Michel Jan Marinus Bieleveld

Improving species distribution model quality with a parallel linear genetic programming-fuzzy algorithm

São Paulo
2016

Michel Jan Marinus Bieleveld

Improving species distribution model quality with a parallel linear genetic programming-fuzzy algorithm

Tese apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Doutor em Ciências

Área de Concentração:
Computer Engineering

Supervisor: Antonio Mauro Saraiva

São Paulo
2016

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, _____ de _____ de _____

Assinatura do autor: _____

Assinatura do orientador: _____

Catalogação-na-publicação

Bieleveld, Michel Jan Marinus

Improving species distribution model quality with a parallel linear genetic programming-fuzzy algorithm / M. J. M. Bieleveld -- versão corr. -- São Paulo, 2016.

140 p.

Tese (Doutorado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Algortimos genéticos 2.Algoritmos úteis e específicos 3.FUZZY (Inteligência artificial) 4.Ecologia de populações 5.Bioclimatologia I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

This dissertation is dedicated to my lovely wife and best friend,
Lygia
You light up my world and always know how to bring a smile on my face

Acknowledgements

First and foremost, I express my deepest gratitude to my advisor Antonio Mauro Saraiva. I have been very fortunate to be given the opportunity and freedom to pursue my own research, while his guidance and patience helped me to succeed and complete this work.

I thank the laboratório de automação agrícola at the University of São Paulo, and all of the faculty, staff and fellow students at the department of computation and digital systems. They provided a safe, fun and dynamic place to study, argue and research. I thank Allan Koch Veiga and Wilian França Costa, my close fellow doctoral students, for their friendship, continuous support and positive criticism to help me stay focused on my work.

I acknowledge the valuable input from André Luis Acosta and Cris Giannini, who joined in stimulating discussions and helped me shape the biological aspect of this work.

I am thankful to the staff, in particular Edson de Souza and Suzano Luiz Bitencourt da Rosa, for their technical expertise and enthusiasm to keep all equipment in continuous working order, and Lourdes Keico Nagae Takigami for her administrative support and care for the people in the lab, but most of all for our pleasant talks.

I deeply appreciate the CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) foundation for supporting me financially to pursue my studies here at the polytechnic school of the university of São Paulo.

All this work would not have been possible without the constant source of love, encouragement and strength of my family, my parents Gerard & Margreet, Marina & Piet and Heliara, and my good friends Arnoud and Paul.

What is a scientist after all?

It is a curious man looking through a keyhole,
the keyhole of nature,
trying to know what's going on.

(Jacques Yves Cousteau, Christian Science Monitor 21 July 1971)

Abstract

Biodiversidade, a variedade de vida no planeta, está em declínio devido a alterações climáticas, mudanças nas interações da populações e espécies, bem como nas alterações demográficas e na ecologia da paisagem. Avaliações integradas baseadas em modelos desempenham um papel fundamental na compreensão e na exploração destas dinâmicas complexas e tem o seu uso comprovado no planejamento de conservação da biodiversidade. Os objetivos deste estudo de doutorado foram investigar; (1) o uso de técnicas de *genetic programming* e *fuzzy* para construir modelos de alta qualidade que lidam com presença e ausência de dados com ruído, (2) a extensão desta solução para explorar o paralelismo inerente à programação genética para acelerar tomadas de decisão e (3) um *framework* conceitual para compartilhar modelos, na expectativa de permitir a síntese de pesquisas. Subsequentemente, a qualidade do método, avaliada com a *true skill statistic*, foi examinada com dois estudos de caso. O primeiro utilizou um conjunto de dados fictícios obtidos a partir da definição de uma espécie virtual, e o segundo utilizou dados de uma espécie de pomba (*Zenaida macroura*) obtidos do *North American Breeding Bird Survey*. Nestes estudos, os modelos foram capazes de predizer a distribuição das espécies de maneira correta mesmo utilizando bases de dados com até 30% de erros nas amostras de presença e de ausência. A implementação paralela utilizando um *cluster* de vinte nós c3.xlarge Amazon EC2 StarCluster, mostrou uma aceleração linear devido a arquitetura de múltiplos *demes* de granulação grossa. O algoritmo de programação genética e *fuzzy* gerado em determinadas condições durante os estudos de caso, foram significativamente melhores na transferência do que os algoritmos do BIOMOD.

Palavras-chave: Algortimos genéticos, Algoritmos úteis e específicos, FUZZY (Inteligência artificial), Bioclimatologia.

Abstract

Biodiversity, the variety of life on the planet, is declining due to climate change, population and species interactions and as the result of demographic and landscape dynamics. Integrated model-based assessments play a key role in understanding and exploring these complex dynamics and have proven use in conservation planning. Model-based assessments using Species Distribution Models constitute an efficient means of translating limited point data to distribution probability maps for current and future scenarios in support of conservation decision making. The aims of this doctoral study were to investigate; (1) the use of a hybrid fuzzy genetic programming to build high quality models that handle noisy real-world presence and absence data, (2) the extension of this solution to exploit the parallelism inherent to genetic programming for fast scenario based decision making tasks, and (3) a conceptual framework to share these models in the hope of enabling research synthesis. Subsequent to this, the quality of the method, evaluated with the true skill statistic, was examined with two case studies. The first with a dataset obtained by defining a virtual species, and the second with data extracted from the North American Breeding Bird Survey relating to the mourning dove (*Zenaida macroura*). In these studies, the produced models effectively predicted the species distribution up to 30% of error rate in both presence and absence samples. The parallel implementation based on a twenty-node c3.xlarge Amazon EC2 StarCluster showed a linear speedup due to the multiple-deme coarse-grained design. The hybrid fuzzy genetic programming algorithm generated under certain conditions during the case studies significantly better transferable models.

Keywords: Genetic Algorithms, Applied and specific algorithms, Fuzzy logic, Species Distribution Modeling, ecological niche models.

List of Figures

List of Tables

List of Listings

Contents

1 Introduction

Contents

The Millennium Ecosystem Assessment (2005) involved over 1360 experts and 95 countries who identified significant contributions of biodiversity in natural ecosystems to human life and well-being. Yet biodiversity, the variety of life on the planet, is declining due to complex causes, such as climate change, population and species interactions (STUART, 2004; MCCANN, 2000; BELLWOOD et al., 2004) and as a result of demographic and landscape dynamics (KEITH et al., 2008). There is clearly a need for well-informed political decision-making in front of the global challenges of biodiversity loss, climate change, and associated food security concerns.

In this context, integrated model-based assessments play a key role in understanding the causes of biodiversity decline, to explore and assess their relations and impact (SIEBER et al., 2010). The conservation of species is not a scientific choice, but depends heavily on the values, mission, or legal mandate of the organisation producing the conservation plan (MURDOCH et al., 2007). Policy makers use conservation tools, such as Marxan (BALL; POSSINGHAM; WATTS, 2009), to evaluate the impact of their choices. Planning by major organisations has already been aided by these tools (PRESSEY et al., 2007), for example, the planning and rezoning of the Great Barrier Reef (FERNANDES et al., 2005). Sarkar et al. (2006) defines a conservation planning tool as a software that at the very minimum identifies either sets of complementary sites needed to achieve quantitative targets for biodiversity features or the complimentary contribution that individual sites make to biodiversity conservation within a region. To that end it is essential to know which features and species are present at sites.

1.1 Problem statement

The intent of this study is to improve the quality of species distribution models using a proposed hybrid algorithm with genetic programming and fuzzy operators. The problem statement is defined as:

Thesis: Not just ecological models, hybrid fuzzy - evolutionary models in extendible environments allow for better predictions.

1.2 Objectives

The above thesis will be demonstrated by fulfilling the following research objectives:

- Design an algorithm that uses genetic programming and fuzzy operators.
- Implement the algorithm and make it available using open source solution.
- Apply the solution to real-world data and verify the thesis statement.
- Consider how models can be re-used for research synthesis.

1.3 Contribution

The contribution to knowledge by answering the thesis is threefold. First, the research will provide a comparative analysis of the impact of fuzzy genetic programming on Species Distribution Model (SDM).

Second, the research provides predictable and understandable models and discusses a conceptual framework and the importance to share models, perhaps even more so than underlaying data.

Third, a demonstration of the architecture on a fully operable modelling solution in R is made available as open source so that others can assess the validity and performance of this methodology.

1.4 Structure

The remainder of this thesis is structured as follows.

- Chapter 2 provides an overview of a research survey concerning relevant background concepts and work related to SDM.
- Chapter 3 includes the research methodology of this dissertation.
- Chapter 4 investigates a new algorithm which combines the concepts of linear genetic programming and fuzzy rule-based systems, which is preceded by a short literature review on genetic programming. It also discusses the interfaces that are used to execute both case studies.
- Chapter 5 Discusses two case studies that are performed with the new algorithm. Experiments are performed to compare the performance of the new algorithm to those provided in the popular BIOMOD package. A theoretical analysis is per-

formed with the first case study, while the second experiment is used to obtain more empirical results.

- While not part of the overall subject of this thesis, Chapter 6 investigates a concept architecture to increase sharing of reproducible models, an important issue that arose during my work.
- Finally Chapter 7 presents the major conclusions of this work and potential future work directions.

2 Species Distribution Models

Contents

The niche is a central concept in ecological thinking and is related to the presence of species in a certain environment, the relations and requirements it has for that environment, and its relationship with other species. Using this concept ecologists try to answer many questions, for instance: Why is a species present in a geographic area? What is its relation with other species? How will the species distribution change over space and time? What is the impact of a species on its local environment? What is a species function? Perhaps, equally important: What is a niche exactly? Since the term *niche* is used differently and in distinct contexts. This chapter discusses the niche and the models that describe them.

2.0.1 The Grinnellian and Eltonian niche

Chase and Leibold (2003), Colwell and Rangel (2009) note that ecologists are avoiding the term *niche*, unlike most articles in the 1960s and 70s where the term was used in a quarter of the publications in the journal *Ecology*. However, researchers started to use the term in different contexts, causing *niche* to become ambiguous and fall into disuse. The most

cited origin of the concept *niche* in a scientific journal is *The Niche-Relationships of the California Thrasher* by Grinnell (1917). However, he already used the term, although not in the title, in his PhD dissertation in 1913 (GRIESEMER, 1994). The term was used in various publications and in similar context as early as the 19th century and probably even as early as 1833 (GIBSON-REINEMER, 2015). Grinnell used *niche* in his early articles to describe the various circumstances and the adaptations in physical structure and temperament of a species to those circumstances, where each species has its own and unique niche. A niche is then the place that a species occupies in its environment and is based on the idea of Darwin that improvements by natural selection means the better species will competitively exclude other species from occupying that same place. Limiting circumstances that Grinnell later considered in his work are biological factors — such as, food, shelter, competition, parasitism, and overpopulation — and abiotical factors — such as, temperature, rainfall and soil conditions —.

Elton (1927) in his book *Animal Ecology* concerned himself similarly and independently of Grinnell with the niche. Elton focused on the impact of the species on his environment to understand the distribution of species. Due to his focus on food chains he defined the *niche* relating to who eats who. Elton wrote "and the 'niche' of an animal means its place in the biotic environment, its relations to food and enemies". However, he did not limit himself to just food as the limiting factor of the niche. He often mentioned suitable soil types as limiting factors for nesting birds and feeding places(GRIESEMER, 1994).

Many texts make a strict distinction between Grinnellian and Eltonian niches (SOBERÓN, 2007; AUSTIN, 2002; PETERSON, 2011), each representing a different end of a spectrum. On one side, the Grinnellian niche representing the habitat idea which is defined by non-interactive factors and conditions. On the other side, the Eltonian niche that is defined by the biotic interactions and the function of species within its environment. While this distinction is useful as it helps to better define what is meant with the niche and the required data to define it, this distinction is not that clear in their actual works as both authors considered biotic and abiotic factors. The difference is mainly in their approach and the problem that they tried to solve and the question whether multiple species are able to occupy the same niche (GRIESEMER, 1994).

Cox (1980) points out that both Elton and Grinnell in their initial publications mainly used the niche as a metaphor and have not explored the term in detail. Only in later work both gave the term more context. Since its initial appearance, the metaphor has been used in various ways. Even before Grinnell and Elton made the term popular, *niche* was already used in a review published in *Nature* of a book written by Grant Allen, titled *Vignettes from Nature* (1881). As both Gibson-Reinemer (2015) and Cox (1980) wrote, the long history of the meaning of *niche* prior to Grinnell and Elton emphasises the importance

of the conceptual richness of the term and that it might be more important than who conceived it.

2.0.2 The concept of niches

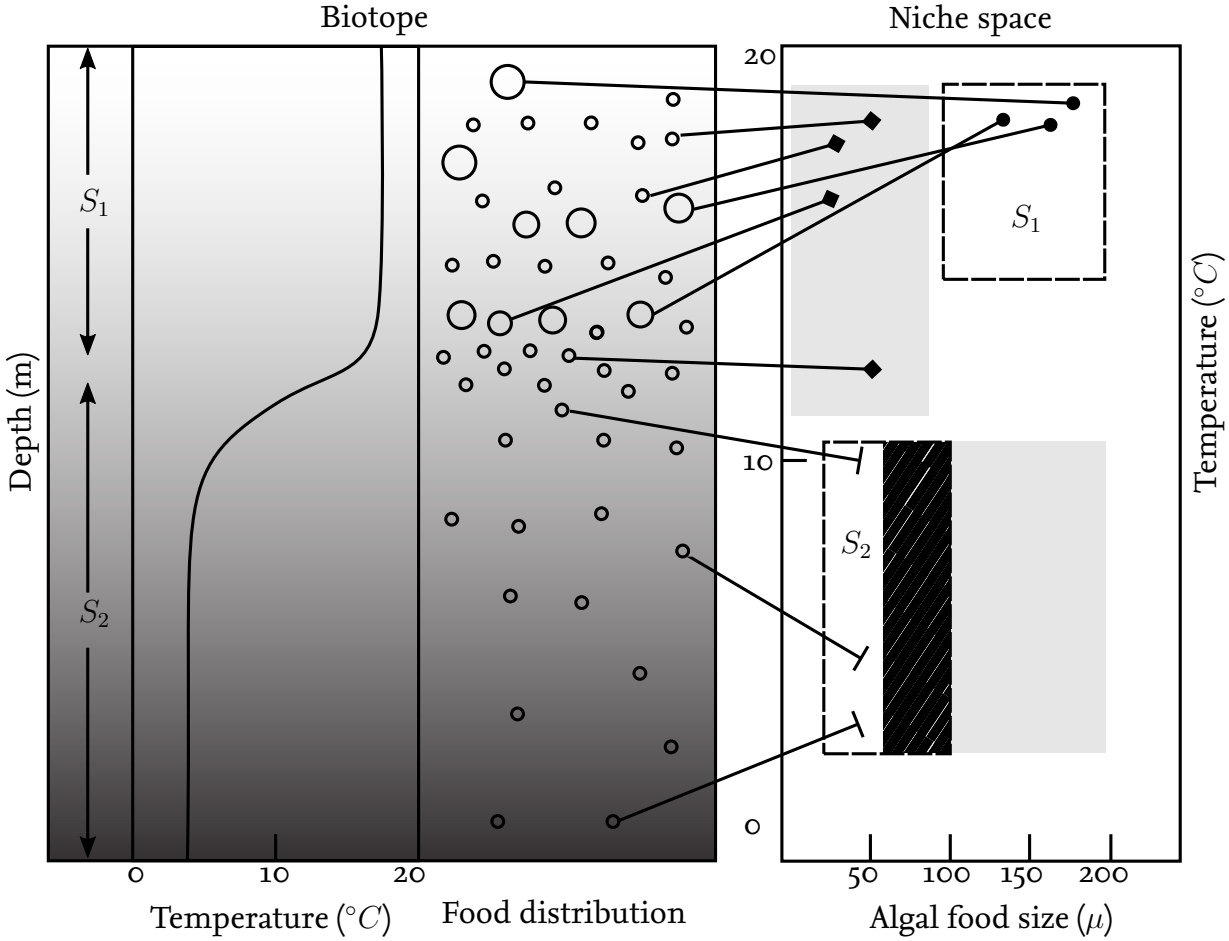
The concept of a niche has been shown and recognised to be a confusing one as there is no single concept of a niche and the word is interchangeably used in different contexts by different authors. This unclarity leads to further confusion when terms depending on the niche concept are defined (WHITTAKER; LEVIN; ROOT, 1973; KEARNEY, 2006; CHASE; LEIBOLD, 2003). Whittaker, Levin and Root (1973) lists the following common current uses for *niche*: to indicate the position or role of a species within a given community, or, as the distributional relation of a species to a range of environments and communities, or to indicate and mixture both previous concepts.

In this work, and in almost all others that concern Species Distribution Models (SDMs), the definition used is based on the work of Hutchinson (1957), who defined it as: "We may now introduce another variable x_3 and obtain a volume, and then further variables $x_4 ... x_n$ until all of the ecological factors relative to S_1 have been considered. In this way an n-dimensional hyper-volume is defined, every point in which corresponds to a state of the environment which would permit the species S_1 to exist indefinitely". This view describes the niche as a place and where the distributional relation of a species (S_1) depends on a range of variables ($x_0 ... x_n$).

This definition still leaves questions that need to be answered, such as: What are the variables? How do they relate to each other and to the state of the environment? and When does a species "exist indefinitely"? To answer the first question, Hutchinson (1978) divided these variables into two categories: dynamic, interacting, resource-related variables for which there is competition; and variables that define environmental conditions for which there is no competition. The first category was named *bionomic* and the second *scenopoetic*. These two distinct categories have been again and again used, although with different names (SOBERÓN, 2007), i.e., *direct/resource* variables (AUSTIN, 2002), and *condition/resource* variables (BEGON; TOWNSEND; HARPER, 2006). While local interactions definitely have impact on population size on small scales, broader distribution patterns depend largely on scenopoetic variables and other variables are often considered as noise, or just ignored (CHASE; LEIBOLD, 2003; PETERSON, 2011).

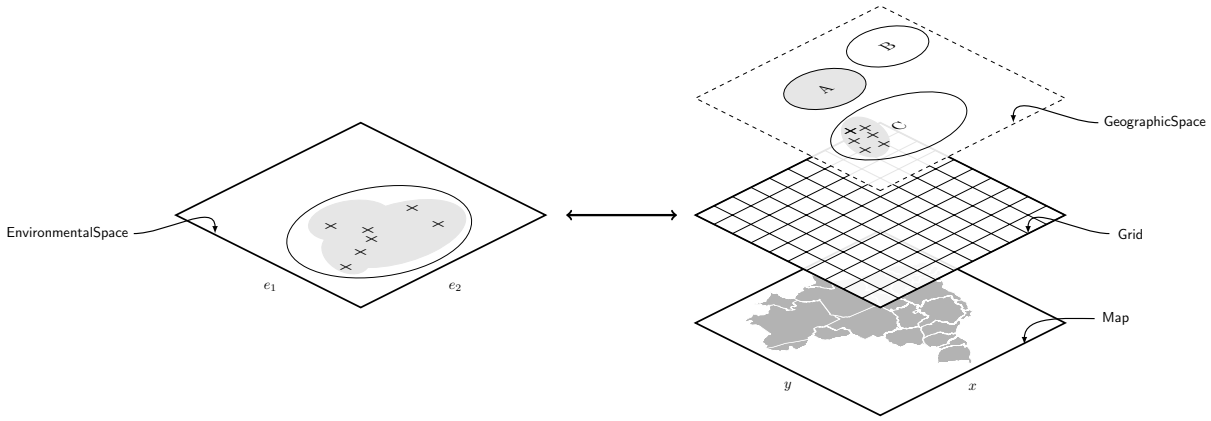
This duality is further illustrated by Hutchinson in Figure 1 which shows many features characteristic of real-world applications. The figure depicts a lake with a niche space with two variables: temperature (i.e., a scenopoetic variable) and the algal food size (i.e., a bionomic variable). The niche space contains two niches of two hypothetical micro algae-eating species: one of S_1 a larger warm-water species that feeds on large algae, and S_2 a cool-water species that feeds on smaller algae. The figure demonstrates the reciprocal

Figure 1 – An illustration used by Hutchinson to show the niche-biotope duality for a lake with a temperature gradient and two species (S_1 and S_2) that consume algae. The biotope is defined by two factors that vary over the depth: the temperature and the distribution of food. The larger algae (circles) correspond with the niche of S_1 in the niche space, while the smaller algae (dots) of S_2 correspond with another niche defined in niche space. The distribution of the algae in the biotope depends on their niche. One sees that the smaller algae also exist on the top half of the biotope as that part of the niche is not utilised by S_1. On the other hand the larger algae of S_2 are not present at low temperatures as the niche is unavailable.



Source: Adapted from Hutchinson (1978)

Figure 2 – ● Occupied niche (left) - Actual distribution (right), + Observed species occurrence record, ○ Fundamental niche (left) - Potential distribution (right). Illustration of the mapping between the geographical space and environmental space for a species. The environmental space (left) is as Hutchinson defined it, see Section 2.0.2, and consists here out of two factors, the axis e_1 and e_2. The $+$-ses represent occurrence records. The ● areas in the environmental niche represent the occupied niche, while they represent those areas that are occupied by the species in geographic space. The mapping of the environmental space to geographical space is one-to-many, shown by a single region of environmental space mapped onto three distinct regions (A,B,C) in geographic space. In region B there is an area that is suitable but not occupied, maybe because the species hasn't been able to disperse to there. On the other hand region A is completely suitable and also occupied. Region C is only partially occupied, maybe because there is competition or maybe other factors are limiting the species distribution.



Source: Author

mapping of the niche with the biotope as discussed in the last paragraph. Furthermore, it shows that niches have limiting factors, for example, algae food size and temperature, or in general: resources and conditions.

It is important to realise that the concept of a niche as defined by Hutchinson (1957), Hutchinson (1978) is distinct from its habitat. While the niche describes the environmental space where both scenopoetic and bionomic variables are favourable, the habitat is then the physical spaces in which species live. This one-to-many duality concept, as in Figure 1, enables to: think of the niche, define and project a niche to a physical space, infer the niche from distribution in the world, realise that the mapping is likely non-linear, have parts of the niche un-utilised or unavailable in geographic space, or to have those parts simply not used (COLWELL; RANGEL, 2009). It is this distinction and these insights that are the foundation of species distribution modelling.

2.0.3 The relation between environmental and geographic spaces

Figure 2 shows the mapping of the environmental space $\boldsymbol{E}$ consisting out of n environmental variables, in the figure illustrated by e_1 and e_2. As discussed, these environmental variables are factors such as: rainfall, coldest month, topography, average temperate. All variables typically defined with a maximum resolution of 1 km² per grid cell, which is mapped on real world data, see Section 2.1.3. On the other side there is the geographic space $\boldsymbol{G}$, equal to Hutchinson's biotope, that typically is composed of a two dimensional grid of a certain region, such as a country or continent, with its own resolution, and often different compared to the $\boldsymbol{E}$ grids. The mapping between $\boldsymbol{E}$ and $\boldsymbol{G}$ is *Hutchinson's Duality* from Section 2.0.2.

More formal, at every cell $g \in \boldsymbol{G}$ all n layers of the environmental factors are measurable to obtain $\vec{e_g}$, where $\vec{e_g} = (e_1, e_2, \ldots, e_n)_g$. This vector is obtainable for every cell and is not necessarily unique for each cell, e.g., multiple locations in geographic space can have the same temperature and altitude combination assuming only those two factors are considered. The space of all vectors $\vec{e_g}$ combined is the actual environmental space $\boldsymbol{E}$ of which a species typically only occupies a subset. Due to effects, such as climate change and invasive species, $\boldsymbol{E}$ is not a static space, but a dynamic one that changes over time. Since $\boldsymbol{E}$ is build from vectors e_g the rank of $\boldsymbol{G}$ and $\boldsymbol{E}$ is typically the same. However, if there are a large number of factors it is common, e.g. in Giannini et al. (2013), for scientist to apply Principal Component Analysis (PCA) in which case $\text{rank}(\boldsymbol{E}) \leq \text{rank}(\boldsymbol{G})$.

Grid resolution, the size of the cells in $\boldsymbol{G}$, impacts the predicted distribution. Scientists need to be aware of the Modifiable Areal Unit Problem (OPENSHAW; TAYLOR, 1979) that is a source of statistical bias that radically affects the results of spatial phenomena when grid sizes are varied. The use of varied grid sizes is common when different sources of environmental data is used and significantly impacts the produced results (SEO et al., 2009). Seo et al. (2009) therefore recommends the use of a grid size of 1 km². This result is in contrast to an extensive study by Guisan et al. (2007) which has shown no severe impact on model quality when the grid resolution had a ten fold coarsening. They did find an overall model quality degradation, but only noticeable for models that already had sufficient performance and/or with initial data that have an intrinsic error smaller than the coarser grid cell size.

The sets $\boldsymbol{G}$ and $\boldsymbol{E}$ have till so far been sets defined and limited solely by a geographic extent. For the concepts to be useful it is necessary to define subset $\boldsymbol{G'} \subset \boldsymbol{G}$ to indicate the geographic area where the species is present. The environment factors present on a single cell of that region is denoted as $\eta(g) = \vec{e_g}$ and for all occupied cells $\eta(\boldsymbol{G'}) = \{\eta(g) | g \in \boldsymbol{G'}\}$. The inverse, mapping environment space on to geographic space is denoted as $\eta^{-1}(\boldsymbol{E'}) = \{g \in \boldsymbol{G} | \vec{e_g} \in \boldsymbol{E'}\}$. Both operations are easily performed with Geographic Information System (GIS) tools, e.g. ArcMap and OpenGIS, and are a first step

in preparing data for model training, evaluation and testing.

2.0.4 The BAM diagram

Equation 2.1 based on ideas of Vandermeer (1972) shows the relationship and factors that affect population growth in an Eltonian grid. It describes the relationship between the distribution of a species in space and its niches, based on the environments that it occupies to predict where it actually or potentially occurs. The first factor of the formula $\frac{1}{x_{i,g}}\frac{\mathrm{d}x_{i,g}}{\mathrm{d}t}$ represents the growth rate of species i in area g, and where $x_{i,g}$ is the density of the species at that particular area. Assuming there is no migration between cells, then the growth rate is determined by the intrinsic growth rate r and φ that represents the densities of all other species that affect the species.

$$\frac{1}{x_{i,g}}\frac{\mathrm{d}x_{i,g}}{\mathrm{d}t} = r_{i,g}(\vec{e}_g - \varphi_{i,g}(\vec{e}_g, \vec{R}_{i,g}; \vec{x}_g) + \Psi(\boldsymbol{T}_i; \vec{x}_i) \tag{2.1}$$

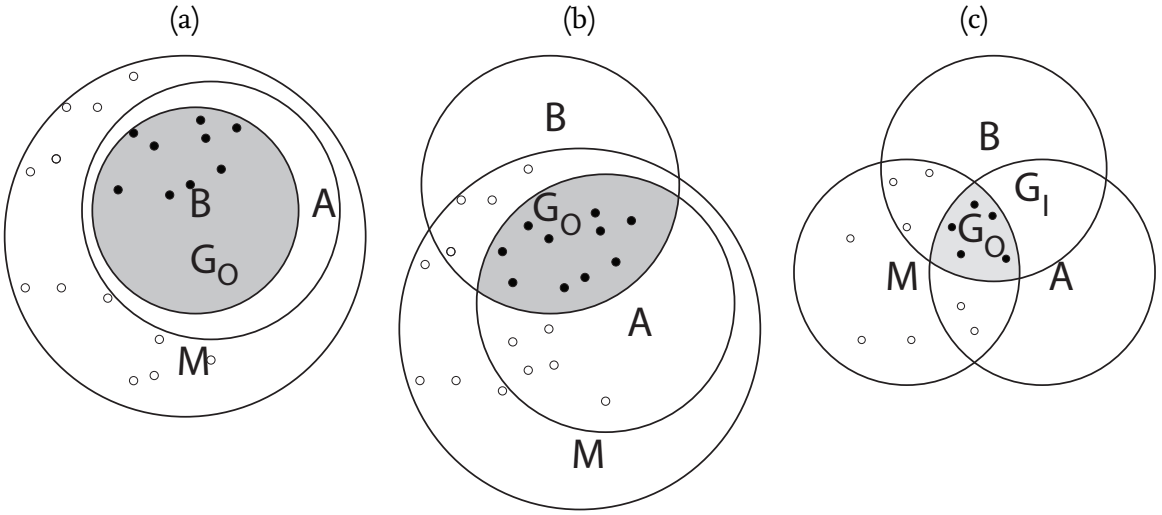
In general the growth rate of the species is determined by the scenopoetic variables (vector $\vec{e}_g$) and the biotic factors φ, such as its interactions with other species, resources and diseases (vector $\vec{R}_{i,g}$). While growth is one factor that determines a population density another factor is dispersal, migration or movement, of the species in its environment. This factor is represented by $\boldsymbol{T}$ in the equation and it is a transition matrix that describes possible movements from area to area.

These three factors combined are detrimental to discuss and find the area of distribution of the species. A graphical representation of the ideas behind this formula is shown in Figure 3. In this figure Set $\boldsymbol{A}$ characterises the geographic area where the scenopoetic variables are such that there is a positive growth rate for that area. The Set $\boldsymbol{B}$ represents the area where the biotic factors are beneficial to the presence of the species. The third area shown in the figure area $\boldsymbol{M}$ is analogous to the last factor in Equation 2.1 and are the regions where the species historically over the generations has been able to disperse to.

Figure 3 show a Venn diagram of the areas $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{M}$ and where these areas overlap. The area ($\boldsymbol{G_o}$) $\boldsymbol{A} \cap \boldsymbol{B} \cap \boldsymbol{M}$ is the *occupied distributional area.* This is the region where all the variables are favourable for the species and where it currently resides. The area ($\boldsymbol{G_I}$) $\boldsymbol{A} \cap \boldsymbol{B} \cap \boldsymbol{M}^C$ in contrast is the area where the species could survive but has not yet spread out to, this is the *invadable distributional area.* Both areas together $\boldsymbol{G_P}$, or ($\boldsymbol{G_o} \cup \boldsymbol{G_i}$), is the *potential distributional area.*

This last area, $\boldsymbol{G_P}$, is the to be discovered area when invasive species are investigated by modelling a species' niche. It is important to note however that the models are constructed with data points obtained from those places where the species currently resides, in other words from the area enclosed by $\boldsymbol{M}$ this is indicated by the black dots in

Figure 3 – Set A represents regions in space where scenopoetic, non interacting, variables are favourable for a species. Region B represents areas where biological conditions, e.g, competitors, predators, diseases, are favourable. M represents regions to which species has access due to movement, barriers and distances. Go, with black dots as presences, is the actual distribution while Gi, with white dots as not yet occupied presences, is the potential area distribution currently limited by the movement. The challenge is finding out what restricts the species, nothing as in (a), by biotic factors (b) or by movement (c)



Source: Author

the figure indicating places where the species occurs. Places where the species is absent are indicated by white circles.

The region encompassed by the circle $\boldsymbol{A}$, $\boldsymbol{G_A}$, is the region where the scenopoetic variables are favourable for the species and which would be populated if there was no competition and assuming perfect dispersal abilities of the species. The environment that exists in this area is $\boldsymbol{E_A}$. $\boldsymbol{E_A}$ is the *existing fundamental niche* and is a subset of the *fundamental niche* $\boldsymbol{E}$, representing the entire spectrum of environmental conditions that are favourable, and obtainable through the operation $\eta(\boldsymbol{G_A})$ discussed in Section 2.0.3. The challenge is determining the ecological niche of a species $\boldsymbol{E_A}$ that hopefully is representative of $\boldsymbol{E}$.

2.1 Geographic areas and ecological niches

The aim of any species distribution modelling algorithm is to identify and estimate the regions of $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{M}$ of Figure 3. Algorithms find similarities between geographic areas and environmental space and use this relationship to interpolate and extrapolate spatially to other regions. A model is said to interpolate when it projects in a region similar to as what it was trained on, as seen in the grey area in region $\boldsymbol{C}$ with many observed occurrences (+) in Figure 2. For interpolation on coarse scales SDMs are not necessary

as it suffices to just predict based on the proximity of other presences (BAHN; MCGILL, 2007). Extrapolation, on the other hand, is to areas that are statistically independent of the training region, i.e into region $\boldsymbol{A}$ and $\boldsymbol{B}$ in Figure 2 with no known occurrences and with environmental conditions not available in the calibration data, see Section 2.2.1. Spatial extrapolation, or projection outside known regions with observed presences, is also known as spatial transferability and requires distribution models that are capable to generalise their training data set, but unlike with extrapolation the environment is similar to that of the calibration region.

Any modelling algorithm that maps between $\boldsymbol{G}$ and $\boldsymbol{E}$ will map between the geospatial range of $\boldsymbol{G_P}$ and $\boldsymbol{G_O}$ and the environmental range $\boldsymbol{E_P}$ and $\boldsymbol{E_O}$. In other words, the model predicts between what is currently occupied and that what potentially is occupied by the species. Whether it is closer to Potential or Occupied will depend on from which region the training data was sampled, the uniformity of the sampling, the overlap of the regions $\boldsymbol{A}$,$\boldsymbol{B}$, $\boldsymbol{M}$ (see Figure 3), and the properties of the algorithm. It is worth noting that even though a model is trainable with results close to $\boldsymbol{E_O}$ based on sampled data from $\boldsymbol{G_O}$, the reverse action will not result in a direct one-to-one mapping back to $\boldsymbol{G_O}$, but instead to the larger area of $\boldsymbol{G_P}$. So while an environmental niche model is obtainable from the area the species occupies, the inverse, to determine the area the species occupies is not possible from that same model. The reason is that there are possibly many regions with the same environmental conditions and there is no way knowing which one is occupied and which one is not.

Figure 3 shows three configurations to illustrate the point that the sampling area, indicated by the occurrences in the figure, and the circumstances of the species, indicated by B, A and M, determine in the end what region of the species is predicted. In Figure 3a the species is a great disperser and as such $\boldsymbol{M}$ encapsulates the intersection of $\boldsymbol{A}$ and $\boldsymbol{B}$. In this case potential and occupied areas are almost the same as $\boldsymbol{A}$, or $\mu(\boldsymbol{G}, \boldsymbol{E}) = \boldsymbol{G_A} \simeq \boldsymbol{G_O}$. In Figure 3c the situation is quite different as all regions only partially overlap. In this case, due to historical reasons and the dispersion rate, a species has not been able to disperse into all suitable invadable areas, indicated with $\boldsymbol{G_I}$. This case is different as only a part of the known environmental space of the species can and has been sampled. Any useful model prediction $\mu(\boldsymbol{G}, \boldsymbol{E})$ will have to be able to predict into the $\boldsymbol{G_I}$ region and at the same time the sampled niche $\boldsymbol{E_O}$ might be far from complete, the species possibly occurs in many more geographic spaces than assumed. This is important to realise when predicting future scenarios. For example, a species might now be assumed not to occur in regions higher than 20 °C and thus to be extinct in the future due to global warming. In reality, it has never been able to disperse to warmer areas due to other constraints in $\boldsymbol{M}$ and $\boldsymbol{B}$. As a result its environmental space was not completely known and could not be sampled, causing this inaccurate prediction.

Therefore, with estimating geographic areas and ecological niches it is important to understand the BAM framework and the consequences of it on the predictions. As Peterson (2011) states a modelling algorithm μ is desired that predicts $\boldsymbol{G_O} \subseteq \mu(\boldsymbol{G_{DATA}}, \boldsymbol{E}) \subseteq \boldsymbol{A}$, where $\boldsymbol{G_{DATA}}$ is the set of presence and absence points. The idea is that the model predicts presence location similar to $\boldsymbol{G_+}$, the set of true presences. The set of predicted presences will most likely have omission and commission errors. Omission errors are those locations where the species should have been predicted to occur, while commission errors on the other hand are locations where the species is predicted to occur, but in reality do not occur. Section 2.2.4.2 discusses how these errors occur in spatial predictions and the effects that they may have on the model.

2.1.1 Steps to build niche models

SDMs have already been shown to contribute to biodiversity conservation management, discussed in Section 2.3, and help understand the optimum conditions for dispersal and to see impact of species populations with possible future climate changes. All these studies follow, although not explicitly, the same steps to make these predictions. This process to compute and present meaningful species distributions has been described in detail in Guisan and Zimmermann (2000), Hirzel and Lay (2008) and in Santana et al. (2008). Summarised, these are the steps:

Step 1. Problem definition

Before anything else it is important to know the objectives and reason for the modelling and the questions that need answering with the resulting outputs, because Peterson and Vieglais (2001) has shown that different goals combined with the wrong modelling technique impacts the quality of the outcomes.

Step 2. Select and prepare scenopoetic variables

The preparation of scenopoetic layers is one if the most time consuming and computationally intensive tasks. However, the task has become easier with the release and use of standardised layers, such as WorldClim (HIJMANS et al., 2005; CHAPMAN; MUÑOZ; KOCH, 2005). Even with the use of standardised layers, combining several sources requires a common geographic coordinate system and/or projection. Coordinate system transformations and resampling are frequently a required preprocessing task for the environmental layers.

Step 3. Select the area under study

Anderson and Raza (2010) show that the definition of the study region and calibrating inside that region plays an important role in obtaining realistic predictions. In addition, models produced with an overly large study area are likely to show low

transferability, in other words reducing their prediction quality in different biological contexts.

Step 4. Identify suitable presences and absences

Species occurrence records will need to be obtained if not a mechanistic approach is used to correlate the species occurrences with the environmental variables.

Step 5. Select a suitable modelling technique(s)

There are many different modelling techniques and algorithms available that differ in their distribution of the response, in their used fitness functions, and their use of occurrence data (ELITH et al., 2006). This is even more so complicated, because the amount of choice here is enormous as Araujo and Correa (2007) argues that a combination of several approaches should be used to create more robust models.

Step 6. Understand the response of the modelling technique

According to the no free lunch theorem (WOLPERT; MACREADY, 1997) if an optimisation algorithm performs well for certain applications then it necessarily performs worse for all other applications (WOLPERT; MACREADY, 1997). Therefore, to select the best modelling technique it is necessary to understand the response of the model in the context of the defined problem.

Step 7. Select an appropriate threshold

For continuous model outputs often a threshold is selected to transform the output to either a presence or an absence. The selection of this threshold has a large impact on the output of the model. For example, with climate change studies the impact of threshold values causes a 1.7- to 9.9 fold differences in the proportions of species projected to become threatened (NENZÉN; ARAÚJO, 2011).

Step 8. Test the sensitivity and robustness of the model

Using statistics and established performance metrics models are objectively analysed and compared to establish model quality, where the definition of quality will depend on the problem domain.

Step 9. Interpret the model results for ecological correctness

The effect of small differences in performance metrics may still be meaningful in the biological sense (PHILLIPS; ANDERSON; SCHAPIRE, 2006; AUSTIN, 2002). For this reason model projections should be visually inspected and not just statistically.

Step 10. Evaluate the predictions

There is variance in predictive power and model output. It is therefore important to understand the uncertainty in model predictions in the context that they are used (WATLING et al., 2015).

Step 11. Show confidence levels for the model output

Hirzel and Lay (2008) suggest to show geographically the areas where the model is

either interpolated or extrapolated to get a better understanding of the output.

Step 12. Reclassify the prediction in meaningful classes

While continuous output might give more information it is misleading due to uncertainty. Hirzel et al. (2006) therefore suggests reclassifying the model output in fewer classes, e.g., unsuitable, marginal, suitable and optimal, giving a better understanding and representation of the contained information.

Step 13. Consider home range, exclude small isolated regions

The appropriate resolution of the projecting, and training, might correlate with the home range of the species and how it uses the surrounding resources. Small patches of suitable areas smaller than the home range might be better excluded (GUISAN; THUILLER, 2005).

Step 14. Expert validation

In the end, only with careful examination and evaluation by experts can a model be asserted as useful, because there are many more aspects and variables to species distributions not covered by SDM.

2.1.2 Occurence data

Knowing where a species is found is an integral ingredient for species distribution modelling as the process often relies on presence-only occurrence records. Although, for some modelling techniques systematic presence-absence occurrence records are needed. Historically, such records have been collected opportunistically as often the more accessible areas such as river beds and roads were investigated for taxonomic studies in museums and herbaria (WILLIAM et al., 1996). Nowadays, there are many sources for these records and with various qualities.

2.1.2.1 Primary and Secondary Occurrence Data

Peterson, Stockwell and Kluza (2002) divide occurrences into primary and secondary information. Primary information is those occurrences that are obtained through direct observation or documentation. Primary occurrences often come in the form of collected specimens for museums. Secondary occurrence are those obtained through range maps, species geographies, description of the species niche or even projections of SDMs.

It may be tempting to use secondary data for modelling as it is already cleaned and comes often in the convenient form of polygons, however, there are some complications. First, one needs to trust that the expert who made the map is truly an expert in knowing which non-sampled regions contain populations. Secondly, secondary data needs to be published and thus lags behind current knowledge. Finally, a publication is static and as

a consequence its quality will degrade over time; new insights are gained, niche stability might not have been achieved, and species distributions continuously change.

Primary occurrence points do not have these drawbacks. Even if they are obtained from old museum collections, some hundred of years old, and therefore are inaccurately, if at all, georeferenced they do not have the earlier mentioned drawbacks. With the proper methodology even inaccurate data is useful for predictions of species distributions with the right robust modelling techniques (GRAHAM et al., 2008). While secondary data gets more inaccurate over time, primary data gets increasingly more accurate as more knowledge is gained.

Not all occurrence records, even if primary, are suitable for generating SDMs. An occurrence record might seem just like a simple recording of a sighting of a particular species at a given point in place and time, but how should one be interpreted?

2.1.2.2 Sampling

To understand that not all occurrence records are created equal it is necessary to see what happens when a species is observed. Typically an observer goes to a place and spots a species. Now, if that species is there because it is an abiotically and biotically suitable place than it means it is a true presence and all is well. However, if the observer spots the species in a place that is abiotically and/or biotically not suitable for long term survival due to dispersal than it is a more complicated case. These type of occurrences are undistinguishable from the true presence points and cause errors in the generated model. Another source of errors in occurrence records is due to identification errors made by the observer, although these points might show up as outliers when analysed together with environment variables.

Not only errors, but also sampling bias gives a distorted view of the true distributional patterns of species. An observer can spot a species, but he might have chosen a physically easier accessible location to investigate. For example, in specific habitat types such as grasslands, or next to railways, rivers, or roads (REDDY; DÁVALOS, 2003). This introduces bias in the distribution of sampling effort and the primary occurrence data will misrepresent the ecological tolerances of the species. Furthermore, bias is also introduced due to the fact that some species are simply detectable in some habitat types more than in others.

Absence records have the same problems. Even if a location is abiotically and biotically suitable and accessible for a species, there is still chance that it goes undetected by an observer. This will result in false absence record. Other false absences occur because species have temporal variations in their presence. Due to bird migration an area might be considered unsuitable while in fact it is only partly so. The BAM diagram, Figure 3, shows another possibility, namely, a location is biotically and abiotically suitable, but the

species has not dispersed to there yet.

Due to the complications mentioned above, occurrence records should be carefully evaluated and inspected for bias and data quality assured before they are used to train models (ELITH et al., 2006; CHAPMAN, 2005). The use of absence points is even more complicated. True absence points are difficult to obtain, but should be used when available (BRAMEIER; BANZHAF, 2007), and the use of pseudo/background data has its own complications (ANDERSON; RAZA, 2010; PHILLIPS et al., 2009; SMITH; FRANKLIN, 2013) and controversions (HASTIE; FITHIAN, 2013).

2.1.2.3 Occurrence Data Content and Availability

Species distribution modelling is commonly hard as useful occurrence records are scarce and difficult to obtain. Models are typically trained with a small number of occurrences, even as few as twenty-five records (JACKSON; ROBERTSON, 2011; PEARSON et al., 2007).

Historically specimens were stored in museums and herbaria which have build up an extensive collection over the years that also represents an irreplaceable legacy information about our biosphere and biodiversity loss. Ariño (2010) estimates that the total museum collection size is in the order of one to two giga specimens of which only about three percent is web-accessible. This number is even higher as not all collections might have emerged through literature and were unknown to the authors.

That makes these collections the main source of occurrence records (PONDER et al., 2001). Other sources are: new field studies, data extracted from earlier articles, data extracted from published range maps, the sharing of unpublished data among collaborators and citizen science. Most sources, however, do not have data in digitalised formats available with adequate metadata that facilitate use and synthesis. Many institution simply do not have the financial, technological and staffing resources to mobilise this enormous amount of information. The complete digitisation of all natural history collections may cost as much as €150,000 million, and take as long as 1,500 years (BLAGODEROV et al., 2012). As a solution, projects are becoming popular that use citizen science, or volunteers, to play a role in making this data accessible for the research community (HILL et al., 2012).

Even so, just a decade ago one of the biggest challenges would have been obtaining any of the information in these collections. If one was lucky a collection had its own portal to query, or if less lucky, one had to visit the museums and visit the collection in person. Now, after many years of endeavour and advancement there are internationals effort to make all this data accessible and searchable through a single portal.

One successful implementation of such a portal is the Global Biodiversity Information Facility (GBIF), that distributes data primarily on plants, animals, fungi, and microbes for the world, and scientific names data (EDWARDS, 2000). The portal offers cur-

rently more than 640 million occurrences of over 1.6 million species by 782 different data publishers. Data quality is an issue for these records but cleaning that data is far less of an investment of resources than reproducing and documenting the samples held in natural history museums.

Data quality is an issue as most data are managed and published through a wide range of heterogeneous databases. Standardisation is needed for guaranteed, stable and persistent access to each data record. Mesibov (2013) asserted that there are many errors and just aggregating data is not enough; data errors needs to be corrected and aggregators may need to do more effective data checking or provide ways to annotate and return feedback to data providers. Koch Veiga, Cartolano and Saraiva (2014) discusses efforts and mechanisms to prevent some of these types of data errors.

All things considered, these aggregators provide a tremendous amount of species records at the touch of a finger tip [1]. Careful planning and rigorous data cleaning, for example such as in Mesibov (2013), lead to useable occurrence records to build species distribution models, as seen in this work and many many others.

2.1.3 Scenopoetic data

Ecological niche models are based upon two sources of data; the occurrences of species as discussed in Section 2.1.2, and environmental data such as described in this section. As earlier discussed and shown in Equation 2.1 population densities are correlated to favourable scenopoetic and bionomic variables within the environmental space.

In general, species respond to environmental variables in different ways. The variation in those variables often forms the limiting factors for species and operate on different spatial and temporal scales. As a result different variables are relatively more or less influential depending on the scale domain and according to biological and environmental contexts (SOBERÓN, 2007; PEARSON; DAWSON, 2003). For example, macro climatic variables such as the average yearly temperature, work on courser scales than the pH of local plot of land.

Austin (1980) suggested a division of the environmental variables into three types; direct, indirect and resource. Indirect means in this context that the variable has no physiological effect on growth or competition. Examples are altitude, latitude or longitude. Their only impact is due to indirect correlation with other variables. Direct variables on the other hand do have a direct physiological impact on the growth and are not consumed by the species. Examples are temperature and pH. Resources are similar to direct variables except that they are consumed in order for the species to grow; think of water and sunlight.

[1] <http://www.gbif.org/occurrence>GBIF occurrence records

Another subdivision of environmental variables is that one into proximal and distal. The idea here is that the proximal environmental variables have the greatest impact on the population densities of the species, while more distal variables have a lesser impact. For example, local pH will have a greater impact than the average pH in soil. Of the division types, indirect variables are more distal than direct variables.

Models based on proximal variables will be better applicable and more precise than those on distal variables. However, models like that are harder to obtain due to the nature of the precise local data required. Austin (1980) correctly argues that the use of proximal variables for predictive mapping of species distribution is for that reason impractical. For that reason, the training of models involves more distal variables that correlate with the causal variables.

As mentioned before, depending on the scale domain different environmental variables play a role in explaining population densities. The question is then to discover exactly which variables are the deciding factors and which ones are not. Models are frequently based on publicly available dataset that provides accurate values for several variables over large spans of geographical areas and in some case even over periods of time. The case studies, see Section 5, use two such publicly available datasets, namely, *WorldClim* and *Hydro1k*.

The *WorldClim* database developed by Hijmans et al. (2005) is such a dataset and is available for download from <http://www.worldclim.org>. The dataset provides interpolated climate surfaces for global land areas, excluding Antarctica, and was generated through interpolation of average monthly climate data from weather stations on a 30 arc-second grid, also commonly referred to as *1 km* resolution grid. The dataset contains variables that describe the monthly total precipitation and monthly mean, minimum and maximum temperature and also nineteen derived bioclimatic variables. See Table 1 for an overview of all the variables.

The data layers were compiled from several sources, such as the Global Historical Climate Network Dataset, the World Meteorological Organization climatological normals and the United Nations Food And Agriculture Organization global climate database (FAO-CLIM 2.0). The resulting data was generated through interpolation of average monthly climate data from weather stations. After cleaning of the data, precipitation record from 47,554 locations, mean temperature from 24,542 locations, and minimum and maximum temperature for 14,835 locations were obtained to generate variables layers.

The *WorldClim* database also provides historical datasets and datasets with forecasted projections for several possible climate change scenarios. Future datasets are based on general circulation models which implement a mathematical model of the general circulation of the earth's atmosphere or ocean. The output of these models are dependent on the assumed atmospheric concentration of greenhouse gases. *WorldClim* provides the

datasets for four emission scenarios that are used in the Fifth Assessment IPCC report (Intergovernmental Panel on Climate Change, 2014). Species Distribution Models generated and based on current environmental variables are often applied on future projections of climate data for the prediction of invasive species pathways (ACOSTA, 2015; ACOSTA et al., 2016) and biodiversity conservation and planning (GIANNINI et al., 2012).

The *Hydro1k* database (U.S. Geological Survey, 2000) provides a 1 km resolution global coverage of common topographically derived data used in hydrologic analysis to predict movement of water across the earths surface. The digital elevation model data set provides five additional raster data layers: aspect in radial degrees, flow direction, flow accumulation, slope and a compound topographic index. Table 1 describes these layers.

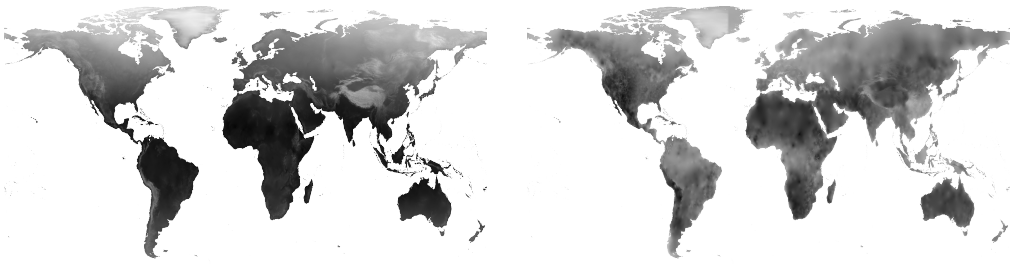
Both databases are used together in this work. Combining *Hydro1k* with *Worldclim* provides a more detailed information as variables have synergy and alter the species' response. As Peterson, Pape and Eaton (2007) point out, the slope of a region, for example, impacts how species experience the climate as the average temperature (obtained from the *WorldClim* database) is higher on south-facing slopes in the Northern Hemisphere. This effect has, according to Acosta (2015), impact on the micro climate and especially that of *Bombus terrestris* colonies, either by impacting the availability of resources, e.g., food and nesting, or its interactions with other species, e.g., plants, predators and diseases. In the end, which variables to use and which ones not, depends highly on the application of the obtained models and their aims and requirements (PETERSON, 2006).

2.2 Species Distribution Models

Modelling algorithms try to find a function, a rule, a procedure, or a program that gives an estimation of relative suitability, $\mu(\boldsymbol{G}, \boldsymbol{E}) = f$ (FERRIER, 2002). This estimation is in many publications considered a probability of occurrence (PEARCE; FERRIER, 2000; KEATING; CHERRY, 2004; PHILLIPS; ANDERSON; SCHAPIRE, 2006), while in others it is a dimensionless suitability index that, for example, ranges from zero to one (HIRZEL et al., 2002). There are many modelling algorithms for SDM. Elith et al. (2006) discuss a comparison of sixteen modelling methods, while new ones, such as the one discussed in this thesis, is ongoing research.

The algorithms differ in their approach (i.e., regression, machine-learning, statistics), in the generated output (continuous, binary, ordinal), in the required input data/maps, their suitability for applications (interpolation, transferability) and their use, if required, of absence data. While some algorithms only require presence data to make predictions (BUSBY, 1991), the majority needs some form of absence data. Absence data comes mainly in three forms; (i) true absence, (ii) background , and (iii) pseudo-absence data. Those algorithms that require absence data contrast the environment at the pres-

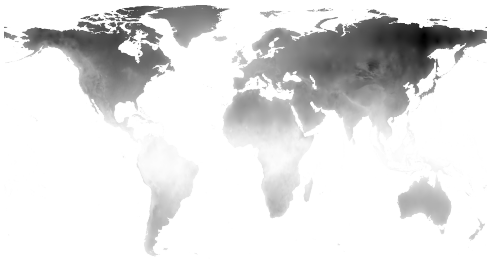
Figure 4 – Plots of the worldclim data layers



(a) Bio1, Annual Mean °C

(b) Bio2,Mean Diurnal Range

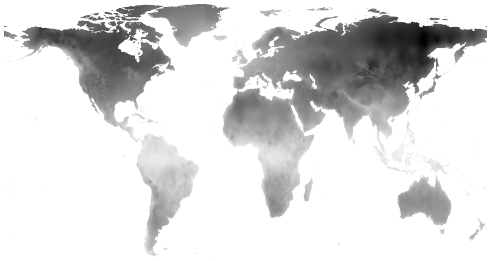
(c) Bio3, Precipitation Wettest Month



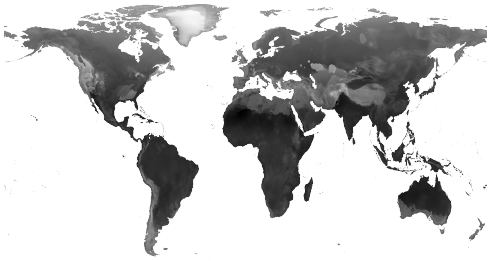
(d) Bio4,Precipitation Driest Month

(e) Bio5,Max °C Warmest Month

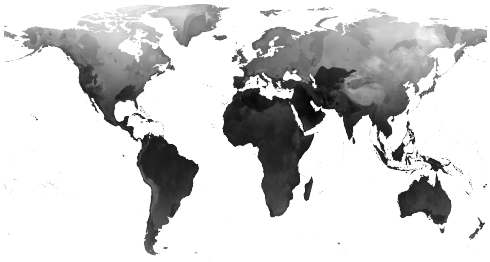
(f) Bio6,Min °C Coldest Month



(g) Bio7,°C Annual Range



(h) Bio8,Mean °C Wettest Quarter



(i) Bio9,Mean °C Driest Quarter

(j) Bio10,Mean °C Warmest Quarter

Source: Adapted from Hijmans et al. (2005)

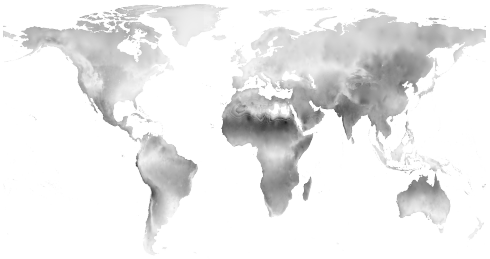
Plots of the worldclim data layers

(k) Bio11, Mean °C Coldest Quarter

(l) Bio12,Annual Precipitation

(m) Bio13,Isothermality

(n) Bio14,Temperature Seasonality



(o) Bio15,Precipitation Seasonality

(p) Bio16,Precipitation Wettest Quarter

(q) Bio17,Precipitation Driest Quarter

(r) Bio18,Precipitation Warmest Quarter

(s) Bio19,Precipitation Coldest Quarter

Source: Adapted from Hijmans et al. (2005)

Table 1 – Overview of the layers provided in the worldclim dataset

Type	Label	Climate Layer	Source
Bioclimatic	BIO1	Annual Mean Temperature	(HIJMANS et al., 2005)
	BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))	
	BIO3	Isothermality (BIO2/BIO7) (* 100)	
	BIO4	Temperature Seasonality (standard deviation *100)	
	BIO5	Max Temperature of Warmest Month	
	BIO6	Min Temperature of Coldest Month	
	BIO7	Temperature Annual Range (BIO5-BIO6)	
	BIO8	Mean Temperature of Wettest Quarter	
	BIO9	Mean Temperature of Driest Quarter	
	BIO10	Mean Temperature of Warmest Quarter	
	BIO11	Mean Temperature of Coldest Quarter	
	BIO12	Annual Precipitation	
	BIO13	Precipitation of Wettest Month	
	BIO14	Precipitation of Driest Month	
	BIO15	Precipitation Seasonality (Coefficient of Variation)	
	BIO16	Precipitation of Wettest Quarter	
	BIO17	Precipitation of Driest Quarter	
	BIO18	Precipitation of Warmest Quarter	
	BIO19	Precipitation of Coldest Quarter	
Climatic	TMEAN	Average monthly mean temperature (°C * 10)	
	TMIN	Average monthly minimum temperature (°C * 10)	
	TMAX	Average monthly maximum temperature (°C * 10)	
	PREC	Average monthly precipitation (mm)	
	ALT	Altitude (elevation above sea level) (m)	
Hydro1k	ASPECT	Direction of maximum rate of change in the elevations	U.S.GeologicalSurvey
	FLOW DIRECTION	The direction to its steepest down-slope neighbour	
	FLOW ACCUMULATION	The amount of upstream area draining into the cell	
	SLOPE	Maximum change in the elevations between cells	
	CTI	The Compound Topographic Index or the Wetness Index	

Source: Author

ence locations with those of the absence data. Ideally true absence data is available, but more often than not it is unavailable to the researcher. For this reason the background and pseudo-absence methods are popular as they artificially create the absences. In the background approach a large sample is uniformly taken from the entire region under study without regard whether a species is present or absent in those samples. The pseudo-absence method differs in that last aspect as samples are only taken from sites where it known not to occur, for example in regions that are too hot/cold, too high/low altitude, too wet/dry. Absence data should optimally still be picked from the region limited by M in the BAM diagram (Figure 3) so that the model is only trained on conditions explainable by the known environmental data from A for the model to have meaning (BARVE et al., 2011).

The construction of models require often adjustment to the parameters of the algorithm, and are algorithm specific. For example, for machine learning methods the maximum number of generations is usually set, and also the acceptable error, and the

population size. FIELDING and BELL (1997) discuss three general principles of model calibration: data splitting, variable selection, and threshold selection. A common practice in machine-learning and other supervised modelling methods is to split the data into three parts when constructing models. To prevent overfitting of the model the available data is split into three sets: training, validation and test. The training dataset is used to train and build the prediction model with the algorithm to fit the parameters of the modelling as best as possible to the data. The validation dataset is used to evaluate independently the trained model to adjust algorithm parameters of the classifier, for example the maximum program length in genetic programming or the maximum number of rules. The test data set is used to pick among the generated models the one that performs and generalises the best on unseen data, and thus is not overfitted to noise in the training dataset.

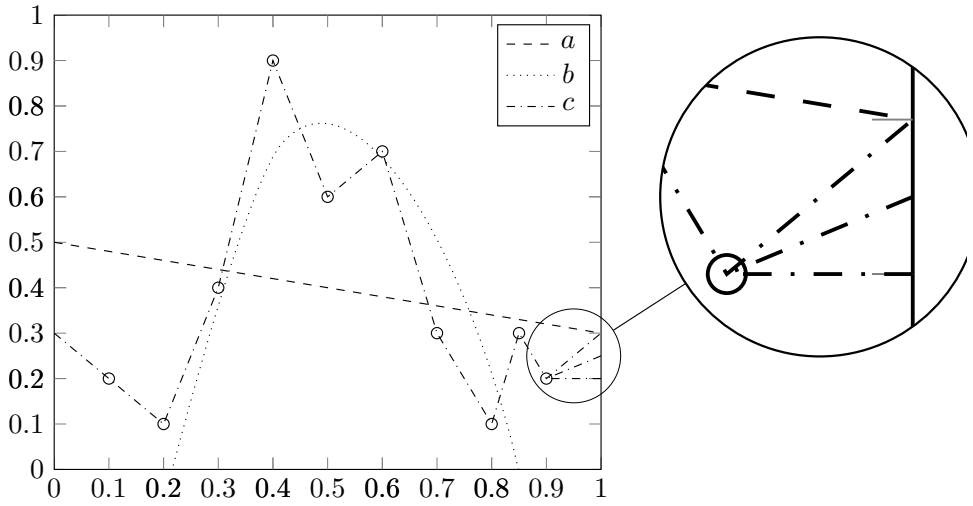
2.2.1 Model complexity and overfitting

Species distribution modelling algorithms operate by searching the environmental space for a function that fits the sampled species environment $\boldsymbol{E_O}$. However, since there is an exponential growing number of different ways to fit the data as the number of environmental variables increases, not all functions are actually be tested to have the smallest error of all possible functions. To emphasise this, Blum and Rivest (1992) show that fitting even a simple two-layer neural network, with only two hidden nodes and one output, to data is an NP-complete problem. Therefore heuristics are successfully used to find a reasonable, but probably suboptimal, function to explain the function.

The aim of any modelling method is to be applied on unseen data and to maximise the predictive accuracy on these data points. However, if the algorithm works too hard to minimise the error between the training dataset and the function f then it is very likely that the model will be overfitted (DIETTERICH, 1995; HAWKINS, 2004; BABYAK, 2004). Hawkins (2004) defines overfitting as "Overfitting is the use of models or procedures that violate parsimony that is, that include more terms than are necessary or use more complicated approaches than are necessary." Figure 5 illustrates and example of overfitting. Using a complex function (c) to match the data points in the graph will lead to zero error in the training set but most likely will be a poor generalised model. The linear line (a) is a too simple model with large errors for all data points. The quadratic function (b) is a better choice, containing enough to explain the data points without adding any more complexity.

There are many reasons why overfitting is undesirable: (i) using more predictors to build the model, means for future predictions these predictors need to be collected again, (ii) using unneeded features leads to additional unneeded complexity and to worse decisions, (iii) portability and understanding of the model becomes worse when the model is more complex and uses more features, (iv) an overfitted model will predicted a small

Figure 5 – Example of overfitting, three lines fitted to the shown data points. Line (a) is probably too simplistic of a model with large error, line(c) will have the least error, but might in some conditions be overfitted and have a large error for unseen data, line (b) is arguably the best model that has the least error while still being able to generalise. The zoomed out area illustrates extrapolation of the model; since there is no data point with an $x < 0.2$ the algorithm has no way in knowing what way the trend continues, the three lines indicate three different responses.



Source: Author

region G that is very similar to the training data and poorly into unknown regions.

Strategies to limit overfitting are: (i) collect more data, especially for some species there is just not enough data available and no method can fix that simple fact, (ii) reduce the degree of freedom, either by reducing the number of features or performing PCA to combine dependent features into one and remove features that do not explain the presence of the species. However, in the end the best way to make sure that a model is not overfitted is by evaluating the models on unseen separately kept data and pick the model that generalises best. However this is not a guarantee that the model predicts well for all unseen data.

2.2.2 Study region extent

As mentioned in Section 2.2 absence samples are sampled from: the entire study region $\boldsymbol{G}$, background samples; or from parts of where the species is assumed to not exist, pseudo-absence samples. Anderson and Raza (2010) makes the argument that while there is a conceptual difference between the two, in practice the samples will be very much alike. The reason is that for most studies very few locations/pixels will have presences, thus the regions where the species is assumed to not exist and the entire background area are close to equal. For either way of selecting 'absences' it is clear that the study region extent

has influence on the selected samples, and on the resulting model too (HIRZEL et al., 2002; ANDERSON; RAZA, 2010).

The work by Anderson and Raza (2010) suggests that models are better transferable and more representative of the true niche when the extent of region G that is sampled for calibration is small and only covers areas limited by M of the BAM diagram, Figure 3. The reason is simple, because the better the sampled region represents the species' potential distribution the more accurate is the model. For this reason sampling from areas where the species is absent due to biological factors introduce noise and false signals about the absence of the species due to scenopoetic factors. Even though selecting a smaller region will possibly mean that a species model will have to extrapolate into regions that have not been sampled during calibration this does not have to result in a lower quality model than when large extent is sampled (ANDERSON; RAZA, 2010).

When models are calibrated it is important to consider the BAM framework and to take into account other factors such as, the species' dispersal ability, the topographic complexity of the study region and the distributional patterns of related species. In addition, models calibrated with an overly large study region extent are likely to show low transferability (ANDERSON; RAZA, 2010). The transferability of a model is especially important for studies on invasive species and climatic change (ACOSTA, 2015) and model transferability (PETERSON; PAPE ; EATON, 2007) where models will be used to predict presences for unseen regions.

2.2.3 Model Extrapolation and Transferability

Model quality is not only affected by the study region, but also by the environmental conditions that are present there. Extrapolation in E-space and transferability in G-space greatly affects the model performance when predictions are projected on areas outside the area where they were calibrated on. Figure 5 shows in the zoomed in area an example of extrapolation in E-space. The model has been trained with the data points shown in the figure. Beyond the known region, the three lines in the zoomed region, the model can be fitted in many ways without affecting the model performance on the calibration data. These different predictions have a large impact on model projections (CARNEIRO et al., 2016). In general the trends shown in the current data do not need hold for future scenarios and great care should be taken to make these kind of projections (DORMANN, 2007).

Model transferability, or extrapolation in G space, is however a simple consequence of Hutchinson's duality principle, see Figure 1. Where a single area in E space is mappable onto many in areas in G space. Still care needs to be taken when projecting into not sampled regions or future and past times. Torres et al. (2015), Huang and Frimpong (2016), for example, have shown that sound ecological basis for evaluating model

outputs is essential, because the transferability of SDMs is severely limited due to factors such as: the type taxa (i.e., terrestrial or marine), climate- or landscape-change effects, natural barriers and how close a species is to equilibrium with current environmental conditions.

Summarising, a biologist with extensive knowledge about the modelled species always needs to verify model outputs for ecological soundness.

2.2.4 Evaluating model performance and significance

All species distribution models, as all other models, have predictions errors as they are simplified, but are useful views of reality that approximate the true distribution. Errors are introduced in every step (Section 2.1.1) of the model building process. Table 2 by Franklin and Miller (2009) gives an overview of some of those sources of error, the step where they are introduced and articles where more information is found over these errors. The most common error discussed in almost all published articles relating to SDMs is the focus on how well a model predicts, see *Model evaluation - Validity* in the table. Equally important, as discussed in the previous section, but given far less attention are other criteria as ecological realism and credibility. Probably the reason that many articles focus on just model evaluation is that it the easiest error to quantify without deep knowledge about the contemporary species distribution. Another reason for focussing on model evaluation is there are many different modelling methods and there is no consensus on which model should be applied for which application. In addition, modelling algorithms have shown to produce varied results even calibrated on the same data with different methods (THUILLER, 2004). Combined with interdisciplinary data and researchers from different backgrounds the only objective way to compare those models is for now the use of the metrics such as those discussed in this section.

2.2.4.1 Calibration and Evaluation Dataset

To evaluate the quality of SDMs there are two main types of techniques: cross-validation that uses a single dataset; and holdout that separates the dataset into a training, validation and test dataset where the test set is not used for calibrating. Both techniques train and evaluate model performance with different data to prevent overfitting. Cross Validation techniques to evaluate the predictability of a species distribution model are, for example; Akaike Information Criteria (AIC) (AKAIKE, 1974), and jackknife and bootstrap (EFRON, 1979).

One reason to use Cross Validation (CV) is that usually only a very small number of presences are known for a given species and the number of explanatory variables is so high that all samples need to be used to calibrate the model. To obtain an unbiased estimate of the SDM performance k-fold cross-validation is used, where frequently k=10

Table 2 – Criteria for evaluating species distribution models that address different kinds of uncertainty arising during model formulation and calibration,

Modelling step	Criterion	Description	Reference
Conceptual formulation	Precision	Ability to replicate system parameters	(MORRISON; MARCOT; MANNAN, 2006)
	Specification	Does the model address the problem?	(BARRY; ELITH, 2006)
		Does it describe the true relationship?	(BARRY; ELITH, 2006)
	Ecological realism	Is conceptual formulation consistent with ecological theory?	(AUSTIN, 1980; AUSTIN, 2002; BARRY; ELITH, 2006)
Statistical formulation	Realism	Account for relevant variables and relationships	(MORRISON; MARCOT; MANNAN, 2006)
	Verification	Is the model logic correct?	(RYKIEL, 1996)
Model calibration	Calibration	Parameter estimation or model fitting and selection	(RYKIEL, 1996; CHATFIELD, 1995)
Model evaluation	Validity, performance	Capability to produce empirically correct predictions to a degree of accuracy that is acceptable given the intended application of the model	(BARRY; ELITH, 2006; RYKIEL, 1996; MORRISON; MARCOT; MANNAN, 2006)
	Appeal, credibility	Accepted by users, matches user intuition, sufficient degree of belief to justify use for intended application	(MORRISON; MARCOT; MANNAN, 2006; RYKIEL, 1996)

Source: Adapted from Table 9.1, p210 in Franklin and Miller (2009)

(FIELDING; BELL, 1997). In k-fold cross-validation the dataset is divided into k equal sized subsets which reduces the high variance that might occur when data is only divided into two subsets. Models are then build k-times, where every time one of the k-subsets is not used to train the model but only used as a test set. Reporting the average performance of the k-models on the k different subsets gives an estimated performance measure.

CV techniques by themselves do not guarantee that the test set is independent from the training set as samples can still be spatially auto-correlated due to sampling bias. An alternative is to spatially split occurrences into training and testing sets to reduce overfitting to sampling bias (PETERSON, 2011; ARAÚJO; GUISAN, 2006). Spatially structured datasets are formed in multiple ways: by countries, continents, checkerboards (PETERSON; PAPE ; EATON, 2007) or just any polygon or shape drawn in GIS software. While spatially dividing the dataset might introduce a bias in $\boldsymbol{G}$, models that actually perform well even with this bias might be highly desired as it signifies they are transferable.

All things written in Section 2.2.2 about considering the BAM framework when training the model also holds for evaluating it. A SDM should also be validated with samples that capture details of $\boldsymbol{G}$ and $\boldsymbol{E}$ and that are representative for the species.

Table 3 – The confusion matrix, as used for model fitness evaluations

Predicted	Actual	
	Presence	Absence
Presence	True Positive (TP)	False Positive (FP)
Absence	False Negative (FP)	True Negative (TN)

Source: Author

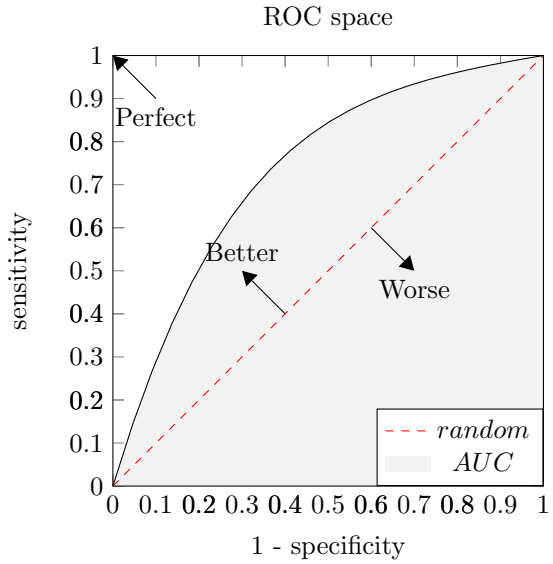
2.2.4.2 Assessing Model Significance

The problems considered in this work are classification problems with two classes; presence (positive) and absence (negative). This means formally that each instance I is mapped to one element of the set $\{p, n\}$ class labels. A SDM is then a classifier that chooses, after training for each instance, the right element from the class label set. For these classifiers, a threshold is usually chosen to divide the continuous values into a binary value to be mapped onto the class label set. Classifiers with discrete outputs and that predict more than two classes will similarly also have to be mapped on a binary set.

For an instance I and a classifier model output $C(I)$, assuming that both are binary sets there are four possible outcomes. The instance and the classifier output can; a) both predict a positive thus signifying a true positive prediction; b) the model prediction is negative while the instance is positive, then it is a false negative; c) the model prediction is negative and the input is negative too, signifying a true negative; or d) the model prediction is positive for a negative instance which is called a false positive. Table 3 show these outcomes in a table form. This two by two confusion matrix or contingency table consists of the above outcomes and its values are the basis for many commonly used model significance metrics.

The main diagonal of the confusion matrix represent the correct predictions made by the classifier, and the numbers on the anti-diagonal represent the errors made by the classifier or the confusion that there is between the two classes. Formulas 2.2 are equations derived from the matrix. The sensitivity, Equation 2.2a, refers to the proportion of instances which have been observed to be present and also have been predicted to be present. Specificity, Equation 2.2b, is the proportion of instances that are truly absent and are also predicted by the classifier to be absent. The accuracy, Equation 2.2c, is the proportion of the total number of instance predictions that are correct. Precision is defined by Equation 2.2d and is the proportion of the predicted positive instances that were correct. See Section 2.2.4.3 for a discussion of Equation 2.2e and 2.2f.

Figure 6 – The Receiver Operating Characteristic curve



Source: Author

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP + FN}} \tag{2.2a}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN + FP}} \tag{2.2b}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{2.2c}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.2d}$$

$$\kappa = \frac{\frac{TP+TN}{TP+FP+FN+TN} - \frac{(TP+FP)(TP+FN)+(FN+TN)(TN+FP)}{(TP+FP+FN+TN)^2}}{1 - \frac{(TP+FP)(TP+FN)+(FN+TN)(TN+FP)}{(TP+FP+FN+TN)^2}} \tag{2.2e}$$

$$\text{TSS} = \text{Sensitivity} + \text{Specificity} - 1 \tag{2.2f}$$

2.2.4.2.1 ROC

Figure 6 shows a Receiver Operating Characteristic (ROC) graph. A ROC graph is a visualisation tool to show and compare the performance of classifiers (FIELDING; BELL, 1997). They have a long history of being used in signal detection theory to visualise the trade off between hit rates and false alarm rates of classifiers. The use of ROC graphs for medical diagnostic testing and a resulting article in Scientific American is what made the technique popular and known to a wider audience (SWETS; DAWES; MONAHAN, 2000; FAWCETT, 2006).

Since the 1980s, ROC analysis has been a popular and become a standard metric in a suit of metrics to evaluate machine learning algorithms and other areas of cost-sensitive

learning algorithms. One of the earliest uses of ROC graphs to compare models and algorithms was by Spackman (1989). He demonstrated that the use of ROC curves is not only useful for medical diagnostics, but also for assessing machine learning algorithms. A further detailed introduction to ROC analysis in research is found in Fawcett (2006).

The graphs are two-dimensional with the sensitivity (Equation 2.2a) plotted on the X-Axis and the false positive rate, which equals to $1 - SPC$ (2.2b). The graph shows the relative tradeoff between predicting true positives and predicting false positives. The ROC space is divided in half by a striped line in Figure 6. All points on this line signify that the classifier performs randomly, meaning it is as often wrong as it is right. Points above the line represent classifiers that predict more often right than wrong, while points below the line represent classifiers that perform worse more than half of the time. The latter type of classifier is easily made useful by inverting their predictions, thus the lower area is often empty.

Other interesting features of the graph are: the point at (0,0) where never a presence is predicted and thus no mistakes are made in predicting a presence; the point (1,1) where everything is predicted to be present while every absence is predicted wrong; and the point (0,1) in the left top corner with only correct predictions and no false ones, the perfect classifier. While a perfect classifier is not realistic to train it represent the goal of any classification algorithm. It is important to note that all discrete binary classifiers produce a single point in ROC space.

To construct a ROC curve of a classifier, first the predictions $C(I)$ are sorted according to the probabilistic scores for each prediction from high to low. Then drawing the graph starts at point (0,0) and all prediction are evaluated. If for a prediction its true class is present then a point is drawn one unit higher then the previous one, if it is negative a point is drawn one unit to the right. At the end of the process the point (1,1) is drawn and all points can be connected with a line.

ROC curves of different classifiers are comparable. If a line dominates all others, meaning it is above the others in the graph, than that classifier is the superior one. However, lines will frequently cross, signifying that a classifier is only superior for some context. Comparing lines when no dominant classifier is present is difficult through looking at the ROC curves. For this reason ROC analysis also defines another metric to measure model classification performance: The area under the ROC curve, or the Area Under Curve (AUC).

2.2.4.2.2 AUC

The area under curve (HANLEY; MCNEIL, 1982) is equal to the area of the grey area in Figure 6. Since the area of the square is always less than one the value of the AUC is in

the range [0,1]. Any useful classification model will have its ROC curve above the random line and therefore an AUC greater than 0.5. The AUC is the same as the probability that the model will rank a randomly chosen present instance higher than a randomly chosen absent instance. The AUC is a popular metric to compare modelling results in ecology articles (THUILLER, 2003; MANEL; WILLIAMS; ORMEROD, 2001; BROTONS et al., 2004).

There are however some disadvantages in using the AUC (LOBO; JIMÉNEZ-VALVERDE; REAL, 2008). Five reasons why it is not recommended: (1) it ignores the predicted probability values as the values are just sorted by the score but then the score is no longer considered; (2) it summarises the test performance over regions in ROC space which would never be considered to be actually used for setting thresholds; (3) while not necessarily so, but conventionally omission and commission errors are weight equally which is not optimal and dependents on the application of the model; (4) the score itself, nor the ROC curve, give information about the spatial distribution of model errors; and most importantly according to Lobo, Jiménez-Valverde and Real (2008) the total extent to which models are carried out highly influences the rate of well-predicted absences and thus the AUC scores.

2.2.4.3 TSS

Allouche, Tsoar and Kadmon (2006) introduced, as an alternative metric, the True Skill Statistic (TSS) , Equation 2.2f, also know as the Hanssen–Kuipers discriminant and the Pierce skill score, to measure species prediction model performance. The TSS score considers both omission, commission errors and a lucky hit due to random guessing. The score ranges from -1 to +1, where +1 indicated perfect predictions and values of zero or less indicate that the predictions are no better than random choice. One reason to use the TSS score is that unlike the AUC and κ (Cohen's Kappa, Equation 2.2e) score, the TSS is not affected by prevalence, the proportion of sampled sites where a species is present (ALLOUCHE; TSOAR; KADMON, 2006). For a literature review of how prevalence can impact AUC and κ scores see Table 1 in Santika (2011). In addition, the TSS score gives better results when used with binary models, such as those produced by the algorithm discussed in this thesis, that show a clear distinction between the two classes, *presence* and *absence* (ALLOUCHE; TSOAR; KADMON, 2006).

There are no official minimal metric scores to state when a classifier is making usable predictions. However, Table 4 shows a general guide for classifying the accuracy of classifiers based on the academic point system and earlier publications. Acceptable scores for all models in this work are *Good, Very Good* and *Excellent*.

Table 4 – Evaluation of metric results. To evaluate the values of the different metrics a general guide for classifying the accuracy of classifiers is the traditional academic point system.

Metric	Fail/Poor	Fair	Good	Very Good	Excellent
TSS	<0.4		0.4-0.75		>0.75
κ	<0.4	0.4-0.55	0.55-0.7	0.7-0.85	>0.85
AUC	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	>0.9

Source: Ranges from Swets (1988) (AUC), Landis and Koch (1977) (KAPPA, κ), Eskildsen et al. (2013) (TSS).

2.3 Conservation planning

This chapter discussed SDM and the relevant background to understand what is actually being modelled. The big question now is why to make these predictions of species distributions in the first place? The most simple reason is that it helps understand the relationship between environment and species. Understanding this relationship helps predicting outcomes and form theories about species distributions and environmental impacts. However, the most useful and essential function is transforming limited point data to range predictions. Not only for current conditions but also for other points in time with likely different climate conditions, to understand the effect on species distributions over time. These range maps help to see where effort is well spent to survey for species populations and helps to understand which areas should be considered for conservation.

One example of the use of SDMs is for conservation planning. Despite the recent proliferation of data driven approaches, algorithms, and software packages for use in systematic conservation planning, most of these techniques share a common purpose. Because in any given region the total amount of land that can be managed for conservation is limited by various social and economic factors, the basic purpose of systematic conservation planning is to establish a system of conservation areas that maximises long-term conservation of biodiversity, subject to socioeconomic constraints.

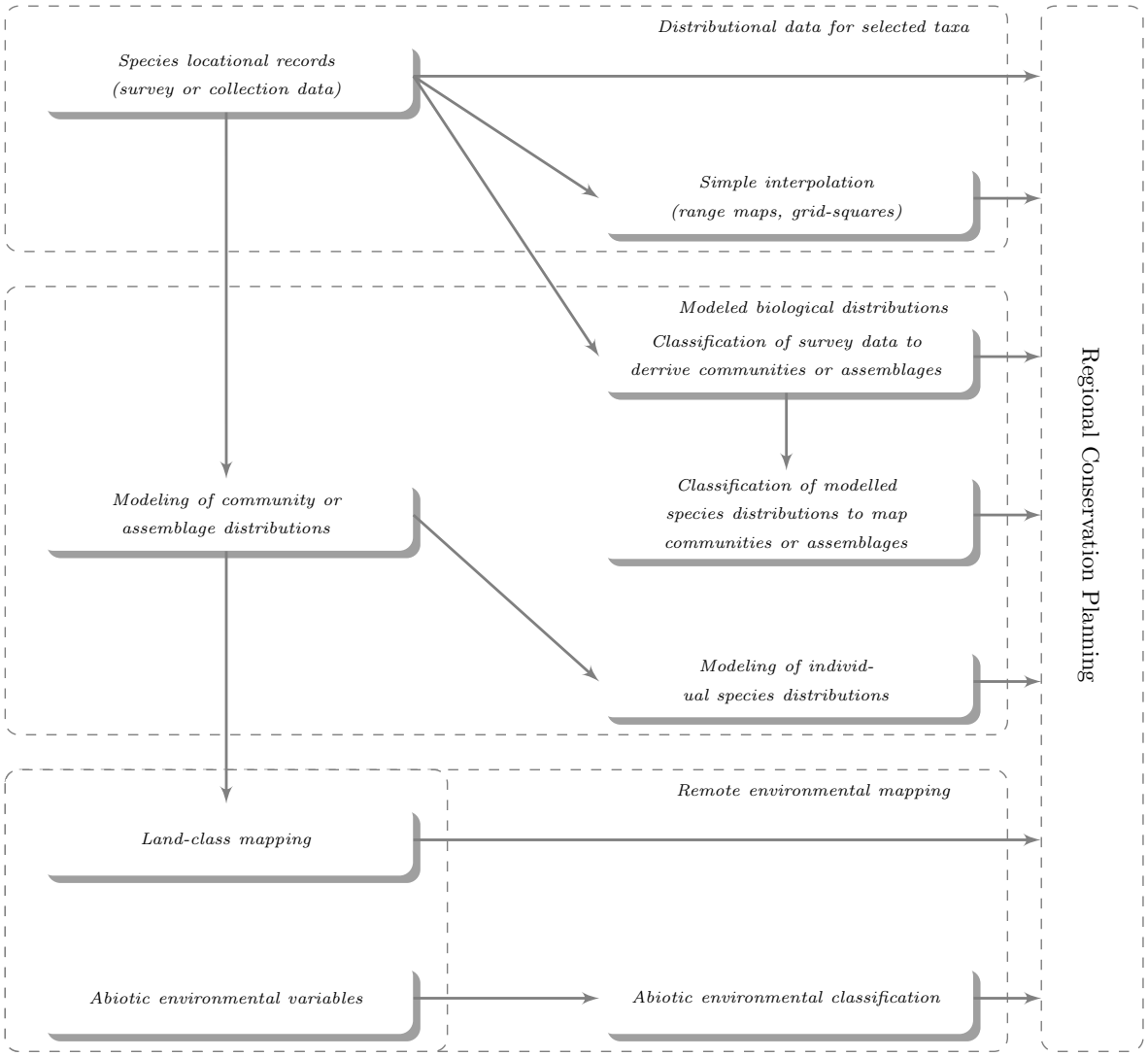
The goal of conservation planning is to protect and maintain indefinitely biological diversity at regional and global scales (MARGULES; PRESSEY, 2000). This involves complex ecological processes and model-based scenarios for possible future circumstances to illustrate consequences of decisions made by governments, the industry and the general public. The use of plausible scenarios has an enormous impact. In the 1980s and 1990s scientific advances and new environmental scenario projections made change in international policies within 24 months (WATSON, 2005; BALMFORD et al., 2005). Due to this complexity and because conservation by itself is not a scientific choice, conservation plans will have global and local challenges, because in any given region the total amount of land used for conservation is constrained by social and economic factors (MURDOCH et

al., 2007).

There is a large choice of tools and algorithms to plan and design representative system of conservation areas. Almost certainly, however, the first essential component of all these tools is information about the spatial distribution of species for the region of interest (FERRIER, 2002). Without this information quantitative assessment of policies and alternative conservation scenarios is impossible as this information is needed to prioritise areas for protected status, to assess threats to those areas and to design the actual conservation areas (FRANKLIN; MILLER, 2009). In practice, however, due to budget and time constraints it is not feasible to have complete knowledge of all species, populations, ecosystems and interspecies relations occurring within the region of interest. For practical purposes, under these limited conditions surrogates, or indicator, species are used within the area to represent the biodiversity of a region as a whole, because single species are easier to study than communities, landscapes or genes, especially when considering that many species still need to be discovered (NOSS, 1990).

Predictive modelling of species distributions is an important tool in conservation biology and climate change research (GUISAN; THUILLER, 2005). Its relationship with species locational records, modelling of community distributions and that of individual species for regional conservation planning is illustrated in Figure 7. While more research is needed to find the best way how SDMs can help to establish priorities and policies (MARSHALL; GLEGG; HOWELL, 2014), it has been used before for non toy-problems. One of the earliest use has been to address protection gaps as part of the U.S. National Gap Analysis Program where spatial interpolations based on land-use patterns and existing conservation reserve network information was partially successfully used to identify gaps in the protection network (SCOTT et al., 1993; PETERSON; KLUZA, 2003). For other applications see the article by Esfandeh, Kaboli and Eslami-Andargoli (2015) that discusses 67 other articles written between 2005-2015 concerning systematic conservation planning.

Figure 7 – Diagram illustrating the function of modelling in general, including species distribution modelling, as a tool for regional conservation planning.



Source: Adapted from Ferrier (2002)

3 Methodology

To defend the thesis the objectives are further broken down into five steps that form the core of the methodology and are discussed in this chapter.

3.1 First step - Fuzzy - Linear Genetic Programming

The objective to design and implement an algorithm that has three aspects: the use of linear genetic programming, the use of fuzzy operators, and parallelisation. Genetic algorithms are typically based on several distinct genotypes. A genotype defines the way of notating operators, variables, and functions. Linear Genetic Programming (LGP) is such a genotype. It is fast due to its limited memory footprint and typical simple instructions that operate on just a handful of registers. Important to note is that *linear* only refers to the structure of the program representation. The program is a sequence of instructions interpreted by a programming language, or directly executed as machine language. LGP in no way limits the method to linearly separable problems. On the contrary, LGP algorithms are able to represent highly-complex and non-linear solutions.

The aim of the algorithm is to give the likelihood of species presence in a geographic regions. Fuzzy rule-based systems have been successfully applied to similar classification and regression problems (BÁRDOSSY; DUCKSTEIN, 1995; CORDON; HERRERA; VILLAR, 2001; ISHIBUCHI et al., 1994). They are known to deal well with noise, impreciseness, uncertain and incomplete information. Perhaps, most importantly, because they describe complex behaviour without the need of precisely defined system models.

The fuzzy system themselves are simple to design and implement, but the identification of the system is a complex undertaking. To use fuzzy operators and rules in the algorithm it is necessary to: (1) define and converse the input and output factors, (2) establish the rules governing the system, (3) define membership functions to determine points in the input space that are mapped to membership values, and (4) define the methods to generate fuzzy rules.

This research aims to do symbolic regression classification by using linear genetic programming to build complex fuzzy rules that can handle the noisy data that inherently rule SDMs. At the same time, the presentation of the model as a program can give answers to the relationship of variables which other classifier approaches can not.

3.2 Second step - Open source software for the use of the system

To benefit the research community the software is released as open source. This has the advantage of giving the full access of the software code and its implementation to others so they can: understand what the solution does, fix possible bugs even when the main author is no longer able to maintain the project, explore and expand with new research methods. Open source is far from new and used in many other work. It is hoped that this project takes advantage from similar benefits as those works. However, the most important reason to release all the code is to promote reproducible research.

SDM and the processing of species data is a challenging area of scientific exploration and involves researchers from various backgrounds. It is therefore essential to have a good methodology that uses established techniques and methods. For this reason, the open source R language and environment for analysis, together with the vast number of available packages is a popular choice. Therefore, R will be adopted as the target language for development and C++ when fast execution is needed.

BIOMOD is *the* software package to use within the R environment and provides a distribution modelling methodology. The BIOMOD methodology supports: the preparation of species data, the generation of models and ensembles, the evaluation of those models and it supports the projection of distributions onto geographic maps. Many popular algorithms for SDM are available, therefore this work will be made compatible with a locally modified version of BIOMOD to evaluate the difference in obtained model quality.

To execute this second step, algorithm defined in the first step is implemented in C++ and tested by using a command line interface. Then, if the solution works to satisfaction and to prepare for the next step, changes are made in the BIOMOD source code to enable the use of the algorithm inside R and with the BIOMOD package.

3.3 Third step - Case studies

While a case study won't proof that the algorithm is usable in every situation, since only a limited number of cases are studied, they will give an insight if the thesis is defendable for those cases and raise questions for future research. In addition, the algorithm is bound to perform inferior to some other ones for certain cases if the no-free-lunch theorem (WOLPERT; MACREADY, 1997) is taken into consideration, still it is of interest how near the algorithm performance is to the best one.

There are three datasets used in the analysis of the algorithm; climate data, hydrologically correct DEM, and species data. Similar to the work of Peterson, Pape and Eaton (2007), the abiotic data set is formed by joining the nineteen bioclimatic variables of the WorldClim data set (HIJMANS et al., 2005) with four layers (elevation, slope, aspect

and the compound topographic index) from the digital elevation model Hydro1k (U.S. Geological Survey, 2000) data set, see Table 1. To reproduce the data set PCA is used to reduce the amount of possibly correlated variables and thus the search space. The eleven most significant components together account for 97,9% of the variance. A difference with Peterson, Pape and Eaton (2007) is that the Hydro1k data set is not resampled to 10' resolution, but only projected to WGS84 and aligned with the 30 arc-seconds WorlClim data set. The reason is that occurrence point data (*Longi* and *Lati* fields) from North American Breeding Bird Survey (BBS) are used and not the route paths shape files. Thus avoiding the effect of resampling as, for example, slope calculations are resolution dependent.

The species data originate from the BBS (SAUER; J. E. Hines; J. Fallon, 2001). These data were obtained in digital form and contains a database with presence/absences for over 400 bird species measured over thousands of routes for more than five decades. From this database stable population data of the species *Zenaida macroura*, the mourning dove, were obtained. Presences were defined as the routes where the species was spotted during eight, not necessarily consecutive, years during the period 1991-2000. Absences where defined by not having a single specimen spotted during that same period. All other routes were not used for this study. This resulted in a data set with 1155 true absence points and 1003 true presence points of *Zenaida macroura* in North America (westlimit=-169.5; southlimit=24.5; eastlimit=-52.0; northlimit=76.5; projection=WGS84).

3.4 Fourth step - Parallelisation

If the algorithm functions satisfactory then, and only then, it is worth to pay attention to parallelisation. There are three main reasons to focus on parallelisation. First, and foremost, parallelisation enables to cover heuristically a larger search space in the same amount of time. Given that the design of conservation area networks and the definition of SDMs are NP-hard problems (SARKAR et al., 2006) and heuristics are used a larger search space can be covered possibly resulting in models with a higher accuracy.

Second, decision support systems are used in real-time negotiations during which stakeholders need rapid information about the implication of policies in alternative ways (SARKAR et al., 2006). Therefore, it is important to have fast algorithms to evaluate the options under conditions of highly variable biodiversity data and the large number of tasks that need to be examined. In addition, Thuiller et al. (2009), Terribile, Diniz-Filho and Marco (2010), Parviainen et al. (2009) recommend to combine multiple SDM to obtain higher quality predictions and better assessments of species richness trends and hotspot patterns. This significantly increases the amount of time spend on building models.

Finally,McBride et al. (2010), Foerster et al. (2010), Larocque et al. (2011), Giuliani et al. (2011), Johannes et al. (2009), Yang et al. (2011), Luong, Talbi and Melab (2010) state

that the further development of decision support tools, and associated algorithms, will depend on the availability of computational power for solving complex applications. According to Christophe, Michel and Inglada (2011) the increase of this computational power for the coming years goes through a parallel approach and therefore parallelisation should be an integral part of algorithm design.

The following solutions are used to parallelise and develop the software: OpenMP to optimise for a single multi-core system; OpenMPI to enable clustering; Starcluster to define and initialise an Amazon EC2 cluster; GCC to compile the algorithm and CMake to link and build the executable.

3.5 Fifth step - Consider research synthesis

In order to solve climate change problems and change policies that affect biodiversity it is necessary to create analytical tools that can do more than just plotting range maps of species. However, those tools are likely to use generated SDM as a building block. For this reason it is extremely important to consider the sharing of the models to enable research synthesis. While not providing an answer as the matter is too complicated to be part of this research, an initial conceptual model is discussed in Chapter 6. In retrospect this matter might be more important than more accurate models.

4 Genetic Programming

Contents

Genetic Algorithms (GA) and Genetic Programming (GP) are based on the biological analogy of evolution, genetics and on the principle of *survival of the fittest* (SPENCER, 1864–1867) as described by Darwin (1871) as natural selection. In GA a population of strings is bred selectively to create new generations and offspring with better characteristics, where the characteristics are often defined by a monotonic function that minimises a metric error, e.g., the Root Mean Square Error (Equation 4.1), where n is the number of samples, y_i is known output and $\bar{y}_i$ is the predicted output. In GP populations are similar, except that the individuals do not represent the solution themselves, but a program that computes the required solutions. GP was proposed, invented, and made popular by (KOZA, 1992) in the nineties and founded on the research done by Holland (1992). For both methods, depending on the genotype, there is typically a structure, such as a string or tree, that acts like a biological chromosome and bind the individual parts together. These structure usually belong to a single individual, although variations on this exist, where the individual represents a point in search space and possibly a solution to the problem at hand. Using the digital analogies of genetic mutation and crossover, the chromosomes are recombined to breed new and hopefully fitter individuals. Breeding is an

iterative generational process; improving solutions from one generation to the next. This iterative process continues until a stopping criteria has been met. Typical stopping criteria are: an individual is found with the required fitness, the population has reached a certain number of generations, a certain amount of time has passed, or no considerable better individual has been found and the fitness function has little changed for some generations.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{4.1}$$

Even though the idea of GA and GP is simple to understand and implement, the theoretical background as to why and how they work is far more complicated and is still an open research question. The theoretical background is important to understand the process of generation and to know for which kind of problem they are suitable. In the mid-seventies John Holland developed his schema theorem (HOLLAND, 1975). A schema by his definition is a string of symbols, where the symbols are elements of an alphabet $0, 1, \#$, where the symbol # signifies a *don't care* and the symbols $0, 1$ the respective numbers of the base-2 numeral system. Thus, the schema *11x0x* represents four string instances: *11000*, *11001* ,*11100*, and *11101*. There are several properties defined for schemas: the order (H), the number of non-# symbols; and the defining length $\mathscr{L}(H)$, the distance between the two furthest away non-# symbols in the string. Using these definitions the schema theorem predicts the variation in the number of particular strings over time in a population of strings. The theorem is formulated as follows (POLI; LANGDON, 1998):

$$E[m(H,t+1)] \geq m(H,t) \cdot \underbrace{\frac{f(H,t)}{f(\bar{t})}}_{\text{Selection}} \cdot \underbrace{(1-p_m)^{\mathscr{O}(H)}}_{\text{Mutation}} \cdot \underbrace{[1 - p_c \frac{\mathscr{L}(H)}{N-1}(1 - \frac{m(H,t)f(H,t)}{M\bar{f}(t)}]}_{\text{Crossover}} \tag{4.2}$$

In Equation 4.2 $m(H,t)$ is the number of strings matching the schema H at time *t*, $f(H,t)$ is the fitness of those strings in the population, M the total number of strings in the populations, $f(t)$ the mean of all those strings, p_m the probability that a symbol mutates, p_c the probability on crossover, N the number of bits in "the length" of the strings, and finally $E[m(H,t+1)]$ is the expected new number of of strings matching schema H at the next time interval. The meaning of this equation is that over time the schema H with above average fitness and shorter length will become more dominant in the population. One might notice that both mutation and crossover are destructive in the equation. However, it is commonly believed that crossover is actually the source of the power of genetic algorithms (HOLLAND, 1975; HOLLAND, 1992; MITCHELL, 1998, c1996; LUKE; SPECTOR, 1998), because the operator combines good performing schemas in new equally, or better, per-

forming schemas. This assumption is known as the *Building Block Hypothesis* (GOLDBERG; HOLLAND, 1988).

Holland's theorem combined with Price's Covariance and Selection Theorem, that describes how a gene change in frequency over time, explicitly shows how changes in different macroscopic properties of population in a genetic algorithm can be derived by using the microscopic dynamics of the GA and GP combined with an appropriate fitness function (ALTENBERG, 1995) leads to the automatic programming of working computer programs that are able to solve a variety of problems, including the accurate prediction of species distributions (see Chapter 5).

The benefits of GA and GP are many fold, they:

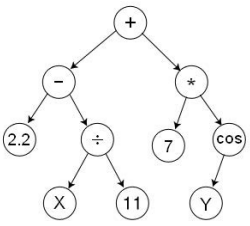
- make no assumptions about the underlaying problem nor is there a need for experts to define *how* the problem should be solved. This is important as there often is not a mathematical model of the objective function available for many real world problems.
- allow expert information about the system to be incorporated to focus the search and limit the exploration of the search space, however, no such information is required to find suitable solutions.
- function inherently and embarrassingly parallel.
- have a good chance of not staying stuck in local optimal solutions as the search space is explored in parallel, unlike traditional methods which typically search from a single point in search space.
- are not deterministic, answers tend to improve (slowly) over time.
- work even when analytical approaches do not work, for example, due to the data or when the objective function itself is not smooth.

4.1 Genotype

The search space that can be covered to find possible solutions is limited and defined by the genotype of the genetic algorithm. The genotype defines the data structure that is operated on by the reproduction, crossover and mutation operators of Equation 4.2. For GP the genotype tends to directly translate to an executable program, although there are a few cases for which this is not true (KELLER; BANZHAF, 1996). The three frequently used genotypes are: tree-based (KOZA, 1996), Figure 8a; linear (OLTEAN et al., 2009), Figure 8b; and Cartesian (MILLER, 2011), Figure 8c.

Each genotype has its benefits and drawbacks (WILSON; BANZHAF, 2008). These drawbacks come frequently to light during the implementation of the algorithm. For example, while trees are easily mutated when just considering nodes from a high level point of view, it is more computation intensive and requires syntax validation of the tree when

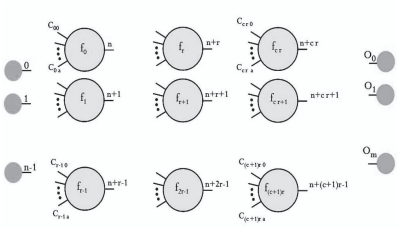
Figure 8 – Genetic programming genotypes



(a) Tree

```
v[0] = v[5] + 73;
v[7] = v[3] – 59;
if (v[1] > 0)
if (v[5] > 21)
v[4] = v[2] * v[1];
v[2] = v[5] + v[4];
```

(b) Linear



(c) Cartesian

Source: Author

implemented. For example, a node with two children mutated to a node with three children needs to have a third child to be executable. Besides these three main types there are many variations, such as, Multi Expression Programming (OLTEAN; GROSAN, 2003) where not only the fitness of the root node is considered, but also the fitness of the best sub-tree.

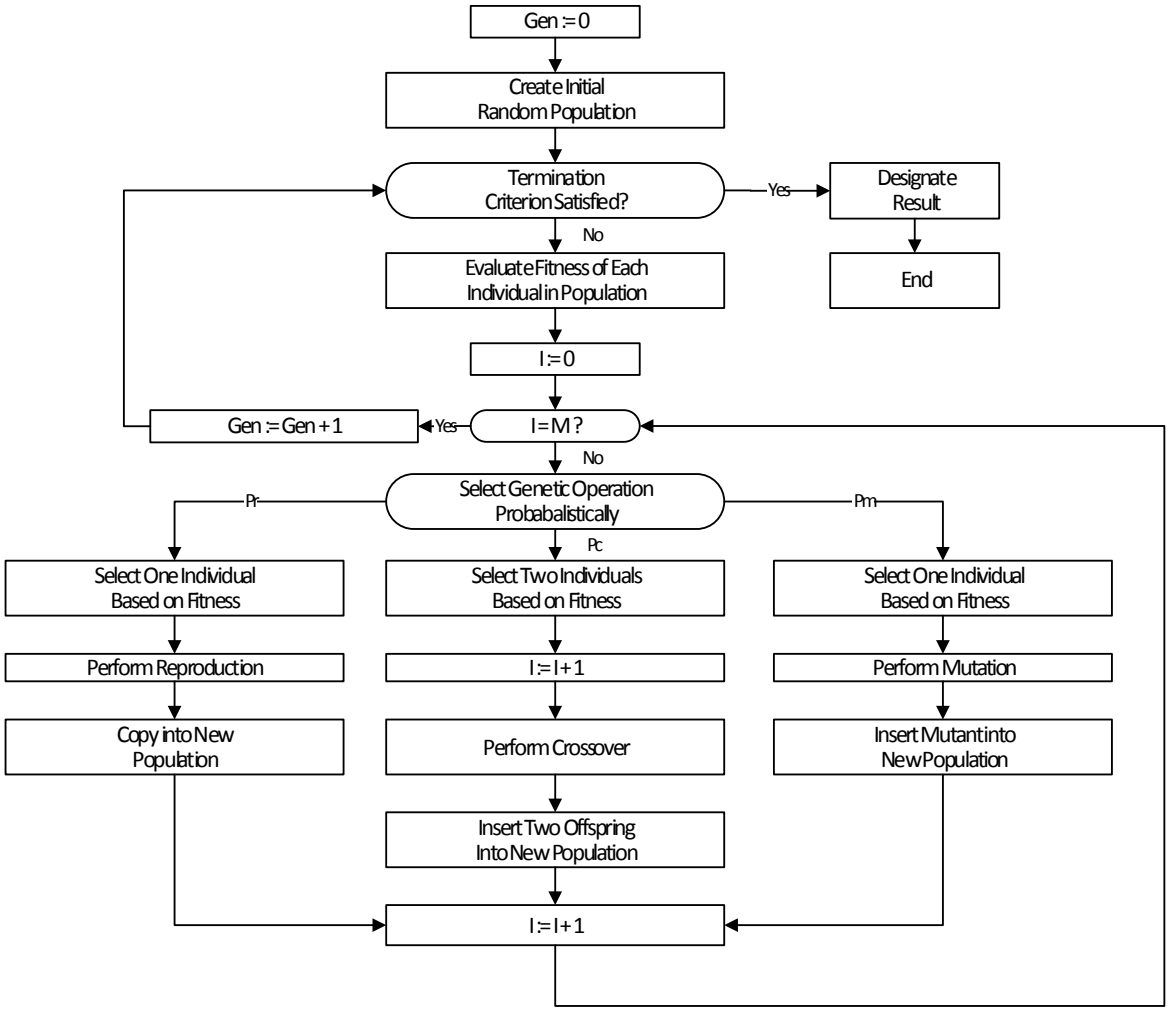
In GP, individuals are generated according to a certain genotype and form a population of typically small computer programs. Using the principles of Holland's Theorem this population is then evolved and improved from one generation to another by using a genetic algorithm in combination with a suitable fitness function. The programs can vary in size and structure and are expressed by software building blocks called genes. Each gene represents a function that is typically a simple unary or binary operator out of a preselected function set, such as $\mathcal{F} \longrightarrow \{+, -, \div, \times, !\}$. Although more complex functions composed out of a set of genes or even other individuals are possible (KOZA, 1994).

The algorithm implemented in this work uses the LGP (OLTEAN et al., 2009) genotype to classify the absences and presences as discussed in Chapter 2. The LGP algorithm is used to discover symbolic expressions that accurately characterise the non-linear relationships that are presumably buried in feature rich datasets of scenopoetic data. Discovering this relationship enables the classification of new regions that are not defined in the training dataset. No assumptions are made about the models for LGP to work. To start, initial expressions are randomly created and evolved to find the function, thus symbolic regression, that most accurately describes the relationship(s) between the training data and the supervisory signal, that is the presence and absence data. Using the genetic algorithm every generation new and possibly better performing functions are found that describe this relationship.

4.2 Evolution

The field of evolutionary algorithms is a relative new area that studies non-deterministic search algorithms that are loosely based on aspects of Darwin's theory of evolution by

Figure 9 – The methodology to evolve populations in genetic programming



Source: Adapted from Koza (1994)

natural selection. John Holland proposed evolutionary algorithms and showed that an evolutionary process could be applied to artificial systems (HOLLAND, 1992). This led to evolving program code, so called genetic programming, that became widely known after the publication of a series of books by John Koza (KOZA, 1994). Genetic Programming is ultimately, as Koza described it, a method for getting a computer to solve a problem by telling it what needs to be done instead of how to do it. The methodology to evolve populations that forms the basis for this work is depicted in Figure 9.

Genetic Algorithms work by creating populations consisting of individuals. Each individual represents a possible solution for the problem at hand. At the start all individuals are initialised randomly, meaning that each gene of an individual is set to a random element operator, variable or function of a programming language. The individuals are then evaluated and compared to stopping criteria; a typical stopping criteria is a pre-set acceptable error rate. If the criteria are met the algorithm is finished and the best individual is a solution to the problem, if not met the population evolves and a next generation

is created.

There are a number of ways to evolve individuals within a population. First, to the left of the diagram, there is elitism, a method with which the best individual of the population is copied to the next-generation of the population to maintain the current best found solution. Second, there is crossover in which two individuals are selected to mate. One strategy frequently used is to create two offspring where each offspring contains a part of each parent. This way new individuals are created that are more likely a better solution than a completely random solution. Last, there is mutation, in which a single individual is selected and then one, or more, genes are modified to represent other genes.

After creating the new generation the above described process repeats itself until a best solution is found. With genetic programming it is not guaranteed that one finds the best solution, just that a solution is found that performed better than the other solutions of that particular population. Therefore performance of a genetic algorithm cannot be measured by the best-found solution in a single run, but measured by how often and how quick, in number of generations, it finds those solutions.

Therefore, CV procedures are an integral part of the methodology to evaluate the risk of reporting a performance that does not represent the performance that is obtained on average (ARLOT; CELISSE, 2010) and is useful when scarcity of field data is an underlying problem. A ten-fold cross-validation was applied to the classifiers used in this work. In a ten-fold cross-validation the data is divided into ten complementary subsets and ten rounds of analysis are performed. In each round, one subset is chosen as the test set, used to measure the performance of the model on events not represented in the data set, and the other nine are used as the training set to build the classifier.

4.3 Fuzzy

SDMs are created from mainly two sources of data: occurrence records and environmental indicators. They predict for each point in a given geographical area typically a value in the range $[0, 1]$ that is often interpreted as the probability that a species potentially occurs at a region, or the relative likelihood in case no true absence data is used. Often these continuous values are converted to a binary *presence* or *absence* for ease of interpretation and for measures to evaluate the model. Conversion of the value to a binary one is conventionally done by setting a threshold, where values above the threshold indicate a *presence* and values below an *absence*.

However, what needs to be realised is that a *presence* and an *absence* are not bivalent conditions. They are perceptions. The following examples illustrate this: *a*) an area is suitable, but has never been visited due to dispersal limitations. Thus the *absence* here does not signify that the abiotic factors are not part of the fundamental niche; *b*) an area is

unsuitable, but a specie is detected resulting in an *occurrence* record. However, the population is in (rapid) decline as the abiotic factors are not part of the fundamental niche of the specie; *c*) samples can be collected over a large time span and mismatch temporally with environmental conditions. A once *occurence* does not signify that the area is still suitable; and *d*) an *absence* or *presence* can have different meanings across spatial scales.

Consequently, model outputs should not be interpreted as probabilities, but as imprecise perceptions with a degree of truthfulness. Fuzzy-set theory, introduced by (ZADEH, 1965), works on problems with a continuum of grades membership and perceptions such as those just discussed. Therefore, genetic programming can benefit from fuzzy-set theory operators.

There exist a large number of algorithms for creating SDMs. Elith et al. (2006) have done an extensive study to model comparison of sixteen SDM modelling methods. The only evolutionary computing based model of the sixteen is the Genetic Algorithm for Rule-set Production Stockwell and Peters (1999). GARP is an algorithm that uses presence and background absence points to generate IF-THEN rules and uses a genetic algorithm to select the methods, i.e. logistics regression, bioclimatic rules and range, to use for a particular rule. The here proposed algorithm works in a similar fashion, but uses linear genetic programming (BRAMEIER; BANZHAF, 2001) and fuzzy operators to generate programs that are more flexible than IF-THEN rules as no predefined structure is assumed.

Finding the characteristic functions of the fuzzy sets that indicate the membership of the variables to *absence* and *presence* is the objective. These fuzzy sets are defined by the program described by the individual and optimised by applying a genetic algorithm. As a result the fuzzy rules will classify species occurrence based on global climate datasets.

4.4 The algorithm

The fuzzy genetic modelling algorithm proposed in this thesis and its operators are discussed in this section and where applicable using SDM as an example. Each individual contains a block of memory and a block of instructions. Both blocks contain an equal number of items: floats for the memory block and a struct for the instruction, with an opcode (int), operand (int) and four constants (float). Each generation consists out of a new list of individuals on which selection, cross-over and mutation are applied to build the next generation, and so on.

4.4.1 Initialisation of model population

Suppose there are n abiotic factors denoted as $e_1, e_2..e_n \in E$ which affect the distribution of the species. Then the program population is initialised by generating x individuals randomly with instruction set I, see Table 5, resulting in group of individuals where each

represent a particular solution to a supervised learning problem. The maximum program length and populations size x are restricted and set before the run of the algorithm and determine the complexity of the models obtained.

The instructions implemented in this solution, Table 5, consists out of the standard fuzzy operators (ZADEH, 1965), two binary operators: AND (the intersection, the minimum of both values), and OR (the union, the maximum of both values), and one unary operator: NOT (the complement, one minus the value). The exact implementation is shown in the Explanation column.

To obtain the random numbers that are used to initialise the individuals, defining the operators used in the program and the membership functions, and to pick the individuals during the selection phase a pseudo random number generator is used. For the implementation of the here discussed algorithm the *Mersenne Twister* (MATSUMOTO; NISHIMURA, 1998) algorithm is used. *Mersenne Twister* has passed many random statistical tests, including the die-hard test (MARSAGLIA, 1995) and the load test (LEEB; WEGENKITTL, 1997). The 32 bits implementation that is used is implemented in the C++11 *random* library and accessible through *std::mersenne_twister_engine*. Meaning that it can generate a lot of random data in a short amount of time.

Each single individual completely defines all the membership functions for all selected attributes that are fuzzified and used to build the complex rules. These definitions are randomly generated during the initialisation of the population, and will be adjusted between generations. It is important to note that all attributes are expressed as floats, and so categorical attributes have to be converted to numerical values, by converting a *presence* to 1.0 and an *absence* as 0.0.

The so called 'Pittsburgh'-style method (SMITH, 1980) is followed to form the fuzzy rule sets. In this approach each chromosome/individual represents a set of rules and not a single rule ('Michigan'-style method; (HOLLAND; REITMAN, 1977)). The main difference is that not the rules themselves are evaluated and scored, but only the set of the rules as a whole. In this way, not only the rules are scored, but also how they are glued and used together.

4.4.2 Fitness evaluation of models

Given that there are n factors, such as longitude, rainfall, altitude, which are assumed to explain the realised niche, see Chapter 2, then the goal of the algorithm is to minimise the error between the model prediction and the training *presence* and *absence* data. However, not all error is considered equal. Consider the confusion matrix in Table 3, elements True Positive (TP) and True Negative (TN) denote correctly predicted samples, whereas False Negative (FN) and False Positive (FP) denote the erroneously predicted samples. With this

Table 5 – Instruction set

Instruction	Explanation
PUSH	push a
AND	pop b, if (b > a) a := b
OR	pop b, if (b < a) a := b
LOAD	a := load(ram[pc])
TRAPMF	a := trapmf(data[op],ram[pc],ram[pc+1],ram[pc+2],ram[pc+3])
NOT	a := 1 - a
IF	if (a > ram[pc]) a := pop

Source: Author

matrix the following metrics are determined: the sensitivity (2.2a), the proportion of positives correctly identified; and the specificity (2.2b), the proportion of negatives correctly identified. For this algorithm implementation, first the sensitivity of the model is considered and only when the models have equal values then also the specificity. The reason for this is that it is better to have models that predict zero (*absence*) when not sure, so that multiple models can be summed together to form a single higher quality model.

4.4.3 Caching of fitness values

The computation of the fitness of the individuals during each iteration is the bottleneck in GP. This is in accordance with the results publish by Santos and Santos Jr (2000). The proposed solution is to cache the fitness values of each individual to significantly reduce run times. The reason caching is effective is because in GA and GP much of the gene/instruction sequence is preserved after cross-over and mutation as they only impact fractions of the individual. Another reason is that typically as the algorithm progresses from generations to generation more and more solutions will be similar. This is even more true because of semantic- and structural 'introns' (BRAMEIER; BANZHAF, 2007).

Introns are sections of instructions that when removed do not alter the outcome. For example, in the mathematical expression $a + b - b$ there is no change in outcome if the instructions $+b$ followed by $-b$ are removed. Now it is possible that during crossover that mathematical expression has changed to, for example, $a + 0$ or $a + c - c$ or perhaps to $a + 1 - 1$. In all those case the result will still be just a.

To not evaluate the 'same' individual twice, all introns are first removed before the individual's fitness is calculated. Then the hash of the simplified individual is determined for each present in that generation. The hash of the individual is calculated by combining the hash of all the instructions, operands, and constants defined in the individual with the *boost::hash_combine* function (section 6.3 of the C++ Standard Library Technical Report (ISO/IEC, 2007-11-15)). If the hash value is present in a look up table then it's matching value

is assumed to be the fitness of the individual. If not found, the fitness is calculated and then stored in that same table for future look ups. This significantly increases the speed of the algorithm, in the study of (SANTOS; Santos Jr, 2000) a speed up of 58% was attained.

4.4.4 Fuzzy parameter optimisation of models

Each generation the fuzzy membership functions are evaluated. Depending on the function will either be optimised or left in its current state. The fuzzyfication function trapmf, see Equation 4.3 is chosen because it allows the definition of an optimum region and configurable slopes to indicate less suitable areas.

$$TRAPMF(x; a, b, c, d) = \begin{cases} 0 & x \leq a, \\ \dfrac{x-a}{b-a} & a \leq x \leq b, \\ 1 & b \leq x \leq c, \\ \dfrac{d-x}{d-c} & c \leq x \leq d, \\ 0 & d \leq x. \end{cases} \tag{4.3}$$

4.4.5 Demes

The term *deme* was first introduced in the field of biology by Gilmour and Gregor (1939) to replace the then popular terms *local intrabreeding populations* and the cumbersome *populations occupying a specific ecological habitat*. The benefit of this idea has been introduced by Wright (1932) in his shifting balance theory and is now later also applied to GA and GP to indicate a separate population where selection, mutation, and crossover are only applied within the local separate population. The deme, or island (WRIGHT, 1932), population can be seen as spatial distinctive of other populations and which as a result independently. The implementation discussed in Section 4.6 uses the *multiple-deme coarse grained* method, meaning that one can generate many separate populations that occasionally exchange individuals through migration. This approach has been shown to maintain diversity over a longer period of time, thus mitigating against premature convergence, even with smaller populations and results in higher quality solutions (POLI; PAGE, 2000; LANGDON, 1995; SKOLICKI; JONG, 2005; WRIGHT, 1932).

4.4.6 Recombination of model population

After the fitness is evaluated the individuals that will make it to the next generation are selected. These individuals are then recombined through crossover or mutation or copied unmodified to the next generation using an elitism strategy. The individuals are laid out in a rectangular grid and offspring is created by using tournament selection with sample size five, thus the cell and all its direct neighbour. This means that every time two

individuals, in case of crossover, are selected of the cell and its neighbours based on the best fitness scores and a probability. The anisotropy degree (SIMONCINI et al., 2006) $\alpha \in [-1, 1]$ used is configurable and automatically adjustable during the run and determines the probability of the centre selection and the take over time of the best individual. This way the algorithm can tune the selective pressure and first focus on exploration, while in later generations focus on exploitation.

The delicate balance between these recombination operators is summed up by Wright (1932) by stating that evolution depends on a delicate balance between these operators. Gene mutation is necessary, but if there is too much there won't be evolution just the creation of random individuals. There must be selection, but again if too much there won't be any variability and no optimum solution will be found. Inbreeding within the demes is advantageous, but only inbreeding and no crossbreeding leads again not to optimum solutions. The optimum values are often application dependent and thus the parameters for these operators are set and experimented with by the user when creating the models.

4.4.6.1 Crossover

Crossover is performed by choosing two random points in the chromosome, both parents will have the same length, then the code between those to points will be exchanged. Unlike with a tree structure, there is no need to worry in LGP about the validity or depth of the offspring after crossover. However, due to the nature of the stack machine implementation it is needed to safeguard against popping more data from the stack than is available. This is done by guaranteeing that the same number of of pop and push instructions are in the exchanged part of the chromosome.

4.4.6.2 Mutation

Mutation is performed point wise on a single parent where every point has a (low) probability of being mutated. The resulting program is the produced offspring. The idea of mutation is to source and maintain a degree of variability and diversity within the population. To maintain valid programs push instructions are replaced with other push instructions, and pop instructions with other instructions that pop.

4.4.6.3 Elitism

Elitism is a method used to conserve one or more of the most fit individuals from one generation to the next. This means that those individuals are copied without modification to the next generation. In this work elitism is implemented as the best individual being copied over a random element after mutation and crossover already generated the next generation. This unlike, for example, in the *preselection scheme* (MAHFOUD, 1992; CAV-

ICCHIO, 1970) where elitism is implemented by allowing offspring only to replace their weakest parent when that child is fitter. Elitism is a guaranteed to make sure that the best individuals discovered are not discarded and are usable in future generations for improvements by either crossover or mutations. The use of the elitist method is generally adopted in GA and GP to let the algorithm focus on a global optimum solution, while still allowing the algorithm to diversify and find multiple other optimal solutions. A disadvantage using elitism is that in the later generations there will be less diversity as many closely related, and likely redundant, individuals will make up a large fraction of the population. A study by Villalobos-Arias, Coello, Carlos A Coello and Hernández-Lerma (2006) shows that for some situations, elitism is required to have an algorithm converge.

4.4.6.4 Migration

There are several factors that define migration: the size of migration, the frequency or interval that individuals move from one deme to another, the topology or configuration of the demes with respect to each other, and the migration policy. Skolicki and Jong (2005) show that small migration sizes (far) less than 10% should be used although the size has far less impact than the frequency has, at least for high frequencies. Frequencies are suggested to be around every 5-10 generations.There are many topologies imaginable, for example, a simple ring topology, a 2-D toroidal grid as is often used in cellular evolutionary algorithms (ALBA; DORRONSORO, 2008), or a full mesh with all nodes connected to each other. The migration policy dictates which individual of the source sub population is selected and which individual of the target population is replaced. Several strategies are imaginable: randomly selecting and replacing; selecting the best and replacing a random individual, or the one used in the algorithm implemented here; selecting the best and replacing a random individual. The migration policy where both source and target individual are selected based on fitness can significantly increase the selection pressure and result in faster convergence (CRUZ; TESHIMA; CETRA, 2013).

4.4.7 Projection

The implementation uses the Geospatial Data Abstraction Library (GDAL) open source library with support for raster and various geospatial data formats (GDAL Development Team, 2016). This library is commonly used, also in R, for reading, writing and manipulating raster data. Rasters are used to represent climate layers where each cell represents a 'small' area on earth, e.g., 1 square kilometer. The GeoTIFF file format (RITTER; RUTH, 1997) is used as intermediate format between the rasters supplied by the user in R and the algorithm. In this format each pixel value represents one geographic area. Multiple bands can be combined into one file, each representing the same geographic area, but a different environmental factor. To make projections a specific cell location across all layers is read,

resulting in a vector of values that will be fed into the model. The output, a single float, will be written back to a single band GeoTIFF raster file, where each cell represents the cell's suitability for the species. This GeoTIFF file is read back into R and can be plotted as is done in Section 5.1.3.

4.5 The Command Line Interface

A Command Line Interface (CLI) forms the boundary between two entities, e.g. a human and a program, and provides a bidirectional but solely textual interaction with each other. This section discusses the CLI implemented in a console application and its exposure to R and the BIOMOD packages.

4.5.1 Console interface

The algorithm is implemented as a console application that interacts with the user, or BIOMOD, through a command-line interface, where the commands are issued to the program in the form of options supplied to the algorithm when the program is started. Listing 1 shows all arguments than can be supplied to the application.

Listing 1 Command Line Interface

```
fuzzy.exe [--print] [-d <int>] [-g <int>] [-i <int>] [-p <int>] [--shard <int>] [-b <int>] [-r <int>]
          [--output-directory <string>] [--project-out-csv <string>] [--project-out <string>]
          [--project-in-csv <string>] [--project-in <string>] [-l <string>] [-o <string>] [-v <string>]
          [--background-samples <string>] [-s <string>] [--version] [-h]

Where:
   --print                                                          Print the rules of the model
   -d <int>,  --dependent <int>                                     The dependent to optimize for
   -g <int>,  --generation <int>                                    Generations
   -i <int>,  --instruction <int>                                   Instructions per individual
   -p <int>,  --population <int>                                    Population size
   --demes <int>                                                    How many demes to use
   -b <int>,  --boost <int>                                         How many boosts to do
   -r <int>,  --runs <int>                                          How many runs to do
   --output-directory <string>                                      Directory to place models and projections
   --project-out-csv <string>                                       Filename of CSV out
   --project-out <string>                                           File name of geotiff out
   --project-in-csv <string>  (accepted multiple times)             File name(s) of CSV in
   --project-in <string>                                            File name of geotiff in
   -l <string>,  --load <string>                                    Filename of model to load
   -o <string>,  --output <string>                                  Filename to write model to
   -v <string>,  --validation <string>  (accepted multiple times)   Validation file(s)
   --background-samples <string>                                    Background Samples file
   -s <string>,  --samples <string>                                 Samples file
   --version                                                        Displays version information and exits.
   -h,  --help                                                      Displays usage information and exits.
```

4.5.2 R Interface

Changes have been made to a local copy of the BIOMOD2 package (version 3.1-26), permissible with the GPLv2 (Free Software Foundation, June 1991) license under which the package is

released. The complete tutorial by Georges and Thuiller (2013) can be followed. The only difference is that in step 3 *Computing the models*, the here discussed algorithm needs to be added in the list of models. The line `models = c('SRE','CTA','RF','MARS','FDA'),` needs to be replaced with `models = c('SRE','CTA','RF','MARS','FDA','THESIS'),`.

Depending on the species under study the options in step 2 for this algorithm can be modified. Similarly as is possible for the other algorithms present in BIOMOD. The options supported in the R interface are shown in Listing 2. The most likely parameter that a user would like to change is the path of the executable that is obtained after compiling the algorithm's implementation. Other options that are modifiable are: (1) the number of runs, i.e. the number of models that will be generated and then combined by adding the outputs of each model and then scaling the result back between zero and one by dividing the result by the number of runs to obtain an averaged model, (2) to skip model generation and load a particular (set) of models by providing the location to them in 'use_model', (3) the number of generations, specifying the maximum number of iterations before the algorithm halts, (4) the maximum number of instructions each individual can have, (5) the number of each individuals present in each generation, or in other words the population size, (6) the number of boosts to execute (FREUND; SCHAPIRE, 1997), and (7) the number of demes to use, i.e. the number of separate populations to evolve in parallel, and to ultimately pick the best individual among the demes for each boost.

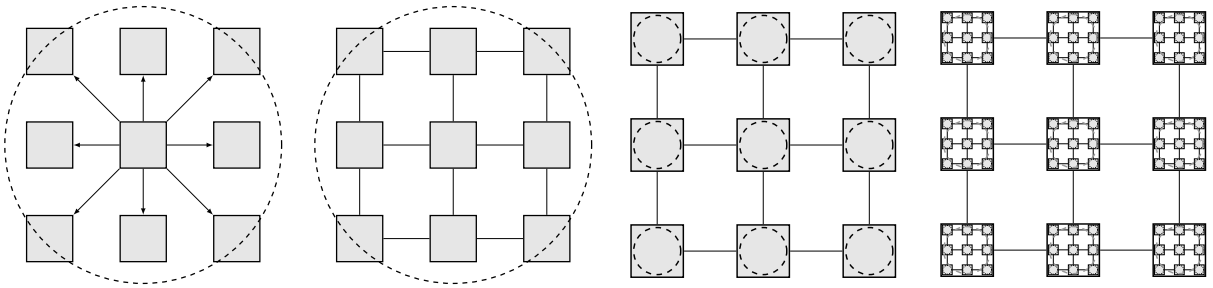
Listing 2 R BIOMOD Interface

```
> Print_Default_ModelingOptions()

 Defaut modeling options. copy, change what you want paste it as arg to BIOMOD_ModelingOptions


-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=  'BIOMOD.Model.Options'  -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=
...
THESIS = list( path_to_rclr = 'C:/Users/michelbieleveld/GitHub/biomod',
               runs = 1,
               use_model = ,
               maximum_generations = 100,
               maximum_instructions = 1024,
               population_size = 100,
               boost = 5,
               demes = 50),
...
-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=
```

4.6 Case study: cloud computing

Generally, in genetic programming parallelisation is achieved in three ways (CANTÚ-PAZ, 1998): (i) global single-population master-slave (Figure 10a, (ii) single-population fine-grained (Figure 10b), and (iii) multiple population coarse-grained (Figure 10c). In the case of a global single-population there is only one population on which the recommendation operators, Section 4.4.6, are executed. What is done in parallel is the fitness evaluation of

Figure 10 – Topology configurations of parallel genetic programs



(a) Master-Slave (b) Fine-grained (c) Coarse-grained (d) Hybrid

Source: Author

Figure 11 – Starcluster configuration file

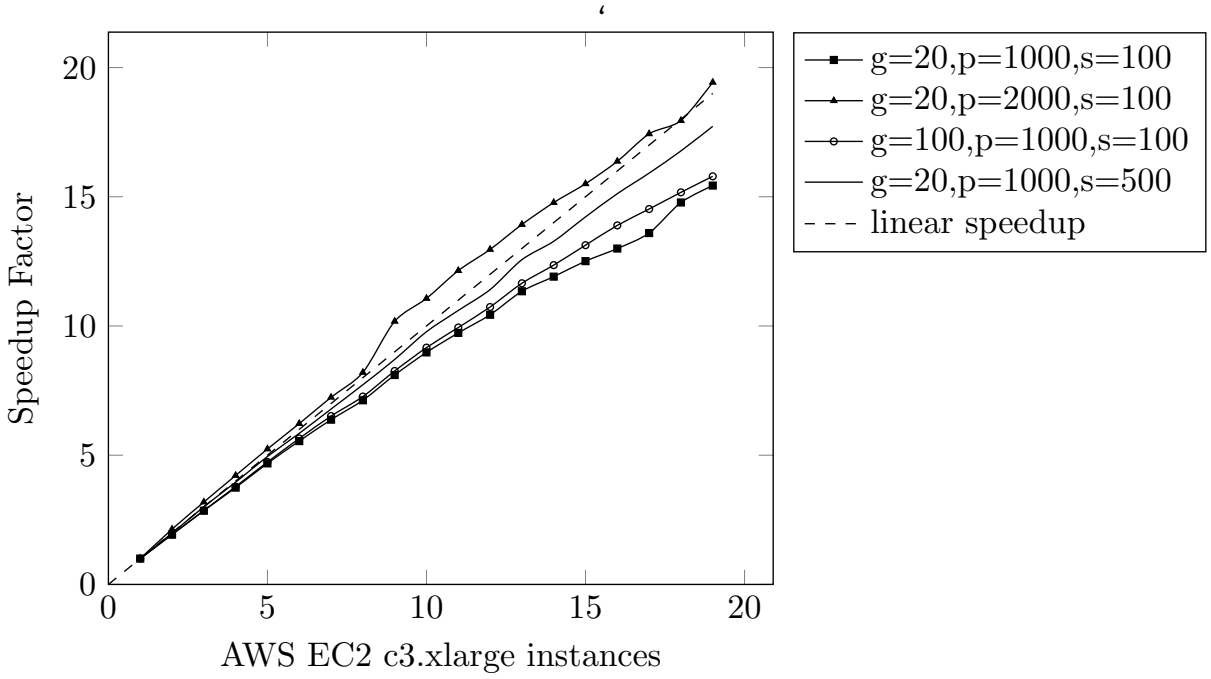
```
[global]
DEFAULT_TEMPLATE=smallcluster

[aws info]
AWS_ACCESS_KEY_ID = ..................
AWS_SECRET_ACCESS_KEY = ......................................
AWS_USER_ID= ....-....-....
AWS_REGION_NAME = us-east-1
AWS_REGION_HOST = ec2.us-east-1.amazonaws.com

[key mykeyABC]
KEY_LOCATION=~/.ssh/mykeyABC.rsa

[cluster smallcluster]
FORCE_SPOT_MASTER=True
KEYNAME = mykeyABC
CLUSTER_SIZE = 19
CLUSTER_USER = sgeadmin
CLUSTER_SHELL = bash
NODE_IMAGE_ID = ami-765b3e1f
NODE_INSTANCE_TYPE = c3.xlarge
AVAILABILITY_ZONE = us-east-1b
SPOT_BID = 0.10
```

Source: Author

the individuals. In the fine grained method, everything is done in parallel but individuals belong to spatially structured population. The recombination operators only select individual from that local population and its direct neighbours. In this way processors do not need to be aware of all neighbourhood, just their own. The last method, the population is divided into several sub-populations (demes) each with individuals only interacting with those of the same sub-population. However, on specific moments some of the best individuals are allowed to migrate from one deme to another.

In the THESIS implementation, parallelisation is obtained by using a hybrid of the coarse-grained method with coarse-grained method (Figure 10d). In this hybrid, nodes are allocated to a specific deme and each node works on a small sub population of the deme.To achieve this Open MPI (HUTCHINSON, 1957) is used in combination with

Figure 12 – Speedup for varying parameters for the THESIS algorithm



Source: Author

the StarCluster[1] (IVICA; RILEY; SHUBERT, 2009) configuration platform to configure Spot Instances in the Amazon Elastic Compute Cloud (EC2) [2]. The used configuration is shown in Figure 11. The implementation also uses OpenMP (DAGUM; MENON, 1998) to parallelise the implementation for a single multi-core machine. The implementation is easily compiled on multiple platforms with the use of CMAKE (MARTIN; HOFFMAN, 2007). The implementation that is used here is compiled with GCC 5.3 (STALLMAN, 2016). The implementation is started on multiple machines and each will be assigned to a deme. There can be a number of demes up and equal to the number of machines, in the testing done here only a single deme is used. When the program is started, the training and validation datasets are distributed from the master machine to all others. Then all machines will train and evaluate independently their solutions. When moving from one generation to the next there is a migration of one or more individuals from all machines to the machine next to it. At the end the best individual of all machines is selected as the solution.

The main advantage of the coarse grained approach is that the compute intensive part of the algorithm, namely calculating the fitness of each individual, is done independently for each individual on each processing node. This means that this approach delivers an increase of speedup that is close to linear with the number of instances that the implementation is run on. There is almost no need to communicate between the processes, except for the migration and picking the best individual after during the last

[1] http://web.mit.edu/star/cluster
[2] https://aws.amazon.com/ec2/

generation. As a result, this method is suitable for low bandwidth parallel computation.

To test the performance of the THESIS algorithm, an experiment was run on the Amazon Elastic Compute Cloud where up to twenty virtual high compute optimised machines were instantiated and configured through StarCluster. The configuration type of these machines was *c3.xlarge*, a configuration that features four high Frequency intel Xeon E5-2680 v2 (Ivy Bridge) processors, 7.5 GiB of memory, and two times fourty GB of storage. The Amazon Machine Image (AMI) *ami-765b3e1f* was used and is based on Ubuntu[3] 12.04. The only changes made to the installations were the addition of *gcc-5* and *cmake* packages on all machines through apt-get. The reason a newer compiler was required was because of the use of the experimental filesystem support that is present in the 5.x and newer releases.

Figure 12 depicts the result of this experiment. The graph shows the amount of speedup relative to running the algorithm on a single instance. The dashed line represents a theoretical linear speedup where the execution time is halved for each doubling of the amount of machines used. In the legend g represents the number of generations executed, p the total size of the population, and s the number of samples used to train the models. On each instance the algorithm also runs in parallel, configured with Open MP, by creating threads up to the number of available cores similar to the global single-population master-slave method. Therefore the x-axis could also represent the number of cores, where the amount of cores used is equal to four times the number of instances used.

The graph shows that for the selected parameters the amount of speedup by adding instances is close to linear due to the fact that the algorithm is embarrassingly parallel. While actual execution time for the experiment with $p = 2000$ is double that of $p = 1000$ the achievable speedup is still linear. Which is also true for the amount of generations used; using $g = 100$ will take five times the amount of time as $g = 20$, but the achievable speedup is still about the same. To train a single model as used in the case studies would take about 10 minutes on a single instance and close to 30 seconds when a cluster of 20 machines is used. Interestingly, the graph also shows a set of parameters where the implementation runs more than the number of machines used faster than the solution runs on a single machine, in other words a super-linear speedup. Alba and Dorronsoro (2008) discussed that this effect is real and has been reported in other work. Two explanations are: better caching and efficient memory usage (BELDING, 1995), and (CANTÚ-PAZ, 2001) theorises that the speedup is due to a reduced convergence by a higher select pressure caused by fitness based migration policies. Concluding it can be said the implementation scales well.

[3] http://www.ubuntu.com

5 Case studies

Contents

This chapter presents two case studies of species distribution modelling with the fuzzy THESIS algorithm and ten models made available through the BIOMOD package that is available in the R project for statistical computing. The case studies in this chapter draw on the data, algorithms, and implementation as discussed in the previous chapter and links it with the biological aspects as discussed in Chapter 2 by applying the algorithm on species to model niches based on selected occurrence and absence points. This chapter discusses first a case study in which a virtual species is created to evaluate the implementation of the THESIS algorithm in a more analytical approach. The second case study discusses the evaluation of the algorithm with real species data. For each case study the pattern of evaluation is described as well as the results and a separate conclusion.

5.1 Case study I - virtual ecology

Numerous studies have compared the performance of SDM techniques and their resulting models (ELITH et al., 2006; MEYNARD; QUINN, 2007) and often leading to an assortment of recommendation and, not infrequently, even conflicting ones (ELITH et al., 2006; PETERSON; PAPE ; EATON, 2007). Comparing modelling techniques is often difficult as the comparisons are made with models trained on real species data. This poses a problem as modelling species distributions is a complex task with many steps (Section 2.1.1) during which there are many sources of uncertainty (Section 2.2.4). The interpretation of the collected species occurrence records is even more complex as: data can be collected from

a place where the species is not in equilibrium with its environment, species might be difficult to detect in their habitat, and sampling can be biased. Biased, because collecting closer to roads and rivers is easier than in the middle of the forrest. The effect of these matters is even more complicated for rare species with low prevalence, and as a result often with small sample sizes. Furthermore, the BAM framework needs to be considered and with it: the dispersal of the species, the geologic history, the biotic interactions, the time interval and special extent, and simply mistaking one species for another during field research.

To create high quality reliable SDMs one needs to consider all the above factors and probably many more. It is practically impossible to know all the facts and circumstances regarding a species because of limited resources at the disposal of researchers and the above issues and biases (HIRZEL et al., 2002; AUSTIN et al., 2006). Even more so, even if it was possible to know all occurrences at a given time, one still needs to understand the relationship with the environmental factors and complex biological processes such as interspecies relationship and diseases.

The Virtual Ecologist (VE) approach in contrast evaluates methods of data sampling, analysis and modelling methods with the use of virtual data by simulating the ecological processes involved and also the sampling processes and biases (ZURELL et al., 2010; THIBAUD et al., 2014). The advantage of this approach is that it gives complete knowledge of the underlying species distribution and its relationship to environmental factors, and possibly biological ones as well. The downside is that this relationship might be oversimplified and not capture the complexity of living real-world species (AUSTIN et al., 2006; HIRZEL; HELFER; METRAL, 2001). While the VE approach tends to simulate the entire process , earlier work by Hirzel, Helfer and Metral (2001) already proposed the use of a virtual species which nature is completely described by its ecological niche and the $\boldsymbol{A}$ region of the BAM diagram. In essence the VE approach removes all uncertainty and present a known truth. Austin et al. (2006) puts forward the view that SDMs should be capable of recovering this artificial "truth" if they are to be of any use on empirical 'complex' real-word data.

According to Meynard and Quinn (2007), Zurell et al. (2010), Hirzel, Helfer and Metral (2001), Barbet-Massin et al. (2012) virtual species simulation proved useful to assess the predictive quality and analysis of SDMs and subscribed to the fact that simulation allows for more accurate model evaluation and better control of the experiment parameters. The simulation of virtual species distributions is increasingly applied as described in an extensive literature review of the use of virtual models and the issues that they tried to clarify, see Miller (2014). There are many advantages of the VE approach, however, after reviewing many works Miller concluded that some studies introduced more problems as a result of the method to generate virtual distributions, and that the correct approach is

determined by the research objectives. A given example of such an introduced problem is the consistently good performance by the BIOCLIM (BUSBY, 1991) method in Saupe et al. (2012) as the method to generate the virtual distribution was probably very similar to BIOCLIM.

5.1.1 Virtual species

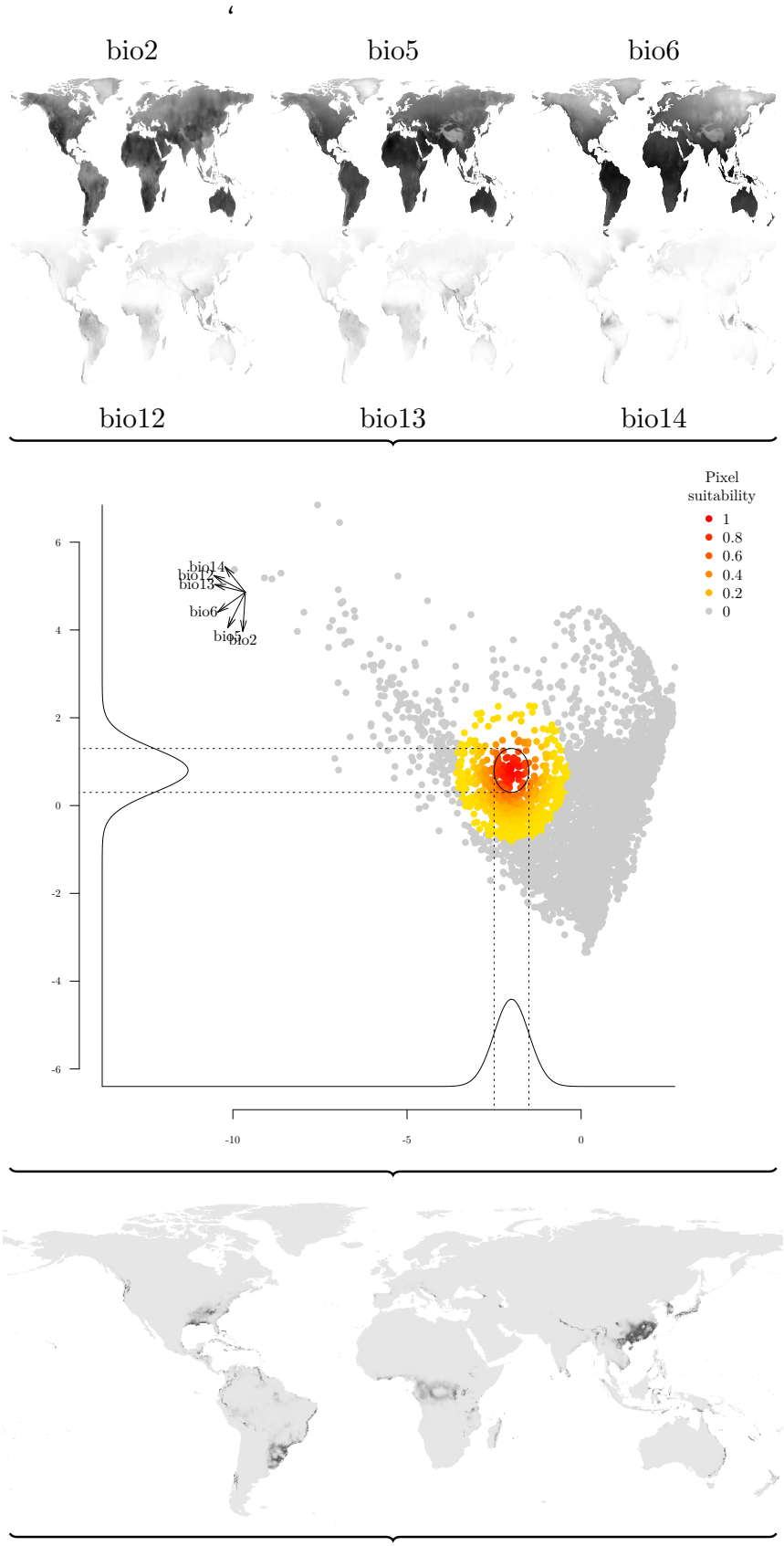
In this case study a virtual species is generated to evaluate the performance of the genetic fuzzy solution described in Chapter 4. This case study focuses on three questions: (1) How does the algorithm perform compared to other popular modelling techniques available in the BIOMOD package in the ideal situation knowing true absences and presences?, (2) What is the impact of knowing only presences and using background data?, and (3) What is the impact of sampling errors on the prediction quality?

A framework for virtual ecology modelling including the generation of virtual species through various customisable species - environment relationships and selecting distribution and sampling biases is made available in an open-source package for the R environment (R Core Team) named *virtualspecies* (LEROY et al., 2015). Other solution exists such as *SDMvspecies* (DUAN et al., 2015), but they do not provide a probabilistic sampling feature, the probability of detecting an occurrence at a site where a species is present (REESE et al., 2005), or the probabilistic approach to convert environmental suitability to presences and absences (MEYNARD; KAPLAN; SILMAN, 2013).

Figure 13 shows the four steps followed with the *virtualspecies* package to build the virtual species used in this case study. The global climate dataset WorldClim (Section 2.1.3) is used as the source for the environmental variables. For this species, six bioclimatic environmental variables derived from the monthly temperature and rainfall values were extracted from the dataset: (1) mean diurnal range, (2) the maximum temperature of the warmest month, (3) the minimum temperature of the coldest month, (4) the annual precipitation, (5) the precipitation of the wettest month, and (6) the precipitation of the driest month. Bucklin et al. (2015) has shown that four of these variables already are highly predictive for the species that were modelled, while using eight variables produces only a slight increase in quality. Taking the middle ground, six environmental variables are chosen to define this virtual species, but the exact number does not matter as long as the variable importance for the species is high, meaning the variables explain well the distribution, as is the case for this virtual species.

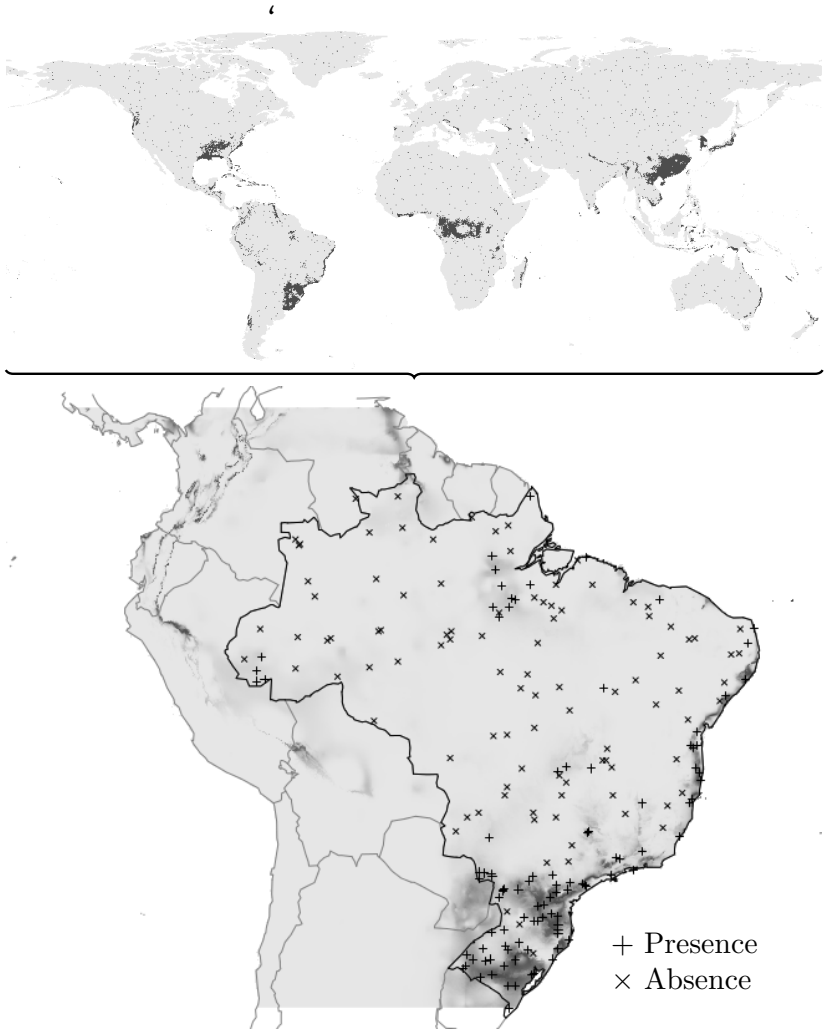
In the first step, the environmental space $\boldsymbol{E}$ of the species (see Section 2.0.3) is defined. For this, PCA (RAO, 1964) was performed to obtain and reduce the variation into two non-correlated environmental variables that explain the distribution of the species. Other methods that do not use PCA to define species suitability exist, but they likely lead to virtual species with unrealistic environmental conditions (LEROY et al., 2015). With the

Figure 13 – The steps taken to build the virtual species with the virtualspecies R framework.



Source: Author

Figure 13 – The steps taken to build the virtual species with the virtualspecies R framework.



Source: Author

PCA 83.5% of the variability of the environmental variation was captured into two independent axes. Similar to the workflow mentioned in Jiménez-Valverde and Lobo (2007), except the mean value is not chosen for the axes, but an arbitrarily one. The environmental range inhabited by the species was arbitrarily chosen as: axis 1, [min=-9.96; max=2.69] : dnorm (mean=-2; sd=0.5), and for axis 2 [min=-3.34; max=6.85] : dnorm (mean=0.8; sd=0.5), see Step 1 in Figure 13.

The second step uses Hutchinson's duality principle to map the defined environmental space of the first step onto the geographic space G. The entire world is here selected as the region of interest and the suitability of each place is indirectly dependent on the earlier selected bioclimatic variables. The resulting map is a suitability map where each cell value is obtained by multiplying the densities of the normal curves of both axes for a given point.

The third step converts the obtained suitability map into presence and absence points. This simulates field research where professional or citizen scientist visit a location and spot a particular species. A simple approach of differentiating presences from absences is by applying a fixed threshold where all values at one side of the threshold are considered presences and at the other side absences (HIRZEL; HELFER; METRAL, 2001; JIMÉNEZ-VALVERDE; LOBO, 2007; PETERSON, 2011). One complication of this approach is that it does not simulate the random processes acting on species occupancies and will lead to misleading data to train SDMs (LEROY et al., 2015). An alternative approach is a probabilistic approach; Meynard, Kaplan and Silman (2013) discusses five reasons why a probabilistic approach is preferred: (1) ecological theory supports a dynamic occupancy pattern, (2) a threshold approach can give an incomplete answer for a set of questions, (3) the capacity to discriminate between presences and absences is lower with a probabilistic approach and SDMs need to be able to handle such ambiguity, (4) standard statistical modelling techniques based on logistic curves, e.g., GLM and GAM, may not converge well as the slope of the logistic curve at the threshold value is infinite, (5) using a single threshold value eliminates all variability and will always result in the same distribution map, as such repeated experiments to separate the effects of prevalence and sample bias can not be run. The environmental suitability is converted to a probability of occurrence with a logistic function using a probabilistic approach. A random draw then determines if a cell is marked as a presence, e.g., a cell value of 0.1 has a one in ten chance to be marked as such. The result is the presence-absence map shown in Step 3 of Figure 13.

The fourth step is to optionally limit the region and then draw samples. One reason to limit the region is to be able to test model transferability, see Section 2.2.2. To test this training samples are randomly drawn from the spatial area defined by the borders of the country Brazil. While testing sample are randomly drawn from the entire world. To answer the three question asked for this case study, an error probability is defined to simulate misidentifications. The testing of the generate SDMs will always be done on the sampled true absences and presences, while models will be trained on either true presences and absences or true presences and background data depending on the question asked.

5.1.2 Evaluation of SDM performance

To explore SDM performance, a training dataset for model calibration needs to be prepared. There is no consensus on the sample size n of presences required to effectively train models. Earlier work suggests that model performance is negatively affected if $n < 30$ for all tested methods (WISZ et al., 2008). In another work $n > 70$ is required to make model reliability independent of the sample size (JIMÉNEZ-VALVERDE; LOBO; HORTAL, 2009). Other work suggested that useful models with as few as five to ten positive obser-

vations and that models trained with $n = 50$ did not outperform models with $n > 100$ (HERNANDEZ et al., 2006). However, in general it is assumed that the quality of the model is increased when more samples are used (JIMÉNEZ-VALVERDE; LOBO; HORTAL, 2009; PEARCE; FERRIER, 2000; STOCKWELL; PETERSON, 2002). Due to the absence of a clear guide, a presence sample size of $n = 80$ was chosen. Well above the recommendations of previous research, but still low enough to simulate the lack of presences for many species in the real world. There is also no consensus on the correct prevalence, or the proportion of presences to absences in presence-absence datasets. According to Jiménez-Valverde, Lobo and Hortal (2009) in the absence of noise the effect of prevalence is not noticeable for $n > 50$. Others suggest that the proportion of presences and absences should be balanced and so a prevalence of 50% is chosen (LANTZ; NEBENZAHL, 1996; HOEHLER, 2000; MCPHERSON; JETZ; ROGERS, 2004).

Overfitted SDMs are good at explaining the training samples and possibly the training region, but likely perform poorly on unseen data and new regions. To check for overfitting ten-fold stratified cross validation is the more robust and popular method (BREIMAN, 1993; BORRA; CIACCIO, 2010). In ten-fold stratified CV the data is randomly divided into ten almost equal sized blocks while maintaining a 0.5 prevalence. The CV process involves creating ten times a model on nine of the blocks and testing the model on the reserved block. Each time a different reserved block is used for testing. Using CV will not only signify the goodness of the fit of the trained models, it is also an indication that a heuristics based algorithm, such as implemented for this thesis consistently converges to a usable model.

To answer the first question the algorithm performance is compared to other modelling techniques available in the BIOMOD package. BIOMOD provides support for: Generalised Linear Model (MCCULLAGH; NELDER, 1989, GLM), Generalised Additive Model (HASTIE; FITHIAN, 2013, GAM), Multiple Adaptive Regression Splines (FRIEDMAN, 1991, MARS), Generalised Boosting Model (FRIEDMAN, 2001, GBM), Classification Tree Analysis (BREIMAN et al., 1984, CTA), Artificial Neural Network (HORNIK; STINCHCOMBE; WHITE, 1989, ANN), Surface Range Envelop or also known as BIOCLIM (BUSBY, 1991, SRE), Flexible Discriminant Analysis (HASTIE; TIBSHIRANI; BUJA, 1994, FDA), Random Forest (HO, 1995, RF), and Maximum Entropy modelling (PHILLIPS; ANDERSON; SCHAPIRE, 2006, MAXENT). After modification BIOMOD also supports the implemented genetic programming SDM that is discussed in this thesis named 'THESIS' in the results. For the ideal situation a training sample set is constructed of 100 true absences and 100 true presences, both restricted to Brazil to test whether the models can extrapolate to other regions in the world. This dataset of 200 samples and a prevalence of 0.5 has no errors in its observations and is depicted in Step 4 of Figure 13. In the figure presences are marked as '$+$' and absences marked as '$\times$'. The main attraction of SDMing is that models make meaningful predictions for new regions and times. To test this ability the testing data set is obtained by

randomly sampling 20.000 locations on the world wide suitability map. It is important to note that the prevalence of the testing set is far from 0.5 and is proportional to the prevalence of the species. For this reason the model evaluation metric TSS is used, because it is readily applied for presence–absence predictions and is shown to be not affected by prevalence (ALLOUCHE; TSOAR; KADMON, 2006).

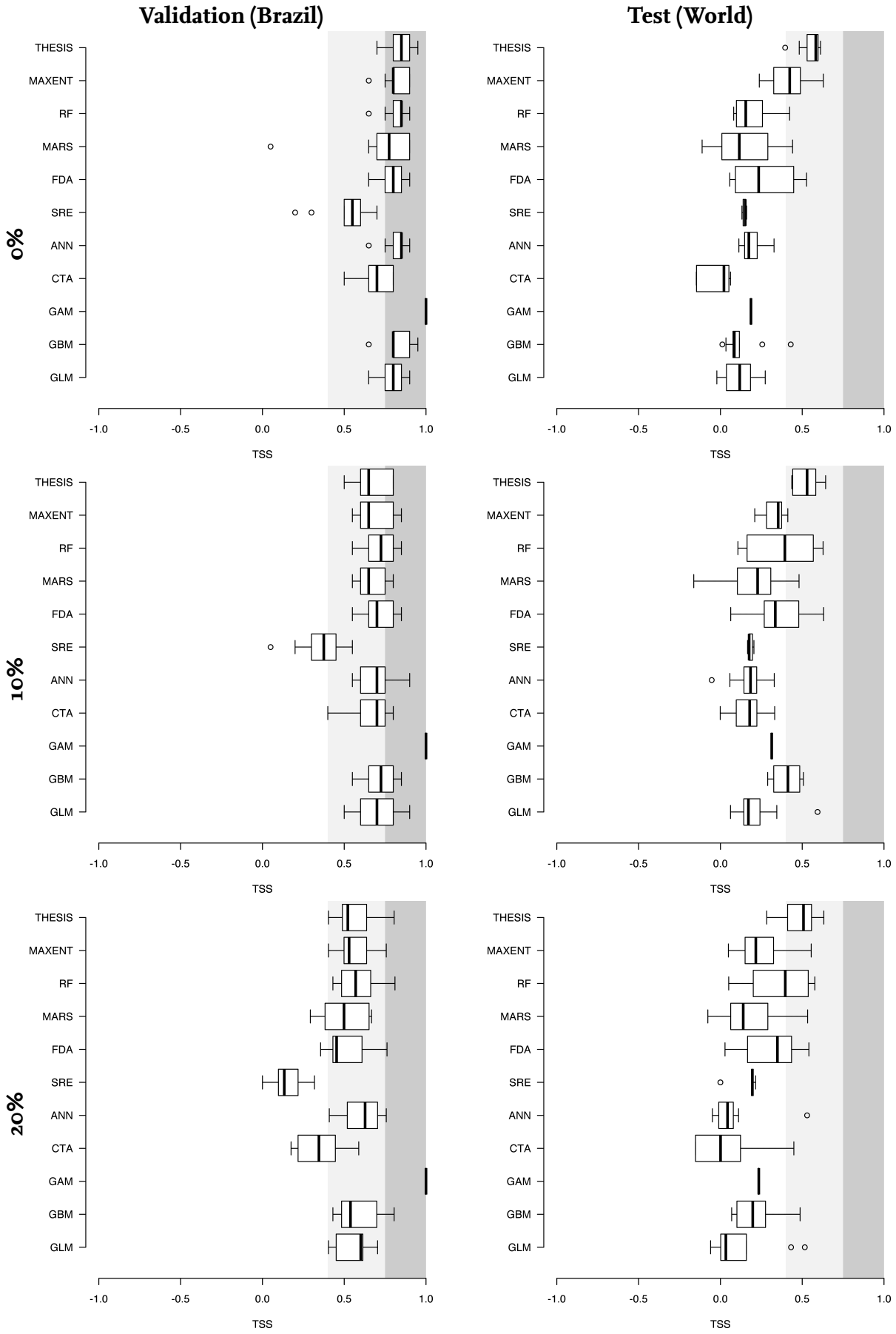
5.1.3 Results

Models are first trained on the discussed training dataset with BIOMOD (default settings for all algorithms) and then evaluated in two ways: on the held out validation data during CV, the left **Validation** column in Figure 14, and on the global test dataset, the right **Test** column in Figure 14. The diagrams graphically illustrate the numerical distribution of the TSS values for the different classifiers, using the smallest and largest observations for the whiskers, and the lower quartile, median and the upper quartile for the boxes. Each row in the figure shows the result of training and validating the models after more and more error is introduced by equally increasing false presences and false absences by a certain percentage of the total samples. During the experiment the training dataset is fixed at 200 samples and a prevalence of 0.5, the test dataset is kept the same for all experiments and does not contain introduced errors. The idea is to test the predictive power under different noise conditions, which is an indication of how well the models will perform when pseudo-absences or background sampling is used as both can be seen as introducing noise (marking a presence as an absence).

The first row (0% error) of Figure 14 shows the evaluation when no noise is introduced. All presences and absences used to train and evaluate are 'true' in this situation. As is expected for all models, part of the box and whiskers lay in the dark grey area, signifying a TSS > 0.75, and are considered 'excellent' models. Projecting all models to an unknown area immediately shows a big drop in the quality of almost all algorithms. Not one algorithm produces 'excellent' results and only the algorithms *THESIS,MAXENT,MARS,FDA* produce 'good' results ($0.4 < TSS < 0.75$).

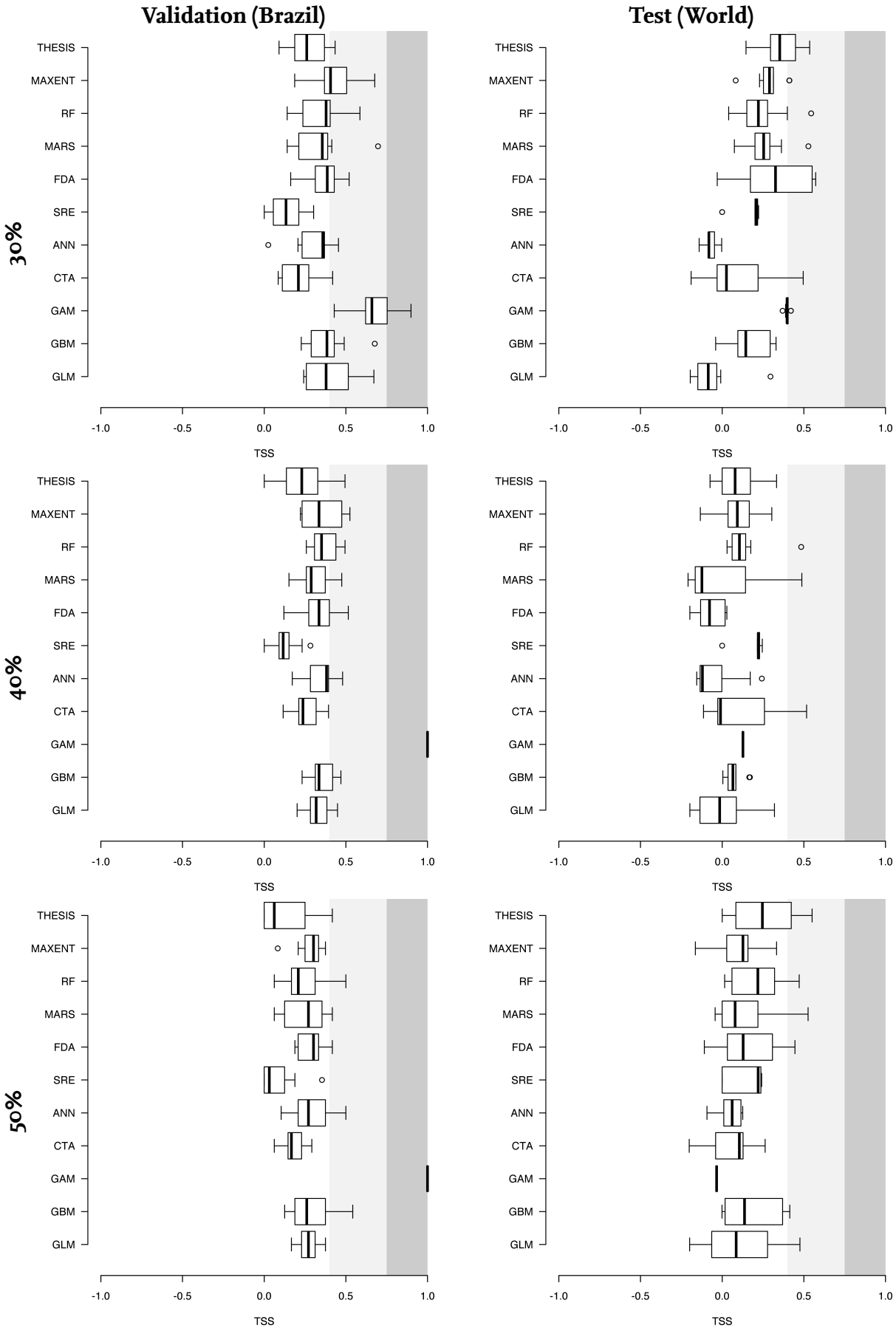
A projection is made for each algorithm with the model that obtained the highest TSS score on the validation dataset. These projections are shown in Figure 15. Figure 15a shows the actual distribution of the species. Ideally model projections should be as similar to this image as possible. Note that this is only possible if the bioclimatic factors present in Brazil are representative of the species' niche. In the case of this virtual species it is known to be true, but this assumption does not necessarily hold for real word projections into challenging new regions and times. The algorithms *FDA,MAXENT,SRE, THESIS* produce results quite similar to the original distribution. The other algorithms all predict good in Brazil, but they incorrectly predict the species to be overly extensive on a very large part of the northern hemisphere.

Figure 14 – TSS score of BIOMOD model projections for increasingly higher presence- and absence error for both validation and test datasets. The light grey area indicates models that are considered 'good' (see See Section 2.2.4.3), while the dark grey area indicates 'excellent' models.



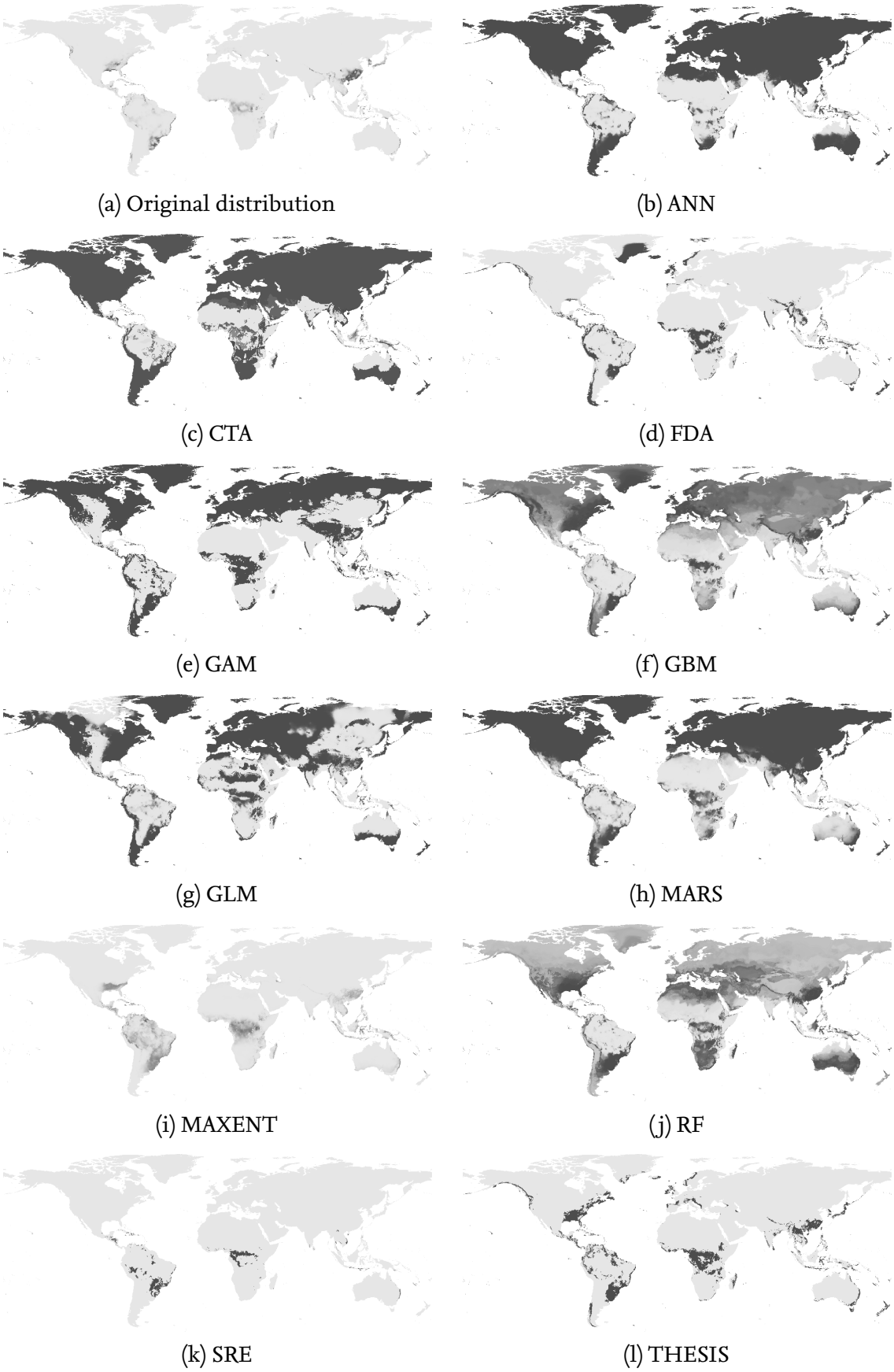
Source: Author

Figure 14 – TSS score of BIOMOD model projections for increasingly higher presence- and absence error for both validation and test datasets. The light grey area indicates models that are considered 'good' (see See Section 2.2.4.3), while the dark grey area indicates 'excellent' models.



Source: Author

Figure 15 – Virtual species model projection for several algorithms



(a) Original distribution (b) ANN

(c) CTA (d) FDA

(e) GAM (f) GBM

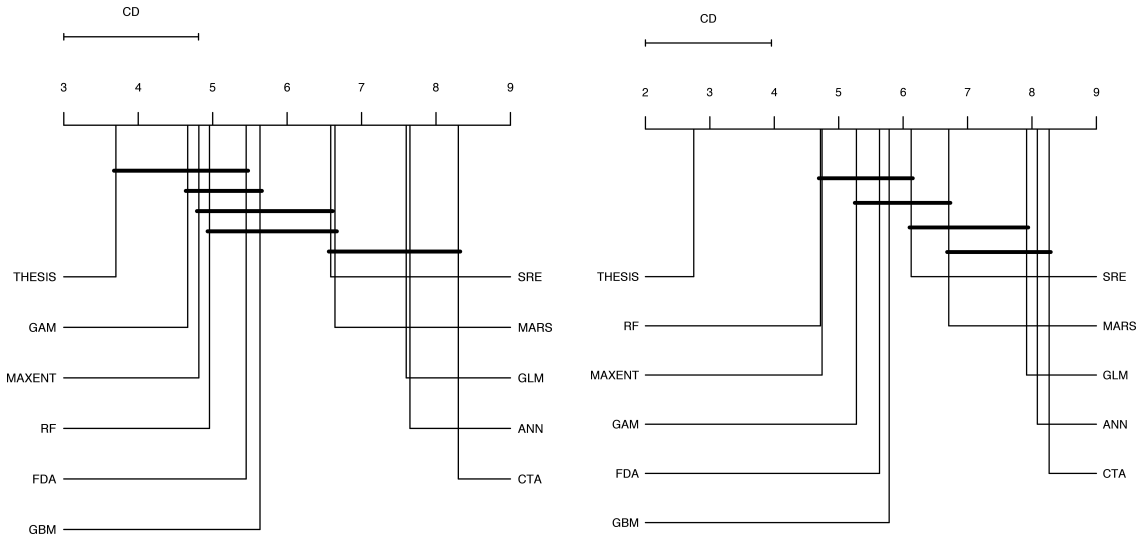
(g) GLM (h) MARS

(i) MAXENT (j) RF

(k) SRE (l) THESIS

Source: Author

Figure 16 – Nemenyi critical difference diagrams comparing the overall performance of the THESIS method with other BIOMOD approaches using the TSS score as performance metric and considering all 0%-50% error experiments. Tested on (a) CV validation datasets, (b) test dataset. Average ranks of the examined SDMs are depicted. Bold lines indicate groups of SDMs which are not significantly different (at $p = 0.05$ their average ranks differ less than the Critical Distance (CD) value/distance indicated by CD line on the top.)



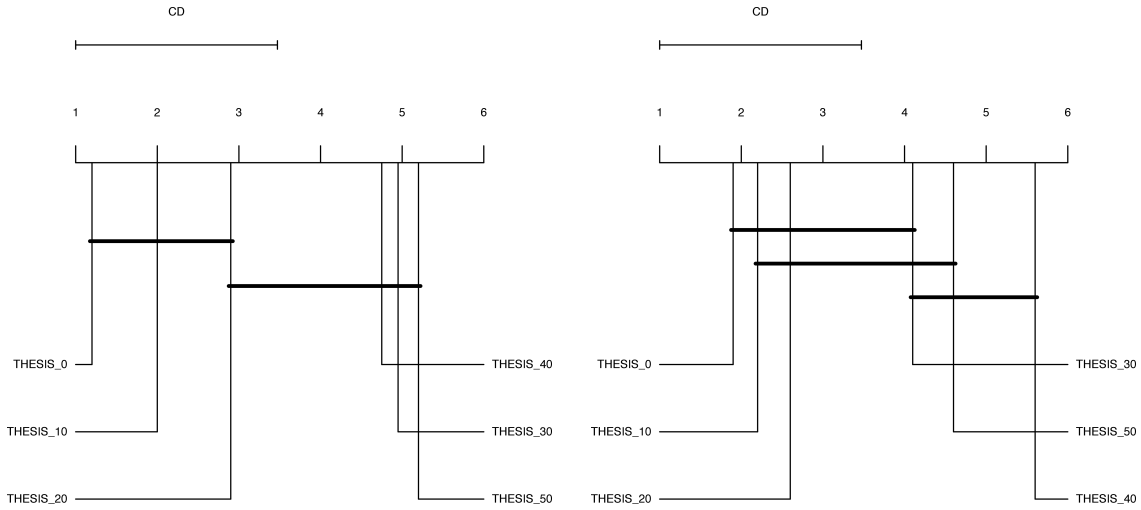
(a) Validation (b) Test

Source: Author

According to Demšar (2006), the Wilcoxon signed rank test (WILCOXON, 1945) is the test to compare two methods: the Friedman test (FRIEDMAN, 1940) when comparing more than two methods and the Nemenyi test (NEMENYI, 1962) to post-hoc compare all classifiers over multiple datasets when the Friedman test indicates that there is a significant difference between the methods. These three tests are the non-parametric equivalents of the paired t-test, the ANOVA analysis of variance, and the Tukey test. Non-parametric tests have the desirable quality of not requiring that the variables are distributed normally. These kind of tests are less powerful than the parametric tests that use data with a particular distribution. As a consequence non-parametric tests are less reliable to reject the null hypothesis when it is false and lead to erroneous conclusions. The three tests, Friedman, Nemenyi and Wilcoxon, are executed in R (R Development Core Team, 2006) in combination with the *scmamp* package (CALVO; SANTAFE, 2015).

The Friedman test is applied to the underlaying data of the graphs depicted in Figure 14. The test confirmed that there is a statistically significant difference in the TSS values obtained from the SDMs. To investigate this difference the post-hoc Nemenyi test was applied. The results of this test are graphically depicted in Figure 16 with a Critical Difference diagram (DEMŠAR, 2006). In the figure the SDMs are ranked with the best overall classifier at the left to the least performing classifier at the right. The Critical Differ-

Figure 17 – Nemenyi critical difference diagrams comparing the overall performance of the THESIS method with increasingly higher error rates using the TSS score. Tested on (a) CV validation datasets, (b) test dataset. Average ranks of THESIS are depicted for each error rate. Bold lines indicate groups of models with specific error rates which are not significantly different (at $p = 0.05$ their average ranks differ less than the Critical Distance (CD) value/distance indicated by CD line on the top.)



(a) Validation (b) Test

Source: Author

ence (CD) of the Nemenyi test is indicated in the top left of the figure. When two classifiers differ less in their ranking than the CD then there is no significant difference with 95% confidence level ($p = 0.05$), else there is a significant difference between the two. SDMs connected with a bold horizontal line are closer to each other than the CD and thus do not differ significantly with 95% confidence interval. Similarly, a CD diagram, Figure 17, is created to compare the THESIS models with themselves that were generated for the 0,10,20,30,40 and 50% error rate experiments (shown in Figure 14). Figure 16 shows that the implemented algorithm ranks among the most highly predictive algorithms for this particular species. Figure 17 shows that the impact of noise does not significantly impact the quality of the predictions for this species for at least up to twenty percent on the validation data, and thirty for the test data.

5.1.4 Conclusion

The results of this case study show that modelling algorithms function differently depending on the application and the data quality of the input data. Concluding the questions that were asked are answered as follows:

How does the algorithm perform compared to other popular modelling techniques available in the BIOMOD package in the ideal situation knowing true absences and presences?

The developed THESIS algorithm produces 'good' models for this species for all experiments up to 30% of error in the input data. Even for the 40% and 50% error experiments 'good' models are produced, although only for a few of the folds. Statistically, the algorithm performs comparable to those present in BIOMOD. Experimental results show that the algorithm generalises well, generating even up to 50% error 'good' models that project well into the unseen region. Based on the CD diagram (Figure 16) the following observations are made:

- The SDM performance ranking for projecting into the unseen region is from best to worst: THESIS, RF, MAXENT, GAM, FDA, GBM, SRE, MARS, GLM, ANN, and CTA.
- The performance ranking on the validation data is from best to worst: THESIS, GAM, MAXENT, RF, FDA, GBM, SRE, MARS, GLM, ANN, and CTA.
- The THESIS algorithm does not show a significant difference (95% confidence level; $p = 0.05$) with the algorithms GAM, MAXENT, RF and FDA on the evaluation data.
- The THESIS algorithm rank as best and shows a significant difference (95% confidence level; $p = 0.05$) with all other algorithms on the test data.

What is the impact of sampling errors on the prediction quality?

The answer is that, as expected, an increase of error causes a deterioration in model quality. This trend is easy to see in the Validation column, Figure 14, by the continuous decrease of the TSS value of the models for each graph with a higher error rate. Almost all algorithms produce 'excellent' models for the ideal situation, then the models deteriorate to 'good' after introducing 10% error, and typically ending up worse than 'good' for 20% error. Note that these errors are really worst case scenarios as it also includes the percentage of errors in the presence data, that in practice may be filtered out with careful data preparation. Based on the CD diagram (Figure 17) the following observations are made regarding the impact of sampling errors:

- For the validation tests, there is no significant impact (95% confidence level; $p = 0.05$) on the model projection quality up until and including 20% of error had been introduced.
- For projecting into unseen regions, there is no significant impact (95% confidence level; $p = 0.05$) on the model projection quality up until and including 30% of error had been introduced.

Concluding, for this species and selection of folds, up to 30% error in both presence

and absence data, does not significantly (95% confidence level; $p = 0.05$) impact model quality. However, as shown in Figure 14 part of the box is no longer in the 'good' area of the graph and thus care needs to be taken to select the correct model based on the validation data.

What is the impact of knowing only presences and using background data?

The comparison of SDMs in Elith et al. (2006) shows that presence-absence models with pseudo-absences or background data have a tendency to outperform presence only-models and are therefore increasingly used (BARBET-MASSIN et al., 2012). With background and pseudo-absence data locations that fall within G_P can erroneously be marked as absence leading to commission errors. The amount of error will depend on the size of the study region and the relative extent of the potential distributional area of the species with that region (ANDERSON; LEW; PETERSON, 2003; PHILLIPS; ANDERSON; SCHAPIRE, 2006; PETERSON, 2011). In general though these commission errors are comparable to what is modelled by introducing the error in the experiments for this case study. The species modelled here has a species prevalence of 0.043, meaning the fraction of the area a species occupies divided by the total area of the region (the world in this case) equals to 0.043. Assuming that background and pseudo-absence locations are randomly chosen, the amount of presence cells selected as absence will be far lower than the 30%, as shown in the literature where models for species with low prevalence have higher accuracy than for species that occupy large areas (KADMON; FARBER; AVINOAM, 2003; SEGURADO; ARAÚJO, 2004; HERNANDEZ et al., 2006; FRANKLIN; MILLER, 2009). As discussed the THESIS algorithm generates good models up to 30% for this species, therefore, there is no significant impact of using background data.

In general this case study suggests that the THESIS algorithm should be used for applications where transferability is required, such as invasion and global change biology. A negative consequence of this is that the algorithm limits the detail that is captured necessary for interpolation.

5.2 Case study II - *Zenaida macroura*

In this study the quality of the algorithms to model a living real world species *Zenaida macroura* is examined. The experiment to train and test the algorithms overlaps with the experiments discussed in Peterson, Pape and Eaton (2007). The focus of the experiments is to answer two questions: How do the SDM algorithms in BIOMOD and the THESIS algorithm perform on a real species? How well do the generated models transfer to new unseen regions?

Figure 18 – *Zenaida Marcroura*

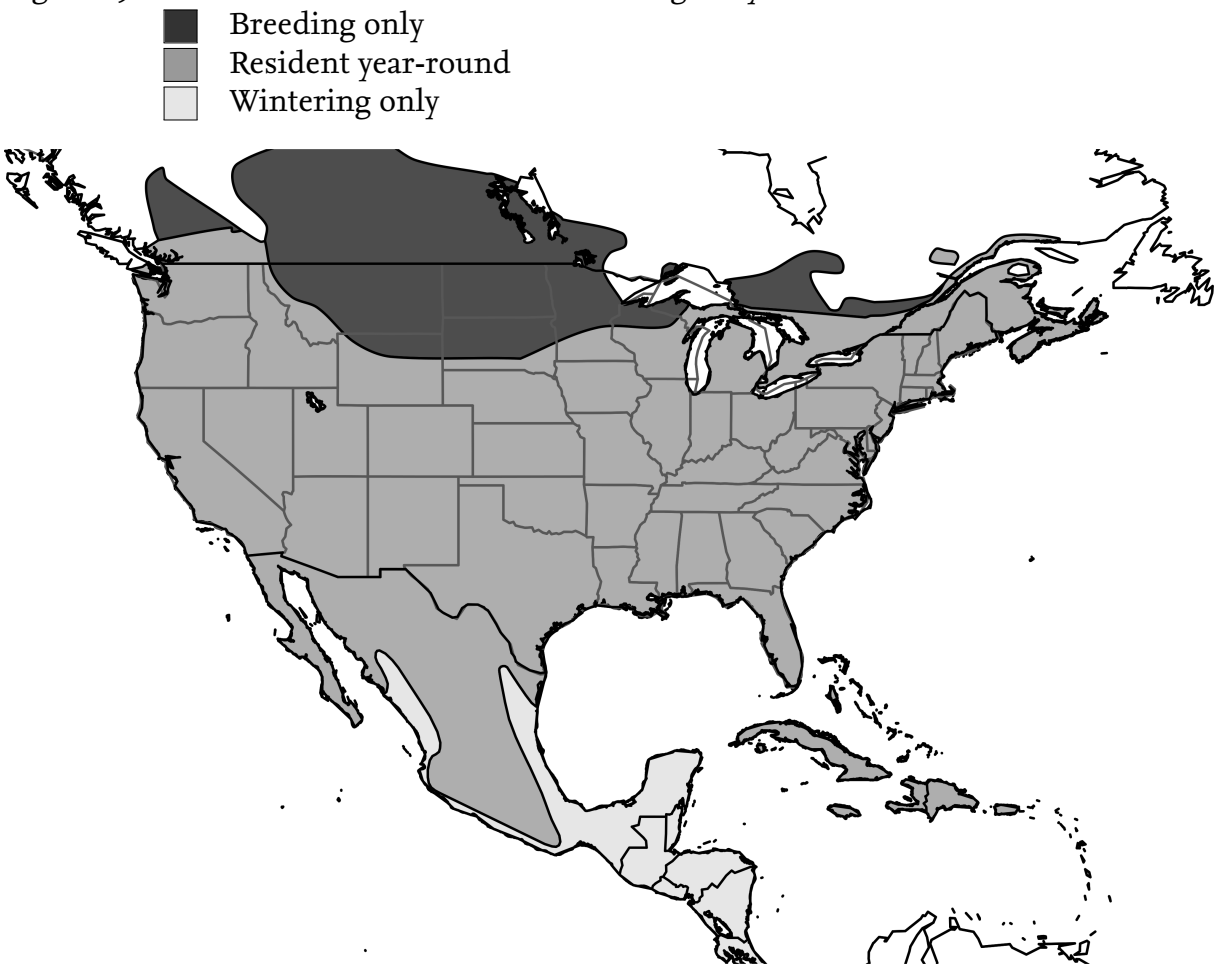
Source: Photo "Zenaida Macroura" by Menke Dave, U.S. Fish and Wildlife Service is licensed under CC0 (Menke Dave, U.S. Fish and Wildlife Service,)

5.2.1 *Zenaida macroura*

The mourning dove (*Zenaida macroura*, see Figure 18) is one of the most abundant birds in northern america and is widely distributed. It has been spotted to be present in various habitats, such as urban areas, woodland edges, bushy fields, meadows and scrubland (RUMELT, 2016).

Northern populations that breed up until the south of Canada migrate in winter to the southern of Mexico and Central America, see Figure 19. Survey data of the species was obtained from the BBS (SAUER; J. E. Hines; J. Fallon, 2001). The species was chosen because of the large geographic distribution and large sample size in the BBS and the interesting model projections for this species in (PETERSON; PAPE ; EATON, 2007). Since the species is considered and known to be a good disperser, the distributions of the species will not be significantly affected by dispersal and historical factors ($\boldsymbol{M}$ in the BAM framework, Section 2.0.4), but instead be largely limited to the ecological factors $\boldsymbol{B}$ and $\boldsymbol{A}$. Normally, in the context of using SDM, the $\boldsymbol{B}$ factor is considered (Eltonian) noise and is hypothesised to play mostly a role in fine-grained spatial regions in contrast to the $\boldsymbol{A}$ factor that has long-ranged autocorrelations with the species distribution (SOBERÓN; PETERSON, 2005; SOBERÓN, 2007). Stable population occurrences were obtained by selecting from the BBS survey routes where the species has been detected during yearly surveys, taken at the height of the breeding season, in eight, not necessarily consecutive, years during the period 1991 - 2000. Absences were defined by selecting survey routes where the species has not been detected by the bird watcher for any of the years during this same period. All

Figure 19 – *Zenaida Marcroura* distribution during the year.



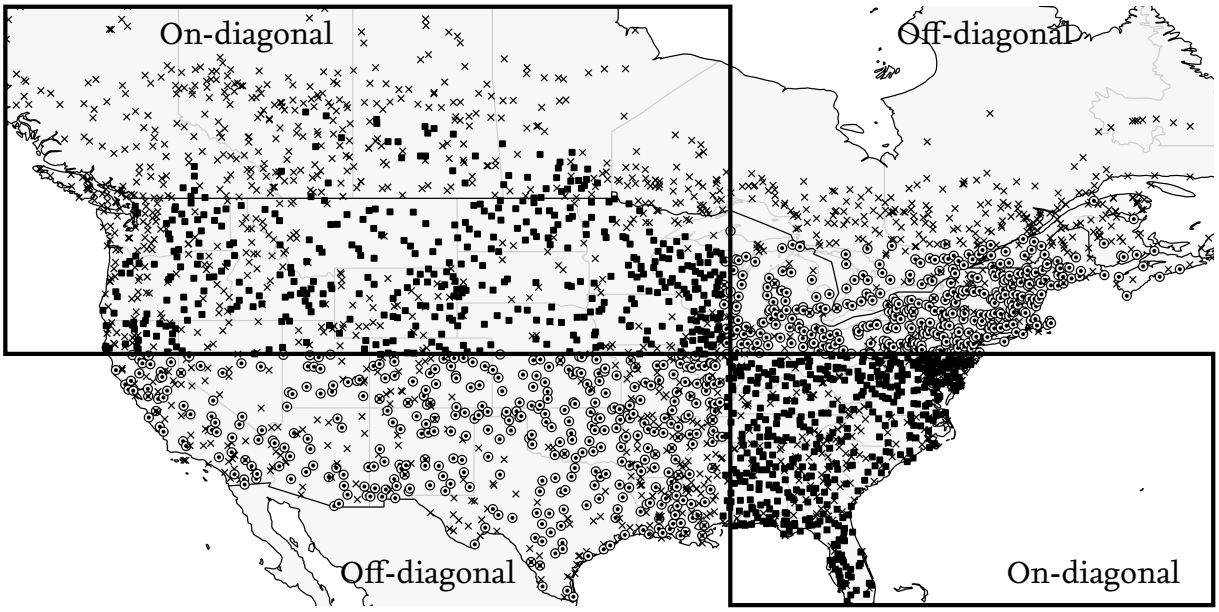
Source: Ranges from BirdLife International and NatureServe (2015)

other routes are not used for this study. This resulted in a data set with 1155 true absence points and 1003 true presence points of *Zenaida macroura* in North America (westlimit=-169.5; southlimit=24.5; eastlimit=-52.0; northlimit=76.5; projection=WGS84) as shown as in Figure 20.

5.2.2 Evaluation of SDM performance

To test if the SDMs extrapolates and transfers into not sampled geographic areas the BBS data is divided into two stratified subsets by separating the samples into geographic quadrants divided by the median longitude and latitude of the samples and then grouping the two diagonally opposing quadrants as shown by the boxes in Figure 20. For this test the models were trained with the data on the on-diagonal and tested with the presence-absence data from the off diagonal. The position of the quadrants was chosen so that all quadrants contain about the same number presence points. A ten-fold CV is used to fairly compare the algorithms by making sure that consistently good models are generated and to evaluate the models' prediction quality on the calibration dataset.

Figure 20 – The geographic distribution of *Zenaida macroura*, obtained from the BBS dataset (1991-2000). Absence data marked by small crosses, presence data as black squares (main diagonal) and by dotted circles (antidiagonal). Models were trained with one diagonal and tested with the other.

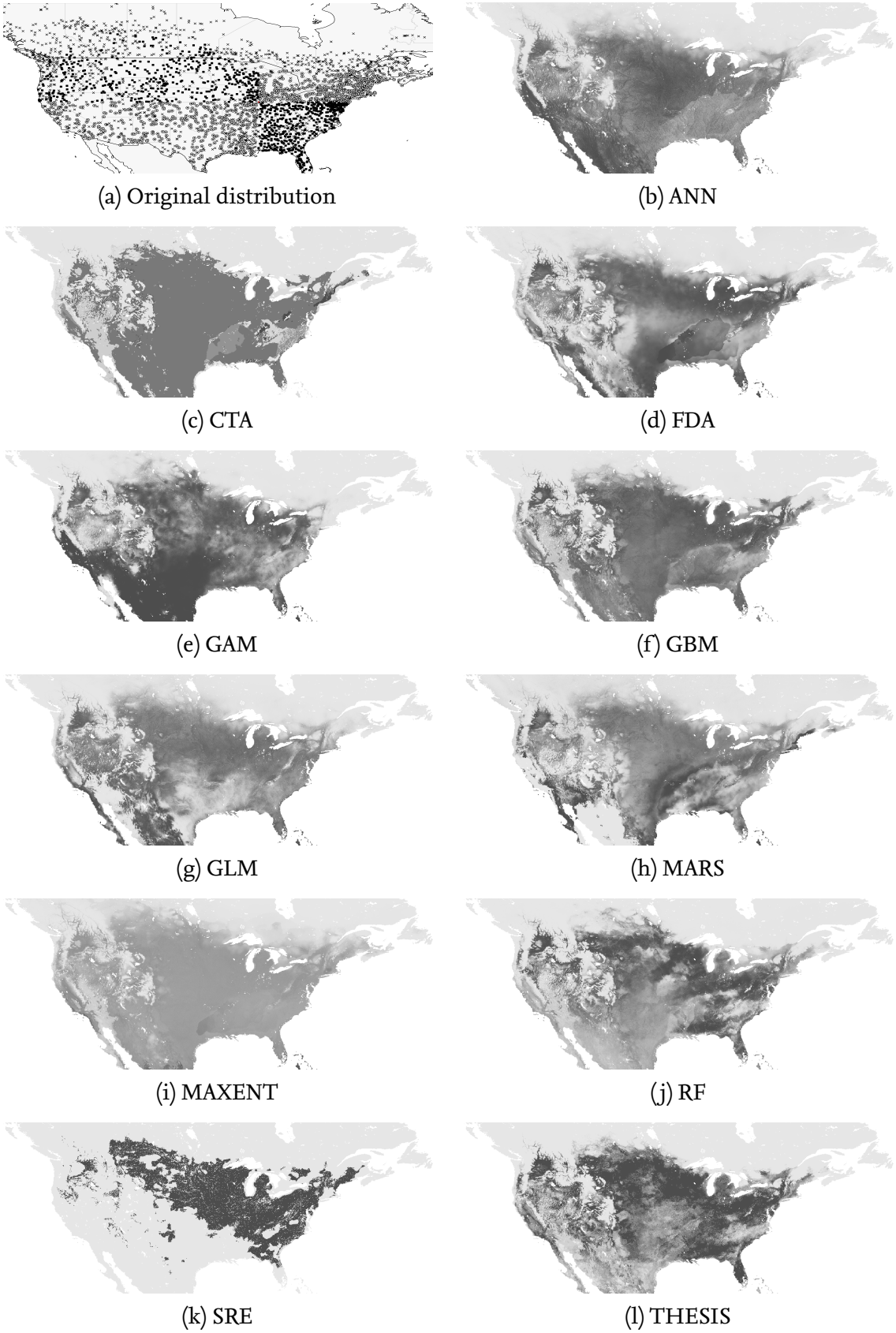


Source: Author

To test how good the SDMs in general predicts the distribution of *Zenaida macroura* the data of both diagonals is merged into one data set. Models are trained for with ten-fold CV. For each fold 50% of the data points is randomly selected, without any regard for geographic location, for calibration and the other 50% is held out for model testing. The idea is that in contrast to the experiment described in the previous paragraph here the environmental space $\boldsymbol{E}$ is more extensive sampled and thus more representative for the species.

Similar to the work of Peterson, Pape and Eaton (2007), the abiotic data set is formed by joining the nineteen bioclimatic variables of the WorldClim data set (HIJMANS et al., 2005) with four layers (elevation, slope, aspect and the compound topographic index) from the digital elevation model Hydro1k (U.S. Geological Survey, 2000) data set. To reproduce the data set PCA is used to reduce the amount of factors. The first eleven components together account for 97,9% of the variance. A difference with the reference work is that the Hydro1k data set is not resampled to 10' resolution, but only projected to WGS84 and aligned with the 30 arc-seconds WorlClim data set. The reason is that occurrence point data (*Longi* and *Lati* fields) from BBS are used and not the route paths shape files. Thus avoiding the effect of resampling as, for example, slope calculations are resolution dependent.

Figure 21 – Zenaida Macroura model projection with stratified sampling



(a) Original distribution (b) ANN

(c) CTA (d) FDA

(e) GAM (f) GBM

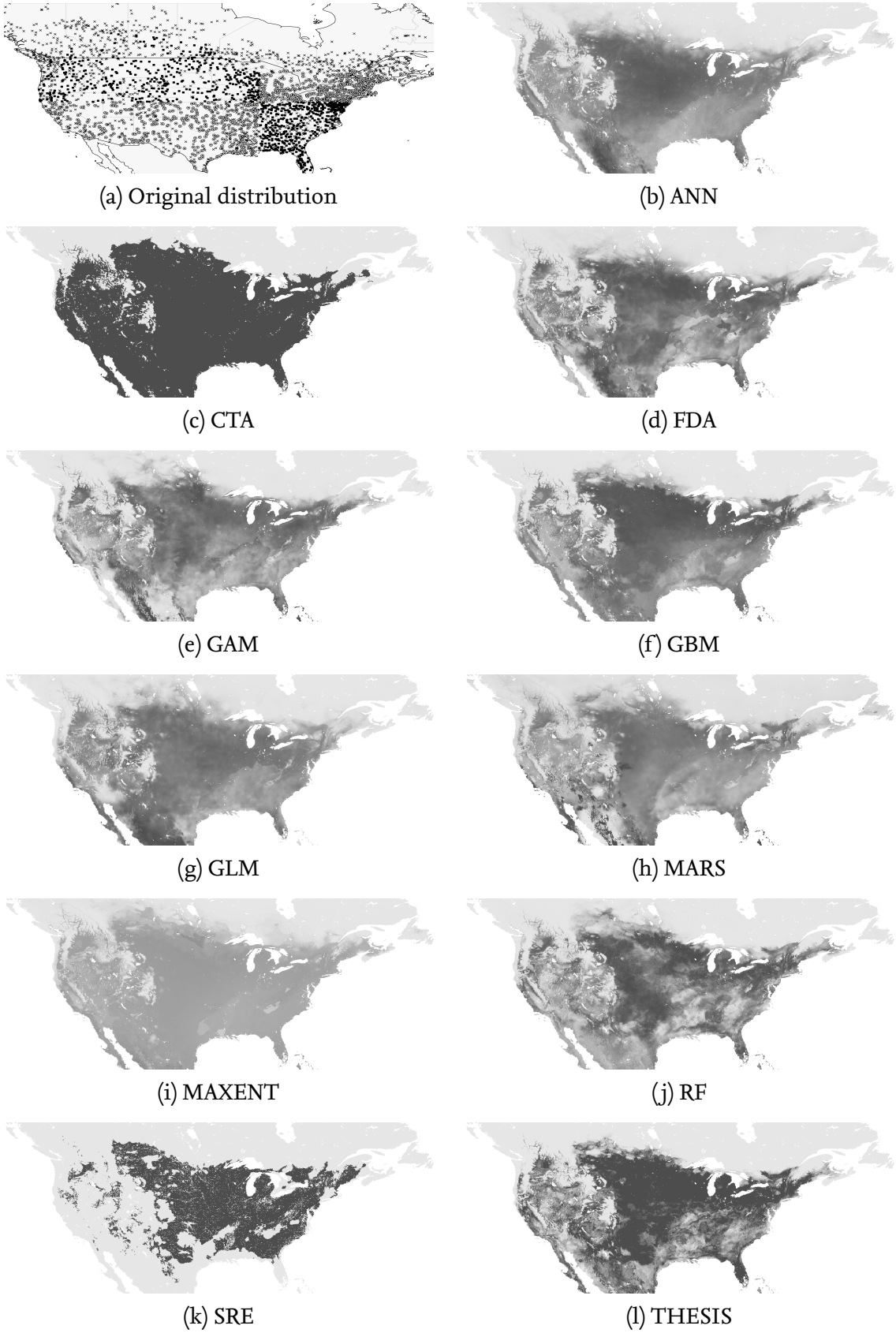
(g) GLM (h) MARS

(i) MAXENT (j) RF

(k) SRE (l) THESIS

Source: Author

Figure 22 – Zenaida Macroura model projection with random sampling



(a) Original distribution (b) ANN

(c) CTA (d) FDA

(e) GAM (f) GBM

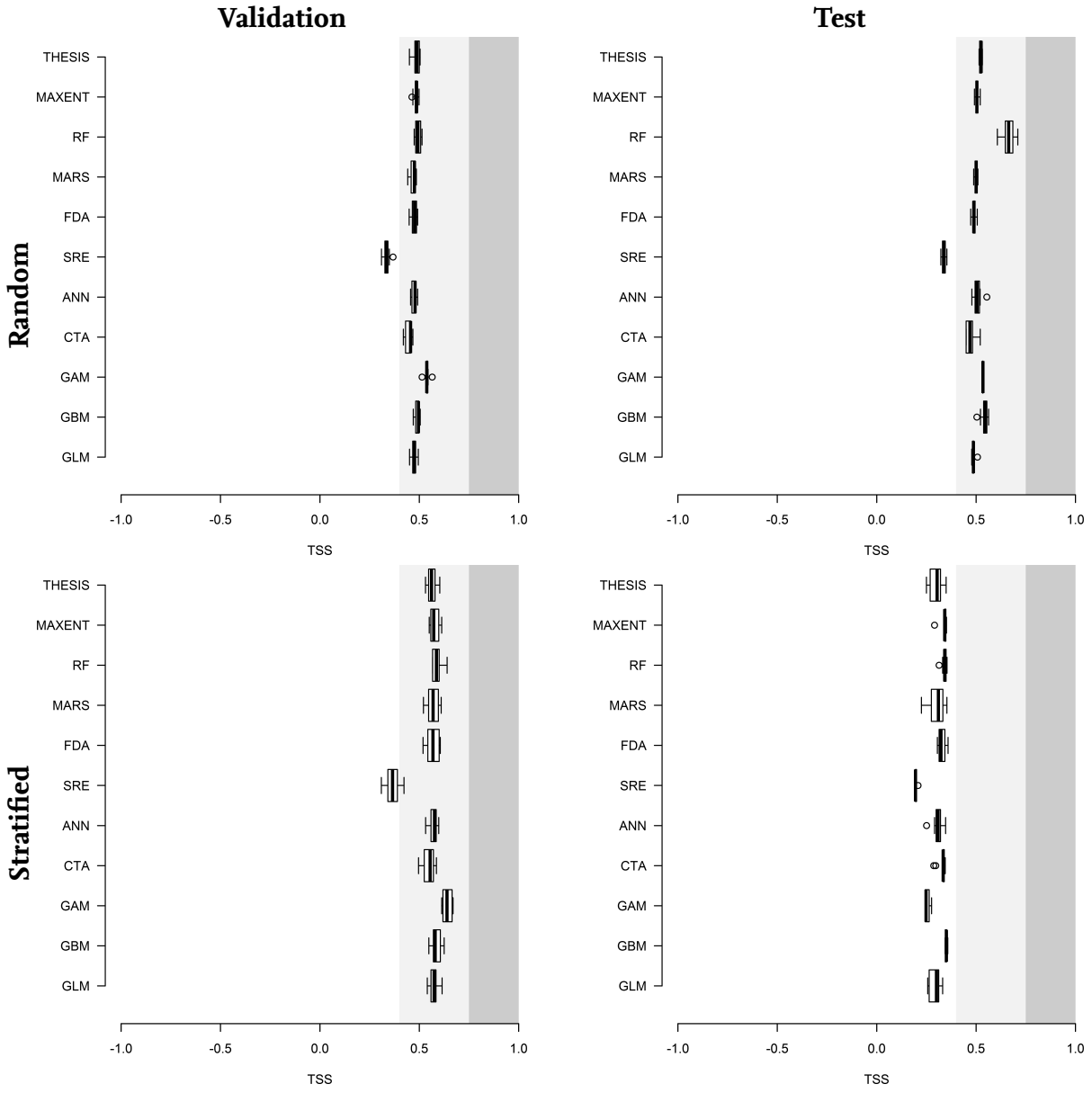
(g) GLM (h) MARS

(i) MAXENT (j) RF

(k) SRE (l) THESIS

Source: Author

Figure 23 – TSS scores of BIOMOD models and THESIS projections for stratified and random datasets, and both with their respective validation- and test datasets. The light grey area indicates models that are considered 'good' (see See Section 2.2.4.3), while the dark grey area indicates 'excellent' models.



Source: Author

5.2.3 Results

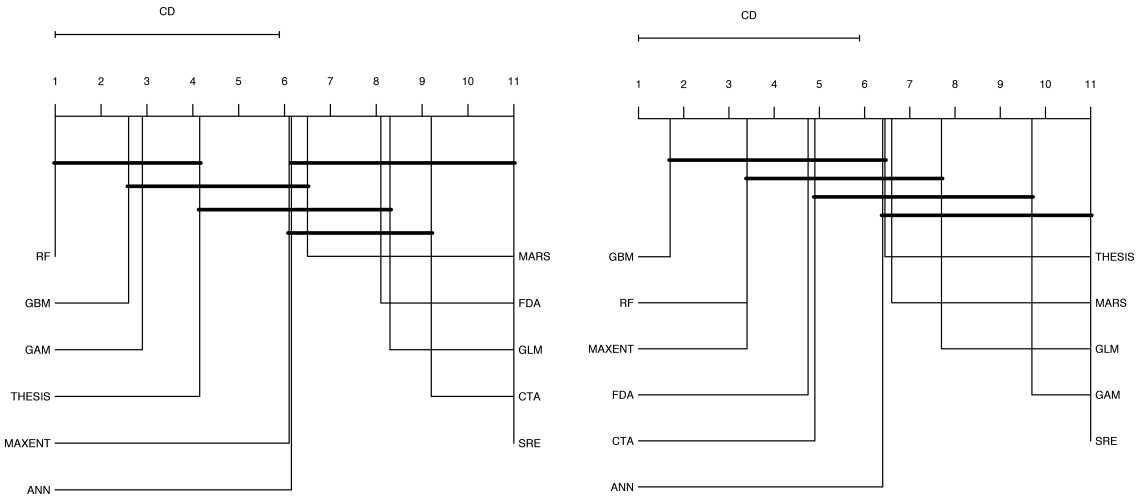
The results of the experiments are summarised with the use of box and whisker plots shown in Figure 23. After training with BIOMOD (default settings for all algorithms) the models are evaluated in two ways: on the held out validation data during CV, results depicted in the left **Validation** column of Figure 23, and on the test dataset, depicted in the right **Test** column of Figure 23. The first row of the figure shows the results obtained for the random sampling experiment, the second row displays the results for the stratified experiment.

One thing to notice is that the models trained in the stratified experiment are significantly (p-value < 2.2e-16) better performing on the evaluation datasets than those in the random experiment. This suggests that in the stratified experiment there is less complexity that needs to be captured by the algorithms to make good predictions. In addition, it suggests that the models score better due to less variety in the ecological factors between the training and evaluation data. As expected, this effect is reversed when the models are tested on unseen data. Models trained with the random data perform significantly better on the unseen data than those trained with the stratified dataset. This further suggests that the models obtained with stratified data are overfitted to the diagonal that they were calibrated on.

The models of the fold with the average highest TSS value for the stratified- and random datasets are projected and shown in Figure 21 and Figure 22 respectively. The most remarking result is that with a few exception, e.g., SRE, all projections look quite similar. Most variation in the maps seems to be in Mexico, at the south of the map. This observation also holds true when comparing the random- (Figure 22) with the stratified (Figure 24b) projection for the same algorithm. One thing that becomes apparent is that it is virtually impossible to pick the better map merely on a single number metric and without extensive knowledge of the species. For this reason, as discussed in Chapter 6, it is of the utmost important that maps are evaluated and hand picked by experts with extensive knowledge of the modelled species. Even with experts it is a difficult task according to Murray et al. (2009) as there was an obvious difference of estimates of distribution between experts from different regions as well as within regions and they concluded it might be dangerous to use only one or two experts.

The Friedman test was applied to the underlaying data of the graphs depicted in Figure 22 and 21. The test confirmed that there is a statistically significant difference in the TSS values obtained from the SDMs. To investigate this difference the post-hoc Nemenyi test was applied. The results of this test are graphically depicted in Figure 24 with a CD diagram. In the figure the SDMs are ranked with the best overall classifier at the left to the least performing classifier at the right. The CD of the Nemenyi test is indicated in the top left of the figure. When two classifiers differ less in their ranking than the CD

Figure 24 – Nemenyi critical difference diagrams comparing the overall performance of the THESIS method with other BIOMOD approaches with each other using the TSS score as performance and considering the random and stratified experiments. Tested on (a) CV validation with 50% of data used for testing, (b) tested on the off diagonal dataset. Average ranks of the examined SDMs are depicted. Bold lines indicate groups of SDMs which are not significantly different (at $p = 0.05$ their average ranks differ less than the Critical Distance (CD) value/distance indicated by CD line on the top.)



(a) Random (b) Stratified

Source: Author

then there is no significant difference with 95% confidence level ($p = 0.05$), else there is a significant difference between the two. SDMs connected with a bold horizontal line are closer to each other than the CD and thus do not differ significantly with 95% confidence interval.

5.2.4 Conclusion

The results of this and the first case study show that modelling algorithms behave differently between (virtual) species and that the ranking among the algorithm vary strongly for the two studies. A result that has been underscored in the literature, where it is has been shown that no method is superior in all circumstances (SEGURADO; ARAÚJO, 2004). All algorithms, but SRE, produced 'good' models for the random experiment, while less acceptable models for the stratified experiment. The projections generated by all models give the impression that they are potentially correct maps of distributions. Without placing the maps under critical examination it nearly impossible to select the map that best represents the species. The questions regarding this case study are answered as follows:

How do the SDM algorithms in BIOMOD and the THESIS algorithm perform on a real species?

Overall, the eleven SDMs showed good ability to predict the observed distributions for the random experiment, and close to good for the stratified experiments. Based on the CD diagram (Figure 24) the following observations are made:

- The SDM performance ranking trained on random data is from best to worst: RF, GBM, GAM, THESIS, MAXENT, ANN, MARS, FDA, GLM, CTA, and SRE (Figure 24a).
- The performance ranking of the models trained on the stratified data is from best to worst: GBM, RF, MAXENT, FDA, CTA, ANN, THESIS, MARS, GLM, GAM, and SRE (Figure 24b).
- Models generated by the THESIS algorithm do not show a significant difference (95% confidence level; $p = 0.05$) with the best ranked algorithms RF, GBM, and GAM on for the random experiment.
- The models trained by the THESIS algorithm do not show a significant difference (95% confidence level; $p = 0.05$) with the best ranked algorithms GBM, RF, MAXENT, FDA, CTA, and ANN for the stratified experiment.

How well do the generated models transfer to new unseen regions?

Unlike in the first case study, all algorithms have more difficulty projecting into the new unseen diagonal. The reason for this is likely due to the shape, nature and complexity of responses to the environmental factors. Even though these models individually do not have the desired accuracy and vary between their predictions that does not make them useless. There is increasing support for ensemble modelling (ARAÚJO et al., 2005; ARAUJO; CORREA, 2007) to combine and aggregate the result of several methods to reach a consensus, an already stablished technique outside of ecological modelling (DIETTERICH, 2000). In the literature it has been shown that ensembles significantly improve predictions (ARAÚJO et al., 2005; MARMION et al., 2009). Grenouillet et al. (2011) even goes as far as recommending not to use any single SDM for predictions, especially for species large environmental ranges.

Summarising, this case study suggests that the THESIS algorithm predicts comparable (95% confidence level; $p = 0.05$) to the top ranked algorithms in BIOMOD.

6 Model sharing

Contents

6.1 Introduction

So far in this thesis, the importance of datasets and Species Distribution Model (SDM) for benefit indicators and synthetic studies has been discussed in Chapter 1 and Chapter 2, where also a theoretical (BAM) framework for modelling ecological niches has been set out and the many complexities of errors and model validation. Chapter 2 also introduced SDMs that play a key role in integrated model-based assessments to understand and explore the complexity and the provision of scenario analyses to make well-informed political decision-making (SIEBER et al., 2010) and to reduce pressure upon biodiversity.

However, while not directly connected to the work presented thus far, true integration of biodiversity data requires theoretical deep models and hypotheses about the processes by which organisms evolve and interact with the environment and other species.

While working on the algorithm and learning about SDM I found it more and more important to consider to store and retrieve the models in some form. This for mainly three reasons; the first the development of the algorithm and the comparison with other algorithms has been troubling because of a lack of models and dataset to compare it with, in other words to perform a kind of benchmark to see where and how the the proposed algorithm operates well. Secondly, there is a significant loss of information and efforts undertaken by researches to not include the models, unlike point data that does not say much, a (published) model predicts the species density and provides a complete and fine-scale spatial coverage of potential distributions for an entire geographic range, and more importantly includes the evaluation of the model by typically an expert of the species. Third, the re-utilization of models of species can help build analytical tools to respond to questions about species interactions (PETERSON et al., 2010; SANTANA; SARAIVA, 2010). This not only increases our knowledge about interactions, but also likely increases the quality of the species predictions as species are only a part of a larger whole. Enabling the use of the models themselves to answer more complex scenario based analysis (STOCKWELL; PETERS, 1999). For this reason, I discovered for myself that this is perhaps the most important issue that should be focussed on and have therefore described in the chapter a possible temporary solution to the problem. Temporary because during the study I discovered that this problem warrants many other doctoral research projects and is an extremely complex, but worthwhile, problem to solve.

Reliable models for occurrences and richness of species assist policy-makers in meeting objectives and are used, for example, to value areas for nature conservation (PARVIAINEN et al., 2009). While there is a lot of discussion about what exactly is modelled and the correct applicable terminology, in essence, they forecast a part of the species niche (SILLERO, 2011). The models forecast the occupied niche when true absences are used from all suitable habitats to train the model, or the realised niche, when the samples do not cover all presence areas or when pseudo- or background absences are used. Knowing where a species occurs is a first step in making decisions.

There are several major problems in predictive modelling studies that affect model performance, e.g., incompleteness of data (ARAÚJO et al., 2005), dependence on the selection of the appropriate study region (ANDERSON; RAZA, 2010), the spatial resolution as landscape patterns and processes are scale-dependent and their response non-linear (YANG et al., 2011; WU et al., 2002), the training of models with sample locations from different temporal periods (ROUBICEK et al., 2010; STANKOWSKI; PARKER, 2010), the choice of threshold to convert probability values into presence or absence (NENZÉN; ARAÚJO, 2011), and the modelling method used best suited for the aim of the modelling (PETERSON, 2011; ELITH et al., 2006). Therefore, evaluate the usefulness of a model it is necessary for an ecologist with expert knowledge of the modelled species to cautiously interpret its predictions.

Policy questions can only be answered when cross-domain interactions are considered and not solely ecological models, as they are incomplete by themselves since they fail to account for important processes that influence extinction outcomes (KEITH et al., 2008). To maximise the benefits of the ecological models it is important to include other factors, such as: (i) costs (MURDOCH et al., 2007), (ii) the human socio-economic system (YUE; JORGENSEN; LAROCQUE, 2011), (iii) demographic and landscape dynamics (WINTLE et al., 2005), and (iv) spatio-temporal interactions and animal movement (LAROCQUE et al., 2011). Considering the broad spectrum of possible cross-scale interactions it is likely that the required expertise goes beyond that of ecologists. To make the fullest use of the scientific value of the generated and evaluated models by ecologist it is necessary that models are available to other scientists and not just the underlaying data. To not share those is an inefficient use of funds and a loss of opportunity to accelerate breakthrough research with an impact beyond ecology.

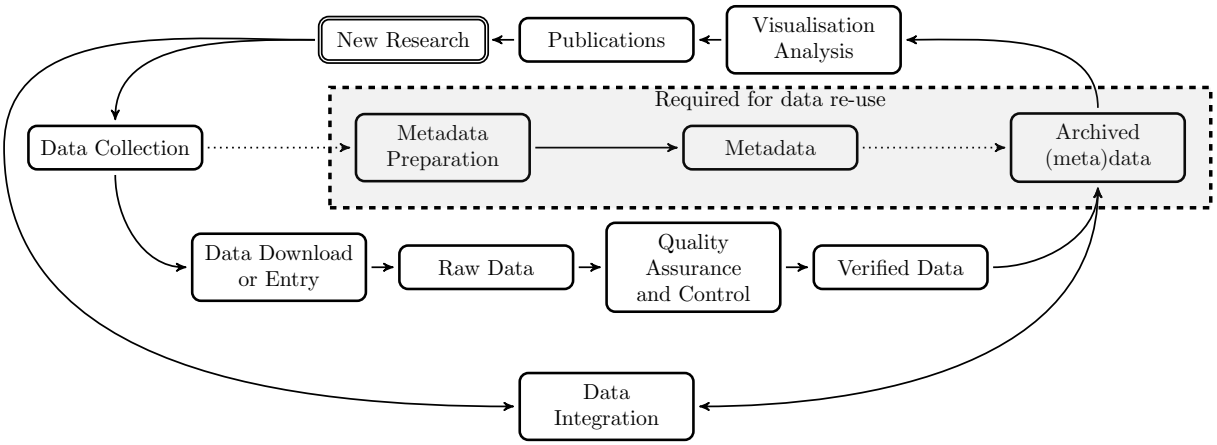
This chapter stresses the need to share and re-use of SDMs through the use of an ecosystem for modelling species distributions. An architecture of a system is described that is comprised of readily available cloud computing technologies and the R language (R Development Core Team, 2006) to enable meaningful storage, retrieval and publishing of data models in a seamless automated fashion in an attempt to achieve integrated environmental modelling. A generic format for model encoding or a specific infrastructure is not proposed nor required for this to work as the infrastructure primarily functions as a general storage method. The main idea behind this conceptual framework is to provide researchers a place to collaborate, share data and to utilise earlier published models. This way model outputs are turned into conservation management tools to allow policy makers to manage the loss of biodiversity in the region. Therefore, turning enabling the use of model outputs for more than just static geographical maps.

The remainder of this chapter is organised as follows: first, other solutions are presented that relate to this problem and that contextualises this chapter in Section 6.2. The requirements of the proposed ecosystem are described in Section 6.3. Section 6.4 provides the overall design and architecture of the framework and describes the high-level details of all components. Since the proposed solution is conceptual, the last section of this chapter will discuss potential problem areas and suggestions how to implement the proposed solution.

6.2 Model sharing

This section will describe briefly four different aspects of the proposed solution, namely: (i) the standardisation of data, (ii) the utilisation of cloud computing technologies, (iii) the Species Distribution Modelling ecosystems themselves, and (iv) the data sharing aspect,

Figure 25 – A view of the data life cycle



Source: Adapted from Michener et al. (2011)

which is the main reason for this ecosystem.

6.2.1 Data Life Cycle

Figure 25 illustrates a view of the data life cycle.Two cycles are shown that lead to new research questions and knowledge. The inner cycle starts with data collection and involves acquisition, data exploration, quality control, analysis and visualisation and, in the end, a publication. The outer cycle starts with integration of archived data and extracting knowledge. The latter cycle facilitates reproducibility and multi-disciplinary collaborations: opening the way to new scientific insights by testing new or alternative hypotheses and methods of analysis and exploration of topics not envisioned by the original researcher (WRUCK; PEUKER; REGENBRECHT, 2014).

Re-usable archived data and metadata is a requisite for the outer cycle. To catalyse scientific progress the data needs an accurate description and be publicly available, illustrated in Figure 25 by the steps inside the dashed box: (i) metadata preparation, (ii) metadata, and (iii) archived (meta) data.

Collectively ecologists produce an enormous amount of data, however only a fraction is shared. The majority works individually and pursues constrained spatial and temporal scales with limited resources for storage, analysis and sharing of data (HEIDORN, 2008). In a study of a hundred randomly selected articles produced by projects funded through NSF's Division of Environmental Biology (2005-2009) only 43% of the papers that produced data shared some or all it (HAMPTON et al., 2013). Nonetheless, 83% of that shared data is related to genetics. Only 8% made non-genetic data available. The fact is, even if more models are published they will only be useful if their access and reuse is easy for all that are involved.

6.2.2 Standardisation of data

Computer science forms a central role in creating new models for the way results are published. As it is common to include references in articles, so it should be routine to include complete computational methods. These results should be extensively tested, cross-referenced and encoded in community-supported and standardised formats. Modelling standards should allow and publish data in an automated way to obtain integrative models as just quantifying losses of biodiversity will not be enough for policy change and loss reduction (BALMFORD et al., 2005).

There are other initiatives that aim to improve this, for example: MetaCat (BERKLEY et al.,), EcoTrend (SERVILLA et al., 2008), LifeWatch (HERNÁNDEZ ERNST; POIGNÉ; LOS, 2010), WBCMS (FOOK, 2009), DataOne and the EML and DarwinCore standards. However, these initiatives do not specifically address the standardisation of the model representation as for example the Systems Biology Markup Language (SBML) does for systems biology (HUCKA et al., 2003).

Even though a standard, such as SBML, for SDMs is not defined, and defining such as standard goes far beyond the scope of this work, the field has a *de facto* SDM standard in BIOdiversity MODelling (BIOMOD) (THUILLER et al., 2009) that supports many if not all popular SDM algorithms, such as MaxEnt (PHILLIPS; ANDERSON; SCHAPIRE, 2006), GARP (STOCKWELL; PETERS, 1999) and Random Forests. BIOMOD allows for the storage and retrieval of earlier created models and, consequently, provides a *de facto* standard to share those models. For that reason, this framework takes advantage of R and BIOMOD to save the models and to document the process by which they are generated and utilised.

6.2.3 Species Distribution Modelling ecosystems

For SDM voluminous bionomic and scenopoetic data are used to discover meaningful insights into the potential niche of the species under study. These insights are revealed through modelling and analysis using machine learning and statistical methods. Various techniques and algorithms are developed specifically with SDMs in mind, focusing on data sets where limited data is available and where real species absence points are almost non-existent due to the labor intensity of obtaining them by biologists.

The proposed SDM ecosystem supports managing, integrating and visualising of the models and their relevant data. The greatest challenges of such an ecosystem is dealing with: the complex and heterogeneous nature of the data, the lack of common standards and the difficulty to obtain funding for long term storage, as that is only sensible if data re-use is proven to work.

The short and limited scope project carried out in this research does not try to solve these problems nor tries to define a standard for workflow, model or meta-data.

Instead, the focus is on version control, citability and reproducibility. The platform does not achieve this by imposing an accepted format for important meta-data, but rather by providing a platform where models are created on-line in a similar fashion as is now customarily done by researchers on their own desktop computers. As an initial step the proposed solution provides a web-based R front-end and a repository to save and retrieve the models.

6.3 SDM framework

The Species Distribution Modelling Framework for an SDM Ecosystem should consider the necessities discussed in the remainder of the section.

6.3.1 Cloud based architecture

A unified cloud based architecture for SDMs provides a single, centralised, consistent, reusable and reproducible modelling service that enables research validation and integration. The cloud brings a self-scalable global access to on-demand long running computing resources such that individual researchers do not have to worry about budgeting for dedicated machines. In principle lowering the overall cost for all that are involved as computing resources are only sporadically required per user.

Researchers access models stored in the cloud and share their own models to collaborate with other scientists for further synthetic studies. It permits others to verify published results of past investigation and to extend and integrate and generalise the models into new unexplored areas.

The cloud should be the preferred way to store long-term generated models instead of on private hard drives or in local repositories. Studies have shown (CAETANO; AISENBERG, 2014; HEIDORN, 2008) that local data is often lost from disuse and that in fact only about eight percent of articles have their data available upon direct request after a period of just five years. This platform needs to facilitate a way to include the full computational methods used in publications. Others can experiment with those methods while reading along with the article.

6.3.2 Data from repositories

Biodiversity data with detailed information for each data set laboriously compiled and made available by and through organisations such as the Global Biodiversity Information Facility (GBIF; *http://www.gbif.org/*) and the Atlas of Living Australia (ALA; *http://www.ala.org.au/*) must be capitalised on by the framework as this data is an essential ingredient for the model building process.

6.3.3 Data quality

Even though GBIF and ALA provide access to large data sets, the compilation and cleaning of that data for building the model is still a very time consuming process. Cleaning and improving the quality of the data requires a profound understanding of the species over large taxonomic, spatial- and temporal scales. While a complete data refinement workflow, as for example discussed in Mathew et al. (2014), is not enforced and outside the scope of this work. The platform merely ought to address the need to store the result of the data cleaning process and the identification and exclusion of erroneous or irrelevant records so that other models are trained and compared with the exact same records thus making the model creation reproducible.

6.3.4 Analytics engine

The projection of models over large geospatial areas, which requires large volumes of static scenopoetic data together with ever-changing species distribution as they continuously are gathered by biologists all over the world, benefits from cloud computing principles to overcome the high computational efforts needed to build, project and analyse models. The analytics engine is cloud based and allows for the projection by using proven supervised data learning algorithms. By keeping the analytic engine open for public changes, computer- and other scientists will have a framework to develop and test new algorithms and strategies. Eventually, the platform is used to write easily model algorithm benchmarks to evaluate new algorithms to determine their quality, since data sets and predicted model outcomes will be verified and possibly peer reviewed.

6.3.5 Interface dashboard

An internationally accessible interface to all involved and interested in the model predictions should be made available, such that research institutes, government agencies and also the general public can utilise the platform. From the perspective of the researcher the cloud based interface provides a way to access model predictions and study the environmental impact on their distributions. The interface facilitates a way to reproduce, validate and analyse earlier published models, opening up the niche models for conservation planning: to give attention and using valuable limited resources for those species and places that need it the most. In addition, through the interface the models can be used for integrated model-based assessments to understand and explore the complexity and the provision of scenario analyses.

Beyond that, the framework should address the reasons why researchers do not share their data, such as: (i) the time needed to prepare data (SWAN; BROWN, 2008), (ii) the loss of control of data (TENOPIR et al., 2011), (iii) not receiving credit and acknowledg-

ment (GROTH; GIBSON; VELTEROP, 2010), (iv) data taken out of context (BECHHOFER et al., 2013), and (v) highly sensitive and/or restricted for public access data, e.g., extinct species (MEIJAARD; NIJMAN, 2014) and (vi) licensing concerns (KLUMP et al., 2006).

6.4 System architecture

The proposed system architecture makes effective use of a cloud based architecture by providing global access, centralised data storage and self-management for processing resources for its users. Figure 26 shows the overview of the SDM framework in a cloud based SDM ecosystem. The architecture is composed of several modules that interact together to provide the required functionality. The following sections describe the following modules in more detail:

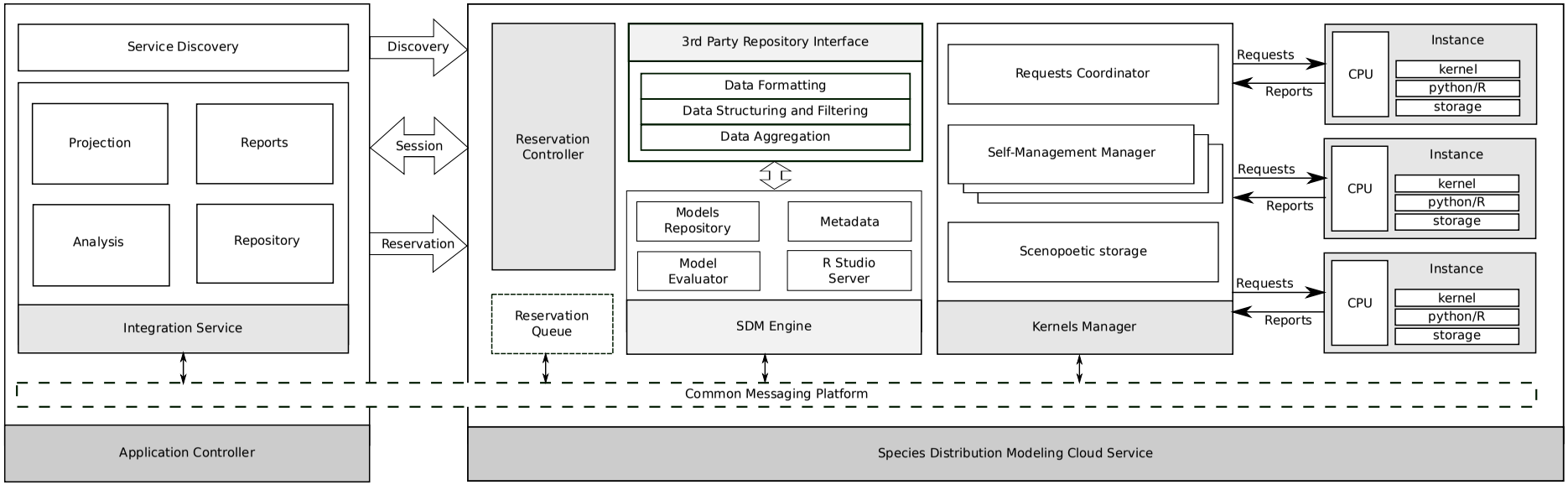
- The Application Controller
- Third Party Repository Interface
- Species Distribution Modelling Cloud Service
  - Reservation Controller
  - SDM Engine
  - Kernels Manager
  - Processing Instances

6.4.1 The application controller

The primary functions of the application controller is to control and direct user requests to the appropriate modules and to provide the web service with the required information to access the models. It is responsible for storing and serving static projections, reports and model quality analysis. The controller synchronises information on load and resource availability with the reservation controller to enable users to track the status of the system.

The application controller enforces modularity by dispatching the requests by sending and receiving messages to the other modules and instances, so to exploit the elasticity of the cloud architecture by increasing the number of active instances proportional to the amount of incoming requests. A single web interface forwards user commands to virtual instances that are allocated by the kernel manager. Returned messages are identified by their host id and displayed in the appropriate user view of the SDM environment. This lays a foundation for a more robust design where the clients and servers are independent modules and may fail separately without one failure affecting others in the system.

Figure 26 – Architecture overview



Source: Author

6.4.2 Third party repository interface

The technical interoperability to obtain heterogeneous data from different sources is complex. The Third Party Repository Interface module provides an internal standardised access to species repositories (e.g., GBIF and ALA) to obtain occurrences with a programmatic interface that is made available by existing R packages in the modelling environment.

6.4.3 Species Distribution Modelling Cloud Service

The SDM cloud service consists out of several components that all support the SDM Kernel. The reservation controller's purpose is requesting and allocating resources, the SDM engine's to maintain the model repository, and lastly the kernels manager function is to self-manage the required number of instances to support parallel task based execution of the models.

6.4.3.1 Reservation Controller

The reservation controller provides an interface to request resource reservations for the SDM kernel application. Its main service is to provide contract negotiation for the required virtual resources for a given time frame. The aim is to implement a resource reservation scheduling algorithm that supports the execution of the SDM algorithms. However, it tries to limit the number of active instances as they are billed per hour and also incur a setup cost and time, delaying the execution of the process. Therefore, not every reservation request should directly result in an immediate launch of a virtual instance.

6.4.3.2 SDM Engine

The Species Distribution Modelling Engine supplies the storage for commonly used scenopoetic data for the reason that researchers often use the same voluminous data set, e.g., the WorldClim Global Climate Data set (HIJMANS et al., 2005), to build their models. However, other data may be uploaded directly into the Cloud Storage.

The researcher has a choice of various open source software packages and modelling algorithms by using BIOMOD (THUILLER et al., 2009) in a general R-Studio Server (setup to create models, or even ensembles of models); just as in any other R instance that is run in a local desktop environment.

6.4.3.3 Kernels Manager

When the reservation controller processes the request information it checks whether the requested resources are currently already available. Each solicited resource is then com-

pared with all instances that are currently active to find an instance that matches all the requested features. Typical features are: architectures, instance sizes and the available number of CPUs and GPUs. If no matching instance is found by the Kernels Manager it will bare-metal provision a new instance, for example: booting the machine to a fresh system image, attaching volumes and configuring the required packages. The new instance is added to the resource inventory and its lifetime subsequently managed by the kernels manager. A confirmation message is returned and a user is granted access. In this way the framework acts in a similar fashion as serving ordinary thin client, thus allowing multiple user applications to run in separate and parallel processes.

6.4.3.4 Processing Instances

The processing instances, alike the other modules, are build on top of a cloud computing service that provides scalable on-demand usage-based ”pay as you go” compute capacity. Each instance makes use of the fixed and removable storage services.

All instances are preconfigured identically with both the Python and R packages available and with access to the shared storage and the scenopoetic datasets maintained by the kernel manager. It is on these virtual instances that the actual SDMs are developed, executed and analysed. Models can be saved to long term storage and retrieved at a later time by either the same user or by any other researcher if the model is shared or made publicly available. Thus providing a way to reproduce data cleaning, model construction and projection at a later time for re-evaluation or to be build upon for new research.

6.5 Final Remarks

This chapter discussed SDM and model sharing for biodiversity ecosystems. It briefly touches the requirements, design aspects and possible architecture to enable research synthesis and utilise the generated models in a broader context and leveraging cloud-computing, self-management by auto-scaling the required resources in the cloud for addressing many issues in traditional non-centralised modelling systems. A system of this type can bring the following benefits that currently are not available:

- a platform for model storage for various species using well known *de facto* standards. Model developers using the proposed system would benefit from this interoperability for data sharing.
- a simple yet effective structured approach to store large amounts of models and layer data using existing technologies for querying and managing the large data sets in a data warehouse solution. Allowing the proposed solution to scale dynamically as it is build on cloud technologies making it easy to increase storage and computational power on demand.

- an interface for SDM that allows generic and unstructured data from various data providers to be accessed and processed in the system. Data stored in such records is easily visualised in solutions such as Google Maps for further investigation.
- a modelling engine that is based on the BIOMOD package and consequently implements various SDM modelling algorithms on the cloud using python and R. With the proposed framework in many cases there is no need to further analyse the models themselves as they will already be evaluated by expert ecologist and cross-referenced to a publication; making them available to be used as an integral part for new composite models and further interdisciplinary research.

Most importantly, a tremendous benefit would be to no longer allow SDM to go in disuse after they are used for publication and as a consequence potentially open up new areas of research.

7 Conclusions

Contents

This thesis proposed a new algorithm to create models of species niches to identify and predict species distributions. Two methods have been combined to present a hybrid solution that utilises both linear genetic programming and fuzzy rule systems. This approach differs from the other models present in BIOMOD in that it is based on a process that mimics biological evolution. The algorithm continuously adapts individual solutions within a population of possible solutions. During each generation, or adaptation round, the solutions "evolve" toward better and more optimal models. This chapter re-visits the results and examines them in light of the original thesis that was stated in Chapter 1. In addition, the chapter discusses the strengths and weaknesses of the new algorithm. It concludes with an exploration of possible future directions for research.

7.1 Thesis revisited

This section reviews the thesis stated in Chapter 1, verifying that it has been addressed.

> *Thesis:* Not just ecological models, hybrid fuzzy - evolutionary models in extendible environments allow for better predictions.

The results of the two case studies presented in this thesis confirm that the implementation of the objective (Section 3.1) regarding a hybrid of fuzzy and evolutionary algorithm produces models that are able to predict the complex non-linear relationship between environmental factors and species presence and absences. Whether these models perform better than other methods will depend on the species under study; regarding the first case study, the algorithm produces significantly ($p = 0.05$) better models that transfer well to unseen regions. With respect to the second case study, the algorithm has not produced better models. However, the generated models were significantly ($p = 0.05$) comparable with the best ones produced. In addition, the algorithm has shown a linear speedup on an cloud computing solution in Section 4.6.

With regard to the objectives mentioned in Section 3.2 the needed modifications are published on GitHub [1] and involves: the BIOMOD package, the source code of the algorithm, and the compiled command line executable that implements the algorithm. A general approach to add algorithms to BIOMOD should be standardised to easily verify new algorithms. The source code is made available under the GPLv2 license (Free Software Foundation, June 1991), permitting commercial use, modification, distribution of the software, without being held liable; in accord with the philosophy of Chapter 6. As a result the experiments described in this thesis are reproducible, although with slightly different results due to the heuristic nature of the algorithm.

As stated previously in Section 2.3, SDMs are used to shape policies for conservation planning. Arguably, increased accuracy and higher quality predictions will result in more effective maps to understand the likely changes in species distributions based on, for example, future emission scenarios such as those defined by Working Group III of the Intergovernmental Panel on Climate Change (op. 2000). Concluding, the thesis and all objectives to demonstrate it have been met.

7.2 Strengths of this approach

The nature of linear genetic programming and fuzzy rule based system make this approach particularly suited for noisy occurrence records and the lack of true absence data, as previously seen in Section 5.1. For the analysis of species distribution modelling to be useful under these circumstance the modelling algorithm is required to handle this well.

The robust nature of the algorithm has been demonstrated by the application of both case studies. While the algorithm does not transfer as well in the second case study, it does not perform significantly worse than other conventional modelling algorithms. The strength of this approach is that it is not based on ideal mathematical models, but instead is problem agnostic. Genetic Programming does not 'know' anything about species distributions, nor the environmental factors that affect them. In fact, no explicit domain knowledge to achieve the goal is used to create the models. Instead, the algorithm simply tries to find this relationship, in this case with symbolic regression, and successively tries better iterations to get improved results. A strength of this approach is therefore that it can be used to create small programs that describe the niche of a species.

The genetic programming approach, even more so with fuzzy rules, differs from many other algorithms of modelling species niches in that it operates by doing a probabilistic search and keeping a diverse set of solutions, that might even be contradictory, to solve the problem. This way the technique does not rely on greedy hill climbing to move from one point to the more optimal in search space. Greedy hill climbing may work well

[1] <https://github.com/michelbieleveld/genetic-fuzzy>

for simple problems, but for non-trivial problems where certain areas in search space are inaccessible to hill climbing this will not work. As a consequence of the probabilist search, the algorithm can spend some portion of its population on sub optimal solutions, while it searches with another portion in different more adventurous areas of the search space.

7.3 Weaknesses of this approach

A general weakness of all approaches that use species distribution modelling is that there are many, many details that need clarification and testing. Sample size, effects of scale, time and resolution, whether model projections are valid for future predictions, sample efforts that can introduce bias, optimal parameters for each algorithm, and how are all these aspects, and many more, are relevant for the training and application of the SDM. Unfortunately, but not unexpected, the work presented in this thesis also does not have an answer to those questions for the proposed algorithm. It's main weakness, as with the other algorithms, is that there is no clear guideline when to use this algorithm as is the case for all algorithms used in SDM. The main reason being that it will depend on the application and species under study. The famous No Free Lunch theorem by Wolpert and Macready (1997) states that any two optimisation algorithms are equivalent when their performance is averaged across all possible problems. A result of this theorem is that if an algorithm is better at one thing, it must be worse at something else. Based on the limited results of just two case studies, the discussed algorithm seems to perform less well when interpolating.

A major weakness in the algorithm is that its search is computationally intensive. While by nature the algorithm is embarrassingly parallel when using multiple demes and shards or even the evaluation of fitness of the individuals, as shown in Section 4.6, but that does not negate the fact that a lot of processing power is required to go through search space. Building a single model for the species in the case studies, often took in the order of ten minutes on a personal computer to build ten models for the cross validation using serially a single core and up to 30 seconds when using the cluster with twenty nodes. However, when in the end a suitable model is found little processing power (in the order of tens of seconds) is required to apply this model to the data (in the order of giga bytes of climate data) to make projections.

7.4 Future directions

Perhaps, now that many other aspects are in place and are proven to be highly useful, the most important improvement needed for species distribution models is not the algorithms, nor the quality of the data points (as shown there is no significant difference

up to 30% error in the virtual species case study) or making presence points available to the public with solutions such as GBIF, but the sharing of SDMs, verified by experts, that can be used for research synthesis and policy making. Ideally, shared through a system of sorts as discussed in Chapter 6, but any framework that makes the models themselves available and citable just like other data will enable research synthesis.

Other future directions in this particular research area of data sharing involves defining formal model representations that are useful for the visual presentation and to allow the models to be computable to automatically simulate and analyse them, similar as is done with the Systems Biology Markup Language. Defining such as language that defines a common intermediate format will enable: (1) researchers to define algorithms and use them without rewriting the tools, and vice-versa, (2) the sharing and publishing of the models for reproducible and synthesis research, and (3) the survival of the generated models for current publications beyond the lifetime of submitting those publications.

Many steps in the algorithms presented here, both for those present in BIOMOD and the algorithm discussed in this thesis, use parameters with ad hoc set values. To understand the effect of those values, good training datasets are needed to identify and measure specific impacts of parameter changes. Optimising these parameters will require a de facto standardised benchmark for species distribution modelling, similar as the *proben1* benchmark (PRECHELT et al., 1994) for neural network learning. This benchmark should contain several real world species with various degrees of prevalence and data for specific applications that either require transferability or interpolation.

Genetic Programming and Fuzzy Rule Systems may prove useful in incorporating other sources of data, such as species relationships and population densities. Algorithms such as discussed in this thesis are problem agnostic and optimise for various non-linear problems. There is good promise in applying this approach to these new sources of data, as it is relatively easy to adapt the instructions of genetic programming to include more complex relationships based on expert knowledge.

Finally, future research in species distribution modelling algorithms and the modelling process should start focussing on how these tools are useful for conservation planning research to support management, and not just to generate maps. This could open up decision support tools to use species distribution maps to identify areas that are suitable for conservation based on a set of biodiversity metrics for minimal economic and social cost. It is possible that the properties of the discussed algorithm can best identify a particular set of target metrics due to the inherent multi-objective process of genetic programming. Future research can consider this question as well.

Bibliography

ACOSTA, A. L. *Bombus terrestris chegará ao Brasil? Um estudo preditivo sobre uma invasão em potencial*. PhD thesis (PhD) — Universidade de São Paulo, São Paulo, 2015. Available from Internet: <http://www.teses.usp.br/teses/disponiveis/41/41134/tde-22092015-080256/>.

ACOSTA, A. L. et al. Worldwide alien invasion: A methodological approach to forecast the potential spread of a highly invasive pollinator. *PloS one*, v. 11, n. 2, p. e0148295, 2016. ISSN 1932-6203. http://doi.org/10.1371/journal.pone.0148295.

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, v. 19, n. 6, p. 716–723, 1974. ISSN 0018-9286. http://doi.org/10.1109/TAC.1974.1100705.

ALBA, E.; DORRONSORO, B. *Cellular genetic algorithms*. Berlin: Springer, 2008. ORCS 42. (Operations research/computer science interfaces series, ORCS 42). ISBN 9780387776101.

ALLOUCHE, O.; TSOAR, A.; KADMON, R. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (tss). *Journal of Applied Ecology*, v. 43, n. 6, p. 1223–1232, 2006. ISSN 00218901. http://doi.org/10.1111/j.1365-2664.2006.01214.x.

ALTENBERG, L. The schema theorem and price's theorem. In: WHITLEY, L. D.; VOSE, M. D. (Ed.). *Foundations of Genetic Algorithms 3*. San Francisco, Calif.: Morgan Kaufmann, 1995, (Foundations of genetic algorithms, v. 3). p. 23–49. ISBN 978-1558603561.

ANDERSON, R. P.; LEW, D.; PETERSON, A. Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modelling*, v. 162, n. 3, p. 211–232, 2003. ISSN 03043800. http://doi.org/10.1016/S0304-3800(02)00349-6.

ANDERSON, R. P.; RAZA, A. The effect of the extent of the study region on gis models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus nephelomys) in venezuela. *Journal of Biogeography*, v. 37, n. 7, p. 1378–1393, 2010. ISSN 03050270. http://doi.org/10.1111/j.1365-2699.2010.02290.x.

ARAUJO, J.; CORREA, P. A framework for species distribution modeling: A performance evaluation approach. *Proceedings of the 6th International Information and Telecommunication Technologies Symposium*, p. 111–118, 2007.

ARAÚJO, M. B.; GUISAN, A. Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, v. 33, n. 10, p. 1677–1688, 2006. ISSN 03050270. http://doi.org/10.1111/j.1365-2699.2006.01584.x.

ARAÚJO, M. B. et al. Reducing uncertainty in projections of extinction risk from climate change. *Global Ecology and Biogeography*, v. 14, n. 6, p. 529–538, 2005. ISSN 1466822X. http://doi.org/10.1111/j.1466-822x.2005.00182.x.

ARIÑO, A. H. Approaches to estimating the universe of natural history collections data. *Biodiversity Informatics*, v. 7, n. 2, 2010. http://doi.org/10.17161/bi.v7i2.3991.

ARLOT, S.; CELISSE, A. A survey of cross-validation procedures for model selection. *Statistics Surveys*, v. 4, n. 0, p. 40–79, 2010. ISSN 1935-7516. http://doi.org/10.1214/09-SS054.

AUSTIN, M. P. Searching for a model for use in vegetation analysis. *Vegetatio*, v. 42, n. 1-3, p. 11–21, 1980. ISSN 0042-3106. Available from Internet: <http://dx.doi.org/10.1007/BF00048865>. http://doi.org/10.1007/BF00048865.

AUSTIN, M. P. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, v. 157, n. 2–3, p. 101–118, 2002. ISSN 03043800. http://doi.org/10.1016/S0304-3800(02)00205-3.

AUSTIN, M. P. et al. Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory. *Ecological Modelling*, v. 199, n. 2, p. 197–216, 2006. ISSN 03043800. http://doi.org/10.1016/j.ecolmodel.2006.05.023.

BABYAK, M. A. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, v. 66, n. 3, p. 411–421, 2004. ISSN 0033-3174. http://doi.org/10.1097/01.psy.0000127692.23278.a9.

BAHN, V.; MCGILL, B. J. Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography*, v. 16, n. 6, p. 733–742, 2007. ISSN 1466822X. http://doi.org/10.1111/j.1466-8238.2007.00331.x.

BALL, I. R.; POSSINGHAM, H. P.; WATTS, M. E. Marxan and relatives: Software for spatial conservation prioritization. In: MOILANEN, A.; WILSON, K. A.; POSSINGHAM, H. P. (Ed.). *Spatial conservation prioritization*. Oxford [England] and , New York: Oxford University Press, 2009. p. 185–195. ISBN 978-0-19-954777-7.

BALMFORD, A. et al. The 2010 challenge: data availability, information needs and extraterrestrial insights. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, v. 360, n. 1454, p. 221–228, 2005. ISSN 0962-8436. http://doi.org/10.1098/rstb.2004.1599.

BARBET-MASSIN, M. et al. Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, v. 3, n. 2, p. 327–338, 2012. ISSN 2041210X. http://doi.org/10.1111/j.2041-210X.2011.00172.x.

BÁRDOSSY, A.; DUCKSTEIN, L. *Fuzzy rule-based modeling with applications to geophysical, biological, and engineering systems*. Boca Raton, FL: CRC Press, 1995. (Systems engineering series). ISBN 9780849378331.

BARRY, S.; ELITH, J. Error and uncertainty in habitat models. *Journal of Applied Ecology*, v. 43, n. 3, p. 413–423, 2006. ISSN 00218901. http://doi.org/10.1111/j.1365-2664.2006.01136.x.

BARVE, N. et al. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, v. 222, n. 11, p. 1810–1819, 2011. ISSN 03043800. http://doi.org/10.1016/j.ecolmodel.2011.02.011.

BECHHOFER, S. et al. Why linked data is not enough for scientists. *Future Generation Computer Systems*, v. 29, n. 2, p. 599–611, 2013. ISSN 0167739X. http://doi.org/10.1016/j.future.2011.08.004.

BEGON, M.; TOWNSEND, C. R.; HARPER, J. L. *Ecology: from individuals to ecosystems*. 4th. ed. Malden, MA: Blackwell Pub., 2006. ISBN 978-1405111171.

BELDING, T. C. The distributed genetic algorithm revisited. In: *Proceedings of the 6th International Conference on Genetic Algorithms*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 1995. p. 114–121. ISBN 1-55860-370-0.

BELLWOOD, D. R. et al. Confronting the coral reef crisis. *Nature*, v. 429, n. 6994, p. 827–833, 2004. ISSN 00280836. http://doi.org/10.1038/nature02691.

BERKLEY, C. et al. Metacat: a schema-independent xml database system. p. 171–179. http://doi.org/10.1109/SSDM.2001.938549.

BirdLife International and NatureServe. *Bird species distribution maps of the world. BirdLife International, Cambridge, UK and NatureServe, Arlington, USA*. 2015. Available from Internet: <http://www.birdlife.org>.

BLAGODEROV, V. et al. No specimen left behind: industrial scale digitization of natural history collections. *ZooKeys*, n. 209, p. 133–146, 2012. ISSN 1313-2970. http://doi.org/10.3897/zookeys.209.3178.

BLUM, A. L.; RIVEST, R. L. Training a 3-node neural network is np-complete. *Neural Networks*, v. 5, n. 1, p. 117–127, 1992. ISSN 08936080. http://doi.org/10.1016/S0893-6080(05)80010-3.

BORRA, S.; CIACCIO, A. D. Measuring the prediction error. a comparison of cross-validation, bootstrap and covariance penalty methods. *Computational Statistics & Data Analysis*, v. 54, n. 12, p. 2976–2989, 2010. ISSN 01679473. http://doi.org/10.1016/j.csda.2010.03.004.

BRAMEIER, M.; BANZHAF, W. A comparison of linear genetic programming and neural networks in medical data mining. *IEEE Transactions on Evolutionary Computation*, v. 5, n. 1, p. 17–26, 2001. ISSN 1089778X. http://doi.org/10.1109/4235.910462.

BRAMEIER, M.; BANZHAF, W. *Linear genetic programming*. New York: Springer, 2007. (Genetic and evolutionary computation series). ISBN 978-0387-31029-9.

BREIMAN, L. *Classification and regression trees*. New York: Chapman & Hall, 1993. ISBN 978-0412048418.

BREIMAN, L. et al. *Classification and regression trees*. [S.l.]: CRC Press, 1984.

BROTONS, L. et al. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, v. 27, n. 4, p. 437–448, 2004. ISSN 09067590. http://doi.org/10.1111/j.0906-7590.2004.03764.x.

BUCKLIN, D. N. et al. Comparing species distribution models constructed with different subsets of environmental predictors. *Diversity and Distributions*, v. 21, n. 1, p. 23–35, 2015. ISSN 13669516. http://doi.org/10.1111/ddi.12247.

BUSBY, J. R. Bioclim - a bioclimatic analysis and prediction system. In: *Nature Conservation: cost effective biological surveys and data analysis*. [S.l.: s.n.], 1991. p. 64–68.

CAETANO, D. S.; AISENBERG, A. Forgotten treasures: the fate of data in animal behaviour studies. *Animal Behaviour*, v. 98, p. 1–5, 2014. ISSN 00033472. http://doi.org/10.1016/j.anbehav.2014.09.025.

CALVO, B.; SANTAFE, G. scmamp: statistical comparison of multiple algorithms in multiple problems. *The R Journal*, Accepted for publication, 2015.

CANTÚ-PAZ, E. A survey of parallel genetic algorithms. *CALCULATEURS PARALLELES, RESEAUX ET SYSTEMS REPARTIS*, v. 10, 1998.

CANTÚ-PAZ, E. Migration policies, selection pressure, and parallel evolutionary algorithms. *Journal of Heuristics*, v. 7, n. 4, p. 311–334, 2001. ISSN 13811231. http://doi.org/10.1023/A:1011375326814.

CARNEIRO, L. R. et al. Limitations to the use of species-distribution models for environmental-impact assessments in the amazon. *PloS one*, v. 11, n. 1, p. e0146543, 2016. ISSN 1932-6203. http://doi.org/10.1371/journal.pone.0146543.

CAVICCHIO, D. J. *Adaptive search using simulated evolution.* PhD thesis (PhD) — University of Michigan, Ann Arbor, 1970. Available from Internet: <http://hdl.handle.net/2027.42/4042>.

CHAPMAN, A. D. *Principles and methods of data cleaning: primary species and species-occurrence data.* version 1.0. Copenhagen: Global Biodiversity Information Facility, 2005.

CHAPMAN, A. D.; MUÑOZ, M. E.; KOCH, I. Environmental information: placing biodiversity phenomena in an ecological and environmental context. *Biodiversity Informatics*, v. 2, n. 0, 2005. http://doi.org/10.17161/bi.v2i0.5.

CHASE, J. M.; LEIBOLD, M. A. *Ecological niches: Linking classical and contemporary approaches.* Chicago: University of Chicago Press, 2003. (Interspecific interactions). ISBN 9780226101804.

CHATFIELD, C. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, v. 158, n. 3, p. 419, 1995. ISSN 09641998. http://doi.org/10.2307/2983440.

CHRISTOPHE, E.; MICHEL, J.; INGLADA, J. Remote sensing processing: From multicore to gpu. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, v. 4, n. 3, p. 643–652, 2011. ISSN 1939-1404. http://doi.org/10.1109/JSTARS.2010.2102340.

COLWELL, R. K.; RANGEL, T. F. Hutchinson's duality: the once and future niche. *Proceedings of the National Academy of Sciences of the United States of America*, v. 106 Suppl 2, p. 19651–19658, 2009. ISSN 1091-6490. http://doi.org/10.1073/pnas.0901650106.

CORDON, O.; HERRERA, F.; VILLAR, P. Generating the knowledge base of a fuzzy rule-based system by the genetic learning of the data base. *IEEE Transactions on Fuzzy Systems*, v. 9, n. 4, p. 667–674, 2001. ISSN 10636706. http://doi.org/10.1109/91.940977.

COX, D. L. A note on the queer history of "niche". *Bulletin of the Ecological Society of America*, v. 61, n. 4, p. 201–202, 1980. ISSN 0012-9623.

CRUZ, B. B.; TESHIMA, F. A.; CETRA, M. Trophic organization and fish assemblage structure as disturbance indicators in headwater streams of lower sorocaba river basin, sã£o paulo, brazil. *Neotropical Ichthyology*, v. 11, n. 1, p. 171–178, 2013. ISSN 1679-6225. Available from Internet: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1679-62252013000100171&nrm=iso>.

DAGUM, L.; MENON, R. Openmp: an industry standard api for shared-memory programming. *IEEE Computational Science and Engineering*, v. 5, n. 1, p. 46–55, 1998. ISSN 10709924. http://doi.org/10.1109/99.660313.

DARWIN, C. *On the origin of species*. New York: D. Appleton and Co., 1871.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, v. 7, p. 1–30, 2006.

DIETTERICH, T. Overfitting and undercomputing in machine learning. *ACM Computing Surveys*, v. 27, n. 3, p. 326–327, 1995. ISSN 03600300. http://doi.org/10.1145/212094.212114.

DIETTERICH, T. G. Ensemble methods in machine learning. In: *Proceedings of the First International Workshop on Multiple Classifier Systems*. London, UK, UK: Springer-Verlag, 2000. (MCS '00), p. 1–15. ISBN 3-540-67704-6.

DORMANN, C. F. Promising the future? global change projections of species distributions. *Basic and Applied Ecology*, v. 8, n. 5, p. 387–397, 2007. ISSN 14391791. http://doi.org/10.1016/j.baae.2006.11.001.

DUAN, R.-Y. et al. Sdmvspecies: a software for creating virtual species for species distribution modelling. *Ecography*, v. 38, n. 1, p. 108–110, 2015. ISSN 09067590. http://doi.org/10.1111/ecog.01080.

EDWARDS, J. L. Interoperability of biodiversity databases: biodiversity information on every desktop. *Science*, v. 289, n. 5488, p. 2312–2314, 2000. ISSN 00368075. http://doi.org/10.1126/science.289.5488.2312.

EFRON, B. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, v. 7, n. 1, p. 1–26, 1979. ISSN 0090-5364. http://doi.org/10.1214/aos/1176344552.

ELITH, J. et al. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, v. 29, n. 2, p. 129–151, 2006. ISSN 09067590. http://doi.org/10.1111/j.2006.0906-7590.04596.x.

ELTON, C. S. *Animal ecology*. 2011. ed. [S.l.]: Nabu Press, 1927. ISBN 978-1175412454.

ESFANDEH, S.; KABOLI, M.; ESLAMI-ANDARGOLI, L. A chronological review on application of marxan tool for systematic conservation planning in landscape. *International Journal of Engineering and Applied Sciences (IJEAS)*, v. 2, n. 12, 2015.

ESKILDSEN, A. et al. Testing species distribution models across space and time: high latitude butterflies and recent warming. *Global Ecology and Biogeography*, v. 22, n. 12, p. 1293–1303, 2013. ISSN 1466822X. http://doi.org/10.1111/geb.12078.

FAWCETT, T. An introduction to roc analysis. *Pattern Recognition Letters*, v. 27, n. 8, p. 861–874, 2006. ISSN 01678655. http://doi.org/10.1016/j.patrec.2005.10.010.

FERNANDES, L. et al. Establishing representative no-take areas in the great barrier reef: Large-scale implementation of theory on marine protected areas. *Conservation Biology*, v. 19, n. 6, p. 1733–1744, 2005. ISSN 0888-8892. http://doi.org/10.1111/j.1523-1739.2005.00302.x.

FERRIER, S. Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic biology*, v. 51, n. 2, p. 331–363, 2002. ISSN 1063-5157. http://doi.org/10.1080/10635150252899806.

FIELDING, A. H.; BELL, J. F. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, v. 24, n. 01, p. 38–49, 1997. ISSN 1469-4387. Available from Internet: <http://journals.cambridge.org/article_S0376892997000088>.

FOERSTER, T. et al. Geoprocessing in hybrid clouds: presented at the geoinformatik 2010. *Die Welt im Netz*, p. 13–19, 2010.

FOOK, K. D. *WBCMS - a service oriented Web architecture for enhancing collaboration in biodiversity: the case of species distribution modelling community*. PhD thesis (PhD) — Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2009. Available from Internet: <http://mtc-m16c.sid.inpe.br/col/sid.inpe.br/mtc-m18@80/2008/03.17.15.17.24/doc/mirrorget.cgi?metadatarepository=sid.inpe.br/mtc-m18@80/2009/03.13.21.33&choice=full&languagebutton=pt-BR>.

FRANKLIN, J.; MILLER, J. A. *Mapping species distributions: spatial inference and prediction*. Cambridge and New York: Cambridge University Press, 2009. (Ecology, biodiversity and conservation). ISBN 978-0-521-87635-3.

Free Software Foundation. *GNU General Public License: GPLv2*. June 1991. Available from Internet: <http://www.gnu.org/licenses/gpl.html>.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, v. 55, n. 1, p. 119–139, 1997. ISSN 00220000. http://doi.org/10.1006/jcss.1997.1504.

FRIEDMAN, J. H. Multivariate adaptive regression splines. *The Annals of Statistics*, v. 19, n. 1, p. 1–67, 1991. ISSN 0090-5364. http://doi.org/10.1214/aos/1176347963.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, v. 29, n. 5, p. 1189–1232, 2001. ISSN 0090-5364. http://doi.org/10.1214/aos/1013203451.

FRIEDMAN, M. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, v. 11, n. 1, p. 86–92, 1940. ISSN 0003-4851. http://doi.org/10.1214/aoms/1177731944.

GDAL Development Team. *GDAL - Geospatial Data Abstraction Library, Version x.x.x,[[BR]]*. 2016. Available from Internet: <http://www.gdal.org>.

GEORGES, D.; THUILLER, W. *An example of species distribution modeling with biomod2*. 2013. Available from Internet: <http://finzi.psych.upenn.edu/usr/share/doc/library/biomod2/doc/Simple_species_modelling.pdf>.

GIANNINI, T. C. et al. Improving species distribution models using biotic interactions: a case study of parasites, pollinators and plants. *Ecography*, v. 36, n. 6, p. 649–656, 2013. ISSN 09067590. http://doi.org/10.1111/j.1600-0587.2012.07191.x.

GIANNINI, T. C. et al. Desafios atuais da modelagem preditiva de distribuição de espécies. *Rodriguésia*, v. 63, n. 3, p. 733–749, 2012. ISSN 2175-7860. http://doi.org/10.1590/S2175-78602012000300017.

GIBSON-REINEMER, D. K. A vacant niche: How a central ecological concept emerged in the 19th century. *Bulletin of the Ecological Society of America*, v. 96, n. 2, p. 324–335, 2015. ISSN 0012-9623. http://doi.org/10.1890/0012-9623-96.2.324.

GILMOUR, J. S.; GREGOR, J. W. Demes: a suggested new terminology. *Nature*, v. 144, n. 3642, p. 333, 1939. ISSN 00280836. http://doi.org/10.1038/144333a0.

GIULIANI, G. et al. Wps mediation: an approach to process geospatial data on different computing backends. *Computers & Geosciences*, 2011. ISSN 00983004.

GOLDBERG, D. E.; HOLLAND, J. H. Genetic algorithms and machine learning. *Machine learning*, v. 3, n. 2, p. 95–99, 1988.

GRAHAM, C. H. et al. The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*, v. 45, n. 1, p. 239–247, 2008. ISSN 00218901. http://doi.org/10.1111/j.1365-2664.2007.01408.x.

GRENOUILLET, G. et al. Ensemble modelling of species distribution: the effects of geographical and environmental ranges. *Ecography*, v. 34, n. 1, p. 9–17, 2011. ISSN 09067590. http://doi.org/10.1111/j.1600-0587.2010.06152.x.

GRIESEMER, J. R. Niche: Historical perspectives. In: KELLER, E. F.; LLOYD, E. A. (Ed.). *Keywords in evolutionary biology*. Cambridge, Mass.: Harvard University Press, 1994. p. 231–240. ISBN 9780674503137.

GRINNELL, J. The niche-relationships of the california thrasher. *The Auk*, v. 34, n. 4, p. 427–433, 1917. ISSN 0004-8038. http://doi.org/10.2307/4072271.

GROTH, P.; GIBSON, A.; VELTEROP, J. The anatomy of a nanopublication. *Inf. Serv. Use*, v. 30, n. 1-2, p. 51–56, 2010. Available from Internet: <http://dl.acm.org/citation.cfm?id=1883685.1883690>. http://doi.org/10.3233/ISU-2010-0613.

GUISAN, A. et al. Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions*, v. 13, n. 3, p. 332–340, 2007. ISSN 13669516. http://doi.org/10.1111/j.1472-4642.2007.00342.x.

GUISAN, A.; THUILLER, W. Predicting species distribution: offering more than simple habitat models. *Ecology letters*, v. 8, n. 9, p. 993–1009, 2005. ISSN 1461-0248. http://doi.org/10.1111/j.1461-0248.2005.00792.x.

GUISAN, A.; ZIMMERMANN, N. E. Predictive habitat distribution models in ecology. *Ecological Modelling*, v. 135, n. 2-3, p. 147–186, 2000. ISSN 03043800. http://doi.org/10.1016/S0304-3800(00)00354-9.

HAMPTON, S. E. et al. Big data and the future of ecology. *Frontiers in Ecology and the Environment*, v. 11, n. 3, p. 156–162, 2013. ISSN 1540-9295. http://doi.org/10.1890/120103.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, v. 143, n. 1, p. 29–36, 1982. ISSN 0033-8419. http://doi.org/10.1148/radiology.143.1.7063747.

HASTIE, T.; FITHIAN, W. Inference from presence-only data; the ongoing controversy. *Ecography*, v. 36, n. 8, p. 864–867, 2013. ISSN 09067590. http://doi.org/10.1111/j.1600-0587.2013.00321.x.

HASTIE, T.; TIBSHIRANI, R.; BUJA, A. Flexible discriminant analysis by optimal scoring. *Journal of the American statistical association*, v. 89, n. 428, p. 1255–1270, 1994.

HAWKINS, D. M. The problem of overfitting. *Journal of chemical information and computer sciences*, v. 44, n. 1, p. 1–12, 2004. ISSN 0095-2338. http://doi.org/10.1021/ci0342472.

HEIDORN, P. B. Shedding light on the dark data in the long tail of science. *Library Trends*, v. 57, n. 2, p. 280–299, 2008. ISSN 1559-0682. http://doi.org/10.1353/lib.0.0036.

HERNÁNDEZ ERNST, V.; POIGNÉ, A.; LOS, W. Lifewatch — a large-scale escience infrastructure to assist in understanding and managing our planet's biodiversity. *Geophysical Research Abstracts*, v. 12, 2010.

HERNANDEZ, P. A. et al. The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, v. 29, n. 5, p. 773–785, 2006. ISSN 09067590. http://doi.org/10.1111/j.0906-7590.2006.04700.x.

HIJMANS, R. J. et al. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, v. 25, n. 15, p. 1965–1978, 2005. ISSN 0899-8418. http://doi.org/10.1002/joc.1276.

HILL, A. et al. The notes from nature tool for unlocking biodiversity records from museum records through citizen science. *ZooKeys*, n. 209, p. 219–233, 2012. ISSN 1313-2970. http://doi.org/10.3897/zookeys.209.3472.

HIRZEL, A. H. et al. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology*, v. 83, n. 7, p. 2027–2036, 2002.

HIRZEL, A. H.; HELFER, V.; METRAL, F. Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, v. 145, n. 2, p. 111–121, 2001. ISSN 03043800.

HIRZEL, A. H.; LAY, G. L. Habitat suitability modelling and niche theory. *Journal of Applied Ecology*, v. 45, n. 5, p. 1372–1381, 2008. ISSN 00218901. http://doi.org/10.1111/j.1365-2664.2008.01524.x.

HIRZEL, A. H. et al. Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, v. 199, n. 2, p. 142–152, 2006. ISSN 03043800. http://doi.org/10.1016/j.ecolmodel.2006.05.017.

HO, T. K. Random decision forests. In: *3rd International Conference on Document Analysis and Recognition*. [S.l.: s.n.], 1995. p. 278–282.

HOEHLER, F. K. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology*, v. 53, n. 5, p. 499–503, 2000. ISSN 08954356. http://doi.org/10.1016/S0895-4356(99)00174-2.

HOLLAND, J. H. *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: University of Michigan Press, 1975.

HOLLAND, J. H. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. Cambridge (Mass.) and London: The MIT Press, 1992. (A Bradford book). ISBN 9780262581110.

HOLLAND, J. H.; REITMAN, J. S. Cognitive systems based on adaptive algorithms. *ACM SIGART Bulletin*, n. 63, p. 49, 1977. ISSN 01635719. http://doi.org/10.1145/1045343.1045373.

HORNIK, K.; STINCHCOMBE, M.; WHITE, H. Multilayer feedforward networks are universal approximators. *Neural Networks*, v. 2, n. 5, p. 359–366, 1989. ISSN 08936080. http://doi.org/10.1016/0893-6080(89)90020-8.

HUANG, J.; FRIMPONG, E. A. Limited transferability of stream-fish distribution models among river catchments: reasons and implications. *Freshwater Biology*, p. n/a, 2016. ISSN 00465070. http://doi.org/10.1111/fwb.12743.

HUCKA, M. et al. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, v. 19, n. 4, p. 524–531, 2003. ISSN 1367-4803. http://doi.org/10.1093/bioinformatics/btg015.

HUTCHINSON, G. E. Concluding remarks. In: *Cold Spring Harbor symposia on quantitative biology*. [S.l.: s.n.], 1957. v. 22, p. 415–427.

HUTCHINSON, G. E. *An introduction to population ecology*. New Haven: Yale University Press, 1978. ISBN 9780300021554.

Intergovernmental Panel on Climate Change. *Climate Change 2013 - The Physical Science Basis*. Cambridge: Cambridge University Press, 2014. ISBN 9781107415324.

ISHIBUCHI, H. et al. Construction of fuzzy classification systems with rectangular fuzzy rules using genetic algorithms. *Fuzzy Sets and Systems*, v. 65, n. 2-3, p. 237–253, 1994. ISSN 01650114. http://doi.org/10.1016/0165-0114(94)90022-1.

ISO/IEC. *Technical Report on C++ Library Extensions*. 2007–11–15.

IVICA, C.; RILEY, J. T.; SHUBERT, C. Starhpc — teaching parallel programming within elastic compute cloud. In: *Proceedings of the ITI 2009 31st International Conference on Information Technology Interfaces (ITI)*. [S.l.: s.n.], 2009. p. 353–356.

JACKSON, C. R.; ROBERTSON, M. P. Predicting the potential distribution of an endangered cryptic subterranean mammal from few occurrence records. *Journal for Nature Conservation*, v. 19, n. 2, p. 87–94, 2011. ISSN 1617-1381. http://doi.org/10.1016/j.jnc.2010.06.006.

JIMÉNEZ-VALVERDE, A.; LOBO, J.; HORTAL, J. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. *Community Ecology*, v. 10, n. 2, p. 196–205, 2009. ISSN 1585-8553. http://doi.org/10.1556/ComEc.10.2009.2.9.

JIMÉNEZ-VALVERDE, A.; LOBO, J. M. Threshold criteria for conversion of probability of species presence to either–or presence–absence. *Acta Oecologica*, v. 31, n. 3, p. 361–369, 2007. ISSN 1146609X. http://doi.org/10.1016/j.actao.2007.02.001.

JOHANNES, B. et al. Towards a research agenda for geoprocessing services. *12th AGILE International Conference on Geographic Information Science*, v. 1, p. 1–12, 2009.

KADMON, R.; FARBER, O.; AVINOAM, D. A systematic analysis of factors affecting the performance of climatic envelope models. *Ecological Applications*, v. 13, n. 3, p. 853–867, 2003. Available from Internet: <http://www.jstor.org/stable/4134701>.

KEARNEY, M. Habitat, environment and niche: what are we modelling? *Oikos*, v. 115, n. 1, p. 186–191, 2006. ISSN 00301299. http://doi.org/10.1111/j.2006.0030-1299.14908.x.

KEATING, K. A.; CHERRY, S. Use and interpretation of logistic regression in habitat-selection studies. *Journal of Wildlife Management*, v. 68, n. 4, p. 774–789, 2004.

KEITH, D. A. et al. Predicting extinction risks under climate change: coupling stochastic population models with dynamic bioclimatic habitat models. *Biology Letters*, v. 4, n. 5, p. 560–563, 2008. ISSN 1744-9561. http://doi.org/10.1098/rsbl.2008.0049.

KELLER, R. E.; BANZHAF, W. Genetic programming using genotype-phenotype mapping from linear genomes into linear phenotypes. In: *Proceedings of the 1st Annual Conference on Genetic Programming*. Cambridge, MA, USA: MIT Press, 1996. p. 116–122. ISBN 0-262-61127-9. Available from Internet: <http://dl.acm.org/citation.cfm?id=1595536.1595551>.

KLUMP, J. et al. Data publication in the open access initiative. *Data Science Journal*, v. 5, p. 79–83, 2006. ISSN 1683-1470. http://doi.org/10.2481/dsj.5.79.

Koch Veiga, A.; CARTOLANO, E. A.; SARAIVA, A. M. Data quality control in biodiversity informatics: The case of species occurrence data. *IEEE Latin America Transactions*, v. 12, n. 4, p. 683–693, 2014. ISSN 1548-0992. http://doi.org/10.1109/TLA.2014.6868870.

KOZA, J. R. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. [S.l.]: Bradford, 1992. (A Bradford book). ISBN 9780262111706.

KOZA, J. R. *Automatic discovery of reusable programs*. Cambridge and Mass. [u.a.]: MIT Press, 1994. v. 2. (Genetic programming, v. 2). ISBN 978-0262111898.

KOZA, J. R. *Genetic programming*. 5. ed. Cambridge and Mass. [u.a.]: MIT Press, 1996. ISBN 978-0262111706.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. *Biometrics*, v. 33, n. 1, p. 159, 1977. ISSN 0006341X. http://doi.org/10.2307/2529310.

LANGDON, W. B. Evolving data structures using genetic programming. In: ESHELMAN, L. (Ed.). *Genetic Algorithms: Proceedings of the Sixth International Conference (ICGA95)*. Pittsburgh, PA, USA: Morgan Kaufmann, 1995. p. 295–302. ISBN 1-55860-370-0.

LANTZ, C. A.; NEBENZAHL, E. Behavior and interpretation of the kappa statistic: Resolution of the two paradoxes. *Journal of Clinical Epidemiology*, v. 49, n. 4, p. 431–434, 1996. ISSN 08954356. http://doi.org/10.1016/0895-4356(95)00571-4.

LAROCQUE, G. et al. Common challenges for ecological modelling: Synthesis of facilitated discussions held at the symposia organized for the 2009 conference of the international society for ecological modelling in quebec city, canada, (october 6–9, 2009). *Ecological Modelling*, v. 222, n. 14, p. 2456–2468, 2011. ISSN 03043800. http://doi.org/10.1016/j.ecolmodel.2010.12.017.

LEEB, H.; WEGENKITTL, S. Inversive and linear congruential pseudorandom number generators in empirical tests. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, v. 7, n. 2, p. 272–286, 1997.

LEROY, B. et al. virtualspecies, an r package to generate virtual species distributions. *Ecography*, p. n/a, 2015. ISSN 09067590. http://doi.org/10.1111/ecog.01388.

LOBO, J. M.; JIMÉNEZ-VALVERDE, A.; REAL, R. Auc: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, v. 17, n. 2, p. 145–151, 2008. ISSN 1466822X. http://doi.org/10.1111/j.1466-8238.2007.00358.x.

LUKE, S.; SPECTOR, L. A revised comparison of crossover and mutation in genetic programming. In: KOZA, J. R. et al. (Ed.). *Genetic Programming 1998: Proceedings of the Third Annual Conference*. University of Wisconsin, Madison, Wisconsin, USA: Morgan Kaufmann, 1998. p. 208–213.

LUONG, T. v.; TALBI, E.-G.; MELAB, N. Parallel hybrid evolutionary algorithms on gpu. *IEEE Congress on Evolutionary Computation'10*, p. 1–8, 2010. http://doi.org/10.1109/CEC.2010.5586403.

MAHFOUD, S. W. Crowding and preselection revisited. In: MÄNNER, R.; MANDERICK, B. (Ed.). *Parallel Problem Solving from Nature 2*. Amsterdam: North-Holland, 1992. p. 27–36.

MANEL, S.; WILLIAMS, H. C.; ORMEROD, S. Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, v. 38, n. 5, p. 921–931, 2001. ISSN 00218901. http://doi.org/10.1046/j.1365-2664.2001.00647.x.

MARGULES, C. R.; PRESSEY, R. L. Systematic conservation planning. *Nature*, v. 405, n. 6783, p. 243–253, 2000. ISSN 00280836. http://doi.org/10.1038/35012251.

MARMION, M. et al. The performance of state-of-the-art modelling techniques depends on geographical distribution of species. *Ecological Modelling*, v. 220, n. 24, p. 3512–3520, 2009. ISSN 03043800. http://doi.org/10.1016/j.ecolmodel.2008.10.019.

MARSAGLIA, G. *The Marsaglia Random Number CDROM including the Diehard Battery of Tests of Randomness*. Tallahassee, FL, USA: Department of Statistics, Florida State University, 1995. Available from Internet: <http://stat.fsu.edu/pub/diehard/>.

MARSHALL, C. E.; GLEGG, G. A.; HOWELL, K. L. Species distribution modelling to support marine conservation planning: the next steps. *Marine Policy*, v. 45, p. 330–332, 2014. ISSN 0308597X. http://doi.org/10.1016/j.marpol.2013.09.003.

MARTIN, K.; HOFFMAN, B. An open source approach to developing software in a small organization. *IEEE Software*, v. 24, n. 1, p. 46–53, 2007. ISSN 0740-7459. http://doi.org/10.1109/MS.2007.5.

MATHEW, C. et al. A semi-automated workflow for biodiversity data retrieval, cleaning, and quality control. *Biodiversity data journal*, n. 2, p. e4221, 2014. ISSN 1314-2828. http://doi.org/10.3897/BDJ.2.e4221.

MATSUMOTO, M.; NISHIMURA, T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, v. 8, n. 1, p. 3–30, 1998. ISSN 10493301. http://doi.org/10.1145/272991.272995.

MCBRIDE, M. F. et al. Mathematical problem definition for ecological restoration planning. *Ecological Modelling*, v. 221, n. 19, p. 2243–2250, 2010. ISSN 03043800. http://doi.org/10.1016/j.ecolmodel.2010.04.012.

MCCANN, K. S. The diversity–stability debate. *Nature*, v. 405, n. 6783, p. 228–233, 2000. ISSN 00280836. http://doi.org/10.1038/35012234.

MCCULLAGH, P.; NELDER, J. A. *Generalized linear models*. 2nd ed.. ed. London and New York: Chapman and Hall, 1989. v. 37. (Monographs on statistics and applied probability, v. 37). ISBN 978-0412317606.

MCPHERSON, J. M.; JETZ, W.; ROGERS, D. J. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, v. 41, n. 5, p. 811–823, 2004. ISSN 00218901. http://doi.org/10.1111/j.0021-8901.2004.00943.x.

MEIJAARD, E.; NIJMAN, V. Secrecy considerations for conserving lazarus species. *Biological Conservation*, v. 175, p. 21–24, 2014. ISSN 00063207. http://doi.org/10.1016/j.biocon.2014.03.021.

Menke Dave, U.S. Fish and Wildlife Service. *Free photograph; close, mourning, dove, bird, standing, rock, zenaida, macroura*. Available from Internet: <http://www.public-domain-image.com/free-images/fauna-animals/birds/dove-birds-pictures/close-up-of-mourning-dove-bird-standing-on-rock-zenaida-macroura/attachment/close-up-of-mourning-dove-bird-standing-on-rock-zenaida-macroura>.

MESIBOV, R. A specialist's audit of aggregated occurrence records. *ZooKeys*, n. 293, p. 1–18, 2013. ISSN 1313-2970. http://doi.org/10.3897/zookeys.293.5111.

MEYNARD, C. N.; KAPLAN, D. M.; SILMAN, M. Using virtual species to study species distributions and model performance. *Journal of Biogeography*, v. 40, n. 1, p. 1–8, 2013. ISSN 03050270. http://doi.org/10.1111/jbi.12006.

MEYNARD, C. N.; QUINN, J. F. Predicting species distributions: a critical comparison of the most common statistical models using artificial species. *Journal of Biogeography*, v. 34, n. 8, p. 1455–1469, 2007. ISSN 03050270. http://doi.org/10.1111/j.1365-2699.2007.01720.x.

MICHENER, W. et al. Dataone: Data observation network for earth preserving data and enabling innovation in the biological and environmental sciences. *D-Lib Magazine*, v. 17, n. 1/2, 2011. ISSN 1082-9873. http://doi.org/10.1045/january2011-michener.

Millennium Ecosystem Assessment. *Ecosystems and Human Well-being: biodiversity Synthesis*. Washington, DC: Island Press, 2005. v. 5.

MILLER, J. A. Virtual species distribution models: Using simulated data to evaluate aspects of model performance. *Progress in Physical Geography*, v. 38, n. 1, p. 117–128, 2014. ISSN 0309-1333. http://doi.org/10.1177/0309133314521448.

MILLER, J. F. *Cartesian Genetic Programming*. Berlin and Heidelberg: Springer Berlin Heidelberg, 2011. ISBN 978-3-642-17309-7.

MITCHELL, M. *An introduction to genetic algorithms*. Cambridge and Mass: MIT Press, 1998, c1996. ISBN 978-0262631853.

MORRISON, M. L.; MARCOT, B. G.; MANNAN, R. W. *Wildlife-habitat relationships: concepts and applications*. 3rd. ed. Washington: Island Press, 2006. ISBN 1597260959.

MURDOCH, W. et al. Maximizing return on investment in conservation. *Biological Conservation*, v. 139, n. 3-4, p. 375–388, 2007. ISSN 00063207. http://doi.org/10.1016/j.biocon.2007.07.011.

MURRAY, J. V. et al. How useful is expert opinion for predicting the distribution of a species within and beyond the region of expertise? a case study using brush-tailed rock-wallabies petrogale penicillata. *Journal of Applied Ecology*, v. 46, n. 4, p. 842–851, 2009. ISSN 00218901. http://doi.org/10.1111/j.1365-2664.2009.01671.x.

NEMENYI, P. Distribution-free multiple comparisons. In: *Biometrics*. [S.l.: s.n.], 1962. v. 18, p. 263.

NENZÉN, H.; ARAÚJO, M. Choice of threshold alters projections of species range shifts under climate change. *Ecological Modelling*, v. 222, n. 18, p. 3346–3354, 2011. ISSN 03043800. http://doi.org/10.1016/j.ecolmodel.2011.07.011.

NOSS, R. F. Indicators for monitoring biodiversity: a hierarchical approach. *Conservation Biology*, v. 4, n. 4, p. 355–364, 1990. ISSN 0888-8892.

OLTEAN, M.; GROSAN, C. Evolving evolutionary algorithms using multi expression programming. In: *Proceedings of The 7 th European Conference on Artificial Life*. [S.l.]: Springer-Verlag, 2003. p. 651–658.

OLTEAN, M. et al. Genetic programming with linear representation: A survey. *International Journal on Artificial Intelligence Tools*, v. 18, n. 02, p. 197–238, 2009. ISSN 0218-2130. http://doi.org/10.1142/S0218213009000111.

OPENSHAW, S.; TAYLOR, P. J. A million or so correlation coefficients: three experiments on the modifiable areal unit problem. *Statistical applications in the spatial sciences*, v. 127, p. 144, 1979.

PARVIAINEN, M. et al. Using summed individual species models and state-of-the-art modelling techniques to identify threatened plant species hotspots. *Biological Conservation*, v. 142, n. 11, p. 2501–2509, 2009. ISSN 00063207. http://doi.org/10.1016/j.biocon.2009.05.030.

PEARCE, J.; FERRIER, S. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, v. 133, n. 3, p. 225–245, 2000. ISSN 03043800. http://doi.org/10.1016/S0304-3800(00)00322-7.

PEARSON, R. G.; DAWSON, T. P. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography*, v. 12, n. 5, p. 361–371, 2003. ISSN 1466822X. http://doi.org/10.1046/j.1466-822X.2003.00042.x.

PEARSON, R. G. et al. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in madagascar. *Journal of Biogeography*, v. 34, n. 1, p. 102–117, 2007. ISSN 03050270. http://doi.org/10.1111/j.1365-2699.2006.01594.x.

PETERSON, A. T. Uses and requirements of ecological niche models and related distributional models. *Biodiversity Informatics*, v. 3, n. 0, 2006. http://doi.org/10.17161/bi.v3i0.29.

PETERSON, A. T. *Ecological niches and geographic distributions*. Princeton and N.J: Princeton University Press, 2011. ISBN 978-0-691-13686-8.

PETERSON, A. T. et al. The big questions for biodiversity informatics. *Systematics and Biodiversity*, v. 8, n. 2, p. 159–168, 2010. ISSN 1477-2000. http://doi.org/10.1080/14772001003739369.

PETERSON, A. T.; PAPE , M.; EATON, M. Transferability and model evaluation in ecological niche modeling: a comparison of garp and maxent. *Ecography*, v. 30, n. 4, p. 550–560, 2007. ISSN 09067590. http://doi.org/10.1111/j.2007.0906-7590.05102.x.

PETERSON, A. T.; STOCKWELL, D.; KLUZA, D. A. Distributional prediction based on ecological niche modeling of primary occurrence data. In: SCOTT, J. M. (Ed.). *Predicting species occurrences*. Washington [u.a.]: Island Press, 2002. p. 617–623. ISBN 1559637870.

PETERSON, A. T.; VIEGLAIS, D. A. Predicting species invasions using ecological niche modeling: New approaches from bioinformatics attack a pressing problem. *BioScience*, v. 51, n. 5, p. 363, 2001. ISSN 0006-3568. http://doi.org/10.1641/0006-3568(2001)051[0363:PSIUEN]2.0.CO;2.

PETERSON, T. A.; KLUZA, D. A. New distributional modelling approaches for gap analysis. *Animal Conservation*, v. 6, n. 1, p. 47–54, 2003. ISSN 1367-9430. http://doi.org/10.1017/S136794300300307X.

PHILLIPS, S. J.; ANDERSON, R. P.; SCHAPIRE, R. E. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, v. 190, n. 3-4, p. 231–259, 2006. ISSN 03043800. http://doi.org/10.1016/j.ecolmodel.2005.03.026.

PHILLIPS, S. J. et al. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, v. 19, n. 1, p. 181–197, 2009.

POLI, R.; LANGDON, W. B. *Genetic programming with one-point crossover*. [S.l.]: Springer, 1998.

POLI, R.; PAGE, J. Solving high-order boolean parity problems with smooth uniform crossover, sub-machine code gp and demes. *Genetic Programming and Evolvable Machines*, v. 1, n. 1/2, p. 37–56, 2000. ISSN 1389-2576. http://doi.org/10.1023/A:1010068314282.

PONDER, W. F. et al. Evaluation of museum collection data for use in biodiversity assessment. *Conservation Biology*, v. 15, n. 3, p. 648–657, 2001. ISSN 0888-8892. http://doi.org/10.1046/j.1523-1739.2001.015003648.x.

PRECHELT, L. et al. Proben1: A set of neural network benchmark problems and benchmarking rules. *Fakultät für Informatik, Univ. Karlsruhe, Karlsruhe, Germany, Tech. Rep*, v. 21, p. 94, 1994.

PRESSEY, R. L. et al. Conservation planning in a changing world. *Trends in Ecology & Evolution*, v. 22, n. 11, p. 583–592, 2007. ISSN 01695347. http://doi.org/10.1016/j.tree.2007.10.001.

R Development Core Team. *R: a Language and Environment for Statistical Computing*. Vienna, Austria: [s.n.], 2006. ISBN 3-900051-07-0. Available from Internet: <http://www.R-project.org>.

RAO, C. R. The use and interpretation of principal component analysis in applied research. *Sankhy : The Indian Journal of Statistics, Series A (1961-2002)*, v. 26, n. 4, p. 329–358, 1964. ISSN 0581572X. Available from Internet: <http://www.jstor.org/stable/25049339>.

REDDY, S.; DÁVALOS, L. M. Geographical sampling bias and its implications for conservation priorities in africa. *Journal of Biogeography*, v. 30, n. 11, p. 1719–1727, 2003. ISSN 03050270. Available from Internet: <http://dx.doi.org/10.1046/j.1365-2699.2003.00946.x>. http://doi.org/10.1046/j.1365-2699.2003.00946.x.

REESE, G. C. et al. Factors affecting species distribution predictions: a simulation modeling experiment. *Ecological Applications*, v. 15, n. 2, p. 554–564, 2005. http://doi.org/10.1890/03-5374.

RITTER, N.; RUTH, M. The geotiff data interchange standard for raster geographic images. *International Journal of Remote Sensing*, v. 18, n. 7, p. 1637–1647, 1997. ISSN 0143-1161. http://doi.org/10.1080/014311697218340.

ROUBICEK, A. et al. Does the choice of climate baseline matter in ecological niche modelling? *Ecological Modelling*, v. 221, n. 19, p. 2280–2286, 2010. ISSN 03043800. http://doi.org/10.1016/j.ecolmodel.2010.06.021.

RUMELT, R. B. *Zenaida macroura - Brief natural history summary of Zenaida macroura. Smithsonian's National Museum of Natural History, Washington, D.C.* 2016. Available from Internet: <http://eol.org/data_objects/22710235>.

RYKIEL, E. J. Testing ecological models: the meaning of validation. *Ecological Modelling*, v. 90, n. 3, p. 229–244, 1996. ISSN 03043800.

SANTANA, F. et al. A reference business process for ecological niche modelling. *Ecological Informatics*, v. 3, n. 1, p. 75–86, 2008. ISSN 15749541. http://doi.org/10.1016/j.ecoinf.2007.12.003.

SANTANA, F. S.; SARAIVA, A. M. Challenges in ecological niche modelling. *XVIIth World Congress of the International Commission of Agricultural and Biosystems Engineering (CIGR)*, 2010.

SANTIKA, T. Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. *Global Ecology and Biogeography*, v. 20, n. 1, p. 181–192, 2011. ISSN 1466822X. http://doi.org/10.1111/j.1466-8238.2010.00581.x.

SANTOS, E. E.; Santos Jr, E. Cache diversity in genetic algorithm design. In: *FLAIRS Conference*. [S.l.: s.n.], 2000. p. 107–111.

SARKAR, S. et al. Biodiversity conservation planning tools: present status and challenges for the future. *Annual Review of Environment and Resources*, v. 31, n. 1, p. 123–159, 2006. ISSN 1543-5938. http://doi.org/10.1146/annurev.energy.31.042606.085844.

SAUER, J. R.; J. E. Hines; J. Fallon. *The North American Breeding Bird Survey, Results and Analysis 1966–2013. Version 01.30.2015*. 2001.

SAUPE, E. E. et al. Variation in niche and distribution model performance: The need for a priori assessment of key causal factors. *Ecological Modelling*, v. 237-238, p. 11–22, 2012. ISSN 03043800. http://doi.org/10.1016/j.ecolmodel.2012.04.001.

SCOTT, M. J. et al. Gap analysis: A geographic approach to protection of biological diversity. *Wildlife Monographs*, n. 123, p. 3–41, 1993. ISSN 00840173, 19385455. Available from Internet: <http://www.jstor.org/stable/3830788>.

SEGURADO, P.; ARAÚJO, M. B. An evaluation of methods for modelling species distributions. *Journal of Biogeography*, v. 31, n. 10, p. 1555–1568, 2004. ISSN 03050270. http://doi.org/10.1111/j.1365-2699.2004.01076.x.

SEO, C. et al. Scale effects in species distribution models: implications for conservation planning under climate change. *Biology Letters*, v. 5, n. 1, p. 39–43, 2009. ISSN 1744-9561. http://doi.org/10.1098/rsbl.2008.0476.

SERVILLA, M. et al. The ecotrends web portal: an architecture for data discovery and exploration. In: *Proceedings of the Environmental Information Management Conference*. [S.l.: s.n.], 2008. p. 139–144.

SIEBER, S. et al. Model-based systems to support impact assessment—methods, tools and applications. *Ecological Modelling*, v. 221, n. 18, p. 2133–2135, 2010. ISSN 03043800. http://doi.org/10.1016/j.ecolmodel.2010.07.002.

SILLERO, N. What does ecological modelling model? a proposed classification of ecological niche models based on their underlying methods. *Ecological Modelling*, v. 222, n. 8, p. 1343–1346, 2011. ISSN 03043800. http://doi.org/10.1016/j.ecolmodel.2011.01.018.

SIMONCINI, D. et al. Anisotropic selection in cellular genetic algorithms. In: *Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation*. New York, NY, USA: ACM, 2006. (GECCO '06), p. 559–566. ISBN 1-59593-186-4. Available from Internet: <http://doi.acm.org/10.1145/1143997.1144098>.

SKOLICKI, Z.; JONG, K. d. The influence of migration sizes and intervals on island models. In: O'REILLY, U.-M.; BEYER, H.-G. (Ed.). *GECCO'05 Genetic and evolutionary computation*. [S.l.: s.n.], 2005. p. 1295.

SMITH, A. B.; FRANKLIN, J. On evaluating species distribution models with random background sites in place of absences when test presences disproportionately sample suitable habitat. *Diversity and Distributions*, p. n/a, 2013. ISSN 13669516. http://doi.org/10.1111/ddi.12031.

SMITH, S. F. *A Learning System Based on Genetic Adaptive Algorithms*. PhD thesis (PhD), Pittsburgh, PA, USA, 1980.

SOBERÓN, J. Grinnellian and eltonian niches and geographic distributions of species. *Ecology letters*, v. 10, n. 12, p. 1115–1123, 2007. ISSN 1461-0248. http://doi.org/10.1111/j.1461-0248.2007.01107.x.

SOBERÓN, J.; PETERSON, A. T. Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics*, v. 2, p. 1–10, 2005.

SPACKMAN, K. A. Signal detecion theory: valuable tools for evaluating inductive learning. In: *Proceedings of the Sixth International Workshop on Machine Learning*. [S.l.]: Elsevier, 1989. p. 160–163. ISBN 9781558600362.

SPENCER, H. *The principles of biology*. London: Williams & Norgate, 1864–1867.

STALLMAN, R. M. *Using The Gnu Compiler Collection: A Gnu Manual For Gcc Version 5.3*. [S.l.: s.n.], 2016.

STANKOWSKI, P. A.; PARKER, W. H. Species distribution modelling: Does one size fit all? a phytogeographic analysis of salix in ontario. *Ecological Modelling*, v. 221, n. 13-14, p. 1655–1664, 2010. ISSN 03043800. http://doi.org/10.1016/j.ecolmodel.2010.03.016.

STOCKWELL, D.; PETERS, D. The garp modelling system: problems and solutions to automated spatial prediction. *int. j. geographical information science*, v. 13, n. 2, p. 143–158, 1999. http://doi.org/10.1080/136588199241391.

STOCKWELL, D. R.; PETERSON, A. Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, v. 148, n. 1, p. 1–13, 2002. ISSN 03043800. http://doi.org/10.1016/S0304-3800(01)00388-X.

STUART, S. N. Status and trends of amphibian declines and extinctions worldwide. *Science*, v. 306, n. 5702, p. 1783–1786, 2004. ISSN 00368075. http://doi.org/10.1126/science.1103538.

SWAN, A.; BROWN, S. *To share or not to share: Publication and quality assurance of research data outputs. A report commissioned by the Research Information Network.* 2008. Available from Internet: <http://eprints.soton.ac.uk/266742/1/Published_report_-_main_-_final.pdf>.

SWETS, J. Measuring the accuracy of diagnostic systems. *Science*, v. 240, n. 4857, p. 1285–1293, 1988. ISSN 00368075. http://doi.org/10.1126/science.3287615.

SWETS, J. A.; DAWES, R. M.; MONAHAN, J. Better decisions through science. *Scientific American*, v. 283, p. 82–87, 2000.

TENOPIR, C. et al. Data sharing by scientists: practices and perceptions. *PloS one*, v. 6, n. 6, p. e21101, 2011. ISSN 1932-6203. http://doi.org/10.1371/journal.pone.0021101.

TERRIBILE, L. C.; DINIZ-FILHO, J. A.; MARCO, P. d. J. How many studies are necessary to compare niche-based models for geographic distributions inductive reasoning may fail at the end. *Brazilian journal of biology Revista brasleira de biologia*, v. 70, n. 2, p. 263–269, 2010. Available from Internet: <http://www.ncbi.nlm.nih.gov/pubmed/20549059>.

THIBAUD, E. et al. Measuring the relative effect of factors affecting species distribution model predictions. *Methods in Ecology and Evolution*, v. 5, n. 9, p. 947–955, 2014. ISSN 2041210X. http://doi.org/10.1111/2041-210X.12203.

THUILLER, W. Biomod - optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology*, v. 9, n. 10, p. 1353–1362, 2003. ISSN 1354-1013. http://doi.org/10.1046/j.1365-2486.2003.00666.x.

THUILLER, W. Patterns and uncertainties of species' range shifts under climate change. *Global Change Biology*, v. 10, n. 12, p. 2020–2027, 2004. ISSN 1354-1013. http://doi.org/10.1111/j.1365-2486.2004.00859.x.

THUILLER, W. et al. Biomod - a platform for ensemble forecasting of species distributions. *Ecography*, v. 32, n. 3, p. 369–373, 2009. ISSN 09067590. http://doi.org/10.1111/j.1600-0587.2008.05742.x.

TORRES, L. G. et al. Poor transferability of species distribution models for a pelagic predator, the grey petrel, indicates contrasting habitat preferences across ocean basins. *PloS one*, v. 10, n. 3, p. e0120014, 2015. ISSN 1932-6203. http://doi.org/10.1371/journal.pone.0120014.

U.S. Geological Survey. *HYDRO1k Elevation Derivative Database, Cent. for Earth Resour. Obs. and Sci., Sioux Falls, S. D.* 2000.

VANDERMEER, J. H. Niche theory. *Annual Review of Ecology and Systematics*, v. 3, p. 107–132, 1972. Available from Internet: <http://dx.doi.org/10.2307/2096844>. http://doi.org/10.2307/2096844.

VILLALOBOS-ARIAS, M.; Coello, Carlos A Coello; HERNÁNDEZ-LERMA, O. Asymptotic convergence of metaheuristics for multiobjective optimization problems. *Soft Computing*, v. 10, n. 11, p. 1001–1005, 2006.

WATLING, J. I. et al. Performance metrics and variance partitioning reveal sources of uncertainty in species distribution models. *Ecological Modelling*, v. 309-310, p. 48–59, 2015. ISSN 03043800. http://doi.org/10.1016/j.ecolmodel.2015.03.017.

WATSON, R. T. Turning science into policy: challenges and experiences from the science-policy interface. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, v. 360, n. 1454, p. 471–477, 2005. ISSN 0962-8436. http://doi.org/10.1098/rstb.2004.1601.

WHITTAKER, R. H.; LEVIN, S. A.; ROOT, R. B. Niche, habitat, and ecotope. *American Naturalist*, p. 321–338, 1973.

WILCOXON, F. Individual comparisons by ranking methods. *Biometrics Bulletin*, v. 1, n. 6, p. 80, 1945. ISSN 00994987. http://doi.org/10.2307/3001968.

WILLIAM, P. et al. A comparison of richness hotspots, rarity hotspots, and complementary areas for conserving diversity of british birds. *Conservation Biology*, v. 10, n. 1, p. 155–174, 1996. ISSN 0888-8892. Available from Internet: <http://www.jstor.org/stable/2386953>.

WILSON, G.; BANZHAF, W. A comparison of cartesian genetic programming and linear genetic programming. In: *Proceedings of the 11th European conference on Genetic programming*. Berlin and Heidelberg: Springer-Verlag, 2008. (EuroGP'08), p. 182–193. ISBN 3-540-78670-8. Available from Internet: <http://dl.acm.org/citation.cfm?id=1792694.1792711>.

WINTLE, B. A. et al. Utility of dynamic-landscape metapopulation models for sustainable forest management. *Conservation Biology*, v. 19, n. 6, p. 1930–1943, 2005. ISSN 0888-8892. http://doi.org/10.1111/j.1523-1739.2005.00276.x.

WISZ, M. S. et al. Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, v. 14, n. 5, p. 763–773, 2008. ISSN 13669516. http://doi.org/10.1111/j.1472-4642.2008.00482.x.

WOLPERT, D.; MACREADY, W. No free lunch theorems for optimization. *Evolutionary Computation, IEEE Transactions on*, v. 1, n. 1, p. 67–82, 1997. ISSN 1089-778X. Available from Internet: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=585893>. http://doi.org/10.1109/4235.585893.

Working Group III of the Intergovernmental Panel on Climate Change. *Emissions scenarios: Summary for policymakers: a special report of IPCC Working Group III*. Geneva: WMO (World Meteorological Organization) and UNEP (United Nations Environment Programme), op. 2000. (IPCC special report). ISBN 92-9169-113-5.

WRIGHT, S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth International Congress of Genetics*, v. 1, p. 356–366, 1932.

WRUCK, W.; PEUKER, M.; REGENBRECHT, C. R. Data management strategies for multinational large-scale systems biology projects. *Briefings in bioinformatics*, v. 15, n. 1, p. 65–78, 2014. ISSN 1477-4054. http://doi.org/10.1093/bib/bbs064.

WU, J. et al. Empirical patterns of the effects of changing scale on landscape metrics. *Landscape Ecology*, v. 17, n. 8, p. 761–782, 2002. ISSN 09212973. http://doi.org/10.1023/A:1022995922992.

YANG, J. et al. An innovative computer design for modeling forest landscape change in very large spatial extents with fine resolutions. *Ecological Modelling*, v. 222, n. 15, p. 2623–2630, 2011. ISSN 03043800. http://doi.org/10.1016/j.ecolmodel.2011.04.032.

YUE, T.-X.; JORGENSEN, S. E.; LAROCQUE, G. R. Progress in global ecological modelling. *Ecological Modelling*, v. 222, n. 14, p. 2172–2177, 2011. ISSN 03043800.

ZADEH, L. A. Fuzzy sets. *Information and Control*, v. 8, n. 3, p. 338–353, 1965. ISSN 00199958. http://doi.org/10.1016/S0019-9958(65)90241-X.

ZURELL, D. et al. The virtual ecologist approach: simulating data and observers. *Oikos*, v. 119, n. 4, p. 622–635, 2010. ISSN 00301299. http://doi.org/10.1111/j.1600-0706.2009.18284.x.