Predicting IUCN conservation status through machine learning: a case study on reptiles

Nádia Filipa de Jesus Soares
Mestrado em Ciência de Dados
Departamento de Ciência de Computadores
2021

Orientador
Rita Paula Almeida Ribeiro, Professora Auxiliar, Faculdade de Ciências da Universidade do Porto

Coorientadores
João Francisco Gonçalves, Investigador Doutorado, Centro de Investigação em Biodiversidade e Recursos Genéticos

Raquel Vasconcelos, Investigadora Doutorada, Centro de Investigação em Biodiversidade e Recursos Genéticos

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

Abstract

Halting biodiversity loss is the ultimate challenge of our era. We are losing species at a high rate, even before assessing their extinction risk. The International Union for Conservation of Nature (IUCN) Red List is the most complete assessment of species conservation status, yet it only covers around 138,374 out of about 1.2 million species identified so far. Additionally, many of those assessments are outdated, due to the ever-evolving nature of taxonomy or lack of resources. These assessments rely on long, mostly manual processes performed by experts. Species conservation would gain by automating the identification of species where experts and financing should focus on.

This work proposes a pipeline to derive datasets out of open data and obtain conservation status predictions, through machine learning, of Not Evaluated (NE) and Data Deficient (DD) species. This pipeline was applied to different groups within the Reptilia class, one of the most under-assessed taxonomic group of vertebrates, and the performance of different machine learning algorithms was compared using five datasets. The first included only two variables (the most used in IUCN assessments) – extent of occurrence (EOO) and area of occupancy (AOO). The second used only ecological variables, the third used all available variables, and the fourth and fifth variations of the latter two using variable selection. Additionally, SMOTE was used to balance the datasets where the class of interest was the minority, and improve the results for the species in that class. The predictions were mapped to locate threatspots and needed conservation assessments, and to infer global patterns.

The results showed that EOO and AOO are the most relevant predictors of threat status achieving good results when used alone in models. Nevertheless, most groups benefited from the inclusion of ecogeographical variables together with those two. Random Forest was the best method for most groups and variable selection improved results. The predictions, when compared to the most recent assessments, achieved good sensitivity and specificity results. The threatspot maps showed that the areas with threatened species may be larger than previously thought, especially in Africa and Brazil.

Overall, results support that the developed pipeline can be used to perform a general model assisted conservation status assessment across species to more efficiently allocate conservation efforts, which more focused studies can then complement.

Resumo

Travar a perda de biodiversidade é o grande desafio da nossa era. Estamos a perder espécies a um ritmo elevado, ainda antes de conseguirmos avaliar o seu risco de extinção. A Lista Vermelha da União Internacional para a Conservação da Natureza e dos Recursos Naturais (IUCN) é a avaliação mais completa do estado de conservação das espécies, no entanto ainda abrange apenas 138,374 das cerca de 1.2 milhões de espécies identificadas até à data. Adicionalmente, muitas dessas avaliações estão desastualizadas devido à permanente evolução da taxonomia e à falta de recursos. Elas dependem de processos demorados, e maioritariamente manuais, realizados por especialistas. A conservação de espécies ganharia com a automatização da identificação de espécies nas quais os especialistas e o financiamento se devam focar.

Neste trabalho, é proposta uma *pipeline* para gerar conjuntos de dados a partir de fontes abertas e obter previsões do estado de conservação, através de aprendizagem automática, para espécies Não Avaliada (NE) e Informação Insuficiente (DD). Esta *pipeline* foi aplicada a diferentes grupos dentro da classe Reptilia, um dos grupos taxonómicos de vertebrados com menos avaliações, e comparou-se a performance de diferentes algoritmos de aprendizagem automática utilizando cinco datasets. O primeiro incluiu apenas duas variaveis (as mais utilizadas nas avaliações da IUCN) - extensão de ocorrência (EOO) e área de ocupação (AOO). O segundo usou apenas variáveis ecológicas, o terceiro usou todas as variáveis disponíveis e o quarto e quinto eram variações dos últimos dois, usando seleção de variáveis. Foi ainda usado SMOTE para balancear os datasets onde a classe de interesse era a minoritária e melhor os resultados para espécies dessa classe. As previsões foram mapeadas para localizar *threatspots* e a necessidade de avaliação de estado de conservação e para inferir padrões globais.

Os resultados mostraram que o EOO e AOO são as variáveis mais relevantes para prever o estado de ameaça, tendo tido bons resultados quando usados como únicas variáveis nos modelos. Ainda assim, a maioria dos grupos beneficiou da inclusão de variáveis ecogeográficas. *Random Forest* foi o algoritmo com melhor performance para a maioria dos grupos e a seleção de variáveis melhorou os resultados. As previsões, quando comparadas com as avaliações de estado de conservação mais recentes, obtiveram boa sensibilidade e especificidade. Os mapas de *threatspots* mostraram que as áreas com mais espécies ameaçadas podem ser mais extensas do que se pensava, especialmente em África e no Brasil.

No geral, os resultados suportam que a *pipeline* desenvolvida pode ser utilizada para realizar

uma avaliação genérica e automática do estado de conservação de múltiplas espécies, de forma a permitir uma alocação mais eficiente dos esforços de conservação, que estudos mais restritos podem depois complementar.

Agradecimentos

Agradeço todo o apoio dos meus orientadores. Sem os seus conhecimentos nas diversas áreas de de estudo que este trabalho abrange, ele teria sido possível.

À minha família, à mãe e ao meu marido, em particular, pela paciência e compreensão durante este ano. À Inês e à Bárbara.

Contents

Abstract i

Resumo ii

Agradecimentos iv

Contents vii

List of Tables ix

List of Figures xi

Acronyms xii

1 Introduction 1

1.1 Motivation and context 1

1.1.1 The reptile case study 2

1.1.2 Data science for conservation 3

1.2 Objectives 3

1.3 Organization of the dissertation 4

2 Background and literature review 6

2.1 Background knowledge 6

2.1.1 IUCN Red List 6

2.1.2 Reptile taxonomy 8

2.2 Machine learning . . . 9
2.2.1 Problem setup . . . 9
2.2.2 Imbalanced domain learning . . . 10
2.2.3 Evaluation metrics . . . 12
2.2.4 Performance estimation . . . 13
2.3 State of the art . . . 14

3 Dataset curation **16**
3.1 Dataset of presences . . . 16
3.1.1 Importing and merging the source data . . . 16
3.1.2 Pre-processing . . . 17
3.1.3 Handle species with a low count . . . 18
3.2 Select and generate ecological variables . . . 18
3.2.1 Relevant ecological variables . . . 18
3.2.2 Extract and pre-process . . . 19
3.2.3 Variables selection . . . 20
3.3 Species-level dataset . . . 21
3.4 Experimental datasets . . . 24

4 Experimental study **28**
4.1 Modelling pipeline . . . 28
4.1.1 Predictive modelling . . . 29
4.1.2 Threat status predictions . . . 30
4.1.3 Maps of threatspots . . . 30
4.2 Experimental setup . . . 31

5 Results and discussion **35**
5.1 Results . . . 35
5.1.1 Variable importance . . . 37

5.1.2 Handling Imbalance . . . 38
5.1.3 Predictions for DD and NE Species . . . 39
5.1.4 Maps of threatspots . . . 41
5.2 Discussion . . . 42
5.2.1 Predictive modelling . . . 42
5.2.2 Species conservation . . . 45

6 Conclusions **46**
6.1 Contributions . . . 47
6.2 Limitations and future work . . . 47

A Model results **49**

B Bibliographic note **56**

References **58**

List of Tables

3.1 List of the geographic and ecological variables used in this study by type, source, name and description. All raster layers used for the ecological variables have a 1 km resolution. 22

3.2 List of ecological variables (EcoV) selected for each of the taxonomic groups by the PCA step of the pipeline. Check Table 3.1 for details on the variables. 23

3.3 Number of presences, total number of species (Tsp) with presence data, number of species (Nsp) with threat status available at IUCN, and ecogeographical variables (Nvr) for each group, the imbalance ratio (IR), calculated as the size of minority class over the size of majority class, and the minority class of each reptile group. 25

3.4 Dataset characterization containing the number of training examples (species with threat status available at IUCN), and variables for each dataset. The imbalance ratio (IR), calculated as the size of minority class over the size of majority class, and the minority class of each reptile group are shown for context. 27

4.1 Parameter values of methods used in grid search for tuning. 33

5.1 Best model trained for each group and their predictions for Data Deficient (DD) and Not Evaluated (NE) species. The table shows the total number of species (Tsp) with presence data, number of species (Nsp) with threat status available at IUCN, and ecogeographical variables (Nvr), the number of unassessed and total taxa predicted to be threatened, and the percentage of threatened taxa in respect to Tsp, the dataset and method that performed better for each group, selected based on $F_{0.5}$ results, and their performance estimates with respect to $F_{0.5}$, AUC and TSS metrics. The imbalance ratio (IR), calculated as the size of minority class over the size of majority class, and the minority class of each reptile group (non-threatened, nT; threatened, T) is also given. 37

5.2 Results comparing the two version of the best algorithm for each of the groups where threatened was the minority class, one using the dataset without SMOTE and another with SMOTE. Values marked with a bold color represent models where performance, was, significantly, improved relative to the original dataset (without SMOTE), used as the baseline. Values of the Wilcoxon tests marked with one or two asterisks (*, **) indicate significance values higher than 95%, or 99%, respectively. 39

5.3 List of errors made by the best model for the amphisbaenians, crocodiles, lizards, snakes and turtles datasets. False negatives (top) are species evaluated as non-threatened (nT, this is, Near Threatened, NT or Least Concern, LC) which were later assessed as threatened (T; this is, Endangered, EN or Vulnerable, VU), and false positives (bottom) vice-versa. Species that were false negatives/positives in both the all species and group-specific models were marked in bold. 40

5.4 List of errors made by the best model for all reptile species dataset. False negatives (top) are species evaluated as non-threatened (nT, this is, Near Threatened, NT or Least Concern, LC) which were later assessed as threatened (T; this is, Endangered, EN or Vulnerable, VU), and false positives (bottom) vice-versa. Species that were false negatives/positives in both the all species and group-specific models were marked in bold. 41

A.1 Results of final models for each combination of model and dataset, and per evaluation metric. Values marked with a bold color represent models where performance, for that particular metric and algorithm, was according to the Wilcoxon test, significantly, improved relative to the *AOO_EOO* dataset. Values marked with one or two asterisks (*, **) indicate a confidence level higher than 95%, or 99%, respectively. 50

List of Figures

2.1 IUCN categories. All existent species, are subsetted into categories with an IUCN evaluation of their conservation status or under Not Evaluated (NE). Among the evaluated, some species lack appropriate data for a risk assessment, and thus fall under Data Deficient (DD). Categories with adequate data have an associated extinction risk level. Adapted from [33]. 7

2.2 Criteria used by IUCN to place species under a category. Adapted from [33]. . . 8

2.3 Taxonomy used to classify all existing species. Adapted from [34]. 9

2.4 Confusion matrix in the context of threatened/non-threatened species. 12

2.5 Example of a ROC curve. 13

3.1 Example of the difference between EOO and AOO. (a) shows the spatial distribution of presences of a given taxon. (b) shows the minimum convex polygon which can be used to estimate EOO, and (c) shows how AOO can be estimated using the sum of the occupied grid squares. Adapted from [33]. 24

3.2 The workflow used for selecting the ecological and geographic variables for each dataset. This was applied to each of the groups and resulted in two datasets per group: one with a set of ecological variables specifically picked for the group using PCA, plus all geographic variables, and one that further drills down on the most relevant variables using recursive feature elimination (RFE). 26

4.1 Diagrams of the modelling pipeline showing the six high-level steps implemented. Namely, three steps for the dataset curation, resulting in several labeled datasets at the species level, containing different sets of variables; the training and evaluation of the models and the generation of outputs. 29

5.1 Performance of the models trained for each combination of taxonomic group, method and dataset, according to three metrics: ($F_{0.5}$, AUC and TSS. n stands for sample sizes. Each dot represents the mean of the results of a model across the 10 folds of the 2x5-fold cross validation setup. The standard deviation is represented by the vertical line. Results of models shown in an opaque color significantly outperformed the same algorithm over the *AOO_EOO* dataset. 36

5.2 Variable importance plot for the group that included all species, using the *All EGVs* dataset. The plot shows the cumulative variable importance across five distinct algorithms, Random Forest (RF), XGBoost, C5.0, GLM, and rpart, shown in different colors. Only the top ten variables obtained with each of the five methods were considered. On the variable names, P stands for precipitation and T for temperature. 38

5.3 Threatspot gap maps showing the predicted areas with threatened species (in red), versus the areas occupied by species already classified as threatened by IUCN (in blue). Darker colours indicate higher species richness. Maps for (a) amphisbaenians, (b) crocodiles, (c) lizards, (d) snakes, (e) turtles, and (f) all species, respectively. Some small details referent to predicted (red outline) and known (blue outline) threatened species are highlighted using zoom. 43

5.4 New threatspot maps showing the ratio of the sum of the known and predicted by the total of species in the group dataset. Darker colours indicate higher ratios of threatened species. Maps for (a) amphisbaenians, (b) crocodiles, (c) lizards, (d) snakes, (e) turtles, and (f) all species, respectively. 44

B.1 Proof of submission to the DSAA conference. 56

B.2 Proof of submission to the MEE journal. 57

Acronyms

AOO Area of occupancy

AUC Area Under the ROC Curve

CBD Convention on Biological Diversity

COP Conference of the Parties

CR Critically Endangered

CV Cross Validation

DCC Departamento de Ciência de Computadores

DD Data Deficient

EBV Essential Biodiversity Variable

EGV Ecogeographical Variables

EN Endangered

EOO Extent of occurrence

EW Extinct in the Wild

EX Extinct

FCUP Faculdade de Ciências da Universidade do Porto

GBIF Global Biodiversity Information Facility

GDP Gross Domestic Product

GEF Global Environment Facility

GLM Generalized Linear Model

IPBES Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services

IPCC Intergovernmental Panel on Climate Change

IUCN International Union for Conservation of Nature

k-NN k-Nearest Neighbors

LC Least Concern

LPI Living Planet Index

LOOCV Leave One Out Cross Validation

NE Not Evaluated

NGO Non-Governmental Organisations

NN Neural Network

NT Near Threatened

PACA Preliminary Automated Conservation Assessments

PCA Principal Component Analysis

RF Random Forest

RFE Recursive Feature Elimination

ROC Receiver Operating Characteristic

SDG Sustainable Development Goal

SDM Species Distribution Model

SMOTE Synthetic Minority Over-sampling Technique

SVM Support Vector Machine

TSS True Skill Statistic

UN United Nations

UNDP United Nations Development Programme

UNEP United Nations Environment Programme

UNFCCC United Nations Framework Convention on Climate Change

VU Vulnerable

Chapter 1

Introduction

This first chapter explains the motivation and context for this work, from the arguments that make this such a relevant area of study in the current days, to the choice of Reptiles as the case study for the dissertation and why the usage of data and machine learning techniques was a good fit for this application, in Section 1.1. It then details the objectives for this dissertation, in Section 1.2, and the organization of the rest of the document, in Section 1.3.

1.1 Motivation and context

Species have been disappearing at an alarming rate [1]. As more studies focus on biodiversity loss and the conservation status of different groups of species, the more we understand the true dimension of this problem [2]. Biodiversity loss is causing several emerging problems to humankind. Health issues [3], reduced food security [4, 5], increased contact with diseases [6], and more unpredictable weather events [7] have been reported to be be related to the biodiversity loss and estimated to be worsened by it in the future. There are also economic costs related to impacts of biodiversity loss in ecosystem services, including pollination [8], irrigation, soil reclamation, pest control, to name a few [9]. The value of global biodiversity and related benefits for human economies and livelihoods has been estimated in the trillions of dollars [9].

In the face of global environmental change, it is vital to understand how much biodiversity is being lost. Knowing which species are most at risk of becoming extinct is essential to guide decision making and to establish priorities for conservation efforts and resource allocation, allowing to more efficiently place the time of researchers and the money of countries and institutions to the species that most need them [10–12]. Not knowing the conservation status of a species may hinder the urgent need for action and miss out on opportunities and resources when devising policies [13]. Additionally, assessments must be reevaluated periodically to account for changes in the conservation status and estimate the effect of conservation measures [14].

There is a growing awareness of the importance of policies that can compensate for the current loss of biodiversity and prevent future losses. Multiple entities have recognized that

and have established sets of goals. This includes the agenda set by the Parties of the United Nations (UN) Convention on Biological Diversity (CBD), which defined the 20 Aichi Biodiversity Targets [15], and the 17 Sustainable Development Goals (SDGs) identified by the UN 2030 Agenda for Sustainable Development [16]. The Global Environment Facility (GEF), United Nations Development Programme (UNDP) and United Nations Environment Programme (UNEP) committed to support governments implement the post-2020 global biodiversity framework [17]. Moreover, in December 2020, a workshop organized by Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) and Intergovernmental Panel on Climate Change (IPCC) explored the interaction and synergies between climate and biodiversity. The goals was for its scientific outcome to provide an input into IPCC and IPBES assessments, and into discussions on the Conference of the Parties (COP) 15 of CBD and COP 26 of United Nations Framework Convention on Climate Change (UNFCCC) [18].

Organizations struggle to generate assessments for the vast biodiversity on the planet. These assessments need to be performed species by species, in a very rigorous and time-consuming way, and by experts [12]. The International Union for Conservation of Nature (IUCN) is the largest organization tracking conservation assessments [19], yet it only covers around 138,374 [20] out of the approximate 1.2 million species identified so far, due to the multitude of existing species and the time and monetary costs of acquiring and analysing data manually, species by species [12]. Moreover, there is not only the need to establish the conservation status of more species, but to do re-assessments in order to keep information up to date, and to evaluate the results of conservation actions. Currently, many of the existing assessments are outdated, either due to the ever-evolving nature of taxonomy, or to the lack of periodic reassessments [14]. This increases the need for fast, frequent and reliable assessment methods. Despite being highly unlikely that automated methods will substitute a full field assessment done by IUCN experts, they can be a valuable tool for faster decision making and prioritization, by generally improving the knowledge on a broader number of species. Having a better distribution of the available resources, could lead to a more focused effort in attributing threat levels to unassessed species and more resources could potentially be spent to reevaluate older assessments.

1.1.1 The reptile case study

The tetrapod taxonomic group is the superclass that aggregates all four-limbed animals, including amphibians, reptiles, birds, and mammals. Global distribution [21] and drivers of threat of tetrapods [22] have been largely assessed. Conservation assessments have typically used distributions of amphibians, birds and mammals, however, the global distribution of reptiles is still poorly known despite the high diversity of this class of terrestrial vertebrates [21], making them one of the most under-assessed group of tetrapods [12]. This may be problematic since hotspots of reptile species, such as some endemic lizards, have little overlap with those of other taxonomic groups, biasing conservation priorities [21]. For instance, reptiles may be more vulnerable to habitat degradation, since they have a limited dispersal ability when compared to birds and some mammals, making them more particularly susceptible to habitat changes [23].

Moreover, reptiles depend on the environmental temperature to regulate internal temperature and many species depend on it for sex determination. As such, due to global climate change there is a growing concern that reptiles will be heavily affected by range contractions and the destabilization of sex ratios [24]. As reptiles play an important role in their ecosystems, helping to keep them healthy and balanced by serving, for instance, as predators and preys, pollinators and seed dispersers, assessing their conservation status urges [25–28].

As of May 2021, there were 11,570 reptile species described in The Reptile Database [29]. According to IUCN [20], 9,100 reptile species were evaluated so far.

1.1.2 Data science for conservation

Massive amounts of data have been collected on species occurrence, abundance, and their habitats [30]. In addition, environmental data on the climate, geomorphology, vegetation and land cover, offer insights about species ecological requirements, their habitats, and their spatiotemporal change. Biologists use it to study species evolution, habitat distribution, perform conservation studies, among other things. Historically, this data is analyzed by species, or by taxonomic groups, in order to keep the amount of data manageable using analytic techniques. Data science and machine learning are providing new tools and techniques to natural sciences to develop strategies to address the biodiversity crisis [31]. By providing faster ways of exploring these data and taking advantage of its volume, these methods may provide more accurate predictions of biodiversity loss [31]. With them we can try to speed up or complement the analysis and prediction processes that biologists and conservationists aim to perform, uncover patterns, improve the understanding of biological or ecological processes and obtain faster predictions.

A characteristic of this data is its heterogeneity, since it comes from a variety of sources, namely, it was collected for different studies, by a multitude of individuals in separate teams, and across a span of many years. That leads to a number of considerations and possible biases that need to be payed attention to when merging the different data sources. A careful curation process needs to happen before it can be used in modelling.

1.2 Objectives

This study aimed at taking advantage of the vast amount of data on species presences to predict the conservation status of unassessed species at a global scale, using several machine learning techniques.

Since data is scattered across many public repositories, the first goal was to create a valid and representative dataset for as many reptile species as the current data sources allowed. That is in itself a valuable deliverable since it implies a careful curation of data from a variety of sources. The result was a representative dataset of non-correlated reptile data points and environment

information that can be used for multiple future studies.

This dissertation aimed at developing a process and a working pipeline for data preparation, modelling and evaluation. The intended result was to model the conservation status of species, using terrestrial reptiles as a case study. This taxonomic group was used to demonstrate the validity of the method, without loss of generality, due to the large number of unassessed species in this group and the vast amount of presence data, so that the resulting models could be used to generate predictions for species not yet evaluated.

Since not all of the IUCN categories have a big enough number of species with available data, this dissertation formulates the modelling task as a binary classification problem, where species were labeled as either threatened or non-threatened according to IUCN categories. Additionally, modelling was done for different taxonomic groups within the Reptilia class in order to better model their characteristics and needs. To understand which algorithms perform better in this data, different methods were tested.

An additional goal was to understand if the use of ecological and geographical variables, not typically used in IUCN assessments, benefits the classification of species into threatened or non-threatened. This information can potentially be used to improve models by including relevant predictors, or to understand if species could be assessed using other variables as surrogates when the Area Of Occupancy (AOO) and the Extent Of Occurrence (EOO) are unavailable. For that, five sets of predictor variables were compared for each group of species to evaluate the impact of using ecogeographical variables (EGV), both by themselves and with other predictors more commonly used in IUCN assessments.

The best models for each group were used to generate predictions for the unassessed species. Since the beginning of this work, IUCN continued the assessment of reptile species. In order to understand how the models performed, their predictions were matched with the latest IUCN assessments, to compare them with the threat status for the species where a manual assessment was recently made available.

Another goal was to try to improve the model results using a pre-processing technique, namely SMOTE [32], to balance the datasets where the class of interest was the minority class.

To have a geographic view of the results, this study mapped the areas with more threatened species (based on model predictions and IUCN assessments), to potentially guide better conservation planning and resource allocation by IUCN, conservationists, researchers, and the involved countries.

1.3 Organization of the dissertation

The remaining of this dissertation is organized as follows. Chapter 2 presents some preliminary concepts on IUCN assessment criteria and categories and on the biologic taxonomy, along with several relevant machine learning concepts needed for understanding this work and reviews the

state of the art of studies conducted in this area. Chapter 3 describes the steps done during the curation of the datasets to gather and pre-process the occurrence data from multiple sources, aggregating it into a species dataset, and presents the five experimental datasets generated. Chapter 4 introduces the pipeline produced in this dissertation to building and evaluating the model and generating threatspot maps and explains the experimental setup used in the modelling process. Chapter 5 presents and discusses the results and, finally, Chapter 6 concludes by exploring the outcomes of this work, its limitations and potential future improvements and applications.

Chapter 2

Background and literature review

This chapter introduces a set of concepts needed to better understand the work done in this dissertation, in Section 2.1, namely it presents the IUCN Red List and its different categories and criteria, and it explains the basic concepts about reptile taxonomy. Then, in Section 2.2 the fundamental machine learning concepts used on this dissertation are introduced and explained. Finally, Section 2.3 introduces several works that preceded this dissertation in applying machine learning or other techniques to automatically place species under a threat level.

2.1 Background knowledge

2.1.1 IUCN Red List

The International Union for Conservation of Nature's Red List of Threatened Species is the most complete source of information on the conservation status of species, and it serves as an indicator on the state of biodiversity [19]. At the time of writing, 138,374 species had been assessed, including vertebrates, invertebrates, plants, fungi, and others, 28% of which are threatened with extinction. IUCN's goal is to reach 160.000 assessed species with its Barometer of Life project [20]. Besides the assessment itself, the list contains information on range, population size, habitat and ecology, use and/or trade, threats, and conservation actions of each species. It is used by government agencies, wildlife departments, conservation-related Non-Governmental Organisations (NGOs), natural resource planners, and many others, to inform and improve policy and conservation decisions, and for education purposes. Using a method other than the IUCN Red List for a study like this, would incur the risk of having limited impact in policy-making and definition conservation actions, since this is the *de facto* standard.

2.1.1.1 Categories

The IUCN assessment consists in a set of nine categories, seven of which representing different levels of extinction risk. In decreasing level of concern: Extinct (EX), Extinct in the Wild (EW), Critically Endangered (CR), Endangered (EN), Vulnerable (VU), Near Threatened (NT) and Least Concern (LC). Figure 2.1 shows this categories, how they are grouped and whether they are associated to a risk level.

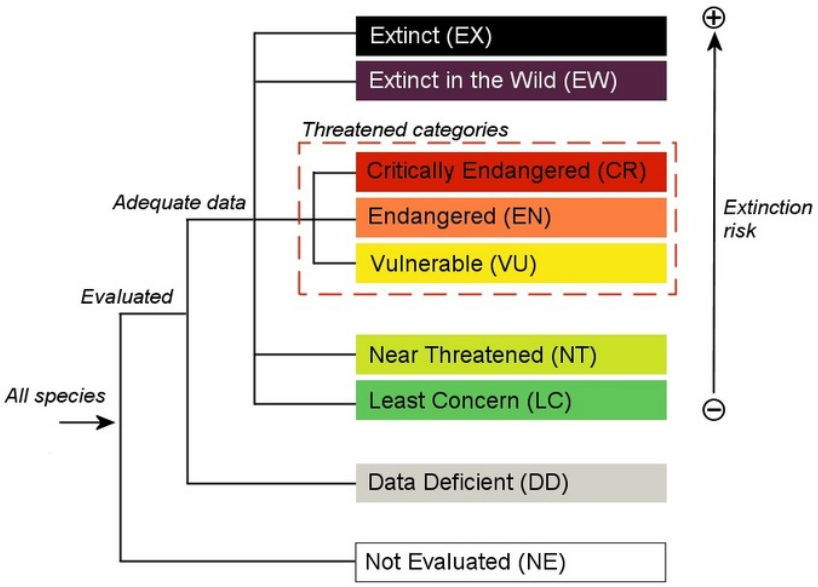


Figure 2.1: IUCN categories. All existent species, are subsetted into categories with an IUCN evaluation of their conservation status or under Not Evaluated (NE). Among the evaluated, some species lack appropriate data for a risk assessment, and thus fall under Data Deficient (DD). Categories with adequate data have an associated extinction risk level. Adapted from [33].

Extinct is the highest level of risk and indicates an assumed global extinction of the species. Extinct in the Wild means that the species no longer exists in its natural habitat, but is present in captivity and may still be reintroduced in the wild, or outside its historic range, as a naturalized population, due to massive habitat loss. Species under Critically Endangered (CR), Endangered (EN) or Vulnerable (VU) categories are considered as Threatened, while Near Threatened (NT) and Least Concern (LC) are considered Non-Threatened species.

Data Deficient (DD) includes species that were evaluated, but there was not enough data to place them under a leveled category. DD and Not Evaluated (NE) represent species were there is no assessment of their conservation status. It is not known whether they are threatened or not, so IUCN advises caution to assume they may have some risk of extinction. Species may be very close to extinction, but due to lack of data or formal evaluation, fall into one of these two

categories [33].

2.1.1.2 Criteria

IUCN Red List uses five criteria (A-E) which are meant to be usable and comparable across different taxa. This comparability ensures proper prioritization decisions and standardizes the process. Criteria take into account many factors including the number of populations, population trends, fragmentation level, EOO, AOO, number of locations, and probability of becoming EW (for which data is rarely available) [33].

Criterion A is based on estimates of population (number of mature individuals) reduction over 10 years or three generations, Criterion B is based on geographic range, Criterion C is based on population size, Criterion D mainly concerns very small or restricted populations based on the number of mature individuals and AOO, and Criterion E is based on a quantitative analysis of extinction probability within a given number of years.

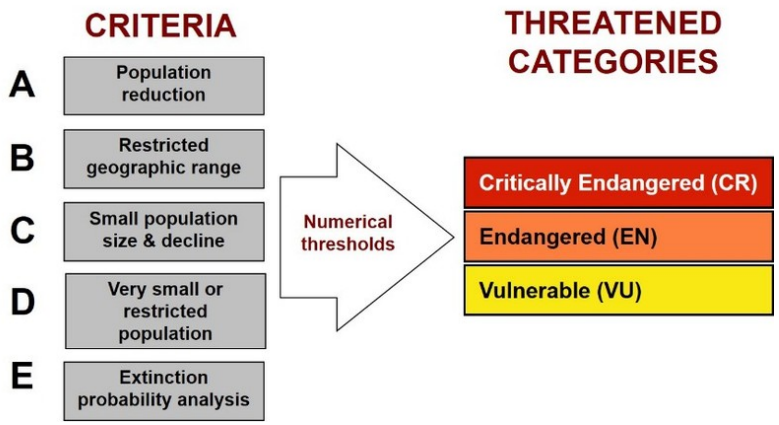


Figure 2.2: Criteria used by IUCN to place species under a category. Adapted from [33].

2.1.2 Reptile taxonomy

In Biology, species are organized according to a taxonomy (Figure 2.3), which organizes them in the tree of life according to their common characteristics and ancestors.

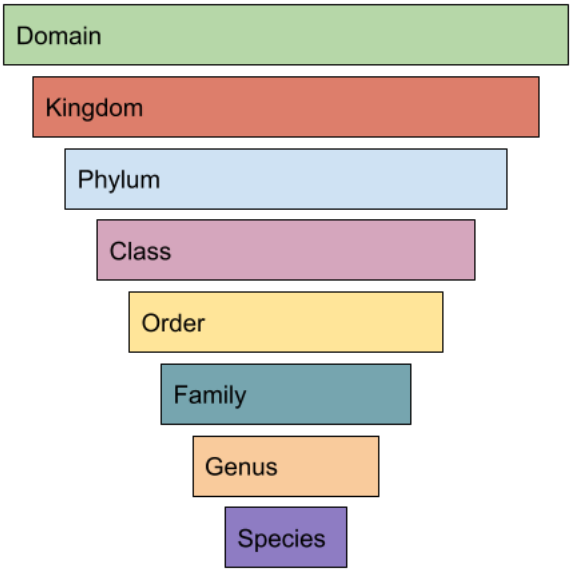


Figure 2.3: Taxonomy used to classify all existing species. Adapted from [34].

The reptilia class, to which all species used in this dissertation are part of, is a class within the tetrapod superclass. Tetrapods refers to all four-limbed animals, and includes other four classes — amphibians, reptiles, birds, and mammals. Within the Reptilia class, there are six distinct clades: amphisbaenians, crocodiles, lizards, snakes, turtles, and tuataras.

2.2 Machine learning

2.2.1 Problem setup

Typical machine learning problems may fall under one of four types of tasks, namely, supervised learning, semi-supervised learning, unsupervised learning, and reinforcement learning [35].

Supervised learning consists of learning a function that maps inputs to outputs using a set of labeled training examples [36]. In some domains, labeled data is hard to obtain; it may require trained humans work to label it, physical experiments, or simply be too expensive or time consuming. Semi-supervised learning uses small set of labeled data and large set of unlabeled data to train models [35]. The premise is that, semi-supervised algorithms, can make usage of more data by using some unlabeled examples, which are typically easier and less expensive to obtain than labeled data. Unsupervised learning uses only unlabeled data and tries to find patterns on the data and categorize the inputs based on those patterns [36]. Essentially, they try to discover the inherent structure of unlabeled data. The output of these models is not as easy to control as supervised learning models since there is no external guidance on the task that should be performed. For instance, they may separate the examples by any of their characteristics to maximizes the distance between clusters, but that may not be the most relevant split for the

task at hand. Lastly, reinforcement learning creates agents that interact with their environment and receive feedback from their actions. Their goal is to maximize a cumulative reward [37].

The problem this dissertation proposed to solve was framed as a supervised learning problem, so a dataset of labeled data had to be put together. Moreover, it was set as a classification problem since the target variable, the conservation status of species, was categorical. Due to the low number of records and imbalance of some of species under the IUCN categories, there was too few data points available to train the models to discriminate between the five categories of risk (CR, EN, VU, NT, and LC). For this reason, it was turned into a binary classification problem by binarizing the target variable into threatened and non-threatened classes. The datasets used contain a number of species (rows) and their their characteristics as features and threatened/non-threatened label as target variable.

2.2.2 Imbalanced domain learning

In some domains it is typical for distribution of the target variable to be skewed. In the case of classification problems, this leads to a much higher number of examples of one class (majority class) than the other (minority class). This becomes an issue when the class with fewer examples is precisely the one which we are interested in. For example, when trying to predict whether a transaction is fraudulent, or a patient has a given disease, there are typically much fewer examples of fraudulent transactions, and sick patients than normal ones, providing fewer training examples of that class for the models to learn from. Ultimately, this yields to models that are biased for the prediction of the most frequent class, which according to standard evaluation metrics (e.g. accuracy), maximizes its performance over all the target variable domain. Since we are interested in the rare positive cases, the cost of misclassification of these cases is typically larger. Moreover, depending on the application domain, sometimes a higher cost is assigned to positive cases incorrectly classified as negatives (false negatives — FN) than negative cases incorrectly classified as positives (false positives — FP). This setting makes standard machine learning methods and evaluation metrics not suitable for tackling these type of problems.

These are called imbalanced domain learning problems [38] and four types of solutions have been proposed: via data pre-processing, special-purpose learning methods, prediction post-processing and hybrid methods.

Data pre-processing applies transformations to change the data distribution and make learning algorithms focus on the cases of the most relevant class. Since they are applied to the data, they have the advantage of working with any existing learning algorithm, the reason why this is the most common approach for tackling imbalanced domains. Also, the models trained using the pre-processed data are biased towards the user's goal, which should make the model more interpretable with relation to those goals. A drawback is the difficulty to map the original distribution to one that fulfills the user's goals. Examples of pre-processing approaches include re-sampling the data to force the model to focus on the most important examples, and setting different weights to the training data in order to penalize the misclassification of important

examples. Within re-sampling, three important strategies should be highlighted: random under-sampling [39], random over-sampling [39] and SMOTE [32]. The first randomly removes examples of the majority class, reducing the sample size. However, this may cause relevant examples to be dropped from the dataset and, consequently, can lead to a worse performance [38]. The second creates new examples by copying those of the minority class, increasing the sample size. This technique may lead to overfitting, specially when using high over-sampling rates [40]. Finally, SMOTE, which stands for Synthetic Minority Over-sampling Technique, incorporates over-sampling of the minority class by generating new synthetic data with a percentage of random under-sampling of the majority class defined by a parameter. The synthetic examples are interpolated at a random point on the lines that join two existing data examples - a seed and one of its k minority class nearest neighbours. A given over-sampling percentage defines the number of examples to be generated for each example of the minority class [32].

Special-purpose learning methods are adaptations of existing algorithms to make them able to learn from imbalanced data. These techniques try to incorporate the involve user's goals into the model itself, which should create more comprehensible models. A clear disadvantage of these methods is that it restricts the choice of learning algorithms, since they need to be modified to optimise the goal or developed from scratch for the task. Moreover, changes to the target loss function force the model to be retrained and, possibly, the algorithm to be further adapted. The developer needs to have a deep understanding of the learning algorithms to be adapted. In the literature there are multiple examples of algorithms adapted to deal with imbalance data (decision tree [41], Support Vector Machine (SVM) [42], RF [43], k-NN [44]).

Post-processing approaches alter the resulting predictions instead of the data distribution or learning methods. It has several advantages: the needed bias does not need to be known at learning time or can be easily changed later without adjusting the incoming data or retraining the models. The trained model can be applied in different deployment scenarios, and standard learning algorithms can be used. However, they are not as interpretable and do not reflect the user preferences as the other methods. Two examples are the threshold method [41], which involves generating multiple models by varying the threshold for examples to fall under one class or the other and picking the most performant of the models, and cost-sensitive post-processing, which attributes costs to prediction errors and minimizes the total expected cost [45].

Lastly, hybrid methods may make usage of techniques from different approaches to obtain a more complex solution.

This work used a hybrid solution using pre and post-processing techniques, namely, re-sampling using SMOTE, and the threshold method. Not using special-purpose learning methods allows this work to test and compare different models trained using different - not adapted - algorithms.

2.2.3 Evaluation metrics

In imbalanced class problems, like this one, where FN errors have a higher cost than FP, accuracy is not an effective measure of the model's performance [38]. This is due to the fact that accuracy focus on minimizing both types of errors, not accounting with their different costs, and since the class of interest is, typically, much less represented than the others, a model would have a high accuracy if it only focus on correctly classifying the examples of the other class. For that reason, other metrics are better suited for these types of problems.

Several metrics were used in this work some because they are a better fit for imbalanced datasets, specifically sensitivity, specificity, precision and F-measure with $\beta = 0.5$ ($F_{0.5}$), others because they were used in other studies of this nature, namely, Area Under the Curve (AUC) and True Skill Statistic (TSS). These metrics are explained in detail below, including their formulas in terms of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). In the case of this work, the positive class is threatened. Figure 2.4 shows a confusion matrix of these four possible results of classification of an example in the context of threatened/non-threatened species.

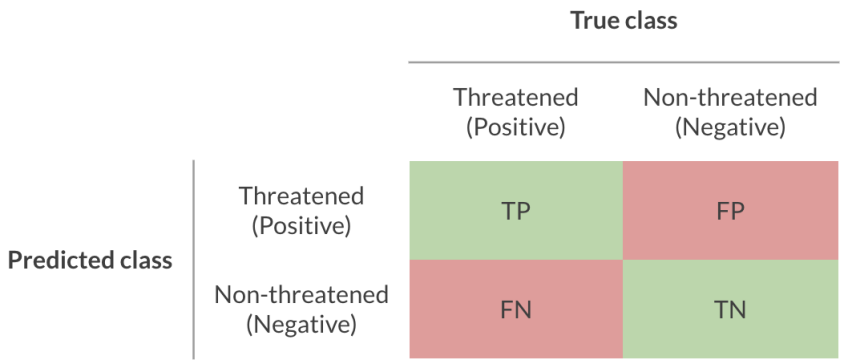


Figure 2.4: Confusion matrix in the context of threatened/non-threatened species.

- **Sensitivity** (recall, true positive rate): Fraction of threatened species which were correctly classified as threatened.

$$\frac{TP}{TP + FN}$$

- **Specificity** (true negative rate): Fraction of non-threatened species which were correctly classified as non-threatened.

$$\frac{TN}{TN + FP}$$

- **Precision**: Fraction of the species that were classified as threatened which were actually of the threatened class.

$$\frac{TP}{TP + FP}$$

- F_β: Measure of the precision and sensitivity (recall). A β of one is the harmonic mean of precision and sensitivity, balancing the two. The value of β can be adjusted to put a higher weight on precision or sensitivity and, consequently, to give a higher cost to false positives or false negatives, respectively.

$$F_\beta = (1 + \beta^2) \times \frac{precision \times sensitivity}{(\beta^2 \times precision) + sensitivity}$$

- **AUC**: Threshold independent metric obtained by calculating the area under the ROC curve (receiver operating characteristic) which plots the sensitivity (true positive rate) against the false positive rate. Figure 2.5 presents an example of a ROC curve.

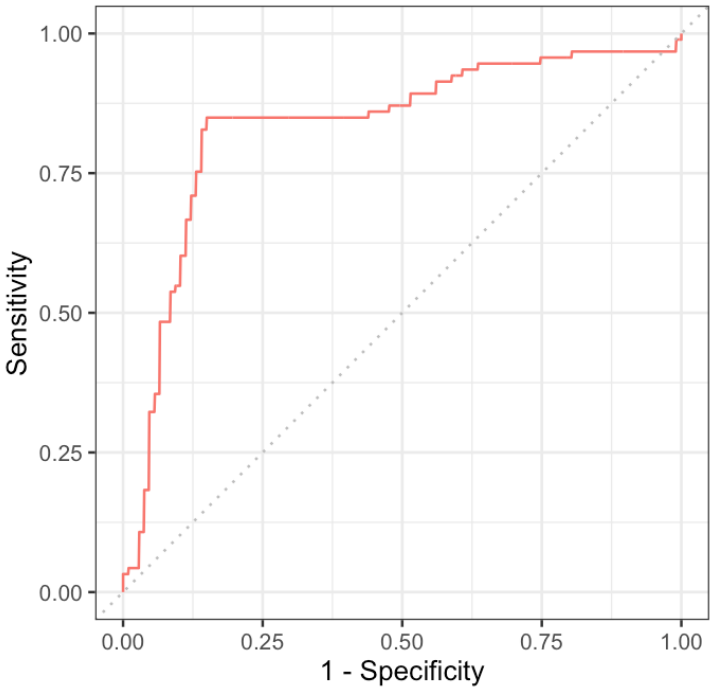


Figure 2.5: Example of a ROC curve.

- **TSS**: Threshold dependent metric based on sensitivity and specificity and independent of the prevalence of positive examples.

$$sensitivity + specificity - 1$$

2.2.4 Performance estimation

As previously described, a supervised learning model is trained on a set of labeled training data, but the goal is for it to be able to generalize and correctly predict the output given new, unseen, inputs. For that reason, in order to validate the performance of the model, one should not use the same data examples that were already seen in the model during training. Those records

were used to obtain the model's underlying function and using them for testing could inflate its performance to unrealistically values [36].

K-fold Cross Validation (CV) [46] is a technique used for splitting the dataset into train and test by creating k partitions (folds) of data without repetition. The model training is done on the data from $k-1$ folds and tested on the remaining one. Each fold is left out for testing once and the average score of the k generated models is taken as an estimation of the model's performance.

When data is not a problem, CV can be used to estimate parameters, and a separate hold out test set is then used to obtain an estimation of the model's performance from data that was never seen by the model. However, CV is particularly useful when there is a limited amount of data to train and test the model and it helps avoiding overfitting since all data points end up being used in training. In the end, a new model should be trained with all training data and CV provides the estimation of its performance [36].

2.3 State of the art

Even though the application of data science to ecology and conservation is still in its early stages, and standards and guidelines are still being defined [31], several works in ecology have emerged in recent years across different scales from 'macro' applications at a global sphere (e.g., [47, 48]) to 'micro' applications intended for fine scale assessments (e.g., [49–51]). In particular, machine learning techniques have been applied to predict conservation status at a global or continental scale [52, 53].

In [52], the authors focused on the flora of the African continent and created a complementary method, the Preliminary Automated Conservation Assessments (PACA). It used large numbers of species data from RAINBIO and GBIF and estimates thereof and, based on the IUCN Criteria A (estimation of population reduction) and B (estimation of geographic distribution), automatically categorized species using six preliminary levels. Those levels could then be used to prioritize more extensive and detailed conservation assessments.

The authors of [53] also devised a workflow to facilitate the process of predicting conservation status of multiple taxa. Like the work presented in [52], this method was applied to land plants (over 150,000 species), but at a global level. It used geographic, environmental and morphological trait data and applied a machine learning method (Random Forest [54]), to predict the conservation status. Comparing the results from two datasets, one with spatial data only and another with spatial and morphological, the authors concluded that the spatial-only performed better, possibly due to the lack of morphological data, which lead them to use fewer data points.

Random Forest (RF) have been extensively used in these type of studies, due to its ability to deal with overfit issues and to handle large quantities of both data points and predictor variables [53]. The algorithm has been showing promising results for certain clades [22, 53, 55–57]. In particular, a study used it to assess the threat status of terrestrial vertebrates, and separately

for four taxonomic groups (amphibians, reptiles, birds and mammals) [22], and another focused on reptiles [58]. This work stands out from previous works by avoiding assuming that RF is the best algorithm for this type of application or data, and tested other machine learning algorithms.

Using global data from a very large number of terrestrial plant species and RF, recent studies showed the utility of coarse-scale data in assessing the extinction risk of plant species [57] by predicting its threat level [53]. The authors of [59] used herbarium specimens data to compare machine learning methods, and specifically RF, with more traditional approaches like directly assessing the AOO and EOO, or the specimen count, against IUCN criteria. Random Forest outperformed all other approaches in terms of accuracy, using these data. In [60], the authors used decision trees and regression methods to predict the extinction risk of thousands of species of plants of the Brazilian Atlantic Forest, and to determine the relationship between predictor variables and extinction risk. Other authors [61] used deep learning to automate the threat assessment of plant species globally. Finally, a study [62] used Cumulative Link Mixed Models (CLMM) to evaluate the threat level of species of a subfamily of fish, the groupers, discriminating six of the IUCN threat categories.

Conservation data is imbalanced since the percentage of threatened species is typically lower than that of non-threatened species, respectively, 28% and 72% for all currently assessed species. That being said, when looking individually into taxonomic groups, there are some where the number of threatened species surpasses that of non-threatened. Previous works have tackled this imbalance problem by under-sampling the majority class to match the number of species in the minority class [53]. Other works did not try to balance the dataset, but used appropriate metrics for performance estimation, namely, sensitivity and specificity [57].

Chapter 3

Dataset curation

This project started by the curation of a dataset that would allow answering the question of which species should be given a higher priority for conservation measures. The first step of the proposed pipeline was to retrieve occurrence data from several sources. Since data was scattered across many public repositories, the first goal was to create a valid and representative dataset for as many reptile species as the data sources allowed. That was in itself a valuable deliverable since it implies data curation from a variety of sources.

This chapter introduces the sources used for this work and all the data transformations done to them in order to create the actual datasets that were used. It starts by explaining the steps to import and pre-process the source data, curating it into a dataset of reptile presences in a geographic locations and at a point in time, in Section 3.1. It then details the process of choosing the relevant ecological variables, in Section 3.2, and, in Section 3.3, the aggregation of presences into a species-level dataset and where the geographic variables were generated. Lastly, in Section 3.4, it introduces the different experimental datasets created in order to test whether some variables can be used to improve or replace the typical predictors used by IUCN, namely, AOO and EOO.

3.1 Dataset of presences

3.1.1 Importing and merging the source data

In total, 18,608,087 data points were extracted from four online open databases GBIF [63], PREDICTS [64], BioTIME [65], and the Living Planet Index (LPI) [66].

After importing the data, each dataset was cleaned and transformed so their columns and values were matched, which had the same meaning but different naming across the different sources. Additionally, date formats had to be converted to the same standard. These datasets were filtered to contain only species of the class Reptilia and to include only data points with a taxonomic rank of species or subspecies. After each dataset was pre-processed, a subset of

columns was selected. The datasets were then merged to create the presence dataset from where duplicated records were removed.

3.1.2 Pre-processing

Simply merging the data from these datasets yielded more than 15 million data points, but with a variety of possible issues. It needed to undergo a process of curation that trimmed down the number of points. The desired result was a dataset that could be trusted to contain only accurate and relevant occurrences, and that could be matched against the current reptile taxonomy.

3.1.2.1 Synonymizing

Species taxonomic classification is ever evolving [67]. New findings from biologic and genetic studies (e.g. genome sequencing) may lead species to be split or synonymized, subspecies to be promoted to species or vice-versa [68]. Since this data spanned over decades, one needed a way to map older records to the most up to date taxonomy. To tackle this challenge, the Reptile Database (or ReptileDB) [29], which is the most complete reference for reptile taxonomy, containing taxonomic information of 11,440 reptiles, in December 2020, including synonyms of each taxon, and is frequently updated (every six months), was used to harmonize the species names on the dataset.

A taxon was referred to using the binomial nomenclature, composed by its genus name and the specific name. The scientific name of each species on the presence dataset was matched to the species scientific name in ReptileDB or to a list of synonyms derived from it. If it matched the synonym it was renamed to the most current name. Records that could not be matched to the taxonomy, either directly or by synonymizing, were removed, since there was no automatic way to accurately match them to a taxon.

3.1.2.2 Cleaning

Some of the records represented occurrence data which conveyed the presence/absence of a species in geographic (or cartographic) coordinates for a given point in time; others, abundance data, which depicted the number of observed individuals at a given location and date of sighting. Here, only presence records were of interest, since one could not be sure if an absence indicated that a species did not exist in that location or habitat, or if it simply was not detected at observation time [69]. However, abundance data with counts higher than zero was also considered as presences. The dataset contained other information about each data point, whenever present in the data source. It included, among others, fields related to the basis of the record (describing if the point was acquired via human observation, machine observation, preserved specimen, etc), a measure associated with the coordinate uncertainty, and any notes about data issues (e.g. the coordinates were rounded, the datum was inferred).

Reptile taxa resided mostly in tropical or temperate climates, so coordinates falling in extreme latitudes (larger than 70ºN or 70ºS) were manually verified by experts and removed from the dataset [21]. Moreover, presences without the observation year, or before the year 2000 were discarded due to the uncertainty associated with older positioning systems such as GPS or GLONASS [70]. Inaccurate coordinates could lead to improper values being extracted for the predictor variables for those species, introducing errors in the data. After applying this filter, the presence dataset spanned two decades, from 2000 to 2020.

3.1.3 Handle species with a low count

At this point, only 1,275,311 of the original 18,608,087 data points were kept in the dataset, corresponding to 7266 species and a great number of them did not have enough data points for an analysis on it to be meaningful. Thus, only species with at least 20 presences should be kept, in order to have a minimum amount of information for each species [71, 72].

One option was to drop the records of those species that did not meet this criteria, but it would mean loosing information on 4651 species. There was also the hypothesis that these species could have low numbers because they were already under threat or near it. Moreover, many of those were endemic species, which means they only existed in a particular region of the globe, putting them more at risk of becoming threatened. These were the reasons for including those species in this work.

Another option, was to use polygon data available in IUCN [19], and extract the predictor variables for each cell on the entire range of the polygon. The issue with that was the species may not occupy the entire range, leading to some false positives in terms of occurrences. A trade-off between those two was to use the polygon data only for species that had been assessed by IUCN, and not found to be of least concern (LC). These species were either known to be threatened (categories: CR, EN and VU), at risk (NT), or had been assessed but the lacked of data lead to no category being assigned (DD). Like the data of individual presences, the polygon data also went through the process of synonymizing the scientific name.

By the end of this pre-processing step, the dataset contained 72,289,380 records, including the values extracted from the polygons.

3.2 Select and generate ecological variables

3.2.1 Relevant ecological variables

Some studies focused solely on the geographic distribution of species. These failed to capture the relation between taxa and their habitat and the fact that the same environmental characteristics, which were relevant for a given specie, may occur across many locations. Biodiversity and conservation models assume a correlation between the number of individuals of a species and the

extent of their habitat, and the existence of certain conditions on the world around them. For most species, there is evidence of the effectiveness of environment variables as predictions. For that reason, this work considered three categories of predictor variables that could, potentially, affect the distribution of a species. Namely, bioclimatic, topographic and habitat heterogeneity variables, together called ecological variables from now on. These provided information about species habitat, climate, vegetation and topography linked to presence data. Predictor variables used for model development were decided with the help of experts and literature review. Nevertheless, these variables had the potential to be highly correlated. So, a variable selection process was carried through an automatic selection process, described below, in Section 3.2.3, and followed by expert revision as well.

Bioclimatic variables represented annual and seasonal trends in temperature and precipitation (e.g., Annual Mean Temperature, Precipitation Seasonality), as well as extremes (temperature of the coldest and warmest month). They were derived from the monthly mean, max, mean temperature, and mean precipitation values and were created precisely for ecological applications. Two distinct datasets providing this information were tested, WorldClim and Chelsa. Chelsa was chosen to move forward since it had a better coverage of the points in the presence dataset, leading to less missing values (see Subsection 3.2.2). It has also been reported to have better accuracy [73].

Topographic variables were used because terrain topography create the finer-scale variations in climate and can have a big impact in nutrient availability, and water flows which influence the number of individuals in a location. Values related to the terrain elevation, roughness, aspect eastness, and aspect northness are correlated to the amount of sun light that can reach a given spot. Since reptiles are ectotherms (cold blooded), and cannot retain heat as well as endotherms (warm blooded) animals do, they depend on the air and soil temperatures to regulate their own as well as for sex determination [24].

Habitat heterogeneity variables portrait the amount of distinct vegetation present. Species with the capacity to inhabit distinct types of habitats should have more ease adapting in the case of changes to their habitat.

3.2.2 Extract and pre-process

These variables existed on raster layers with values corresponding to coordinates on the globe. The 29 rasters had the same resolution of 1 km and they were aligned in order of all cells in order for their cells to overlap geographically, using the *bio_1* variable as the baseline.

Due to the valid representation extent of most of the ecological variables used (e.g. land temperature), their rasters only had values for coordinates on land. For that reason, marine species, like marine turtles and sea snakes were excluded from the analysis. Even though many of them spend a part of their time on land, one would need to use a different set of variables that explains their marine habits and ecosystem better. Without them, the models would not be

able to learn from a significant part of their habitats. In order to include them other variables would have to be used, for instance, ocean temperature, salinity, and others. In total, 7 species of marine turtles and 71 sea snakes were left out. After excluding these species, there were still multiple records with missing values for the ecological variables. Those could be caused by wrong coordinated that fell on the water, or by the fact that the raster layers did not have enough resolution to represent some of the islands and coastal coordinates.

Once again, to avoid loosing information on entire species, these records for species which would have less than 20 presences after this step and which had a conservation status of CR, EN, VU, NT or DD were imputed. For that, it was assumed presences of terrestrial species located on water were due to imprecise coordinates and the points were imputed by looking for the nearer land cell in a radius incremented by 1 up to 10 km. These data points were savaged because they belong from species of categories with most interest to conservation efforts. LC records with missing values, were not imputed and left out of the dataset to avoid the introduction of too many imputed values.

Duplicates records can happen due to the same record existing in multiple datasets, or to different samples being done on the same grid cell, thus having the same values in all the predictor variables. Additionally, species may occur in geographically distant places, but those environments may be characterized by very similar conditions. If the values of predicted variables are correlated, even for very geographically distant individuals, the amount of information contained in the second observation may be small. It has been proven that correlated records can cause biases in the dataset [74]. To avoid this, presences were filtered using a process of re-sampling them onto a lower resolution on the predictor space, and by dropping duplicates on the newer resolution. To maximize variance on the relevant dimensions a Principal Component Analysis (PCA) was run before the re-sampling, and the top three dimensions were used.

3.2.3 Variables selection

The result from the previous steps was a representative dataset of non-correlated reptile data points and environment information that can be used for multiple future studies. The final presence dataset had 70,550,648 data points. In total, the data contains details on 3260 reptile species which were divided into five groups by their clades, namely, amphisbaenians, crocodiles, lizards, snakes and turtles. Tuataras were intentionally left out, since this clade comprises only one species with very unique characteristics and does not fit into any of the other groups. A sixth group contained the data points of all taxa.

The original set of ecological variables contained 29 variables from the three groups detailed before (19 climatic, four topographic and six habitat heterogeneity predictors, check Table 3.1). This was a large number of variables for the amount of species present in each group, and many of these variables were known to be highly correlated among them. Moreover, each reptile group has different characteristics, which may lead to different environmental requirements. To better model the conservation status of each group, this work selected different sets of variables for each

one and minimized the correlation among each set.

To obtain such subset of variables, a Principal Component Analysis (PCA) was run on the presences for each group. From these, this dissertation kept the variables that maximize the variance of each of the top PCA dimensions, according to the Pearson criteria by retaining 80% of the variance [75]. This predictor selection step allowed each of the groups to be modelled using the most adequate variables for their characteristics and behaviours.

The ecological variables selected by the PCA provide a subset of variables with low correlation, that represent the environmental factors tuned to each group. The list of variables for each group can be seen in Table 3.2.

3.3 Species-level dataset

At this point, presence data was aggregated from individual spatial presences to the species level to match the response variable (i.e., IUCN conservation status attributed for each species). This was done by calculating the mean and the standard deviation of each of the ecological predictors selected by the PCA step, to express how much a species is able to occupy habitats with different range of ecological conditions aiming to describe species realized niches (i.e., a n-dimensional hypervolume [81]). Moreover, the geographical variables were extracted from the presence data and added to the species datasets. The variables were grouped as follows:

- Geographic
  - AOO and EOO (two variables)
  - Latitude/Longitude-derived variables (four variables)
- Ecological
  - Climate (19 variables)
    - Temperature (T)
    - Precipitation (P)
  - Topographic (four variables)
  - Habitat heterogeneity (six variables)

The complete description of these 35 variables can be found in Table 3.1. The variables were represented in raster layers which allowed to extract values from each one using the geographic location of data points.

EOO is the area within the shortest imaginary boundary which encompasses all presences of the given species. According to the IUCN guidelines [33], it can obtained from the minimum convex polygon, which is the smallest polygon containing all presences of the species and which has no internal angle over 180 degrees. AOO represents the area within the EOO that is actually

Table 3.1: List of the geographic and ecological variables used in this study by type, source, name and description. All raster layers used for the ecological variables have a 1 km resolution.

Type (Source)	**Name**	**Description (unit)**
Geographic		
(GBIF [63], PREDICTS [64], BioTIME [65], LPI [66])	AOO	Area of occupancy of a species (km^2)
	EOO	Extent of occurrence of a species (km^2)
	min_lat	5th percentile of latitude (°)
	max_lat	95th percentile of latitude (°)
	med_dist_eq	Median distance to the equator (°)
	med_long	Median longitude (°)
Ecological		
Climatic (CHELSA [76])	bio_1	Annual mean temperature (°C)
	bio_2	Mean diurnal range (°C) mean of monthly (max temp - min temp)
	bio_3	Isothermality bio_2/bio_7 (×100)
	bio_4	Temperature seasonality standard deviation ×100
	bio_5	Max Temperature of warmest month (°C)
	bio_6	Min Temperature of coldest month (°C)
	bio_7	Temperature annual range (°C) bio_5-bio_6
	bio_8	Mean Temperature of wettest quarter (°C)
	bio_9	Mean Temperature of driest quarter (°C)
	bio_10	Mean Temperature of warmest quarter (°C)
	bio_11	Mean Temperature of coldest quarter (°C)
	bio_12	Annual precipitation (mm)
	bio_13	Precipitation of wettest month (mm)
	bio_14	Precipitation of driest month (mm)
	bio_15	Precipitation seasonality (mm)
	bio_16	Precipitation of wettest quarter (mm)
	bio_17	Precipitation of driest quarter (mm)
	bio_18	Precipitation of warmest quarter (mm)
	bio_19	Precipitation of coldest quarter (mm)
Topographic (EarthEnv [77])	eastness	Aspect eastness
	elevation	Elevation (meters a.s.l.)
	northness	Aspect northness
	roughness	Roughness
Habitat heterogeneity (EarthEnv [78])	contrast	Exponentially weighted difference in Enhanced Vegetation Index (EVI) between adjacent pixels
	cv	Coefficient of variation. Normalized dispersion of EVI
	homogeneity	Similarity of EVI between adjacent pixels
	maximum	Dominance of EVI combinations between adjacent pixels
(Copernicus [79] [80])	fcover	Fraction of green vegetation cover
	ndvi	Normalized Difference Vegetation Index

Table 3.2: List of ecological variables (EcoV) selected for each of the taxonomic groups by the PCA step of the pipeline. Check Table 3.1 for details on the variables.

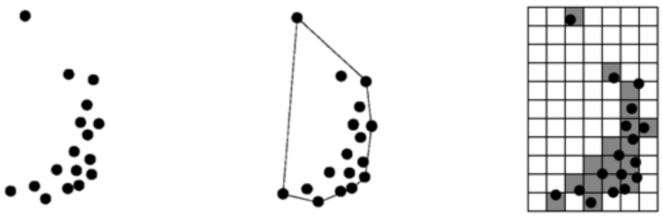
EcoV	Amphisbaenians	Crocodiles	Lizards	Snakes	Turtles	All species
bio_1			V		V	
bio_2		V				
bio_3			V			
bio_4						
bio_5	V					V
bio_6						V
bio_7				V	V	
bio_8		V			V	
bio_9	V	V			V	
bio_10				V		
bio_11						
bio_12		V	V			
bio_13	V					V
bio_14						
bio_15				V		
bio_16						
bio_17						V
bio_18				V		
bio_19						
eastness		V	V		V	V
elevation						
northness	V		V	V		
roughness		V				
contrast	V	V			V	V
cv			V	V	V	V
homogeneity		V	V	V		
maximum	V				V	V
fcover						
ndvi	V					

occupied by the species. This reflects the fact that a taxon may not occupy all of its EOO, since it may contain unsuitable or unoccupied habitats. Its estimation was obtained, in accordance with the IUCN guidelines [33], by counting the number of occupied cells in the grid that covers the range of the taxon and multiplying it by the area of a cell, as shown in the following formula:

$$\text{AOO} = \text{no. occupied cells} \times \text{area of individual cell}$$

The cell size used for estimating these metrics, was the recommended reference scale by IUCN [33], $4km2$ ($2x2km$). Figure 3.1 depicts the difference between the two metrics.

The labels corresponding to the IUCN threat level (Critically Endangered, CR; Endangered, EN; Vulnerable, VU; Near Threatened, NT; Least Concern, LC; Data Deficient, DD) were extracted from the IUCN Red List webpage (https://www.iucnredlist.org/) and matched with



(a) (b) (c)

Figure 3.1: Example of the difference between EOO and AOO. (a) shows the spatial distribution of presences of a given taxon. (b) shows the minimum convex polygon which can be used to estimate EOO, and (c) shows how AOO can be estimated using the sum of the occupied grid squares. Adapted from [33].

the corresponding species. The target variable was obtained by binarizing to threatened/non-threatened classes. IUCN categories CR, EN and VU were considered as threatened and, NT and LC categories were considered non-threatened. DD and NE have an unassessed extinction risk category, so they were not used for training. The final models were used to make predictions for species in those two latter classes. DD and NE species are of particular interest for this application given that the lack of information may imperil them [12].

The result of this step was a set of six species-level datasets, with threat status label, corresponding to the five clades of species, plus the entire species pool. The characteristics of these datasets are presented in Table 3.3, including the number of examples used for training (corresponding to the number of assessed species) and the number of variables. Neither of the datasets had a perfect balance between threatened and non-threatened species, but the imbalance was not too steep, see Table 3.3 for details. Note that, for the most imbalanced groups, there were more non-threatened than threatened species, while the groups where threatened was the majority class tended to be less imbalanced. Also note that there was a large difference in the number of examples across the groups of species. Independently of which is the minority class for a given group, *Threatened* was always used as the positive class.

While the amount of presence data was very substantial, the models will be trained on the dataset of species, which has much less data points. In total, the data contains details on 3260 reptile species.

3.4 Experimental datasets

AOO and EOO are traditionally used by IUCN experts to generate assessments, but often, it is not possible to calculate them accurately or these depict a shallow representation of the species niche hypervolume. In those cases, it would be good to rely on a more complete proxy set of variables. This work set the goal of understanding whether ecological variables, which are more

Table 3.3: Number of presences, total number of species (Tsp) with presence data, number of species (Nsp) with threat status available at IUCN, and ecogeographical variables (Nvr) for each group, the imbalance ratio (IR), calculated as the size of minority class over the size of majority class, and the minority class of each reptile group.

Group	Presences	Tsp	Nsp	Nvr	IR	Minority class
Amphisbaenians	1,901,717	44	19	21	0.7	Non-threatened
Crocodiles	1,044,135	23	21	23	0.9	Non-threatened
Lizards	29,323,862	2716	2103	21	0.5	Threatened
Snakes	25,078,835	1582	970	21	0.4	Threatened
Turtles	13,202,099	203	147	23	0.6	Non-threatened
All species	70,550,648	4568	3260	24	0.5	Threatened

easily available, would have a satisfying performance and could be used without the presence of AOO and EOO. Moreover, if adding other geographical variables would improve performance over a dataset of only AOO and EOO.

In order to evaluate if there is a gain from using variables other than AOO and EOO, this work created five distinct datasets for each of the six groups of species, totalling 30. All five datasets contain the same set of species, only the variables vary.

The six groups/datasets, one for each taxonomic group plus the one for the group of all species, had up to eight ecological variables chosen using PCA, plus six geographic variables, totalling up to 14 variables, but some of the groups were small (e.g.: $Nsp_{amphisbaenians} = 19$; $Nsp_{crocodiles} = 21$) for such a number of variables. For this reason, an additional variable selection step, using recursive feature elimination (RFE) [82], a backwards selection method, was tested and implemented for the datasets of ecological and of ecogeographical variables, generating the last two of the five datasets created per group. The workflow used for choosing the most relevant variables for each group is depicted in Figure 3.2.

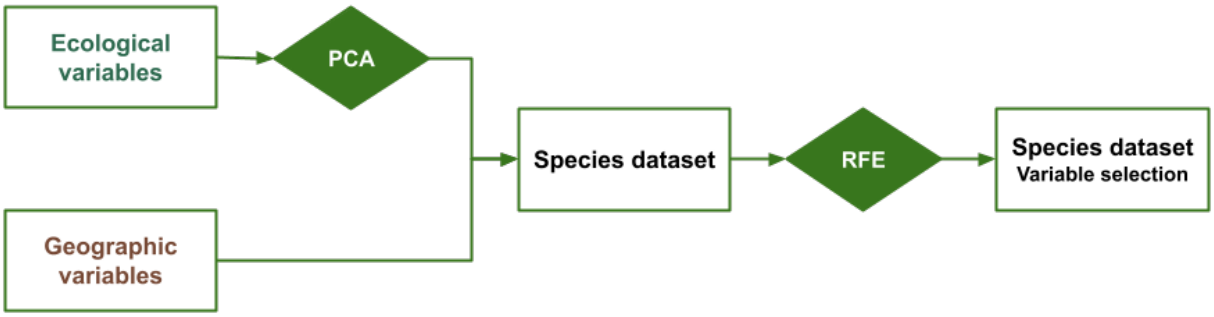


Figure 3.2: The workflow used for selecting the ecological and geographic variables for each dataset. This was applied to each of the groups and resulted in two datasets per group: one with a set of ecological variables specifically picked for the group using PCA, plus all geographic variables, and one that further drills down on the most relevant variables using recursive feature elimination (RFE).

The five datasets, summarized below, were used in the experimental study performed in this dissertation to evaluate the effect of including different types of variables.

- *AOO_EOO* - the baseline — dataset containing only AOO and EOO, the two most important variables considered in IUCN assessments;
- *EcoV* - dataset containing only ecological variables;
- *EcoV_FS* - *EcoV* dataset post-processed by a variable selection step, using backwards selection;
- *All EGVs* - datasets containing both AOO and EOO, and ecological variables, plus other geographical variables derived from the presences dataset: minimum and maximum latitudes, latitude length, median longitude and median distance to the equator;
- *All EGVs_FS* - *All EGVs* dataset post-processed by a variable selection step, using backwards selection.

Given the interest in comparing performance of classification models to predict the threat status of species using the environmental datasets and the *AOO_EOO* datasets typically used in IUCN assessments, the *AOO_EOO* dataset was used as the baseline, for each group. The result of this step was a set of six species-level datasets, with threat status label, corresponding to the five clades, plus the entire species pool. Table 3.4 shows the characterization of the 30 datasets generated.

The next chapter presents the experimental setup in detail.

Table 3.4: Dataset characterization containing the number of training examples (species with threat status available at IUCN), and variables for each dataset. The imbalance ratio (IR), calculated as the size of minority class over the size of majority class, and the minority class of each reptile group are shown for context.

Group	Dataset	Examples	Variables	IR	Minority class
Amphisbaenians	*AOO_EOO*	19	2	0.7	Non-threatened
	EcoV		14		
	EcoV_FS		3		
	All EGVs		21		
	All EGVs_FS		4		
Crocodiles	*AOO_EOO*	21	2	0.9	Non-threatened
	EcoV		16		
	EcoV_FS		2		
	All EGVs		23		
	All EGVs_FS		2		
Lizards	*AOO_EOO*	2103	2	0.5	Threatened
	EcoV		14		
	EcoV_FS		6		
	All EGVs		21		
	All EGVs_FS		16		
Snakes	*AOO_EOO*	970	2	0.4	Threatened
	EcoV		14		
	EcoV_FS		11		
	All EGVs		21		
	All EGVs_FS		19		
Turtles	*AOO_EOO*	147	2	0.6	Non-threatened
	EcoV		16		
	EcoV_FS		6		
	All EGVs		23		
	All EGVs_FS		16		
All species	*AOO_EOO*	3260	2	0.5	Threatened
	EcoV		17		
	EcoV_FS		10		
	All EGVs		24		
	All EGVs_FS		16		

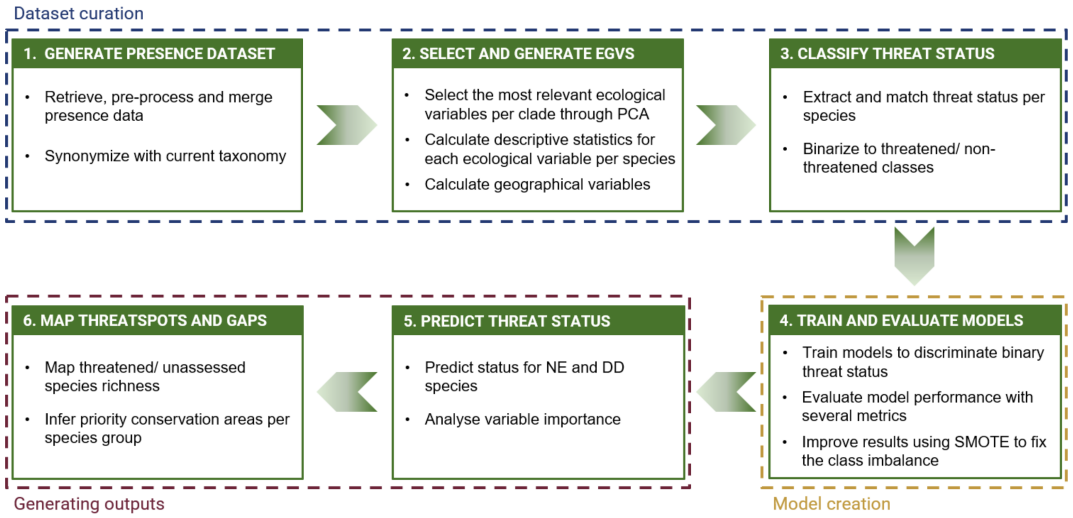
Chapter 4

Experimental study

This chapter presents the modelling pipeline created to generate and pre-process the datasets used for this dissertation, model the threat level and evaluate the models, and to generate the outputs, namely, the predictions for unassessed species and the maps of threatspots, in Section 4.1. It then details the experimental setup used for the modelling task and to compare the models using different datasets, in Section 4.2.

4.1 Modelling pipeline

The created pipeline has six main steps to address the gathering of terrestrial reptile presences, construction of a tidy dataset of species, including the relevant predictors, and the modelling and evaluation of the resulting predictions (cf. Figure 4.1). It started from the raw presence data available in online databases. Afterwards, the pipeline cleaned, pre-processed, performed variable analysis, and aggregated the data into a species-level dataset. These three steps are detailed in Chapter 3. The dataset, aggregating all variables at species-level, was the input of the modeling step, in which models were trained and evaluated using the experimental setup described in Section 4.2. The best models were used to generate predictions for the species without a known conservation status (i.e., DD and NE). Five sets of predictor variables were compared for each group of species to evaluate the impact of using ecogeographical variables (EGV), both by themselves and with other predictors more commonly used in IUCN assessments and the predictions were matched to newly available assessments. Lastly, those predictions were mapped globally by generating the convex hull of presence records of each species, thus revealing potentially unknown threatspots.



(a)

Figure 4.1: Diagrams of the modelling pipeline showing the six high-level steps implemented. Namely, three steps for the dataset curation, resulting in several labeled datasets at the species level, containing different sets of variables; the training and evaluation of the models and the generation of outputs.

4.1.1 Predictive modelling

In this domain, the cost of a false negative was higher, as that may lead to the lack of conservation efforts allocated to threatened species. On the other hand, a false positive may mean that the already scarce funding would be distributed to species that do not need it. In imbalanced classification problems, like this one, with different costs for the two types of errors, accuracy is not an effective measure, since it does not reflect the higher cost of a false negative. For these reasons, this work selected models that maximize sensitivity (also known as true positive rate), but without loosing too much specificity (true negative rate).

The models were trained and evaluated using reptile species where the conservation status is known, using the IUCN Reptiles assessments available in December 2020 [83]. The binary classification models were trained to discriminate between threatened and non-threatened species, according to IUCN criteria [33], for each group, using different algorithms implemented in R [84], namely, RF [85], XGBoost [86], decision trees (C5.0-[87] and rpart [88], General Linear Model (GLM) [89] and k-NN [90].

To understand whether the use of other variables have an effect on the classification performance of species into threatened or non-threatened, three groups of datasets were generated: containing (i) only variables based on IUCN assessment criteria — AOO and EOO, (ii) only ecological variables, and, lastly, (iii) ecological and geographical variables (EGVs). To assess if the performance of the models using different datasets was significantly better with respect to the baseline dataset (*AOO_EOO*), this dissertation used the paired one-sided Wilcoxon test. It

also applied this test to the pairs of the *EcoV* and *EcoV_FS* and the *All EGVs* and *All EGVs_FS* datasets, allowing us to test how the ecological variables alone would perform against the latter datasets as a baseline.

Lastly, to try to improve the performance of the models for the groups where the class of interest was the minority class, this study tried to address the class imbalance problem [38] by applying SMOTE [32] as a pre-processing step on those datasets, an comparing the results. As previously mentioned, some works balanced the dataset by under-sampling the majority class to match the minority class [53]. This had the drawback of reducing the number of species available to train the model. With only 970 labeled species on the smallest of the three groups, the snakes, and 3260 on the group of all species, it was important to keep as many data points as possible. For this reason, SMOTE was used to generate synthetic examples from the existing data, but without doing any under-sampling, balancing the datasets to a 50/50 distribution of threatened/non-threatened species.

4.1.2 Threat status predictions

To understand which factors contributed the most to predicting the threat status of species, this dissertation ranked variable importance scores, using the cross validation (CV) dataset. For that the varImp function of the caret package [82] was used, which is a generic wrapper function that calls the specific methods for calculating variable importance of different types of models produced with caret. Also, in this step of the pipeline, previously trained models which performed best for each of the five reptile clades were used to obtain predictions for previously unassessed species, labelled as DD and NE.

Some reptile species have been very recently re-assessed by IUCN in June 2021 after the predictions were generated [14], making them the perfect blind dataset for evaluating the models and estimating how well the models performed. To compare the predictions with these new assessments, this study extracted the new threat status of those species, synonymized their scientific names against the taxonomy used to generate the models, from Reptile Database [29] in December 2020, and matched them against the species to which the models generated predictions. 88 matches were retrieved.

4.1.3 Maps of threatspots

Using the dataset of presences gathered and cleaned in step one, the conservation status extracted from IUCN assessments and the predictions from the models, this work generated a dataset containing the range of the each species, using the convex hull, as it is used by IUCN, and the label – threatened or not threatened.

For each reptile group, the total and threatened species richness was calculated in each 10×10km cell based on rasterized convex hulls, merging geographic records by species.

Predictions were mapped in order to analyse how the threatened species of each group were distributed at global level and outline the hotspots of threatened and unassessed species, thus revealing potentially unknown threatspots, this is, gap areas previously unknown by IUCN.to have high threatened species richness. Thus, this dissertation plotted the percentage of threatened species to all reptile species, and of the threatened species predicted by the models to the threatened species assessed by IUCN. In the first case, higher values highlight the threatspots where conservationists should focus on, due to the relatively high number of threatened species. In the second case, high values on the map show gap areas that have a high incidence of predicted threatened species, but low incidence of species known to be threatened. So they are potential unknown threatspots.

4.2 Experimental setup

As previously mentioned, different classification models were trained, for each group, using different algorithms implemented in R [84]. Namely, Random Forest (RF) [85], XGBoost [86], decision trees (C5.0 [87] and rpart [88]), Generalized Linear Model (GLM) [82, 89] and k-NN [82, 90]. RF is an ensemble method based in trees, and used by other recent studies in prediction conservation status of species [53, 57, 59]. XGBoost is another ensemble strategy that has shown to perform better than RF in some settings [91]. In opposition to RF, which uses bagging to improve performance by using the vote of multiple, parallel trees, XGBoost uses a boosting to minimize the classifier error each iteration [92]). Decision trees generated by the CART algorithm implemented by rpart, is the base algorithm used by Random Forest. Both CART and C5.0 decision trees are simple algorithms, which may have better results in smaller datasets due to their smaller natural tendency to overfit. GLM is also extensively used in this domain [93], and it creates a linear decision boundary, which may be desirable for groups with a low species count. k-NN is a simple instance-based machine learning algorithm. The fit to the data can be tuned with a single parameter k, which establishes the number of neighbour instances to consider to obtain a classification.

In the case of k-NN, the variables were normalized using Min-Max normalization. This was done because, k-NN is distance-based, making it sensitive to having variables with different scales, and Min-Max transforms all values into decimal between 0 and 1. The formula to perform Min-Max normalization is as follows:

$$\frac{value - min}{max - min}$$

The feature selection step used to generate the *EcoV_FS* and *All EGVs_FS* datasets described in Chapter 3 was done using Recursive Feature Elimination (RFE) [82], a backwards selection method.

In a first approach, for the datasets of smaller groups, namely, amphisbaenians, crocodiles and turtles, which have 19, 21 and 147 examples respectively, training and test sets were split using

Leave One Out Cross Validation (LOOCV) and for bigger groups, lizards, snakes and all species, with 2103, 970 and 3260 examples respectively, a standard 10-fold Cross Validation (10-CV) estimation method was used to split the dataset in training and test datasets. This allowed for more training examples to be available in the smaller groups, where data is too scarce for a bigger testing set. This had the drawback of yielding non-comparable results between groups. Moreover, the estimations obtained for the smaller groups were highly variable. Additionally, in a first approach, there was a separate testing set containing 25% of the examples, where no tuning of parameters was done, to allow for the performance estimation to use a set of data that was not seen by the model in any stage of training. Again, due to the low number of species available for training, even in the group of all species (only 2445 data points would be available for training after this split), this approach was dropped. In the final setup CV was used both to tune the parameters and to estimate the models' performance, which is also a typical approach in other similar studies [57].

Each dataset was split into a training and testing set using a 2x5-fold cross validation (CV). In this context, and given the small number of examples, the option taken was to split the dataset into five random partitions. Then, one partition was left out to be used for testing and performance evaluation, and the model was trained in the remaining four partitions [94]. Each partition was used for testing once, which means that five different models were trained and tested. To achieve a more reliable performance estimate, this process was run twice [46]. Thus, the performance of the models was assessed based on ten different metrics.

For the lizards, snakes and all species groups/datasets, we could face an imbalanced classification problem [38]. That is, they had a, comparatively, higher imbalance and, their minority class was precisely the class of interest (i.e., threatened class). In conservation, a false negative may lead to the lack of conservation efforts for threatened species and their irreversible extinction. On the other hand, a false positive may lead to scarce funding to be distributed to non-threatened species. As previously mentioned, in imbalanced problems like this one, with different costs for the two types of errors, accuracy is not an effective measure, since it does not reflect the higher cost of a false negative [38]. For these reasons, the models that maximized sensitivity (true positive rate), but without losing too much specificity (true negative rate [95] were selected.

To obtain a good model for each dataset, a set of parameters of each algorithm were tuned using a grid search approach (Table 4.1) which consists in testing and comparing multiple combinations of parameter values [96]. Each set of parameter values was used to train one model and its performance was estimated on the 2x5-fold. The parameters of the best model, according to the sensitivity and specificity metrics, were the ones used to train the model on the entire training set. It should be noted that the range of values for the tuning of k in k-NN and max_depth in rpart were chosen as to allow a good set of values to be tested without compromising too much on performance [97]. Since the performance of the models was not improving for any groups for values approaching 20 and 14 for k and max_depth, respectively, values higher than those were not tested. Additionally, the classification threshold was also tuned

by testing each value from 0.10 to 1, with increments of 0.05 [98]. The tuning process maximized the sensitivity of the models, without letting specificity drop to values lower than 0.7. Over the six algorithms used, a total of 1086 parameter values were tested for each of the five datasets of each of the six groups of species, leading to a total of 195,480 models trained using this tuning process.

Table 4.1: Parameter values of methods used in grid search for tuning.

Method	Parameter	Definition	Values	R Package
RF	mtry	Number of variables randomly sampled at each split.	2:7	randomforest [85]
	ntree	Number of trees to train.	{5,10,20,50,100,200,500}	
	nodesize	Minimum size of terminal nodes. Large numbers lead to smaller trees.	1:5	
XGBoost	nrounds	Number of passes through the data. Each new round further reduces the difference between ground truth and prediction.	{5, 10, 20, 50, 75, 100, 200}	xgboost [86]
	eta	Size of each boosting step.	{0.01, 0.1, 0.3, 0.5}	
	max_depth	Maximum depth of the trees.	2:4	
	colsample_bytree	Ratio of column under-sampling.	{0.5, 1}	
	min_child_weight	Minimum sum of instance weight needed in a child in order to continue partitioning that node further.	1:5	
k-NN	k	Number of neighbours to consider.	1:20	caret [82]
rpart	max_depth	Maximum depth of the tree.	1:14	caret [82]

The performance of models was assessed by the evaluation metrics sensitivity, specificity, precision and three metrics that combine others. Namely, Area Under Curve (AUC), True Skill Statistic (TSS), and F-measure with $\beta = 0.5$ ($F_{0.5}$). These metrics are common and appropriate when dealing with imbalanced data [99]. AUC and TSS, in particular, are used by other applications in this domain [100, 101]. $F_{0.5}$ was choose since it aligns with the domain objectives.

The five datasets resulting from the data integration explained in Chapter 3, namely, *AOO_EOO*, *EcoV*, *EcoV_FS*, *All EGVs* and *All EGVs_FS*, were split keeping the folds constant in all five datasets. Meaning when comparing one of the 10 folds of the 2x5-fold CV, between the five datasets, the records were the same and only the columns varied. This makes the set of estimated performance values comparable by pairwise comparison test.

To assess if the performance of the models using a different dataset is significantly better with respect to the baseline dataset (*AOO_EOO*), the paired one-sided Wilcoxon signed rank test was used over the results obtained by 2x5-fold CV. This test is a non-parametric statistical hypothesis test, so it does not assume that the data is normally distributed. It tests the null hypothesis that the distribution of pairwise difference between the two samples is symmetric around 0. Being a one-sided test, it allows to checking if the performance of a given model is significantly improved against a baseline, instead of testing for a significant difference between the two, in any direction. The statistical tests were carried using both confidence levels of 95% and 99%, so p-values smaller than 0.05 and 0.01, respectively. This process allow us to assess if the models trained with the dataset being tested had a significantly better performance than the one trained with the baseline dataset. This test was also applied to the pairs of the 'ecological'

and the 'all' variables datasets, allowing us to test how the ecological variables alone would perform against the all variables as a baseline. For implementing the test, the function wilcox.test of the stats R package was used.

The datasets responsible for generating the best models for the groups where threatened was the minority class, namely, lizards, snakes and all species, were augmented using SMOTE to generate synthetic examples. For that, the smote function of the DMwR package [102] was used with the parameters $perc.over = 300$, $k = 3$, and $perc.under = 0$, and a random selection of the synthetic examples was appended to the dataset in order to obtain a 50/50 balance of threatened and non-threatened examples. The best algorithms for each of these groups were then trained using these datasets. A Wilcoxon signed rank test was then run to compare the original models with the ones generated with the SMOTE datasets, using the first as the baseline.

Chapter 5

Results and discussion

This chapter starts, in Section 5.1, by presenting the results of the experiments done, namely, it shows the results for the models created, specially, for the best performing model of each taxonomic group, selected based on $F_{0.5}$, the variable importance from the *All EGVs* models, the new predictions generated for the unassessed species and the maps of threatspots. Then, in Section 5.2, these results are analysed from a modelling and from a conservation perspectives.

5.1 Results

Figure 5.1 summarizes the comparison of the performance of the models for the five datasets and for each of the six groups of species using the three metrics previously mentioned, $F_{0.5}$, AUC and TSS. For a given group, some models performed significantly better (alpha=0.05) when comparing to the results of the same algorithm over the *AOO_EOO* dataset. The results show that most groups of species benefited from using ecological and geographical variables, since, for multiple groups, *All EGVs* and *All EGVs_FS* datasets performed better than when using only the traditional predictors from IUCN, at least for some of the methods.

Overall, models trained for smaller groups, namely amphisbaenians and crocodiles, showed a larger performance variation across algorithms. With such a small number of cases the mean is a weak estimate and statistical significance differences are harder to obtain. This reinforces the importance of gathering more species presence (or abundance) data and including a broader set of species for improving training conditions for the models. In addition, these results suggest that when deciding on how to group species for modelling, one should balance between having groups that reflect species with similar taxonomic, phylogenetic and/or functional characteristics (so these can be well represented by the same set of variables), and also, having enough data points for the methods to be able to properly learn and generalize. The results of the best models for each of the six reptile groups, according to the three performance metrics are shown in Table 5.1.

A complete view of the results of the modelling, including the results for sensitivity, specificity, precision, $F_{0.5}$, AUC and TSS, for the six groups, six methods and five datasets, and of the

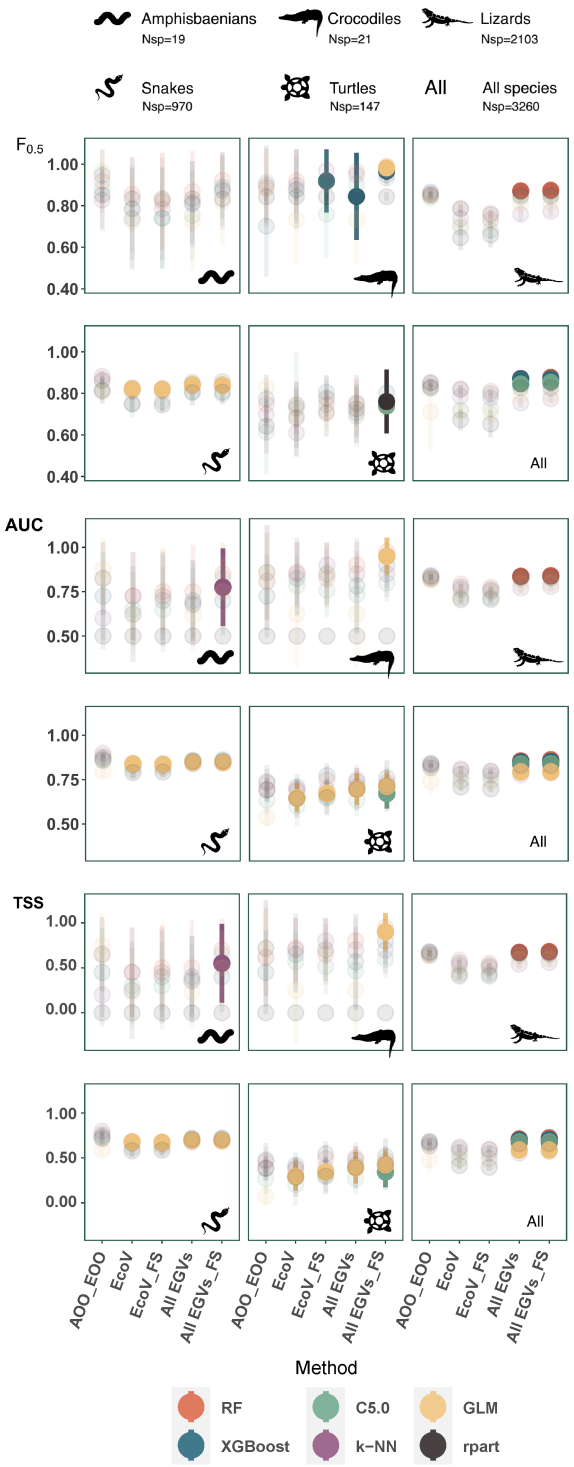


Figure 5.1: Performance of the models trained for each combination of taxonomic group, method and dataset, according to three metrics: ($F_{0.5}$, AUC and TSS. n stands for sample sizes. Each dot represents the mean of the results of a model across the 10 folds of the 2x5-fold cross validation setup. The standard deviation is represented by the vertical line. Results of models shown in an opaque color significantly outperformed the same algorithm over the *AOO_EOO* dataset.

Table 5.1: Best model trained for each group and their predictions for Data Deficient (DD) and Not Evaluated (NE) species. The table shows the total number of species (Tsp) with presence data, number of species (Nsp) with threat status available at IUCN, and ecogeographical variables (Nvr), the number of unassessed and total taxa predicted to be threatened, and the percentage of threatened taxa in respect to Tsp, the dataset and method that performed better for each group, selected based on $F_{0.5}$ results, and their performance estimates with respect to $F_{0.5}$, AUC and TSS metrics. The imbalance ratio (IR), calculated as the size of minority class over the size of majority class, and the minority class of each reptile group (non-threatened, nT; threatened, T) is also given.

Group	Tsp	Nsp	Nvr	DD+NE to T	Total T	%T	Dataset	Method	$F_{0.5}$	AUC	TSS	IR	Minority class
Amphisbaenians	44	19	21	10	21	47.7	*AOO_EOO*	GLM	0.97	0.88	0.75	0.7	nT
Crocodiles	23	21	23	1	12	52.2	*All EGVs_FS*	RF	0.99	0.98	0.95	0.9	nT
Lizards	2716	2103	21	404	1125	41.4	*All EGVs_FS*	RF	0.87	0.84	0.68	0.5	T
Snakes	1582	970	21	125	382	24.1	*AOO_EOO*	k-NN	0.88	0.90	0.80	0.4	T
Turtles	203	147	23	34	124	61.1	*EcoV_FS*	XGBoost	0.88	0.77	0.54	0.6	nT
All species	4568	3260	24	854	1945	42.6	*All EGVs_FS*	RF	0.88	0.86	0.72	0.5	T

statistical tests, can be found in the Table A.1. RF showed very good results across all groups, and was the most performant model for these three metrics on three out of six groups (check Table A.1), Additionally, less complex algorithms, like GLM and k-NN which are not based in ensembles, performed better with the *AOO_EOO* dataset, which has few variables, while RF and XGBoost tended to work better with the remaining datasets (Table 5.1). Moreover, in general, the additional variable selection step implemented in the pipeline yielded better results in comparison to its counterpart without variable selection.

5.1.1 Variable importance

To better understand the role of each EGV in the predictions by the different models, this work also evaluated their importance. Figure 5.2 shows the ranking of importance for RF, XGBoost, C5.0, GLM, and rpart, using only the top ten variables obtained with each method, and considering the dataset with *All EGVs* and the group of all species. K-NN was left out of this analysis because there was no wrapper for it in the varImp function.

It is clear that both EOO and AOO had an important role in the classification models (Figure 5.2). Additionally, other variables contributed to improving model performance, especially, geographic and bioclimatic variables. Results also suggest that variables based on statistical dispersion measures (e.g., standard deviation, SD) perform generally better since these portray species ecological tolerance ranges.

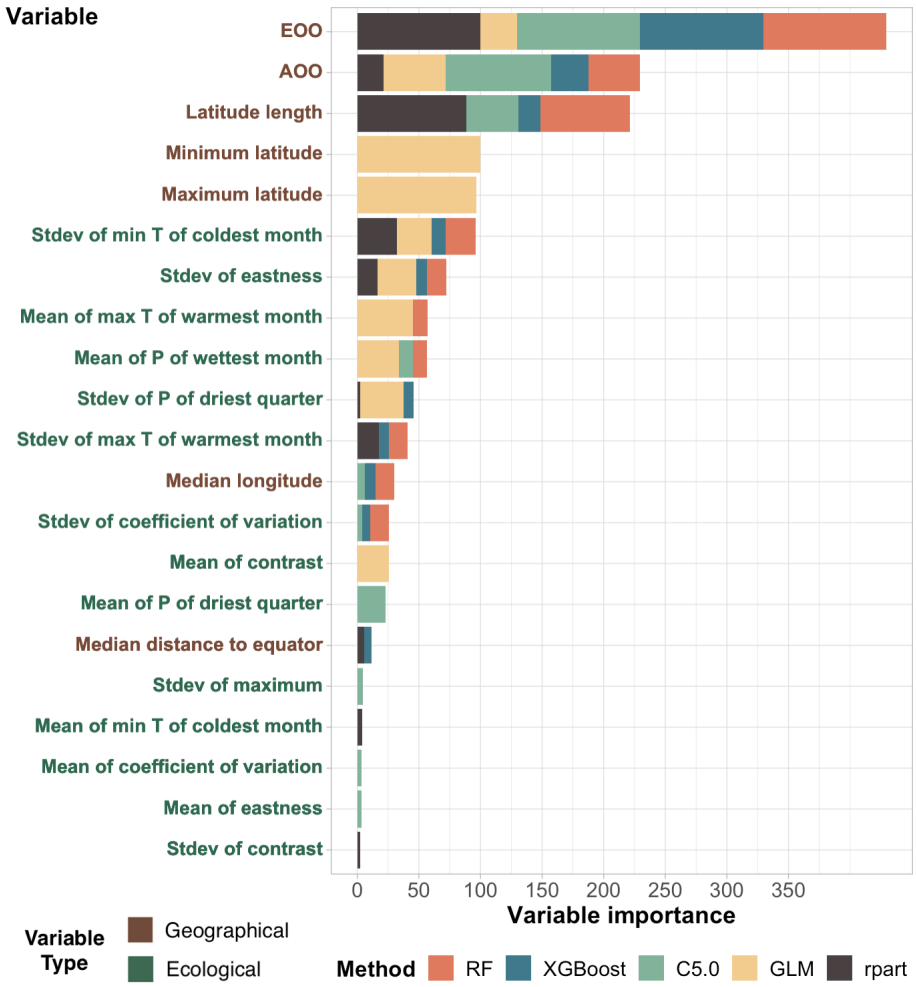


Figure 5.2: Variable importance plot for the group that included all species, using the *All EGVs* dataset. The plot shows the cumulative variable importance across five distinct algorithms, Random Forest (RF), XGBoost, C5.0, GLM, and rpart, shown in different colors. Only the top ten variables obtained with each of the five methods were considered. On the variable names, P stands for precipitation and T for temperature.

5.1.2 Handling Imbalance

When comparing the results of the original models to the ones using SMOTE to obtain a 50/50 balance of threatened and non-threatened species, all three groups improved their performance significantly, for metrics that put more wait in false negative errors, namely, sensitivity, precision and $F_{0.5}$. The lizards group, in particular, improved in all of the metrics using SMOTE, even though the improvement in specificity was not significant, as can be seen in Table 5.2.

Table 5.2: Results comparing the two version of the best algorithm for each of the groups where threatened was the minority class, one using the dataset without SMOTE and another with SMOTE. Values marked with a bold color represent models where performance, was, significantly, improved relative to the original dataset (without SMOTE), used as the baseline. Values of the Wilcoxon tests marked with one or two asterisks (*, **) indicate significance values higher than 95%, or 99%, respectively.

Sensitivity

	NO SMOTE		SMOTE		
	mean	sd	mean	sd	p-value
Lizards	0.963	0.017	**0.977**	0.004	0.024*
Snakes	0.945	0.032	**0.983**	0.013	0.001**
All species	0.945	0.015	**0.985**	0.005	0.001**

Specificity

	NO SMOTE		SMOTE		
	mean	sd	mean	sd	p-value
Lizards	0.717	0.025	0.731	0.030	0.053
Snakes	0.851	0.032	0.719	0.054	1.000
All species	0.778	0.022	0.704	0.024	0.998

Precision

	NO SMOTE		SMOTE		
	mean	sd	mean	sd	p-value
Lizards	0.640	0.022	**0.785**	0.019	0.001**
Snakes	0.699	0.045	**0.779**	0.032	0.001**
All species	0.683	0.023	**0.769**	0.015	0.001**

F1 0.5

	NO SMOTE		SMOTE		
	mean	sd	mean	sd	p-value
Lizards	0.874	0.016	**0.931**	0.004	0.001**
Snakes	0.882	0.028	**0.934**	0.010	0.001**
All species	0.877	0.016	**0.933**	0.006	0.001**

AUC

	NO SMOTE		SMOTE		
	mean	sd	mean	sd	p-value
Lizards	0.840	0.017	**0.854**	0.014	0.010**
Snakes	0.898	0.022	0.851	0.025	0.999
All species	0.862	0.015	0.845	0.013	1.000

TSS

	NO SMOTE		SMOTE		
	mean	sd	mean	sd	p-value
Lizards	0.680	0.034	**0.854**	0.027	0.010**
Snakes	0.797	0.044	0.703	0.049	0.999
All species	0.723	0.030	0.689	0.026	1.000

5.1.3 Predictions for DD and NE Species

Using the most performant model for each group (Table 5.1), this dissertation generated predictions for the DD and NE species. These predictions indicated that about a fifth of unassessed species of snakes (125 out of 612) and a half of amphisbaenians (10 out of 25) and crocodiles (1 out of 2) were considered threatened. Turtles were predicted to have more than half of threatened species (34 out of 56). For lizards, this value was almost three quarters (404 out of 613). The model for all species predicted 854 out of 1308 unassessed species as threatened,

which is more worrying view compared to 574, the sum of threatened species detected by the models of the five groups. Since the beginning of this work, IUCN continued to generate new assessments for reptile species [19]. From the 88 species that were a match between the model's predictions and the latest IUCN assessments – 47 lizards, 32 snakes, seven turtles, one crocodile, and one amphisbaenian.

Using the best group-specific models, 70 were correctly classified as non-threatened and two as threatened. There were eight false positives and eight false negatives. An analysis of the eight false negatives showed that mostly VU species were misclassified (5 out of 8) and the remaining were assessed as EN. Turtles was the group with more false negatives, followed by snakes and lizards. Seven of the eight species that the model misclassified as false positives were lizards assessed as LC and one was a turtle assessed as NT (check Table 5.3).

Table 5.3: List of errors made by the best model for the amphisbaenians, crocodiles, lizards, snakes and turtles datasets. False negatives (top) are species evaluated as non-threatened (nT, this is, Near Threatened, NT or Least Concern, LC) which were later assessed as threatened (T; this is, Endangered, EN or Vulnerable, VU), and false positives (bottom) vice-versa. Species that were false negatives/positives in both the all species and group-specific models were marked in bold.

Prediction	**Scientific name**	**Group**	**True assessment**	
nT	Aldabrachelys gigantea	turtles	VU	T
nT	**Cachryx defensor**	lizards	VU	T
nT	Chelonoidis denticulatu	turtles	VU	T
nT	Emys blandingii	turtles	EN	T
nT	**Hydrodynastes gigas**	snakes	EN	T
nT	Lissemys ceylonensis	turtles	VU	T
nT	Nawaran oenpelliensis	snakes	VU	T
nT	Trachemys dorbigni	turtles	EN	T
T	Abronia monticola	lizards	LC	nT
T	**Aspidoscelis neotesselatus**	lizards	LC	nT
T	**Cyclemys gemeli**	turtles	NT	nT
T	**Hemidactylus smithi**	lizards	LC	nT
T	**Lerista chalybura**	lizards	LC	nT
T	**Pseuderemias striatus**	lizards	LC	nT
T	**Scincopus fasciatus**	lizards	LC	nT
T	Uromastyx dispar	lizards	LC	nT

On the other hand, the best model using the group of all reptile species achieved a sensitivity of 0.80 and a specificity of 0.85, with only two false negatives and 12 false positives. The false negatives were a VU lizard and an EN snake, and the false positives were mostly LC (10 out of 12) and mostly lizards and snakes (5 species each), followed by a crocodiles and a turtle (check Table 5.4).

Table 5.4: List of errors made by the best model for all reptile species dataset. False negatives (top) are species evaluated as non-threatened (nT, this is, Near Threatened, NT or Least Concern, LC) which were later assessed as threatened (T; this is, Endangered, EN or Vulnerable, VU), and false positives (bottom) vice-versa. Species that were false negatives/positives in both the all species and group-specific models were marked in bold.

Prediction	**Scientific name**	**Group**	**True assessment**	
nT	**Cachryx defensor**	lizards	VU	T
nT	**Hydrodynastes gigas**	snakes	EN	T
T	Amblyodipsas rodhaini	snakes	LC	nT
T	Atractaspis reticulata	snakes	LC	nT
T	**Aspidoscelis neotesselatus**	lizards	LC	nT
T	Contia longicaudae	snakes	LC	nT
T	Crocodylus johnsoni	crocodiles	LC	nT
T	**Cyclemys gemeli**	turtles	NT	nT
T	**Hemidactylus smithi**	lizards	LC	nT
T	**Lerista chalybura**	lizards	LC	nT
T	Letheobia stejnegeri	snakes	LC	nT
T	**Pseuderemias striatus**	lizards	LC	nT
T	**Scincopus fasciatus**	lizards	LC	nT
T	Suta flagellum	snakes	LC	nT

5.1.4 Maps of threatspots

According to the models' predictions, the threatspot maps showed the area occupied by threatened species could be larger than what is currently known. Figure 5.3 shows the maps for amphisbaenians, crocodiles, lizards, snakes, turtles, and all reptile species. Maps showing the ratio of known and predicted threatened species in respect to the total of species in the group dataset are displayed in Figure 5.4. Predictions showed no new threatspots for amphisbaenians (Figure 5.3a), but highlighted two concerning areas in Guinea and on the Central American islands (Figure 5.4a). They also showed there may be unassessed threatened crocodiles in a large range of Africa (Figure 5.3b), with a hotspot for threatened species ratio in India (Figure 5.4b). Lizards showed new predicted threatened species over Africa and India (Figure 5.3c) and a threatspot on the eastern South American coast, New Zealand and other south-west Pacific Ocean islands (Figure 5.4c). Snakes showed little new threatspots (Figure 5.3d), and areas of most concern in Madagascar, several south-west Pacific Ocean islands (Figure 5.4d) and western South American coast. Turtles have a much larger distribution which is reflected in the vast area showing threatened species. Predictions showed that their range may be even greater, extending to Central and South America, southern Africa and Australia (Figure 5.3e). Their new threatspots are over Central and Eastern Africa, Brazil, and several countries on the Pacific including Vietnam, Malaysia, Indonesia, Philippines and parts of Australia (Figure 5.4e). Lastly, the model contained all species had predictions consistent to the group-specific models, showing most of South America, specially Brazil, North America, specially Florida, Africa, India and

the Pacific Islands, to have most predicted threatened species. New Zealand and several smaller islands on the South Pacific, like the Fiji archipelago and New Caledonia, have the highest ratio of threatened species, according to the all species model.

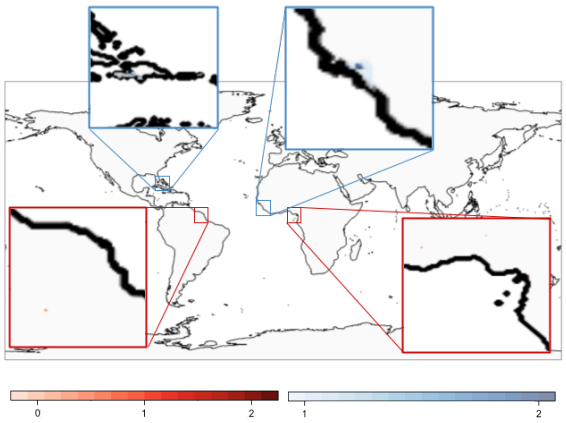
5.2 Discussion

5.2.1 Predictive modelling

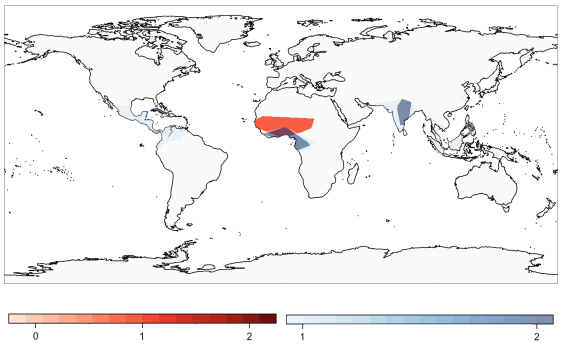
Both the specific models for species groups and the model using all species achieved very promising results. In the new predictions that were matched to IUCN new assessments, the model trained using all the reptile species available had few mistakes, and, specially, few false negatives (only 2 out of 10 threatened species) showing this is indeed a valid method to help boost conservation assessment efforts. As for the modelling technique, it has been demonstrated that RF generated the best models for three of the six groups, which is consistent with the fact that many studies on this domain used this algorithm, due to its ability to deal with overfit issues and to handle large quantities of both data points and predictor variables, and it typically performed well [53]. RF has been shown to out-perform other machine learning methods for certain clades [56]. Previous authors devised a pipeline and applied it to the global flora [53]. Unlike us, they generated different models for each continent. This dissertation created a model for each taxonomic subgroup, so the variables that better described each group could then be picked considering ecological differences across groups. They reported that the accuracy of their classifiers was in the range of 73-82%. In this study, the best models of all six groups achieved a $F_{0.5}$ higher than 0.87 and only one group, the turtles, had an AUC lower than 0.80. In another study, the authors did not use machine learning techniques, but were able to generate preliminary assessments of the risk level of 22,036 plant species at a continental scale, using IUCN Criteria A and B [52].

Additionally, variable selection was an important step included in the proposed pipeline. For four out of the six groups, the use of a variable selection step generated the best models. Using global data from a very large number of terrestrial plant species and RF, recent studies showed the utility of coarse-scale data and variable selection for assessing the extinction risk of plant species [53, 57]. The authors of [59] used herbarium specimens' data to compare machine learning methods, and specifically RF, with more traditional approaches like directly assessing AOO and EOO, or the specimen count, against IUCN criteria, and RF outperformed all other approaches in terms of accuracy.

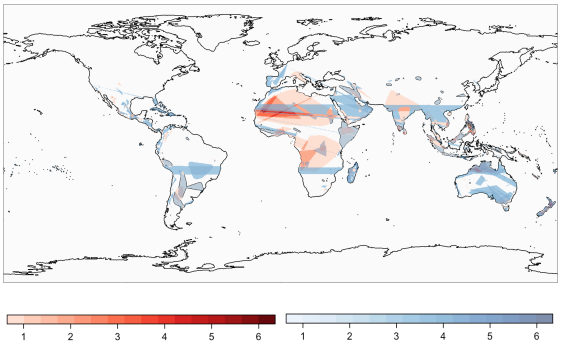
Regarding the variable importance, results are also consistent with studies performed with flora species where the authors compared the results from two datasets, one with spatial data only and another with spatial and morphological and concluded that the spatial-only performed better [53]. Results also suggest that variables based on statistical dispersion measures (e.g., standard deviation) perform also generally better than other ecological ones since these portray species ecological tolerance ranges [53].



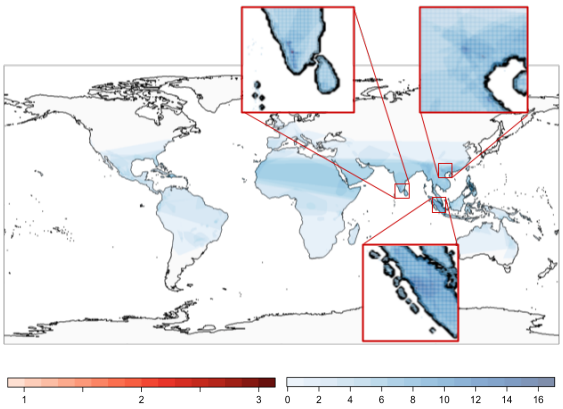
(a) Amphisbaenians



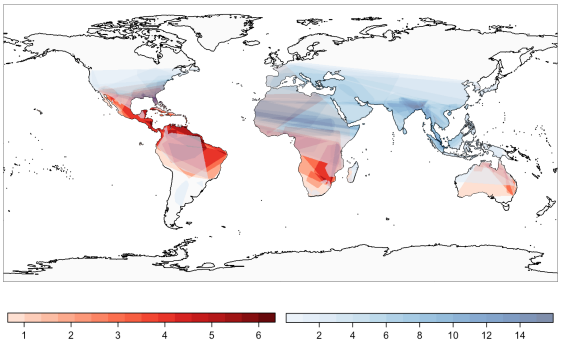
(b) Crocodiles



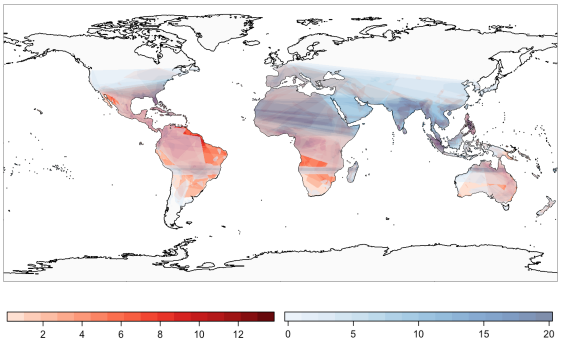
(c) Lizards



(d) Snakes

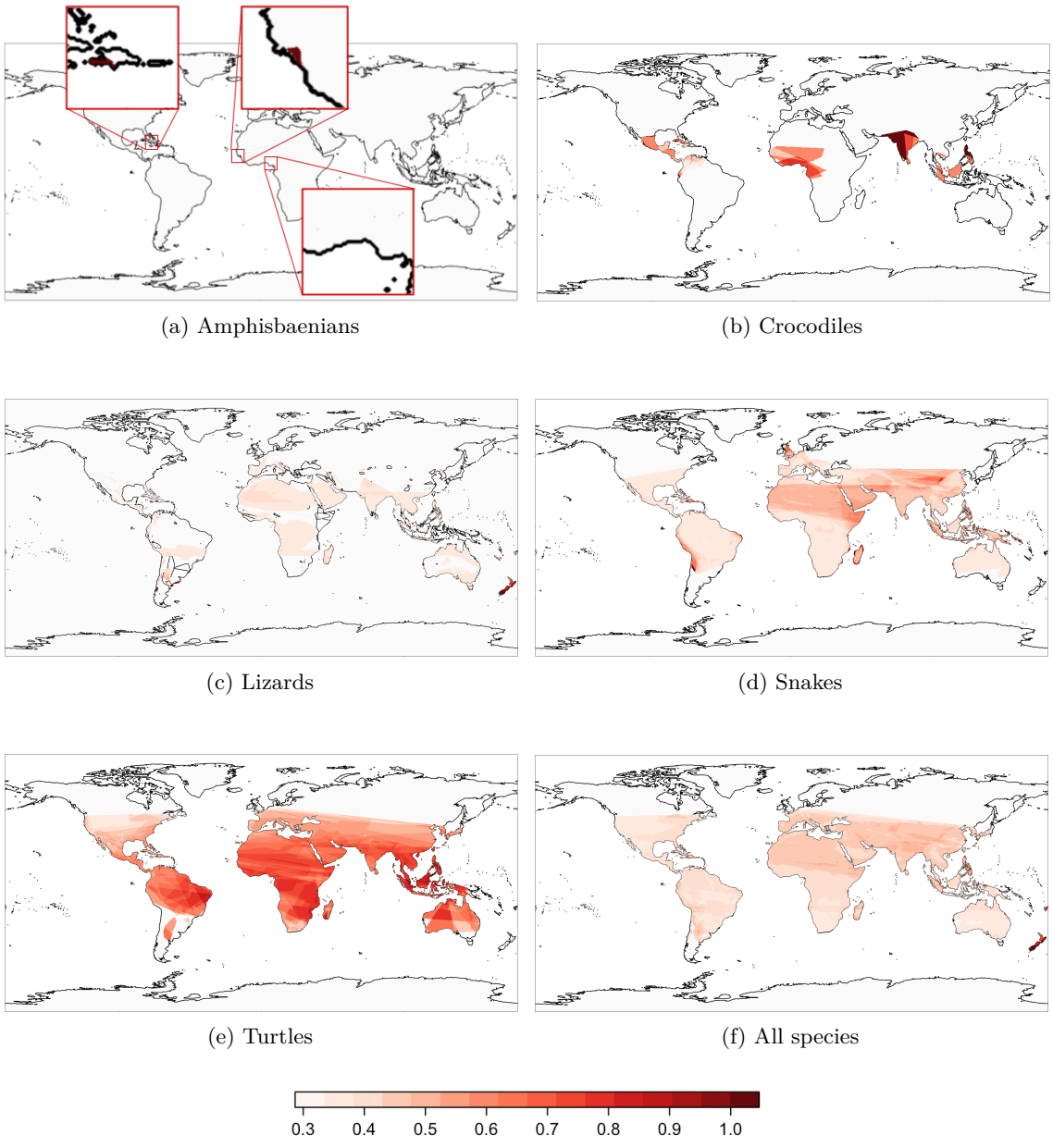


(e) Turtles



(f) All species

Figure 5.3: Threatspot gap maps showing the predicted areas with threatened species (in red), versus the areas occupied by species already classified as threatened by IUCN (in blue). Darker colours indicate higher species richness. Maps for (a) amphisbaenians, (b) crocodiles, (c) lizards, (d) snakes, (e) turtles, and (f) all species, respectively. Some small details referent to predicted (red outline) and known (blue outline) threatened species are highlighted using zoom.



(a) Amphisbaenians (b) Crocodiles

(c) Lizards (d) Snakes

(e) Turtles (f) All species

Figure 5.4: New threatspot maps showing the ratio of the sum of the known and predicted by the total of species in the group dataset. Darker colours indicate higher ratios of threatened species. Maps for (a) amphisbaenians, (b) crocodiles, (c) lizards, (d) snakes, (e) turtles, and (f) all species, respectively.

Previous state of the art works predicting the conservation status of species did not try to tackle the imbalance class problem, or they used simple techniques, like under-sampling. SMOTE was able to significantly improve model results, specially when using the metrics that focus on the class of interest, showing the potential of advanced pre-processing techniques to balance conservation data.

In the new predictions that were matched to IUCN new assessments, the model trained using all the reptile species available had less mistakes, and, specially, less false negatives (only two against eight from the group-specific models). Specifically, the turtles model, generated many false positives that the model of all species was able to classify correctly. This indicates that certain groups may benefit from a specific model tuned to its data and with adjusted set of features, but others may benefit from more data being used in the training process. The turtles group had only 147 records that could be used for training, so this group probably benefited from the remaining 3113 species present in the all species dataset.

5.2.2 Species conservation

When looking into other predictions made for DD and NE species, using the most complete dataset with spatial variables only the authors of [53] placed 3.4 to 13.6% plant species as non-Least Concern (non-LC) at a probability of >0.80 and 18.7 to 41.9% at a probability of >0.60. The authors of [52], when applying Criterion A predicted 22.8% of species to be threatened, 22.1% using Criteria B and 31.7% using both. These estimates look conservative in comparison to this work's predictions for reptile species, where only one group (snakes) presented less than half of unassessed species as threatened. The predictions suggested that lizards are the reptile group with the highest percentage of threatened species, followed by turtles, amphisbaenians and crocodiles, and lastly, snakes. This advises for more conservation action to be put forward for these groups, lizards in particular [12]. On the other hand, the smaller percentage of predicted threatened species for a group like snakes (only a fifth of the DD and NE species) versus the lizards and turtles (almost three quarters and more than half, respectively) may indicate biases in the manual evaluation process caused by experts being cautious when assessing endemic and island restricted species, as previously documented for the vascular flora of Cabo Verde [103].

It is important to note that the errors presented by the models, especially the false positives, could indicate that these species are indeed threatened, or very close to be. This is especially relevant due to the very high specificity of the models, even in the blind dataset, and for the species that were misplaced by both the group specific and the all species models.

Comparing the threatspots generated using the predictions to the maps of threatened species richness for tetrapods, and reptiles in particular, generated by the authors of [22] the same regions are highlighted. Particularly, India, Central Africa and Madagascar, some regions of South America and several Pacific islands. The models also predicted a strong presence of threatened species in Central America, Australia, New Zealand and across a larger range of South America and Africa.

Chapter 6

Conclusions

This chapter started by summarizing the work done in this dissertation and the conclusions of the study. It then presents the main contributions and the limitations and future work, in Sections 6.1 and 6.2, respectively.

This work strived to create a time and cost-efficient tool that provided preliminary threat status over a large set of unassessed species, providing the first picture of the conservation status of entire groups and allowing governmental and non-governmental organizations to better prioritize conservation efforts. The solution is in the form of a pipeline, which was applied to the Reptilia class, without loss of generality. The incredible volume of presence records from a large number of species and freely available EGVs can be used to perform general model-assisted assessments across other under-assessed taxonomic groups following the proposed framework, which can then be complemented by detailed studies. The proposed pipeline included several steps to tackle very relevant issues in the application domain such as data curation and imputation, variable generation and selection, hyperparameter selection as well as algorithm and dataset comparison [53].

Overall, the models performed well according to standard metrics and metrics that consider the higher cost of false negatives. On the blind dataset, the model for all species was balanced, and the models for the amphisbaenians, crocodiles, lizards, snakes and turtles performed very well in terms of specificity, but were limited on their sensitivity. This shows how data quality issues may lead to overly positive outcomes, which needs to be a consideration when applying machine learning models to this domain. Results showed that some models may be improved by using more variables besides those traditionally used by IUCN evaluators, which reside mostly on the AOO and EOO. Those variables are, in decreasing order, geographic (mainly the latitudinal range of species, but also their longitudinal distribution), bioclimatic indices, habitat, and topographic heterogeneity. However, to benefit from the inclusion of such variables, more data both quantity and quality-wise are needed as the scale of assessment becomes finer and more detailed.

6.1 Contributions

In sum, the main contributions of this work were: (i) a curated dataset of ecogeographical and conservation data for reptiles, (ii) a processing pipeline applicable to multiple taxa, (iii) a large scale prediction model to automate the assessment of conservation status of one of the most under-assessed taxonomic groups of vertebrates, the reptiles, (iv) a comparison and comment on predictor variable importance, including experimental results and discussion on the usage of different groups of variables to augment two traditional predictors used by IUCN, (v) predictions for the threat status of 1,308 DD and NE reptile species, (vi) threatspot and gap maps for terrestrial herpetological richness and for each reptile clade, and, (vii) demonstrating the potential of techniques like SMOTE to balance the dataset and improve performance.

6.2 Limitations and future work

In the future, other algorithms could be tested, for instance, Neural Networks (NN) since they had promising results in similar studies [61].

Improvements to the pipeline can also open up the possibility of forecasting species threat status and addressing climate change impacts by feeding models data from climate scenarios provided by United Nation's (UN) Intergovernmental Panel on Climate Change (IPCC). These scenarios are a set of projections of how the climate may evolve in the future, according to different greenhouse gas emission pathways. Such application is deemed critical for domain specialists to anticipate species vulnerabilities in the face of global environmental change [31]. Moreover, other sets of variables that depict anthropogenic disturbances on species habitats, such as land use change, landscape fragmentation, wildfires, invasive species, which influence the amount, the connectivity and the ability of species finding suitable environments could be tested [57].

Some of the taxonomic groups had very few species, leading to a larger performance variation across algorithms. This reinforces the importance of gathering more species presence (or abundance) data and including a broader set of species for improving training conditions for the models [57]. In addition, these results suggest that when deciding on how to group species for modelling, one should balance between having groups that reflect species with similar taxonomic, phylogenetic and/or functional characteristics (so these can be well represented by the same set of variables), and also, having enough data points for the methods to be able to properly learn and generalize. Additionally, some species presented data points very far from their typical habitat, which could correspond to species present in museums, or purely, to location errors. Models are very dependent on the quality of the data since it is the raw material of any data product [59]. Without data quality there is no way to achieve good modelling results. For this reason, data repositories need to provide proper data curation, not only by ensuring taxonomic correctness, but also that the coordinates are precise and make sense given the species distribution. Additionally,

the class imbalance problem [38] could concern the lizards, snakes and all species groups/datasets, because they had a, comparatively, higher imbalance and, their minority class was precisely the class of interest (i.e., Threatened class), as was also seen in other studies [53]. And as previously mentioned, in a modelling task, the performance of the models should be estimated on a testing set where no parameter tuning has been done. The low amount of species data limited this work, and is likely to limit most models of the threat status of species. For that reason, it should be noted that the performance metrics estimated using CV may be slightly optimistic.

The models were based mostly in the IUCN Criteria B (restricted distribution), but Criteria A (declining distribution) is also a very important basis to determine the threat status (IUCN, 2012). There was a limitation that prevented us from including this criteria: the lack of abundance data periodically added to repositories for each species. For the models to include a temporal dimension, one would need to have a much higher amount of data for each species, distributed over the years. Without it, a possible approach could leverage species distribution models (SDMs) to predict the distribution range of each species over time [93].

Using the convex hulls was a good option to approximate the species distribution. It is the technique used by IUCN themselves, but it generates polygons sharp polygons, while the actual distribution of species is likely to be much more even. A possible improvement would be to use the information on the species presences and, possibly, absences, and SDMs, to estimate their potential range, and use that output to generate the maps.

Also, in the future, the threat maps, combined with socioeconomic knowledge of the underlying countries, could be used to study the correlation between the threat level and countries gross domestic product (GDP) to understand if there is any correlation between a country's economic status and its conservation efforts.

Lastly, SMOTE was able to significantly improve model results for this data, specially for the metrics that focus on the class of interest, showing the potential of techniques to balance the dataset in this domain, other parameters and methods could be explored and compared. Due to the chronology followed by this dissertation, the predictions for DD and NE species were done using the original models. SMOTE was an attempt to improve model results even further, after the fact. Since it did improve them for some of the groups, so those models should be the ones used to generate the final predictions and outputs.

Appendix A

Model results

The following tables show the results of final models for each combination of model and dataset. Each table details the values for the six metrics used, sensitivity, specificity, precision, $F_{0.5}$, AUC and TSS, for each of the six groups of species. They also contain the results of the paired one-sided Wilcoxon tests done using the *AOO_EOO* dataset as baseline, or the *All EGVs* or *All EGVs_FS* in the case of the tests done between the *EcoV* and *All EGVs* datasets, see eco-all columns. Values marked with a bold color represent models where performance, for that particular metric and algorithm, was, significantly, improved relative to the *AOO_EOO* dataset. Values of the Wilcoxon tests marked with one or two asterisks (*, **) indicate significance values higher than 95%, or 99%, respectively.

Table A.1: Results of final models for each combination of model and dataset, and per evaluation metric. Values marked with a bold color represent models where performance, for that particular metric and algorithm, was according to the Wilcoxon test, significantly, improved relative to the *AOO_EOO* dataset. Values marked with one or two asterisks (*, **) indicate a confidence level higher than 95%, or 99%, respectively.

Amphisbaenians

Sensitivity

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.950	0.158	0.800	0.350	0.978	0.800	0.350	0.978	0.579	0.750	0.354	0.986	0.950	0.158	1.000	0.986
xgboost	0.800	0.422	0.600	0.394	0.940	0.750	0.354	0.807	0.978	0.750	0.264	0.672	0.650	0.474	0.963	0.294
c50	0.900	0.211	0.750	0.354	0.907	0.700	0.350	0.960	0.412	0.600	0.394	0.988	0.900	0.211	0.579	0.977
knn	0.550	0.497	0.800	0.258	0.112	0.667	0.408	0.361	0.259	0.800	0.258	0.112	**0.900**	0.211	0.050*	0.970
glm	1.000	0.000	0.683	0.364	0.986	0.683	0.364	0.986	1.000	0.750	0.354	0.981	0.750	0.354	0.981	0.579
rpart	0.800	0.422	0.600	0.516	0.970	0.600	0.516	0.970	1.000	0.600	0.516	0.970	0.600	0.516	0.970	1.000

Specificity

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.700	0.350	0.650	0.412	0.707	0.700	0.350	0.681	0.661	0.750	0.354	0.386	0.750	0.354	0.500	0.681
xgboost	0.650	0.337	0.650	0.412	0.579	0.650	0.337	0.681	0.681	0.650	0.412	0.579	0.750	0.354	0.173	0.862
c50	0.750	0.354	0.550	0.438	0.931	0.650	0.412	0.977	0.862	0.700	0.422	0.814	0.750	0.354	1.000	0.814
knn	0.650	0.474	0.650	0.337	0.556	0.700	0.350	0.445	0.977	0.650	0.337	0.556	0.650	0.412	0.543	0.607
glm	0.750	0.354	0.550	0.438	0.843	0.550	0.438	0.843	1.000	0.700	0.422	0.814	0.750	0.354	0.681	0.715
rpart	0.200	0.422	0.400	0.516	0.173	0.400	0.516	0.173	1.000	0.400	0.516	0.173	0.400	0.516	0.173	1.000

Precision

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.825	0.194	0.758	0.325	0.865	0.758	0.325	0.865	0.542	0.824	0.222	0.789	0.875	0.163	0.500	0.789
xgboost	0.713	0.309	0.694	0.358	0.099	0.725	0.314	0.186	0.605	0.775	0.261	0.138	0.821	0.170	0.500	0.607
c50	0.858	0.193	0.713	0.232	0.956	0.787	0.232	0.814	0.572	0.781	0.263	0.814	0.875	0.163	0.500	0.789
knn	0.694	0.245	0.742	0.237	0.977	0.731	0.333	0.574	0.707	0.742	0.237	0.977	0.825	0.194	0.133	0.901
glm	0.875	0.163	0.731	0.220	0.948	0.731	0.220	0.948	1.000	0.806	0.257	0.789	0.843	0.197	0.814	0.769
rpart	0.542	0.077	0.500	0.000	1.000	0.500	0.000	1.000	1.000	0.500	0.000	1.000	0.500	0.000	1.000	1.000

F1 0.5

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.916	0.152	0.856	0.179	0.962	0.868	0.197	0.969	0.664	0.822	0.231	0.950	0.922	0.136	0.500	0.950
xgboost	0.947	0.045	0.739	0.242	0.972	0.808	0.207	0.963	0.969	0.741	0.244	0.979	0.889	0.152	0.977	0.972
c50	0.881	0.191	0.785	0.209	0.899	0.754	0.216	0.973	0.500	0.739	0.242	0.969	0.878	0.174	0.715	0.889
knn	0.827	0.149	0.785	0.247	0.795	0.816	0.199	0.312	0.708	0.785	0.247	0.795	0.872	0.186	0.500	0.899
glm	0.966	0.044	0.725	0.236	0.993	0.725	0.236	0.993	1.000	0.817	0.239	0.969	0.819	0.215	0.969	0.606
rpart	0.852	0.035	0.833	0.000	1.000	0.833	0.000	1.000	1.000	0.833	0.000	1.000	0.833	0.000	1.000	1.000

AUC

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.825	0.206	0.725	0.249	0.883	0.750	0.264	0.978	0.661	0.750	0.236	0.907	0.850	0.175	0.500	0.940
xgboost	0.725	0.249	0.625	0.270	0.972	0.700	0.230	0.807	0.978	0.700	0.258	0.807	0.700	0.197	0.814	0.556
c50	0.825	0.206	0.650	0.211	0.981	0.675	0.206	0.962	0.725	0.650	0.242	0.979	0.825	0.169	0.579	0.967
knn	0.600	0.175	0.725	0.249	0.084	0.683	0.242	0.213	0.375	0.725	0.249	0.084	**0.775**	0.219	0.032*	0.771
glm	0.875	0.177	0.617	0.209	0.989	0.617	0.209	0.989	1.000	0.725	0.275	0.969	0.750	0.204	0.972	0.644
rpart	0.500	0.000	0.500	0.000	1.000	0.500	0.000	1.000	1.000	0.500	0.000	1.000	0.500	0.000	1.000	1.000

TSS

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.650	0.412	0.450	0.497	0.883	0.500	0.527	0.978	0.661	0.500	0.471	0.907	0.700	0.350	0.500	0.940
xgboost	0.450	0.497	0.250	0.540	0.972	0.400	0.459	0.807	0.978	0.400	0.516	0.807	0.400	0.394	0.814	0.556
c50	0.650	0.412	0.300	0.422	0.981	0.350	0.412	0.962	0.725	0.300	0.483	0.979	0.650	0.337	0.579	0.967
knn	0.200	0.350	0.450	0.497	0.084	0.367	0.483	0.213	0.375	0.450	0.497	0.084	**0.550**	0.438	0.032*	0.771
glm	0.750	0.354	0.233	0.417	0.989	0.233	0.417	0.989	1.000	0.450	0.550	0.969	0.500	0.408	0.972	0.644
rpart	0.000	0.000	0.000	0.000	1.000	0.000	0.000	1.000	1.000	0.000	0.000	1.000	0.000	0.000	1.000	1.000

Crocodiles

Sensitivity

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.917	0.180	0.950	0.158	0.500	1.000	0.000	0.186	0.977	0.950	0.158	0.500	1.000	0.000	0.186	0.977
xgboost	0.700	0.270	0.900	0.211	0.080	**0.867**	0.219	0.049*	0.500	**0.950**	0.158	0.027*	**1.000**	0.000	0.016*	0.977
c50	0.817	0.337	0.917	0.180	0.293	0.867	0.322	0.293	0.500	0.767	0.251	0.717	1.000	0.000	0.091	0.988
knn	0.917	0.180	0.917	0.180	0.574	1.000	0.000	0.186	0.963	1.000	0.000	0.186	1.000	0.000	0.186	1.000
glm	0.917	0.180	0.600	0.402	0.972	0.600	0.402	0.972	1.000	0.950	0.158	0.500	1.000	0.000	0.186	0.977
rpart	1.000	0.000	1.000	0.000	1.000	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000	0.000	1.000	1.000

Specificity

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.800	0.422	0.750	0.264	0.604	0.800	0.258	0.556	0.725	0.700	0.258	0.696	0.950	0.158	0.186	0.976
xgboost	0.750	0.354	0.750	0.354	0.556	0.700	0.422	0.666	0.391	0.750	0.264	0.579	0.800	0.258	0.412	0.807
c50	0.900	0.211	0.650	0.242	0.958	0.600	0.211	0.995	0.383	0.750	0.264	0.890	0.600	0.211	0.995	0.074
knn	0.700	0.422	0.800	0.258	0.304	0.700	0.258	0.500	0.173	0.800	0.258	0.294	0.700	0.258	0.546	0.242
glm	0.700	0.422	0.650	0.412	0.672	0.650	0.412	0.672	1.000	0.750	0.264	0.396	**0.900**	0.211	0.036*	0.890
rpart	0.000	0.000	0.000	0.000	1.000	0.000	0.000	1.000	1.000	0.000	0.000	1.000	0.000	0.000	1.000	1.000

Precision

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.883	0.249	0.833	0.188	0.633	0.875	0.163	0.706	0.709	0.800	0.185	0.714	0.967	0.105	0.186	0.958
xgboost	0.800	0.258	0.852	0.201	0.294	0.825	0.250	0.500	0.446	0.833	0.188	0.446	0.875	0.163	0.221	0.769
c50	0.850	0.337	0.758	0.178	0.784	0.675	0.273	0.991	0.333	0.825	0.194	0.633	0.750	0.136	0.911	0.136
knn	0.817	0.254	0.850	0.200	0.387	0.808	0.167	0.458	0.290	0.883	0.153	0.276	0.808	0.167	0.664	0.133
glm	0.817	0.254	0.683	0.404	0.828	0.683	0.404	0.828	1.000	0.833	0.188	0.416	**0.933**	0.141	0.049*	0.829
rpart	0.520	0.042	0.520	0.042	1.000	0.520	0.042	1.000	1.000	0.520	0.042	1.000	0.520	0.042	1.000	1.000

F1 0.5

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.900	0.185	0.919	0.152	0.400	0.966	0.044	0.337	0.709	0.910	0.150	0.472	0.991	0.029	0.091	0.958
xgboost	0.702	0.242	0.878	0.190	0.054	**0.845**	0.210	0.026*	0.337	**0.919**	0.152	0.026*	**0.966**	0.044	0.010*	0.865
c50	0.913	0.181	0.874	0.163	0.803	0.906	0.097	0.667	0.802	0.759	0.209	0.915	0.933	0.037	0.602	0.974
knn	0.886	0.187	0.898	0.175	0.444	0.948	0.045	0.358	0.821	0.969	0.041	0.134	0.948	0.045	0.337	0.133
glm	0.882	0.178	0.733	0.216	0.882	0.733	0.216	0.882	1.000	0.919	0.152	0.311	**0.982**	0.038	0.029*	0.829
rpart	0.843	0.021	0.843	0.021	1.000	0.843	0.021	1.000	1.000	0.843	0.021	1.000	0.843	0.021	1.000	1.000

AUC

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.858	0.267	0.850	0.175	0.633	0.900	0.129	0.543	0.802	0.825	0.169	0.715	0.975	0.079	0.091	0.977
xgboost	0.725	0.233	0.825	0.206	0.114	0.783	0.267	0.304	0.500	0.850	0.175	0.061	0.900	0.129	0.053	0.885
c50	0.858	0.267	0.783	0.172	0.782	0.733	0.211	0.967	0.412	0.758	0.154	0.837	0.800	0.105	0.810	0.748
knn	0.808	0.275	0.858	0.197	0.362	0.850	0.129	0.375	0.425	0.900	0.129	0.303	0.850	0.129	0.295	0.242
glm	0.808	0.258	0.625	0.297	0.927	0.625	0.297	0.927	1.000	0.850	0.175	0.471	**0.950**	0.105	0.027*	0.912
rpart	0.500	0.000	0.500	0.000	1.000	0.500	0.000	1.000	1.000	0.500	0.000	1.000	0.500	0.000	1.000	1.000

TSS

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.717	0.533	0.700	0.350	0.633	0.800	0.258	0.543	0.802	0.650	0.337	0.715	0.950	0.158	0.091	0.977
xgboost	0.450	0.465	0.650	0.412	0.084	0.567	0.534	0.304	0.500	0.700	0.350	0.061	0.800	0.258	0.053	0.885
c50	0.717	0.533	0.567	0.344	0.782	0.467	0.422	0.967	0.412	0.517	0.309	0.837	0.600	0.211	0.810	0.748
knn	0.617	0.550	0.717	0.393	0.362	0.700	0.258	0.375	0.425	0.800	0.258	0.303	0.700	0.258	0.295	0.242
glm	0.617	0.516	0.250	0.594	0.927	0.250	0.594	0.927	1.000	0.700	0.350	0.471	**0.900**	0.211	0.027*	0.912
rpart	0.000	0.000	0.000	0.000	1.000	0.000	0.000	1.000	1.000	0.000	0.000	1.000	0.000	0.000	1.000	1.000

Lizards

Sensitivity

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.933	0.019	0.836	0.034	0.998	**0.961**	0.017	0.003**	0.998	0.811	0.027	0.998	**0.963**	0.017	0.003**	1.000
xgboost	0.942	0.023	0.851	0.022	0.998	**0.956**	0.015	0.006**	0.998	0.812	0.025	0.998	**0.958**	0.018	0.009**	0.998
c50	0.914	0.027	0.741	0.057	1.000	0.909	0.034	0.637	1.000	0.718	0.059	0.998	0.912	0.034	0.528	1.000
knn	0.945	0.022	0.777	0.050	0.998	0.805	0.039	1.000	0.986	0.777	0.032	0.998	0.823	0.034	0.998	0.998
glm	0.922	0.023	0.743	0.020	0.998	0.881	0.033	0.999	0.998	0.757	0.023	1.000	0.884	0.030	0.997	1.000
rpart	0.915	0.015	0.668	0.082	1.000	0.914	0.014	0.963	1.000	0.690	0.085	1.000	0.914	0.014	0.963	1.000

Specificity

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.715	0.034	**0.750**	0.033	0.007**	0.711	0.026	0.639	0.005**	**0.740**	0.029	0.038*	0.717	0.025	0.383	0.026*
xgboost	0.738	0.031	0.703	0.028	0.991	0.703	0.023	0.998	0.439	0.712	0.032	0.976	0.701	0.028	0.998	0.116
c50	0.721	0.048	0.695	0.085	0.882	0.733	0.054	0.500	0.958	0.719	0.057	0.472	0.742	0.042	0.363	0.904
knn	0.713	0.025	**0.746**	0.029	0.007**	**0.738**	0.038	0.046*	0.120	0.730	0.035	0.078	**0.737**	0.037	0.023*	0.784
glm	0.718	0.069	0.732	0.031	0.238	0.711	0.037	0.784	0.033*	0.735	0.024	0.312	0.713	0.033	0.722	0.012*
rpart	0.757	0.025	0.744	0.067	0.652	0.756	0.025	0.963	0.652	0.723	0.073	0.923	0.756	0.025	0.963	0.923

Precision

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.632	0.030	0.637	0.031	0.312	0.635	0.023	0.246	0.423	0.620	0.024	0.839	0.640	0.022	0.097	0.995
xgboost	0.653	0.032	0.600	0.024	1.000	0.628	0.021	1.000	1.000	0.597	0.026	1.000	0.626	0.023	1.000	0.999
c50	0.633	0.043	0.566	0.057	1.000	0.643	0.042	0.500	1.000	0.574	0.038	1.000	0.650	0.035	0.277	1.000
knn	0.632	0.024	0.615	0.035	0.903	0.617	0.038	0.839	0.423	0.602	0.031	0.999	0.621	0.034	0.935	0.968
glm	0.635	0.055	0.592	0.031	0.997	0.615	0.032	0.958	0.995	0.599	0.023	0.986	0.617	0.030	0.947	0.993
rpart	0.664	0.025	0.582	0.042	1.000	0.662	0.024	0.969	1.000	0.571	0.048	1.000	0.662	0.024	0.969	1.000

F1 0.5

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.852	0.019	0.787	0.028	1.000	**0.872**	0.018	0.001**	1.000	0.763	0.022	1.000	**0.874**	0.016	0.001**	1.000
xgboost	0.865	0.025	0.785	0.019	1.000	0.866	0.017	0.246	1.000	0.757	0.020	1.000	0.866	0.017	0.246	1.000
c50	0.839	0.029	0.695	0.036	1.000	0.839	0.024	0.500	1.000	0.682	0.041	1.000	0.843	0.027	0.318	1.000
knn	0.859	0.022	0.738	0.044	1.000	0.759	0.034	1.000	0.976	0.734	0.027	1.000	0.772	0.027	1.000	1.000
glm	0.845	0.025	0.707	0.021	1.000	0.810	0.028	0.999	1.000	0.719	0.020	1.000	0.813	0.027	0.999	1.000
rpart	0.851	0.016	0.646	0.059	1.000	0.849	0.015	0.969	1.000	0.659	0.061	1.000	0.849	0.015	0.969	1.000

AUC

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.824	0.021	0.793	0.023	0.998	**0.836**	0.018	0.014*	1.000	0.775	0.017	1.000	**0.840**	0.017	0.005**	1.000
xgboost	0.840	0.025	0.777	0.018	1.000	0.830	0.017	0.986	1.000	0.762	0.018	1.000	0.829	0.018	0.986	1.000
c50	0.817	0.032	0.718	0.031	1.000	0.821	0.024	0.318	1.000	0.718	0.025	1.000	0.827	0.024	0.062	0.998
knn	0.829	0.020	0.762	0.033	1.000	0.772	0.029	1.000	0.920	0.754	0.022	1.000	0.780	0.023	1.000	1.000
glm	0.820	0.035	0.738	0.021	1.000	0.796	0.026	0.998	1.000	0.746	0.017	1.000	0.798	0.024	0.993	1.000
rpart	0.836	0.017	0.706	0.026	1.000	0.835	0.016	0.969	1.000	0.706	0.028	1.000	0.835	0.016	0.969	1.000

TSS

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.648	0.042	0.586	0.045	0.998	**0.672**	0.037	0.014*	1.000	0.551	0.035	1.000	**0.680**	0.034	0.005**	1.000
xgboost	0.680	0.049	0.554	0.035	1.000	0.660	0.035	0.986	1.000	0.524	0.036	1.000	0.659	0.036	0.986	1.000
c50	0.635	0.063	0.436	0.061	1.000	0.643	0.048	0.318	1.000	0.437	0.050	1.000	0.654	0.049	0.062	0.998
knn	0.657	0.041	0.523	0.065	1.000	0.543	0.059	1.000	0.920	0.507	0.044	1.000	0.560	0.046	1.000	1.000
glm	0.641	0.070	0.475	0.042	1.000	0.592	0.052	0.998	1.000	0.492	0.034	1.000	0.596	0.049	0.993	1.000
rpart	0.673	0.034	0.412	0.051	1.000	0.670	0.033	0.969	1.000	0.413	0.056	1.000	0.670	0.033	0.969	1.000

Snakes

Sensitivity

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.942	0.032	0.914	0.040	0.968	**0.967**	0.028	0.007**	0.997	0.910	0.044	0.984	**0.973**	0.028	0.007**	0.998
xgboost	0.949	0.025	0.930	0.039	0.975	**0.965**	0.029	0.011*	0.993	0.926	0.040	0.993	**0.967**	0.024	0.007**	0.995
c50	0.856	0.077	0.797	0.088	0.916	0.838	0.052	0.807	0.923	0.809	0.067	0.946	0.821	0.051	0.899	0.779
knn	0.945	0.032	0.908	0.051	0.986	0.959	0.033	0.276	0.998	0.914	0.035	0.984	0.947	0.033	0.500	0.997
glm	0.971	0.025	0.918	0.037	0.995	0.953	0.035	0.979	0.995	0.920	0.036	0.995	0.949	0.037	0.986	0.993
rpart	0.846	0.081	0.825	0.084	0.763	0.838	0.083	0.663	0.762	0.809	0.085	0.931	0.840	0.081	0.534	0.967

Specificity

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.820	0.036	0.756	0.041	0.997	0.722	0.036	0.998	0.005**	0.756	0.042	0.997	0.710	0.037	1.000	0.003**
xgboost	0.794	0.042	0.750	0.038	0.989	0.750	0.041	1.000	0.637	0.752	0.028	0.989	0.757	0.039	0.997	0.736
c50	0.867	0.043	0.818	0.050	0.984	0.878	0.028	0.181	0.995	0.820	0.061	0.976	0.881	0.029	0.193	0.997
knn	0.851	0.032	0.751	0.045	0.998	0.716	0.051	0.998	0.009**	0.741	0.043	1.000	0.745	0.043	0.998	0.620
glm	0.621	0.071	**0.759**	0.041	0.001**	**0.743**	0.046	0.001**	0.071	**0.751**	0.043	0.003**	**0.745**	0.047	0.002**	0.305
rpart	0.870	0.033	0.754	0.123	0.995	0.867	0.030	0.825	0.997	0.776	0.135	0.995	0.868	0.032	0.778	0.991

Precision

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.657	0.044	0.577	0.041	1.000	0.558	0.032	1.000	0.032*	0.576	0.042	0.999	0.549	0.035	1.000	0.014*
xgboost	0.627	0.048	0.575	0.036	0.998	0.585	0.040	1.000	0.812	0.574	0.028	0.999	0.591	0.041	0.999	0.935
c50	0.706	0.056	0.618	0.054	1.000	0.715	0.042	0.312	1.000	0.626	0.066	0.998	0.717	0.048	0.278	0.999
knn	0.699	0.045	0.571	0.051	1.000	0.552	0.044	1.000	0.042*	0.562	0.041	1.000	0.575	0.038	1.000	0.935
glm	0.484	0.044	**0.581**	0.038	0.001**	**0.576**	0.041	0.001**	0.264	**0.574**	0.040	0.001**	**0.576**	0.043	0.002**	0.652
rpart	0.705	0.048	0.567	0.081	1.000	0.696	0.043	0.853	1.000	0.592	0.096	1.000	0.700	0.044	0.819	0.998

F1 0.5

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.866	0.024	0.818	0.034	0.999	0.843	0.023	0.993	0.997	0.815	0.037	0.999	0.842	0.025	0.990	1.000
xgboost	0.860	0.022	0.827	0.031	1.000	0.853	0.026	0.947	0.995	0.824	0.031	1.000	0.857	0.025	0.784	1.000
c50	0.818	0.052	0.751	0.066	0.999	0.809	0.037	0.688	0.990	0.762	0.049	0.995	0.797	0.038	0.754	0.920
knn	0.882	0.028	0.811	0.044	0.999	0.835	0.026	1.000	0.990	0.811	0.029	1.000	0.838	0.023	1.000	0.999
glm	0.806	0.026	**0.822**	0.029	0.032*	**0.841**	0.021	0.002**	0.994	**0.820**	0.028	0.032*	**0.839**	0.024	0.003**	0.993
rpart	0.812	0.062	0.748	0.038	0.997	0.804	0.062	0.779	0.993	0.744	0.030	0.999	0.806	0.059	0.528	0.998

AUC

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.881	0.019	0.835	0.028	1.000	0.845	0.021	0.999	0.958	0.833	0.030	1.000	0.841	0.023	1.000	0.862
xgboost	0.872	0.021	0.840	0.026	1.000	0.858	0.024	0.984	0.990	0.839	0.023	1.000	0.862	0.024	0.981	0.999
c50	0.862	0.029	0.808	0.041	1.000	0.858	0.021	0.539	0.999	0.814	0.036	0.999	0.851	0.023	0.722	0.986
knn	0.898	0.022	0.830	0.036	1.000	0.837	0.025	1.000	0.884	0.827	0.025	1.000	0.846	0.020	1.000	0.998
glm	0.796	0.034	**0.839**	0.025	0.002**	**0.848**	0.019	0.001**	0.988	**0.836**	0.025	0.003**	**0.847**	0.021	0.002**	0.976
rpart	0.858	0.038	0.789	0.035	1.000	0.853	0.037	0.853	1.000	0.793	0.034	1.000	0.854	0.035	0.583	0.998

TSS

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.762	0.039	0.670	0.056	1.000	0.689	0.042	0.999	0.958	0.666	0.061	1.000	0.682	0.046	1.000	0.862
xgboost	0.743	0.042	0.680	0.051	1.000	0.715	0.048	0.984	0.990	0.678	0.046	1.000	0.724	0.047	0.981	0.999
c50	0.723	0.057	0.616	0.083	1.000	0.716	0.043	0.539	0.999	0.629	0.072	0.999	0.702	0.046	0.722	0.986
knn	0.797	0.044	0.659	0.072	1.000	0.675	0.050	1.000	0.884	0.655	0.051	1.000	0.692	0.039	1.000	0.998
glm	0.591	0.067	**0.678**	0.050	0.002**	**0.696**	0.038	0.001**	0.988	**0.671**	0.050	0.003**	**0.694**	0.041	0.002**	0.976
rpart	0.716	0.075	0.579	0.071	1.000	0.705	0.074	0.853	1.000	0.585	0.068	1.000	0.708	0.070	0.583	1.000

Turtles

Sensitivity

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.736	0.123	0.742	0.140	0.276	0.742	0.131	0.453	0.500	0.769	0.136	0.077	0.758	0.125	0.180	0.547
xgboost	0.769	0.136	0.720	0.150	0.800	0.735	0.114	0.780	0.698	0.802	0.106	0.117	0.802	0.112	0.223	0.500
c50	0.608	0.263	0.539	0.435	0.730	0.691	0.176	0.109	0.797	0.713	0.274	0.142	**0.741**	0.142	0.032*	0.639
knn	0.687	0.153	0.576	0.125	0.970	0.687	0.129	0.571	0.916	0.742	0.073	0.102	0.748	0.104	0.262	0.633
glm	0.896	0.080	0.687	0.136	0.997	0.721	0.143	0.997	0.763	0.710	0.114	0.998	0.737	0.115	0.996	0.724
rpart	0.616	0.151	0.675	0.173	0.238	0.719	0.185	0.086	0.858	0.703	0.135	0.091	**0.758**	0.171	0.029*	0.955

Specificity

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.723	0.142	0.644	0.146	0.914	0.761	0.107	0.135	0.989	0.724	0.183	0.453	0.707	0.156	0.583	0.178
xgboost	0.708	0.182	0.698	0.139	0.600	0.742	0.082	0.363	0.886	0.742	0.161	0.337	0.716	0.104	0.541	0.383
c50	0.661	0.382	0.663	0.272	0.571	0.577	0.250	0.797	0.264	0.600	0.244	0.730	0.602	0.182	0.762	0.524
knn	0.698	0.179	**0.795**	0.083	0.046*	0.733	0.117	0.311	0.060	0.642	0.143	0.781	0.661	0.092	0.696	0.617
glm	0.180	0.105	**0.606**	0.145	0.003**	**0.670**	0.156	0.003**	0.854	**0.634**	0.107	0.003**	**0.687**	0.123	0.003**	0.930
rpart	0.768	0.158	0.611	0.183	0.971	0.673	0.181	0.940	0.898	0.592	0.202	0.988	0.673	0.193	0.947	0.939

Precision

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.816	0.080	0.778	0.050	0.857	0.834	0.074	0.118	0.985	0.830	0.076	0.312	0.814	0.074	0.385	0.278
xgboost	0.818	0.088	0.800	0.061	0.652	0.824	0.036	0.594	0.882	0.845	0.077	0.216	0.824	0.052	0.577	0.342
c50	0.806	0.123	0.701	0.101	0.922	0.759	0.140	0.938	0.852	0.755	0.082	0.819	0.765	0.104	0.882	0.752
knn	0.799	0.080	0.821	0.063	0.207	0.805	0.089	0.423	0.143	0.777	0.065	0.797	0.784	0.042	0.652	0.615
glm	0.640	0.031	**0.742**	0.071	0.002**	**0.786**	0.081	0.001**	0.839	**0.762**	0.046	0.001**	**0.796**	0.066	0.001**	0.935
rpart	0.825	0.091	0.743	0.104	0.958	0.791	0.119	0.884	0.970	0.746	0.090	0.991	0.793	0.126	0.797	0.967

F1 0.5

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.747	0.108	0.744	0.113	0.318	0.756	0.115	0.318	0.779	0.775	0.110	0.080	0.765	0.100	0.246	0.423
xgboost	0.773	0.117	0.729	0.125	0.839	0.749	0.094	0.797	0.784	0.806	0.083	0.053	0.803	0.091	0.278	0.461
c50	0.614	0.203	0.746	0.252	0.234	0.690	0.147	0.097	0.289	0.780	0.086	0.082	**0.739**	0.121	0.019*	0.500
knn	0.699	0.127	0.610	0.112	0.968	0.706	0.119	0.577	0.839	0.746	0.060	0.078	0.752	0.084	0.246	0.615
glm	0.828	0.057	0.694	0.114	0.999	0.729	0.122	0.990	0.722	0.716	0.093	1.000	0.745	0.100	0.981	0.812
rpart	0.642	0.131	0.682	0.147	0.246	0.725	0.163	0.097	0.839	0.706	0.114	0.118	**0.761**	0.154	0.024*	0.978

AUC

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.729	0.082	0.693	0.055	0.905	0.751	0.086	0.142	0.971	0.746	0.079	0.270	0.733	0.072	0.461	0.361
xgboost	0.738	0.094	0.709	0.071	0.821	0.739	0.043	0.500	0.947	0.772	0.071	0.161	0.759	0.058	0.500	0.278
c50	0.635	0.081	0.601	0.116	0.920	0.634	0.090	0.581	0.784	0.657	0.096	0.423	**0.671**	0.085	0.033*	0.778
knn	0.692	0.064	0.686	0.059	0.500	0.710	0.106	0.342	0.577	0.692	0.068	0.547	0.704	0.052	0.305	0.601
glm	0.538	0.052	**0.646**	0.079	0.002**	**0.696**	0.088	0.001**	0.839	**0.672**	0.046	0.001**	**0.712**	0.083	0.001**	0.923
rpart	0.692	0.069	0.643	0.109	0.903	0.696	0.118	0.216	0.981	0.648	0.099	0.978	0.715	0.143	0.246	0.984

TSS

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.458	0.164	0.386	0.111	0.914	0.502	0.171	0.143	0.971	0.493	0.157	0.278	0.465	0.144	0.461	0.348
xgboost	0.476	0.188	0.417	0.143	0.821	0.478	0.086	0.500	0.935	0.544	0.141	0.161	0.518	0.116	0.500	0.278
c50	0.270	0.161	0.202	0.233	0.920	0.267	0.180	0.581	0.784	0.313	0.192	0.423	**0.343**	0.171	0.033*	0.784
knn	0.385	0.128	0.372	0.117	0.500	0.421	0.212	0.342	0.577	0.384	0.136	0.547	0.408	0.103	0.305	0.581
glm	0.076	0.103	**0.293**	0.159	0.002**	**0.392**	0.177	0.001**	0.839	**0.344**	0.093	0.001**	**0.424**	0.166	0.001**	0.920
rpart	0.384	0.139	0.286	0.218	0.903	0.393	0.235	0.216	0.982	0.296	0.199	0.978	0.431	0.286	0.246	0.984

All species

Sensitivity

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.913	0.014	0.896	0.019	0.986	**0.943**	0.017	0.003**	0.998	0.886	0.018	0.997	**0.945**	0.015	0.003**	0.998
xgboost	0.917	0.015	0.895	0.024	0.986	**0.949**	0.012	0.003**	1.000	0.864	0.017	1.000	**0.943**	0.013	0.005**	1.000
c50	0.885	0.020	0.769	0.049	1.000	**0.908**	0.034	0.038*	0.998	0.777	0.038	1.000	**0.925**	0.021	0.004**	0.998
knn	0.916	0.014	0.854	0.023	1.000	0.795	0.024	0.998	0.003**	0.844	0.023	1.000	0.824	0.019	1.000	0.005**
glm	0.783	0.258	0.748	0.034	0.615	0.851	0.025	0.539	0.998	0.741	0.030	0.652	0.861	0.022	0.539	0.998
rpart	0.879	0.023	0.717	0.085	1.000	0.879	0.027	0.500	1.000	0.692	0.095	1.000	0.879	0.027	0.500	0.998

Specificity

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.717	0.032	0.724	0.018	0.500	**0.767**	0.022	0.001**	0.998	0.709	0.020	0.884	**0.778**	0.022	0.001**	0.998
xgboost	0.760	0.030	0.710	0.023	0.997	0.749	0.025	0.986	1.000	0.726	0.021	0.997	0.760	0.025	0.561	1.000
c50	0.752	0.022	0.715	0.029	0.997	0.768	0.037	0.323	0.999	0.717	0.052	0.981	0.751	0.025	0.620	0.938
knn	0.766	0.029	0.706	0.022	1.000	0.760	0.025	0.682	0.998	0.709	0.024	1.000	0.749	0.022	0.981	1.000
glm	0.695	0.171	0.748	0.019	0.161	0.736	0.020	0.188	0.007**	0.726	0.020	0.312	0.727	0.023	0.278	0.594
rpart	0.783	0.025	0.699	0.095	0.990	0.780	0.029	0.814	0.989	0.704	0.071	0.990	0.780	0.029	0.814	0.990

Precision

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.620	0.028	0.620	0.018	0.688	**0.671**	0.023	0.001**	1.000	0.606	0.019	0.935	**0.683**	0.023	0.001**	1.000
xgboost	0.658	0.029	0.608	0.024	1.000	0.656	0.023	0.862	1.000	0.614	0.021	1.000	0.665	0.023	0.278	1.000
c50	0.643	0.023	0.576	0.021	1.000	0.665	0.037	0.065	1.000	0.582	0.039	1.000	0.652	0.024	0.161	1.000
knn	0.664	0.030	0.594	0.023	1.000	0.625	0.028	0.997	1.000	0.594	0.024	1.000	0.623	0.024	0.999	1.000
glm	0.592	0.091	0.599	0.022	0.385	0.619	0.021	0.188	1.000	0.576	0.022	0.754	0.614	0.021	0.312	1.000
rpart	0.672	0.026	0.554	0.056	1.000	0.669	0.027	0.814	1.000	0.544	0.029	1.000	0.669	0.027	0.814	1.000

F1 0.5

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.834	0.016	0.823	0.017	0.947	**0.872**	0.016	0.001**	1.000	0.811	0.018	0.990	**0.877**	0.016	0.001**	1.000
xgboost	0.850	0.017	0.818	0.024	0.999	**0.871**	0.012	0.001**	1.000	0.799	0.017	1.000	**0.870**	0.011	0.002**	1.000
c50	0.823	0.020	0.720	0.035	1.000	**0.845**	0.025	0.007**	1.000	0.727	0.028	1.000	**0.853**	0.015	0.003**	1.000
knn	0.851	0.017	0.785	0.022	1.000	0.754	0.023	1.000	0.001**	0.778	0.022	1.000	0.774	0.019	1.000	0.246
glm	0.711	0.188	0.712	0.029	0.577	0.791	0.022	0.461	1.000	0.700	0.026	0.577	0.796	0.019	0.423	1.000
rpart	0.828	0.018	0.672	0.050	1.000	0.827	0.019	0.814	1.000	0.653	0.065	1.000	0.827	0.019	0.814	1.000

AUC

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.815	0.019	0.810	0.015	0.862	**0.855**	0.016	0.001**	1.000	0.798	0.017	0.981	**0.862**	0.015	0.001**	0.998
xgboost	0.838	0.019	0.802	0.022	1.000	**0.849**	0.014	0.007**	1.000	0.795	0.016	1.000	**0.852**	0.012	0.005**	1.000
c50	0.819	0.018	0.742	0.020	1.000	**0.838**	0.021	0.005**	1.000	0.747	0.025	1.000	**0.838**	0.014	0.005**	1.000
knn	0.841	0.019	0.780	0.021	1.000	0.777	0.020	0.998	0.216	0.776	0.021	1.000	0.786	0.018	1.000	0.995
glm	0.739	0.073	0.748	0.020	0.348	**0.793**	0.017	0.019*	1.000	0.733	0.019	0.577	**0.794**	0.017	0.019*	1.000
rpart	0.831	0.015	0.708	0.022	1.000	0.830	0.014	0.963	1.000	0.698	0.020	1.000	0.830	0.014	0.963	1.000

TSS

	AOO & EOO only (base)		Ecological features only			All features			eco-all	Ecological features only FS			All features FS			eco-all
	mean	sd	mean	sd	p-value	mean	sd	p-value	p-value	mean	sd	p-value	mean	sd	p-value	p-value
rf	0.630	0.038	0.620	0.031	0.862	**0.710**	0.031	0.001**	1.000	0.596	0.034	0.981	**0.723**	0.030	0.001**	1.000
xgboost	0.676	0.038	0.605	0.043	1.000	**0.698**	0.027	0.007**	1.000	0.590	0.032	1.000	**0.704**	0.025	0.005**	1.000
c50	0.637	0.036	0.484	0.040	1.000	**0.676**	0.041	0.005**	1.000	0.493	0.050	1.000	**0.676**	0.028	0.005**	1.000
knn	0.682	0.037	0.560	0.041	1.000	0.554	0.041	1.000	0.216	0.553	0.041	1.000	0.572	0.036	1.000	0.997
glm	0.478	0.146	0.496	0.041	0.348	**0.587**	0.035	0.019*	1.000	0.466	0.038	0.577	**0.587**	0.033	0.019*	1.000
rpart	0.662	0.031	0.416	0.044	1.000	0.660	0.029	0.963	1.000	0.396	0.039	1.000	0.660	0.029	0.963	1.000

Appendix B

Bibliographic note

The work developed in this dissertation allowed the production and submission of two scientific articles, one for a core A international conference in data science and another in an Ecology journal. Figures B.1 and B.2 shows the proof of submission to the two.

- Soares, N., Gonçalves, J., Vasconcelos, R., Ribeiro, R. - *Combining Multiple Data Sources to Predict IUCN Conservation Status of Reptiles* submitted to 8th IEEE International Conference on Data Science and Advanced Analytics (DSAA) 2021 - Application Track

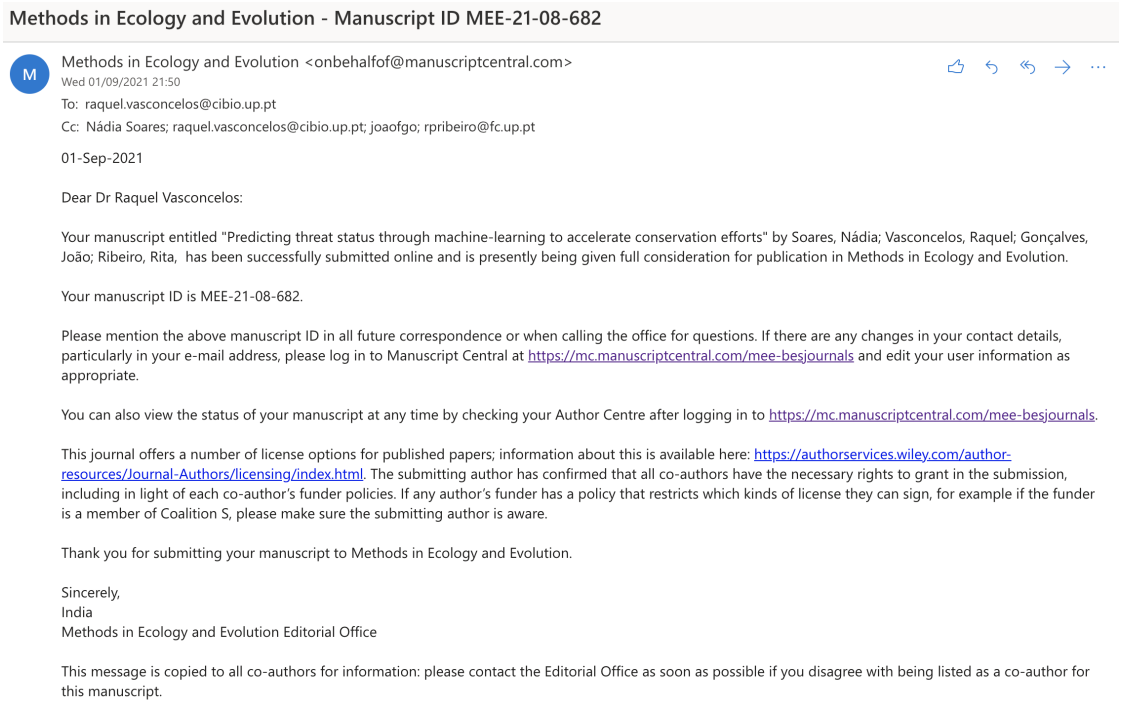
Submissions — Help Center

Submission Summary

Conference Name	The 8th IEEE International Conference on Data Science and Advanced Analytics
Track Name	Applications track
Paper ID	131
Paper Title	Combining Multiple Data Sources to Predict IUCN Conservation Status of Reptiles
Abstract	We are losing species at a high rate, even before their extinction risk being assessed. The International Union for Conservation of Nature (IUCN) Red List is the most complete assessment of all species conservation status, yet it only covers around 134,400 out of the approximate 1.2 million species identified so far. Additionally, many of the existing assessments are outdated. The assessment of the conservation status of a species is a long, mostly manual process, carefully done by experts. The conservation field would gain by having ways of automating this process, for instance, by prioritizing the species where experts and financing should focus on.In this paper, we present a pipeline used to derive a conservation dataset out of openly available data and obtain predictions, through machine learning techniques, on which species are most likely to be threatened. We applied this pipeline to the different groups within the Reptilia class, as a model of one of the most under-assessed taxonomic groups.Additionally, we compared the performance of three main datasets. The first one used only two predictors (the most relevant ones used by IUCN in their assessments) – the extent of occurrence (EOO) and the area of occupancy (AOO). The second dataset used only ecological variables (related to climate, vegetation, geomorphology), and the third used all available features – AOO plus EOO, and ecogeographical variables (EGV). Our results show that EOO and AOO are the most relevant predictors of threat status and can achieve good results when used alone in a model. Nevertheless, most groups benefit from the usage of EGV in conjugation with these two variables. Random Forest appeared to be the best method for most species groups and feature selection was shown to improve results. Results support that the developed pipeline can be used to perform a general model-assisted assessment across many species (such as reptiles), which more focused studies can then complement.
Created on	04/06/2021, 23:35:36
Last Modified	26/07/2021, 12:40:23
Authors	**Nádia Soares** (Faculdade de Ciências da Universidade do Porto) < [redacted] > João Gonçalves (Research Centre in Biodiversity and Genetic Resources) < [redacted] > Raquel Vasconcelos (Research Centre in Biodiversity and Genetic Resources) < [redacted] > Rita Ribeiro (INESC TEC) < [redacted] >
Submission Files	DSAA___Predicting_the_conservation_status_of_species.pdf (1.7 Mb, 06/06/2021, 01:30:28)
Submission Questions Response	**1. Keywords** *(comma-separated, max 125 characters)* data integration, extinction risk assessment, machine learning, multi source

Figure B.1: Proof of submission to the DSAA conference.

- Soares, N., Vasconcelos, R., Gonçalves, J., Ribeiro, R. - *Predicting threat status through machine learning to accelerate conservation efforts* submitted to Methods in Ecology and Evolution (MEE) - Special Feature in Methods in Ecology and Evolution: Realising the promise of large data and complex models



Methods in Ecology and Evolution - Manuscript ID MEE-21-08-682

M

Methods in Ecology and Evolution <onbehalfof@manuscriptcentral.com>
Wed 01/09/2021 21:50

To: raquel.vasconcelos@cibio.up.pt
Cc: Nádia Soares; raquel.vasconcelos@cibio.up.pt; joaofgo; rpribeiro@fc.up.pt

01-Sep-2021

Dear Dr Raquel Vasconcelos:

Your manuscript entitled "Predicting threat status through machine-learning to accelerate conservation efforts" by Soares, Nádia; Vasconcelos, Raquel; Gonçalves, João; Ribeiro, Rita, has been successfully submitted online and is presently being given full consideration for publication in Methods in Ecology and Evolution.

Your manuscript ID is MEE-21-08-682.

Please mention the above manuscript ID in all future correspondence or when calling the office for questions. If there are any changes in your contact details, particularly in your e-mail address, please log in to Manuscript Central at https://mc.manuscriptcentral.com/mee-besjournals and edit your user information as appropriate.

You can also view the status of your manuscript at any time by checking your Author Centre after logging in to https://mc.manuscriptcentral.com/mee-besjournals.

This journal offers a number of license options for published papers; information about this is available here: https://authorservices.wiley.com/author-resources/Journal-Authors/licensing/index.html. The submitting author has confirmed that all co-authors have the necessary rights to grant in the submission, including in light of each co-author's funder policies. If any author's funder has a policy that restricts which kinds of license they can sign, for example if the funder is a member of Coalition S, please make sure the submitting author is aware.

Thank you for submitting your manuscript to Methods in Ecology and Evolution.

Sincerely,
India
Methods in Ecology and Evolution Editorial Office

This message is copied to all co-authors for information: please contact the Editorial Office as soon as possible if you disagree with being listed as a co-author for this manuscript.

Figure B.2: Proof of submission to the MEE journal.

References

[1] Gerardo Ceballos, Paul R. Ehrlich, and Peter H. Raven. Vertebrates on the brink as indicators of biological annihilation and the sixth mass extinction. *Proceedings of the National Academy of Sciences of the United States of America*, 117(24):13596–13602, 2020. ISSN: 10916490. doi:10.1073/pnas.1922686117.

[2] Carlos R. Fonseca, Gustavo B. Paterno, Demétrio L. Guadagnin, Eduardo M. Venticinque, Gerhard E. Overbeck, Gislene Ganade, Jean P. Metzger, Johannes Kollmann, Johannes Sauer, Márcio Z. Cardoso, Priscila F.M. Lopes, Rafael S. Oliveira, Valério D. Pillar, and Wolfgang W. Weisser. Conservation biology: four decades of problem- and solution-based research. *Perspectives in Ecology and Conservation*, 19(2):121–130, 2021. ISSN: 2530-0644. doi:10.1016/j.pecon.2021.03.003.

[3] Cristina Romanelli, David Cooper, Diarmid Campbell-Lendrum, Marina Maiero, William B. Karesh, Danny Hunter, and Christopher D. Golden. Connecting Global Priorities: Biodiversity and Human Health, a State of Knowledge Review. *World Health Organization and Secretariat for the Convention on Biological Diversity*, page 360pp, 2015. doi:10.13140/RG.2.1.3679.6565.

[4] FAO Commission on Genetic Resources for Food and Agriculture. The State of the World's Biodiversity for Food and Agriculture, 2019. doi:10.4060/ca3129en.

[5] FAO Commission on Genetic Resources for Food and Agriculture. How the World's Food Security Depends on Biodiversity, 2020. https://www.fao.org/3/cb0416en/CB0416EN.pdf.

[6] Felicia Keesing, Lisa K. Belden, Peter Daszak, Andrew Dobson, Catherine D. Harvell, Robert D. Holt, Peter Hudson, Anna Jolles, Kate E. Jones, Charles E. Mitchell, Samuel S. Myers, Tiffany Bogich, and Richard S. Ostfeld. Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature*, 468(7324):647–652, 2010. ISSN: 00280836. doi:10.1038/nature09575.

[7] Hanno Seebens, Tim M. Blackburn, Ellie E. Dyer, Piero Genovesi, Philip E. Hulme, Jonathan M. Jeschke, Shyama Pagad, Petr Pyšek, Marten Winter, Margarita Arianoutsou, Sven Bacher, Bernd Blasius, Giuseppe Brundu, César Capinha, Laura Celesti-Grapow, Wayne Dawson, Stefan Dullinger, Nicol Fuentes, Heinke Jäger, John Kartesz, Marc Kenis, Holger Kreft, Ingolf Kühn, Bernd Lenzner, Andrew Liebhold, Alexander Mosena, Dietmar

Moser, Misako Nishino, David Pearman, Jan Pergl, Wolfgang Rabitsch, Julissa Rojas-Sandoval, Alain Roques, Stephanie Rorke, Silvia Rossinelli, Helen E. Roy, Riccardo Scalera, Stefan Schindler, Kateřina Štajerová, Barbara Tokarska-Guzik, Mark Van Kleunen, Kevin Walker, Patrick Weigelt, Takehiko Yamanaka, and Franz Essl. No saturation in the accumulation of alien species worldwide. *Nature Communications*, 8(February), 2017. ISSN: 20411723. doi:10.1038/ncomms14435.

[8] Néstor Pérez-Méndez, Georg K.S. Andersson, Fabrice Requier, Juliana Hipólito, Marcelo A. Aizen, Carolina L. Morales, Nancy García, Gerardo P. Gennari, and Lucas A. Garibaldi. The economic cost of losing native pollinator species for orchard production. *Journal of Applied Ecology*, 57(3):599–608, 2020. ISSN: 13652664. doi:10.1111/1365-2664.13561.

[9] Leon C. Braat, Patrick ten Brink with J. Bakkes, K. Bolt, I. Braeuer, B. ten Brink, A. Chiabai, H. Ding, H. Gerdes, M. Jeuken, M. Kettunen, U. Kirchholtes, C. Klok, A. Markandya, P. Nunes, M. van Oorschot, N. Peralta-Bezerra, M. Rayment, C. Travisi, and M. Walpole. The cost of policy inaction, 06 2011.

[10] Diego Juffe-Bignoli, Thomas M. Brooks, Stuart H. M. Butchart, Richard B. Jenkins, Kaia Boe, Michael Hoffmann, Ariadne Angulo, Steve Bachman, Monika Böhm, Neil Brummitt, Kent E. Carpenter, Pat J. Comer, Neil Cox, Annabelle Cuttelod, William R. T. Darwall, Moreno Di Marco, Lincoln D. C. Fishpool, Bárbara Goettsch, Melanie Heath, Craig Hilton-Taylor, Jon Hutton, Tim Johnson, Ackbar Joolia, David A. Keith, Penny F. Langhammer, Jennifer Luedtke, Eimear Nic Lughadha, Maiko Lutz, Ian May, Rebecca M. Miller, María A. Oliveira-Miranda, Mike Parr, Caroline M. Pollock, Gina Ralph, Jon Paul Rodríguez, Carlo Rondinini, Jane Smart, Simon Stuart, Andy Symes, Andrew W. Tordoff, Stephen Woodley, Bruce Young, and Naomi Kingston. Assessing the cost of global biodiversity and conservation knowledge. *PLOS ONE*, 11(8):1–22, 08 2016. doi:10.1371/journal.pone.0160640.

[11] Carlo Rondinini, Moreno Di Marco, Piero Visconti, Stuart Butchart, and Luigi Boitani. Update or outdate: long-term viability of the iucn red list. *Conservation Letters*, 7, 01 2013. doi:10.1111/conl.12040.

[12] David G. Chapple, Uri Roll, Monika Böhm, Rocío Aguilar, Andrew P. Amey, Chris C. Austin, Marleen Baling, Anthony J. Barley, Michael F. Bates, Aaron M. Bauer, Daniel G. Blackburn, Phil Bowles, Rafe M. Brown, S.R. Chandramouli, Laurent Chirio, Hal Cogger, Guarino R. Colli, Werner Conradie, Patrick J. Couper, Mark A. Cowan, Michael D. Craig, Indraneil Das, Aniruddha Datta-Roy, Chris R. Dickman, Ryan J. Ellis, Aaron L. Fenner, Stewart Ford, S.R. Ganesh, Michael G. Gardner, Peter Geissler, Graeme R. Gillespie, Frank Glaw, Matthew J. Greenlees, Oliver W. Griffith, L. Lee Grismer, Margaret L. Haines, D. James Harris, S. Blair Hedges, Rod A. Hitchmough, Conrad J. Hoskin, Mark N. Hutchinson, Ivan Ineich, Jordi Janssen, Gregory R. Johnston, Benjamin R. Karin, J. Scott Keogh, Fred Kraus, Matthew LeBreton, Petros Lymberakis, Rafaqat Masroor, Peter J. McDonald, Sven Mecke, Jane Melville, Sabine Melzer, Damian R. Michael, Aurélien

Miralles, Nicola J. Mitchell, Nicola J. Nelson, Truong Q. Nguyen, Cristiano de Campos Nogueira, Hidetoshi Ota, Panayiotis Pafilis, Olivier S.G. Pauwels, Ana Perera, Daniel Pincheira-Donoso, Robert N. Reed, Marco A. Ribeiro-Júnior, Julia L. Riley, Sara Rocha, Pamela L. Rutherford, Ross A. Sadlier, Boaz Shacham, Glenn M. Shea, Richard Shine, Alex Slavenko, Adam Stow, Joanna Sumner, Oliver J.S. Tallowin, Roy Teale, Omar Torres-Carvajal, Jean-Francois Trape, Peter Uetz, Kanishka D.B. Ukuwela, Leonie Valentine, James U. Van Dyke, Dylan van Winkel, Raquel Vasconcelos, Miguel Vences, Philipp Wagner, Erik Wapstra, Geoffrey M. While, Martin J. Whiting, Camilla M. Whittington, Steve Wilson, Thomas Ziegler, Reid Tingley, and Shai Meiri. Conservation status of the world's skinks (scincidae): Taxonomic and geographic patterns in extinction risk. *Biological Conservation*, 257:109101, 2021. ISSN: 0006-3207. doi:10.1016/j.biocon.2021.109101.

[13] E. C. M. Parsons. Why iucn should replace "data deficient" conservation status with a precautionary "assume threatened" status—a cetacean case study. *Frontiers in Marine Science*, 3:193, 2016. ISSN: 2296-7745. doi:10.3389/fmars.2016.00193.

[14] IUCN. The IUCN Red List of Threatened Species Assessment Process, 2021. Retrieved 21 May 2021, from https://www.iucnredlist.org/assessment/process.

[15] Convention on Biological Diversity. Aichi Biodiversity Targets, 2020. Retrieved 11 May 2021, from https://www.cbd.int/sp/targets/.

[16] United Nations Department of Economic and Social Affairs. THE 17 GOALS. Retrieved 11 May 2021, from https://sdgs.un.org/goals.

[17] Convention on Biological Diversity. First draft of the post-2020 global biodiversity framework, 2021. Retrieved 14 November 2021, from https://www.cbd.int/doc/c/abb5/591f/2e46096d3f0330b08ce87a45/wg2020-03-03-en.pdf.

[18] Hans-Otto Pörtner, Robert J. Scholes, John Agard, Emma Archer, Xuemei Bai, David Barnes, Michael Burrows, Lena Chan, Wai Lung (William) Cheung, Sarah Diamond, Camila Donatti, Carlos Duarte, Nico Eisenhauer, Wendy Foden, Maria A. Gasalla, Collins Handa, Thomas Hickler, Ove Hoegh-Guldberg, Kazuhito Ichii, Ute Jacob, Gregory Insarov, Wolfgang Kiessling, Paul Leadley, Rik Leemans, Lisa Levin, Michelle Lim, Shobha Maharaj, Shunsuke Managi, Pablo A. Marquet, Pamela McElwee, Guy Midgley, Thierry Oberdorff, David Obura, Balgis Osman Elasha, Ram Pandit, Unai Pascual, Aliny P F Pires, Alexander Popp, Victoria Reyes-García, Mahesh Sankaran, Josef Settele, Yunne-Jai Shin, Dejene W. Sintayehu, Peter Smith, Nadja Steiner, Bernardo Strassburg, Raman Sukumar, Christopher Trisos, Adalberto Luis Val, Jianguo Wu, Edvin Aldrian, Camille Parmesan, Ramon Pichs-Madruga, Debra C. Roberts, Alex D. Rogers, Sandra Díaz, Markus Fischer, Shizuka Hashimoto, Sandra Lavorel, Ning Wu, and Hien Ngo. IPBES-IPCC co-sponsored workshop report on biodiversity and climate change, June 2021. doi:10.5281/zenodo.5101133.

[19] IUCN. The IUCN Red List of Threatened Species, 2021. Version 2021-2. Retrieved 27 June 2021, from https://www.iucnredlist.org/.

[20] IUCN. Barometer of Life, 2020. Retrieved 20 May 2021, https://www.iucnredlist.org/about/barometer-of-life.

[21] Uri Roll, Anat Feldman, Maria Novosolov, Allen Allison, Aaron M. Bauer, Rodolphe Bernard, Monika Böhm, Fernando Castro-Herrera, Laurent Chirio, Ben Collen, Guarino R. Colli, Lital Dabool, Indraneil Das, Tiffany M. Doan, Lee L. Grismer, Marinus Hoogmoed, Yuval Itescu, Fred Kraus, Matthew Lebreton, Amir Lewin, Marcio Martins, Erez Maza, Danny Meirte, Zoltán T. Nagy, Cristiano De C. Nogueira, Olivier S.G. Pauwels, Daniel Pincheira-Donoso, Gary D. Powney, Roberto Sindaco, Oliver J.S. Tallowin, Omar Torres-Carvajal, Jean François Trape, Enav Vidan, Peter Uetz, Philipp Wagner, Yuezhao Wang, C. David L. Orme, Richard Grenyer, and Shai Meiri. The global distribution of tetrapods reveals a need for targeted reptile conservation. *Nature Ecology and Evolution*, 1(11): 1677–1682, 2017. ISSN: 2397334X. doi:10.1038/s41559-017-0332-2.

[22] Christine Howard, Curtis Flather, and Philip Stephens. A global assessment of the drivers of threatened terrestrial species richness. *Nature Communications*, 11, 02 2020. doi:10.1038/s41467-020-14771-6.

[23] J. Whitfield Gibbons, David E. Scott, Travis J. Ryan, Kurt A. Buhlmann, Tracey D. Tuberville, Brian S. Metts, Greenem Judith L., Tony Mills, Leiden Yale, Sean Poppy, and Christopher T. Winne. The Global Decline of Reptiles, Déja Vu Amphibians: Reptile species are declining on a global scale. Six significant threats to reptile populations are habitat loss and degradation, introduced invasive species, environmental pollution, disease, unsustainable use, and global climate change. *BioScience*, 50(0):8, 2000. doi:10.1641/0006-3568(2000)050[0653:TGDORD]2.0.CO;2.

[24] Nicole Valenzuela, Robert Literman, Jennifer Neuwald, Beatriz Akemi Mizoguchi, John Iverson, Julia Riley, and Jacqueline Litzgus. Extreme thermal fluctuations from climate change unexpectedly accelerate demographic collapse of vertebrates with temperature-dependent sex determination. *Scientific Reports*, 9, 03 2019. doi:10.1038/s41598-019-40597-4.

[25] Everton B. P. de Miranda. The plight of reptiles as ecological actors in the tropics. *Frontiers in Ecology and Evolution*, 5:159, 2017. ISSN: 2296-701X. doi:10.3389/fevo.2017.00159.

[26] Angela M. Cortés-Gomez, Cesar A. Ruiz, Anyelet Valencia-Aguilar, and Richard J. Ladle. Ecological functions of neotropical amphibians and reptiles: A review. *Universitas Scientiarum*, 20(2):229–245, 2015. ISSN: 20271352. doi:10.11144/Javeriana.SC20-2.efna.

[27] Yi Yang, Yingying Lin, and Lei Shi. The effect of lizards on the dispersal and germination of capparis spinosa (capparaceae). *PLOS ONE*, 16(2):1–16, 02 2021. doi:10.1371/journal.pone.0247585.

[28] Alfredo Valido and Jens Olesen. Frugivory and seed dispersal by lizards: A global review. *Frontiers in Ecology and Evolution*, 7:49, 03 2019. doi:10.3389/fevo.2019.00049.

[29] J. Uetz, P. & Hošek. The reptile database, 2020. Retrieved 02 January 2021, from http://www.reptile-database.org.

[30] Tanja K. Petersen, James D. M. Speed, Vidar Grøtan, and Gunnar Austrheim. Species data for understanding biodiversity dynamics: The what, where and when of species occurrence data collection. *Ecological Solutions and Evidence*, 2(1):e12048, 2021. doi:10.1002/2688-8319.12048.

[31] Gordon S. Blair, Peter Henrys, Amber Leeson, John Watkins, Emma Eastoe, Susan Jarvis, and Paul J. Young. Data Science of the Natural Environment: A Research Roadmap. *Frontiers in Environmental Science*, 7(August):1–14, 2019. ISSN: 2296665X. doi:10.3389/fenvs.2019.00121.

[32] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16: 321–357, Jun 2002. ISSN: 1076-9757. doi:10.1613/jair.953.

[33] IUCN. Guidelines for Using the IUCN Red List Categories and Criteria, 2019. https://nc.iucnredlist.org/redlist/content/attachment_files/RedListGuidelines.pdf.

[34] Britannica. A classification of living organisms, 2021. Retrieved 4 September 2021, from https://www.britannica.com/science/taxonomy/A-classification-of-living-organisms.

[35] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2 edition, 2018. ISBN: 978-0-262-03940-6.

[36] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001. ISBN: 978-0-387-84858-7.

[37] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[38] Paula Branco, Luís Torgo, and Rita P. Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.*, 49(2), August 2016. ISSN: 0360-0300. doi:10.1145/2907070.

[39] Roweida Mohammed, Jumanah Rawashdeh, and Malak Abdullah. Machine learning with oversampling and undersampling techniques: Overview study and experimental results. pages 243–248, 04 2020. doi:10.1109/ICICS49469.2020.239556.

[40] Chris Drummond and Robert Holte. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats oversampling. *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Datasets*, 01 2003.

[41] Marcus Maloof. Learning when data sets are imbalanced and when costs are unequal and unknown. *Analysis*, 21, 07 2003.

[42] Rukshan Batuwita and Vasile Palade. Fsvm-cil: Fuzzy support vector machines for class imbalance learning. *Fuzzy Systems, IEEE Transactions on*, 18:558 – 571, 07 2010. doi:10.1109/TFUZZ.2010.2042721.

[43] Chao Chen and Leo Breiman. Using random forest to learn imbalanced data. *University of California, Berkeley*, 01 2004.

[44] Ricardo Barandela, Josep Sánchez, Vicente García, and E. Rangel. Strategies for learning in class imbalance problems. *Pattern Recognition*, 36:849–851, 03 2003. doi:10.1016/S0031-3203(02)00257-1.

[45] Jose Hernandez-Orallo. Soft (gaussian cde) regression models and loss functions, 2012. Retrieved from https://arxiv.org/abs/1211.1043.

[46] Daniel Berrar. *Cross-Validation*. 01 2018. ISBN: 9780128096338. doi:10.1016/B978-0-12-809633-8.20349-X.

[47] Rafael O. Wüest, Niklaus E. Zimmermann, Damaris Zurell, Jake M. Alexander, Susanne A. Fritz, Christian Hof, Holger Kreft, Signe Normand, Juliano S. Cabral, Eniko Szekely, Wilfried Thuiller, Martin Wikelski, and Dirk N. Karger. Macroecology in the age of big data – where to go from here? *Journal of Biogeography*, 47(1):1–12, 2020. doi:10.1111/jbi.13633.

[48] Scott S. Farley, Andria Dawson, Simon J. Goring, and John W Williams. Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions. *BioScience*, 68(8): 563–576, 07 2018. ISSN: 0006-3568. doi:10.1093/biosci/biy068.

[49] Ryan B. Ghannam and Stephen M. Techtmann. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. *Computational and Structural Biotechnology Journal*, 19:1092–1107, 2021. ISSN: 2001-0370. doi:10.1016/j.csbj.2021.01.028.

[50] Robin C. Whytock, Jędrzej Świeżewski, Joeri A. Zwerts, Tadeusz Bara-Słupski, Aurélie Flore Koumba Pambo, Marek Rogala, Laila Bahaa-el din, Kelly Boekee, Stephanie Brittain, Anabelle W. Cardoso, Philipp Henschel, David Lehmann, Brice Momboua, Cisquet Kiebou Opepa, Christopher Orbell, Ross T. Pitman, Hugh S. Robinson, and Katharine A. Abernethy. Robust ecological analysis of camera trap data labelled by a machine learning model. *Methods in Ecology and Evolution*, 12(6):1080–1092, 2021. doi:10.1111/2041-210X.13576.

[51] Debra P. C. Peters, Kris M. Havstad, Judy Cushing, Craig Tweedie, Olac Fuentes, and Natalia Villanueva-Rosales. Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. *Ecosphere*, 5(6):art67, 2014. doi:10.1890/ES13-00359.1.

[52] T. Stévart, G. Dauby, P. Lowry, A. Blach-Overgaard, V. Droissart, D. J. Harris, A. B. Mackinder, G. E. Schatz, B. Sonké, M. S.M. Sosef, J. C. Svenning, J. Wieringa, and T. L.P.

Couvreur. A third of the tropical African flora is potentially threatened with extinction. *Science Advances*, 5(11), 2019. ISSN: 23752548. doi:10.1126/sciadv.aax9444.

[53] Tara A. Pelletier, Bryan C. Carstens, David C. Tank, Jack Sullivan, and Anahí Espíndola. Predicting plant conservation priorities on a global scale. *Proceedings of the National Academy of Sciences of the United States of America*, 115(51):13027–13032, 2018. ISSN: 10916490. doi:10.1073/pnas.1804098115.

[54] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. ISSN: 1573-0565. doi:10.1023/A:1010933404324.

[55] D. Richard Cutler, Thomas C. Edwards, Karen H. Beard, Adele Cutler, Kyle T. Hess, Jacob Gibson, and Joshua J. Lawler. Random forests for classification in ecology. *Ecology*, 88(11):2783–2792, 2007. ISSN: 00129658. doi:10.1890/07-0539.1.

[56] Lucie M. Bland, Ben Collen, David Orme, and Jon Bielby. Predicting the conservation status of data-deficient species. *Conservation biology : the journal of the Society for Conservation Biology*, 29, 08 2014. doi:10.1111/cobi.12372.

[57] Sarah E. Darrah, Lucie M. Bland, Steven P. Bachman, Colin P. Clubbe, and Anna Trias-Blasi. Using coarse-scale species distribution data to predict extinction risk in plants. *Diversity and Distributions*, 23(4):435–447, 2017. doi:10.1111/ddi.12532.

[58] Lucie M. Bland and Monika Böhm. Overcoming data deficiency in reptiles. *Biological Conservation*, 204:16–22, 2016. ISSN: 0006-3207. Advancing reptile conservation: Addressing knowledge gaps and mitigating key drivers of extinction risk. doi:https://doi.org/10.1016/j.biocon.2016.05.018.

[59] Eimear Nic Lughadha, Barnaby E. Walker, Cátia Canteiro, Helen Chadburn, Aaron P. Davis, Serene Hargreaves, Eve J. Lucas, André Schuiteman, Emma Williams, Steven P. Bachman, David Baines, Amy Barker, Andrew P. Budden, Julia Carretero, James J. Clarkson, Alexandra Roberts, and Malin C. Rivers. The use and misuse of herbarium specimens in evaluating plant extinction risks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1763):20170402, 2019. doi:10.1098/rstb.2017.0402.

[60] Tarciso C. C. Leão, Carlos R. Fonseca, Carlos A. Peres, and Marcelo Tabarelli. Predicting extinction risk of brazilian atlantic forest angiosperms. *Conservation Biology*, 28(5): 1349–1359, 2014. doi:10.1111/cobi.12286.

[61] Alexander Zizka, Daniele Silvestro, Pati Vitt, and Tiffany M. Knight. Automated conservation assessment of the orchid family with deep learning. *Conservation Biology*, 35 (3):897–908, 2021. doi:10.1111/cobi.13616.

[62] Osmar Luiz, Rachael Woods, Elizabeth Madin, and Joshua Madin. Predicting iucn extinction risk categories for the world's data deficient groupers (teleostei: Epinephelidae). *Conservation Letters*, 9:n/a–n/a, 09 2016. doi:10.1111/conl.12230.

[63] GBIF. GBIF presence data download, 2020. Retrieved 23 October 2020, from https://doi.org/10.15468/dl.kkfc5m.

[64] Lawrence N. Hudson, Tim Newbold, Sara Contu, and Samantha L. L. Hill et al. Dataset: The 2016 release of the PREDICTS database. Natural History Museum Data Portal (data.nhm.ac.uk), 2016. Retrieved 23 October 2020, from https://doi.org/10.5519/0066354.

[65] BioTIME. BioTIME presence data download, 2020. Retrieved 23 October 2020, from https://biotime.st-andrews.ac.uk/downloadFull.php.

[66] Living Planet Index. Living Planet Index presence data download, 2020. Retrieved 23 October 2020, from https://www.livingplanetindex.org/.

[67] Amal Y. Aldhebiani. Species concept and speciation. *Saudi Journal of Biological Sciences*, 25(3):437–440, 2018. ISSN: 1319-562X. doi:10.1016/j.sjbs.2017.04.013.

[68] Jen-Pan Huang and L. Lacey Knowles. The species versus subspecies conundrum: Quantitative delimitation from integrating multiple data types within a single bayesian approach in hercules beetles. *Systematic Biology*, 65(4):685–699, 12 2015. ISSN: 1063-5157. doi:10.1093/sysbio/syv119.

[69] Darryl MacKenzie. What are the issues with presence-absence data for wildlife managers? *Journal of Wildlife Management*, 69:849–860, 09 2009. doi:10.2193/0022-541X(2005)069[0849:WATIWP]2.0.CO;2.

[70] Gabriel O. Jerez and Daniele B. M. Alves. Generation and performance analysis of gps and glonass virtual data for positioning under different ionospheric conditions. *Boletim de Ciencias Geodesicas*, 25(2), 2019. ISSN: 19822170. doi:10.1590/s1982-21702019000200007.

[71] A.s.J. Proosdij, Marc Sosef, Jan Wieringa, and Niels Raes. Minimum required number of specimen records to develop accurate species distribution models. *Ecography*, 39:542–552, 05 2016. doi:10.1111/ecog.01509.

[72] M.s Wisz, Robert Hijmans, Jin Li, Andrew Peterson, Catherine Graham, Antoine Guisan, Jane Elith, Miroslav Dudík, Simon Ferrier, Falk Huettmann, John Leathwick, Anthony Lehmann, Lucia Lohmann, Bette Loiselle, Glenn Manion, Craig Moritz, Miguel Nakamura, Yoshinori Nakazawa, Jake Overton, and Niklaus Zimmermann. Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, 14:763–773, 01 2008. doi:10.1111/j.1472-4642.2008.00482.x.

[73] Dirk N. Karger, Olaf Conrad, Jürgen Böhner, Tobias Kawohl, Holger Kreft, Rodrigo W. Soria-Auza, Niklaus Zimmermann, Peter Linder, and Michael Kessler. Climatologies at high resolution for the earth's land surface areas. *Scientific Data*, 4, 09 2017. doi:10.1038/sdata.2017.122.

[74] Sara Varela, Robert P. Anderson, Raúl García-Valdés, and Federico Fernández-González. Environmental filters reduce the effects of sampling bias and improve predictions of ecological

niche models. *Ecography*, 37(11):1084–1091, 2014. ISSN: 16000587. doi:10.1111/j.1600-0587.2013.00441.x.

[75] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi:10.1080/14786440109462720.

[76] Chelsa climate. Climatologies at high resolution for the earth's land surface areas, 2020. Retrieved 20 November 2020, from http://chelsa-climate.org/bioclim/.

[77] EarthEnv. EarthEnv Topography raster layers, 2020. Retrieved 20 November 2020, from http://www.earthenv.org/topography.

[78] EarthEnv. EarthEnv Texture raster layers, 2020. Retrieved 20 November 2020, from http://www.earthenv.org/texture.

[79] Copernicus Land Monitoring Service. Copernicus fcover raster layer, 2020. Retrieved 20 November 2020, from https://land.copernicus.eu/global/products/fcover.

[80] Copernicus Land Monitoring Service. Copernicus NDVI raster layer, 2020. Retrieved 20 November 2020, from https://land.copernicus.eu/global/products/ndvi.

[81] Benjamin Blonder, Christine Lamanna, Cyrille Violle, and Brian J. Enquist. The n-dimensional hypervolume. *Global Ecology and Biogeography*, 23(5):595–609, 2014. ISSN: 14668238. doi:10.1111/geb.12146.

[82] Max Kuhn. *caret: Classification and Regression Training*, 2020. R package version 6.0-86.

[83] IUCN. The IUCN Red List of Threatened Species, 2020. Version 2020-3. Retrieved 13 December 2020, from https://www.iucnredlist.org/.

[84] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. R version 4.0.3.

[85] Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. *R News*, 2(3):18–22, 2002.

[86] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng, and Yutian Li. *xgboost: Extreme Gradient Boosting*, 2021. R package version 1.3.2.1.

[87] Max Kuhn and Ross Quinlan. *C50: C5.0 Decision Trees and Rule-Based Models*, 2020. R package version 0.1.3.1.

[88] Terry Therneau and Beth Atkinson. *rpart: Recursive Partitioning and Regression Trees*, 2019. R package version 4.1-15.

[89] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. ISSN: 00359238. doi:10.2307/2344614.

[90] Evelyn Fix and J. L. Hodges Jr. Discriminatory analysis - nonparametric discrimination: Consistency properties. *International Statistical Review*, 57:238, 1989. doi:10.2307/1403797.

[91] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. pages 785–794, 08 2016. doi:10.1145/2939672.2939785.

[92] Peter Bühlmann. Bagging, boosting and ensemble methods. *Handbook of Computational Statistics*, 01 2012. doi:10.1007/978-3-642-21551-3_33.

[93] Sébastien Bonthoux, Andrés Baselga, and Gérard Balent. Assessing community-level and single-species models predictions of species distributions and assemblage composition after 25 years of land cover change. *PLOS ONE*, 8(1):1–8, 01 2013. doi:10.1371/journal.pone.0054179.

[94] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning, 2020. Retrieved from https://arxiv.org/abs/1811.12808.

[95] Barnaby E. Walker, Tarciso C. C. Leão, Steven P. Bachman, Friederike C. Bolam, and Eimear Nic Lughadha. Caution needed when predicting species threat status for conservation prioritization on a global scale. *Frontiers in Plant Science*, 11:520, 2020. ISSN: 1664-462X. doi:10.3389/fpls.2020.00520.

[96] B. H Shekar and Guesh Dagnew. Grid search-based hyperparameter tuning and classification of microarray cancer data. In *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, pages 1–8, 2019. doi:10.1109/ICACCP.2019.8882943.

[97] Rabi Behera and Kajaree Das. A survey on machine learning: Concept, algorithms and applications. *International Journal of Innovative Research in Computer and Communication Engineering*, 2, 02 2017. doi:10.15680/IJIRCCE.2017.0502001.

[98] Quan Zou, Sifa Xie, Ziyu Lin, Meihong Wu, and Ying Ju. Finding the best classification threshold in imbalanced classification. *Big Data Research*, 5, 01 2016. doi:10.1016/j.bdr.2015.12.001.

[99] Mohamed Bekkar, Hassiba Djema, and Taklit A. Alitouche. Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications*, 3:27–38, 01 2013. ISSN: 2225-0506.

[100] 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. ISSN: 09067590. doi:10.1111/j.1600-0587.2008.05742.x.

[101] Nicolas Loiseau, Nicolas Mouquet, Nicolas Casajus, Matthias Grenié, Maya Guéguen, Brian Maitner, David Mouillot, Annette Ostling, Julien Renaud, Caroline Tucker, Laure Velez, Wilfried Thuiller, and Cyrille Violle. Global distribution and conservation status of ecologically rare mammal and bird species. *Nature Communications*, 11(1):5071, 2020. ISSN: 2041-1723. doi:10.1038/s41467-020-18779-w.

[102] L. Torgo. *Data Mining with R, learning with case studies*. Chapman and Hall/CRC, 2010.

[103] Maria Romeiras, Sílvia Catarino, Ana Filipe, Maria Magalhães, Maria Duarte, and Pedro Beja. Species conservation assessments in oceanic islands: The consequences of precautionary versus evidentiary attitudes. *Conservation Letters*, 11 2015. doi:10.1111/conl.12212.