UNIVERSIDADE
BEIRA INTERIOR

Suitability of Agent-Based Models to Predict Spatio-temporal Distribution of Species

João Holder Dulo Bioco

Tese para obtenção do Grau de Doutor em
Engenharia Informática
(3º ciclo de estudos)

Orientador: Prof. Doutor Paulo André Pais Fazendeiro
Co-orientador: Prof. Doutor Fernando Cánovas Garcia

Júri:
Prof. Doutor Eugénio da Costa Oliveira
Prof. Doutor Luís Filipe Barbosa de Almeida Alexandre,
Prof. Doutor Hugo Pedro Martins Carriço Proença
Prof. Doutor João Carlos Gomes Moura Pires
Prof. Doutora Maria Margarida Chagas de Ataíde Ribeiro
Prof. Doutor Paulo André Pais Fazendeiro

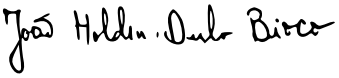
18/06/2024

Declaração de Integridade

Eu, João Holder Dulo Bioco, que abaixo assino, estudante com o número de inscrição 2042 do Doutoramento em Engenharia Informática da Faculdade de Engenharia, declaro ter desenvolvido o presente trabalho e elaborado o presente texto em total consonância com o **Código de Integridades da Universidade da Beira Interior.**

Mais concretamente afirmo não ter incorrido em qualquer das variedades de Fraude Académica, e que aqui declaro conhecer, que em particular atendi à exigida referenciação de frases, extratos, imagens e outras formas de trabalho intelectual, e assumindo assim na íntegra as responsabilidades da autoria.

Universidade da Beira Interior, Covilhã 30 /07 /2024



Dedicatória

Dedico este trabalho aos meus pais Samuel e Ester Bioco que apostaram desde cedo na minha formação. Sem vocês, não teria como chegar a esse momento. À minha amada esposa e ao meu filho que serviram de incentivo para poder concluir esse trabalho. Essa conquista é nossa.

Agradecimentos

Quero agradecer ao Professor Paulo Fazendeiro por me ter sugerido e orientado esse trabalho, por todo o suporte, dedicação e paciência. Agradeço também ao Professor Fernando Cánovas que co-orientou esse trabalho, por todo o contributo, bem com à Professora Paula Prata pelo grande apoio que me deu ao longo desse trabalho, e a todos os meus colegas que contribuiram direta ou indiretamente nesse trabalho.

Ao Instituto de Telecomunicações (Delegação da Covilhã) pelo acolhimento durante grande parte do tempo dedicado à realização dos trabalhos de desenvolvimento.

Agradeço à minha família pelas orações, e por todo o apoio prestado ao longo do meu trajeto.

Presto aqui o meu agradecimento a todos os meus colegas angolanos Bolseiros de doutoramento na UBI pela amizade e união.

Este trabalho foi suportado pela operação Centro-01-0145-FEDER-000019 -C4 - Centro de Competências em Cloud Computing, cofinanciada pelo Fundo Europeu de Desenvolvimento Regional (FEDER) através do Programa Operacional Regional do Centro (Centro 2020), no âmbito do Sistema de Apoio à Investigação Científica e Tecnológica - Programas Integrados de IC&DT.

Este trabalho foi financiado pela FCT/MCTES através de fundos nacionais e quando aplicável cofinanciado por fundos comunitários no âmbito do projeto UIDB/50008/2020.

Esse trabalho teve também o financiamento do Instituto Nacional de Gestão de Bolsas de Estudo (INAGBE-Angola))

Resumo

Os modelos de distribuição de espécies descrevem a relação entre espécies e o ambiente em que elas podem ser encontradas. Estes modelos são amplamente utilizados na modelação ecológica e ambiental principalmente para analisar as causas e os efeitos das alterações climáticas no ecossistema. O aumento das temperaturas causado pela intervenção humana tem contribuído significativamente para vários fenómenos que se têm verificado na natureza, dentre os quais se destacam o abandono do habitat natural por parte de certas espécies, a colonização de espécies invasoras, ou até mesmo a extinção de algumas espécies, entre outros. Mecanismos capazes de permitir analisar e prever tais fenómenos são de grande utilidade para se traçar medidas que garantam a gestão, conservação e preservação da biodiversidade. Apesar de garantirem essa projeção da distribuição da espécie no ambiente, os modelos de distribuição de espécies têm limitações em representar o comportamento da espécie nesse ambiente projetado. De uma forma genérica, há um conjunto significativo de informações úteis relativas ao ciclo de vida da espécie que não são tidas em devida conta, o que pode resultar em previsões pouco específicas acerca da sua reação aos estímulos ambientais. Como meio de colmatar essas limitações, a abordagem de modelos baseados em agentes tem sido adotada com sucesso assinalável. Geralmente esses modelos baseados em agentes são constituídos por indivíduos que incorporam regras comportamentais simples. Contudo, as interações entre estes, observadas numa abstração do ambiente, podem servir para descrever sistemas complexos capazes de oferecer uma (pre)visão mais fidedigna sobre o comportamento da espécie em ambiente real.

Neste trabalho propomos um modelo resultante da combinação dos modelos clássicos de distribuição de espécies com a abordagem de modelação baseada em agentes de forma a garantir uma melhor caraterização da relação entre a espécie e o ambiente. Normalmente, as implementações de modelos baseados em agentes são exigentes em termos de tempo computacional. De modo a minimizar o custo computacional resultante da simulação do modelo, apresentamos uma estratégia de paralelização que ao mesmo tempo garanta a integridade dos resultados.

Um desafio inerente à implementação de modelos baseados em agentes consiste em quantificar a escala de tempo, isto é mapear o tempo computacional com o tempo geológico. Conseguimos identificar claramente o tempo computacional dispendido com uma simulação do modelo; quando se procura estabelecer uma correspodência com o tempo real as dificuldades são acrescidas. Numa tentativa de mapear esse tempo computacional com o tempo geológico, desenvolvemos um método capaz de estimar o tempo geológico de uma simulação para os nossos modelos baseados em agentes. Este método permitiu também que se fizessem previsões da distribuição de espécies em ambientes dinâmicos. Grande parte dos ensinamentos retirados da realização desse trabalho, bem como a nossa abordagem à simulação da distribuição de espécies biológicas, foram integrados numa ferra-

menta computacional de acesso gratuito.

Palavras-chave

Modelos de distribuição de espécies, modelos baseados em agentes, distribuição espaciotemporal, modelação ambiental, temporalidade

Resumo Alargado

Os modelos de distribuição de espécies descrevem a relação entre espécies e o ambiente em que as mesmas podem ser encontradas. Estes modelos são amplamente utilizados na modelação ecológica e ambiental principalmente para analisar as causas e os efeitos das alterações climáticas no ecossistema. O aumento das temperaturas causado pela intervenção humana tem contribuído significativamente para vários fenómenos que se têm verificado na natureza, dentre os quais se destacam o abandono do habitat natural por parte de certas espécies, a colonização de espécies invasoras, ou até mesmo a extinção de algumas espécies, entre outros. Mecanismos capazes de permitir analisar e prever tais fenómenos são de grande utilidade para se traçar medidas que garantam a gestão, conservação e preservação da biodiversidade. A relação espécie-ambiente é efetivada através da projeção da distribuição da espécie no ambiente, representada em um mapa de adequação da espécie. Este mapa de adequação da espécie contém tanto os locais mais adequados como os locais menos adequados para a espécie sob o ponto de vista da sua sobrevivência e colonização. Apesar de garantirem essa projeção da distribuição da espécie no ambiente, os modelos de distribuição de espécies têm limitações em representar o comportamento da espécie, isto é, em como a espécie reproduz-se e expande-se nesse ambiente projetado. De uma forma genérica, há um conjunto significativo de informações úteis relativas ao ciclo de vida da espécie que não são tidas em devida conta, o que pode resultar em previsões pouco específicas acerca da sua reação aos estímulos ambientais. Como meio de colmatar essas limitações impostas pelos modelos clássicos de distribuição de espécies, a abordagem de modelos baseados em agentes tem sido adotada com sucesso assinalável. Geralmente esses modelos baseados em agentes são constituídos por indivíduos que incorporam regras comportamentais simples. Esses modelos apresentam uma perspetiva bottom-up permitindo que o comportamento de um sistema emerja das interações entre os indivíduos e o ambiente, e os indivíduos entre si. Contudo as interações entre estes, observadas numa abstração do ambiente, podem servir para descrever sistemas complexos capazes de oferecer uma (pre)visão mais fidedigna sobre o comportamento da espécie em ambiente real.

Neste trabalho propomos um modelo resultante da combinação dos modelos clássicos de distribuição de espécies com a abordagem de modelação baseada em agentes de forma a caraterizar melhor a relação espécie-ambiente.

Inicialmente, desenvolvemos um modelo que descreve essencialmente o ciclo de vida da espécie, e a forma como cada etapa do ciclo de vida é afetada pelas caraterísticas do ambiente onde a espécie pode ser encontrada. A espécie é representada no ambiente por uma percentagem de ocupação em cada unidade de habitat (também designada por célula). Durante cada ciclo de vida da espécie essa percentagem de ocupação aumenta de acordo com uma taxa de natalidade, reduz com base em uma taxa de mortalidade, e é transferida

para as células vizinhas de acordo com uma taxa de expansão. Entretanto, essas percentagens de ocupação da espécie são também afetadas pela adequação do ambiente; isto é, os locais mais adequados para espécie serão favorecidos, enquanto que os locais menos adequados serão penalizados. Inicialmente foram analisados os efeitos da parametrização das variáveis do ciclo de vida da espécie no output do modelo, para se perceber até que ponto os resultados estão dependentes dos valores dos parâmetros do ciclo de vida da espécie. De acordo com os resultados, os valores dos parâmetros do ciclo de vida têm grande influência nos resultados do modelo.

Normalmente, as implementações de modelos baseados em agentes são bastante exigentes em termos computacionais. De modo a minimizar o tempo computacional associado a uma simulação do modelo, implementamos uma estratégia de paralelização que garantisse simultaneamente o speedup e a integridade dos resultados. Essa estratégia de paralelização consistia em dividir o ambiente em vários blocos que eram executamos simultaneamente, introduzindo sobreposição entre os blocos e reduzindo a sincronização. Os resultados demonstraram existir uma combinação entre a sobreposição entre blocos e periodicidade de sincronização que garante não só o aumento do speedup como também a integridade dos resultados.

No modelo, o ciclo de vida da espécie, corresponde a uma iteração a nível computacional; isto é, em cada simulação realizada, é conhecido o número de iterações (ciclos ou épocas) que durou a simulação. Entretanto essa informação não é suficiente para se tomarem medidas concernentes a gestão, preservação e conservação de espécies, pois não é conhecido o tempo geológico correspondente a essa simulação. Numa tentativa de mapear o tempo computacional com o tempo geológico, desenvolvemos um método capaz de estimar o tempo geológico de uma simulação. Esse método permitiu também que se fizessem predições da distribuição de espécies, bem como análises do comportamento de espécies em cenários de alterações constantes no ambiente, conseguindo uma melhor aproximação com a realidade.

No final, uma ferramenta computacional baseada na Web e de acesso gratuito foi desenvolvida com a finalidade de agregar os principais ensinamentos e contributos deste trabalho. Esta ferramenta permite que qualquer investigador sem competências de desenvolvimento de soluções de software, interessado em analisar a distribuição de espécies de interesse em função das alterações no ambiente ao longo do tempo possa fazê-lo com um mínimo de dificuldades.

Abstract

Species distribution models are used to describe the species-environment relationship. These models are widely applied in ecological and environmental modelling mainly to analyse the causes and effects of climate changes in the ecosystem. Climate changes contribute significantly to several observed phenomena among which stand out the displacement of species from their natural habitat, the colonization of invasive species, and even the extinction of species, for instance. Mechanisms that allow analysing and predicting such phenomena are widely needed in order to adopt measures that ensure the management, conservation and preservation of biodiversity. Despite ensuring the projection of the species distribution in the environment, species distribution models have limitations concerning representing the species' behaviour in this projected environment. In generic terms, there is a set of useful information regarding the species' life cycle that is not taken into account, resulting in predictions less specific concerning the species' reaction to the environmental stimulus. To address these limitations, agent-based models approaches have been successfully adopted. Normally, these agent-based models are composed of individuals that incorporate simple behavioural rules. Therefore, the interaction between them, observed in an abstraction of the environment, could allow the description of complex systems offering a more reliable prediction regarding the species' behaviour in the environment.

In this work we propose a model resulting from the combination of traditional species distribution models with the agent-based models approach to ensure better characterisation of the species-environment relationship. Usually, agent-based models implementations are quite time-consuming and can demand a lot of computer resources. To minimize the computational cost resulting from the models' simulation, we presented a parallelization strategy that allows increased speedups, and at the same time ensures the integrity of the results.

Another challenge inherent in implementing agent-based models concerns the measurement of the time scale, i.e., mapping between computational and geological time. We can easily identify the computational time of a simulation; however, when it comes to establishing a mapping in real time, difficulties are increased. In our attempt to map the computational time with the geological time, we developed a method capable of estimating the geological time of a simulation for our agent-based models. This method also allowed performing predictions of species distribution in dynamic environments. Much of the lessons learned from this study as well as our approach concerning the species distribution simulation, were integrated into an open-access computational tool.

Keywords

Species distribution models, agent-based models, spatio-temporal distribution, environmental modelling, temporality

Contents

List of Figures

List of Tables

List of Algorithms

UBI	Universidade da Beira Interior
SDM	Species Distribution Models
EGV	Eco-geographical Variable
SDSIM	Species Distribution Simulator
ABM	Agent-based Model
ABM&S	Agent-based Modelling and Simulation
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
TSS	True Skill Statistic
AB-SDSim	Agent-based Species Distribution
ODD	Overview, Design concepts, Detail
LR	Logistic Regression
DP	Probability Density Function
GPU	Graphical Processing Unit
PE	Processing Element
sTi	Number of Iteration Until Stabilisation
itp_steps	Interpolation Steps
g_{time}	Geological Time
BP	Before Present
dEGV	Directional Eco-geographical Variable

Chapter 1

Introduction

Species distribution models (SDM) are modelling approaches widely adopted in ecological and environmental modelling to describe the relationship between a species and the environment where that species can exist [1][2]. The environment is composed of a set of eco-geographical variables (EGVs), e.g., temperature, precipitation or slope, and the observations of the species occurrence (species occurrence data) in that environment. Based on this information SDM projects the distribution of species in the environment. There are two types of occurrence data: (1) presence-only data, containing only the locations where the species was observed; and (2) presence-absence data, containing both locations where the species was observed and locations where it was not [3].

For example, given a study area with a set of EGVs (e.g. climatic variables) that constrain the behaviour of certain species and a dataset with the locations where the species has been observed in that study area, SDM can project the species' suitability map (i.e. a map of the entire study area showing both more and lesser suitable locations for the species' survival).

Usually, the performance of SDM is evaluated by using a set of performance measures, e.g. [4, 5, 6, 7], that compare the results collected in the terrain (occurrence data), with the results from the model. Therefore the ability of SDM to project the distribution of species accurately is highly dependent on the quality of the occurrence data [8][9]. Consequently, several studies have adopted virtual species to evaluate and compare the performance of SDM. There are some software packages e.g. [10, 11, 12, 13], developed to create virtual species and allow users to have complete control over species' behaviours, reducing the bias and guaranteeing a more realistic performance comparison between several SDM.

Usually, statistical and machine learning methods are used to create SDM. These methods create a response function for each EGV that describes the species' behaviours. SDM produce as output the habitat suitability, also called the suitability map of the species under study. This suitability map represents locations where the species are more likely to be found, and also the locations where the probability of species occurrences is lower. Despite that, the habitat suitability does not explain the actual behaviour of the species, taking into account that there are many issues regarding species' life cycle and relationships with the environment that cannot be answered.

It is essential to notice that each species presents a distinct behaviour concerning how they interact with the environment. Therefore, in addition to projecting the species' suitability map based on the set of EGVs and the occurrence data, the model must reflect these par-

ticular characteristics by showing how the species adapt to the environment, reproducing and spreading to the places considered more suitable for their survival. In addition to a suitability map, it is essential to have a representation that shows the places where the species can or cannot reach, settle, and populate based on their characteristics.

Another fundamental aspect, not observed in the suitability maps of the species, has to do with how to represent the changes observed in the environment. SDMs consider only abrupt changes in the environment. For example, considering projecting the distribution of a species in a study area in a different period (e.g. after 40 years), SDMs only perform abrupt changes in the values of the environmental variables, i.e., if the mean annual temperature increases 2°C during this 40 years, SDMs do not consider a gradual increment of the temperature. Thus, crucial details about the species' behaviour in changing environmental scenarios, which could more accurately reflect the species' reality, are lost.

For this information to be reflected in the models, it is important to introduce an iterative process representing the species' life cycle. This iterative process should involve simulating the distribution of the species in the environment, and analysing how they reproduce, die and spread in the environment at each iteration. For this purpose, there are different modelling and simulation approaches with an emphasis on agent-based modelling (ABM), a bottom-up modelling approach, in which the behaviour of the system emerges from the local interactions between independent entities (agents) that constitute the system [14][15]. ABM is a modelling approach widely used in ecology to simulate the distribution of species and populations. Several studies have been implementing ABM in the spatio-temporal distribution of species, e.g. [16, 17, 18, 19]. Despite the wide usage of ABM in ecology, several issues related to their implementations have been reported [20]. Hence new methods and concepts capable of addressing these issues are needed.

1.1 Motivation and Scope

Climate change is an important factor in current extinction events and habitat loss and fragmentation. The colonisation processes of invasive species also contribute significantly to the displacement of species from their natural habitat [21].

Researchers in ecology have shown great interest in modelling species distribution for management and conservation purposes. Statistical methods and machine learning models have been extensively studied to predict the spatio-temporal distribution of species and populations accurately.

Current SDM does not give us much information about a species's potential colonisation of an environment. Some factors, such as species' life cycle, would be essential in analysing species distribution. Another significant aspect concerns predicting species in different periods, for example, predicting the distribution of a species twenty years in the future. The traditional SDM predict the distribution of the species using the EGV data for the

future as if it were an abrupt change. In reality, the EGV data changes over time at different rates (smoother or harsher) rather than in an abrupt singular event.

Therefore, analysing the distribution of species in changing environmental scenarios would provide valuable information for species planning, management, and conservation purposes.

1.2 Research Aims

In this dissertation, new concepts and methods at the intersection of evolutionary ecology, mathematics and computational science are studied to address two main goals:

- Develop new mathematical and computational concepts to interpret the time evolution of the spatial distribution of species;
- Combine concepts from the fields of essential ecology, mathematics and informatics as a way to create novel pieces of software focused on the analysis of biodiversity data.

1.2.1 Specific aims and working hypotheses

The specific aims associated with the working hypothesis are described as follows:

1. *Evaluation of common problems when applying classical Spatial Distribution Model (SDM) techniques.*

   Hypothesis: classical SDM techniques present several constraints concerning analytical efficiency. Those techniques are usually limited by critical aspects at different stages of the analytical process, such as i) limited source of species occurrence data, ii) implementation of temporality, and iii) evaluation and classification of results.

2. *Implementation and evaluation of Agent-based Models (ABMs) used to model the spatial distribution of diversity of species and populations.*

   Hypothesis: ABMs are suitable models that can represent biologically complex phenomena, allowing model spatial distribution and variability of species and populations in an ecological framework.

3. *Improvement of the analytical efficiency of ABMs for their use in evolutionary and ecological studies at different scales.*

   (a) *Evaluation of the resolution of provided EGVs in the model's computational cost.*

Hypothesis: Higher resolution maps (EGVs) may significantly increase the computational cost required for species distribution simulations.

(b) *Evaluation of the effects of the implementation of dynamic variables that can virtually represent any EGV.*

Hypothesis: incorporating structural connectivity can provide a considerable advance over traditional techniques, where modellers can only infer relationships. This dynamic approach to non-static environmental variables, such as oceanographic currents, could enable the study of essential processes like dispersion, colonization, and invasion.

4. *Interpretation of temporality into the proposed new modelling environment.*

Hypothesis: Temporal modeling can also be performed. In this context, a new ecological concept should represent an ABM epoch (related to computational time) and its translation into geological terms. For example, this could include the description and interpretation of the time needed to reach equilibrium, programming the running time according to the characteristics of the agent, or considering ecological aspects of the environmental envelope.

1.3 Main Research Contributions

The main contributions of this dissertation can be summarized as follows:

1. Proposal of a new method to model and simulate the spatial distribution of species:

In addition to representing the species-environment relationship, the species' behaviour in spreading the study area is also represented by the species' life cycle parameters. This new method showed that these species' parameters have a significant impact on the way the species spread into the study area; and the projected distribution of the species provided by only the species-environment relationship (also called the suitability map) could be considerably different from the potential distribution of the species when additional factors, such as the species life cycle parameters are considered. Based on the parameterization of the model, including the representation of dynamic variables, several colonisation scenarios of the species can be produced [2, 22, 23, 24].

2. A parallelization strategy of the agent-based species distribution model:

The proposed parallelization strategy for implementing an agent-based species distribution model consists of operating in a multi-stage synchronous distributed memory mode to obtain gains in performance while reducing the need for synchronization. The environment is divided into stage-subsets according to the number of

available processes in the computer. The stage-subset is extended with an overlapping section from each neighbouring stage-subsets to ensure no information loss. This strategy guarantees an optimal trade-off between the level of redundancy and the synchronization frequency; thus, time consumption can be minimized without information loss [25].

3. A proposal of a method capable of estimating the geological time of a simulation, using the agent-based species distribution model:

   This method brings a new approach to performing the mapping between computational and geological time in the agent-based species distribution model. Instead of just simulating the distribution of species in a static environment, this mapping enables predicting the distribution of species over time in a setup of climatic changes. In addition, it allows us to estimate the distribution of species in different intervals of geological time [26].

4. A generalized web-based species distribution simulator (SDSim):

   SDSim allows modellers to simulate and predict movements and colonization patterns of species in different climate scenarios without any programming skill and monitor and analyze the spreading of the species through a visual component. It also guarantees the simulation of species' geographical ranges in a continuum of time in an environment. The parametric style of SDSim allows modellers to use real or potential environmental scenarios, as well as real or virtual species [24].

Some papers of our work have been published in journals ([23, 24]) and conferences ([2, 22, 25, 27, 26]) during this research period. One paper is currently under review process [20]. It is our understanding that these contributions provide new insights and methods regarding the objectives of this research work.

1.4 Outline of the Dissertation

The remaining of this dissertation is organized as follows.

Chapter 2 describes the main challenges of agent-based models (ABM) applied to the analysis of the spatio-temporal distribution of species, including the fundamental concepts related to ABM in ecology and the implementation of ABM in ecological and biogeographical studies. At the end of the chapter, open challenges and future trends related to ABM in the spatio-temporal distribution of species are also discussed.

Chapter 3 describes the proposed agent-based species distribution model (AB-SDSIM), following the main steps of the ODD protocol [28]. After the model's description, an in-depth parametric study is performed to analyse the effect of the parameters on the quality/behaviour of the model. In addition, two types of SDM methods (density probability

function and logistic regression) are compared. Different samples from occurrence data (presence-only data) collected in the terrain are used to perform this comparison.

In Chapter 4 a parallelization strategy of the AB-SDSIM is proposed to deal with the time-consuming issue, reducing the processing time involved in the simulations. This parallel implementation is then compared to the sequential implementation of the model to analyse the speed-up in scenarios of different dimensionality.

In Chapter 5 an approach to solving the temporality issue involving species distribution simulations is proposed by creating a method that tries to explain the mapping between computational and geological time.

Chapter 6 aggregates all these contributions described in the previous chapters into a web-based simulator (SDSIM) that allows users without any programming skills to model and simulate the distribution of species and populations.

In the concluding chapter, the main contributions of the dissertation, some unsolved issues and the future research directions are described.

Chapter 2

An Overview of Agent-based Models for Species Distribution

2.1 Introduction

Agent-based modelling and simulation (ABM&S) has become a suitable alternative to test and analyse different facets of complex adaptive systems [14].
ABM&S approaches are applied in several different scientific domains such as stock market [29, 30, 31] and supply chains assessment [32, 33, 34], electrical energy market [35, 36, 37], epidemiology [38, 39, 40], biology [41, 42], population dynamics of forest pests [43], environmental impacts [44], etc.

Furthermore, ABM&S have been increasingly adopted in the ecological modelling of species distribution, mainly to account for the impact of climate change, the dynamics of invasive species, and species range shifts [45, 46, 47, 48, 49]. The acquisition of adequate knowledge about species distribution has become crucial to developing management programs for conservation and economic reasons [50, 51, 52, 53]. ABM&S approaches have been reasonably used in ecology and conservation biology due to their capacity to simulate the dynamics of ecosystems more realistically. Implementing a species distribution model (SDM) to mimic species behaviour over space and time raises several concerns related to simulation requirements and results. This chapter outlines the main challenges encountered during the development of spatio-temporal species distribution models using ABM&S (Agent-Based Modeling and Simulation) approaches. It addresses a description of surveys that focus on the main challenges of agent-based models (ABMs) applied to analyse the spatio-temporal distribution of species; after a description of the fundamental concepts related to ABMs, the implementation of ABMs in ecological and bio-geographical studies is reviewed. Finally, the leading open challenges and future trends are discussed.

2.2 Agent-Based Modelling and Ecology

System dynamics, individual-based models (IBM), and cellular automata have been addressed in previous reviews about their application in spatio-temporal modelling of ecology systems, e.g. [54]. Geographic information systems (GIS) and remote sensing (RS) have become powerful tools to model the spatial distribution of species, although they have shown limitations when applied to the description of the dynamics of such biologi-

cal systems [54]. Among them, there are three major classes of issues that can be highlighted when dealing with spatial simulation approaches (i) contribution to theory foundations, (ii) model validation and (iii) model communication. These are particularly apparent when the aim is to integrate a geographical space into a computation system to develop more realistic models. An ecological model is a set of state variables changing along space and time. It highlights the value of ABMs, capable of treating each individual (agent) as an entity with their properties, interacting with other agents and their environment (habitat). Unlike conventional mathematical models that deal with collective population properties, ABMs are an alternative for those situations that involve spontaneous motion, such as, for instance, animal movements [54].

Computational models are increasingly used to simulate the spatio-temporal distribution of species. However, surveys focused on such kinds of ABM applications are scarce. ABMs have been previously applied to simulate processes and patterns of animal movements [55]. Four key components of agents are discussed: internal states, external factors, motion capacities, and navigation capacities. All of these components are crucial for modelling animal movement behaviour.

Several challenges are emphasized, including spatial adaptation, multi-scale environment representation, validation, and computational limitations. [55]. Additional key challenges identified in socio-ecological system modelling with ABM&S include modelling agents' behaviour, sensitivity analysis, verification and validation, coupling socio-demographic, ecological, and biophysical models, and spatial representation [56, 57, 58]. Consequently, potential research directions are identified to address these issues.

After decades of application, ABMs have been used in various ecological topics such as the conservation and management of local populations and assessing issues in complex ecological systems [51]. Future trends and challenges in ABMs have been partially addressed, including the standardization of sub-models, the need for ABMs to go beyond the construction phase, and the inclusion of a deconstruction phase [51]. All the key issues described in this section are summarized in Table 2.1.

2.2.1 Agent-based modelling and simulation approaches

ABM&S is a modelling and simulation approach existing for several years. ABMs have been adopted due to the limitation imposed by modelling alternatives such as systems of differential equations and statistical modelling. One fundamental aspect that made ABMs widely adopted was the ability to demonstrate phenomena that emerge from local interactions between individual entities [60]. Additionally, ABMs can provide a more realistic description of a system by defining simple rules (decisions) and behaviours in the individual level entities, ensuring greater flexibility [61, 62].

Nowadays, an increasing number of studies are adopting ABMs to build more efficient

Table 2.1: Issues of ABMs to the spatio-temporal distribution of species found in previous survey papers

Author	*Identified Issue*	*reference*
Tang and Bennett	- Spatial adaptation. - Multi-scale environment representation. - Validation and computation.	[55]
Filatova et al.	- Modelling agents' behaviour. - Sensitivity analysis, verification and validation. - Coupling socio-demographic, ecological, and biophysical models. - Spatial representation.	[56]
DeAngelis and Grimm	- Standardization of sub models. - The need of ABMs to go beyond the construction phase. - Include deconstruction phase.	[51]
Wallentin	- Contribution to theory. - Model validation. - Model communication.	[59]

descriptive models capable of simulating the dynamics of ecological systems.

However, several issues must be considered to build models that accurately represent reality. To achieve this, it is essential to address the generalization of ABMs, design issues, and evaluation techniques.

2.2.1.1 Agents

There are several definitions of agent elsewhere. However, existing definitions have several convergent points [63]. In general, an agent should be capable of performing an autonomous task in a virtual environment, fulfilling a purpose for which it was designed [64, 65, 66]. Usually, agents have two levels of rules: the agent's internal rules (base-level rules); and rules designed to overwrite internal rules according to the variation in the environment (high-level rules). When stated, they demonstrate an independent decision-making mechanism of their own [65].

Agents can represent several entities with behaviours and react according to their states and environment in different granularity (different levels of observation of the environment) [67]. Furthermore, several agents can interact, considering a system as a set of subsystems, i.e., an agent composed of several agents.

There are several shared characteristics among ABMs [63, 68]:

- Discrete or modular in structure, containing a set of rules to build a decision-making ability;

- Autonomous and self-directed, acting independently from the environment;
- Social, interacting with other ABM, driven by an interaction protocol to recognize the particularities of the other ABM;
- Particular allocation in an external environment;
- Driven by goals, allowing a comparison of results;
- Flexibility, showing an ability to learn and adapt based on experience and controlled by a set of results.

The number of characteristics described above that an agent could have depends significantly on the model's purposes and performance.

2.2.1.2 Agent-based modelling

Implementation and analysis of more complex and realistic models have resulted in the broad adoption of ABM&S in studies where the traditional mechanisms do not respond appropriately [69]. Any ABM consists of a set of agents capable of keeping the behaviours of the various individuals that constitute any studied system [70]. Gilbert [71] defines agent-based modelling (AB modelling) as a computational method that allows an investigator to create, analyse and experiment with models composed of agents that interact within an environment. For Kim and McGraw [72], AB modelling consists of agents that can make autonomous and adaptive decisions using local information and rules that trigger social mechanisms and processes.

Software and tools

There are several tools and software designed to implement ABMs. Typically, ABMs are implemented by using programming languages that allow for developing applications. Usually, the limitations of any AB modelling framework led us to implement dedicated applications [69]. Table 2.2 shows some of the most common software solutions for developing ABM&S applications, both desktop and large-scale development environments.

2.2.2 Agent-based modelling in ecology

The practice of AB modeling in ecology involves models that describe the interrelationships between species and their environments by representing individuals as discrete, autonomous entities. [73, 74]. Initial ABMs studies in ecology did not present a precise distinction of ABMs from classical models until four criteria were proposed. [75]: 1) how the complexity of the individual's life cycle is reflected in the model; 2) the explicit representation of the dynamic of resources used by individuals; 3) the type of data (real or integer

Table 2.2: Summary of the most common software solutions for developing ABM&S applications. .

Software	*Desktop Solution*	*Large Scale Solution*	*Programming Language*
Repast Simphony	✓	-	Java
NetLogo	✓	-	NetLogo
StarLogo	✓	-	Logo
Repast for High Performance Computing	-	✓	C++
Swarm	-	✓	Java, Objective-C
MASON	-	✓	Java
AnyLogic	-	✓	Java
FLAME	✓	✓	C
MESA	✓	-	Python

numbers) used to represent the size of a population; 4) the interval to which variability among individuals of the same age is considered.

ABMs in ecology were first applied in forest modelling, and then the application to other areas of ecology began to increase in the 1990s [76]. A particular early use of ABMs was to model the recruitment of fish populations to understand and assess human impacts on the mortality of fish recruitment [51]. However, two main model categories have motivated the development of ABMs in ecology based on purpose [77]. One category is mainly used to model a specific type of species, population, or ecosystem with management purposes; another category aims to help ecologists better understand the factors behind several ecological phenomena [78]. Nowadays, ABMs still receive much attention in ecological studies due to the possibility of simulating complex ecosystem behaviours.

A significant advantage of ABMs in ecology concerns population study by considering the individual's behaviours and relationships, resulting in a more natural way of looking at populations, making simulation results closer to reality [79]. Therefore, new techniques could reduce such differences between a simulation and the reality that it is trying to represent.

Two mutually dependent motivations have been listed when using ABMs: first, essential features of individuals, such as individual variability, are not taken into account in classical models; and second, deficiencies of the theory emerge from state variable approaches [77].

To understand the importance of ABMs in ecology, first, it is necessary to have a look at the way that classical models are built. In-state variable models or classical models, a system is described only with variables representing the state of the whole system. There is no distinction between the elements of those systems. In terms, these models ignore important particularities of species (individuals) and population [80]. Unlike classical models, individuals or agents have heterogeneous behaviours and interact with each other

and their environment. For example, an algorithm should represent the behaviour of an individual agent in a specific environment, searching for suitable places. Each agent has its behaviour according to its state, the environment and its neighbourhood. As seen in the algorithm 1, the agent will move to another location if a suitable place is found; otherwise, it will remain at the exact location.

Algorithm 1: ABM in Ecology: Searching for Suitable Places. Example of algorithm which defines an application of ABMs in Ecology. This algorithm represents how an individual agent moves to find suitable places, which could be considered locations where enough food and optimal environmental conditions occur.

```
1:  Initialize agents, environment
2:  Set time t = t_0, t_max = t_1
3:  while t < t_max do
4:      Set array of newStates[]
5:      for agent in agents do
6:          suitable = findSuitable(neighbours)
7:          if suitable! = empty then
8:              agent.select(suitable)
9:              agent.step(1)
10:         else
11:             agent.step(0)
12:         append agent to newStates
```

ABMs could better reflect the reality of ecological systems but are more complex than classical models and hence more difficult to analyse [81]. Therefore a thorough study and assessment should be done when developing an agent.

2.2.3 Model evaluation

The evaluation of the quality of the model is a crucial step in developing ABMs. After the development of a model, it should be measured, i.e., its performance should be evaluated. Conclusions can be extracted regarding the proximity between the simulation data and the actual data, provided that the latter is available. Some available model performance measures rely primarily on comparing the predictions against the actual data. In the following, the performance measures that seem to have reunited more outstanding practical adhesion are presented.

2.2.3.1 AUC

The area under the receiver operating characteristic curve (AUC) is a widely used measure of model performance for assessing the accuracy of a species distribution model (SDM) (e.g. [46, 82, 83, 84, 85, 4, 86, 87, 88]). AUC is the probability that a classifier ranks a randomly chosen positive instance higher than a randomly chosen negative one. It pro-

vides a single measure of overall accuracy that is not dependent on a specific threshold [5]. AUC values range from 0 to 1. Values close to 1 indicate that the model has good discriminating power, while values around 0.5 suggest that the model cannot distinguish between the two classes. If the AUC value is less than 0.5, the model's output should be swapped as it is producing the reciprocal classes.

2.2.3.2 Cohen's kappa

Cohen's Kappa, or the Kappa coefficient, is also widely used in ecological modelling to assess agreement between two or more observers. [6]. This coefficient analyzes situations where multiple observers classify cases as either present or absent. Kappa coefficient, k is calculated using the relation between the proportion of observed agreement P_o and the expected hypothetical probability of chance agreement P_e given by $k \equiv (P_o - P_e)/(1 - Pe)$ [89]. Kappa coefficient values range from -1 to +1, where +1 indicates perfect agreement [6]. This coefficient has been used alone or in combination with other performance measures, such as AUC and error rate, to evaluate the performance of species distribution models (SDMs) in various ecological research studies involving both terrestrial and aquatic species [90, 91].

2.2.3.3 True skill statistic (TSS)

TSS is a particular case of kappa, traditionally used to assess forecasts' accuracy. It calculates the difference between the number of correct predictions and those attributable to random guessing to a hypothetical set of perfect predictions [7]. Different from Kappa, TSS is not affected by prevalence. TSS has been used to measure the performance of SDM in vegetal and animal species [83, 46, 82, 92].

2.3 Spatio-temporal distribution of species

Spatio-temporal analysis of species distribution targets substantial interest in ecological modelling. It can be verified by the increasing number of research studies conducted with findings recently published, e.g., [93, 94, 95]. Spatio-temporal distribution allows the analysis of species dynamics along space and time with a particular goal (e.g., food search, search for suitable places, reproduction purposes) [24]. Usually, the fundamental variables influencing species' dynamics and their movement along space and time are state, environmental, and input data. ABMs are usually deployed in high-performance computers (HPC) and a Cloud Computing environment. Fig. 2.1 shows the most common components for developing ABMs in the spatio-temporal distribution of species.

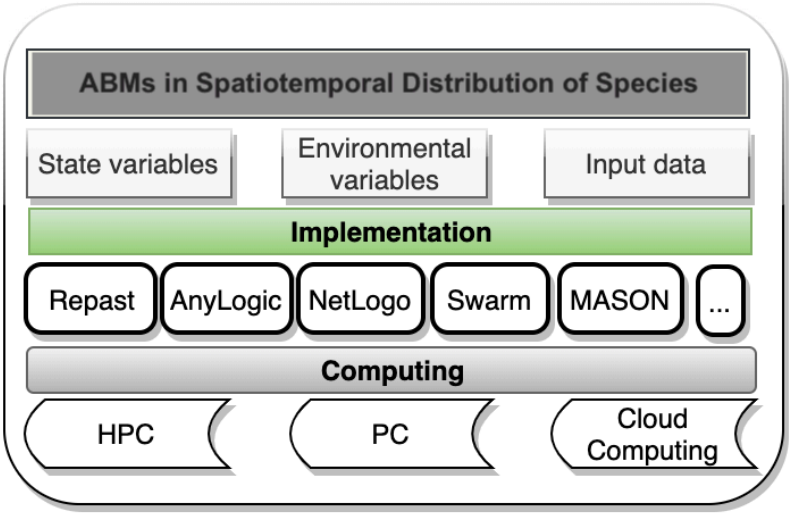


Figure 2.1: Components for the development of ABMs in the spatio-temporal distribution of species.

2.3.1 Impact of space and time in the distribution of species

Space and time are probably the most important factors when studying species distribution. Analysis of species distribution also involves interaction between species or populations. Such interaction could occur when their biological traits allow it, e.g., predators can capture their prey, viruses can overcome host resistance, and invasive species can invade a new habitat. These interactions between species could also vary along with space within different periods [96].

Predicting species distribution involves using a model with a set of environmental variables (EGVs) to simulate past, current, or future conditions. The model aims to identify the relationship between species occurrences and environmental descriptors to project species distribution. The predictive power is linked to the quality of the data used to build the model and the model's characteristics. Additionally, predicting species distribution over time requires simulating changing environmental conditions, such as transitioning from past to current or future scenarios. However, simulating species distribution within the same period but across different spatial scenarios, even if conditions are optimal for survival, may encounter constraints such as geographic barriers.

2.3.2 Overview of studies related to the spatio-temporal distribution of species

Numerous studies focus on investigating the correlation between a species' distribution and a set of environmental descriptors that capture variability, utilizing ABMs.

Pepin et al. [17] developed a stochastic ABM of feral swine population dynamics to find

a better-reducing strategy for this pest species. Feral swine are mainly responsible for the destruction of plantations in several areas. The study was conducted in the Western Czech Republic and Southeastern USA. The impact of different spatio-temporal management strategies on population response (such as culling rate, the timing of culling during the year, the spatial pattern of culling and strength barrier immigration) was tested. According to these results, the spatial culling strategy using spatial zoning showed the most significant impact on control efficiency. Their model predicts that population reduction could be achieved at lower culling intensities but when zonation is applied.

Reuter et al. [97] simulated spatio-temporal distribution of fish schooling when searching for food resources. Some studies with empirical investigations and different modelling approaches motivated authors to study fish schooling behaviour using ABM and understand how schooling can contribute to foraging success. Their model consists of three sub-models to study fish movements: description of individual fish behaviours, description of fish schooling and the environment. Cellular automata were used to simulate a heterogeneous environment where fishes were moving over time and space.

Xing et al. [18] used an object-oriented simulator of marine ecosystems (OSMOZE) to evaluate the performance of an end-to-end multispecies ABM. This end-to-end model was used to simulate the Jiaozhou Bay ecosystem dynamics in situations of limited data used for modelling. The model was calibrated using an evolutionary algorithm, and results were compared with actual data collected in random locations in the Bay. OSMOZE provided insights to improve the understanding of ecosystem dynamics in Jiaozhou Bay.

Parry et al. [98] developed an ABM to simulate the spatio-temporal distribution of moth pests across agricultural landscapes in Australia. That model simulated individual movements of female moths, reacting to landscape configuration and crop attractiveness, and therefore, exploring the influence of landscape dynamics in the movement behaviour. When developing the model, two types of agents were considered: female and habitat patches.

Heinnanen et al. [19] implemented an integrated modelling approach based on hydrodynamic modelling (HDM), SDM and ABM to study the spatio-temporal distribution of Atlantic mackerel, a highly mobile migratory species, in the Norwegian Sea. The previous implementation of single modelling techniques, such as SDM, was not enough to describe movement patterns and migrations for this species. A combination of these three modelling approaches described a more realistic result about the spatio-temporal distribution of this species.

Morales and Perry [99] developed a spatially explicit IBM (SEIBM) to represent the long-term dynamics and viability of podocarp-tawa forest patches in New Zealand. Three simulation scenarios were applied (unfragmented forest, fenced and unfenced patches). Three forest canopy species and three sub-canopy species were considered for simulations. For each time step, a series of ecological routines were sequentially performed, such as restora-

tion, planting, herbivory, and mortality. The obtained model could represent fundamental ecological processes related to the long-term dynamics of this forest.

Fust and Schlecht [100] applied ABMs to develop the Rangeland model in Drylands (RaMDry) to understand rangeland dynamics and analyse the fundamental parameters for sustainable and practical use. This model was developed in a step-wise manner to simplify, and then new specifications were gradually increased. Three entities were modelled: agent types-herds, water resources (water points) and environment (environmental spatial cells). This set of parameters gave this model the power to assess several scenarios for sustainable foraging resources.

Zhang et al. [101] proposed a modelling framework that combines the statistical inferential movements model (IMMS) and approximate Bayesian computation (ABC) to produce more realistic behaviours of ABMs. The movement of black petrels in foraging and searching for prey was used as a case study. This framework improved the calibration and parameterization of this model, resulting in more confident projections of the model outcomes.

Anderson and Dragićević [102] developed a geospatial ABM called EAB-BioCon that allows interactions between two types of insects: emerald ash borer and parasitoid wasp to evaluate the spread of forest infestation. Geospatial data from the City of Oakville, Canada, was used to explore strategies and control infestations. Several scenarios were simulated to reach a stage where borer infestation is eradicated from the study environment using this wasp. More recently, the same authors [16] combined network and complex systems theory to develop a network-ABM (N-ABM) to simulate complex spatial networks. As a result of the simulations, spatio-temporal networks were able to represent patterns of infestation.

Table 2.3 summarises the main characteristics of these studies and enumerates the main issues and challenges that seem to affect their successful development. Due to their widespread general nature, these challenges are discussed in the next section.

2.4 Open Challenges in Spatio-temporal Distribution of Species

Despite the increasing adoption of ABMs in ecological studies, literature shows that some issues remain challenging, occasionally preventing an accurate prediction of the system dynamics.

2.4.1 Modelling agent behaviours

Design and parameterization of agent's decision models are fundamental in developing ABMs to capture behaviour and interaction from reality [56]. There are some agents' be-

Table 2.3: A selection of works on ABMs in the spatio-temporal distribution of species and their significant challenges (c2.3.1 - Modelling agents' behaviour; c2.3.2- Calibration, analysis, evaluation and validation; c2.3.3- Temporality issue; c2.3.4- Spatial representation; c2.3.5- Computational cost. See sec. 2.3)

Authors	Description	Study area	State variables	Input data	Issues	Challenges				
						c2.3.1	c2.3.2	c2.3.3	c2.3.4	c2.3.5
[5]	A stochastic agent-based simulation model (ABM) of feral swine population dynamics in order to found a better reducing strategy of this pest species.	Western Czech Republic and Southeastern USA	Age, sex, family group ID, natal dispersal status, longevity, minimum age at first conception, litter size, reproductive status, gestation time, postnatal time, home range centroid, grid cell ID.	Longevity distribution, weekly conception probability, litter size distribution, sex ration of litters, age at reproductive mature for females, minimum time between farrowing and conception, gestation time, distribution of age of natal dispersal, distribution of dispersal distances, maximum group size.	Quantify the effects of population density and environment on capture costs.	-	✓	-	-	-
[57]	An agent-based model to simulate the spatio-temporal distribution of fish schooling in search food resources.	-	Cell state, fish location.	Number of simulation times steps, star value for random seed, x/y extension of simulation area, boundary values of fish initial area, number of fish, number of fish to be considered for calculation of direction, size of simulation area.	Modelling and simulation of social interactions among individuals in a complex environment; and competition between individuals (agents) in acquisition of resources.	✓	-	-	-	-
[1]	Combination of network and complex systems theory in order to developer a network-agent-based model (N-ABM) to simulate complex spatial networks.	Oakville, Canada.	Age, location, number of offspring produced, fertility, sex, stress, number of larvae.	Maximum flight/day, chance of success fertility, maximum number of offspring, survival rate of eggs, sex ratio, survival rate of larvae, carrying capacity.	Validation of the spatial networks generated by the N-ABM.	-	✓	-	-	-
[8]	Usage of object-oriented simulator of marine ecosystems (OSMOSE) to evaluate the performance of an end-to-end multispecies individual-based model.	Jiaozhou Bay (south of Shandong Peninsula in China).	-	Time steps for moving out of simulated domain, the range of age for migratory species (year), minimum size (cm), maximum size (cm), trophic level.	Several simplifications in the parameterization of the model due to lack of relevant biological information.	✓	✓	-	-	-
[103]	An individual-based model to simulate spatio-temporal distribution of moth pests, across agricultural landscapes	Australia.	Spatial location ,daily distance max, eggs laid per day (max), lifespan total eggs (max), egg laying rate.	-	Model learning. Exploring the effects of learning in the model. Add other factors in the model (environmental variables) e.g wind. Coupling with a migration model.	✓	✓	-	-	-
[4]	An integrated modelling approach based on hydrodynamic modelling (HDM), species distribution models (SDM) and agent-based modelling in order to study spatio-temporal distribution of Atlantic mackerel, a highly mobile migratory species, in the Norwegian Sea.	Norwegian Sea.	Location (x, y coordinates), speed, body length, initial body weight, total body weight.	Bathymetry, initial water level, current velocities, wind speed, boundary conditions.	Fine-scale behaviours are not well described in the model; intra-specific /inter-specific interactions are not included. Lack of data to simulate in other scenarios; the model just fits on data of a short period of time (from July to August).	✓	-	✓	✓	-
[59]	A spatially explicit individuals-based model (SEIBM) to represent long-term dynamics and viability of podocarp-tawa forest fragments.	Waikato, New Zealand.	Proportional abundance, mean dbh, mean basal area.	External species, repro-age, seeding-survivor, sapling-survivor, lld-dispersal, seed-prod, supp-mortality, sup-tolerance.	The model over-predicted class of species with small size. Lack of representation of some species and important variables in the model resulting in discrepancy between the data from the model and the data gathered from the field data.	✓	-	-	✓	-
[100]	An ABM to assess the fundamentals parameters for a sustainable and effective utilization of rangelands.	Southwestern in corner of Madagascar.	Location, number of animals, body weight, body condition score, potential milk production of the representative animal.	Maximum daily vegetation growth rate, maximum daily vegetation death rate, maximum daily vegetation decay rate, maximum potential vegetation biomass, current green biomass area density, current dry biomass area density, total biomass area density, sum of green and dry biomass.	Computing power and lack of suitable validation data. Some simplifications of the model.	✓	✓	-	-	✓
[101]	Modelling framework that combines IMMs and ABC into ABM in order to improve calibration, parameterization and evaluation of ABM.	Great Barrier Island.	Location, speed, wind speed.	Number of behavioural switches per individual, Proportion of time spent in state 1, Proportion of time spent in state 2, Distance displaced, Mean displacement, SEM displacement.	Computational power and lack of some environmental information data.	✓	-	-	-	✓
[102]	Geo-spatial agent-based model to explore strategies to control EAB infestation in forests .	Oakville, Canada.	ID, age, geography, fertility status, health status, predation status.	Maximum number of parasites,carrying capacity, maximum flight distance/day, change of fertility, sex ratio, survival rate of eggs, survival rate of larvae.	Computational efficiency.	-	-	-	-	✓

haviours that are well understood; meanwhile, others are not even observable [101]. The cognitive process fundamental to movement is an example of non-observable behaviours [104]. The agents' spatial distribution also significantly impacts the agent's behaviour definition. The spatial distribution of EGVs is highly variable; therefore, agent behaviours for one set of EGVs could not be suitable for another. Parameterising agents' behaviours to adapt to environmental (spatial) predictors would involve many variables. In the same way that the species' behaviours depend on defined environmental conditions, cognition of agents and consequently its behaviours are quite influenced by the changes observed in the environment [105].

2.4.2 Calibration, analysis, evaluation and validation

Inaccurate data is another crucial aspect to take into account [56] (e.g., false positive or false negative data), causing bias in inference and prediction of species distribution. Analysis, evaluation, and validation depend on the quality of the data used. One of the most

effective ways to analyze, validate, and evaluate a model is by comparing its results with real data collected throughout the study area. Performance can change a lot depending on the quality of available data. However, real data collected from a survey could not be enough to make assumptions about models' performance, resulting in unrealistic conclusions. Besides, missing data is another critical aspect, leading to unrealistic assumptions. Model results could fail to match real observations because of a lack of data to establish a suitable comparison [59].

2.4.3 Temporality implementations

The representation of time in ABMs is a significant challenge. Modellers face issues related to the representation and measurement of time when updating the model due to the changes observed in the environment [106]. Computational time is considered as the amount of time a particular simulation lasts. Such time t is usually set before starting a simulation by defining the number of iterations i that the simulation should run; e.g. if the number of iterations is set to $i = 50$ that simulation should run fifty times, each time (except for the first one) taking as the initial state the output of the previous one. Usually, in ABMs, time is represented by mapping each model iteration to a particular interval of time change. For instance, each iteration in a model can stand for an hour, a day, a month, year (or years) [107]. However, finding a suitable unit of time to be represented in a model is complex. This aspect is still a significant challenge in constructing ABMs applied to ecological studies when it involves computational approaches: mapping geological or biological times on a computational one. This mapping could be performed by analysing any iteration i in the simulation where results are closer to observable data. Therefore, finding an effective way to perform this mapping is necessary.

2.4.4 Spatial representation

An optimal way to develop an ABM for species distribution is by capturing spatial heterogeneity of inputs and outputs across multiple spatial scales [56]. More realistic results can be provided when multiple spatial scales are considered in the ABM definition. Regretfully, the way that space is represented in a model can be quite different from the notion of space, in reality, [59].

2.4.5 Computational cost

Computational cost is another challenge in studies that use ABMs [102, 101, 100]. Some studies represent only some agents' behaviours in their models because the computational capacity to perform a simulation is sometimes unreachable. Generally, ABMs studies involve a set of agents interacting with each other and with the virtual environment. This

situation has a high computational cost. Consequently, many studies choose to represent an agent as a group of agents in a specific location rather than an individual [100]. These representations of agents can hide some essential behaviours, resulting in a less realistic representation of them.

2.5 Future Trends

Despite the many studies showing the potential of ABMs, their development, testing, and analysis remain challenging. Different ways to approach ABMs to obtain better results should be studied and implemented. The performed analysis suggests that future research directions for ABMs may include significant efforts to address aspects such as the spatial and temporal granularity of the resulting model, the coupling and hybridisation of different models, the optimisation of the computational costs and the necessary development of fundamental theory that unites all these endeavours.

2.5.1 Spatial and temporal granularity

The ability to analyse a model according to its data, in different spatial and temporal perspectives, could result in fine/large scale species' behaviours not yet observed [19, 99]. It might be interesting to analyse and observe how species evolve by following these approaches. One way to address this challenge could be by implementing ABMs applications capable of receiving and manipulating spatial data according to the granularity the study wants to achieve. Without a doubt, spatial and temporal granularity tuning is an approach to consider, which incorporation will increase the accuracy of simulation results to actual observations.

2.5.2 Coupling models

Coupling models can obtain more realistic simulation results. Integrated ABMs can describe hidden behaviours by developing specific models for their later combined into a single, integrated model. The performance of ABMs could also be improved, and several additional and relevant factors can be brought into the simulation [103]. However, coupling models could be complex, depending on the number of models to be coupled and the specificity of each model. There are some fundamental aspects and variables to take into account when developing this type of model, such as the spatial representation of each model and the computational cost resulting from communication between ABMs.

2.5.3 Theory development

Theory patterns of ABMs have been a centre of attention that has brought essential reflections [108]. Theory patterns would help modellers build simple ABMs, following the guidance of best practices or patterns of development. However, it is necessary to start thinking outside of a specific system to accomplish that and focus on general concepts related to the development of ABMs. There are several general approaches verified across ABMs studies. The development of AB modelling focusing on general patterns and theory patterns would advance the theory development of ABMs in general. Further challenges should address how to derive theory from specific or multiple case studies [108].

2.6 Remarks and Discussion

This chapter discusses the current concerns regarding implementing ABMs in the spatio-temporal distribution of species. Most of these issues are not new and still being reported by current studies, demonstrating that they are still a challenge for ecological modellers; e.g., the absence of mechanisms to observe simulations in different spatial and temporal perspectives, models unable to fit a variety of real ecological problems, and lack of sound contributions to the theory development. ABMs have been revealed as an excellent approach for studying complex ecological systems, mainly due to their ability to simulate agents' behaviours closely to species' behaviour in ecosystems. Furthermore, modellers have the opportunity to observe virtual system behaviour as well as emerging behaviours during a simulation. However, ABM&S approaches are hard to implement due to the level of detail agents can carry. Several concerns have been raised during this chapter. Therefore, new approaches focused on addressing these issues, partially or as a whole, could provide ABMs with the ability to produce better results, particularly in highly complex systems.

Chapter 3

Parameterization Effects on Model Performance and Quality of Results

3.1 Introduction

In this chapter, an ABM to model and simulate species distribution was developed. This ABM is presented in its essential, focusing on the species' life cycle to facilitate the analysis of the effects of the parameters on the model's quality. These parameters concern what is considered the species' life cycle. Results will depend highly on the species' life cycle parameters and the type of suitability function adopted to describe the environment. These fundamental parameters would contribute to the modellers calibrating the model according to their needs and also analysing the species' behaviour depending on parameter changes. This model provides the first steps in the attempt to address the issues described in chapter 2.

3.2 A Lean Model for Parameter Assessment

The AB-SDSim Model is an agent-based species distribution model developed to study species' spatial and temporal distribution in real and simulated environmental scenarios. It brings a new perspective to species distribution models concerning the analysis of the species' behaviour considering their life cycle.

The model description follows the ODD (Overview, Design concepts, Detail) protocol [109, 28, 110].

3.2.1 Entities, state variables and scales

The model contains two entities: the grid cell and the environment. Grid cells are used to represent the agents. They are characterized by their (x, y) coordinates or latitude and longitude, the suitability value resulting from the aggregation of a set of EGVs, and the percentage of species' occupation in that location. The environment is characterized by its spatial dimensions. The attributes of the entities are numerical. Grid cell attributes are integer (coordinates) and floating point (percentage of species' occupation, suitability value). The percentage of species' occupation varies from 0 to 100%, and the suitability value ranges between $(0, 1)$. The attributes of the environment are integers. Table 3.1

shows the description of entities and state variables. For grid cells, the coordinates and suitability value remain constant, whereas the percentage of species' occupation changes over time.

Regarding spatial scale, grid cells can be represented according to the available data. These data generally range from 30 seconds ($\approx 1\ km^2$) to 10 minutes ($\approx 340\ km^2$) [111].

Table 3.1: Entities, State variables and Scales

Entity	State Variable	Attribute Type
Grid cells	(x, y) *location*	*integer*
	percentage of species' occupation	*floating point*
	suitability value	*floating point*
Environment	$m \times n$ *dimension*	*integer*

3.2.2 Process overview and schedule

Initially, the model read a set of EGVs and a known distribution of the species (presence-only data or presence-absence data); alternatively, this known distribution of the species can be replaced by the knowledge of specialists regarding the optimal values for the species concerning each EGV. These EGVs are standardized in a normal distribution. Then the model generates the habitat suitability (or suitability map). Thus the environment is initialized, followed by the initialization of the population (random grid cells) with random percentages of species' occupation. Species' life cycle defines the behaviour of the agent, considering three independent parameters: birth rate, spread rate and death rate. Figure 3.1 shows the steps of the species' life cycle incorporation in the behaviour of the agent. The model saves the state of the simulation at each time step or at a time step previously defined. The simulation goes on until a specified stopping criterion is observed. Algorithm 2 describes the process overview of the model.

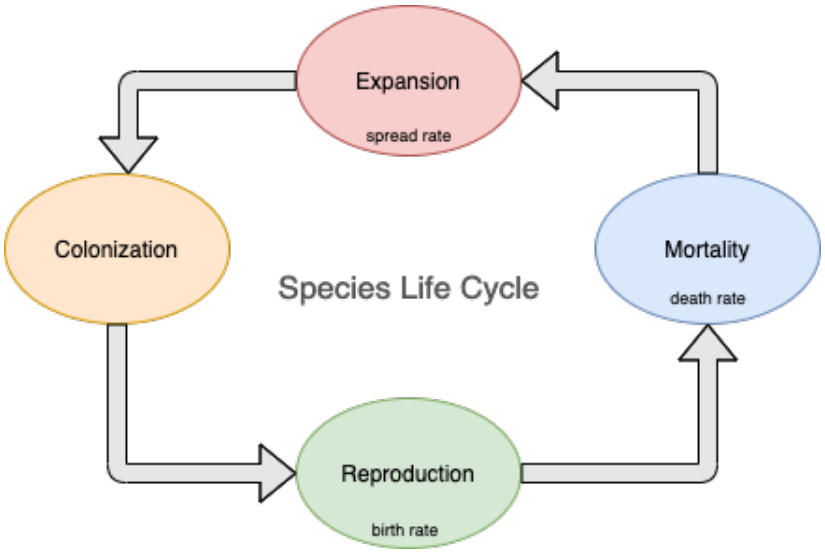


Figure 3.1: The Species Life Cycle.

Algorithm 2: *AB-SDSim* General Algorithm

procedure Simulation
Initialization of the general parameters, see Table 1
EGV_data = readEcoGeographicalVariables($filepath$)
$converted\_EGV$ = convertEGVs(EGV_data, μ, σ)
$suitability\_map$ = generateSuitabilityMap($converted\_EGV, probabilityDensityFunction$)
$normalized\_suitability\_map$ = normalizeValues($suitability\_map, model\_type$)
$patches$ = createPatches($normalized\_suitability\_map$)
Copy $patches$ into $previous\_patches$
generateInitialPopulation($previous\_patches, agents\_quantity$)
repeat
  for patch P of $patches$ **do**
    Find the $neighbours$ of P
    LIFE_CYCLE($P, previous\_P, birth\_rate, death\_rate, spread\_rate, P\_neighbours$)
  if $mod(t, output\_interval) == 0$ **then**
    Create the species distribution file at time *t*
  Calculate cell-by-cell differences between $patches$ and $previous\_patches$
  Copy $patches$ into $previous\_patches$
until stopping criterion

3.2.3 Design concepts

Basic principles

The basic principles of species distribution models are the starting point of this model [112]. This model can be seen as an extension of the species distribution models approach, in which, in addition to the projection of the suitability map, it analyses how the species can behave by spreading into this map.

Emergence

From individual behaviours emerges a general pattern consisting of transferring specimens (species' occupation percentage) between grid cells and spreading across the entire map, colonising the more suitable places and reducing these specimens in locations less suitable for the species.

Interaction

There are no direct interactions between agents. However, during the species' life cycle, the species' occupation percentage in each cell is also affected by its neighbours. The

amount transferred to a grid cell is more or less according to the values of its neighbouring cells.

Stochasticity

The initialisation of grid cells and the initial percentage of species' occupation in those cells are random. However, an initialisation dataset can also specify the initialisation of grid cells.

Observation

The following output is produced to analyse and evaluate the model: (1) the species distribution maps in each time step, containing the (x,y) locations and the percentage of species' occupation; (2) the suitability map.

3.2.4 Initialization

Models' initialization depends on the modeller's needs. It consists of initializing the environment by randomly choosing some grid cells in the environment and adding to each grid cell a random percentage of species' occupation.

3.2.5 Main Functions

As shown in the Algorithm 2 the simulation process involves four main functions as follows:

GenerateSuitabilityMap()

The suitability map for a species is obtained by following a statistical or machine-learning approach. (1) The probability density function is applied in the statistical approach. Based on a species' occurrence dataset, the mean and standard deviation considered suitable for each EGV are calculated, or these values (mean and standard deviation) can be provided by specialists who know the species' behaviour; (2) Alternatively, a logistic regression algorithm is applied.

GenerateInitialPopulation()

An amount of grid cells are randomly chosen and then filled by a random percentage of species' occupation. This percentage varies in the interval of [0, 100].

FindNeighbours()

Based on its location, each grid cell saves its neighbours, following the Moore neighbourhood [113]. However, the quantity of neighbours depends on the location of the grid cell (a minimum of three neighbouring cells).

RunLifeCycle()

Each grid cell increases the percentage of species' occupation according to the current species' occupation percentage, suitability value, and the defined life cycle parameters. Therefore, the grid cell's suitability value constrains the species' occupation percentage growth. In addition, there is no growth if the current occupation is zero. After this reproduction phase, each grid cell reduces the species' occupation percentage according to the current percentage, the suitability value, and defined life cycle parameters. Grid cells with low suitability values (closer to zero) are more penalized. In the next phase, an amount of species' occupation is distributed to the cells' neighbours. There are two spreading approaches: (1) the spreading to the neighbour cells is distributed equally; (2) the neighbour cells receive the amounts according to their suitability value. Algorithm 3 describes this procedure.

Algorithm 3: Species life cycle

procedure Run_Life_Cycle($self, birth\_rate, death\_rate, spread\_rate, neighbours$)

\# reproduction phase
$self.quantity += self.quantity * birth\_rate * self.suitability$
\# death phase
$self.quantity -= self.quantity * death\_rate * (2 - self.suitability)$
\# expansion phase
$spread = self.quantity * spread\_rate$
$self.quantity -= spread$
for $neighbour$ in $neighbours$ **do**
    $neighbour.quantity +=$ $spread*$ suitability($neighbour$) / suitability($neighbours$)

3.2.6 Input data

The model uses as input data the set of EGVs that compose the environment (study area) and also a dataset containing the occurrence of the species in the study when provided.

3.3 Life Cycle Sensitivity Analysis

In the experiments, only one grid cell in the environment is initialised with a random quantity of the species' occupation percentage. The location of the origin is randomly chosen between the grid cells with suitability values closer to one (suitable places).

Three different experiments are reported. The first experiment presented in [114] shows the effects of the main parameters of the model in a setup where only one environmental variable is considered as the determinant for the species' suitability. The environment is assumed to be smoothly changing from an area of high suitability (a level next to one) towards a hostile area (a suitability value next to 0). From a small population in a suitable area, propagation in the environment is compared to the suitability map after an equilibrium state is reached.

The following experiment introduces a second environmental variable to mimic the presence of migratory routes or otherwise corridors propitious to developing a given species.

The third setup shows the combined effect of two environmental variables, each changing from suitability to non-suitability in different directions. It is worth mentioning that these environmental variables are artificial and were created only for experimental purposes; however, their distribution, at least at a local level, is not far away from some situations that occur in real environments. Due to the high number of simulated scenarios, only a selected set of results is presented.

One fundamental aspect of these experiments was the exact time to stop each simulation (stop criteria). Several simulations were performed to find where the system reached the stabilisation - no noticeable change between two consecutive states of the system. The differences between two sequential states of the system (time t and time $t-1$) were analysed. In the model, the difference between one and another state of the system lies in the species' occupation percentage presented in each grid cell. Thus, the sum of the cell-by-cell differences in these sequential states of the system was calculated. The simulation was interrupted when this difference was maintained below a small threshold for several ticks.

3.3.1 Smooth environmental gradation

Figure 3.2 depicts an environment that is gradually changing from an area of high suitability (a level next to 1 at the bottom) towards a non-suitable area (suitability level next to 0 over the top).

After randomly placing the origin of the species in a suitable environment, the simulation starts, and the model evolves according to its life cycle. In the following, the results of the simulations scenarios varying the spread rate for values equal to (A) 0.03, (B) 0.05, (C) 0.07, (D) 0.09 and keeping fixed the birth rate (0.7) and the death rate (0.1) are presented.

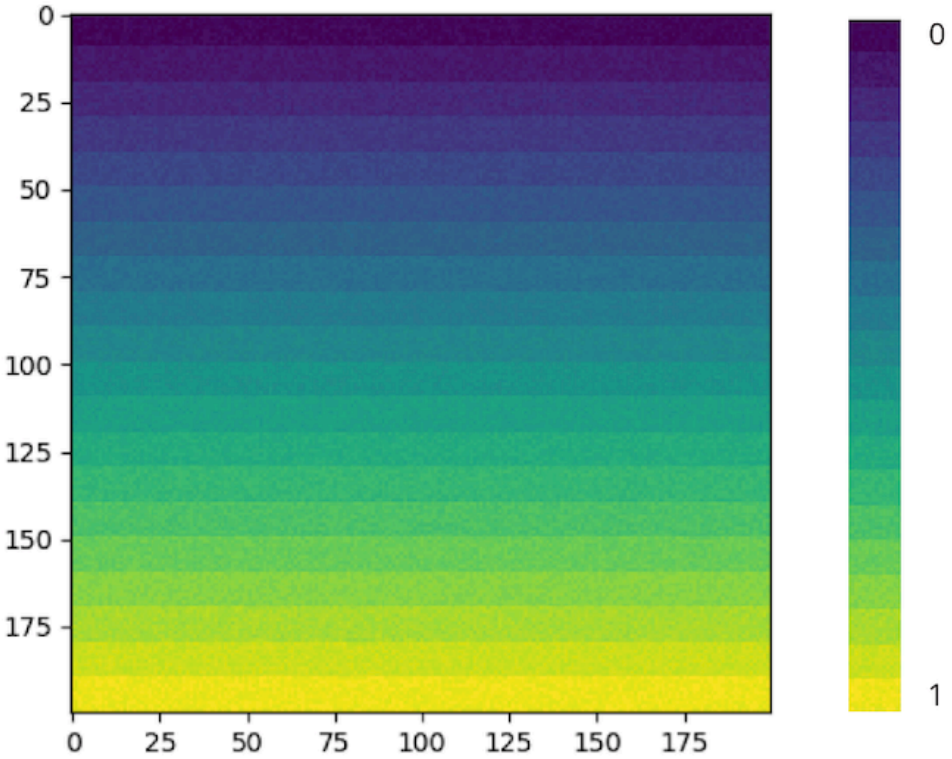


Figure 3.2: Environment map. The colour scale represents the suitability of the environment.

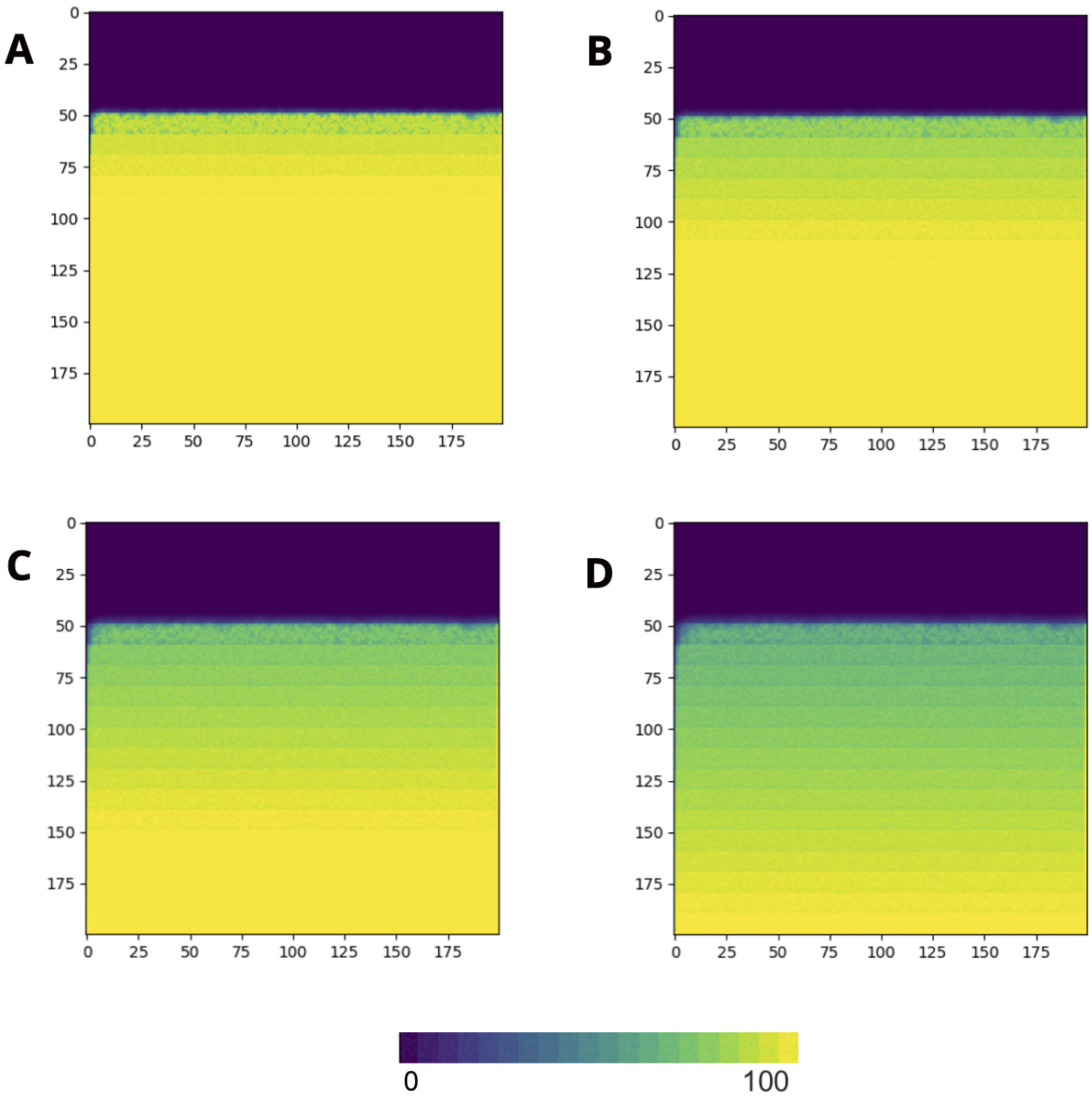


Figure 3.3: Distribution of the species in a smooth gradation environment. The scale of colour represents the species' occupation percentage in each grid cell. Gradation visibility is more evident as the spread rate is higher (Fig C and D), whereas as the spread is lower, gradation tends to disappear (Fig A and B).

As observed in Fig. 3.3, species tend to establish themselves in locations where the envi-

ronmental conditions suit them to survive and reproduce. The model outputs often follow the same pattern, excluding the scenarios where the species cannot survive or reproduce. However, the capacity of the species to spread varies according to the three parameters (birth rate, death rate and spread rate). In the first approach, making a visual comparison between these results (Fig. 3.3), it is possible to verify similarities between them. The model output follows the transition (gradation) presented in the environment map. However, a visual comparison is insufficient to conclude the model's behaviour. Often, species do not survive when the value of the birth rate is less than the death rate. To analyse the output of the model in these different parameters' combinations, Fig. 3.4 depicts the comparison of the model's output in all scenarios with the suitability map (see the environmental map in Fig. 3.2). The model output was converted to the same scale (0,1) of the environmental map to facilitate comparison. The overall comparison technique adapted from [115] was performed for each model output.

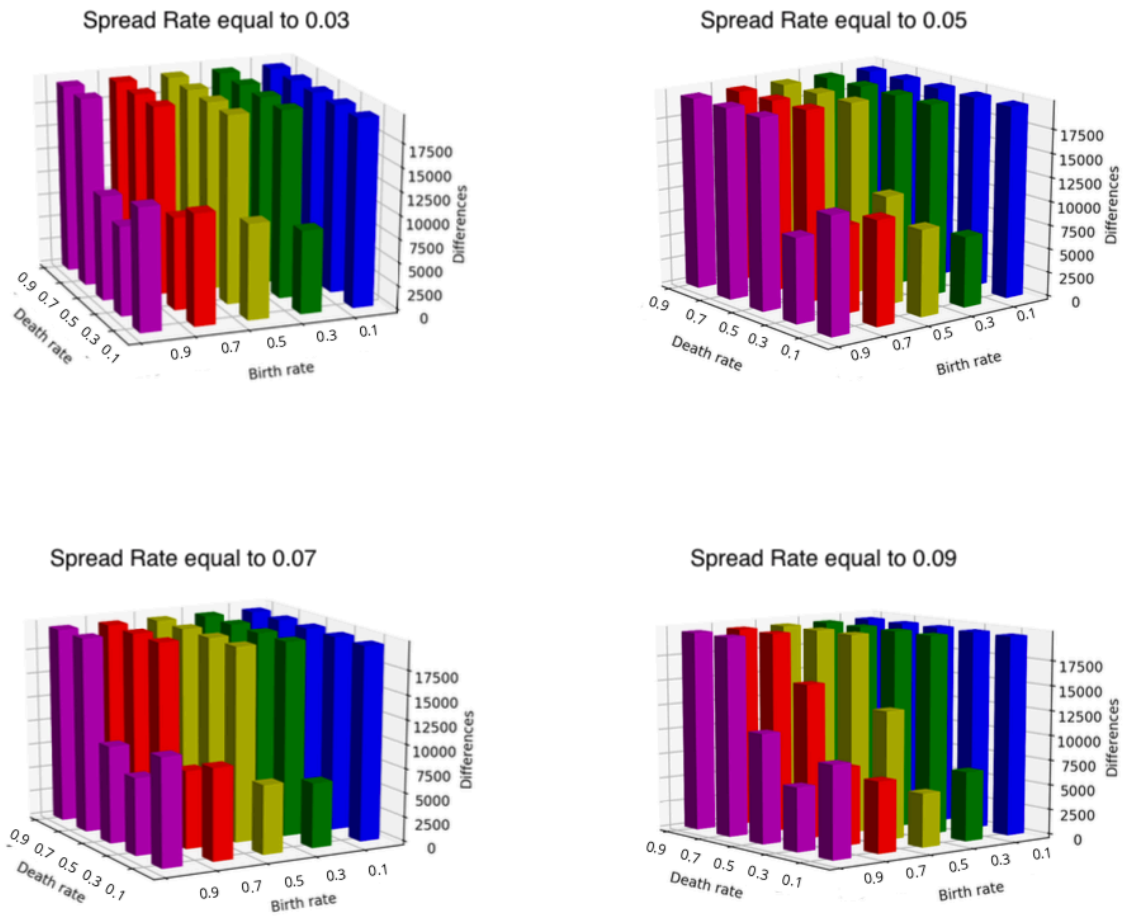


Figure 3.4: Cell-by-Cell comparison between model output resulting from the smooth gradation environment and the environment map. For each spreading rate scenario, the vertical bars depict the result for a particular tuple (Death rate, Birth rate) of parameters.

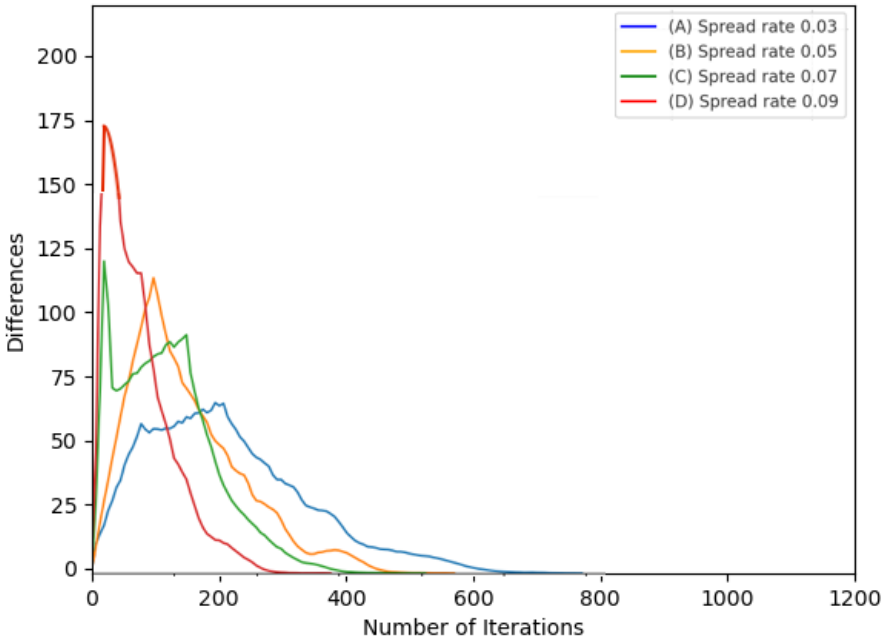


Figure 3.5: Stabilisation of the model for the four simulation scenarios in a smooth environmental gradation. (A) Spread rate equal to 0.03, (B) spread rate equal to 0.05, (C) spread rate equal to 0.07 and (D) spread rate equal to 0.09.

In Figure 3.4, it is possible to identify the scenarios presenting the relative lowest differences towards the suitability map levels. The combination (birth rate = 0.5, death rate = 0.1, spread rate = 0.09) presented the lowest difference, followed by the combination (0.9, 0.3, 0.09), and the combination (0.5, 0.1, 0.07) in the same order of rates. According to Fig. 3.4 when the death rate is greater than or equal to 50%, even with a high birth rate of 90%, the chances of the species surviving are minimal. This indicates that extremely high mortality rates significantly hinder the species' prevalence, making it difficult for the population to sustain itself despite high reproductive efforts. Conversely, when the birth rate is less than 20%, the chances of the species surviving and spreading are also very low. This suggests that very low reproduction rates critically limit the population growth and expansion potential of the species. In summary, both high mortality rates and low reproduction rates severely impact the species' ability to survive and thrive. . In this regard, the model can achieve a more consentaneous filling of the environment for higher spread rates, subsumed to the hypothesis on the suitability of the species.

Figure 3.5 shows the number of iterations necessary to reach a stable state for four different spread rates (everything else being equal).

Observing Fig. 3.5, it is possible to notice that at the beginning of the simulation, the difference between two sequential states increases very quickly. The increase in the difference until a certain number of iterations is verified, and then these differences start to

decrease to the point that it stabilizes. Another interesting finding is that a higher spread rate promotes quicker stability.

3.3.2 Conceptualization of a suitability corridor

For this experiment, the synthetic environmental variable presented in the previous section is considered, see Fig. 3.6-A, and a second environmental map is introduced, Fig. 3.6-B, representing a suitability corridor (we can think of it as a migratory route, for instance). The combined suitability cell values were obtained by summing up the values of the two environmental variables and sequent normalization to the unit interval, see Fig. 3.6-C.

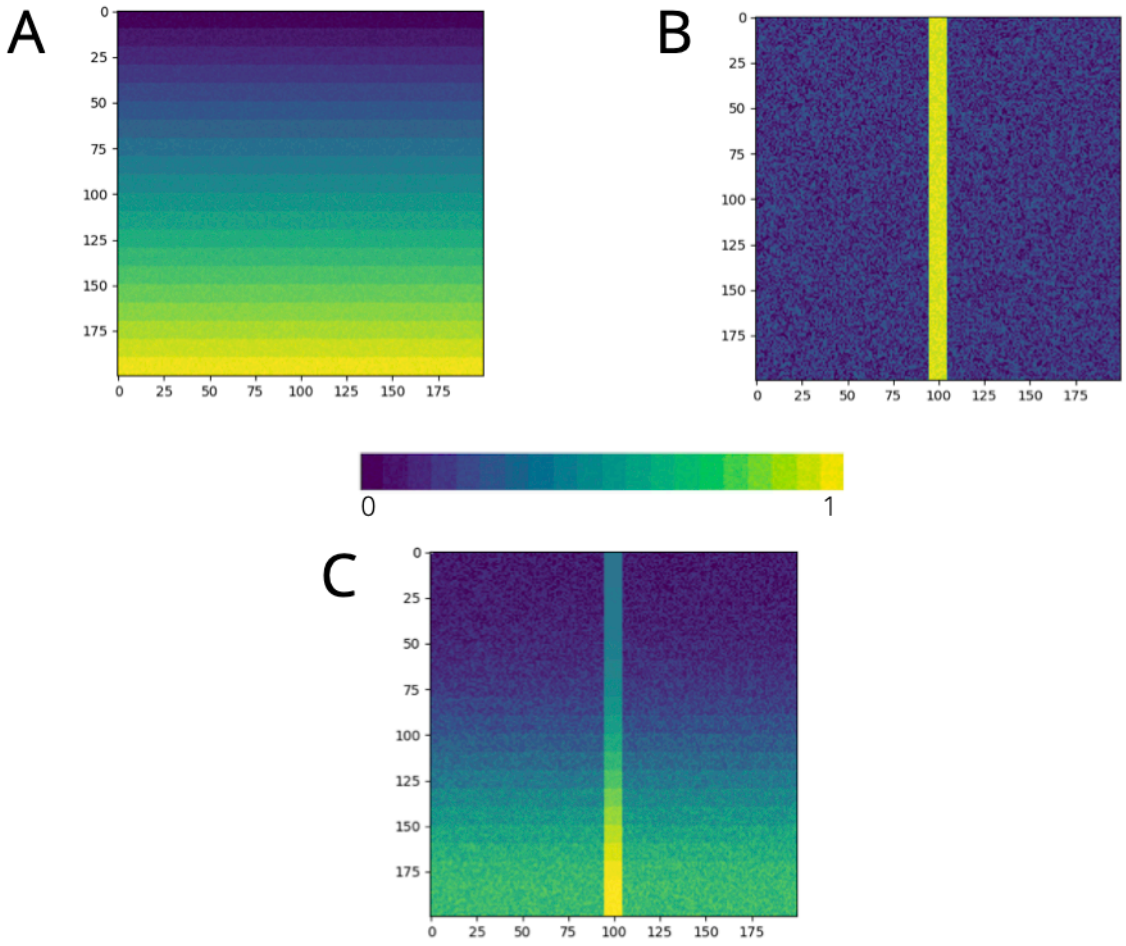


Figure 3.6: Environmental variables (A and B) and the suitability map (C). The obtained suitability map (map C) shows the most propitious places for the given species.

Figure 3.7 shows the final distribution of the species in the environment for the previously chosen four simulation scenarios having the same birth rate (0.7) and death (0.1) rates, varying the spread rate for the values 0.03, 0.05, 0.07 and 0.09.

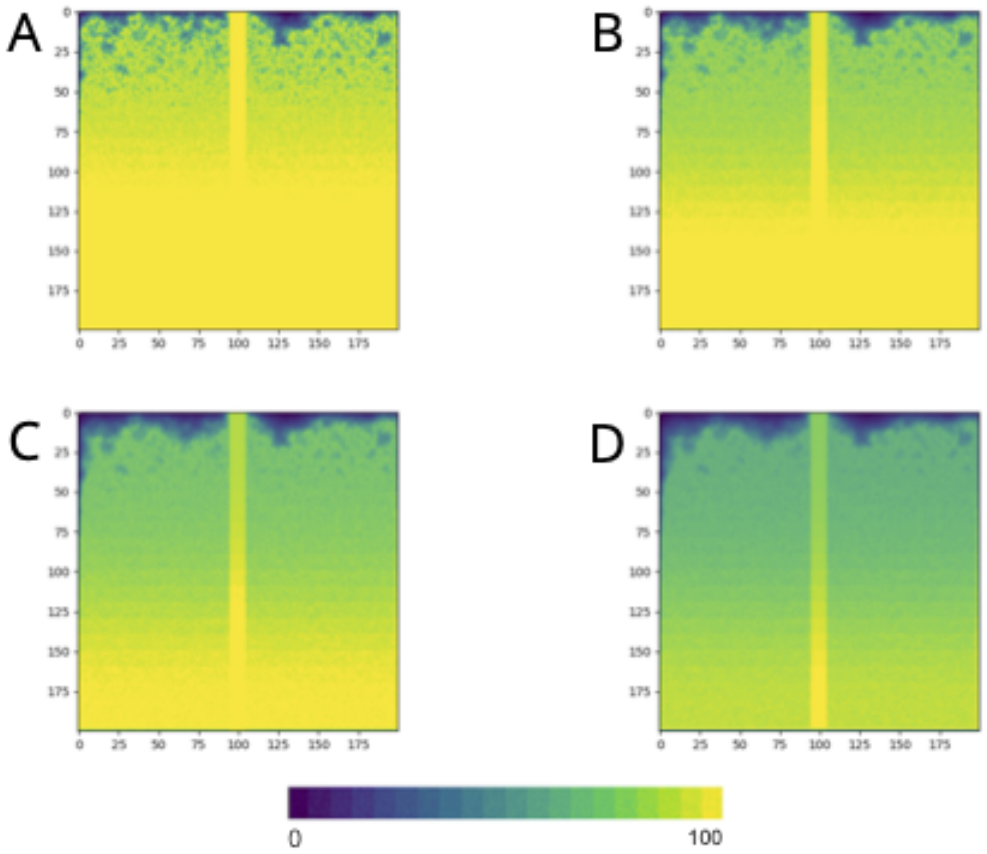


Figure 3.7: Distribution of the species in an environment with a suitability corridor. The scale of colour represents the species' occupation percentage in each grid cell.

According to Fig. 3.7, species tend to colonize all the environment. Unlike the previous maps (Fig. 3.3), where there were no conditions for the species to spread on the top, in this particular case, there is a set of suitable grid cells that allows species to spread. Another factor that influences the spread of the species to the top of the map is the suitability corridor (the vertical line). This corridor allows species to reach less suitable locations. The difference between birth rate and death rate (0.7 and 0.1) also has a significant impact on the colonization effect, and it is possible to observe a larger filling of the map when the spread rate is lower, see Fig. 3.7-A. Comparing Figure 3.7 with the suitability map (Fig. 3.6-C) it is possible to observe the same pattern between them. The model results also observe the transition (gradation) and the vertical line presented in the suitability map.

Figure 3.8 shows the cell-by-cell comparison between the model output (smooth gradation + suitability corridor) and the environmental map (Fig 3.6-C).

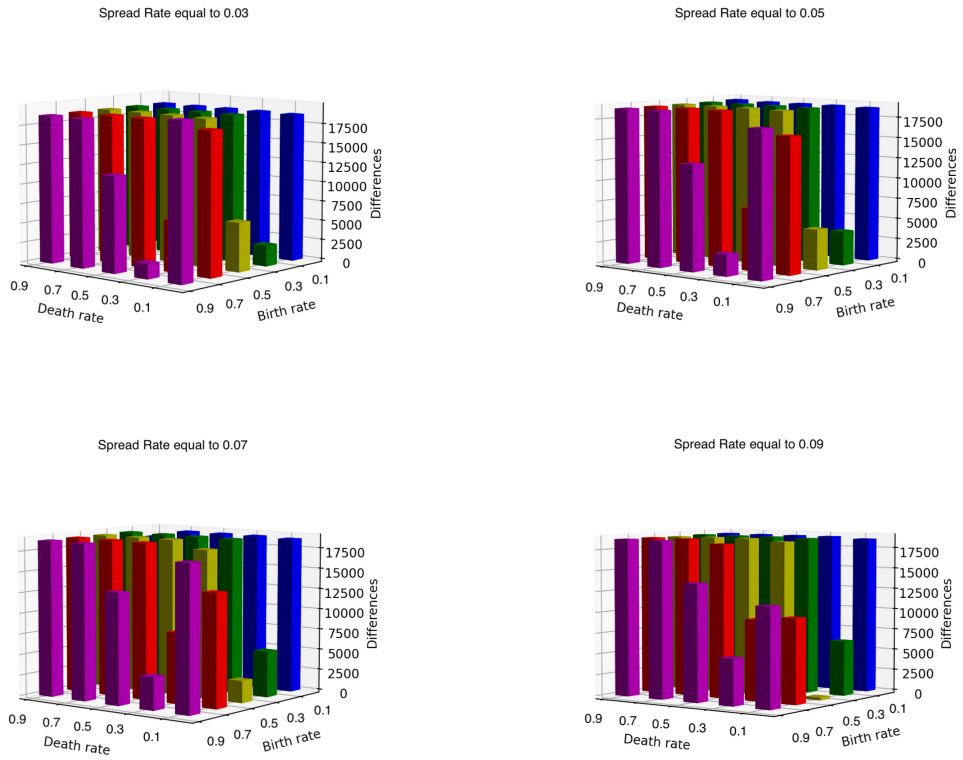


Figure 3.8: Cell-by-Cell comparison between model output (smooth gradation + suitability corridor) and the environment map.

These model results were converted in the scale (0, 1) to facilitate the comparison with the suitability map (Fig. 3.6). Doing the comparison between each simulation results (output) and the suitability map (Fig. 3.8), it is verified that the combination (death rate = 0.1, birth rate = 0.5, spread rate = 0.09), presented the lowest difference, followed by the combination (0.1, 0.3, 0.03), and the combination (0.1, 0.5, 0.07) in the same order of rates. Observing Fig. 3.8, at a death rate greater or equal to 70%, species do not survive, even with a birth rate greater or equal than 90% at the birth rate less than 20% the chances for the species to survive are scarce. Contrary to the first experiments, the three best results were obtained with different spread rates: 0.09, 0.03 and 0.07.

Unlike the first experiment, the best results were obtained in different spread rates (0.09, 0.03, 0.07). Figure 3.9 shows the number of iterations necessary to reach a stable state for four different spread rates.

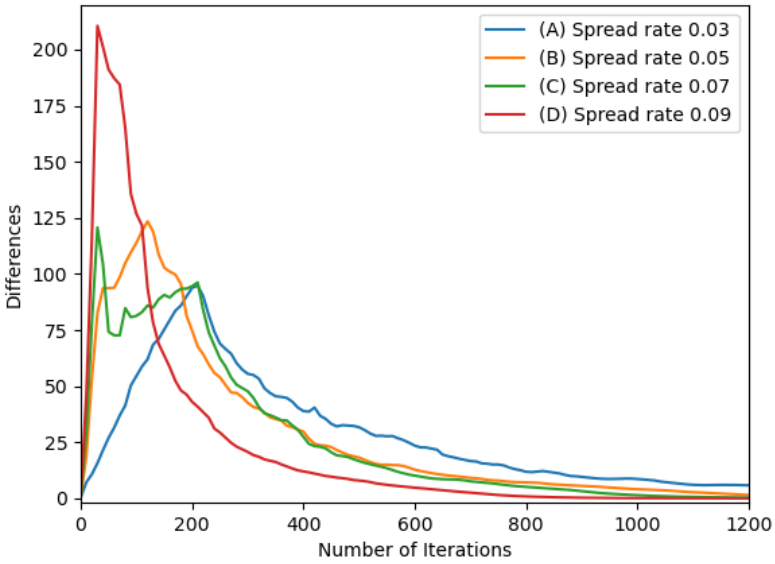


Figure 3.9: Stabilisation of the model for the four simulation scenarios in an environment with a suitability corridor. (A) Spread rate equal to 0.03, (B) spread rate equal to 0.05, (C) spread rate equal to 0.07 and (D) spread rate equal to 0.09.

Like the first experiment, it is observed in Fig. 3.9 that the differences between two sequential states start to grow until they reach their peak. These differences start to decrease until the point of stabilisation. Compared with the previous experiment, this experiment takes longer to converge due to more significant heterogeneity of the environment resulting from the combination of two environment variables. These results show that the simulation with a spread rate equal to 0.03 (A) takes much longer to converge; It allows a more extensive filling of the map when the combination of birth rate and the death rate is suitable for the species (for example birth rate equal to 0.7 and death rate equal to 0.1).

3.3.3 Compound effect of two environmental variables

In this experiment, the environmental variable presented in the first experiment is considered, see Fig. 3.10-A and a second variable is considered by rotating 90° this map, resulting in a similar gradation but with a different orientation, see Fig. 3.10-B. The suitability map was obtained by combining these two environmental variables; see Fig. 3.10-C.

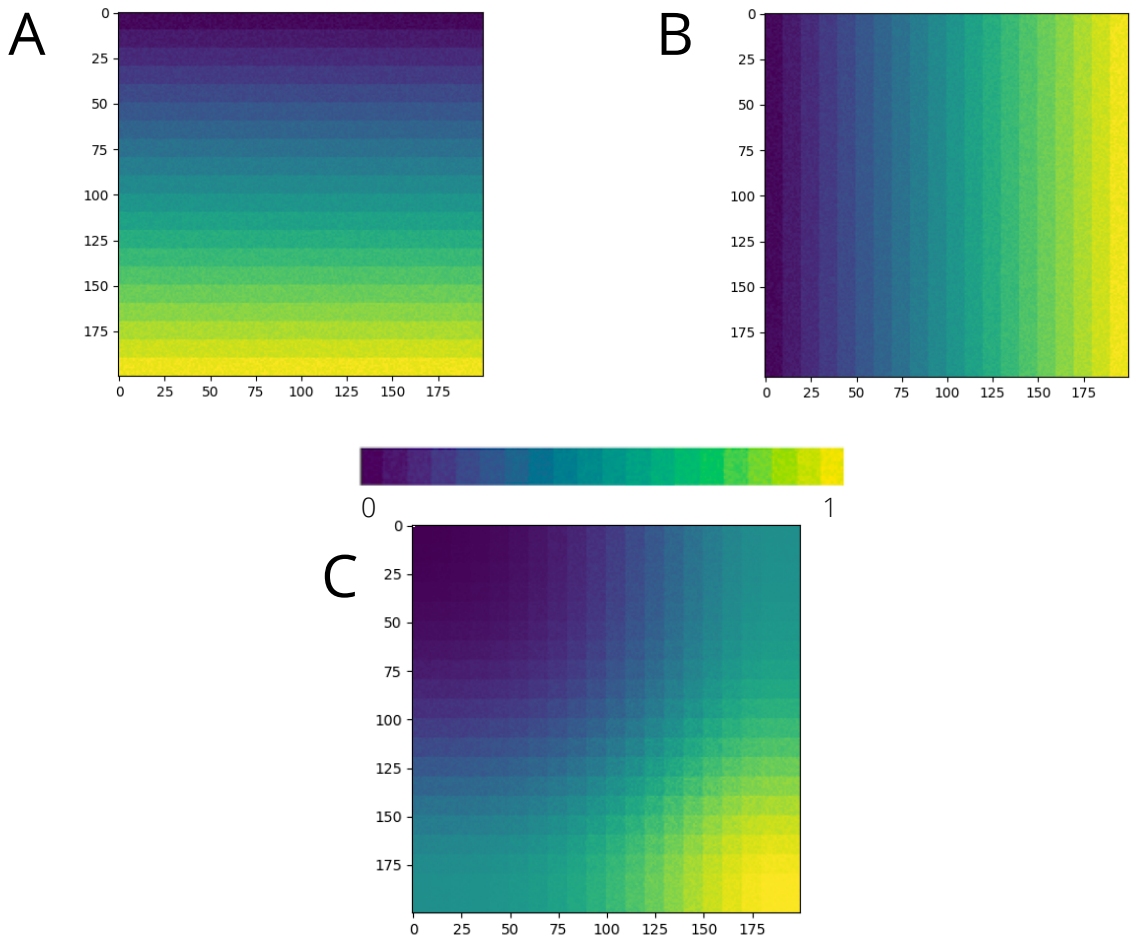


Figure 3.10: Environmental variables (A and B) and the suitability map (C) resulting from a compound effect of A and B. The obtained suitability map (map C) shows where the species will be located in greater abundance.

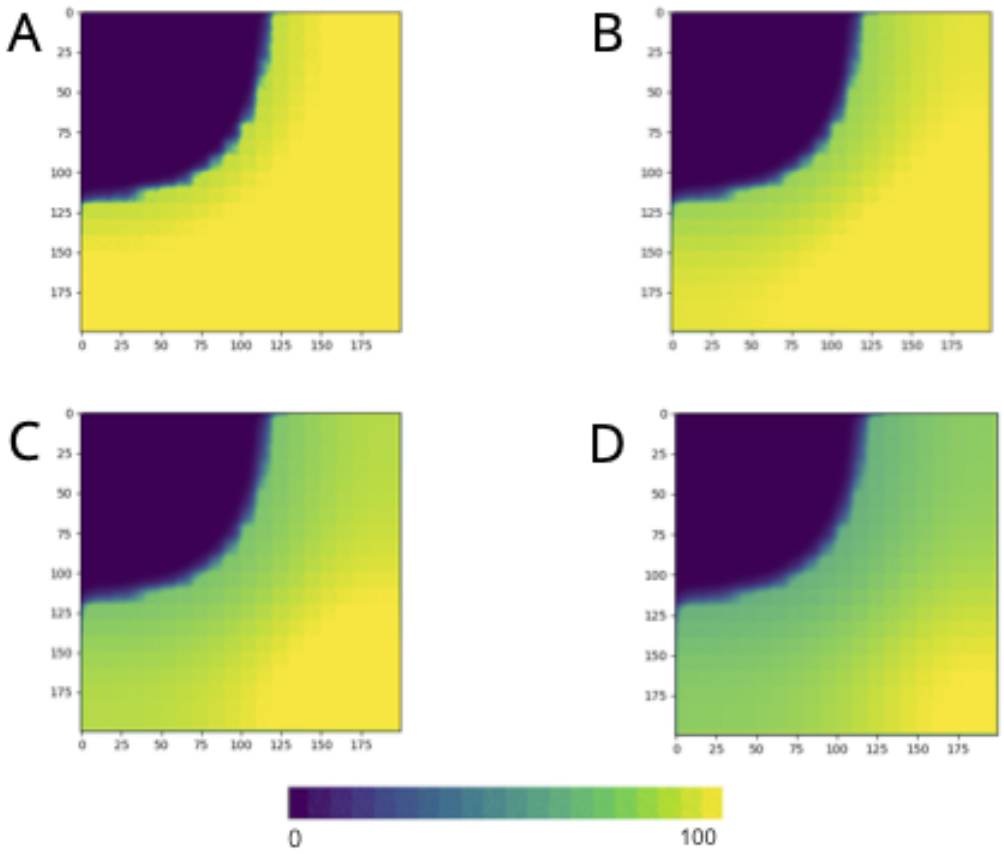


Figure 3.11: Distribution of the species in the environment composed of two environmental variables.

In Fig. 3.11, species occupy the most suitable places to stabilize and reproduce. Species tend to disappear in locations where the suitability values are low. We can observe the gradation pattern presented in the suitability map in each figure (A, B, C and D). The impact of the spread rate is highly noticeable. In the resulting suitability map (Fig. 3.10-C), the least suitable places for the species to survive are located at the top left. Therefore, species do not reach these places. As seen in previous experiments, species colonize in more abundance for the scenarios where the spread rate is lower.

Figure 3.12 shows the cell-by-cell comparison between the model output (compound effect of two environmental variables) and the environmental map (Fig. 3.10-C).

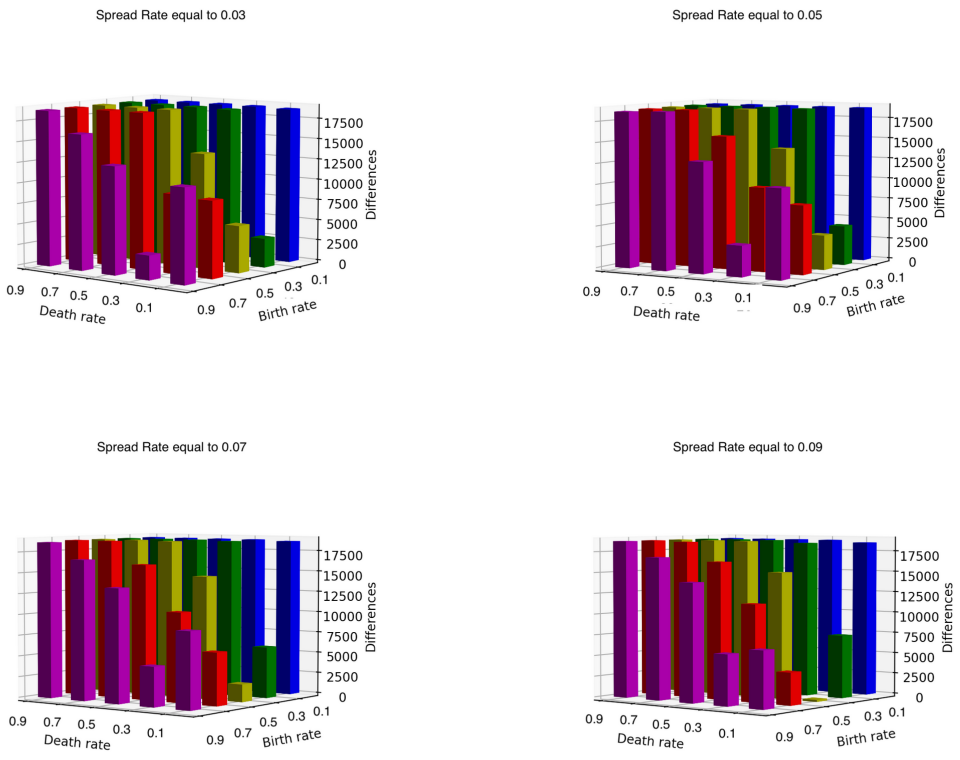


Figure 3.12: Normalized Cell-by-Cell comparison between model output and compound environment map. For each spreading rate scenario, the vertical bars depict the result for a particular tuple (Death rate, Birth rate) of parameters.

The simulation results allow verifying that for this experiment, the combinations (death rate = 0.1, birth rate = 0.5, spread rate = 0.09) presented the lowest difference concerning the suitability map, followed by the combination (0.1, 0.5, 0.7), and the combination (0.3, 0.9, 0.3) in the same order of rates. In Fig. 3.12, it is verified that there are no chances for the species to survive or reproduce at a birth rate equal to the death rate. The lowest differences can be observed for the four spread rates: 0.03, 0.05, 0.07 and 0.09.

Figure 3.13 shows the number of iterations necessary to reach a stable state for four different spread rates.

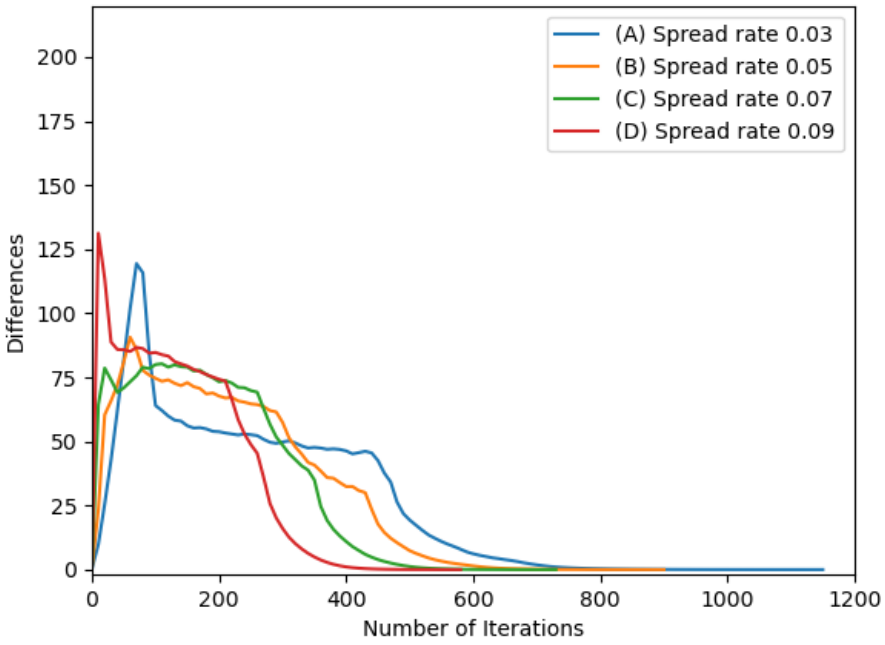


Figure 3.13: Stabilisation of the model for the four simulation scenarios with a compound effect of two environmental variables. (A) Spread rate equal to 0.03, (B) spread rate equal to 0.05, (C) spread rate equal to 0.07 and (D) spread rate equal to 0.09.

As the simulation proceeds, the differences between two sequential states gradually increase until they reach their pick. Then, the differences start to decrease until they reach a point in the simulation where they remain in a very low range corresponding to the stabilisation point; see Fig. 3.13. As observed in the previous experiments, the lower the spread rate, the longer the simulation will take to converge.

3.4 Environmental Suitability in Action

Species distribution models are widely implemented in ecological and biological studies to predict the potential geographical distribution of species. Several methods have been

used in the implementation of species distribution models, e.g. [116][13][117] [9]. Two types of species occurrence data can be used in the parameterization of these methods: (i) presence-only data that contains the locations where the species in the study was observed and (ii) presence-absence data containing the locations where species were not. In the parameterization of the methods that require presence-absence data, whenever the presence data is the only available, background (pseudo-absence) data are generated [118]. Two main approaches can be used to create pseudo-absence data: (i) select pseudo-absence at random in the study area; (ii) use a preliminary approach to restrict the selection of absence data in locations considered less suitable for the species [119].

No method always performs better than the others. Therefore, it is essential to use different methods to evaluate its performance.

The performance of the logistic regression algorithm (LR) and probability density function (DP) regarding the capacity to project the species' environmental suitability are compared. In addition to generating the suitability landscape, the simulation of species distribution in the projected environment is also performed. Departing from the suitability map produced by each approach, the species evolution on different environmental landscapes is simulated and discussed for two case studies of species with an economic interest.

3.4.1 Projecting the environmental suitability

The projection of the environmental suitability (suitability map) for each species was obtained in two ways: LR and DP implementations. Both methods use occurrence data to describe the relationship between the species and a set of EGVs.

Each EGV is a predictor variable in LR, and the occurrence data is the response. Since LR is a classification method, in addition to locations where the species were observed (presence data), it requires absence data. Since there is only presence data available, pseudo-absence data were generated. Pseudo-absence data were chosen randomly from the study area, excluding the locations where the species was observed [8]. To evaluate the effect of pseudo-absence data, five different samples were chosen. These samples include all the available presence-only data and the pseudo-absence data. Pseudo-absence data varies from 200 to 1000 points (the first sample contains 200 pseudo-absence data and the last 1000 points). Each sample is composed of the values of each EGV in each point and the corresponding response variable (0, 1). Based on the predictor variables and the values of the response variables, LR produces a model that predicts the species' probability occurrence, given the set of EGVs values at each point of the study area. For LR implementation, sci-kit-learn (a Python machine learning library) was used [120]. Parameters were defined as follows: *solver='liblinear'*, *random_state=0*, *tol=0.00001*, *max_iter=1500*, *C=0.050*, *penalty='l1'*. The choice of parameters for logistic regression was based on specific considerations for a small dataset, the need to avoid overfitting, and the ability to reproduce results.

For DP, presence-only data is sufficient to project the suitability map for the species. Presence-only data are used to calculate the mean and standard deviation of each EGV (EGVs' optimal suitability values). These EGVs values are then standardized in the form: $x_{i'} = (x_i - \mu)/\sigma$, where x_i is the value of an EGV in that location, μ is the mean and σ the standard deviation for that EGV. The probability density function is applied for each EGV, and the values are normalized in the interval from 0 to 1 (optimal). Then, the aggregation of each EGV value in each point produces the species' suitability map (predicted map).

3.4.2 The case of Apis mellifera honeybee

Apis mellifera is an Iberian honeybee, also presented in other locations. They can be found in both natural and artificial hives. The interest in studying the distribution of this species is related to the production and storage of honey and the construction of colonial nests from wax, widely used in the honey derivatives and cosmetics industries, respectively.

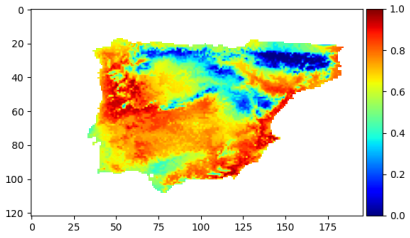
Four EGVs that influence the behaviour of *Apis mellifera* at $10 \times 10\ km$ were used: $BIO1$ (annual mean temperature), $BIO5$ (max temperature of warmest month), $BIO6$ (min temperature of coldest month), and $BIO15$ (precipitation seasonality). These EGVs were obtained from WorldClim, version 1.4 [121]. Presence-only data were collected in terrain.

3.4.2.1 Experimental results

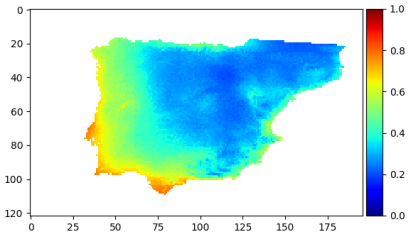
Figure 3.14 presents the suitability map obtained from the LR and DP. Visually, the difference between the suitability maps from LR and DP becomes more pronounced as the sample size increases.

Figure 3.15 shows the AUC for both methods. On average, the value of AUC for the LR is 0.67, whereas, for the DP, the average value is 0.61.

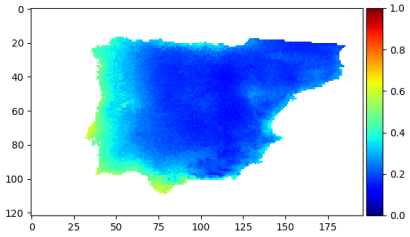
Figure 3.16 shows the distribution map of the species and the corresponding ROC curve for the two methods. The initial population was set to 100 randomly chosen grid cells in the study area. The life cycle parameters were defined as follows: $birth\ rate\ :\ 0.5$, $death\ rate\ :\ 0.2$ and $spread\ rate\ :\ 0.3$. These parameters ensure that the species can both reproduce and expand, as the birth rate exceeds the death rate, and the spread rate facilitates the dispersal of the species across the study area. Additionally, the simulation runs 200 times, which is sufficient to observe the stabilization of the population dynamics and the distribution pattern. The value of AUC for the distribution map of the species using DP ($AUC = 0.60$) is less than the average AUC of the suitability map. In contrast, the AUC for the species distribution using LR ($AUC = 0.72$) is greater than the average AUC of the suitability map with LR.



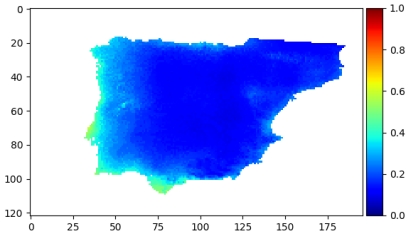
(a) Suitability Map - Probability density function



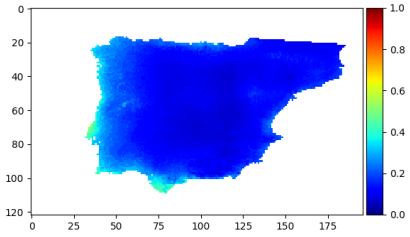
(b) 200 pseudo-absences (LR)



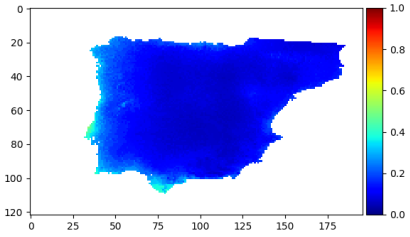
(c) 400 pseudo-absences (LR)



(d) 600 pseudo-absences (LR)



(e) 800 pseudo-absences (LR)



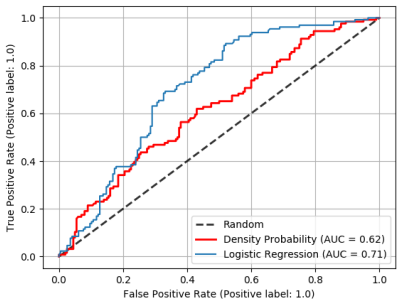
(f) 1000 pseudo-absences (LR)

Figure 3.14: Suitability Map obtained by Probability density function (Fig. a) and Logistic Regression (Fig. b, c, d, e, and f) with different quantity of samples of Apis Mallifera Honeybee. All the figures with the 135 occurrence data, varying the quantity of pseudo-absence data from 200 to 1000.

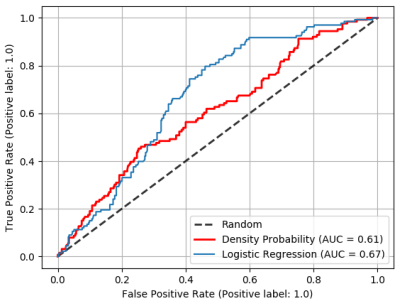
3.4.3 The case of *Arbutus unedo L.*

Arbutus unedo L. is a Mediterranean species found in large quantities in Portugal and the Mediterranean. The study of the distribution of this species has a significant interest in Portugal for economic reasons. Its fruit is used to produce spirit drinks, considered the primary source of revenue for forest owners [122].

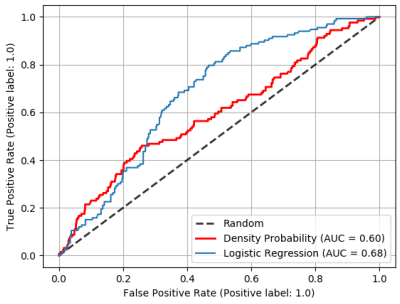
In agreement with the study presented in [122] nine EGVs have the greater influence on



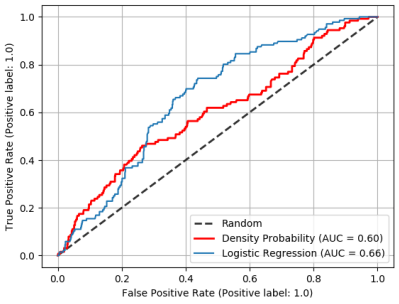
(a) Sample of 200 points



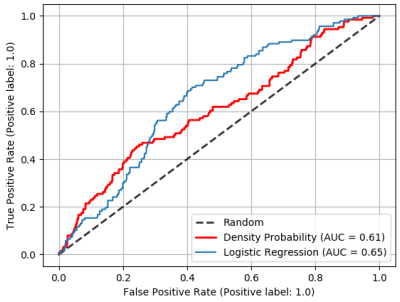
(b) Sample of 400 points



(c) Sample of 600 points



(d) Sample of 800 points

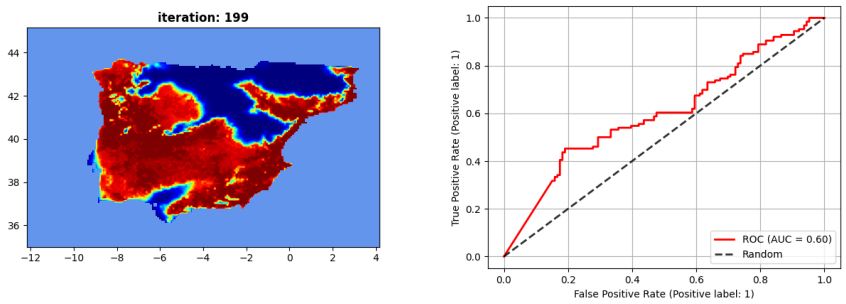


(e) Sample of 1000 points

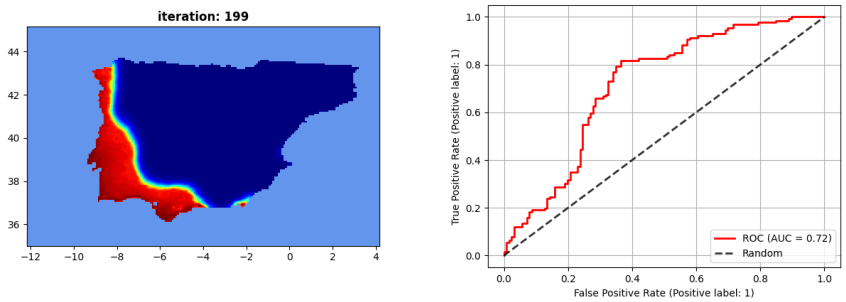
Figure 3.15: ROC Curve - Comparison between Logistic Regression algorithm and Probability density function for *Apis mellifera* Honeybee.

the behaviour of *A. unedo* were selected ($1 \times 1\ km$): seven climatic variables ($BIO1$: annual mean temperature, $BIO2$: mean diurnal range (mean of monthly (max temp - min temp)), $BIO5$: max temperature of warmest month, $BIO9$: mean temperature of driest quarter, $BIO15$: precipitation seasonality, t_{max}: maximum temperature, t_{min}: minimum temperature) from the WorldClim database 1.4 and two topographic variables ($slope$ and $altitude$); altitude was obtained from the Global Multi-resolution Terrain Elevation Data 2010 [123]. The slope was generated from the altitude using the GDAL/OGR library [124].

In this case study, 318 locations where the species was observed were used (presence-only data), with different pseudo-absence data randomly chosen in the study area.



(a) Distribution Map and ROC curve (Density Probability)



(b) Distribution Map and ROC curve (Logistic Regression)

Figure 3.16: Distribution Maps obtained from both logistic regression method and probability density function from SDSim for *Apis mellifera*.

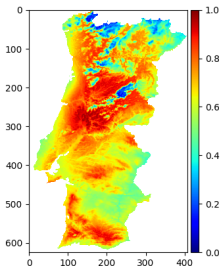
3.4.3.1 Experimental results

Figure 3.17 shows the suitability maps obtained from DP and the suitability maps obtained from each sample from LR. Visually, the difference between the suitability map from DP and the suitability maps from LR increases as the pseudo-absence increases.

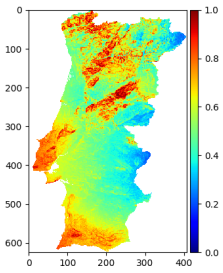
The AUC is calculated with the Receiver Operating Characteristic (ROC) to analyse how these two methods predict the species suitability map.

Figure 3.18 presents the AUC and ROC to compare the classification performance of LR and DP. For LR, the value of AUC is on average 0.70, whereas the value of AUC for DP is on average 0.62.

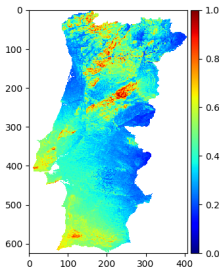
Figure 3.19 shows the distribution map of the species and the corresponding ROC curves for the two methods. The sampling strategy consists of choosing an equal quantity of both presence-only and pseudo-absence data, and these pseudo-absence data are chosen randomly from the study area. In addition, the distribution of the species in the predicted environment was simulated according to the AB-SDSim model. The initial population was set to 100 random grid cells in the study area; the life cycle parameters were defined as follows: $birthrate : 0.5$, $deathrate : 0.2$ and $spreadrate : 0.3$, and the simulation runs 200 times. The value of AUC for the distribution map of the species using DP (AUC=0.62) is equal to the average AUC of the suitability map. In contrast, the AUC for the distribution



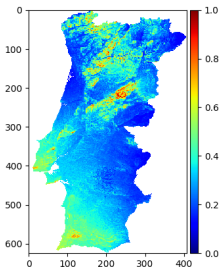
(a) Suitability Map - Probability density function (DP)



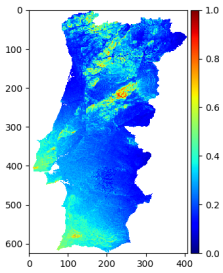
(b) 200 pseudo-absences (LR)



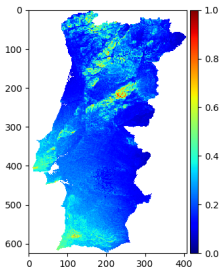
(c) 400 pseudo-absences (LR)



(d) 600 pseudo-absences (LR)



(e) 800 pseudo-absences (LR)

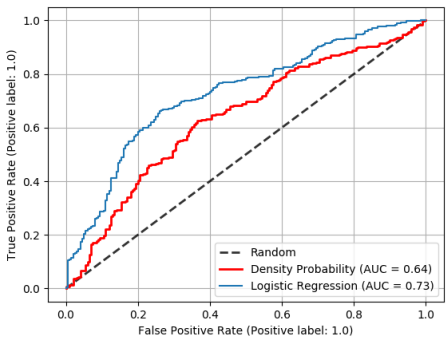


(f) 1000 pseudo-absences (LR)

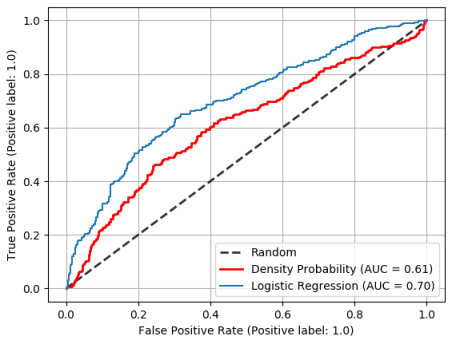
Figure 3.17: Suitability Map obtained by probability density function (Fig. a) and Logistic Regression (Fig. b, c, d, e, and f) with different quantity of samples of *A. unedo*. All the figures with the 318 occurrence data, varying the quantity of pseudo-absence data from 200 to 1000.

of the species using LR (AUC=0.68) is less than the average AUC of the suitability map with LR.

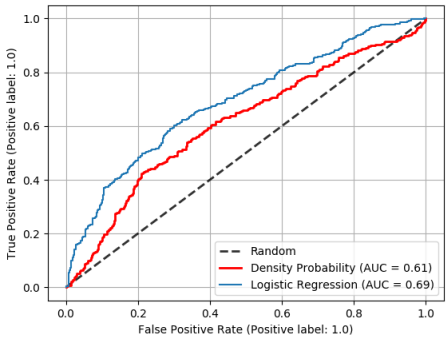
It is possible to notice a high concentration of optimal locations for each species on the maps obtained by DP. Despite both methods ensuring similar patterns on maps, maps obtained by DP produce more suitable regions, allowing the species to reproduce and colonize more quickly. Per the performance measures, DP produced poor results for both case studies. However, from a biological point of view, the DP approach seems to be the one that closely agrees with the real data collected in the field.



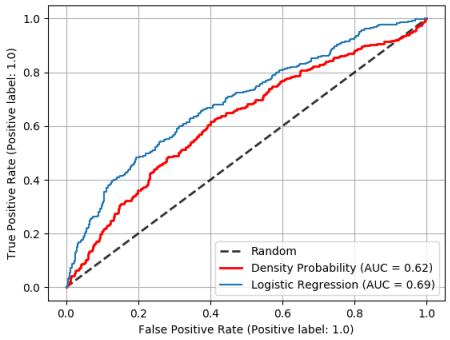
(a) Sample of 200 points



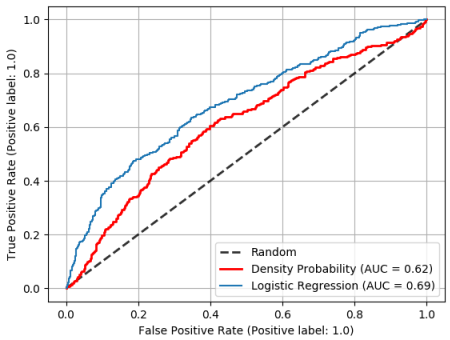
(b) Sample of 400 points



(c) Sample of 600 points



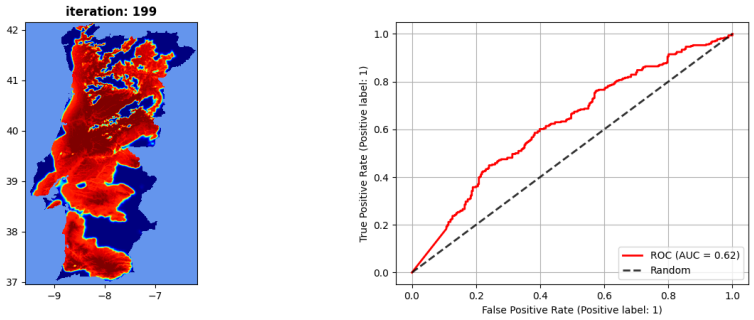
(d) Sample of 800 points



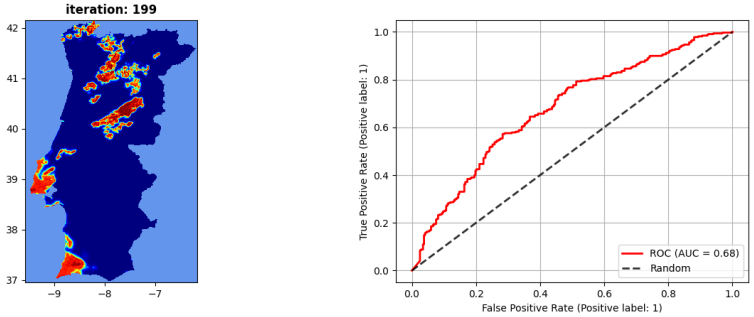
(e) Sample of 1000 points

Figure 3.18: ROC Curve - Comparison between Logistic Regression algorithm and Probability density function for *A. unedo*.

On the other hand, the suitability maps obtained by LR have many more places where the species has difficulty surviving. Values of the suitability maps are lower, causing the species to take longer to spread in the environment. This effect results from the approach used to select pseudo-absence data (randomly selected). When selecting pseudo-absence data, several suitable locations for the species are potentially classified as absences. LR, and any regression approach, fit the model with these data. Therefore, the selected approach used to generate pseudo-absence data greatly influences LR performance. De-



(a) Distribution Map and ROC curve (Density Probability)



(b) Distribution Map and ROC curve (Logistic Regression)

Figure 3.19: Distribution Maps obtained from both logistic regression method and probability density function from SDSim for *A. unedo*.

spite its wide use, one can say that from the biological standpoint, the random selection of pseudo-absence data is not the best approach.

Another factor to consider that impacts LR performance is the sample size (quantity of pseudo-absence data). According to the results, the number of pseudo-absences closer or equal to the number of presence-only data turned out to be a good approach.

Overall, both methods (LR and DP) performed better than the random classifiers (AUC=0.5). However, in these case studies, LR performed better than DP.

3.5 Remarks and Discussion

The effects of an agent-based model's parameters in the spatial distribution of species were analysed by implementing a model to deal with a heterogeneous environment represented by a combination of (environmental) variables of interest. A parametric study was performed to obtain the parameters combination that fits the purpose of the model. The results showed that in addition to the environmental conditions, the combination of the model parameters significantly impacts its results. The study is limited in the sense that the environment of the model was not natural; however, the initial conditions of the presented experiments are well aligned with some real local environmental constraints

that it is intended to explore in future studies for the prediction of the geographical distribution of biological species (both flora and fauna) with an economic interest in a setup of environmental uncertainty.

Model behaviour and outputs are deeply coupled with the chosen parameters and environment. The parameters of the reported model are entirely independent of each other in the sense that any adjustment made to any parameter does not affect the value of the remaining parameters. However, any slight change in a subset of parameters can result in drastic changes in the overall behaviour of the model; The same happens if the environmental conditions change.

It is necessary to perform a specific parameters analysis and verify the environmental variables that compose the environment and its values. It is a well-known fact that comprehensive analysis of the output-to-input variability is an essential step during the development of an agent-based model [125]. Parameters analysed in this study have each an effect on the model. However, it is necessary to consider the effect of the different parameters on the model's output instead of only these parameters individually. Discarding the effect of either one of the parameters will jeopardise the ability to explain the model's output.

One aspect to consider is the distinction between the birth rate and the death rate. To observe reproduction, it is essential to have a significant distinction between these two rates, fixing the values of the birth rate always more significant than the death rate. This is the only case that the species can survive and reproduce. Therefore, without the spread rate, there is no way for the species to spread (and colonise) to other grid cells in the environment. Once the birth and death rates are chosen, the spread rate determines if the species have a propensity to consolidate the occupied places or if they have a greater predisposition to colonise new territories.

The choice of parameters will always constrain the desired results. When using a model like the one described in this study, it is necessary to analyse several scenarios to find the parameters' combinations that answer the purpose of the reference model.

In addition, the performance of LR and DP in projecting species' environmental suitability was compared. Two species were used to perform two case studies: 1) The distribution of *A. unedo* and the distribution of *Apis mellifera*. The performance of both methods was compared. Only presence data were available for both case studies; consequently, pseudo-absence data were generated and used to assess the methods. The cardinality of pseudo-absence data has impacted LR's performance significantly. Considering the usage of presence-only data to project environmental suitability, unlike LR, which uses, in addition to presence-only data, pseudo absence data (when absence data is missing), it was expected that DP would perform better than LR. However, strictly numerically speaking, in both case studies, LR performed better than DP in describing the relationship between occurrence data and environmental conditions. Overall, the results obtained by the two

methods presented similar patterns, as the significant abundance of the species was observed in the same locations.

Chapter 4

Performance Considerations

4.1 Introduction

There are some aspects to consider when using ABM to model and simulate species distribution, such as the environmental conditions that influence the species' life cycle, the available resources (i.e. food, water), and the dimension of the environment where species exist (study area). The area under the study can be small or large, and the coarseness of the simulation stage can be widely heterogeneous from simulation to simulation. When the dimension of the environment is substantial, and the required level of detail for the simulation is high, the simulation can be quite time-consuming and requires considerable computation power, becoming necessary to implement strategies to narrow down the time requirements.

Computational cost is another issue identified (chapter 2) in ABM&S in the spatio-temporal distribution of species that significantly impacts the models' performance. To address this issue, a strategy to parallelize the AB-SDSim model described in the previous chapter is presented. This strategy involves running the model in a multi-stage synchronous mode, reducing processing time while ensuring no significant information loss. The size of the overlapping section sent with each stage subset and the number of iterations between each global synchronization is parameterized. The findings indicate that it is possible to achieve a good trade-off between the size of the internal computation and the data transfer time while maintaining the algorithm's correctness.

4.2 Parallelization Approaches

One parallelization approach concerns dividing the computation at the agent level, where each processing element (PE) is responsible for a set of agents. Another approach consists of dividing the computation at the spatial environment level where each PE is responsible for a set of grid cells [98]. In both cases, handling the interaction between agents and their movements requires communication and synchronization, which constitutes the limiting factor for obtaining a scalable parallel model [98].

Existing proposals for parallel implementations of spatial ABMs range from multi-threaded implementations in shared memory architectures [126][127] to implementations in Graphing Processing Units (GPUs) [128][129][130]. However, most proposals are based on dis-

tributed memory programming models, which can be scaled to thousands of cores. These works include frameworks such as FLAME [131][132] and Repast HPC [133][134] that use MPI for inter-process communication. The work presented in [135] implements an ABM in the Apache Spark framework trying to take advantage of its in-memory computation model. In all these works, the main performance bottleneck remains the communication cost.

4.3 The Parallel Model

The model explores different hypotheses on the space-time distribution of species. It comprises an environment disposed of in a regular rectangular grid, where each grid cell stores its suitability and the number of specimens presented in the grid cell. Figure 4.1 shows in more detail the characterization of the environment.

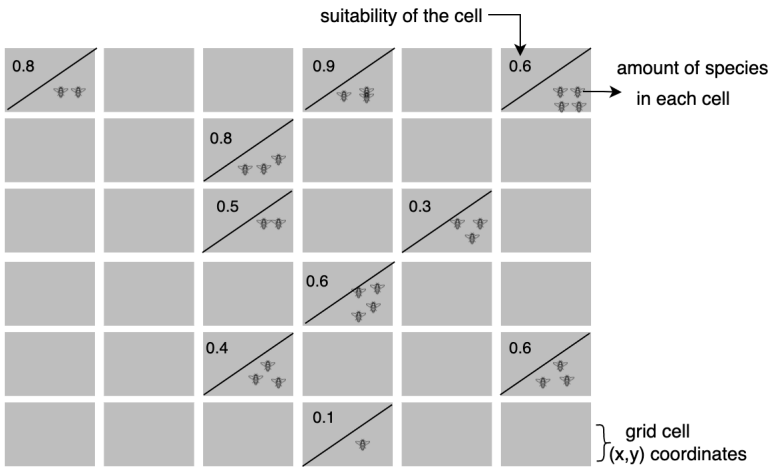


Figure 4.1: Characterization of the environment.

4.3.1 Process overview and schedule

Figure 4.2 presents the main steps of the simulation process. It starts by initializing the environment stage (hereafter referred to as patches) after setting the model's parameters. In each iteration, a birth and death rate affecting the percentage of species' occupation is applied to each cell. These rates are constrained by the suitability value of the grid cell. The spread of the species occupancy percentage occurs through the neighbourhood of the cells. The model implements Moore neighbourhood [113]. Therefore at each iteration, each cell transfers an amount of material to its neighbours, according to a spread rate. Algorithm 4 describes the updating mechanism that simulates the spatial dynamics of the spreading of a natural organism guided towards spatial self-organization.

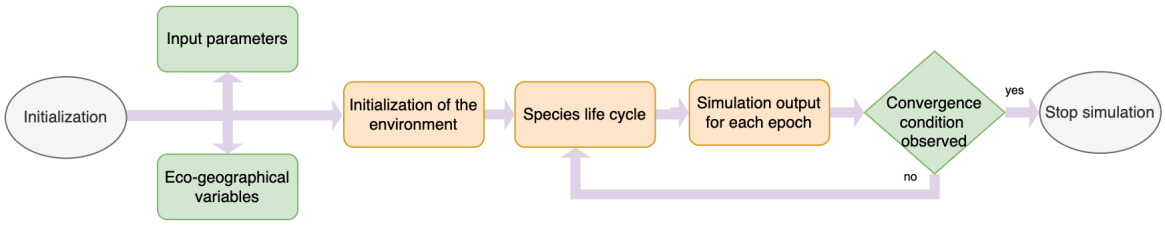


Figure 4.2: General steps of the simulation process

Algorithm 4: Species update algorithm. The reproduction method contains the exchange policy of the cell with its neighbours depending on its birth, death and spread rates.

procedure $Distribution\_updatepatches, previous\_patches, steps$
for k in $steps$ **do**
  for i, row in $patches$ **do**
    for $j, patch$ in row **do**
      $neighbours = patch.find\_neighbours(previous\_patches, i, j)$
      $patch.reproduce(previous\_patches[i][j], birth\_rate,$
                $death\_rate, spread\_rate, neighbours)$
    end for
  end for
end for
return $patches$
end procedure

The sequence of patches (different states) obtained during the iterative operation of the algorithm 4 is called the evolution. As defined in [136], evolution is the result of the simulation task, representing the process under simulation. If the process converges to a stable global state, then the algorithm evolution has a termination. If that is not the case, then the evolution is infinite, exhibiting oscillatory or chaotic behaviour. Algorithm 5 represents the evolution snapshot for a fixed amount t of epochs.

Algorithm 5: Evolution of the simulation task.

Initialize $patches, previous\_patches$, set $epochs$, $output\_interval$ and let $steps \leftarrow 1$
for t in $epochs$ **do**
  $patches$ = $Distribution\_update(patches, previous\_patches, steps)$
  if $mod(t, output\_interval) == 0$ **then**
    $\Delta = sum(abs(previous\_patches - patches))$
    Copy $patches$ into $previous\_patches$

Alternatively, the evolution can be terminated once a convergence condition is observed. In that case, the evolution stops, i.e. for all that matters, it has converged, whenever the cell-by-cell difference, Δ drops below a threshold τ given by

$$\tau = \varepsilon \times N_c, \quad (4.1)$$

where N_c is the number of cells in the environment, and ε is the amount of admissible error between any given pair of corresponding cells (or minimum distinguishability level). This threshold value can also be used as a reference value to analyse if two different evolutions exhibit, or not, a similar convergent behaviour.

4.3.2 Parallelization strategies

The adopted parallelization strategy assigns each PE a set of grid cells. The species distribution map was divided into stage subsets (hereafter referred to as strips, since in our implementation, any subset encompasses all the columns of the main stage), each with a dimension given by the number of rows in the map divided by the number of available processes.

While pursuing high efficiency, breaking the equality between the initial sequential evolution and that of a decomposed parallel algorithm is possible. Therefore, to guarantee the sequential evolution's equality to its parallelized version, a set of correctness conditions, see [136], must be assured during the interaction between processes. Stated, "the problem is to organize the parallel operation in such way that each domain interacts with the adjacent ones by exchanging data that is needed to be used in one of them for computing the next-states in the other". Moreover, any given cell must be updated only once per iteration. Algorithm 6 presents the synchronous parallelization used in reported experiments.

Algorithm 6: Synchronous parallel algorithm.

Initialize $patches, previous\_patches$
Set $number\_processes, overlap, steps, epochs, output\_interval$
for t in $epochs$ **do**
  $strips = Build\_strips(patches, number\_processes, overlap)$
  $previous\_strips = Build\_strips(previous\_patches, number\_processes, overlap)$
  for each $strip\_s$ **do**
    Run process $P_s(Distribution\_update(strip_s, previous\_strip_s, steps))$
  for each process P_s **do**
    Receive and fuse the P_s results into $patches$
  if $mod(t, output\_interval) == 0$ **then**
    $\Delta = sum(abs(previous\_patches - patches))$
  Copy $patches$ into $previous\_patches$

The equality between sequential and parallel versions is only assured if all the cells whose values are necessary for updating any given cell of the strips are available. In this case, using a Moore neighbourhood means that each strip must be extended to perform a parallel iteration by adding a border region (overlap) of at least one row. When dealing with parallel processes, one must also be aware of the inter-process communication costs, which may be reduced by preventing frequent synchronisation between processes. It can be archived if, instead of joining the strips after each iteration, the different stages are al-

lowed to evolve independently in their parallel processes for a given number of evolutive steps. Once again, the equality between versions is only achieved if the overlap between strips (or partial stages) is composed of a number of rows equal to or greater than the number of steps.

The strips are extended with adjacent rows from each neighbouring strip to deal with the border region, t. In what follows, those common regions are referred to as overlap, as shown in figure 4.3 where domain decomposition is emphasised.

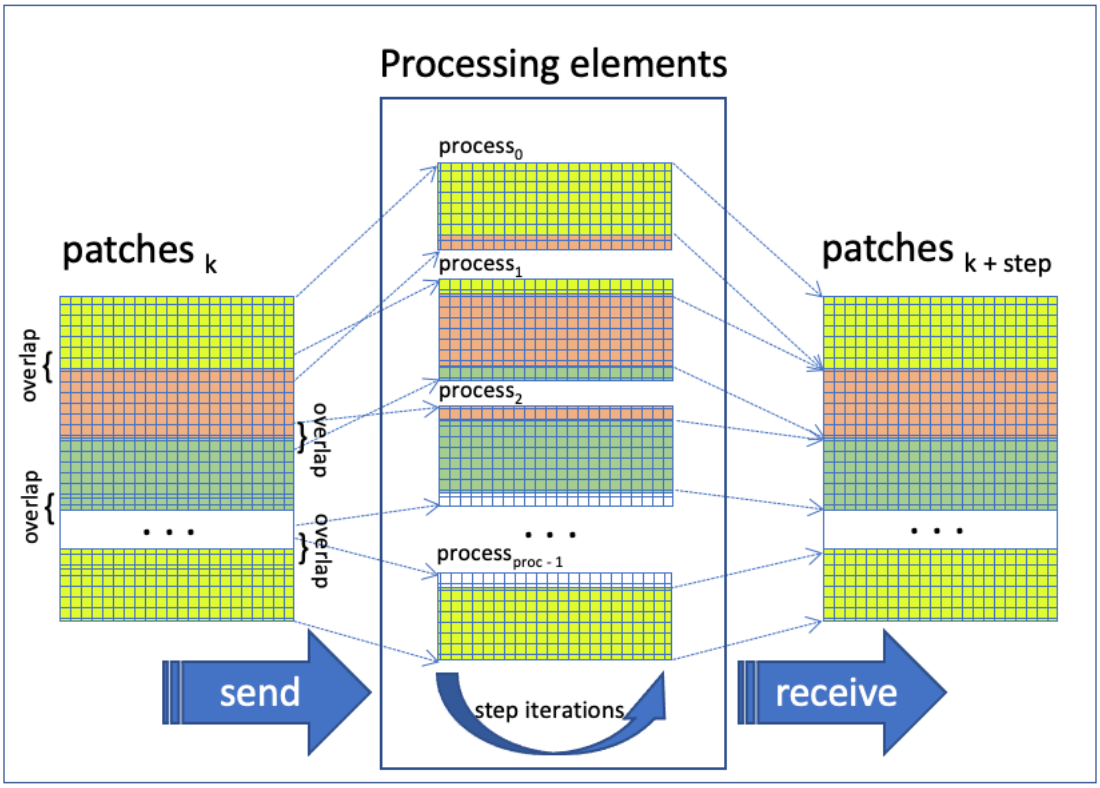


Figure 4.3: Parallelization of the spatial environment. Decomposition of the study area into a set of overlapping strips for parallel processing.

The data is divided into stage subsets in the body of the simulation cycle. Each chunk of data (including a strip of the current distribution map, $patches_k$, and its overlapping sections) is sent to a process running in parallel with the ones responsible for the other data chunks. Each process will perform over its data the number of iterations given by a step parameter. At the end of the parallel phase, the arising strips are sent back to the main process, producing the next map instance ($patches_{k+step}$). It means that, at this phase, the overlapping sections are discarded since their main purpose is to act as a buffer, updating the frontier cells in the starting iterations of each process. The core strips of the distribution map are fused, and the main process continues as described.

4.4 Experimental Results

The reported experiments simulations using both sequential and parallel implementations of the model were performed. Initially, the simulation with the sequential implementation was performed, followed by the simulation with the parallel implementation. Execution times were computed in a machine with the following hardware and software configurations: a) Operating System: Linux Ubuntu Desktop version: 18.04.5 LTS 64bits, b) RAM: 64GB, c) Processor: Intel® Core™ i9-9900X CPU @ 3.50GHz × 20, d) Python version: 3.7.2.

For these experiments, the spatial distribution of the african honeybee Apis mellifera in the Iberian Peninsula was simulated. The set of variables of interest for the environment was composed of four EGVs, which are the input values determining the suitability of each cell.

In order to facilitate the comparison between the two implementations (sequential and parallel), the number of species in each grid cell was initialised using the same seed for each simulation. Table 4.1 shows all the parameters of the model and the values used for both sequential and parallel implementations. For the parallel implementation, 12 parameter combinations were chosen: a step of 10 varying the frontier with the values (4, 6, 8, 10); a step of 50 varying the frontier with the values (20, 30, 40, 50); and step of 100 varying the frontier with values (40, 60, 80, 100). Therefore, for the parallel implementation, 12 different simulations were performed. For the sequential implementation, results at each timestamp were saved; for the parallel implementation, results were saved according to the chosen step. The number of processes was fixed to 12 to analyse the algorithm's behaviour when the processes were subjected to a varying workload directly related to the dimensions of the data used in the experiment.

Table 4.1: Models' parameters.

Parameters	Value
Initial population	200 000
Number of epochs	200
Cells capacity	1000
Output generation interval	according to the step
birth rate	0.9
death rate	0.2
spread rate	0.6
Type of neighbourhood	Moore
Environment dimension	1210×1940

4.4.1 Quasi-equality behaviour

Aiming at the determination of the rate of degradation resulting from the reduction of the border region, the differences between the initial sequential evolution and the parallel evolution with different process steps (period, where each strip evolves independently of the remaining ones, were analysed, see Fig. 4.3). For each fixed number of $steps$, experiments were conducted for different overlap levels: 100%, 80%, 60% and 40% of the performed steps.

For a fixed number of sequential epochs (200 in the reported experiments), the number of steps directly influences the number of synchronizations (with its inherent communication costs). In contrast, the level of overlap (determining the overall number of cells treated by each process) significantly influences the processes' workload.

As previously noted, the equality between versions is only archived if the overlap between strips (for partial stages) is composed of a number of rows equal to or greater than the number of steps.

Table 4.2 shows the sum of differences (cell-by-cell comparison) between each parallel combination (different steps and overlaps) and the sequential implementation in the same set of iterations.

Table 4.2: Cell-by-Cell differences, Δ, between the results of the sequential implementation and the results obtained from the parallel implementation in the iterations: 50, 100, 150 and 200.

Steps	Overlap	50	100	150	200
10	10	0.00067	0.00067	0.00067	0.00065
10	8	0.14313	0.11767	0.10852	0.10361
10	6	5.20148	3.82719	3.41619	3.24228
10	4	58.60432	42.44981	38.00387	36.54078
50	50	0.0	≈ 0	≈ 0	≈ 0
50	40	≈ 0	≈ 0	≈ 0	≈ 0
50	30	≈ 0	≈ 0	≈ 0	≈ 0
50	20	0.00372	0.00170	0.00099	0.000706
100	100	-	0.0	-	0.0
100	80	-	≈ 0	-	≈ 0
100	60	-	≈ 0	-	≈ 0
100	40	-	≈ 0	-	≈ 0

For $\tau < 1E^{-5}$, see (4.1), the error values were denoted as "≈ 0".

As expected, table 6.1 shows that the differences between the sequential and parallel evolution increase with the reduction of the level of overlap between stage subsets. Those differences are more apparent at the simulation's earliest stages due to the simulation's convergence to a stable state. Interestingly enough, a somewhat counter-intuitive observation should be noted. In this model, many parallel steps tend to reduce the overall inequity between the parallel and the sequential evolution. Even when there is a signif-

icant relative gap in the number of rows necessary to guarantee an equal evolution, the local convergence of the algorithm can circumvent this lack of data as long as enough processing steps are allowed. The following section reports some performance indicators for these and other selected experiments.

4.4.2 Performance comparison

As a way to analyse the performance of the proposed parallel strategy, the improvements in the speed of execution for several parallel implementations with different parameter combinations were measured. In the reported experiments, the speedup of each parallel configuration P against the corresponding sequential version S is given by

$$S_P \;=\; T_S \;/\; T_P \;, \quad (4.2)$$

where T_S is the execution time of the sequential evolution and T_P is the execution time of the assessed parallel configuration.

The impact of varying the step and the overlap parameters for a fixed number of processes and several environment dimensions was studied. The speedups of a chosen configuration, when the number of processes varies from 2 to 20, were also calculated. Figure 4.4 shows the speedups obtained when simulating each parameter combination, considering three environment maps dimensions (1210×3880), (1210×1940) and (1210×970), i.e., the initial map (1210×1940) was extended and shrunken along its second dimension. The 12 processes ran simulations over 200 epochs. That number of iterations has shown to be enough to reach convergence.

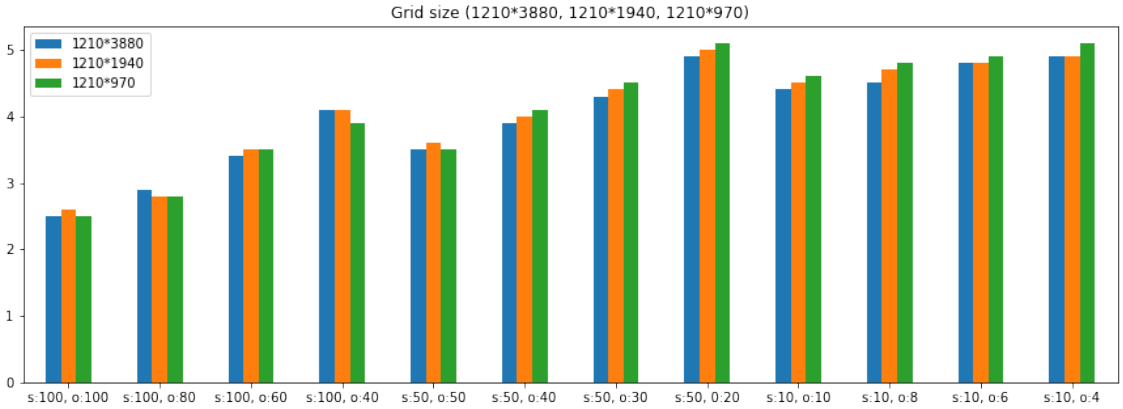


Figure 4.4: Speedups obtained with maps of a different number of columns, varying the number of steps *s* (10, 50 and 100) and for each step, the overlapping (*o*) size varies from 100% to 40% of the number of steps.

In figure 4.4, it is possible to observe that the speedups are approximately the same for the three maps. The small map speedups have insignificantly higher values for the smallest step values. Increasing the number of columns has no impact on the speedup. When

observing the figures with the same step values, as can be expected, the speedup increases when the overlap reduces. Less data redundancy implies less communication and less processing. As seen in the previous section, the overlapping value must be high enough to avoid information loss, but the higher it is, the worse the performance will be. In the studied cases, the best speedups are obtained for the pair $steps = 50$ and $overlap = 20$ (i.e., 40% of the step value). The variation of both dimensions of the map was also studied, considering this pair of values. Figure 4.5 shows the speedups obtained when running simulations along 200 epochs in 12 processes for three maps with dimensions (605×970), (1210×1940) and (2420×3880). Overall parameters stayed the same as before, and the values of step and overlap were fixed at 50 and 20, respectively.

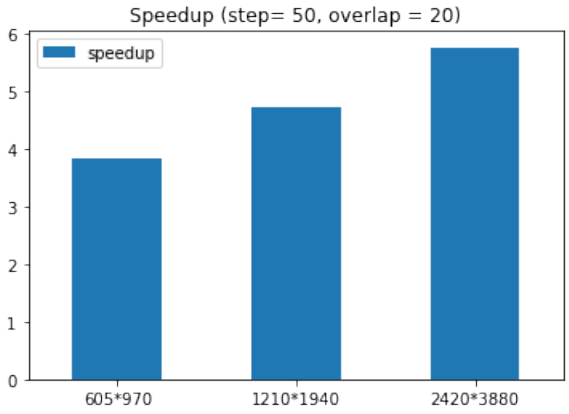


Figure 4.5: Speedups obtained when increased data in both dimensions using step = 50, overlap = 20, 200 iterations and 12 processes.

As can be seen, the speedup increases when the dimension of the map increases. When the first dimension of the map is increased, and the same number of processes is kept, the number of rows in each data strip increases. Keeping the same overlap value means less data redundancy when the size of each strip grows.

Finally, figure 4.6 shows the speedups obtained when the number of processes, p, varies from 2 to 20. The initial map with dimension (1210×1940) was used for step 50 and overlap 20. The values of the remaining parameters were kept.

The speedups increased steadily with the number of the available parallel processes to a value nearly five times faster than the sequential implementation, see Fig. 4.6.

The described parallel decomposition can preserve absolute equality with the sequential evolution, provided that a widely enough border region is shared between contiguous stage subsets. If the border dimension is less than the number of parallel inner-process steps, equality is potentially broken.

Figure 4.7 shows the evolution of the sequential implementation side-by-side with the most different parallel evolution ($step = 10$, $overlap = 4$).

It is worth noting that in Fig. 4.7 some snapshots of the evolution are depicted only in

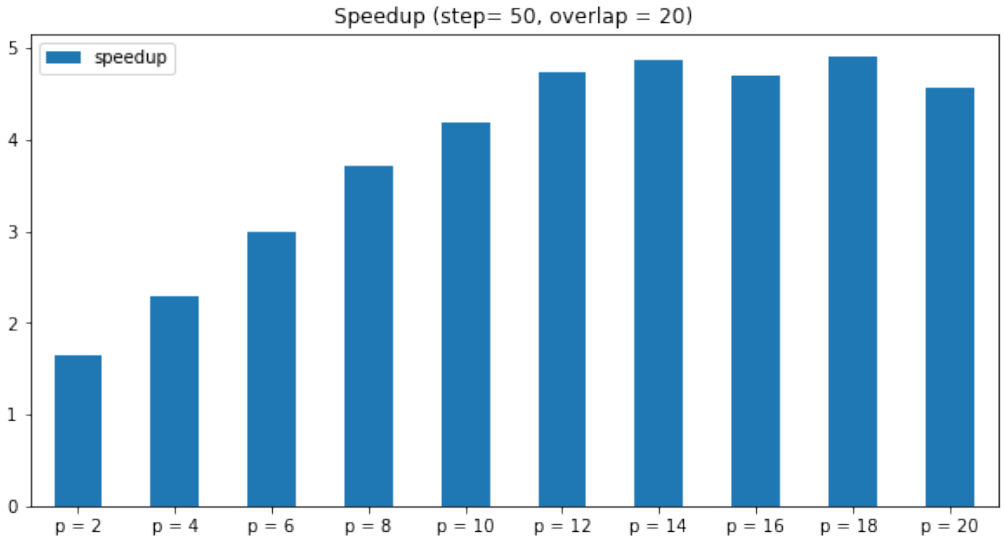


Figure 4.6: Speedups obtained for the map dimension (1210 x 1940) when the number of processes, p, varies from 2 to 20, using step = 50, overlap = 20 and 200 iterations.

the initial steps. During this period, the major differences are accounted for. As a side note, notice that it is difficult to spot any differences in plain sight, even at these earliest stages. However the simulation runs until the convergence criteria (viz. the differences between consecutive states drop below $\tau = 1E^{-5} \times N_c$, see (4.1), for a number of epochs) is satisfied.

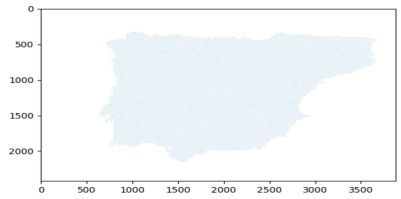
On the other hand, for the same relative level of overlap, given the local convergence characteristics of this particular model, any loss of equality is less noticeable when more inner-process steps are allowed. Thus if one is ready to relax the equality constraint for the parallel version, it is possible to obtain a model presenting indistinguishable results from the sequential version while achieving at least a good speedup.

According to the performance results, the parallel implementation gained a speedup of approximately 5. Different parameters' combinations of the parallel implementation constraint both the model's accuracy and performance.

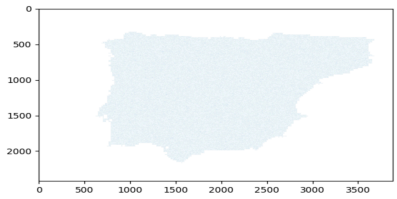
4.5 Remarks and Discussion

In this chapter, a parallelization strategy of an agent-based model of the spatial distribution of species was proposed aiming at a good trade-off between the synchronization requirements and the amount of data redundancy necessary to achieve the equality between the parallel and the sequential evolution.

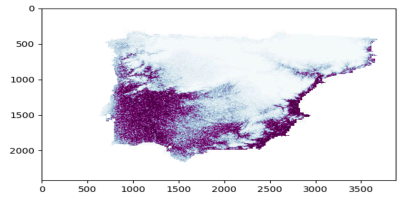
The adopted strategy unveils the effects that the number of parallel evolutive steps and the size of the overlap between stage subsets have on the algorithms' performance. The relation between these two parameters is explored to find the best parameter combination



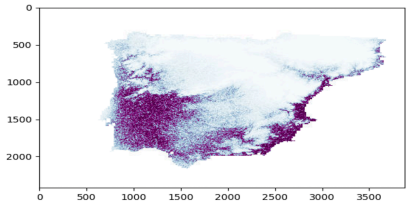
(a) Sequential, $t = 0$



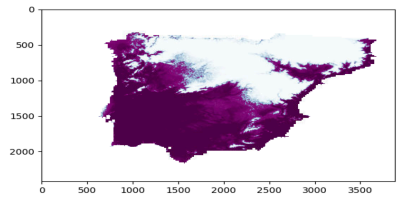
(b) Parallel, $t = 0$



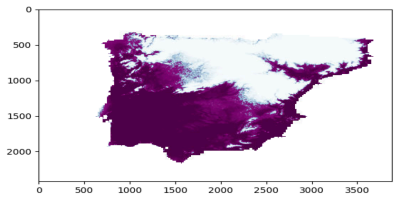
(c) Sequential, $t = 10$



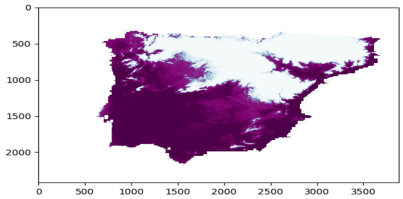
(d) Parallel, $t = 10$



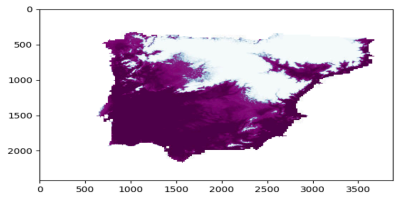
(e) Sequential, $t = 30$



(f) Parallel, $t = 30$



(g) Sequential, $t = 40$



(h) Parallel, $t = 40$

Figure 4.7: Sequential evolution (on the left) and a non-equal parallel evolution (on the right) with 10 inner-process steps and only 4 overlapping rows.

that ensures increased speedups.

According to the empirical results, there is a scale opportunity to model more significant problems with almost negligible errors. Due to the local convergence of the model (resulting from the Moore neighbouring in the composed stages), it is possible to attain almost indistinguishable results from the ones of the sequential version. It is verified even when the synchronizations are scarce, and the overlap is kept at a parsimonious level, hence accomplishing significant performance gains.

Thus, a clear line of future developments is related to the scale-up of the implementation to a distributed setup to tackle more significant cardinality problems. However, for a sounder statistical evaluation and deeper inspection of the proposal, aiming at the generalization to other kinds of models (with different, non-regular types of iterations between stage-subsets), the performance of the algorithm must be assumed over averaged metrics for multiple runs in a set of carefully chosen descriptive parameters.

Chapter 5

A First Approach on the Temporality Issue

5.1 Introduction

Several methods and techniques have been used to implement SDMs [13][117][137][138], showing more and less suitable regions for the species' survival in a map. However, studying such a distribution over time becomes much more difficult.

A changing climate reshapes bio-geographic patterns, and as a consequence, conservation planners require reliable methods to project future distributions of species that allow for prioritising conservation efforts [139]. Spreading a particular species from the projection of an SDM over future environmental conditions could lack important information. For example, a disequilibrium can arise if species' ranges are shaped by biotic interactions that are independent of climate [140][141]. More critical is that time scaling is an issue rarely considered in ecology and conservation biology [142], although it becomes crucial in several contexts in ecology. For example, recently, scientists have investigated the ecological consequences of climatic changes over different temporal and spatial scales. These studies have highlighted confounding methodological issues when the generalisation of biological patterns could be impeded due to inappropriate time scales during analyses [143].

In ecology, time is assumed to scale in years or generations, although the World Conservation Union (IUCN) uses a mixture in its categorisation system for threatened and vulnerable species [144]. Time has been considered a niche dimension along which organisms may segregate to minimise overlap in the use of a resource for competing species [142]. Time can also be another parameter, occasionally included as an exponent or a subscript, in equations describing conceptualisations of patterns. However, timely treatment is inconsistent in ecology (e.g., discrete vs continuous). Sometimes, ecologists and conservation biologists are reluctant to give time such importance as other disciplines, for example, physics, where it is a central focus. A possible explanation would rely on the fact that approaches to treat time as a variable to include in an SDM is a challenging task reserved for modelled ecological processes, such as physical processing (e.g., larval or seed dispersion).

This chapter presents a novel approach to mapping computational and ecological time using the agent-based distribution model approach to address this issue. Reported past environmental conditions are used as the simulation's initialisation point, allowing the system to evolve into current environmental conditions by interpolation. Species distribution responses to climatic changes are registered and evaluated to analyse the number

of iterations needed to reach a dynamic equilibrium of occupancy in such a given climatic scenario.

5.2 Rationale

The model described in chapter 3 was applied to simulate the distribution of Apis mellifera honeybee African lineage and the *A. unedo* in the Iberian Peninsula in two different geological times, namely, the glacial period (approximately 10000 BP) and the current period. The selected EGVs for both *Apis mellifera* and *A. unedo* were the same as the EGVs used in the experiments of the chapter 3, section 3.4.

Three simulation scenarios were performed for each species by fixing the birth and death rates and varying the spread rate. Life cycle parameters were set up as follows: ($birth\ rate = 0.6$, $death\ rate = 0.2$, $spread\ rate = [0.2, 0.3, 0.4]$). The chosen values allow the species to reproduce and spread to suitable locations.

The approach for establishing a reasonable approximation between each epoch of the agent-based model and the corresponding geological time consists of three main phases: 1) initialisation, 2) estimation and 3) prediction. It is assumed that we have access to EGVs measured in two different, sufficiently far apart and distinct, moments in time [1].

In the initialisation phase, the species distribution is simulated in the past environmental conditions until a stable distribution is observed. This stable distribution in the past is then used to initialise the species simulation in the current environmental conditions. Usually, an abrupt change in the environmental conditions occurs, and the species distribution is simulated until the system stabilises. The number of iterations needed to stabilise the distribution of the species in the current environmental conditions, $stabilisation\_present$, is registered. Stabilisation is determined by the cell-by-cell comparison index [115].

An indistinguishability level between them gives the directly matched cells. Therefore, two cells are considered indistinguishable if the following condition holds:

$$|current_{ij} - previous_{ij}| \leq e_{adm}, \quad (5.1)$$

where $current_{ij}$ is one cell in the current state of the simulation, and $previous_{ij}$ is the same cell in the previous state of the simulation. e_{adm} stands for the maximum admissible error. The adopted e_{adm} was given by $maxOcupation \times 1e-3$, where $maxOcupation$ is the cell occupancy maximum value (in this case $e_{adm} = 0.1$).

In the estimation phase, the species distribution is simulated, departing from the past to the current environmental conditions by linear interpolation of the EGVs. The initial

[1] This notion of distance between the environmental conditions is inherently dependent on the species' life-cycle. It suffices to think, for instance, that the notion of time has a very different meaning for trees and bacteria.

number of bins for the interpolation, itp_steps, is fixed as the number of iterations necessary to stabilisation after the abrupt change of the previous phase, i.e., $stabilisation\_present$. At each time step, the environmental conditions change gradually, approaching the current environmental conditions. From there, the EGVs are kept fixed, and the simulation continues until stabilisation. The number of iterations until stabilisation after interpolation is denoted sTi. This is an induction process, and in the next simulation, itp_steps is given by the sum of previous itp_steps with the previous sTi. This procedure is repeated until sTi is less or equal to η (in the reported experiments $\eta = stabilisation\_present \times 0.05$).

Having completed this inductive interpolation process, we can finally hypothesise the quantification of each iteration into the corresponding geological time. The geological time g_{time} of each iteration of the agent-based computational model is approximated by the ratio between the geological time range and the total number of iterations until the final stabilisation of the system:

$$g_{time} = \Delta_T / itp\_steps, \quad (5.2)$$

where Δ_T is the change over time, in the current case expressed in years, and itp_steps is the number of interpolation steps necessary to satisfy the defined condition. Note that this hypothesis depends on the assumption that environmental conditions vary uniformly and are also subject to the agent-based model's constraints on the species' life cycle.

Finally, in the prediction phase, the information regarding the geological time is used to define the itp_steps needed to simulate the species distribution departing from the current to the future environmental conditions in a changing climatic scenario. Figure 5.1 and algorithm 7 describe the proposed approach from the initialisation until the prediction phase.

Algorithm 7: Predict species colonisation time via an agent-based distribution model.

1: $stabilisation\_present$=$initialise\_simulation(past\_environmental\_conditions, current\_environmental\_conditions)$
2: set $\eta = stabilisation\_present \times 0.05$
3: set $itp_{steps} = stabilisation\_present$, $sTi = stabilisation\_present$
4: **while** $sTi > \eta$ **do**
5: $\quad interpolate(current, past)$
6: $\quad sTi = reachStabilization(current)$
7: $\quad itp_{steps} = it_{steps} + sTi$
8: $predict\_distribution(g_{time}, current\_environmental\_conditions, future\_environmental\_conditions, stabilisation_{p}resent)$

This approach is evaluated by simulating the distribution of the species under two different environmental scenarios, although, in this case, knowing the quantification of each

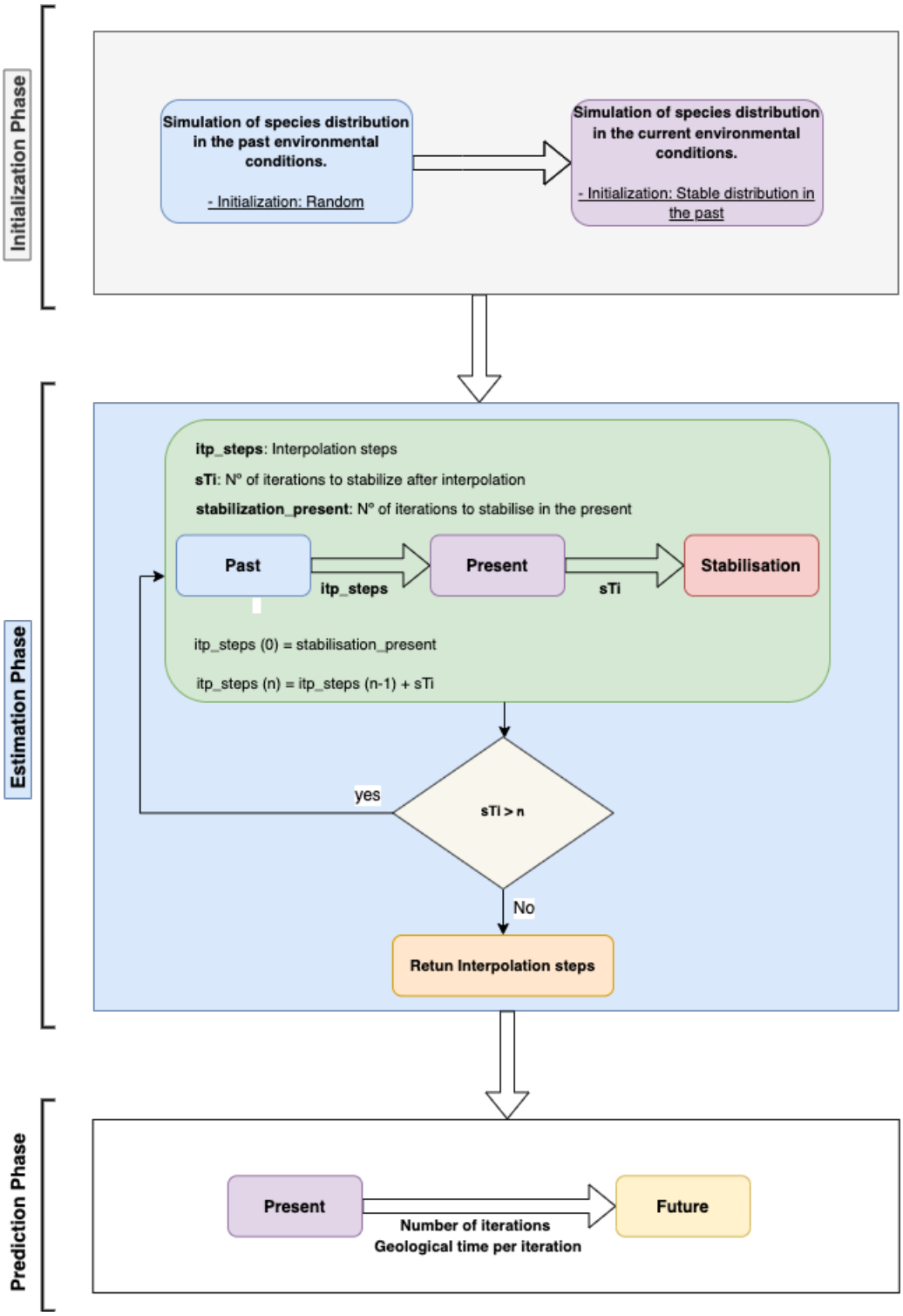


Figure 5.1: Interpolation Method

iteration in geological time. The EGVs under current and future environmental conditions (2070), both from WorldClim (version 1.4), are used.

It is important to note that version 1.4 climate data for current environmental conditions were collected from 1960 to 1990. In this study, two intervals to count geological time are

considered. One from 1960 to 2070 and the other from 1990 to 2070. So the distribution of the species after 80 years and after 110 years.

Apis mellifera case

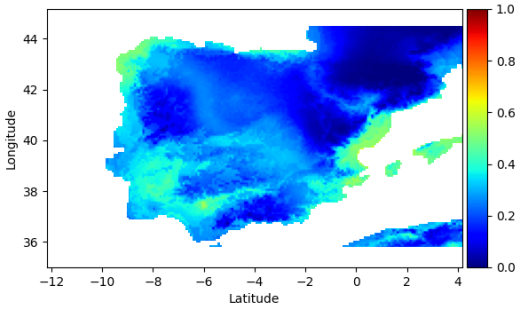
For the African lineage, the number of iterations for the simulation with the past environmental conditions (10000 BP) until the stabilisation was 15 iterations; departing from the stabilisation of the species in the past environmental conditions and making the sudden transition to the current environmental conditions it took 165 iterations until the stabilisation was reached.

Figure 5.2(a,b) shows the suitability map of the African lineage of both past and current environmental conditions. As expected, the values of the suitability map are higher in the map with the current environmental conditions (fig. 5.2(b)), with values in the south region above 0.8 on a scale of (0, 1), whereas the values of the suitability map in the past environmental conditions (fig. 5.2(a)) are bellow 0.5 in almost the entire map. It occurs due to the variation of the differences between the past and current environmental variables. These conditions are less suitable for the survival of the African lineage since the more suitable temperatures for this species vary between 15ºC and 19ºC.

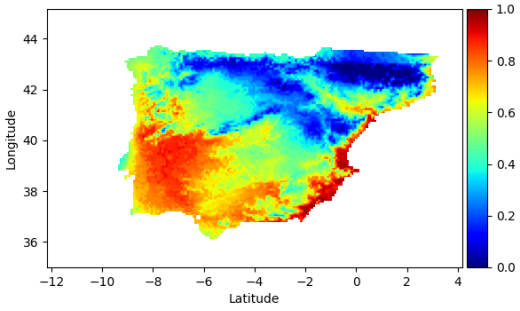
Figure 5.2(c,d) shows the stable distribution map of the African lineage in both 10000 BP and current environmental conditions. With 10000 BP environmental conditions (fig. 5.2(c)), the species cannot spread on the map, whereas with the current environmental conditions (fig. 5.2(d)), the species can spread and occupies a large portion of the map.

Fig. 5.2(e) plots the differences between two sequential maps until the stabilisation under 10000 BP environmental conditions and in current environmental conditions for the African lineage. In 10000 BP environmental conditions, it is possible to verify a higher difference between maps, about one at the beginning of the simulation. Then the values decreased until they reached the stop criterion value. In current environmental conditions, the higher difference between maps was about 0.36 at 1/4 of the simulation, and then the values started to drop.

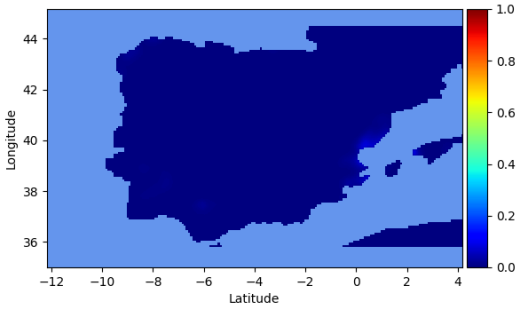
As for the deductive interpolation procedure, for the first interpolation between 10000 BP environmental conditions and the current conditions, $stabilisation\_present$ was used to determine the first number of itp_steps. In this first simulation using the linear interpolation method, 147 more iterations were needed to stabilise the system ($sTi = 147$). After this first simulation, 46 more simulations were run until the stopping criterion was met. From simulation 31, the number of iterations until stabilisation has practically stabilised. The first three steps required until stabilisation sTi were relatively large. Then the values of these differences decreased in the order of 1 and were even continuous in some cases. Fig. 5.2(f) shows the decrease in the values of sTi, indicating that the process is converging. For the African lineage, it took 1283 itp_{steps} for the stopping criterion to be



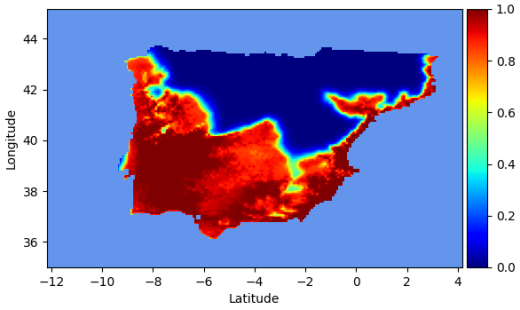
(a) Suitability Map of African lineage (10000 BP).



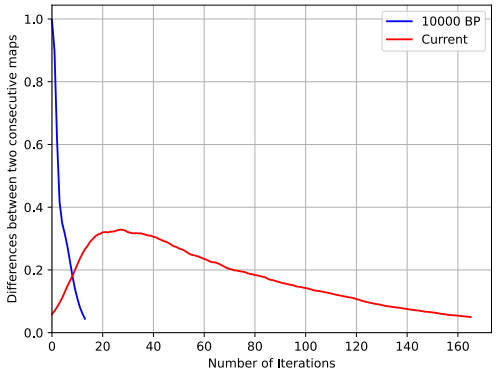
(b) Suitability Map of African lineage (Current).



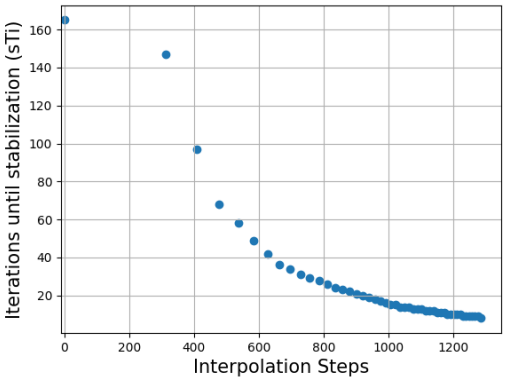
(c) Distribution Map of African lineage (10000 BP)



(d) Distribution Map of African lineage (Current).



(e) Differences between two sequential maps.



(f) Sequence of the number of iterations to reach stabilisation (sTi) during the interpolation phase (African lineage)

Figure 5.2: Suitability and distribution maps of *Apis mellifera* (African lineage), for both 10000 BP and current environmental conditions, interpolation process, and differences between two consecutive maps.

satisfied.

A. unedo case

For the *A. unedo*, it took 17 iterations to stabilise under the environmental conditions of 10000 BP. In comparison, it took 270 iterations ($stabilisation\_present$) to stabilise under the current environmental conditions using the stable distribution of the *A. unedo* in 10000 BP.

Fig. 5.3(a,b) shows the suitability map of *A. unedo* under both 10000 BP and the current environmental conditions. Under the current environmental conditions (Fig. 5.3(b)) has a higher probability of reproducing and spreading (Fig. 5.3(d)), as there are sites (in the west of the peninsula) where suitability values are higher (greater than 0.8). In contrast, at 10000 BP the suitability map values (Fig. 5.3(a)) are lower (less than 0.5), limiting the reproduction and spread of the species (Fig. 5.3(c)).

Fig. 5.3(e) shows the differences between two successive maps until the simulation stabilises. Like the previously analysed species in 10000 BP, where the initialisation of the species was random, the differences between the maps are higher at the beginning of the simulation. Then these differences decrease until the stop criterion is reached. Under the current environmental conditions, where the stable distribution of species in the 10000 BP was used to initialise the species, these differences have lower values.

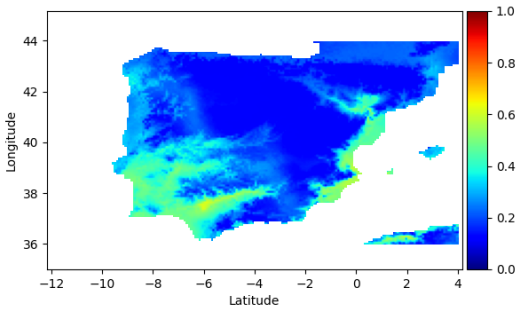
For the linear interpolation between 10000 BP and the current environmental conditions, 96 simulations were required until the stop criterion was met. Fig. 5.3(f) shows the number of iterations needed to achieve stabilisation after each interpolation phase. It took 3030 itp_{steps} until the stop criterion was satisfied.

5.3 Materials and Methods

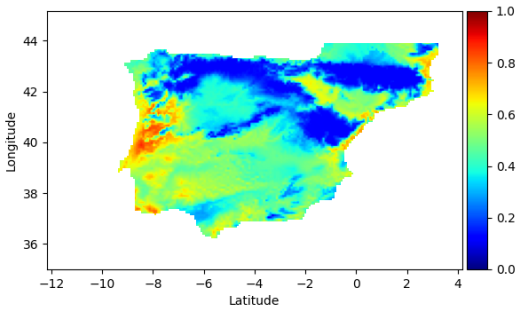
In this section, a validation approach for the interpolation method is presented. For this validation, it is assumed that the result of the mapping between computational and geological time is known. It is an attempt to predict the distribution of the species in the future, considering a changing environmental scenario.

The validation approach consists in:

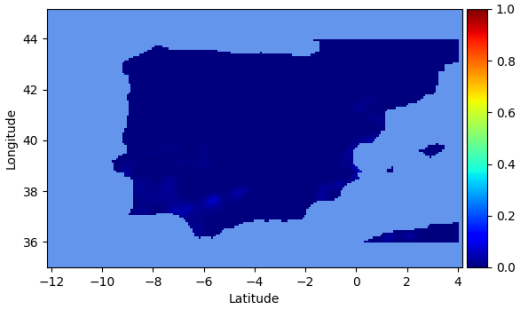
- Simulate the distribution of species departing from a stable distribution with current and future environmental conditions.
- For the interpolation steps (itp_steps), the result of the mapping between computational and geological time is considered; i.e., knowing the unit of geological time corresponding to the unit of computational time (1 $itp_{steps} \Rightarrow$ x $years$). In order to



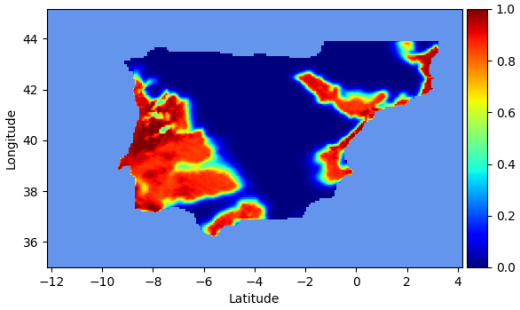
(a) Suitability Map of *A. unedo* (10000 BP).



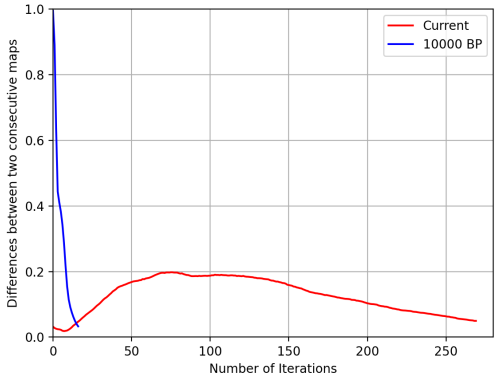
(b) Suitability Map of *A. unedo* (Current).



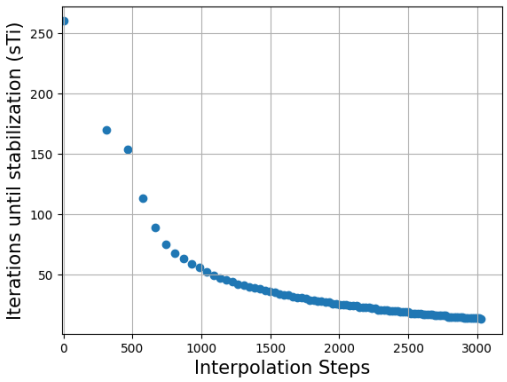
(c) Distribution Map of *A. unedo* (10000 BP).



(d) Distribution Map of *A. unedo* (Current).



(e) Differences between two sequential maps (*A. unedo*)



(f) Sequence of the number of iterations to reach stabilisation (sTi) during the interpolation phase (*A. unedo*)

Figure 5.3: Suitability and distribution maps of *A. unedo*, for both 10000 BP and current environmental conditions, interpolation process, and differences between two consecutive maps.

calculate the itp_{steps} needed to simulate the distribution of species from the current

to the future environmental conditions, the formula below is applied:

$$itp\_steps = (future_{year} - current_{year})/x \quad (5.3)$$

where $future_{year}$ is the years considered for the future environmental conditions, $current_{year}$ is the year that represents the present environmental conditions, and x is the geological time corresponding to 1 itp_steps.

- After reaching the species distribution with the future environmental conditions, the simulation proceeds until it reaches stabilisation.
- The number of iterations from the future environmental conditions until the stabilisation represents the expected time for the species to adapt to the new environmental conditions.

The same species are used in the method validation. Future climate data were obtained from Worldclim version 1.4, 2070 data, rcp26 (gas concentration) [111].

It is important to note that three different results were collected for each species by fixing the birth and death rates and varying the spread rate to (0.2, 0.3 and 0.4). The suitability map of the two species was also projected when using future climate data.

Current environmental conditions were collected from 1960 to 1990; two different simulations are presented: one from 1960 to 2070 and another from 1990 to 2070. To obtain the itp_steps, formula 5.3 is applied.

5.3.1 *Apis mellifera*

According to the results, the simulation times of the three scenarios (spreads: 0.2, 0.3 and 0.4) were close to each other; see Fig. 5.4. The results show that as the spread rate increases, the simulation time decreases.

Figure 5.5 shows the suitability map of *Apis mellifera* according to three environmental scenarios (past, current and future). If compared with the current, future environmental conditions are less suitable for the species survival, and with the past environmental conditions, the species does not survive, see Fig. 5.5(c) and Fig. 5.5(a). However, there are places in Fig. 5.5(c) where the species can survive or spread to neighbouring locations.

Figure 5.6 shows the distribution maps resulting from the simulation in present and future environmental conditions, and the distribution maps resulting from the interpolation between present and future environmental conditions in both scenarios (from 1960 to 2070 - Fig. 5.6(b) and from 1990 to 2070 - Fig. 5.6(e)). All the distribution maps were obtained with a spread rate of 0.4). Considering that for *Apis mellifera*, one iteration corresponds to 7.79 years in the first scenarios (1960-2070) to simulate species distribution from present to future (2070), 14 iterations are needed. In total, 188 iterations are

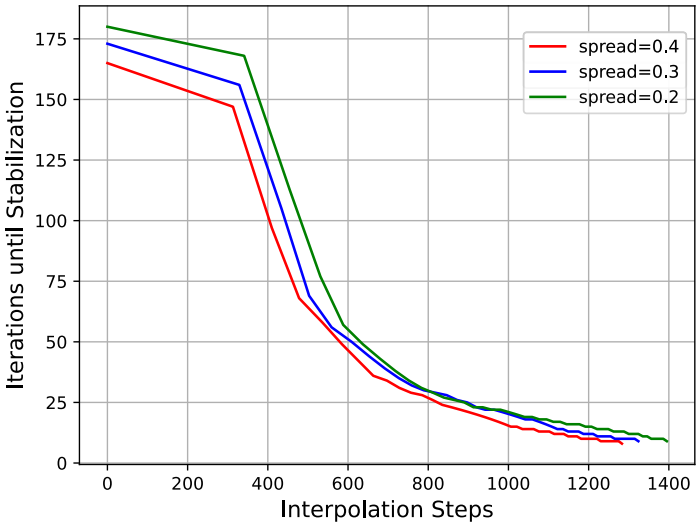
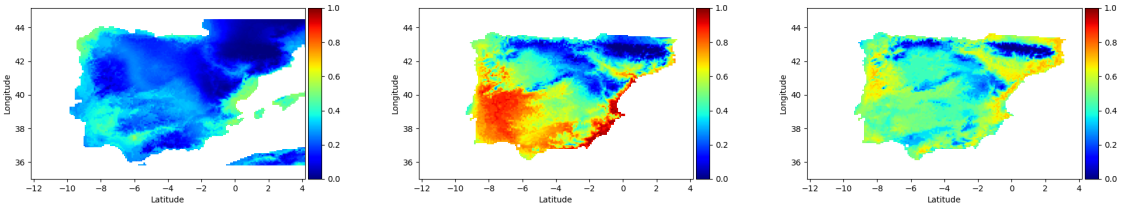


Figure 5.4: *Apis mellifera* (birth rate: 0.6; death rate: 0.2)



(a) Suitability Map (10000 BP).

(b) Suitability Map (Current).

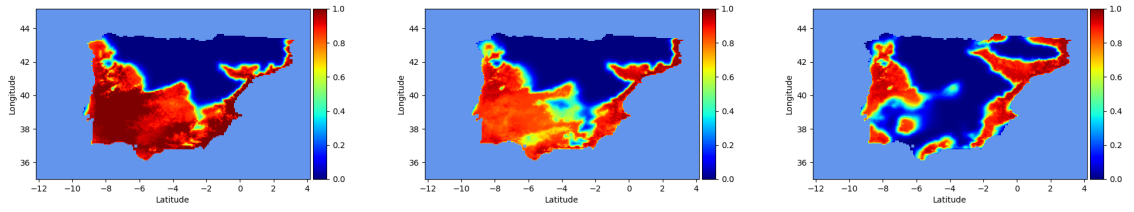
(c) Distribution Map (Future - 2070 rcp26).

Figure 5.5: Suitability maps of *Apis mellifera* - 10000 BP, Current, Future.

needed until they reach stabilisation. For the second scenario (1990-2070), 10 iterations are needed in order to obtain the distribution of the species in 2070 (Fig. 5.6(c)); after that in total 185 iterations are needed to reach the stabilisation (Fig. 5.6(f)). It is possible to notice that for both scenarios, after reaching the environmental conditions of the future (2070), more iterations are needed until the stabilisation.

Figure 5.7 shows the species distribution maps for the present and future environmental conditions and the distribution map resulting from the interpolation between present and future environmental conditions for the same two scenarios but with a spread rate of 0.3. After the simulation reaches future environmental conditions in both scenarios, more iterations are needed to stabilise the distribution.

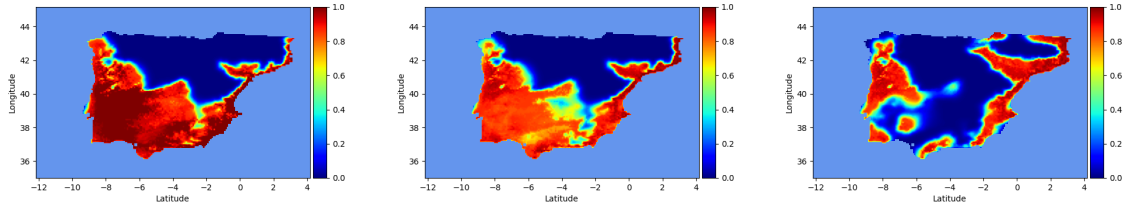
Figure 5.8 shows the distribution maps in the present and future environmental conditions with a spread rate of 0.2 and the distribution map resulting from the interpolation between present and future environmental conditions. Because of the spread rate, there



(a) Distribution Map (Current).

(b) Distribution Map (Future-Interpolation: from 1960 to 2070 rcp26).

(c) Distribution Map (Future Until stabilisation: from 1960 to 2070).



(d) Distribution Map (Current).

(e) Distribution Map (Future-Interpolation: from 1990 to 2070 rcp26).

(f) Distribution Map (Future Until stabilisation: from 1990 to 2070).

Figure 5.6: Distribution maps of *Apis mellifera* - Current, Interpolation and Future. birth rate: 0.6; death rate: 0.2; spread rate: 0.4; $1 itp_{steps} = 7.79 years$; its=1283. years until stabilisation (from 1960 to 2070: 1441; from 1990 to 2070: 1464 years)

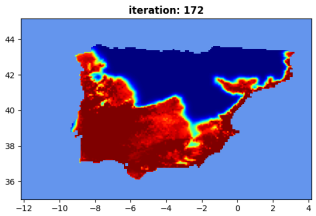
is a high species concentration in suitable locations. The species occupancy percentage is close to one in almost all suitable locations because species reproduce much more than spread. Figure 5.8(b) and Fig. 5.8(e) show that more iterations are needed after reaching the future environmental conditions, to stabilise the distribution of the species.

5.3.2 *A. unedo*

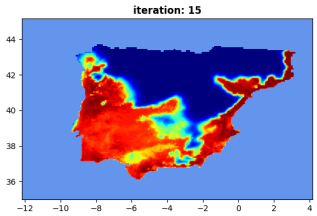
The results of the simulations in the three scenarios (spread rates of 0.2, 0.3 and 0.4) have more considerable differences than the results of *Apis mellifera*. The simulation of *A. unedo* with a spread rate equal to 0.2 lasted much longer than the simulation with a spread rate of 0.4; see Fig. 5.9.

Figure 5.10 shows the suitability maps for *A. unedo* in the past, present and future environmental scenarios. According to Fig. 5.10(c), if compared with the suitability map of the present (Fig. 5.10(b) the suitable locations for the species will change, but the species will not disappear. Species will be more predominant in the northeast region.

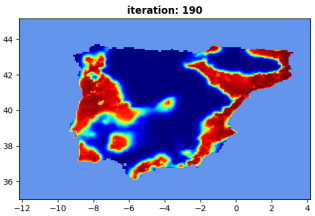
In Fig. 5.11, the distribution map of the species in the present and future environmental conditions is presented, as well as the distribution map resulting from the interpolation between present and future environmental conditions. For the first case (1960-2070, Fig.



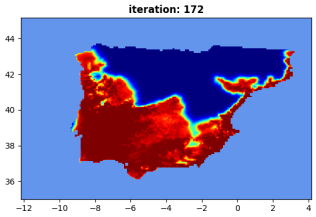
(a) Distribution Map (Current).



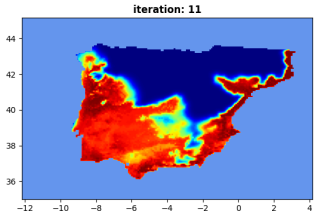
(b) Distribution Map (Future-Interpolation: from 1960 to 2070 rcp26).



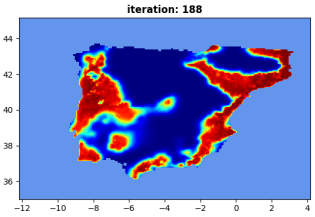
(c) Distribution Map (Future until stabilisation from 1960 to 2070).



(d) Distribution Map (Current).



(e) Distribution Map (Future-Interpolation: from 1990 to 2070 rcp26).



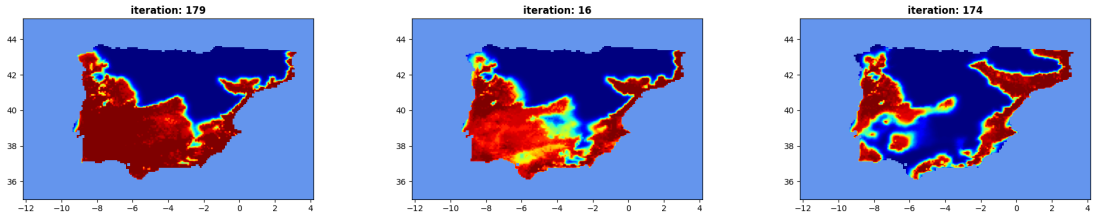
(f) Distribution Map (Future until stabilisation from 1990 to 2070).

Figure 5.7: Distribution maps of *Apis mellifera* - Current, Interpolation and Future. birth rate: 0.6; death rate: 0.2; spread rate: 0.3; $1itp_{steps} = 7.5years$; its=1324. Years until stabilisation (from 1960 to 2070:1425; from 1990 to 2070: 1410 years)

5.11(b)), considering that one iteration corresponds to 3.3 years, to simulate the distribution of the species from the present environmental conditions until the future environmental conditions (from 1960 to 2070), 33 iterations are needed; however, species need 109 iterations until it stabilises. For the second case departing the simulation from the present (1990) to the future (2070), 24 iterations are needed; but more iterations are needed until stabilising the simulation.

Figure 5.12 presents the simulation results of *A. unedo* for a spread rate of 0.3. With that spread rate, one iteration corresponds to 2.85 years. For the first case (Fig. 5.12(b)), doing the simulation, departing from the present environmental conditions (1960) to the future environmental conditions (2070), 39 iterations are needed; In total, 114 iterations are needed in order to stabilise the simulation. In the second case (Fig. 5.12(e)), departing from the present environmental conditions (1990) to the future (2070), 29 iterations are needed; Eighty more iterations are needed in order to stabilise the simulation.

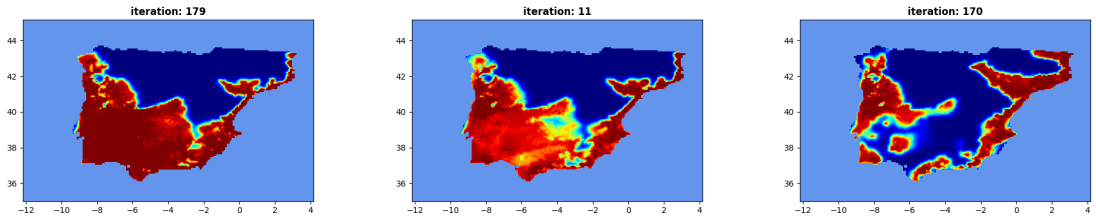
In figure 5.13, the distribution maps in the present and future environmental conditions are presented, as well as the distribution map resulting from the interpolation between present and future environmental conditions. For a spread rate of 0.2, one iteration corresponds to 2.59 years. Therefore, 42 iterations are needed to simulate the distribution of the species from 1960 to 2070 and 31 from 1990 to 2070. In both cases, more iterations



(a) Distribution Map (Current).

(b) Distribution Map (Future-Interpolation: from 1960 to 2070 rcp26).

(c) Distribution Map (Future until stabilisation from 1960 to 2070).



(d) Distribution Map (Current).

(e) Distribution Map (Future-Interpolation: from 1990 to 2070 rcp26).

(f) Distribution Map (Future until stabilisation from 1990 to 2070).

Figure 5.8: Distribution maps of *Apis mellifera* - Current, Interpolation and Future. birth rate: 0.6; death rate: 0.2; spread rate: 0.2; $1itp_{steps} = 7.1years$; its=1395. Years until stabilisation (from 1960 to 2070:1207; from 1990 to 2070: 1235 years)

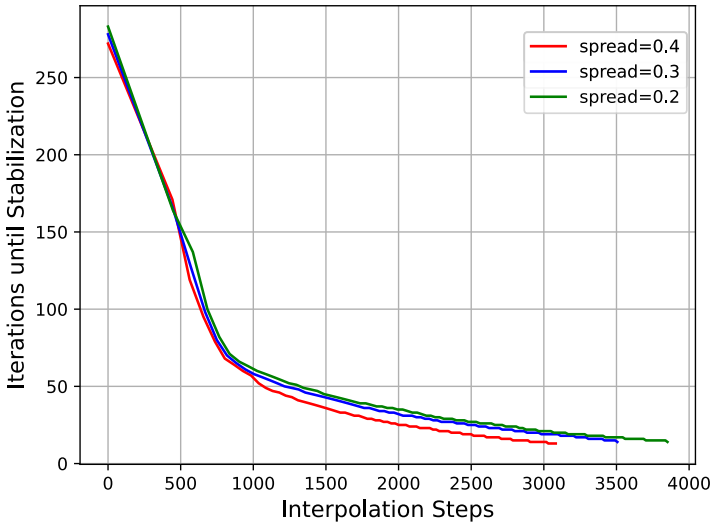
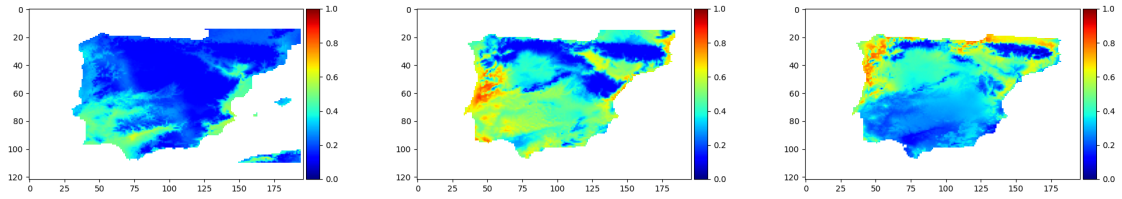


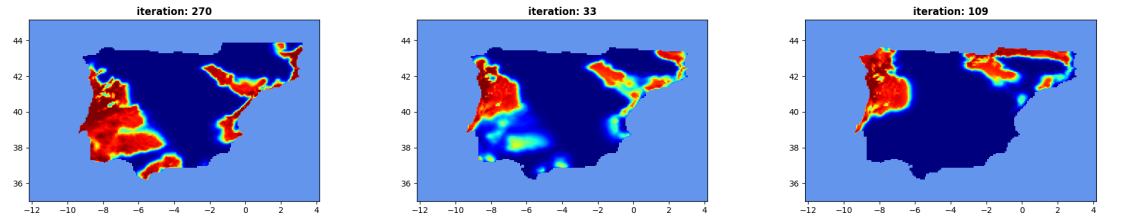
Figure 5.9: *A. unedo* (birth rate: 0.6; death rate: 0.2)

are needed until they reach stabilisation.

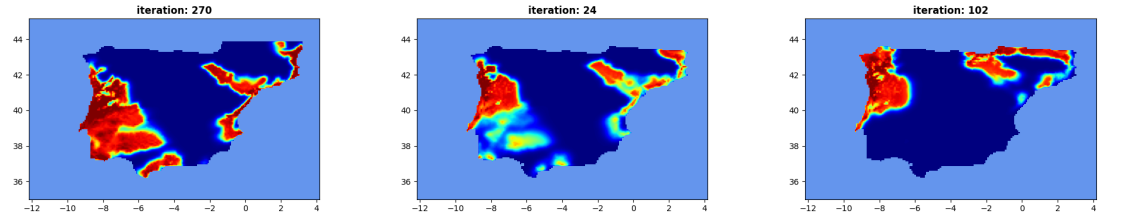


(a) Suitability Map (10000 BP). (b) Suitability Map (Current). (c) Distribution Map (Future - 2070 rcp26).

Figure 5.10: Suitability maps of *A. unedo* - 10000 BP, Current, Future.



(a) Distribution Map (Current). (b) Distribution Map (Future-Interpolation: from 1960 to 2070 rcp26). (c) Distribution Map (Future until stabilisation from 1960 to 2070).

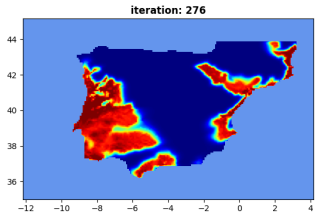


(d) Distribution Map (Current). (e) Distribution Map (Future-Interpolation: from 1990 to 2070 rcp26). (f) Distribution Map (Future until stabilisation from 1990 to 2070).

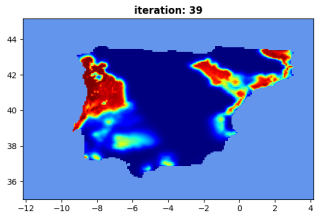
Figure 5.11: Distribution maps of *A. unedo* - 10000 BP, Current, Future. birth rate: 0.6; death rate: 0.2; spread rate: 0.4; $1itp_{steps} = 3.3years$; its=3030. Years until stabilisation (from 1960 to 2070: 363; from 1990 to 2070: 340 years)

5.4 Remarks and Discussion

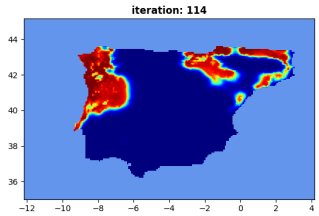
According to the results, the chances of survival of *Apis mellifera* and *A. unedo* at 10000 BP are scarce. It is important to note that in addition to the species' suitability map, the values of the species' life cycle parameters (birth rate, death rate, and spread rate) significantly impact the species' behaviour. With the chosen values for the life cycle parameters in 10000 BP, these species cannot disperse and thus stabilise with few iterations.



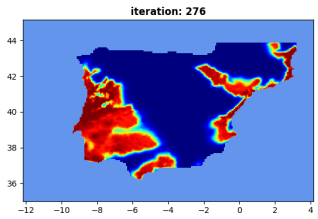
(a) Distribution Map (Current).



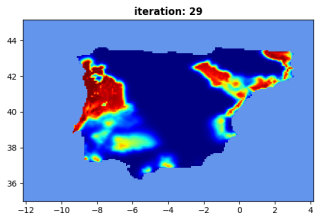
(b) Distribution Map (Future-Interpolation: from 1960 to 2070 rcp26).



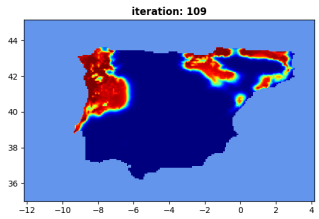
(c) Distribution Map (Future until stabilisation from 1960 to 2070).



(d) Distribution Map (Current).



(e) Distribution Map (Future-Interpolation: from 1990 to 2070 rcp26).



(f) Distribution Map (Future Until stabilisation from 1990 to 2070).

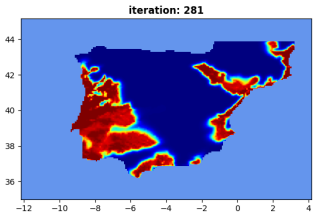
Figure 5.12: Distribution maps of *A. unedo* - 10000 BP, Current, Future. birth rate: 0.6; death rate: 0.2; spread rate: 0.3; $1itp_{steps} = 2.85years$; its=3506. years until stabilisation (from 1960 to 2070: 328; from 1990 to 2070: 314 years)

On the other hand, the current environmental conditions are more suitable for the species' survival. *Apis mellifera* occupies the entire southern region, the east and west coasts of the Iberian Peninsula. *A. unedo* inhabits the east and west coasts, the southeast and part of the northeastern region.

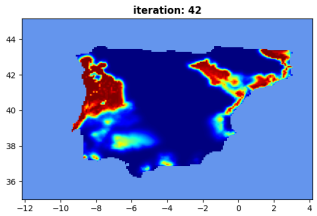
Starting from a random initialisation with many filled cells, the differences between two successive maps are generally significant initially, as several unsuitable cells could be filled with values. In the next phase of the simulation, the values in these unsuitable cells decrease considerably. Departing from a different initialisation than the random one (with many filled cells), the graph of the differences between the maps could show a different pattern; see Fig. 5.2 (e).

The number of iterations to stabilise the simulation under the current environmental conditions ($sHard$) is insufficient to stabilise the simulation with the linear interpolation approach. $sHard$ is significantly less than the interpolation steps (itp_{steps}) required to stabilise the system with the interpolation method. As itp_{steps} increases, (sTi) decreases, even becoming constant for some successive simulations (Fig. 5.2(f)) and then reaches the defined stopping criterion, see algorithm 7.

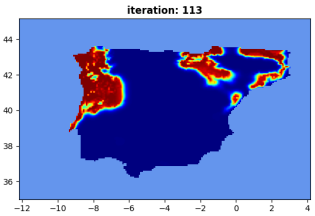
Table 5.1 presents the time per iteration for the two species analyzed.



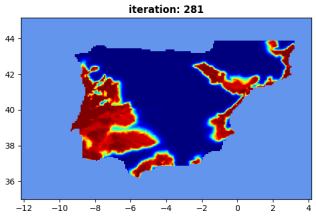
(a) Distribution Map (Current).



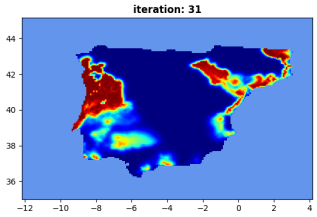
(b) Distribution Map (Future-Interpolation: from 1960 to 2070 rcp26).



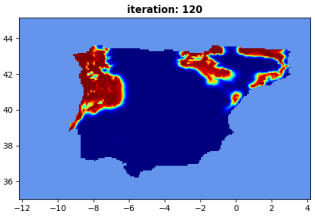
(c) Distribution Map (Future until stabilisation from 1960 to 2070).



(d) Distribution Map (Current).



(e) Distribution Map (Future-Interpolation: from 1990 to 2070 rcp26).



(f) Distribution Map (Future until stabilisation from 1990 to 2070).

Figure 5.13: Distribution maps of *A. unedo* - 10000 BP, Current, Future. birth rate: 0.6; death rate: 0.2; spread rate: 0.2; $1itp_{steps} = 2.59years$; its=3851. Years until stabilisation (from 1960 to 2070: 285; from 1990 to 2070: 293)

Table 5.1: Geological time corresponding to a single iteration of the agent-based model.

Species	Time frame (Δ_T)	Iterations until stabilisation (itp_{steps})	Time per iteration (g_{time})
Apis mellifera	10.000,00 years	1283	7.79 years
A. unedo	10.000,00 years	3030	3.3 years

It is important to note that any results obtained are constrained by the life cycle parameters chosen and the model used to project the suitability maps of these species. Furthermore, this study assumes that environmental conditions change linearly from the 10000 BP to the present. Another critical aspect of this study concerns the variation in the values of the environmental variables between the two-time points (10000 BP and present). If there is no significant variation between the values of the environmental variables in both time points, the interpolation phase is omitted.

Therefore, based on this approach, it is possible to predict species distribution due to climatic changes. For example, suppose one assumes a temperature change of 4°C expected over the next 80 years. Predicting how species will spread over this period is possible if one knows the mapping between computational and geological time. Thus, if one wants to predict the spread of species over a period within these 80 years (e.g. 40 years), knowledge of temporal mapping could be sufficient to simulate the number of iterations corresponding to the geological time.

According to the prediction phase, it is essential to notice that in real scenarios, the average climatic data from one point in time to another, with a significant distance between them, does not change abruptly. Instead, these data change gradually during the time interval between these two points. It is essential to consider a stable initial distribution of the species when simulating the distribution of species from one environmental condition (e.g. present environmental conditions) to another (future). Assuming an initial stable distribution of the species allows us to analyse better the time that the species lasts until it adapts to the new environmental scenarios. Applying the interpolation method by knowing the mapping between the computational and geological time, it is possible to predict the distribution of the species in each period until 2070. The suitability maps of the two species used as a case study, when using the environmental variables of the future, show that both species will decrease their range in 2070 by reducing occupation across the habitat.

The species' initial (stable) distribution and the gradual climatic changes allow the species to survive much longer than expected in unsuitable locations. Thus, the number of iterations corresponding to the future environmental conditions is insufficient for species to stabilise. Therefore, the stabilisation of models could probably take more time.

Chapter 6

An Integrated Tool for Spatio-temporal Prediction

6.1 Introduction

Based on the AB-SDSim model, we propose a generalized user-friendly web ABM system called Species Distribution Simulator (SDSim). SDSim allows any modeller without programming skills to model and simulate the distribution of species and populations in real or potential environmental scenarios. SDSim was designed to study species distribution in a landscape based on a set of parameters entered by the researcher.

In a nutshell, SDSim starts by enabling users to produce an SDM based on the species response functions to each EGV that influence particular species distribution. A Gaussian distribution function was chosen as the response function to each EGV. Therefore, the user provides as input data each EGV as a raster map. For each EGV, a mean and standard deviation defining a normal distribution are hypothesized as optimal for the distribution of that species. Response functions are combined to create an SDM representing the species' environmental suitability. This virtualization of the species distribution proceeds constrained by the environmental suitability but in close agreement with the species' life cycle algorithm, see Algorithm 3. After a defined quantity of virtual species is placed in random or selected prior locations of a loaded landscape, a simulation can start. At each iteration, the virtual species will promote colonization of those locations deemed as suitable. Moreover, using a visual component, SDSim can monitor and analyse the evolution of the species spatio-temporal distribution in a landscape.

6.2 Modelling Software

There are different software packages used to model the distribution of species. To the best of our knowledge, most of them are based on R package [137, 138, 145, 146, 147], but there are also some offering specially tailored graphical user interfaces (GUI), e.g. [148]. Usually, these packages project the past, current and future scenarios of species distribution. The methods applied to predict the distribution of species can be divided into two categories: machine learning and statistical methods. These methods are calibrated by a set of predictor variables and a sample of the known distribution of the species (presence-only or presence-absence data). In order to evaluate the prediction performance, these

software implement some widely used performance measures such as the area under the ROC curve (AUC), True Skill Statistics (TSS), among others. Predicted results are heavily influenced by the data quality (bias, poor data quality). For this reason, several studies adopt virtual species to completely control the relationship between the virtual species and the environment (EGVs) [149].

Several software packages were developed to generate virtual species, e.g. [10, 11, 12, 13]. Generally, these software packages receive as input data the environmental variables related to the species under study and the response function that describes this relationship and produces the environment suitability (suitability map) for the species. This suitability map is converted into a potential presence/absence map of the species in that location [149]. Different techniques are implemented to generate a presence/absence map; see [10]. Usually, a sampling of the presence/absence points is saved from being used in an SDM as presence-only or presence-absence data.

Generally speaking, these software are bound to predict the areas where a species may or may not occur. A biologically guided simulation process is usually absent in which it is possible to observe how the species spreads spatially in the environment over time, following a life cycle. Therefore, obtaining information regarding where a species may or may not occur could be insufficient, particularly for management, species conservation, and resource optimisation. For example, based on the information regarding how species of plants can occupy some locations swiftly, better transplanting strategies could be adopted to optimise the allocated resources.

Our web-based software solution, SDSim, while sharing the common concerns of available software packages for modelling species distribution, provides components that empower the user to visualise and analyse how species spread spatially in the environment at each time interval.

6.3 Simulator Highlights

SDSim is a software tool that allows users to monitor interactively the movement of a species across any real or putative environmental scenarios a researcher introduces. Modellers can easily analyse how an ecological system can spread spatially and behave from the beginning of the simulation. In the SDSim, the environment is defined by a set of EGVs encoded in raster maps made up of a matrix of pixels, also called cells. Simulation outputs depend on a set of parameters that provides a flexible virtual framework to define a colonisation pattern for virtually any species [22]: (1) an initial distribution of patches (grid cells) from which simulation can start, (2) three life cycle parameters (birth rate, death rate and spread rate), (3) any of the multiplicative or additive approaches to model habitat suitability; (4) type of landscape, and (5) a stopping criterion. Table 6.1 summarizes SDSims input data and parameters.

Table 6.1: Description of all input data / parameters of *SDSim*.

Input data / Parameters	Description
Species name	Name of a species.
Initial population	Initial distribution of patches (set of cells) generated randomly across a landscape.
Number of iterations (ticks)	Number of the simulated epochs.
Refresh Frequency	Frequency of generated output (e.g. every ten iterations).
Life cycle parameters	Birth rate or survival percentage.
	Death rate or extinction percentage.
	Spreading capacity rate or colonization percentage.
Type of model	Types of aggregation operators (multiplicative or additive) to model habitat suitability.
Type of neighbourhood	Kind of approach to simulate expansion pattern (Moore or Weighted Moore).
Eco-geographical variables (EGVs)	Set of independent EGVs to fit habitat suitability.
Mean (μ) and standard deviation (σ)	Standard parameters that define a normal distribution for each EGV. Mean values represent optimal conditions for that species.
Stopping criterion	Maximum number of iterations or stability criterion.

6.3.1 Habitat suitability function

The habitat suitability function becomes crucial to provide a realistic simulation scenario. In general terms, species show habitat preferences. For example, the probability of colonisation and spreading should be lower in a location where environmental conditions are not suitable for a certain species. SDSim landscape can be characterised by a set o EGVs, relying on the researchers' knowledge about a species ecology, which one could limit, or otherwise promote a species distribution. From a species' perspective, the overall suitability of the region is determined by the species' probability of occurrence, given the environmental conditions in that location, [112]. It is the result of the aggregation of local EGV values that influence its life cycle. In SDSim, each EGV is characterised by a normal distribution around a hypothesised optimal value for that species. Mean, and standard deviation are provided for each variable according to the species preferences studied in the simulation. Alternatively, these values can be automatically calculated by providing the species occurrence (dataset) observed directly on the terrain.

As a result, habitat suitability would be internally calculated on a map, and values would be normalised in a close interval from 0 (unsuitable) to 1 (optimal). These values are obtained through probability density functions [150] that incorporate EGV in each map location as arguments. SDSim standardizes each EGV map [151]: $x_{i'} = (x_i - \mu)/\sigma$, where x_i is the value of an EGV in that location, μ is the mean (i.e., EGVs' optimal suitability value for a species) and σ represents the standard deviation for that EGV map. SDSim implements

two different model aggregation operators that the researcher can select to compute an overall suitability map for a given species: additive and multiplicative. The additive option is a straightforward implementation of the generalised additive model (GAM). Therefore, a habitat suitability map is obtained by adding EGV values in each raster cell after applying a probability density function. Under the multiplicative option, the habitat suitability map is obtained by a strict Archimedean triangular norm. In this case, habitat suitability is more restrictive because the product t-norm produces a stronger conjunction of probabilistic values. The multiplicative model might become useful in particular circumstances (e.g. invasive species, colonisation of clonal organisms).

6.3.2 General workflow

In order to perform a simulation, users should set all required simulation input data and parameters shown in Table 6.1. SDSim allows users to create real and simulated scenarios by uploading different EGVs as raster maps and species occurrence data (or providing the mean and standard deviation for each EGV).

Users can select a type of neighbourhood from any of Moore [113] or weighted Moore [152] options. Assuming a regular grid of raster maps, by using Moore's neighbourhood, species will spread homogeneously to their eight neighbours. When using weighted Moore, each of the eight neighbouring cells will receive some transferred percentage of the species' occupation directly proportional to its suitability value. SDSim generates a suitability map based on EGV maps and every specific probability density function. The more suitable an adjacent cell for a species, the more likely it will be chosen for its range spreading. Each cell would contain two values representing the percentage of species' occupation and habitat suitability value. Initially, population patches (grid cells) are randomly placed in a landscape before starting the simulation. During a simulation, SDSim calculates the difference between previous and current states at each iteration by the sum of the cell-by-cell differences between successive states of the system, see Algorithm 2. The simulation can be stopped when the system reaches a stable state or when a maximum number of iterations is completed. In this regard, the system is said to be stable when the difference between two consecutive states of the simulation converges to zero.

SDSim saves the output of a simulation according to an interval previously defined by the user at each iteration or after a given number of epochs. At the end of a simulation, SDSim produces as output the species distribution map in three file formats: (i) text file, (ii) ASCII Grid file, (iii) image file; and (iv) video file, see Fig. 6.1. For future reference, the text file containing the values of all parameters used in the simulation is also available. Table 6.2 summarizes the main functions implemented in *SDSim* ABM.

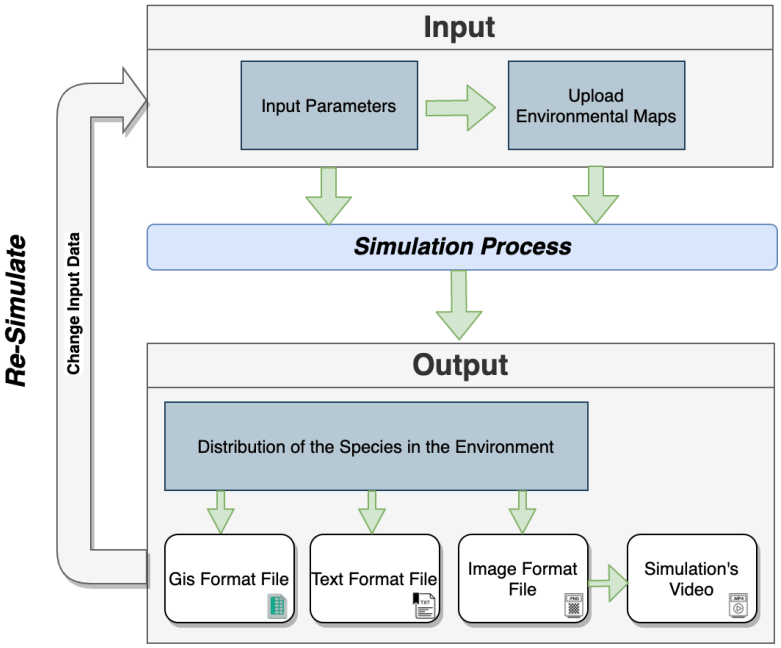


Figure 6.1: *SDSim* General workflow, showing a standard procedure that should be followed by a user to perform a simulation in *SDSim*.

Table 6.2: Description of the main functions implemented in the *SDSim*.

Function	Description
findNeighbours	Identifies neighboring cells.
lifeCycle	Quantifies species capacity in each cell according to a birth rate, death rate and spread rate.
readEcoGeographicalVariables	Loads EGVs.
convertEGVs	Standardization of EGVs.
generateSuitabilityMap	Fits a normal distribution as a response function for each EGV and then combines such response functions by adding or multiplying.
normalizeValues	Normalizes habitat suitability map into a [0,1] interval.
generateInitialPopulation	Generates a defined capacity for a species distributed in random patches.
createDistributionFile	Outputs species distribution every user-defined interval.

6.3.3 Directional constraints

Directional EGV (dEVG) can be used to promote or otherwise limit the movement of a species in certain spatial directions. Typically a dEGV is characterized by two components representing a vector field - magnitude (intensity) and direction. A raster map should represent each component. For intensity preferences of the species, the treatment is similar to the one previously described, i.e. given the intensity optimal parameters (mean and standard deviation) for a given species, the corresponding density value is an input to the model aggregation, which determines the overall suitability map for the species. The direction component will be likewise integrated into the model once suitability is computed. It should include values expressed in degrees clockwise from the geographic North. Thus

each grid cell has a direction value d, $0 \leq d < 360$, representing the flow or directional movement. Species can move from the core cell to any of its eight neighbouring cells. Each neighbour has a relative direction towards the core cell, as depicted in Fig. 6.2.

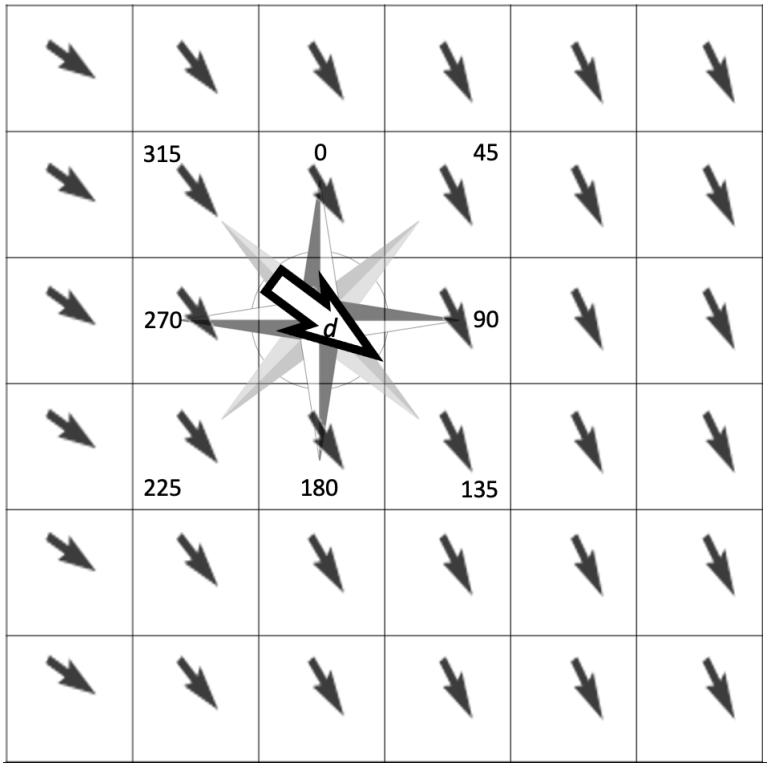


Figure 6.2: Example of a dEGV where arrows represent the direction in degrees clockwise from the geographic North. The highlighted cell has a direction d, see (6.1), and 8 neighbours. Each neighbour has a relative direction towards the cell ranging from 0^o to 315^o.

In order to compute the direction of each neighbour, *SDSim* computes the difference between the relative direction of the neighbour and the direction of movement for that grid cell:

$$\Delta_k = min\left(|d - d_k|, |d + 360 - d_k|\right), \quad (6.1)$$

where $k \in \{1, ..., 8\}$ stands for the neighbour cell, and d_k is the relative direction of such neighbour. In Figure 6.2, those directions are represented as values in each of the neighbours of their core cell. Hence $\forall k \in \{1, ..., 8\} : 0 \leq \Delta_k \leq 180$. Values are then normalized to the unit interval and incorporated into the model aggregation. Species expansion will be favoured by the overall direction information codified in every dEGV, i.e., most of the transferred individuals will migrate from a core cell to neighbouring ones that are well aligned with the direction of movement.

Let us return to the intensity component of an EGV. Besides being treated as a mere EGV representing the preference of a species for a given value (e.g. some species might pre-

fer waters with less current intensity), it is worthwhile to notice that, most times, there is a physical facet attached to it, which should also have implications to the species fixation in a locus or to the speed of range spreading (e.g. some species travel easily if the current speed is higher). In such cases, it is proposed the introduction of a *momentum* (species-dependent) constant, $m_S \in [0, 1]$, constraining the number of individuals that are transferred to neighbouring cells, N_{spread}, in the following way:

$$N_{spread} = N \times (spR + I \times m_S), \quad (6.2)$$

where N is the cell occupation percentage, spR denotes the species spread ratio as introduced before, and $I \in [0, 1]$ is the normalized intensity value observed at the cell level.

Thus the intensity of a directional EGV constrains the spread of the species not only in what regards "preferences" of the species but also by applying the modelled traction forces. Notice that some species heavily depend on such variables to perform their spreading while others can expand their range even if they are positioned in a zero-intensity region, hence the need to introduce a momentum constant.

6.3.4 Web interface

SDSim provides a user-friendly Web interface that allows users to perform their simulations, avoiding local installation. It is available online at `https://sdsim.it.ubi.pt`. To get access to the application, users request an account by filling out a form that will be sent to the SDSim administrator. Based on the information sent, the administrator will create the account. After an account is created, the user receives a notification with the credentials to log in and access available services of the SDSim application presented on the main screen, see Fig. 6.3.

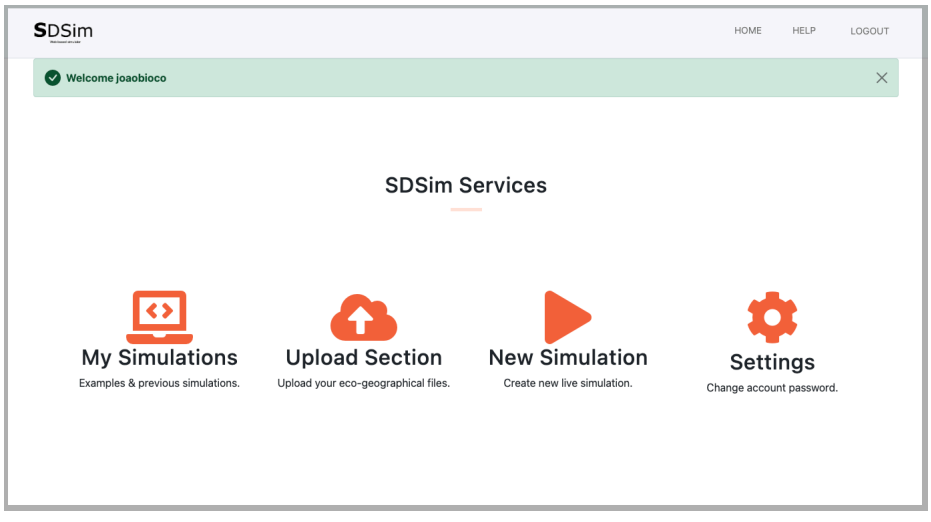


Figure 6.3: *SDSim* available services through main screen.

In the section *"My Simulations"*, users can access details from previous simulations, including visualising the results, simulation parameters and output data (simulation results). Users could download the simulation results, including a set of raster maps containing the state of the simulation at different time steps, figures of each step of the simulation, the video showing the spatial distribution evolution, and performance metric graphs (receiver operating characteristic curve - ROC curve) if a species occurrences file were provided. The user can also perform new simulations based on previous ones, or remove them, see Fig. 6.4. In this section, users can find an initial set of simulation examples to explore and get acquainted with the system.

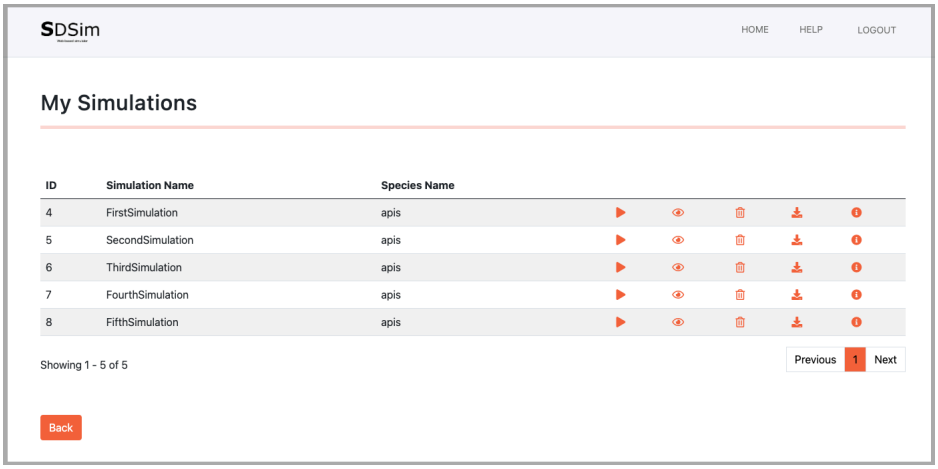


Figure 6.4: Options included in the section "My simulations".

In the "Upload Section", users can upload all the necessary EGVs and dEGVs in the form of raster maps in order to describe the landscape that is intended to simulate. SDSim does not use any specific datum or projection, relying on user needs to make such decisions. Currently, the Web application accepts only ASCII Grid raster format (please see GDAL library for raster and vector geospatial data formats at `https://gdal.org/`).

In addition, the user can also upload a comma-separated values (.csv) file containing the coordinates of the occurrences (and/or absences) of any sampled species. This file facilitates the user's work so that he/she does not need to provide the mean and standard deviation for each EGV to estimate each probability density function. SDSim performs all the necessary calculations to provide the mean and standard deviation for each EGV depending on the occurrences data that the user has uploaded.

Presence/absence data can be uploaded to SDSim, allowing a numerical comparison between the results of different simulations based on the correct classification of presence/absence locations showing a ROC curve and the AUC at the end of the simulation.

In the section "Simulation", users can start a new simulation by filling a form with all the required parameters, as described in table 6.1.

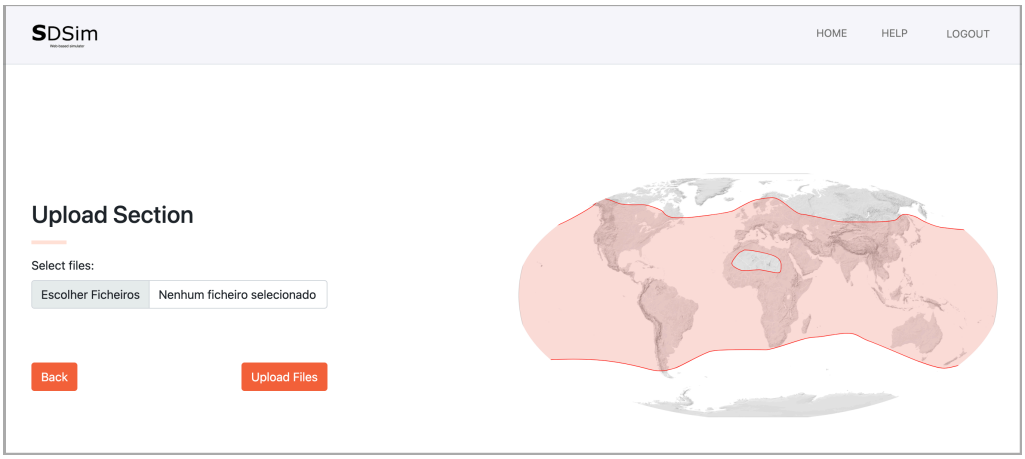


Figure 6.5: *SDSim* "Upload" section.

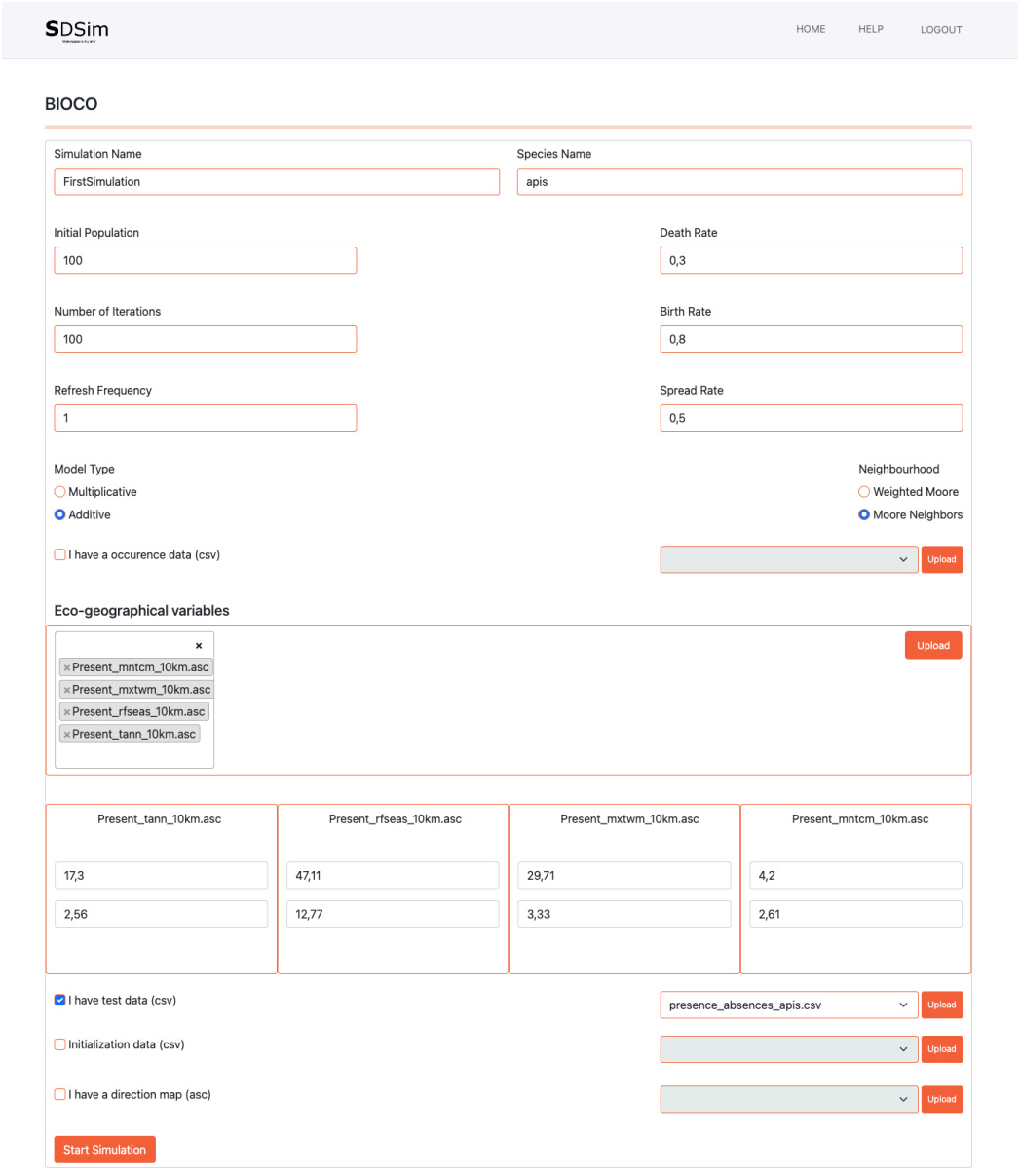


Figure 6.6: Example form to introduce *SDSim* parameters.

Users should follow all these steps after authentication in order to perform the simulation:

1. Access the section *"Upload"* and add EGVs, see Fig 6.5.

2. Return to the main screen and access the *Simulation* section where users should complete all the required parameters, including the selection of the corresponding EGVs and their parameters to estimate each probability density function (see Figure 6.6).

3. Results (species distribution maps, suitability map and ROC curves) can be managed in a gallery of images after the completion of the simulation (see Figure 6.7). Users can also see the video that shows how the species' temporal distribution evolved (see Figure 6.8).

4. A simulation can be saved to reformulate any experiment and refine new ones based on previously stored simulations.

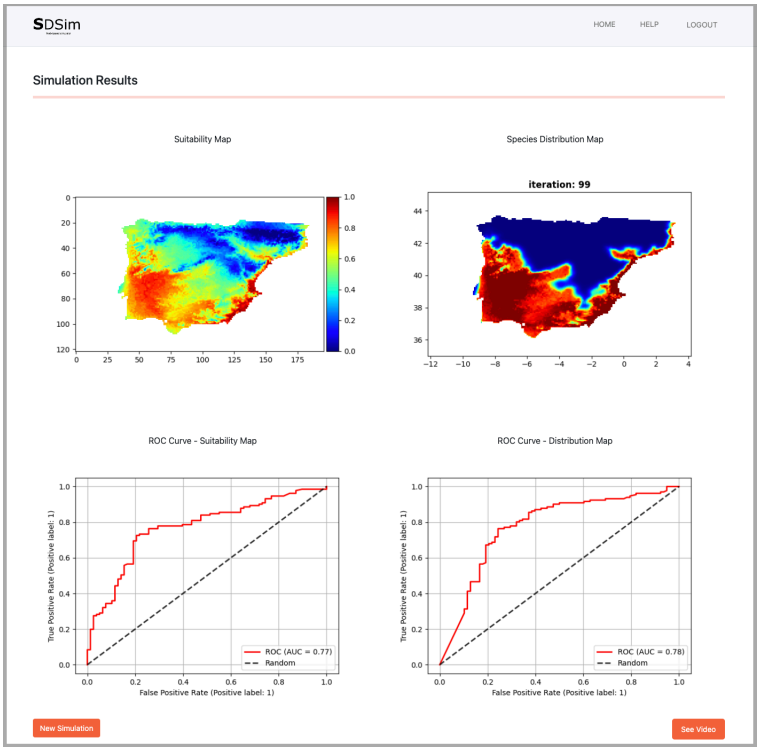


Figure 6.7: Simulation results are shown in the form of images that later can be exported as raster maps.

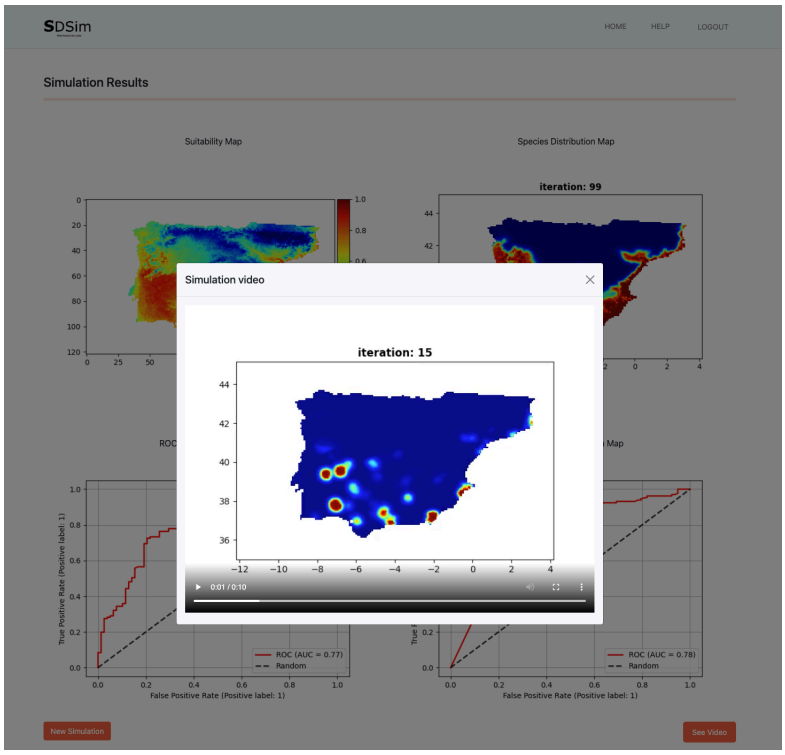


Figure 6.8: Videos enable the visualization of the evolution of a simulation.

6.4 Remarks and Discussion

In this chapter, an ABM system was designed to model and simulate the spatial distribution of biological species by using a set of EGVs. SDSim computes a habitat suitability index, producing an internal species' suitability map. SDSim assumes a standard normal distribution as a probabilistic response function for each EGV. Those are later aggregated using an additive or a multiplicative model, chosen at the users' will. Together with EGVs, SDSim receives a set of input parameters which fit ABM behaviour and can be easily tuned. Therefore, any species could be modelled by using this ABM in different environmental scenarios. A simulation will always start by generating an initial occupation of the available habitat in random or chosen locations. At each iteration of the simulation, species will find more suitable locations to survive, following a defined species life cycle with corresponding parameters. SDSim allows users to study and understand the species-environment relationship by changing the environmental variables' values to verify any simulation scenario's behaviour. If in possession of real experimental data collected on the field, the Web interface provides supporting tools for assessing and comparing a set of distinct hypotheses subsumed by the different experimental setups. Moreover, the flexibility and user-friendliness of the deployed Web ABM system enable the visual analysis of the evolution of species in any hypothetical landscape without requiring any programming skills from the user.

Chapter 7

Conclusion and Further Work

7.1 Main Conclusions

This dissertation presents work and progress relevant to a new insight on the *in silico* characterization of the species-environment relationship. Next, a summary of this dissertation's main results and deliverables is offered.

7.1.1 Agent-based Species Distribution Model

As a rule, the suitability map obtained through SD modelling does not consider the species' behaviour in the environment, making it challenging to analyse how the individuals act and interact as a whole with their ecological environment. On the other hand, the behaviour of autonomous agents reflects better the reality if the environmental constraints for each location are well known. Therefore the following working hypothesis was placed:

- Can the combination of traditional SDM with ABM provide a better understanding of the collective behaviour of a species?

In this regard, the developed algorithms, adding to the traditional SDM the agent-based simulation approach, allow ecologists to analyse how the species spread and colonise the locations considered suitable for the species' survival. Additional information regarding the representation of dynamic variables, such as wind directions and oceanographic currents, and information concerning the species' life cycle approximate the simulation of species distribution to reality.

7.1.2 The Effects Occurrence Data on the Quality of the Model

The prediction of species distribution highly depends on the available species occurrence data. When biased, species occurrence data can be unrealistic in predicting species distribution. It is necessary to consider the following questions:

- How to minimize biased results due to the poor quality or absence of occurrence data?

It is more common to have occurrence data in the form of presence-only data than presence-absence data because absence data should be collected carefully. Usually, it is difficult to know if the species is missing because the location is not suitable or they are just absent

when we collect the data. Because of that, presence-only data are considered more reliable. The problem is that absence data is also needed. Pseudo-absence data take place when there is no available absence data. It is necessary to choose the right approach to select pseudo-absence data to avoid selecting locations with similar environmental conditions to the ones where the species was found. One approach might be to select randomly locations far from locations where the species was observed.

In addition to the available species occurrence data, the expertise of professionals (ecologists, biologists) regarding the characteristics of the environment where a species should be found can be used as an alternative to species occurrence data.

7.1.3 Computational Cost of ABM

Usually, the agent-based species distribution simulation is timing-consuming. Simulations can last seconds or even days, depending on the spatial dimension of the environment (study area) and available hardware. Simulating species distribution in a high-scale environment at a reasonable execution time is challenging and poses the following question:

- How to minimise the species distribution simulation time while ensuring that no information is lost?

Parallelising the tasks by dividing the environment (study area) into stage subsets, and reducing the synchronisation frequency might increase the speedup, solving one part of the issue. However, it is important to notice that there is a lot of data transferred from one location to another during the species life cycle that affects the data integrity. To avoid information loss each stage subset will contain an overlapping section from each neighbouring stage subset.

7.1.4 Representing Time

Time representation is a significant issue in modelling species distribution. At the computation level, it is not difficult to measure the simulation time (the number of epochs/iterations), regardless of the chosen simulation stopping criteria. On the other hand, there is not a consolidated approach that easily adequates the computational time to the geological time. In addition, the environment used to predict species distribution is static, for instance, no climate change is presented during the simulation of species distribution. Thus the following research question emerges:

- How to represent geological time in the agent-based species distribution model, allowing the simulation of the distribution of species in a changing environmental scenario?

The time representation is obtained by implementing an algorithm (iterative interpolation algorithm) that allows the mapping between computational and geological time. A

geological time approximation is obtained, and from there, it is possible to simulate the distribution of species in changing environmental scenarios more accurately predicting what could occur in a real scenario.

7.1.5 Simulation of Species Distribution Made Easy

Existing agent-based species distribution software approaches can be challenging for users without programming skills. These software require the user to implement its model using a programming language. On the other hand, traditional SDM approaches are limited concerning showing the species' dynamics in spreading and colonising suitable locations, an essential aspect when the analysis of the species' behaviour for management and species conservation purposes is needed. Considering the following questions:

- How do we ensure users an easy way to model and simulate species distribution? Moreover, how do we improve the capability of analysing the species distribution for management and preservation purposes?

The web-based species distribution simulator (SDSim) allows users without programming skills to model and simulate species distribution. Users only need to upload the EGVs, the species occurrence data (if it exists), and set the simulation's parameters. On the other hand, in addition to projecting the species-environment relationship, SDSim simulates species distribution, observing species occupancy dynamics in the predicted environment. This species occupancy dynamics can give essential information regarding species' behaviour, such as how fast the species adapt to the environment, considering that they adapt when stabilising, which locations in the study area colonise first and which places the species could reach (helpful information for management purposes).

7.2 Future Research

Some aspects reported in this dissertation can be used as a starting point for further research. Some of these opportunities are presented as follows:

- The presented results obtained by the agent-based species distribution model regarding this combination between traditional SDM and ABM are promising in allowing a better understanding of the species' behaviour. However, regarding this aspect, we would like to have the opportunity to answer the following question:
  - How to couple different species distribution models with our approach to ensure even more realistic results?
- Using the expertise of ecologists and biologists to give information concerning the species' behaviours is a solid alternative, presenting promising results. Concerning this aspect, we would like to have the opportunity to answer the question below:

- How do we implement more reliable methods for validating and evaluating the model?

- Implementing a strategy capable of simulating species distribution in parallel and reducing the need for synchronisation insured gains in the computation cost. Despite the presented parallelization approach, a question remains to be answered:
  - How do we implement more efficient parallelization strategies to improve the time-consuming species distribution simulations?

- Simulating the distribution of the species in changing climatic scenarios by using linear interpolation produced relevant information, with one piece of it being the amount of time the species lasted to spread and colonise the region in the study. Regarding this aspect, we would like to have the opportunity to answer the following question:
  - Can new approaches of climatic change scenarios rather than linear interpolation be more efficient in studying species distribution?

- Implementing ABMs applications capable of receiving and manipulating spatial data according to the granularity the study wants to achieve. Concerning this aspect we would like to have the opportunity to answer the following question:
  - How do we implement different spatial and temporal perspectives in order to analyse behaviours not yet observed?

Bibliography

[1] J. Elith and J. R. Leathwick, "Species distribution models: ecological explanation and prediction across space and time," *Annual Review of Ecology, Evolution and Systematics*, vol. 40, no. 1, pp. 677–697, 2009. 1

[2] J. Bioco, P. Prata, F. Canovas, and P. Fazendeiro, "On the modelling of species distribution: Logistic regression versus density probability function," in *Intelligent Computing*, K. Arai, Ed. Cham: Springer International Publishing, 2022, pp. 378–391. 1, 4, 5

[3] R. M. Chefaoui and J. M. Lobo, "Assessing the effects of pseudo-absences on predictive distribution model performance," *Ecological modelling*, vol. 210, no. 4, pp. 478–486, 2008. 1

[4] A. Jiménez-Valverde, "Insights into the area under the receiver operating characteristic curve (auc) as a discrimination measure in species distribution modelling," *Global Ecology and Biogeography*, vol. 21, no. 4, pp. 498–507, 2012. 1, 12

[5] D. J. Hand and R. J. Till, "A simple generalisation of the area under the roc curve for multiple class classification problems," *Machine learning*, vol. 45, no. 2, pp. 171–186, 2001. 1, 13

[6] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960. 1, 13

[7] O. Allouche, A. Tsoar, and R. Kadmon, "Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (tss)," *Journal of Applied Ecology*, vol. 43, no. 6, pp. 1223–1232, 2006. [Online]. Available: http://dx.doi.org/10.1111/j.1365-2664.2006.01214.x 1, 13

[8] M. Barbet-Massin, F. Jiguet, C. H. Albert, and W. Thuiller, "Selecting pseudo-absences for species distribution models: how, where and how many?" *Methods in Ecology and Evolution*, vol. 3, no. 2, pp. 327–338, 2012. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.2041-210X.2011.00172.x 1, 39

[9] C. Liu, G. Newell, and M. White, "The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo-absences or background sites," *Ecography*, vol. 42, no. 3, pp. 535–548, 2019. 1, 39

[10] B. Leroy, C. N. Meynard, C. Bellard, and F. Courchamp, "virtualspecies, an r package to generate virtual species distributions," *Ecography*, vol. 39, no. 6, pp. 599–607, 2016. 1, 80

[11] R.-Y. Duan, X.-Q. Kong, M.-Y. Huang, G.-L. Wu, and Z.-G. Wang, "Sdmvspecies: a software for creating virtual species for species distribution modelling," *Ecography*, vol. 38, no. 1, pp. 108–110, 2015. 1, 80

[12] H. Qiao, A. T. Peterson, L. P. Campbell, J. Soberón, L. Ji, and L. E. Escobar, "Nichea: creating virtual species and ecological niches in multivariate environmental scenarios," *Ecography*, vol. 39, no. 8, pp. 805–813, 2016. 1, 80

[13] S. J. Phillips, R. P. Anderson, and R. E. Schapire, "Maximum entropy modeling of species geographic distributions," *Ecological Modelling*, vol. 190, no. 3, pp. 231–259, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030438000500267X 1, 39, 61, 80

[14] S. J. Taylor, A. Anagnostou, T. Kiss, G. Terstyanszky, P. Kacsuk, and N. Fantini, "A tutorial on cloud computing for agent-based modeling and simulation with repast," in *Proceedings of the 2014 Winter Simulation Conference*. IEEE Press, 2014, pp. 192–206. 2, 7

[15] C. M. Macal, "Everything you need to know about agent-based modelling and simulation," *Journal of Simulation*, vol. 10, no. 2, pp. 144–156, 2016. 2

[16] T. M. Anderson and S. Dragićević, "Network-agent based model for simulating the dynamic spatial network structure of complex ecological systems," *Ecological Modelling*, vol. 389, pp. 19–32, 2018. 2, 16, 17

[17] K. M. Pepin, A. J. Davis, and K. C. VerCauteren, "Efficiency of different spatial and temporal strategies for reducing vertebrate pest populations," *Ecological Modelling*, vol. 365, pp. 106 – 118, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0304380017304349 2, 14, 17

[18] L. Xing, C. Zhang, Y. Chen, Y.-J. Shin, P. Verley, H. Yu, and Y. Ren, "An individual-based model for simulating the ecosystem dynamics of jiaozhou bay, china," *Ecological Modelling*, vol. 360, pp. 120–131, 2017. 2, 15, 17

[19] S. Heinänen, M. E. Chudzinska, J. B. Mortensen, T. Z. E. Teo, K. R. Utne, L. D. Sivle, and F. Thomsen, "Integrated modelling of atlantic mackerel distribution patterns and movements: A template for dynamic impact assessments," *Ecological Modelling*, vol. 387, pp. 118 – 133, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0304380018302795 2, 15, 17, 19

[20] J. Bioco, P. Prata, F. Canovas, and P. Fazendeiro, "Agent-based modelling applied to the analysis of spatiotemporal distribution of species (under review)," 2023. 2, 5

[21] S. Skowronek, G. P. Asner, and H. Feilhauer, "Performance of one-class classifiers for invasive species mapping using airborne imaging spectroscopy," *Ecological Informatics*, vol. 37, pp. 66–76, 2017. 2

[22] J. Bioco, P. Fazendeiro, F. Cánovas, and P. Prata, "Parameterization of an agent-based model of spatial distribution of species," in *Emerging Technologies in Computing*, M. H. Miraz, P. S. Excell, A. Ware, S. Soomro, and M. Ali, Eds. Cham: Springer International Publishing, 2020, pp. 251–260. 4, 5, 80

[23] J. Bioco, P. Prata, F. Cánovas, and P. Fazendeiro, "Remarks on the behavior of an agent-based model of spatial distribution of species," *Annals of Emerging Technologies in Computing (AETiC)*, vol. 5, no. 2, 2021. 4, 5

[24] J. Bioco, F. Cánovas, P. Prata, and P. Fazendeiro, "Sdsim: A generalized user friendly web abm system to simulate spatiotemporal distribution of species under environmental scenarios," *Environmental Modelling & Software*, vol. 147, p. 105234, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364815221002760 4, 5, 13

[25] J. Bioco, P. Prata, F. Cánovas, and P. Fazendeiro, "Synchronization overlap trade-off for a model of spatial distribution of species," in *International Conference on Computational Science and Its Applications*. Springer, 2021, pp. 296–310. 5

[26] J. Bioco, P. Prata, F. Canovas, and P. Fazendeiro, "Prediction of the arbutus unedo colonization time via an agent-based distribution model," 2022. 5

[27] J. Bioco, J. Silva, F. Canovas, and P. Fazendeiro, "A cellular automata model of spatio-temporal distribution of species," in *International Conference on Advanced Intelligent Systems for Sustainable Development*. Springer, 2018, pp. 118–128. 5

[28] V. Grimm, U. Berger, D. L. DeAngelis, J. G. Polhill, J. Giske, and S. F. Railsback, "The odd protocol: a review and first update," *Ecological modelling*, vol. 221, no. 23, pp. 2760–2768, 2010. 5, 21

[29] C. Gou, "Predictability of shanghai stock market by agent-based mix-game model," in *2005 International Conference on Neural Networks and Brain*, vol. 3, Oct 2005, pp. 1651–1655. 7

[30] V. Krishnamurthy and S. Bhatt, "Sequential detection of market shocks with risk-averse cvar social sensors," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 6, pp. 1061–1072, Sept 2016. 7

[31] H. Takahashi, "Analyzing the role of noise trader in financial markets through agent-based modelling," in *2014 IEEE 38th International Computer Software and Applications Conference Workshops*, July 2014, pp. 444–449. 7

[32] S. Chen, K. Tai, and Z. Li, "Evaluation of supply chain resilience enhancement with multi-tier supplier selection policy using agent-based modeling," in *2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, Dec 2016, pp. 124–128. 7

[33] N. I. Arvitrida, S. Robinson, and A. A. Tako, "How do competition and collaboration affect supply chain performance? an agent based modeling approach," in *2015 Winter Simulation Conference (WSC)*, Dec 2015, pp. 218–229. 7

[34] D. Fang and W. Puqing, "Simulating the structural evolution in agri-food supply chain: An agent-based model," in *2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics*, vol. 1, Aug 2015, pp. 214–219. 7

[35] K. Dehghanpour, H. Nehrir, J. Sheppard, and N. Kelly, "Agent-based modeling of retail electrical energy markets with demand response," *IEEE Transactions on Smart Grid*, vol. PP, no. 99, pp. 1–1, 2016. 7

[36] J. Babic and V. Podobnik, "A review of agent-based modelling of electricity markets in future energy eco-systems," in *2016 International Multidisciplinary Conference on Computer and Energy Science (SpliTech)*, July 2016, pp. 1–9. 7

[37] Y. Guo, H. Zhang, J. Dong, D. Shen, and J. Yin, "Simulation for promotion of solar energy diffusion in residential consumer market with agent-based modeling and random forest," in *2014 Sixth International Conference on Intelligent Human-Machine Systems and Cybernetics*, vol. 2, Aug 2014, pp. 301–304. 7

[38] W. W. L. Wong, Z. Z. Feng, and H. H. Thein, "A parallel sliding region algorithm to make agent-based modeling possible for a large-scale simulation: Modeling hepatitis c epidemics in canada," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 6, pp. 1538–1544, Nov 2016. 7

[39] F. Miksch, P. Pichler, K. J. P. Espinosa, K. S. T. Casera, A. N. Navarro, and M. Bicher, "An agent-based epidemic model for dengue simulation in the philippines," in *2015 Winter Simulation Conference (WSC)*, Dec 2015, pp. 3202–3203. 7

[40] M. Saravanan, P. Karthikeyan, A. Arathi, M. Kiruthika, and S. Suganya, "Mobile agent-based approach for modeling the epidemics of communicable diseases," in *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, Aug 2013, pp. 16–20. 7

[41] J. Pleyer and C. Fleck, "Agent-based models in cellular systems," *Frontiers in Physics*, vol. 10, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fphy.2022.968409 7

[42] K. M. Miller and S. A. Yoon, "Teaching complexity in biology through agent-based simulations: the relationship between students' knowledge of complex systems and metamodeling knowledge," *Frontiers in Education*, vol. 8, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/feduc.2023.1198307 7

[43] M. Suárez-Muñoz, F. Bonet-García, J. A. Hódar, J. Herrero, M. Tanase, and L. Torres-Muros, "Instar: An agent-based model that integrates existing knowledge to simulate the population dynamics of a forest pest," *Ecological Modelling*, vol. 411, p. 108764, 2019. 7

[44] R. S. Beltran, J. W. Testa, and J. M. Burns, "An agent-based bioenergetics model for predicting impacts of environmental change on a top marine predator, the weddell seal," *Ecological Modelling*, vol. 351, pp. 36–50, 2017. 7

[45] R. K. Heikkinen, M. Luoto, M. B. Araújo, R. Virkkala, W. Thuiller, and M. T. Sykes, "Methods and uncertainties in bioclimatic envelope modelling under climate change," *Progress in Physical Geography*, vol. 30, no. 6, pp. 751–777, 2006. 7

[46] Y.-S. Kwon, M.-J. Bae, S.-J. Hwang, S.-H. Kim, and Y.-S. Park, "Predicting potential impacts of climate change on freshwater fish in korea," *Ecological Informatics*, vol. 29, pp. 156–165, 2015. 7, 12, 13

[47] P. D. Moore, "Back to the future: biogeographical responses to climate change," *Progress in physical geography*, vol. 27, no. 1, pp. 122–129, 2003. 7

[48] C. Parmesan and G. Yohe, "A globally coherent fingerprint of climate change impacts across natural systems," *Nature*, vol. 421, no. 6918, pp. 37–42, 2003. 7

[49] G.-R. Walther, E. Post, P. Convey, A. Menzel, C. Parmesan, T. J. Beebee, J.-M. Fromentin, O. Hoegh-Guldberg, and F. Bairlein, "Ecological responses to recent climate change," *Nature*, vol. 416, no. 6879, pp. 389–395, 2002. 7

[50] R. A. Stillman, S. F. Railsback, J. Giske, U. Berger, and V. Grimm, "Making predictions in a changing world: the benefits of individual-based ecology," *BioScience*, vol. 65, no. 2, pp. 140–150, 2015. 7

[51] D. L. DeAngelis and V. Grimm, "Individual-based models in ecology after four decades," *F1000Prime Rep*, vol. 6, no. 39, p. 6, 2014. 7, 8, 9, 11

[52] F. Cánovas, C. Magliozzi, F. Mestre, J. A. Palazón, and M. González-Wangüemert, "Enirg: R-grass interface for efficiently characterizing the ecological niche of species and predicting habitat suitability," *Ecography*, vol. 39, no. 6, pp. 593–598, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.01426 7

[53] F. Mestre, F. Cánovas, R. Pita, A. Mira, and P. Beja, "An r package for simulating metapopulation dynamics and range expansion under environmental change," *Environmental Modelling & Software*, vol. 81, pp. 40–44, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364815216300718 7

[54] Q. Chen, R. Han, F. Ye, and W. Li, "Spatio-temporal ecological models," *Ecological Informatics*, vol. 6, no. 1, pp. 37–43, 2011. 7, 8

[55] W. Tang and D. A. Bennett, "Agent-based modeling of animal movement: A review," *Geography Compass*, vol. 4, no. 7, pp. 682–700, 2010. 8, 9

[56] T. Filatova, P. H. Verburg, D. C. Parker, and C. A. Stannard, "Spatial agent-based models for socio-ecological systems: challenges and prospects," *Environmental modelling & software*, vol. 45, pp. 1–7, 2013. 8, 9, 16, 17, 18

[57] L. An, V. Grimm, A. Sullivan, B. Turner II, N. Malleson, A. Heppenstall, C. Vincenot, D. Robinson, X. Ye, J. Liu, E. Lindkvist, and W. Tang, "Challenges, tasks, and opportunities in modeling agent-based complex systems," *Ecological Modelling*, vol. 457, p. 109685, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030438002100243X 8

[58] C. James and K. Bradshaw, "Agent-based model development of a complex socio-ecological system: Methods for overcoming data and domain limitations," *Ecological Informatics*, p. 102224, 2023. 8

[59] G. Wallentin, "Spatial simulation: A spatial perspective on individual-based ecology—a review," *Ecological Modelling*, vol. 350, pp. 30–41, 2017. 9, 18

[60] S. C. Bankes, "Agent-based modeling: A revolution?" *Proceedings of the National Academy of Sciences*, vol. 99, no. suppl 3, pp. 7199–7200, 2002. 8

[61] D. Helbing, "Agent-based modeling," in *Social self-organization*. Springer, 2012, pp. 25–70. 8

[62] E. Bonabeau, "Agent-based modeling: Methods and techniques for simulating human systems," *Proceedings of the national academy of sciences*, vol. 99, no. suppl 3, pp. 7280–7287, 2002. 8

[63] C. M. Macal and M. J. North, "Tutorial on agent-based modeling and simulation," in *Proceedings of the 37th Conference on Winter Simulation*, ser. WSC '05. Winter Simulation Conference, 2005, pp. 2–15. [Online]. Available: http://dl.acm.org/citation.cfm?id=1162708.1162712 9

[64] N. R. Jennings, "An agent-based approach for building complex software systems," *Commun. ACM*, vol. 44, no. 4, pp. 35–41, Apr. 2001. [Online]. Available: http://doi.acm.org/10.1145/367211.367250 9

[65] J. Nicholas, "On agent-based software engineering," *Artificial Intelligence*, vol. 117, no. 2, pp. 277 – 296, 2000. [Online]. Available: //www.sciencedirect.com/science/article/pii/S0004370299001071 9

[66] M. Wooldridge, "Agent-based software engineering," 1997. 9

[67] M. G. Richiardi, "Agent-based computational economics: a short introduction," *The Knowledge Engineering Review*, vol. 27, no. 2, p. 137–149,

Apr 2012. [Online]. Available: https://www.cambridge.org/core/article/div-class-title-agent-based-computational-economics-a-short-introduction-div/38A8453D740FDCA45E15E335501D8FDF 9

[68] C. Macal and M. North, "Introductory tutorial: Agent-based modeling and simulation," in *Proceedings of the Winter Simulation Conference 2014*, Dec 2014, pp. 6–20. 9

[69] C. M. Macal and M. J. North, "Agent-based modeling and simulation: Abms examples," in *2008 Winter Simulation Conference*, Dec 2008, pp. 101–112. 10

[70] H. Van Dyke Parunak, R. Savit, and R. L. Riolo, *Agent-Based Modeling vs. Equation-Based Modeling: A Case Study and Users' Guide*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 10–25. [Online]. Available: http://dx.doi.org/10.1007/10692956_2 10

[71] N. Gilbert, *Agent-based models*. Sage, 2008, no. 153. 10

[72] Y. Kim and C. McGraw, "Use of agent-based modeling for e-governance research," in *Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance*, ser. ICEGOV '12. New York, NY, USA: ACM, 2012, pp. 531–534. [Online]. Available: http://doi.acm.org/10.1145/2463728.2463850 10

[73] H. Kaiser, "The dynamics of populations as result of the properties of individual animals." *Fortschritte der Zoologie*, 1979. 10

[74] M. Huston, D. DeAngelis, and W. Post, "New computer models unify ecological theory," *BioScience*, vol. 38, no. 10, pp. 682–691, 1988. 10

[75] J. Uchmański and V. Grimm, "Individual-based modelling in ecology: what makes the difference?" *Trends in Ecology & Evolution*, vol. 11, no. 10, pp. 437–441, 1996. 10

[76] D. L. DeAngelis, L. J. Gross *et al.*, *Individual-based models and approaches in ecology*. Chapman & Hall, 1992. 11

[77] V. Grimm, "Ten years of individual-based modelling in ecology: what have we learned and what could we learn in the future?" *Ecological modelling*, vol. 115, no. 2, pp. 129–148, 1999. 11

[78] A. Lomnicki, "Population ecology of individuals." *Monographs in population biology*, vol. 25, pp. 1–216, 1987. 11

[79] B. Breckling, F. Müller, H. Reuter, F. Hölker, and O. Fränzle, "Emergent properties in individual-based ecological models—introducing case studies in an ecosystem research context," *Ecological modelling*, vol. 186, no. 4, pp. 376–388, 2005. 11

[80] S. F. Railsback and V. Grimm, *Agent-based and individual-based modeling: a practical introduction*. Princeton university press, 2011. 11

[81] ——, *Agent-based and individual-based modeling: a practical introduction*. Princeton university press, 2019. 12

[82] C. Chikuruwo, M. Masocha, A. Murwira, H. Ndaimani *et al.*, "Predicting the suitable habitat of the invasive xanthium strumarium l. in southeastern zimbabwe," *Applied Ecology and Environmental Research*, vol. 15, no. 1, pp. 17–32, 2017. 12, 13

[83] L. Hill, A. Hector, G. Hemery, S. Smart, M. Tanadini, and N. Brown, "Abundance distributions for tree species in great britain: A two-stage approach to modeling abundance using species distribution modeling and random forest," *Ecology and Evolution*, vol. 7, no. 4, pp. 1043–1056, 2017. 12, 13

[84] L. Abade, D. Macdonald, and A. Dickman, "Assessing the relative importance of landscape and husbandry factors in determining large carnivore depredation risk in tanzania's ruaha landscape," *Biological Conservation*, vol. 180, pp. 241–248, 2014. 12

[85] N. T. Garavito, A. C. Newton, D. Golicher, and S. Oldfield, "The relative impact of climate change on the extinction risk of tree species in the montane tropical andes," *PloS one*, vol. 10, no. 7, p. e0131388, 2015. 12

[86] Y. Sun, T. Wang, A. K. Skidmore, S. C. Palmer, X. Ye, C. Ding, and Q. Wang, "Predicting and understanding spatio-temporal dynamics of species recovery: implications for asian crested ibis nipponia nippon conservation in china," *Diversity and Distributions*, vol. 22, no. 8, pp. 893–904, 2016. 12

[87] P. J. Mitchell, J. Monk, and L. Laurenson, "Sensitivity of fine-scale species distribution models to locational uncertainty in occurrence data across multiple sample sizes," *Methods in Ecology and Evolution*, vol. 8, no. 1, pp. 12–21, 2017. 12

[88] S. Feldmeier, L. Schefczyk, N. Wagner, G. Heinemann, M. Veith, and S. Lötters, "Exploring the distribution of the spreading lethal salamander chytrid fungus in its invasive range in europe–a macroecological approach," *PloS one*, vol. 11, no. 10, p. e0165682, 2016. 12

[89] F. K. Hoehler, "Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity," *Journal of clinical epidemiology*, vol. 53, no. 5, pp. 499–503, 2000. 13

[90] R. Sor, Y.-S. Park, P. Boets, P. L. Goethals, and S. Lek, "Effects of species prevalence on the performance of predictive models," *Ecological Modelling*, vol. 354, pp. 11–19, 2017. 13

[91] B. Z. Oh, A. M. Sequeira, M. G. Meekan, J. L. Ruppert, and J. J. Meeuwig, "Predicting occurrence of juvenile shark habitat to improve conservation planning," *Conservation Biology*, 2017. 13

[92] S. McIntyre, E. F. Rangel, P. D. Ready, and B. M. Carvalho, "Species-specific ecological niche modelling predicts different range contractions for lutzomyia intermedia and a related vector of leishmania braziliensis following climate change in south america," *Parasites & vectors*, vol. 10, no. 1, p. 157, 2017. 13

[93] M. Ángel Matus-Hernández, R. O. Martínez-Rincón, R. J. Aviña-Hernández, and N. Y. Hernández-Saavedra, "Landsat-derived environmental factors to describe habitat preferences and spatiotemporal distribution of phytoplankton," *Ecological Modelling*, vol. 408, p. 108759, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304380019302674 13

[94] P. J. Dolder, C. Minto, J.-M. Guarini, and J. J. Poos, "Highly resolved spatiotemporal simulations for exploring mixed fishery dynamics," *Ecological Modelling*, vol. 424, p. 109000, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304380020300727 13

[95] S. Sadchatheeswaran, G. M. Branch, L. J. Shannon, M. Coll, and J. Steenbeek, "A novel approach to explicitly model the spatiotemporal impacts of structural complexity created by alien ecosystem engineers in a marine benthic environment," *Ecological Modelling*, vol. 459, p. 109731, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304380021002830 13

[96] T. Poisot, D. B. Stouffer, and D. Gravel, "Beyond species: why ecological interaction networks vary through space and time," *Oikos*, vol. 124, no. 3, pp. 243–251, 2015. 14

[97] H. Reuter, M. Kruse, A. Rovellini, and B. Breckling, "Evolutionary trends in fish schools in heterogeneous environments," *Ecological Modelling*, vol. 326, pp. 23–35, 2016. 15, 17

[98] H. R. Parry and M. Bithell, "Large scale agent-based modelling: A review and guidelines for model scaling," *Agent-based models of geographical systems*, pp. 271–308, 2012. 15, 49

[99] N. S. Morales and G. L. Perry, "A spatial simulation model to explore the long-term dynamics of podocarp-tawa forest fragments, northern new zealand," *Ecological Modelling*, vol. 357, pp. 35 – 46, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0304380016306068 15, 17, 19

[100] P. Fust and E. Schlecht, "Integrating spatio-temporal variation in resource availability and herbivore movements into rangeland management: Ramdry—an agent-based model on livestock feeding ecology in a dynamic, heterogeneous, semi-arid

environment," *Ecological Modelling*, vol. 369, pp. 13 – 41, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0304380016304227 16, 17, 18, 19

[101] J. Zhang, T. E. Dennis, T. J. Landers, E. Bell, and G. L. Perry, "Linking individual-based and statistical inferential models in movement ecology: A case study with black petrels (procellaria parkinsoni)," *Ecological Modelling*, vol. 360, pp. 425 – 436, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0304380016304860 16, 17, 18

[102] T. M. Anderson and S. Dragićević, "Geospatial pest-parasitoid agent based model for optimizing biological control of forest insect infestation," *Ecological modelling*, vol. 337, pp. 310–329, 2016. 16, 17, 18

[103] H. Parry, C. Paull, M. Zalucki, A. Ives, A. Hulthen, and N. Schellhorn, "Estimating the landscape distribution of eggs by helicoverpa spp., with implications for bt resistance management," *Ecological Modelling*, vol. 365, pp. 129–140, 2017. 17, 19

[104] W. F. Fagan, M. A. Lewis, M. Auger-Méthé, T. Avgar, S. Benhamou, G. Breed, L. LaDage, U. E. Schlägel, W.-w. Tang, Y. P. Papastamatiou *et al.*, "Spatial memory and animal movement," *Ecology letters*, vol. 16, no. 10, pp. 1316–1329, 2013. 17

[105] P. Meyfroidt, "Environmental cognitions, land change, and social–ecological feedbacks: an overview," *Journal of Land Use Science*, vol. 8, no. 3, pp. 341–367, 2013. 17

[106] S. Manson, L. An, K. C. Clarke, A. Heppenstall, J. Koch, B. Krzyzanowski, F. Morgan, D. O'Sullivan, B. C. Runck, E. Shook *et al.*, "Methodological issues of spatial agent-based models," *JASSS-THE JOURNAL OF ARTIFICIAL SOCIETIES AND SOCIAL SIMULATION*, vol. 23, no. 1, 2020. 18

[107] D. O'Sullivan and G. L. Perry, *Spatial simulation: exploring pattern and process*. John Wiley & Sons, 2013. 18

[108] I. Lorscheid, U. Berger, V. Grimm, and M. Meyer, "From cases to general principles: A call for theory development through agent-based modeling," *Ecological Modelling*, vol. 393, pp. 153–156, 2019. 20

[109] V. Grimm, U. Berger, F. Bastiansen, S. Eliassen, V. Ginot, J. Giske, J. Goss-Custard, T. Grand, S. K. Heinz, G. Huse *et al.*, "A standard protocol for describing individual-based and agent-based models," *Ecological modelling*, vol. 198, no. 1-2, pp. 115–126, 2006. 21

[110] V. Grimm, S. F. Railsback, C. E. Vincenot, U. Berger, C. Gallagher, D. L. DeAngelis, B. Edmonds, J. Ge, J. Giske, J. Groeneveld *et al.*, "The odd protocol for describing agent-based and other simulation models: A second update to improve clarity,

replication, and structural realism," *Journal of Artificial Societies and Social Simulation*, vol. 23, no. 2, 2020. 21

[111] R. J. Hijmans, S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis, "Very high resolution interpolated climate surfaces for global land areas," *International Journal of Climatology*, vol. 25, no. 15, pp. 1965–1978, 2005. [Online]. Available: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.1276 22, 69

[112] J. Elith and J. R. Leathwick, "Species distribution models: Ecological explanation and prediction across space and time," *Annual Review of Ecology, Evolution, and Systematics*, vol. 40, no. 1, pp. 677–697, 2009. [Online]. Available: https://doi.org/10.1146/annurev.ecolsys.110308.120159 23, 81

[113] E. F. Moore, "Machine models of self-reproduction," in *Proceedings of symposia in applied mathematics*, vol. 14, no. 5. American Mathematical Society New York, 1962, pp. 17–33. 25, 50, 82

[114] J. Bioco, P. Fazendeiro, F. Cánovas, and P. Prata, "Parameterization of an agent-based model of spatial distribution of species," in *Emerging Technologies in Computing*, M. H. Miraz, P. S. Excell, A. Ware, S. Soomro, and M. Ali, Eds. Cham: Springer International Publishing, 2020, pp. 251–260. 26

[115] M. Kuhnert, A. Voinov, and R. Seppelt, "Comparing raster map comparison algorithms for spatial modeling and analysis," *Photogrammetric Engineering & Remote Sensing*, vol. 71, no. 8, pp. 975–984, 2005. 29, 62

[116] L. J. Beaumont, E. Graham, D. E. Duursma, P. D. Wilson, A. Cabrelli, J. B. Baumgartner, W. Hallgren, M. Esperón-Rodríguez, D. A. Nipperess, D. L. Warren *et al.*, "Which species distribution models are more (or less) likely to project broad-scale, climate-induced shifts in species ranges?" *Ecological Modelling*, vol. 342, pp. 135–146, 2016. 39

[117] J. Elith*, C. H. Graham*, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann *et al.*, "Novel methods improve prediction of species' distributions from occurrence data," *Ecography*, vol. 29, no. 2, pp. 129–151, 2006. 39, 61

[118] J. VanDerWal, L. P. Shoo, C. Graham, and S. E. Williams, "Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know?" *Ecological Modelling*, vol. 220, no. 4, pp. 589–594, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304380008005486 39

[119] M. S. Wisz and A. Guisan, "Do pseudo-absence selection strategies influence species distribution models and their predictions? an information-theoretic approach based on simulated data," *BMC ecology*, vol. 9, no. 1, pp. 1–13, 2009. 39

[120] "Scikit-learn. machine learning in python," https://scikit-learn.org/stable/index.html, accessed: 2021-05-30. 39

[121] F. Cánovas, P. De la Rúa, J. Serrano, and J. Galián, "Analysis of a contact area between two distinct evolutionary honeybee units: an ecological perspective," *Journal of insect conservation*, vol. 18, no. 5, pp. 927–937, 2014. 40

[122] M. M. Ribeiro, N. Roque, S. Ribeiro, C. Gavinhos, I. Castanheira, L. Quinta-Nova, T. Albuquerque, and S. Gerassis, "Bioclimatic modeling in the last glacial maximum, mid-holocene and facing future climatic changes in the strawberry tree (arbutus unedo l.)," *PLOS ONE*, vol. 14, no. 1, pp. 1–15, 01 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0210062 41

[123] J. J. Danielson and D. B. Gesch, *Global multi-resolution terrain elevation data 2010 (GMTED2010)*. US Department of the Interior, US Geological Survey, 2011. 42

[124] GDAL/OGR contributors, *GDAL/OGR Geospatial Data Abstraction software Library*, Open Source Geospatial Foundation, 2022. [Online]. Available: https://gdal.org 42

[125] A. Ligmann-Zielinska, P.-O. Siebers, N. Magliocca, D. C. Parker, V. Grimm, J. Du, M. Cenek, V. Radchuk, N. N. Arbab, S. Li *et al.*, "'one size does not fit all': A roadmap of purpose-driven mixed-method pathways for sensitivity analysis of agent-based models," *Journal of Artificial Societies and Social Simulation*, vol. 23, no. 1, pp. 1–6, 2020. 47

[126] N. Fachada, V. V. Lopes, R. C. Martins, and A. C. Rosa, "Parallelization strategies for spatial agent-based models," *International Journal of Parallel Programming*, vol. 45, no. 3, pp. 449–481, 2017. 49

[127] A. Voss, J.-Y. You, E. Yen, H.-Y. Chen, S. Lin, A. Turner, and J.-P. Lin, "Scalable social simulation: investigating population-scale phenomena using commodity computing," in *2010 IEEE Sixth International Conference on e-Science*. IEEE, 2010, pp. 1–8. 49

[128] P. Heywood, S. Maddock, J. Casas, D. Garcia, M. Brackstone, and P. Richmond, "Data-parallel agent-based microscopic road network simulation using graphics processing units," *Simulation Modelling Practice and Theory*, vol. 83, pp. 188–200, 2018. 49

[129] M. K. Chimeh, P. Heywood, M. Pennisi, F. Pappalardo, and P. Richmond, "Parallel pair-wise interaction for multi-agent immune systems modelling," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018, pp. 1367–1373. 49

[130] M. K. Chimeh *et al.*, "Parallelisation strategies for agent based simulation of immune systems," *BMC bioinformatics*, vol. 20, no. 6, pp. 1–14, 2019. 49

[131] S. Coakley, M. Gheorghe, M. Holcombe, S. Chin, D. Worth, and C. Greenough, "Exploitation of high performance computing in the flame agent-based simulation framework," in *2012 IEEE 14th International Conference on High Performance Computing and Communication & 2012 IEEE 9th International Conference on Embedded Software and Systems*. IEEE, 2012, pp. 538–545. 50

[132] R. A. Williams, "User experiences using flame: A case study modelling conflict in large enterprise system implementations," *Simulation Modelling Practice and Theory*, vol. 106, p. 102196, 2021. 50

[133] N. Collier and M. North, "Parallel agent-based simulation with repast for high performance computing," *Simulation*, vol. 89, no. 10, pp. 1215–1235, 2013. 50

[134] N. Collier, J. Ozik, and C. M. Macal, "Large-scale agent-based modeling with repast hpc: A case study in parallelizing an agent-based model," in *European Conference on Parallel Processing*. Springer, 2015, pp. 454–465. 50

[135] Q. Zhang, R. R. Vatsavai, A. Shashidharan, and D. V. Berkel, "Agent based urban growth modeling framework on apache spark," in *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, 2016, pp. 50–59. 50

[136] O. Bandman, "Coarse-grained parallelization of cellular-automata simulation algorithms," in *International Conference on Parallel Computing Technologies*. Springer, 2007, pp. 370–384. 51, 52

[137] R. Muscarella, P. J. Galante, M. Soley-Guardia, R. A. Boria, J. M. Kass, M. Uriarte, and R. P. Anderson, "Enm eval: An r package for conducting spatially independent evaluations and estimating optimal model complexity for maxent ecological niche models," *Methods in ecology and evolution*, vol. 5, no. 11, pp. 1198–1205, 2014. 61, 79

[138] N. Golding, T. A. August, T. C. Lucas, D. J. Gavaghan, E. E. van Loon, and G. McInerny, "The zoon r package for reproducible and shareable species distribution modelling," *Methods in Ecology and Evolution*, vol. 9, no. 2, pp. 260–268, 2018. 61, 79

[139] A. B. Smith, M. J. Santos, M. S. Koo, K. M. Rowe, K. C. Rowe, J. L. Patton, J. D. Perrine, S. R. Beissinger, and C. Moritz, "Evaluation of species distribution models by resampling of sites surveyed a century ago by joseph grinnell," *Ecography*, vol. 36, no. 9, pp. 1017–1031, 2013. 61

[140] L. Pellissier, K. Anne Bråthen, J. Pottier, C. F. Randin, P. Vittoz, A. Dubuis, N. G. Yoccoz, T. Alm, N. E. Zimmermann, and A. Guisan, "Species distribution models

reveal apparent competitive and facilitative effects of a dominant species on the distribution of tundra plants," *Ecography*, vol. 33, no. 6, pp. 1004–1014, 2010. 61

[141] E. M. Rubidge, W. B. Monahan, J. L. Parra, S. E. Cameron, and J. S. Brashares, "The role of climate, habitat, and species co-occurrence as drivers of change in small mammal distributions over the past century," *Global Change Biology*, vol. 17, no. 2, pp. 696–708, 2011. 61

[142] R. Frankham and B. W. Brook, "The importance of time scale in conservation biology and ecology," in *Annales Zoologici Fennici*. JSTOR, 2004, pp. 459–463. 61

[143] B. T. Barton, "Time in ecology: A theoretical framework. eric post. 2019. princeton university press, princeton, new jersey, usa. 224 pp. $40.00 paperback. isbn: 978-0-691-18235-3." 2020. 61

[144] R. Frankham, S. E. J. D. Ballou, D. A. Briscoe, and J. D. Ballou, *Introduction to conservation genetics*. Cambridge university press, 2002. 61

[145] J. M. Kass, B. Vilela, M. E. Aiello-Lammens, R. Muscarella, C. Merow, and R. P. Anderson, "Wallace: A flexible platform for reproducible modeling of species niches and distributions built for community expansion," *Methods in Ecology and Evolution*, vol. 9, no. 4, pp. 1151–1156, 2018. 79

[146] B. Naimi and M. B. Araújo, "sdm: a reproducible and extensible r platform for species distribution modelling," *Ecography*, vol. 39, no. 4, pp. 368–375, 2016. 79

[147] W. Thuiller, B. Lafourcade, R. Engler, and M. B. Araújo, "Biomod–a platform for ensemble forecasting of species distributions," *Ecography*, vol. 32, no. 3, pp. 369–373, 2009. 79

[148] J. L. Brown, J. R. Bennett, and C. M. French, "Sdmtoolbox 2.0: the next generation python-based gis toolkit for landscape genetic, biogeographic and species distribution model analyses," *PeerJ*, vol. 5, p. e4095, 2017. 79

[149] C. N. Meynard, B. Leroy, and D. M. Kaplan, "Testing methods in species distribution modelling using virtual species: what have we learnt and what are we missing?" *Ecography*, vol. 42, no. 12, pp. 2021–2036, 2019. 80

[150] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962. 81

[151] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 1, pp. 99–102, 1974. 81

[152] A. Lipowski and D. Lipowska, "Roulette-wheel selection via stochastic acceptance," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 6, pp. 2193–2196, 2012. 82