 Logical Processors

early in training could have a large impact on downstream performance. The MAP value—which takes an average across all species—between each model in Section 5.3 implies that while performance varied per model, in general performances were within range of each other and so could be that benefits in certain areas of the model necessarily result in worse performance in others. Whether this could be improved by increasing model capacity is left for future work, but it is clear that there is no one solution to this problem, which is true in machine learning across the field.

Single-species vs. Multi-species: A nuanced choice

In summary, multi-species models demonstrate consistency and defense against low observer count bias by utilising collective data strength, albeit with the potential risk of poorer performance on individual species due to loss of focus. Increasing model capacity may help mitigate this issue, but achieving balance between species-specific performance and overall model efficacy remains a challenge, and choice of which model to select depends on the task at hand. We summarise our results in Table 5.8.

Factor	**Single-Species Model**	**Multi-Species Model**
Computational Efficiency	More efficient	Less efficient
Consistency	Less consistent	More consistent
Focus on individual sp.	More focused	Less focused
Potential for bias	More susceptible	Less susceptible

Table 5.8: Summary of trade-offs between single- and multi-species models.

Chapter 6

Conclusions

This chapter reviews the aims of the project and describes how they were achieved. In summary, this dissertation has detailed two deep learning model architecture designs, implementations, and investigations into optimal pseudo-absence generation techniques for optimal model performance.

The first aim of this project was to develop a single-species model architecture, which could produce accurate range maps given presence-only observations. This aim was achieved through a custom feedforward fully-connected multi-layer perceptron, which generated absences and produced accurate and precise range maps.

The second aim was to develop this architecture into a multi-label scenario, where instead of modelling for a single-species, a prediction was made for all species simultaneously. This aim was achieved by adapting the single-species architecture to handle the increased complexity and dimensional by the multi-label case. By producing accurate and precise maps for unseen data, we increased confidence in our model results, which can be used to estimate future species spatial distributions for novel species.

The final aim was to investigate the effect of different pseudo-absence generation techniques on both single- and multi-species model performance. By adapting absence sampling or using a custom loss function, both architectures were developed to perform supervised learning with random and target-group hybrid pseudo-absences, as well as a combination of both. For each pseudo-absence generation technique, we performed a thorough investigation into the effects on both model architectures, finding a balanced dataset with random-only pseudo-absences to be the most effective technique for single-species, and a hybrid approach for multi-species.

6.1 Related Work

Barbet-Massin et al. [13] explored background point generation techniques and their affect on models, but utilised simulated species for the task—which may miss underlying real-world complexities—and considered only ‘shallow’ models. In this study, we explore similar generation techniques, but focus specifically on MLP models and

compare between single-species and multi-species. We also consider adaptations to the architecture which benefit model predictions (Section 4.5).

Cole et al. [27] investigate the effect of varying the number of pseudo-absences, but at lower thresholds than we investigate in this study and without balancing the dataset. This study addresses these limitations by examining a wider range of thresholds with a balanced sampler. Few studies investigate combination techniques for pseudo-absence generation methods, providing inspiration for the study of them in this paper. Other work [64] investigates the effect of including other cofactors in the input dimension, where we limit ourselves to coordinates to isolate the specific effect of pseudo-absence generation techniques.

6.2 Limitations and Future Work

Species distribution modelling is intrinsically a low-resource problem, and machine learning models' fundamental data requirement is undeniably a barrier to producing effective range maps for underrepresented or endangered species. Our minimum presences count was 50, which neglects species whose observation count fall below this threshold. Future work will explore methods to improve model performance for low-resource species, which could involve incorporating known species presence areas to generate additional data points. By creating more accurate range maps for these species, they can be trusted as independent data sources to inform policy decisions.

Another limitation of our study lies in how we treated presence observations. We took each observation to be equal as an assumption for training our model, when in reality, some presence records might represent occurrences in suboptimal areas due to factors like resource scarcity or forced movement. This could lead to our model overestimating the suitability of certain areas or underestimating the potential for marginal areas. Future work could explore incorporating weighted presence reports on habitat suitability.

Our models focus on a global scale, which can provide insightful and large-scale distributional trends, but this may result in less accurate predictions on smaller scales, which are often more relevant for conservation efforts. Policy decisions typically occur at local or regional levels, and these may require information about specific environmental variables or habitat features that might not be captured in a global model. For example, incorporating data on local land-use patterns or specific prey availability could improve model accuracy at a regional scale by increasing information given to a model. Future work could explore techniques to bridge global and local scale approaches, allowing models which provide both large-scale trends and fine-grained detail.

Finally, Our models considered spatial distributions only. By disregarding temporal aspect of SDM, we limit the effectiveness of our method in certain downstream tasks, for example environmental policy decision or migration analysis.

Bibliography

[1] . iNaturalist. Birds (Class Aves). URL https://www.inaturalist.org/taxa/3-Aves. Accessed 31/03/2024.

[2] . iNaturalist. Mammals (Class Mammalia). URL https://www.inaturalist.org/taxa/40151-Mammalia. Accessed 31/03/2024.

[3] . iNaturalist. Reptiles (Class Reptilia). URL https://www.inaturalist.org/taxa/26036-Reptilia. Accessed 31/03/2024.

[4] . iNaturalist. Vertebrates (Subphylum Vertebrata). URL https://www.inaturalist.org/taxa/355675-Vertebrata. Accessed 31/03/2024.

[5] . iNaturalist. Available from https://www.inaturalist.org. Accessed on: 10-03-24.

[6] . IUCN. 2023. The IUCN Red List of Threatened Species. Version 2023-1. https://www.iucnredlist.org. Accessed on 22-03-24.

[7] . BirdLife International. 2016. Cincloramphus cruralis. The IUCN Red List of Threatened Species 2016: e.T22715511A94456565. https://dx.doi.org/10.2305/IUCN.UK.2016-3.RLTS.T22715511A94456565.en. Accessed on 27 March 2024.

[8] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438, 2015.

[9] Riccardo Albertoni, Sara Colantonio, Piotr Skrzypczyński, and Jerzy Stefanowski. Reproducibility of machine learning: Terminology, recommendations and open issues, 2023.

[10] Margaret E Andrew and Elizabeth Fox. Modelling species distributions in dynamic landscapes: The importance of the temporal dimension. *Journal of Biogeography*, 47(7):1510–1529, 2020.

[11] Massimo Aria, Corrado Cuccurullo, and Agostino Gnasso. A comparison among interpretative proposals for random forests. *Machine Learning with Applications*, 6:100094, 2021.

[12] Robert A Barber, Stuart G Ball, Roger KA Morris, and Francis Gilbert. Target-group backgrounds prove effective at correcting sampling bias in maxent models. *Diversity and Distributions*, 28(1):128–141, 2022.

[13] Morgane Barbet-Massin, Frédéric Jiguet, Cécile Hélène Albert, and Wilfried Thuiller. Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in ecology and evolution*, 3(2):327–338, 2012.

[14] Linda J Beaumont, Erin Graham, Daisy Englert Duursma, Peter D Wilson, Abigail Cabrelli, John B Baumgartner, Willow Hallgren, Manuel Esperón-Rodríguez, David A Nipperess, Dan L Warren, et al. Which species distribution models are more (or less) likely to project broad-scale, climate-induced shifts in species ranges? *Ecological Modelling*, 342:135–146, 2016.

[15] Jan Beck, Marianne Böller, Andreas Erhardt, and Wolfgang Schwanghart. Spatial bias in the gbif database and its effect on modeling species' geographic distributions. *Ecological Informatics*, 19:10–15, 2014.

[16] Sara Beery, Elijah Cole, Joseph Parker, Pietro Perona, and Kevin Winner. Species distribution modeling for machine learning practitioners: A review, 2021.

[17] Sara Beery, Elijah Cole, Joseph Parker, Pietro Perona, and Kevin Winner. Species Distribution Modeling for Machine Learning Practitioners: A Review, July 2021. URL `http://arxiv.org/abs/2107.10400`. arXiv:2107.10400 [cs, stat].

[18] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[19] Juliette Boiffin, Vincent Badeau, and Nathalie Bréda. Species distribution models may misdirect assisted migration: insights from the introduction of douglas-fir to europe. *Ecological Applications*, 27(2):446–457, 2017.

[20] Dale Bowman. App-solutely enhancing outdoors experience: Inaturalist, seek and merlin bird id apps; plus stray cast, Jul 2021. URL `https://chicago.suntimes.com/2021/7/8/22568173/appsolutely-enhancing-outdoors-experience-inaturalist-seek-merlin-bird-id`

[21] Sérgio Branco, André G Ferreira, and Jorge Cabral. Machine learning in resource-scarce embedded systems, fpgas, and end-devices: A survey. *Electronics*, 8(11): 1289, 2019.

[22] Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

[23] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[24] Philipp Brun, Thomas Kiørboe, Priscilla Licandro, and Mark R Payne. The predictive skill of species distribution models for plankton in a changing climate. *Global change biology*, 22(9):3170–3181, 2016.

[25] Anaïs Charbonnel, Patrick Lambert, Géraldine Lassalle, Eric Quinton, Antoine Guisan, Lise Mas, Guillaume Paquignon, Marie Lecomte, and Marie-Laure Acolas. Developing species distribution models for critically endangered species using participatory data: The european sturgeon marine habitat suitability. *Estuarine, Coastal and Shelf Science*, 280:108136, 2023.

[26] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[27] Elijah Cole, Grant Van Horn, Christian Lange, Alexander Shepard, Patrick Leary, Pietro Perona, Scott Loarie, and Oisin Mac Aodha. Spatial implicit neural representations for global-scale species mapping, 2023.

[28] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. Supervised learning. In *Machine learning techniques for multimedia: case studies on organization and retrieval*, pages 21–49. Springer, 2008.

[29] Daniele Da Re, Enrico Tordoni, Jonathan Lenoir, Jonas J Lembrechts, Sophie O Vanwambeke, Duccio Rocchini, and Manuele Bazzichetto. Use it: Uniformly sampling pseudo-absences within the environmental space for applications in habitat suitability models. *Methods in Ecology and Evolution*, 14(11):2873–2887, 2023.

[30] Fabio Suzart de Albuquerque, Blas Benito, Paul Beier, Maria José Assunção-Albuquerque, and Luis Cayuela. Supporting underrepresented forests in mesoamerica. *Natureza & Conservação*, 13(2):152–158, 2015.

[31] Paul B De Laat. Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? *Philosophy & technology*, 31 (4):525–541, 2018.

[32] Glenn De'ath. Boosted Trees for Ecological Modeling and Prediction. *Ecology*, 88(1):243–251, 2007. ISSN 0012-9658. URL `https://www.jstor.org/stable/27651085`. Publisher: Ecological Society of America.

[33] Grace J Di Cecco, Vijay Barve, Michael W Belitz, Brian J Stucky, Robert P Guralnick, and Allen H Hurlbert. Observing the observers: how participants contribute data to inaturalist and implications for biodiversity science. *BioScience*, 71(11):1179–1188, 2021.

[34] Hugh Dingle. *Migration: the biology of life on the move*. Oxford University Press, USA, 2014.

[35] Sehaba Mohammed El Amine, Crispim-Junior Carlos, and Tougne Rodet Laure. Embedded plant recognition: a benchmark for low footprint deep neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 670–677, 2023.

[36] J. Elith, J. R. Leathwick, and T. Hastie. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4):802–813, 2008. ISSN 1365-2656. doi: 10.1111/j.1365-2656.2008.01390.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2656.2008.01390.x`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2656.2008.01390.x.

[37] Jane Elith and John R Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics*, 40:677–697, 2009.

[38] Jane Elith, Catherine Graham, Roozbeh Valavi, Meinrad Abegg, Caroline Bruce, Simon Ferrier, Andrew Ford, Antoine Guisan, Robert J Hijmans, Falk Huettmann, et al. Presence-only and presence-absence data for comparing species distribution modeling methods. *Biodiversity informatics*, 15(2):69–80, 2020.

[39] Yoan Fourcade, Jan O Engler, Dennis Rödder, and Jean Secondi. Mapping species distributions with maxent using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. *PloS one*, 9 (5):e97122, 2014.

[40] Janet Franklin. Species distribution modelling supports the study of past, present and future biogeographies. *Journal of Biogeography*, 50(9):1533–1545, 2023. doi: https://doi.org/10.1111/jbi.14617. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jbi.14617.

[41] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[42] William Godsoe and Luke J Harmon. How do species interactions affect species distribution models? *Ecography*, 35(9):811–820, 2012.

[43] Melanie Gogol-Prokurat. Predicting habitat suitability for rare plants at local spatial scales using a species distribution model. *Ecological Applications*, 21(1): 33–47, 2011.

[44] Sunila Gollapudi. *Practical machine learning*. Packt Publishing Ltd, 2016.

[45] Forough Goudarzi, Mahmoud-Reza Hemami, Mansoureh Malekian, Sima Fakheran, and Fernando Martínez-Freiría. Species versus within-species niches: a multi-modelling approach to assess range size of a spring-dwelling amphibian. *Scientific Reports*, 11(1):597, January 2021. ISSN 2045-2322. doi: 10.1038/s41598-020-79783-0. URL https://www.nature.com/articles/s41598-020-79783-0. Number: 1 Publisher: Nature Publishing Group.

[46] Charlène Guillaumot, Alexis Martin, Marc Eléaume, and Thomas Saucède. Methods for improving species distribution models in data-poor areas: example of sub-antarctic benthic species on the kerguelen plateau. *Marine Ecology Progress Series*, 594:149–164, 2018.

[47] Antoine Guisan, Reid Tingley, John B Baumgartner, Ilona Naujokaitis-Lewis, Patricia R Sutcliffe, Ayesha IT Tulloch, Tracey J Regan, Lluis Brotons, Eve McDonald-Madden, Chrystal Mantyka-Pringle, et al. Predicting species distributions for conservation decisions. *Ecology letters*, 16(12):1424–1435, 2013.

[48] Trevor Hastie and Robert Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.

[49] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[50] Douglas M Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12, 2004.

[51] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[52] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[53] Tomislav Hengl, Henk Sierdsema, Andreja Radović, and Arta Dilo. Spatial prediction of species' distributions from occurrence-only records: combining point pattern analysis, enfa and regression-kriging. *Ecological modelling*, 220 (24):3499–3511, 2009.

[54] Hartwig H Hochmair, Rudolf H Scheffrahn, Mathieu Basille, and Matthew Boone. Evaluating the data quality of inaturalist termite records. *PLoS One*, 15 (5):e0226534, 2020.

[55] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[56] Christine Howard, Philip A Stephens, James W Pearce-Higgins, Richard D Gregory, and Stephen G Willis. Improving species distribution models: the value of data on abundance. *Methods in Ecology and Evolution*, 5(6):506–513, 2014.

[57] Maialen Iturbide, Joaquín Bedia, Sixto Herrera, Oscar del Hierro, Miriam Pinto, and Jose Manuel Gutiérrez. A framework for species distribution modelling with improved pseudo-absence generation. *Ecological Modelling*, 312:166–174, 2015.

[58] Benjamin Kellenberger, Elijah Cole, Diego Marcos, and Devis Tuia. Training techniques for presence-only habitat suitability mapping with deep learning. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 5085–5088. IEEE, 2022.

[59] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[60] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324, 1998.

[61] Julia S Stewart Lowndes, Benjamin D Best, Courtney Scarborough, Jamie C Afflerbach, Melanie R Frazier, Casey C O'Hara, Ning Jiang, and Benjamin S Halpern. Our path to better science in less time using open data science tools. *Nature ecology & evolution*, 1(6):0160, 2017.

[62] Lisha Lyu, Flurin Leugger, Oskar Hagen, Fabian Fopp, Lydian M. Boschman, Joeri Sergej Strijk, Camille Albouy, Dirk N. Karger, Philipp Brun, Zhiheng Wang, Niklaus E. Zimmermann, and Loïc Pellissier. An integrated high-resolution

mapping shows congruent biodiversity patterns of Fagales and Pinales. *The New Phytologist*, 235(2):759–772, July 2022. ISSN 1469-8137. doi: 10.1111/nph.18158.

[63] Oisin Mac Aodha, Rory Gibb, Kate E Barlow, Ella Browning, Michael Firman, Robin Freeman, Briana Harder, Libby Kinsey, Gary R Mead, Stuart E Newson, et al. Bat detective—deep learning tools for bat acoustic signal detection. *PLoS computational biology*, 14(3):e1005995, 2018.

[64] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-Only Geographical Priors for Fine-Grained Image Classification, October 2019. URL `http://arxiv.org/abs/1906.05272`. arXiv:1906.05272 [cs].

[65] Charles J Marsh, Yoni Gavish, Mathias Kuemmerlen, Stefan Stoll, Peter Haase, and William E Kunin. Sdm profiling: A tool for assessing the information-content of sampled and unsampled locations for species distribution models. *Ecological Modelling*, 475:110170, 2023.

[66] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[67] Giang Nguyen, Stefan Dlugolinsky, Martin Bobák, Viet Tran, Álvaro López García, Ignacio Heredia, Peter Malík, and Ladislav Hluchỳ. Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey. *Artificial Intelligence Review*, 52:77–124, 2019.

[68] K Denise Kendall Niemiller, Mark A Davis, and Matthew L Niemiller. Addressing 'biodiversity naivety'through project-based learning using inaturalist. *Journal for Nature Conservation*, 64:126070, 2021.

[69] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[70] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[71] Steven J Phillips, Miroslav Dudík, Jane Elith, Catherine H Graham, Anthony Lehmann, John Leathwick, and Simon Ferrier. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological applications*, 19(1):181–197, 2009.

[72] PlantNET. The nsw plant information network system, 2020.

[73] Laura J Pollock, Reid Tingley, William K Morris, Nick Golding, Robert B O'Hara, Kirsten M Parris, Peter A Vesk, and Michael A McCarthy. Understanding co-occurrence by modelling species simultaneously with a joint species

distribution model (jsdm). *Methods in Ecology and Evolution*, 5(5):397–406, 2014.

[74] David Quesada, Maykel Cruz-Monteagudo, Terace Fletcher, Aliuska Duardo-Sanchez, and Humbert González-Díaz. Complex networks and machine learning: from molecular to social sciences, 2019.

[75] Tom Radomski, David Beamer, Alan Babineau, Christa Wilson, Joseph Pechmann, and Kenneth H Kozak. Finding what you don't know: testing sdm methods for poorly known species. *Diversity and Distributions*, 28(9):1769–1780, 2022.

[76] Christophe F Randin, Thomas Dirnböck, Stefan Dullinger, Niklaus E Zimmermann, Massimiliano Zappa, and Antoine Guisan. Are niche-based species distribution models transferable in space? *Journal of biogeography*, 33(10): 1689–1703, 2006.

[77] Satyendra Singh Rawat and Amit Kumar Mishra. Review of methods for handling class-imbalanced in classification problems. *arXiv preprint arXiv:2211.05456*, 2022.

[78] Ramiro Rico-Martínez, Ioannis G. Kevrekidis, and Raymond A. Adomaitis. Noninvertibility in neural networks. *IEEE International Conference on Neural Networks*, pages 382–386 vol.1, 1993. URL `https://api.semanticscholar.org/CorpusID:18477952`.

[79] Rafael Masson Rosa, Daniel Caracanhas Cavallari, and Rodrigo Brincalepe Salvador. inaturalist as a tool in the study of tropical molluscs. *PLoS One*, 17(5): e0268048, 2022.

[80] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[81] David Sánchez-Fernández, Jorge M Lobo, and Olga Lucía Hernández-Manrique. Species distribution models that do not incorporate global data misrepresent potential distributions: a case study using iberian diving beetles. *Diversity and Distributions*, 17(1):163–171, 2011.

[82] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? *Advances in neural information processing systems*, 31, 2018.

[83] Henry Scheffe. *The analysis of variance*, volume 72. John Wiley & Sons, 1999.

[84] Valerie Sessions and Marco Valtorta. The effects of data quality on machine learning algorithms. pages 485–498, 01 2006.

[85] Adam B Smith. On evaluating species distribution models with random background sites in place of absences when test presences disproportionately sample suitable habitat. *Diversity and Distributions*, 19(7):867–872, 2013.

[86] Alexandre Somavilla, Raimundo Nonato Martins Moraes Junior, Marcio Luiz de Oliveira, and José Albertino Rafael. Biodiversity of insects in the amazon: survey

of social wasps (vespidae: Polistinae) in amazon rainforest areas in amazonas state, brazil. *Volume 67, Issue 2, June 2020, Pages 312-321*, 2020.

[87] Diana L Soteropoulos, Caitlin R De Bellis, and Theo Witsell. Citizen science contributions to address biodiversity loss and conservation planning in a rapidly developing region. *Diversity*, 13(6):255, 2021.

[88] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research (JMLR)*, 2014.

[89] Peter Stone and Manuela Veloso. Layered learning. In *European conference on machine learning*, pages 369–381. Springer, 2000.

[90] AT Strathdee and JS Bale. Life on the edge: insect ecology in arctic environments. *Annual review of entomology*, 43(1):85–106, 1998.

[91] Brian L Sullivan, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, and Steve Kelling. ebird: A citizen-based bird observation network in the biological sciences. *Biological conservation*, 142(10):2282–2292, 2009.

[92] Wilfried Thuiller, Bruno Lafourcade, Robin Engler, and Miguel B Araújo. Biomod–a platform for ensemble forecasting of species distributions. *Ecography*, 32(3):369–373, 2009.

[93] Gleb Tikhonov, Øystein H Opedal, Nerea Abrego, Aleksi Lehikoinen, Melinda MJ de Jonge, Jari Oksanen, and Otso Ovaskainen. Joint species distribution modelling with the r-package hmsc. *Methods in ecology and evolution*, 11 (3):442–447, 2020.

[94] John W Tukey. *The relationship between sets of observations with an extension to several groups of correlations*. McGraw-Hill, 1951.

[95] Roozbeh Valavi, Jane Elith, José J. Lahoz-Monfort, and Gurutzeta Guillera-Arroita. Modelling species presence-only data with random forests. *Ecography*, 44(12):1731–1742, December 2021. ISSN 0906-7590, 1600-0587. doi: 10.1111/ecog.05615. URL https://onlinelibrary.wiley.com/doi/10.1111/ecog.05615.

[96] Roozbeh Valavi, Gurutzeta Guillera-Arroita, José J. Lahoz-Monfort, and Jane Elith. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecological Monographs*, 92(1):e01486, 2022. ISSN 1557-7015. doi: 10.1002/ecm.1486. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ecm.1486. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ecm.1486.

[97] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[98] Peter Vermeiren, Peter Reichert, and Nele Schuwirth. Integrating uncertain prior knowledge regarding ecological preferences into multi-species distribution

models: Effects of model complexity on predictive performance. *Ecological Modelling*, 420:108956, 2020. ISSN 0304-3800. doi: https://doi.org/10.1016/j.ecolmodel.2020.108956. URL https://www.sciencedirect.com/science/article/pii/S0304380020300284.

[99] Alexander Vezhnevets and Olga Barinova. Avoiding boosting overfitting by removing confusing samples. In *Machine Learning: ECML 2007: 18th European Conference on Machine Learning, Warsaw, Poland, September 17-21, 2007. Proceedings 18*, pages 430–441. Springer, 2007.

[100] Dani Villero, Magda Pla, David Camps, Jordi Ruiz-Olmo, and Lluís Brotons. Integrating species distribution modelling into decision-making to inform conservation actions. *Biodiversity and Conservation*, 26:251–271, 2017.

[101] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable ai: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, pages 563–574. Springer, 2019.

[102] Jianjin Xu and Zhanxing Zhu. Learning rate schedules for faster stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2022.

[103] Robin Zbinden, Nina van Tiel, Benjamin Kellenberger, Lloyd Hughes, and Devis Tuia. On the selection and effectiveness of pseudo-absences for species distribution modeling with deep learning, 2024.

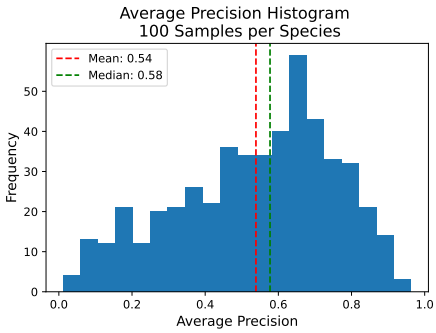
[104] Robin Zbinden, Nina van Tiel, Marc Rußwurm, and Devis Tuia. Imbalance-aware presence-only loss function for species distribution modeling, 2024.

[105] Niklaus E Zimmermann, Thomas C Edwards Jr, Catherine H Graham, Peter B Pearman, and Jens-Christian Svenning. New trends in species distribution modelling. *Ecography*, 33(6):985–989, 2010.

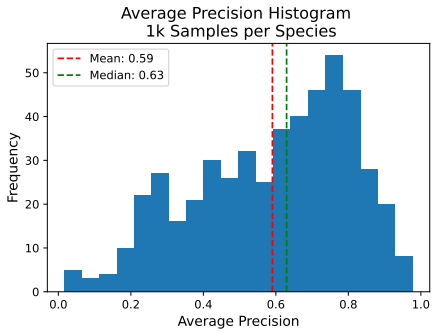
Chapter 7

Appendix

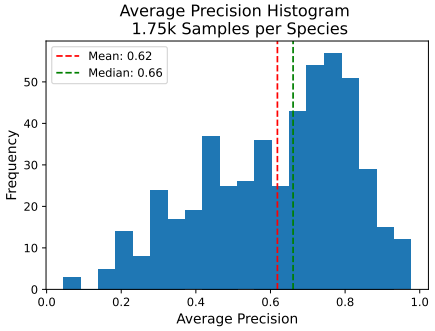
7.1 Omitted Graphs



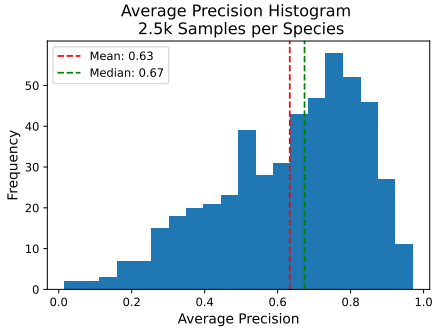
(a) 100 observations per species.



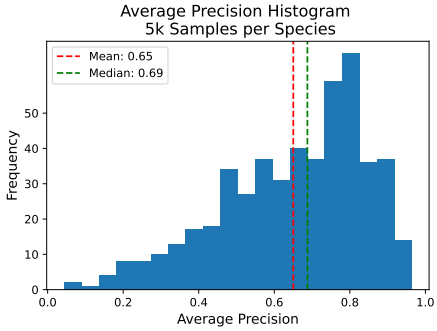
(b) 1k observations per species.



(c) 1.75k observations per species.



(d) 2.5k observations per species.



(e) 5k observations per species.

Figure 7.1: Average precision of all species with varying number of samples per species (includes oversampling when relevant). Includes Figure excluded in main body.