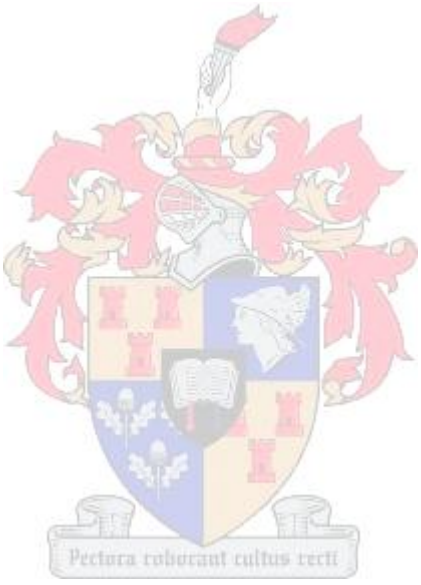
Modelling control strategies for the invasive tree species *Prosopis* in the Northern Cape

Fuzail Dawood



Final year project presented in partial fulfilment of the requirements for the degree of
Bachelor of Engineering (Industrial Engineering)
in the Faculty of Engineering at Stellenbosch University

Supervisor: Mr A Flemming
Co-supervisor: Prof JH van Vuuren

December 2021

Declaration

By submitting this project electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: December 1, 2021

Abstract

Biological *invasive* species have proven to threaten biodiversities and economies worldwide, often resulting in devastating environmental and socio-economic effects. As such, there is a growing importance for humans to understand and effectively manage the complex interactions that arise with biological invasions in order to accurately predict the expected behaviour of the species, and in doing so, improve the effectiveness of the control strategies adopted by decision-makers and invasive species management.

The objective of this project is to capture the response of the invasive tree species, *Prosopis*, in the presence of a modelled control strategy. As such, the proposed model seeks to investigate and predict the extent to which the species spreads when confronted with the strategic implementation of an effective contol method. In particular, the paradigm of *machine learning* is adopted for predicting the habitat suitability of Prosopis. By employing the aforementioned habitat suitability distribution, a *cellular automata* approach within the realm of simulation modelling is adopted across a hexagonally discretised study region, and is deeply-rooted in the mathematical modelling of population growth.

The inputs of the machine learning model comprises *topographical*, *bioclimatic*, and species observation data, while its output is the habitat suitability of Prosopis within the study region. The cellular automata model's inputs includes the habitat suitability scores from the output of the machine learning model, a known density distribution range of Prosopis, a growth rate, a dispersal rate, and a transition rule governing the change in states during the execution of the model. As such, the model outputs the updated state of the discretised areas at each time step for the duration of the observed study period.

The practicality of the developed cellular automata model is illustrated by the implementation of the model in the form of a real-world case study in a municipality within the Northern Cape. This municipality was identified as a region in which Prosopis is densely populated, and is the region in which the cellular automata model is executed, in order to compare the spread of Prosopis with, and without the implementation of a control strategy.

Graduate Attributes Reference

Attribute	Reference	
	Section	**Page**
1. Problem solving: Demonstrate competence to identify, assess, formulate and solve convergent and divergent engineering problems creatively and innovatively.	*All*	*All*
5. Engineering methods, skills and tools, including information technology: Demonstrate competence to use appropriate engineering methods, skills and tools, including those based on information technology.	*2,3,4 & 5*	*9–59*
6. Professional and technical communication: Demonstrate competence to communicate effectively, both orally and in writing, with engineering audiences and the community at large.	*All*	*All*
9. Independent learning ability: Demonstrate competence to engage in independent learning through well developed learning skills.	*2,3,4 & 5*	*9–59*
10. Engineering professionalism: Demonstrate critical awareness of the need to act professionally and ethically and to exercise judgement and take responsibility within own limits of competence.	*All*	*All*

Acknowledgements

The author wishes to acknowledge the following people and institutions for their various contributions towards the completion of this work:

- My supervisor, Alexander Flemming, for being the best supervisor I could've asked for. Thank you for your support, patience, availability, and countless hours spent reviewing my work to ensure that I produce my best possible work. I've learn't a lot from you this year and your dedicaton for conducting thorough research is something which I strive towards.
- Prof JH van Vuuren, for affording me the opportunity to be apart of such an enriching research group. Thank your for believing in me and for your willingness to share your knowledge and wisdom. Your attention to detail and pursuit of excellence is truly remarkable and has inspired me throughout this year.
- *The Stellenbosch Unit for Operations Research in Engineering* (SUnORE) and its members for their invaluable advice, guidance, and willingness to assist.
- My parents and siblings, for their unwavering love and support throughout the last four years of my undergraduate studies. Thank you for understanding the commitments and sacrifices which I've made, particularly in the last year, and I hope to make you proud.
- My friends, for their support and ecouragement throughout the past four years and for helping me maintain a somewhat balanced life.
- The Almighty, for blessing me with the ability to persevere throughout the last four years and for guiding me with each and every step.

Table of Contents

List of Acronyms

AUC: Area under the receiver operating characteristic curve

CA: Cellular automata

CABI: Centre for agriculture and bioscience international

DEA: Department of environmental affairs

DFD: Data flow diagram

GIS: Geographical Information System

GISD: Global invasive species database

KNN: k-Nearest neighbour

ML: Machine learning

NEM:BA: National environmental management: biodiversity act

ROC: Receiver operating characteristic curve

SDM: Species distribution model

SME: Subject matter expert

SUnORE: Stellenbosch Unit for Operations Research in Engineering

WfW: Working for Water

List of Figures

List of Tables

List of Algorithms

CHAPTER 1

Introduction

Contents

1.1 Background

The profound effect of humans on the natural environment has sparked a debate over the age in which humans exist, leading to many calling it the *Anthropocene* era. *Biological invasive species* are examples of the profound and often devastating effects of this era caused by human intervention. The need for humans to understand and effectively manage the complex interactions that arise with biological invasive species is becoming increasingly important [100]. The development of a species having been *introduced* into a non-indigenous environment is widely contentious in literature pertaining to biological invasions. A brief elucidation of a closely related group of terms is imperative towards understanding the nuances associated with species invasions. The introduction of a species refers to the displacement of a *native* species (*i.e.* from its place of origin), by humans, across a significant geographical border. *Naturalisation* occurs when factors impeding the survival and reproduction of the species are overpowered. Finally, the condition of defining a biological invasion is that the naturalised species reproductively produces offspring at far distances from the initial location of introduction.

The classification of a species as invasive is also to reflect the severe environmental and socio-economic effects that these species have within the region in which it has naturalised. Environmental effects include reduced diversity and density of native fauna and flora species [92], as a result of having to compete for the resources of the land. Furthermore, the degradation of an ecosystem will likely result in a decline in soil quality and a disturbed water supply in the region [80]. Socio-economic effects of invasions include consequences on human health, decreased profits of farmers due to lower crop yields, as well as increased costs for farmers and locals as a result of having to repair damaged infrastructure [88]. An alternative interpretation by Richardson *et al.* [74], places more emphasis on non-indigenous species surmounting many

biological hindrances from the time it was introduced, such as survival and reproduction, until it eventually becomes invasive.

Many species have often been displaced out of their natural habitat through human intervention for aesthetic or agricultural reasons, however, few species become naturalised and even invasive [71], spreading uncontrollably, and subsequently unsettling the biodiversity of the environment as well as the livelihood of locals [37, 68]. Invasive plant species have significantly threatened global and local native biodiversities in recent decades by consuming the limited resources available of a region, by effectively competing with indigenous species for these resources [73]. As a result of both the invasions by non-indigenous species and the efforts implemented to combat these invasions, many socio-economic factors in the invaded region are often detrimentally affected. The destruction of local ecosystems, water supplies, livestock, crop production, infrastructure, and human well-being are just some of the ravaging impacts documented by those affected by biological invasions. As a result, the financial implications of having to restore damaged assets and systems are often straining, especially to developing countries [45].

The global distribution of recorded invasive alien species per country in 2016 is depicted in Figure 1.1 and supports the notion that biological invasions are experienced globally and have the potential to alter the biodiversity of the world as a whole [89]. In addition to this, the phenomenon of *climate change* has a compounding effect on the spread of invasive species. This is because climatic events (*e.g.* floods and cyclones) worldwide foster opportunities for the species to be spread to new regions, allowing them to naturalise, and potentially become invasive and destructive within these regions [35].

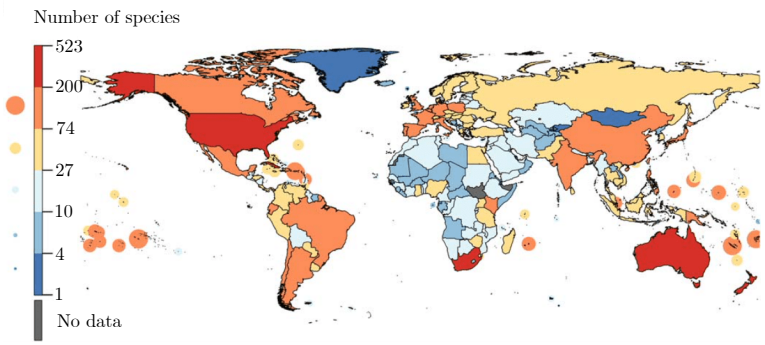


FIGURE 1.1: *Global map illustrating the number of invasive alien species recorded per country as of 2016 in the CABI and GISD databases, adapted from [89].*

South African lawmakers have classified known invasive species within South Africa into three main categories, defined by the *National Environmental Management: Biodiversity Act* (NEM:BA) based on their use and impact. Category 1a consists of identified species that should, by law, be eradicated from the region. Category 1b stipulates that the identified species must be controlled in order to avoid potential invasion. In addition, Category 1b invaders should also, as far as possible, be eradicated from the region as their invasive potential is highly threatening [63]. Trade and further transmission of Category 1 species are strictly forbidden. Category 2 specifies that the use of invasive or potentially invasive species, including commoditised species, requires one to be in possession of a permit [18]. Category 3 rules that existing species that have already been introduced to the land may be left as is, however, further transmission, use, and trade of the species is strictly forbidden.

Over the last two centuries, South Africa and more than 100 other countries and islands worldwide have been invaded by the plant genus, *Prosopis* (also known as *Mesquite*) [45], a group of deciduous and leguminous tree species native to the Americas. In the latter part of the 19th

century, farmers in the arid regions of South Africa were counselled to plant Prosopis on their farms in an attempt to provide a source of fodder, shade, and firewood [81]. After many years of extensive dispersal, Prosopis naturalised in South Africa and was eventually classified invasive once the adverse effects were realised [53]. These adverse effects include a disturbed ecosystem, water-supply contamination, ravaged underground pipes and boreholes, and a reduction in grazing potential [46].

Fluctuating densities of the species can be located in 61 out of the 234 municipalities in the arid and semi-arid inland regions of South Africa [81]. A 1998 survey conducted by Versfeld *et al.* [96] discovered that approximately 1.8 million hectares of South Africa had been invaded by Prosopis. By then, the Northern Cape was identified as the province of concern as it hosted 990 000 hectares (equal to 55% of the total Prosopis invasions across South Africa in 1998) of the invader in this region. According to a remote sensing and GIS investigation, Van den Berg *et al.* [91] uncovered that by 2007, the distribution of Prosopis had increased to 1.473 million hectares in the Northern Cape. Behind the invasive Australian *Acacia* species, Prosopis has grown to be South Africa's second most invasive tree species. Figure 1.2 illustrates a 2010 distribution of Prosopis in South Africa [103], where each dot represents the presence of Prosopis, per quarter degree square [102].

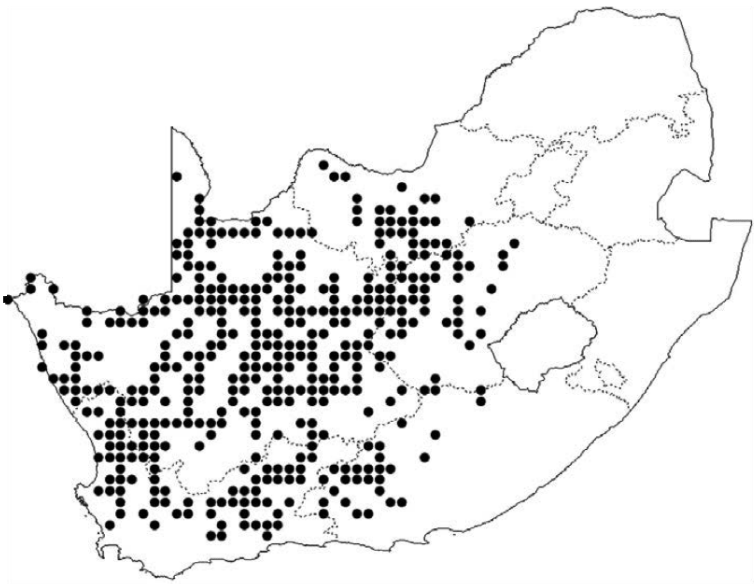


FIGURE 1.2: *The approximate distribution of Prosopis glandulosa* var. *torreyana, Prosopis velutina and their hybrids in South Africa in 2010, adapted from [103].*

As of 2004, Prosopis has been listed by the South African national government as a Category 1b invasive species in the Eastern Cape, Free State, North-West and Western Cape. Throughout the Northern Cape, however, Prosopis is listed as a Category 3 invasive species with the exception of riparian areas, where they are listed as Category 1b invaders [18]. It is important to note that the use of pods which grow on the Prosopis is what causes it to be a Category 3 invasive species (instead of Category 1b) in non-riparian areas. This is primarily due to the fact that the pods serve as a fodder for the livestock of private farmers. This exemption is certainly worth revising, given the vast potential for the invader to spread via the excrement of animals that have consumed the pods and under long distance travelling [81].

A 2001 study conducted by Pimentel *et al.* [69] estimated that global damages caused by invasive species and subsequent effects to manage them, amounted to in excess of US$ 1.4 trillion in 1998,

which was equivalent to 5% of the global economy at the time. South Africa had reportedly suffered a US$ 4.3 billion loss due to the destructive effects of invasive species on crops, pastures, and forests. Despite the data in the 2001 study being more than two decades old, it is likely, given the effects of globalisation and climate change, that the global cost of damages caused by invasive species has also risen [29]. The South African government has spent exorbitant sums of money in an attempt to combat the spread of Prosopis. As of 2019, the government spent an average of approximately ZAR 2 billion annually in the fight against invasive tree species at large, and this figure continues to grow precipitously [92]. In light of the aforementioned financial effects, management strategies have been initiated in parts of the world that experience severe invasions. These management programs are also referred to as investing in ecological infrastructure, or natural resource management [81], and seek to eradicate the destructive species threatening a region.

Current interventions for impeding further spread of invasive species and safeguarding resources includes the use of, or a combination of *mechanical*, *chemical*, and *biological* methods [81]. Of course, each control method will have its strengths and weaknesses and thus evaluating cost-benefit relationships are essential. Briefly, mechanical methods entail removing the invasive species by hand or by using tools [59]. In the case of invasive tree species, this can be achieved by cutting tree stumps or using heavy-duty machines such as bulldozers. Despite being beneficial for areas identified for agricultural purposes, mechanical methods are damaging to the environment, slow, and not ideal for large-scale invasions. Chemical methods entail spraying herbicide on the invasive species, and is typically done by plane. This method is the fastest and most cost effective, however, the environmental consequences of this method remain a grey area. Lastly, biological methods entail the deployment of biological *agents*, a predator species that prey on the invading species. The agents are released into the area with the aim to either hinder the spread, or increase the mortality rate of the invading species. While this control method ensures no human impact on the environment, the research costs associated with identifying the most effective agent are initially high, and much uncertainty exists over a long duration with regard to the success of the deployed agent and strategy [25].

South Africa has utilised seed-feeding beetles as agents against Prosopis in the past, however, they were selected with the purpose to reduce the rate of spread and not contribute to the demise of Prosopis, thereby allowing the invader to persist [103]. While this method has the potential to be very potent, it requires in excess of ZAR 1 million to fund research since the best agent to eradicate Prosopis is yet to be discovered. Stakeholders are, therefore, strongly advised to make an informed decision on which control method (or combination thereof), should be used [45], based on how long the method(s) take to remove the species, the time required, job creation, efficiency of the deployed control method(s), and the subsequent financial implications. Great complexity arises when attempting to manage an introduced species as they are often synchronously advantageous and disadvantageous. The trade-off between the positive and negative contributions of the species often results in conflicts of interest with regards to the most effective control strategy that should be deployed [81].

In spite of a wide range of control methods and resources at their disposal for managing invasive species, many governments have failed to ensure that strategies of senior management are defined and executed on a species-specific level [81]. South Africa is following case-study instances for the invasive Australian Acacia and *Parthenium hysterophorus* species [88, 99] in the hope of developing more species-specific responses to those invaders. In doing so, they attempt to develop general management responses for similar species causing similar problems. However, this generic 'mould' approach has proven futile since the inception of the state-supported programme, Working for Water (WfW) in 1995, because every invasive species has different drivers

and sets of information unique to it. Moreover, between 1997 and 2014, the programme has been criticised for lacking an effective management strategy with regard to the control of invasive species [94].

The two primary objectives of WfW is to: (1) Upskill disadvantaged communities, thereby assisting in reducing unemployment and poverty rates, a long lasting scar of past *Apartheid* policies in South Africa, and (2) restore the damaged environment, and enhance the livelihood of locals through effective management of invasive species. The Department of Environmental Affairs (DEA) manages the programme and contractually works with locals companies for 2–3 months at a time. The contractors are typically from previously disadvantaged communities and work in teams of approximately ten semi-skilled labourers who remove invasive species from identified locations. WfW is exceptionally well supported compared to many other ecosystem-management schemes in South Africa, receiving ZAR 1.1 billion annually from the government to manage biological invasions [81].

With respect to controlling Prosopis, not much success has been experienced largely due to ineffective planning, a skewed focus towards job creation as opposed to ecological successes, and feeble management practices [81]. When reviewing the control strategies attempted, it is clear to see that biological methods in which biological agents are released into the environment have been favoured since the 1980s. Despite this control method not proving to have long term success with controlling Prosopis, improved biocontrol methods are believed by many to be the best solution going forward [103].

Developing a model capable of accurately predicting the extent of biological invasions could improve the quality of stakeholder decisions regarding the control strategies deployed. Growing in popularity, spatial modelling paradigms, together with the *Geographic Information System* (GIS) software and prevalent *machine learning* (ML) algorithms, are being employed to visualise existing, as well as potential distributions of invasive species [97]. These investigative models are commonly referred to as *Species Distribution Models* (SDMs). GIS is a tool used for visually exploring and analysing spatial data in order to better understand spatial patterns associated with a population [4]. ML algorithms entail a computer automatically learning patterns between variables, given historical data. These powerful algorithms are capable of making predictions about future instances, based on the input-output relationships found during the training process [38].

There are two SDM paradigms of interest when attempting to contain the spread of invasive species, namely *correlative* models and *expert-based* models. Correlative models seek to explore the relationship between environmental variables and the presence or absence of the invasive species in that region. This is often accomplished by using many algorithms within the ML paradigm. Since the invader is likely to spread in areas exhibiting similar environmental conditions, the relationships uncovered are important as they are used to predict the extent of invasive species distribution [97]. The expert-based approach, is especially useful in studies where empirical models are unavailable [84]. Here, a model is constructed by seeking the opinions and knowledge of domain experts. The evaluation and judgements of the experts are then used in developing the model that is capable of providing context and insights with respect to a particular focus area. Due to the fact that expert-based models depend more on the subjectivity of the experts than extensively collecting data, expert-based models prove to be effective as SDMs for large-scale studies, in which no single species is focused upon [21]. When employed in a GIS environment, expert models perform well in predicting the locations to which biological invaders will spread [52].

Cellular automata (CA) models have thrived in the branch of correlative species distribution modelling due to their ability of effectively capturing both the spatial and temporal dynamics of a system [77]. Briefly described, CA models are composed of a uniformly discretised grid space

in which all cells contain a discrete variable. The state of each cell is described by the value of the discrete variable, which is determined by its previous state as well as the states of its neighbouring cells. Each iteration pursued results in the updating of cells according to a defined set of transition rules [16].

1.2 Problem Statement

Biological invasions have been found to have an adverse impact on biodiversity, human livelihood, as well as the economy of the invaded region. With this in mind, a transdisciplinary approach consisting of mathematics and science can be deployed in order to assist stakeholders in effectively mitigating the adverse effects of invasive species by use of strategic approach to implementing control methods.

The problem considered in this project is that of modelling control strategies for the invasive tree species, Prosopis, in the Northern Cape region of South Africa. The models developed will focus on capturing the species' response to the implementation of a control strategy.

The fundamental techniques explored are based on mathematical modelling approaches, GIS, ML as well as CA. GIS software will be employed to build an hexagonally discretised spatial data set. Thereafter, an ML model will be developed in order to predict the habitat suitability of Prosopis in the hexagonally discretised study region. The results of the ML model will then be employed as input to the CA, along with the transition rules governing the growth, dispersal, and eradication of Prosopis. By adopting a CA approach with an hexagonally discretised map, the spatio-temporal spread and control of Prosopis will be captured. Finally, the model will be verified by comparing its habitat suitability predictions with existing data, after which it will be validated by being applied to a case study focusing on the Northern Cape province. Comparing the simulated results of the study region with, and without a control strategy implemented may yield valuable insights into the effect which an effective control strategy may have on the spread of Prosopis.

1.3 Problem objectives

The following objectives will be pursued in this project:

I To *conduct* a thorough study of the literature with reference to:

  (a) the impact and spread of the invasive tree species Prosopis,

  (b) the factors driving the invasion of Prosopis,

  (c) prevalent control strategies for the mitigation of invasive species spread,

  (d) using GIS to conduct spatial analysis on the data,

  (e) relevant algorithms within the ML paradigm capable of predicting the distribution of Prosopis, and

  (f) spatio-temporal modelling paradigms capable of modelling the spread and control of Prosopis.

II To *construct* a spatial data set which is representative of the current distribution of Prosopis and its environmental requirements, based on the analysis techniques in Objective I(d).

III To *design* an appropriate prediction model capable of identifying and ranking, based on importance, the environmental features identified in Objective II. The model should extract the desirable habitat of Prosopis, using the algorithms identified in Objective I(e).

IV To *implement* the results from the predictive model designed in Objective III in a spatio-temporal model, inspired by approaches identified in Objective I(f), to simulate the potential spread and ecological impact of Prosopis.

V To *verify* and *validate* the model components outlined in Objectives II–IV, complying with generally accepted modelling guidelines researched in fulfilment of Objective I(e) and (f).

VI To *apply* the verified and validated model of Objective V to a specific case study, focused on the invasive spread of Prosopis in the Northern Cape.

VII To *evaluate* the ability of the model of Objective V to simulate the potential spread of Prosopis in the Northern Cape.

VIII To *reflect* on the project and recommend possible improvements and follow-up work which may be pursued in the future.

1.4 Project scope

Given the complexity and vast number of factors that influence the spread of invasive species, the scope of this project is limited by the following assumptions:

The species of interest. The models developed will consider all species of Prosopis in general, assuming similar environmental requirements and growth habits. Therefore, no distinction between the variants of Prosopis will be made, so as to simplify the modelling of its population growth against the use of different control strategies.

Selecion of control methods. Many methods exist for controlling invasive species, such as mechanical, chemical, and biological methods, as well as concentrated, controlled fires. Rather than employing a specific control method (or combination thereof) throughout the execution of this project, the focus is more geared towards investigating the effect that an effective control strategy may have on the spread of Prosopis over time.

The CA model considerations. The hexagonally discretised area considered in which control strategies are modelled is limited to a specific size, so as to ensure that the size of individual hexagonal cells are not too big as this could adversely affect the performance and accuracy of predicting the response of the specie to the control strategy. Furthermore, the system is assumed to be closed, that is, factors external to the study area are not considered.

1.5 Report organisation

As stated in §1.2, the primary objective of this project is to model the effect of control strategies on the spread of the invasive tree species, Prosopis. This requires the development of a verified and validated model that may be applied to a real-world case study in the Northern Cape. The organisation of the report provides an overview of the structure of the report in pursuit of fulfilling the objectives defined in §1.3.

The report comprises of five additional chapters. Chapter 2 contains an in-depth review of the literature which is of relevance to this project in fulfilment of Objective I.

Inspired by the literature reviewed in Chapter 2, Chapter 3 is devoted to deriving and developing the modelling components relating to the spread and control of Prosopis in a spatio-temporal context. These components include collection of the relevant data, and implementation of ML and CA models in fulfilment of Objectives II, III, and IV.

Chapter 4 seeks to fulfil Objective V and entails the verification and validation of the ML and CA models developed in Chapter 3.

In Chapter 5, the validated and verified model of Chapter 4 will be applied to a real-world case study in the Northern Cape in fulfilment of Objective VI. Furthermore, Chapter 5 will also contain an evaluation of the model applied to the Northern Cape case study in fulfilment of Objective VII.

The report concludes in Chapter 6, fulfilling Objective VIII, with a brief summary and appraisal of the entire project, reflections by the author in terms of what was learnt, as well as recommendations for possible improvements and follow-up work which may be pursued in the future.

1.6 Report timeline

The timeline of this project is provided in Gantt chart form in Appendix A.

CHAPTER 2

Literature Review

Contents

This chapter focuses on the literature pertaining to Prosopis, as well as the relevant modelling components required to successfully model the problem considered. First, the characteristics of Prosopis invasions are investigated in §2.1. Thereafter, the well-known control strategies which are typically employed to inhibit the spread of the species is discussed in §2.2. This is followed by an in-depth discussion on the relevant techniques used to model the problem at hand. In particular, §2.3 deals with the spatial analysis component, §2.4 focusses on the spatio-temporal component, and finally §2.5 details the considerations for the ML component. The chapter concludes in §2.6 with a brief summary of the aforementioned sections.

2.1 Characteristics of the invasive species Prosopis

Invasive plant species have severely threatened global biodiversities and human livelihoods in recent decades [73]. This is certainly the case for the plant species under the name Prosopis, where some of its species are considered to be among the world's most destructive invasive species [45]. Introductions of Prosopis across continents has transpired over many centuries. The earliest report of Prosopis being introduced outside of its origin, the Americas, was in 1822 in Senegal. As of 2014, Prosopis has been observed in approximately 131 countries, with widespread introductions occurring in Africa and Asia between the 1970s and 1990s [66]. The full timeline of Prosopis introductions from the year 1822 to the year2000 is graphically displayed in Figure 2.1.

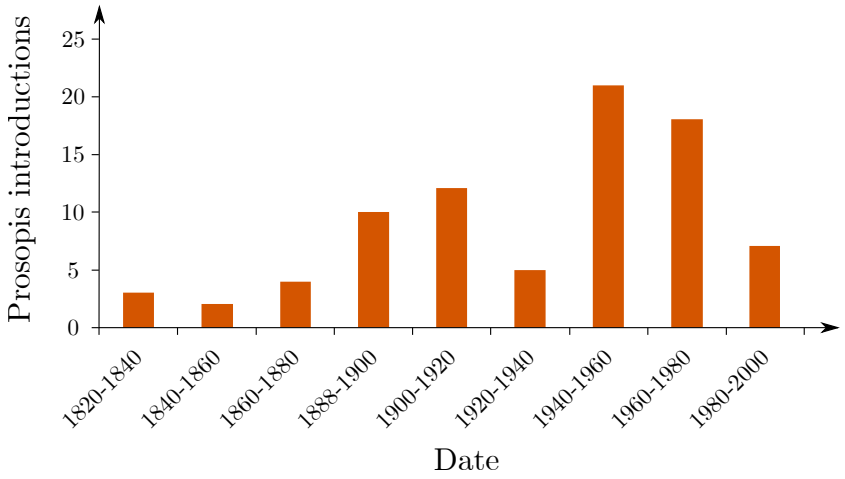


FIGURE 2.1: *Timeline of all Prosopis introductions globally, adapted from [45].*

The primary reason for the widespread introduction of Prosopis was to assist regions in reforestation programmes after experiencing severe droughts. As such, Prosopis was intended to serve as a source of shade and fodder in the case of South Africa and Australia, dune stabilisation and fuel-wood in Sudan, and natural fencing in Malawi [45]. While many countries have intentionally introduced Prosopis for reasons similar to the aforementioned, some inter-border introductions have accidentally happened. Introductions in Botswana, Nigeria, and Yemen to name but a few, were the result of livestock being traded with neighbouring countries [66]. This comes as a result of the livestock feeding on Prosopis as fodder and carrying the seeds into these neighbouring countries via their excrement.

2.1.1 Dynamics of invasive plant species

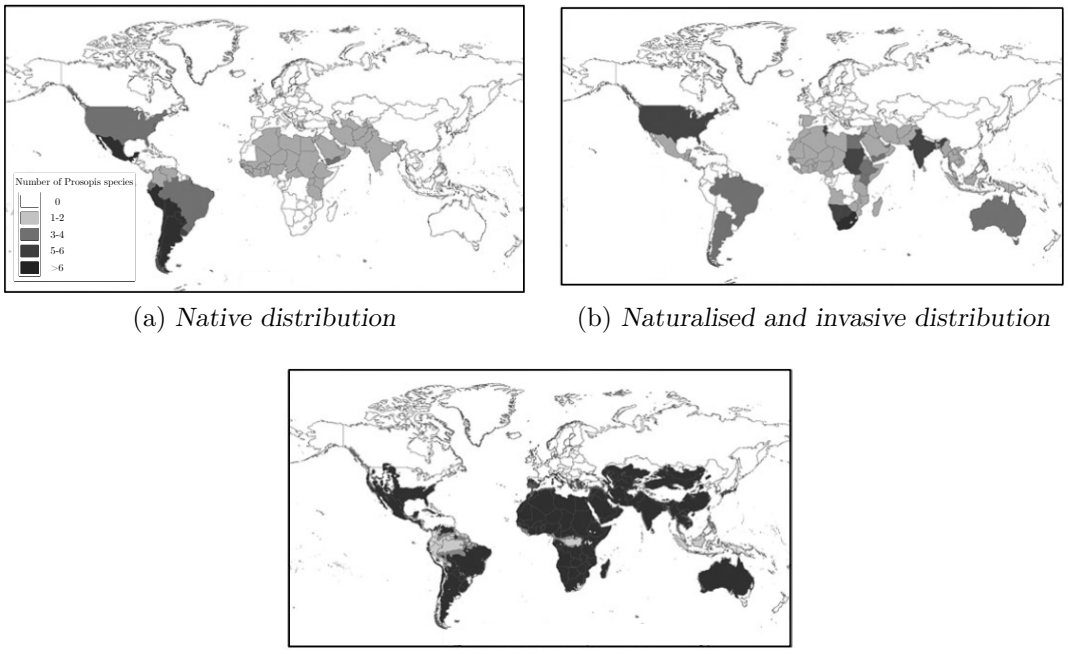
The dynamics of invasive species are vital to understand when attempting to minimise their negative effects, while maximising their benefits. There is unfortunately a shortage in frameworks connecting theory and management pertaining to biological invasions [45], and thus it is necessary to fully grasp the dynamics of the species being studied in order to develop sustainable management plans.

Habitat and distribution

The 44 variants of the Prosopis species is native to hot arid and semi-arid environments of the Americas, and have naturalised in the arid and semi-arid regions of the countries in which it has been introduced. In the south-western region of of the United States, the growth of Prosopis has been observed to be limited to altitudes below 1 676 m above sea level, with the preference of most members of the population being below 1 371 m. Prosopis is well-known in withstanding extreme climatic conditions such as high temperatures and low rainfall [66, 17]. When observed in the desert, Prosopis prefers to grow along drainage passages where rainfall does not exceed approximately 15 cm annually [17].

Prosopis is rarely limited by the condition of the soil, and so is able to thrive in alkaline, saline, or infertile soils [66]. As such, many researchers omit describing the soil of Prosopis' habitat given that it is considered to have adapted to all soil types, irrespective of the moisture level of the soil. Moreover, it is well known that Prosopis is able to grow in most environments regardless of how rocky, broken, flat, or sandy the environment may be. While Prosopis has proven to grow irrespective of its soil conditions, it does prefer regions with medium to fine textured soils [17].

The native, naturalised and invasive, as well as potential (based on climatic suitability) distributions of Prosopis species across 131 territories worldwide is provided in Figure 2.2. Figure 2.2(a) displays the regions in which Prosopis is naturally observed (*i.e.* its native regions). Figure 2.2(b) illustrates the distribution of introduced Prosopis species that have either naturalised or have become invasive. Furthermore, Figure 2.2(c) represents the potential distribution for Prosopis which was determined as a result of assessing climatically suitable regions for Prosopis which currently have no records of the species being present [45].



(a) *Native distribution* (b) *Naturalised and invasive distribution*

(c) *Potential distribution*

FIGURE 2.2: *The number of territories per country containing a native, naturalised, or invasive population distribution of Prosopis, adapted from [45].*

The number of territories per country containing a native, naturalised, or invasive population distribution of Prosopis is tabulated in Table 2.1. The most recent extensive global review completed in 2001 revealed that Prosopis was present in 93 countries and islands [66]. While it is unlikely that Prosopis has been introduced to many more territories since the review, the figure of 131 countries and islands may be due to data availability increasing since then, or that the species has unintentionally been spread due to cross-border livestock trading. Most importantly perhaps is the fact that 79% of introductions recorded have resulted in naturalisation, of which 38% have become invasive [45].

TABLE 2.1: *Global distribution of Prosopis by region, adapted from [45].*

Region	No. of territories containing Prosopis
Africa	40
Americas	19
Asia	26
Caribbean islands	18
Europe	4
Islands, and Australia	24

Spread and growth

Historical observations of Prosopis have estimated its annual rate of spread to range from 3.5–18% in South Africa, implying that the invaded region could double every five to eight years. The Prosopis tree typically grows to heights of between 12 and 20 metres, while some shrub-like species of Prosopis, including *laevigata* and *reptans*, reach only three metres in height. Regardless of the variants of Prosopis, its trunk is often short and crooked and grows to a diameter of approximately 65 cm [92]. Apart from the aforementioned robustness of Prosopis to grow in a wide range of conditions, the primary reasons for which Prosopis is a successful invader lies in their ability to produce a large amount of seeds that stay viable for decades, fast growth rates, early reproductive age, the ability to regrow after being cut, tap root systems reaching depths of 50 metres and deeper in order to utilise both surface and ground water, and *allelopathic*[1] effects [45]. As previously stated, one of the main reasons for any ongoing spread of Prosopis is due to unintentional spread through livestock. In particular, the seed pods that grow on Prosopis contain approximately 25 seeds per pod [92], and so when grazed upon by animals, the potential for germination via their excrement enhances the spread of the invader [81].

Impacts of invasive spread

Prosopis invasions have yielded both benefits and harms for local and global environments, economies, and human livelihoods. As such, its positive and negative impacts have often been contrasted, questioning whether or not the positive impact of the species can outweigh the negative [45]. The benefits resulting from Prosopis introductions is that the pods which grow on the tree serve as a source of fodder for livestock, the wood of the tree may be used as firewood, turned into charcoal, or used as timber products such as poles or boards. Furthermore, the small flowers on the tree serve as a source of forage for bees due to the large quantities of pollen and nectar produced by them [92]. In addition, the physical tree itself provides sources

[1] A biological phenomenon in which a plant hinders the germination of seeds of other plants, thereby discouraging other plant species to grow near to it [67].

of shade for animals during the warmer months [81]. Generally, the negative impacts caused by Prosopis invasions include the disturbance of native ecosystems, a compromised water-supply to the region, and damaged underground pipes and boreholes caused by their roots [46]. From a natural resources point of view, many global studies on the adverse effects of Prosopis invasions all found that Prosopis outcompetes native plant species for the available natural resources such as water and sunlight, thereby reducing the density, richness, and diversity of native species [80]. Moreover, the *allelochemicals* contained within the leaves of Prosopis are known to kill certain insects as well as inhibit the germination and growth of other trees [92]. The socio-economic impact of large scale invasions may lead to consequences of human health due to compromised ecosystem services, decreased profits for farmers resulting from lower crop yields, as well as increased costs for farmers and locals due to repairing damaged infrastructure [88].

2.1.2 Prosopis in South Africa

In the late 1800s, Prosopis was introduced to South Africa from the Americas and was widely distributed in the arid and semi-arid regions to be planted in order to serve as a source of shade and fodder for livestock during drought periods. A 1998 study revealed that 1.8 million hectares of South Africa is estimated to been invaded [80, 96]. As of 2004, Prosopis has been listed by the South African national government as a Category 1b invasive species in the Eastern Cape, Free State, North-West and Western Cape. Throughout the Northern Cape, however, Prosopis is listed as a Category 3 invasive species with the exception of riparian areas, where they are listed as Category 1b invaders [18].

The rate of spread, as previously stated, is estimated to be between 3.5 and 18% per annum. Between 2002 and 2007, the area occupied by Prosopis in the Northern Cape alone had increased by approximately one million hectares, which is equivalent to a spread rate of 27.5% per year. Most recently, it was found that between 2000 and 2015, the public works eradication programme had treated 203 000 hectares of area covered by Prosopis nationally and the cost of this project amounted to ZAR 1.8 billion at the time [92].

The choice of management strategies surrounding the control of Prosopis has been been a contentious issue for many years. Some advocate for control through utilisation (*i.e.* managing the invader in a way that seeks to minimise the negative impacts caused by the invader, while simultaneously benefiting from its positive impacts), however, many believe this approach is inefficient [99] and calls for more conventional methods of control. As such, the government managed WfW programme employs a combination of mechanical, chemical, and biological methods to control Prosopis. Three seed-feeding beetles, namely *Algarobius prosopis*, *Algarobius bottimeri*, and *Neltumius arizonensis* were employed as biological agents to inhibit the spread of Prosopis while ensuring that Prosopis was not harmed. These approaches have, however, proved to be unsuccessful in significantly affecting the spread of Prosopis due to the large-scale of the invasion [103]. Despite biological methods being considered to be the most cost-effective method, the return on investment for Prosopis in particular is low when compared with the strategy employed on other invasive species. The lack of success experienced thus far in controlling Prosopis is likely due to the poor strategy employed by management as well as the prioritisation of projects. As such, there is certainly a need for decision-makers to revise their management plan and prioritisation of projects in order to effectively use the available resources for controlling Prosopis more efficiently than it has been thus far [45].

2.2 Control strategies

A wide range of control methods exist for Prosopis and each method has its own set of strengths and weaknesses. As such, it is necessary to investigate the cost-benefit relationships for each method before a method is implemented in a study. Current strategies in South Africa have been considered to have failed and not reduce the extent of the invasion as a whole due to labour intensive methods being favoured as opposed to identifying a best-suited approach or combination of approaches [81].

2.2.1 Comparing control methods

In 2017, Shackleton *et al.* [81] developed a comprehensive national strategy for the control of Prosopis and evaluated various control methods according to three criteria, the cost to clear the species, the area cleared per day on average, and the number of people employed by each method. The results are tabulated in Table 2.2.

TABLE 2.2: *Control methods evaluation with respect to cost, time, and employment, adapted from [81].*

Method	Cost to clear (ZAR)	ha cleared/day	Employment
Mechanical	±5000–7000	0.33	11
Chemical	±1000	<1000	1–2
Biological	Several millions	-	Researchers, lab assistants

Mechanical methods entail removing the invasive species by hand or by using tools [59]. This can be achieved by cutting tree stumps or using heavy-duty machines such as bulldozers. Despite this method being the slowest for clearing Prosopis, the WfW programme employs this method as its standard approach in the case of South Africa as it fulfils the objectives of removing the invader and employing as many workers as possible. While mechanical methods are ideal for areas used for agricultural reasons, it is important to note that they are damaging to the environment and not effective for large-scale invasions [81].

Chemical methods entail spraying herbicide on the invasive species, and is typically done by plane. This method is the fastest and most cost effective, however, the environmental consequences of this method remain a grey area. Moreover, although spraying herbicide will require ground teams to be employed for following up, the employment required is very low [81], thus making it an unpopular choice in regions with mandates for high employment.

Biological methods entail the deployment of biological agents to prey on the invading species. The agents are released into the area with the aim to either hinder the spread, or increase the mortality rate of the invading species. While this control method ensures no human impact on the environment, the research costs associated with identifying the most effective agent are initially high [81], and much uncertainty exists over a long duration with regard to the success of the deployed agent and strategy [25].

2.3 Spatial analysis

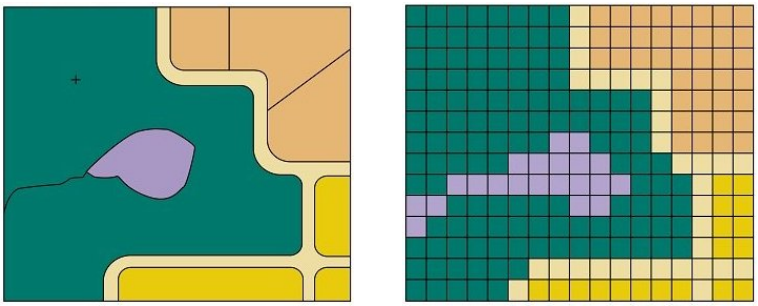
Spatial analysis is an integrated analysis technique used to solve problems geographically [87]. The analysis enables the user to identify and understand complex spatial patterns and relationships that exist within the environment. This is typically achieved through visualisation,

allowing users to make inferences about the system being studied [2]. Moreover, the data obtained from the analysis may serve as an input for further applications such as building an ML model. Many fields of research such as ecology, archaeology, geodesy and landscape architecture share a strong interest in spatial analysis, and so spatial analysis techniques have been applied in many studies within these fields [62]. Moreover, spatial analysis has become the fastest growing field of analyses within the study of ecology. Ecologists have recognised the importance of including spatial considerations into their ecological thought processes and decision making. Furthermore, the ease of access to software capable of performing spatial analyses has enabled users to build and maintain models that stay up to date with the rapid changes of the environment [24].

The primary reason for performing a spatial analysis in an ecological modelling setting is to study the *biotic*[2] and *abiotic*[3] relationships and influences [2] that exist within a region. A spatial analysis is only justified if *spatial dependence* exists in the study. Spatial dependence, also known as *spatial autocorrelation*, is the statistical perception that the space of a system is distributed non-stochastically. Similar to the well-known correlation coefficient, *Moran's coefficient* is a metric bounded in the range of $[-1, 1]$, and is employed to quantify the strength of spatial autocorrelation, measuring how similar or dissimilar entities are to one another. High positive coefficients indicate a strong spatial autocorrelation (clustering of similar values), a value of zero indicates no autocorrelation (complete randomness), and high negative coefficients indicate weak autocorrelation (patterns exist within a complex clustering) [98].

The spatial realm of the real world can be represented as either discrete or continuous data. A discrete representation stores data about attributes at an exact location, such as the landcover of an area. Continuous data are represented using grid-based systems in which each cell of the grid assumes a continuous value [87]. Environmental attributes of a location such as elevation, temperature, and precipitation are typically continuous data and are often represented as grid-based systems.

Two fundamental data models arise in the approach to represent and simplify data models when performing spatial analysis, namely the *vector model* and the *raster model*. Figure 2.3 illustrates the graphical comparison between the vector and raster model representations [79]. Raster data models represent the grid-based continuous components of the environment such as



(a) *Vector data representation* (b) *Raster data representation*

FIGURE 2.3: *Graphical comparison between vector and raster model representations, adapted from [79].*

precipitation, temperature, and humidity. The structure of a raster model is a matrix of rows and columns containing cells. The length and width of each cell is defined in surface units of the real world so that the square cells in Figure 2.3b may be defined as having sides of 100 metres

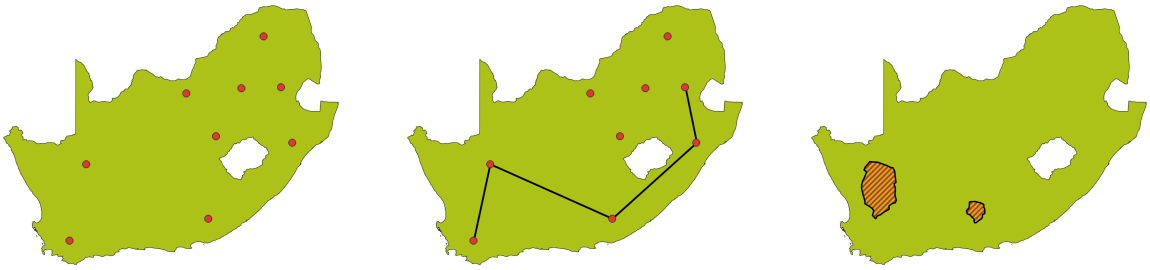
[2]Living components in an ecosystem such as animals, plants, and bacteria.

[3]Non-living components in an ecosystem such as water, soil, and atmosphere.

in length [90]. Furthermore, each cell within the matrix contains data about an environmental feature at that location [98].

Vector data models represent discrete geometric components of the environment such as transport routes, locations of trees, and networks of rivers [70]. In order to depict these geometric components, three elements are used individually or in combination with one another. These vector data elements are known as *point*, *line*, and *polygon* data types and are illustrated in Figure 2.4 [98]. Points are dimensionless and indicate the x, y coordinate location of a feature, such as a building. The nine points plotted in Figure 2.4(a) denote an arbitrary location in each province of South Africa. Lines connect points along a path and typically represent linear features and networks such as roads or rivers. Often, the line thickness of a road or river is dependent on the type of road or river. Figure 2.4(b) displays an arbitrary route between five provinces. Polygons, illustrated by a confined set of lines, represent defined areas and may assume any shape. The polygons illustrated in Figure 2.4(c) denotes two arbitrarily selected municipal regions in South Africa.

Choosing the appropriate data model depends largely on the operations required to be performed, the form of the data, the effect of the data model on data quality, and the skill level of the analyst [90]. Both conceptualisations have their own strengths and weaknesses. The raster model is simpler to implement, and less data-intensive, whereas the vector model is more versatile in its application and can represent discrete entities such as buildings more effectively [86].



(a) *Points* (b) *Lines* (c) *Polygons*

FIGURE 2.4: *Illustration of the three types of geographical data elements.*

2.3.1 Geographical Information Systems

The literature considers many quantitative tools and software packages focused on integrating spatial analysis and data visualisation. The combination of the most popular analysis tools into a single software toolbox has resulted in the development of the spatial analysis software environment known as a GIS. A GIS aids spatial analysts in processing new information, handling and storing data, as well as generating useful visualisations [62].

The GIS software environment facilitates the visual exploration, manipulation, and analysis of spatial data in order to better understand the spatial patterns that exist between the spatial data [4]. GIS automates the procedures required for performing a spatial analysis, thus making it a highly effective tool. However, if spatial dependence is not present within the system, then GIS would prove to be an irrelevant tool for that particular application [98]. GIS software typically employs the principal of *layering* in order to combine different sets of data in order to access these layers simultaneously and perform analyses. The concept of layering can be

understood as vertically 'stacking' spatial data set layers that represent geographical attributes of the Earth's surface [49] and is illustrated schematically in Figure 2.5 [87].

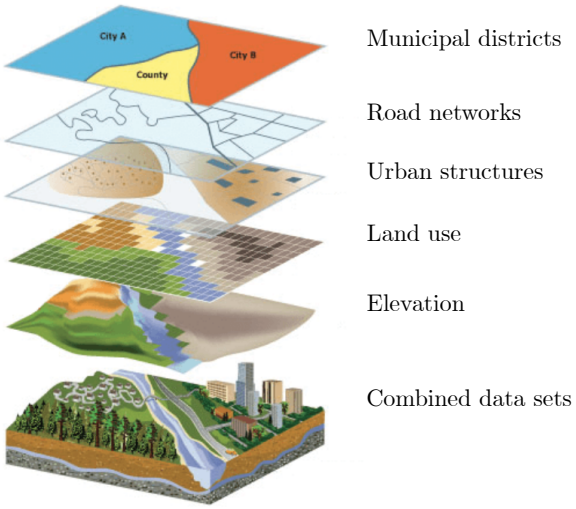


FIGURE 2.5: *Conceptual model of layering geographical attributes of the Earth's surface, adapted from [87].*

Map generation and *map comparison* are two widely used applications in GIS. Once the data have been obtained and added into the GIS software, many individual maps may be generated depending on which layers of data are selected to be included. A typical use of GIS involves mapping attributes of the natural environment with species activity and investigating any relationships that may exist [54]. Map comparison in GIS is often employed to determine the degree of similarity between maps and to study the change in natural phenomena over time, such as the growing and shrinking of glaciers in polar regions. Furthermore, map comparison is a fundamental consideration in the validation of many models. Most commonly, comparisons are carried out by considering either the composition (content) or the configuration (arrangement) of the attributes depicted on the maps. However, given the multifaceted nature of comparing maps, it is often unlikely that one single approach will perfectly represent the similarity between two or more maps [23].

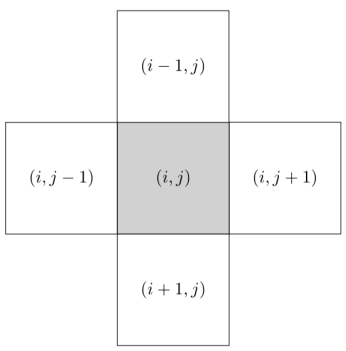
2.3.2 Cellular automata

Von Neumann [61] formally introduced the paradigm of CA modelling in 1966 after studying the behaviour of highly complex systems. Inspired by self-reproduction and evolution, CA models were conceptualised by Von Neumann as a result of considering the possibility of setting up a machine shop in which each machine is capable of replicating itself, given sufficient raw materials and time. In particular, if the design of a machine is defined by a pattern and a set of rules detailing the duplication of the pattern, then the machine can effectively deploy the set of rules to create a copy of itself.

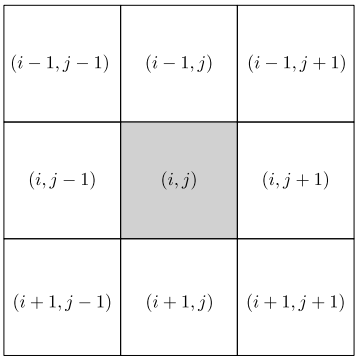
CA models exploit discrete, uniform grid structures, similar to that of raster-based models described in §2.3, and consist of four fundamental concepts, namely a *cell*, *state*, *neighbourhood*, and *transition rule* [16]. The spatial domain, or *cellular space* of the model is an n-dimensional lattice structure, and is discretised by the cells of the model which are typically square-shaped. Each cell in the discretised space contains a variable which specifies the state of the cell. Furthermore, the time-varying state of each cell depends on the state of itself, as well as the states

of all adjacent cells in the neighbourhood of that specific cell at time step t. Moreover, the states of each cell may be updated in discrete time steps or iterations according to a set of basic transition rules. CA models are capable of exhibiting highly complex, dynamic behaviours, even though the transition rules are simple in construction.

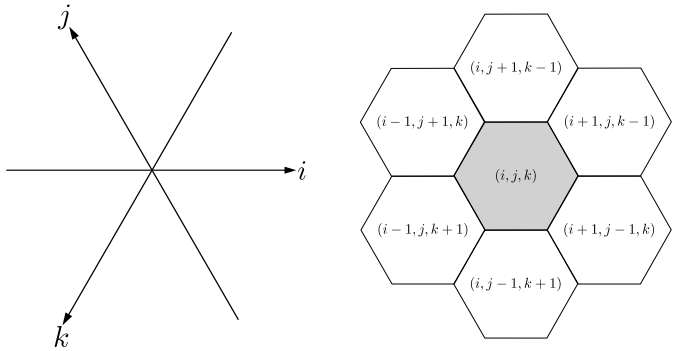
The neighbourhoods of individual cells may differ in geometric structure and size depending on the required application of the model. The two most common neighbourhood structures for square lattices are the *Von Neumann* and *Moore* neighbourhood configurations, visualised in Figure 2.6(a) and 2.6(b) respectively, where the shaded cell denotes the central cell. When an hexagonal lattice structure is employed by the CA model, the neighbourhood structure configuration visualised in Figure 2.6(c) is assigned to each cell in the grid [104]. The notable difference between the square and hexagonal neighbourhoods is that each square cell shares a common boundary with only four of the other cells in the two square neighbourhood structures, as opposed to six cells for the hexagonal neighbourhood structure. Furthermore, the centre-to-centre distances between all of the neighbours within a hexagonal structure is the same, unlike the square lattice structures.



(a) *Von Neumann neighbourhood*



(b) *Moore neighbourhood*



(c) *Hexagonal neighbourhood*

FIGURE 2.6: *Common neighbourhood structures employed in CA models.*

Considering the Moore neighbourhood structure comprising eight neighbours around the focal cell, the set of neighbours to cell c in row i, column j, denoted as $C_{(i,j)}$ can be written as

$$\Omega_{(i,j)} = C_{(i,j-1)}, C_{(i,j+1)}, C_{(i-1,j)}, C_{(i+1,j)}, C_{(i-1,j-1)}, C_{(i-1,j+1)}, C_{(i+1,j-1)}, C_{(i+1,j+1)}. \quad (2.1)$$

In 1997, Karafyllidis and Thanailakis [39] formulated a CA model for predicting the spread of fire in forests. By discretising the forest into a lattice of identical square cells and employing

the Moore neighbourhood structure, the state of each cell $S^t_{(i,j)}$ at discrete time intervals t represented the ratio of the burned area of the cell, A_B, to the total area of the cell, A_T, expressed mathematically as

$$S^t_{(i,j)} = \frac{A_B}{A_T}. \tag{2.2}$$

As a result, unburned cells will assume a state value of 0 and burning cells will assume a state value in the interval between 0 and 1, until a cell is fully burned out, in which case it will assume a value of 1. The transition rule defining the changing of a cell from its state at time step t to time step $t+1$ was given as a function f of the state of cell $C_{i,j}$, as well as the state of its neighbouring cells, as per (2.1), at time step t. Mathematically, the transition rule may be fully expressed as

$$S^{t+1}_{(i,j)} = f(S^t_{(i,j)}, S^t_{(i,j-1)}, S^t_{(i,j+1)}, S^t_{(i-1,j)}, S^t_{(i+1,j)}, S^t_{(i-1,j-1)}, S^t_{(i-1,j+1)}, S^t_{(i+1,j-1)}, S^t_{(i+1,j+1)}), \tag{2.3}$$

and by employing the notation of (2.1), the transition rule in (2.3) may be re-written as

$$S^{t+1}_{(i,j)} = f(S^t_{(i,j)}, S^t_{\Omega_{(i,j)}}). \tag{2.4}$$

CA models have become a popular technique in the field of ecology, and biological invasions in particular, and have shown to be capable of handling complex biological processes in spite of its simplicity [34]. Invasive species exhibit similar spread patterns when compared to other disturbance agents, like the aforementioned spread of fire [101]. As such, the presence or absence of a species at a specific location can be represented by discrete cell states that can change over time as a result of competition with native species, availability of resources, and employed control methods.

In 2016, Yoshimoto *et al.* [101] employed a CA model as part of an optimisation framework that captured the spatial dynamics of invasive species and returned the optimal control strategy that should be employed. Two states were considered for each cell within the CA model, namely *colonised* or *uncolonised*. A binary variable, $z_{(i,j)}(t) \in \{0,1\}$, may be introduced to formally define the aforementioned states at time t for the cell $C_{(i,j)}$. Consequently, if a cell is colonised, it assumes a state value of $z_{(i,j)}(t) = 1$, and if uncolonised, it assumes a state value of $z_{(i,j)}(t) = 0$. Once a cell is colonised, the invader will attempt to spread to uncolonised neighbouring cells, with a colonisation probability $P\{z_{(i,j)}(t) = 1\}$. By deriving a probability rule for the colonisation of a cell, the CA model is linearised, allowing for a $(0,1)$ integer programming problem to be evaluated, thereby assisting decision makers to either 'not implement any treatment' or 'implement a treatment.' The colonisation probability can be expressed as

$$P\{z_{(i,j)}(t) = 1\} = 1 - \gamma^{S_{(i,j)}(t)}, \tag{2.5}$$

where $\gamma \in (0,1)$ is a species-specific parameter, and $S_{(i,j)}(t)$ is the sum of invaders coming from colonised cells to cell $C_{(i,j)}$ at time t. Furthermore, colonisation of a cell only occurs when $S_{(i,j)}(t)$ exceeds a threshold value $\boldsymbol{p}$ when $P\{z_{(i,j)}(t) = 1\} \geq \boldsymbol{p}$.

GIS integrated with CA

Over the last two decades, the integration of GIS within the realm of CA modelling has been to the benefit of the CA modelling paradigm, making it a suitable choice for qualitative predictions [34]. By combining the data management and visualisation capabilities of GIS technology with the spatio-temporal capabilities of CA, the result is a powerful tool for data rich modelling. Furthermore, GIS is often employed commercially to display and manipulate the behaviours

exhibited by complex systems resulting from CA models. Studies concerned with modelling the spread of invasive species typically employ GIS as well as some dynamic modelling technique. However, these studies have focused more on a system that displays the results of a predetermined model than the development of the model itself. In the past, ecological models were mathematically-rooted and iteratively derived from approximate data that estimated or averaged values of unmeasurable parameters. The high degree of complexity and expensive computational cost hindered the development and exploration of spatio-temporal models in the context of species modelling for a long time [16]. In this regard, CA models became an ideal technique due to the ability of representing complex, spatially distributed, dynamic models using a relatively simple set of transition rules. Despite the fact that CA models can exhibit similar dynamics to partial differential equations, they use cell states, transition rules, and parameters to describe the model as opposed to non-spatial methods that employ approximate data. As a result, CA is a suitable dynamic modelling technique for GIS to be accompanied with due to its 'bottom-up' approach and conservation of data relating to individual components.

2.4 Mathematical modelling of population growth

Across all subdisciplines of ecology, population biology could be the most mathematically driven, and problems relating to population dynamics has long piqued the curiosity of scholars [72]. Interest into the study of population growth was formally introduced as early as 1798 by Thomas R. Malthus in a study titled *An essay on the principle of population* [50] which discussed population dynamics and the role it plays in affecting the betterment of society. Throughout history, a great concern has been that of the ability to sustain a growing population within existing environments and systems. Mathematical models serve as a suitable tool that may be employed to address the questions that arise from complex behaviours of populations [10] and functions as a guide for sustainable decision making by future generations.

It is well known that some form of stochastic behaviour is inherent in all biological populations [64]. Consequently, it is imperative that environmental noise (*e.g.* climate or natural disasters) is reflected in dynamic population growth models. Generally, dynamic population growth models consist of deterministic and stochastic components that work simultaneously. The deterministic component is responsible for ensuring that the target variable is predictable for a given set of initial conditions, while the stochastic component is mainly responsible for mirroring demographic dynamics (*e.g.* factors affecting reproduction or mortality) as well as the aforementioned environmental noise.

Historical approaches towards modelling population growth were rooted in the assumption that under positive population growth rates, a population increases exponentially in what is known as a population *explosion*. Moreover, the degree of the explosion is determined by the resource limitations of the environment. The validity of such models are, however, naturally limited due to the fact that the exhibits a *carrying capacity*[4]. As such, population growth models cannot be developed over a temporal dimension, but instead as a function of both spatial and temporal dimensions [10]. Various equations have been derived to describe population dynamics and the choice thereof depends on the species being studied, as well as the aim of the study [26].

[4]The maximum population size that a specific environment can support given the available natural resources in that environment [10].

2.4.1 Modelling population growth over time

Much of the literature pertaining to modelling population growth focusses on the population growth of a species over time, disregarding the movement of species populations in the spatial dimension. It is important to note, however, that simple models do not consider the stochastic elements discussed in §2.4. Despite this limiting the usefulness of these models, they still deliver fundamental insights to aid our understanding of complex processes and serve as a natural point of departure in the study of growth models for biological populations [10].

Malthusian growth model

The exponential growth model, commonly referred to as the *Malthusian growth model*, was developed by Malthus between 1798 and 1826 and is widely accepted as one of the most influential works on population dynamics. Malthus theorised his model on the groundings of the disparity between a biological population and resource production, stating that population members would increase at an exponential rate, while production of resources increases arithmetically [50]. For a population size of $N(t)$ at time t, growing at a constant rate r, the Malthusian population growth model may be expressed mathematically by the initial value problem

$$\left.\begin{aligned} \tfrac{\mathrm{d}N}{\mathrm{d}t}(t) &= rN(t), \quad t > 0 \\ N(0) &= \alpha. \end{aligned}\right\} \tag{2.6}$$

The solution to the differential equation in (2.6) may be determined by employing the separation of variables technique and evaluating the integral

$$\int_{\alpha}^{N} \frac{\mathrm{d}N'}{N'} = r \int_{0}^{t} \mathrm{d}t',$$

at the initial condition $N(0) = \alpha$ at time $t = 0$, which yields

$$\log \left| \frac{N}{\alpha} \right| = rt,$$

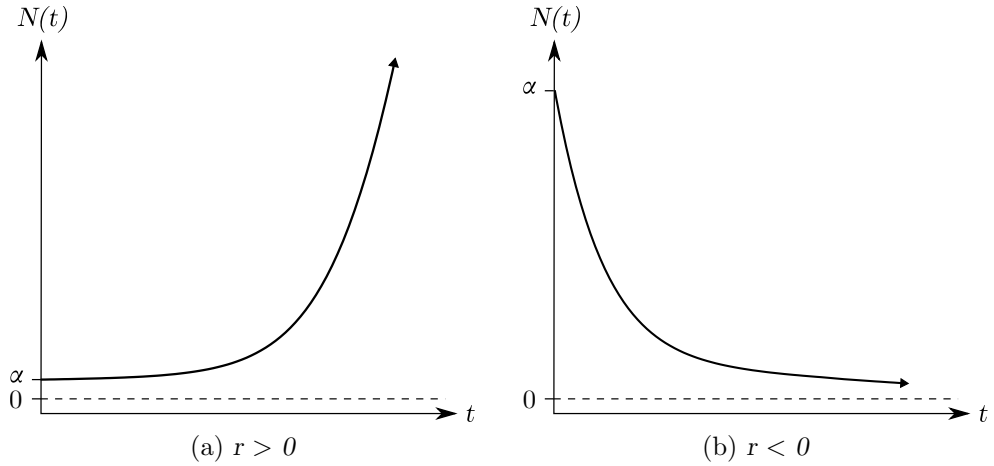
and can be expressed more succinctly for $t \geq 0$ as

$$N(t) = \alpha e^{rt}. \tag{2.7}$$

The assumption of a constant population growth rate which is proportional to its size leads to the model in (2.7) and predicts population explosion if $r > 0$ as illustrated by Figure 2.7(a), extinction if $r < 0$ as illustrated by Figure 2.7(b), and an unchanged population α if $r = 0$ [10]. Furthermore, if $\alpha = 0$, the system will exhibit a steady state solution of $N(t) = 0$ for all $t > 0$. The modern interpretation of the model described in (2.6) is that it is only viable for small populations over a short period of time in an idealised environment containing infinite resources and the absence of competing species [33].

Logistic growth model

If the growth rate of a population is proportional to its size, as is the case with the Malthusian model, then the population size would expand exponentially, without being bounded to some limit. In reality, as the population size increases, the demand for resources in the environment will naturally increase as well. The *logistic growth* model accommodates this adjustment from the



(a) $r > 0$ (b) $r < 0$

FIGURE 2.7: *The Malthusian model illustrated for a population explosion and a population extinction.*

Malthusian model by incorporating an asymptotic self-regulating mechanism (*i.e.* the carrying capacity), which is induced when a population grows too large, thereby taking into account the phenomenon of *population overcrowding* [64]. As such, the model represents the population growth rate, previously r, as a function of the species birth rate b and its death rate d resulting from overcrowding. The logistic growth model may be expressed mathematically as

$$\left.\begin{aligned} \tfrac{\mathrm{d}N}{\mathrm{d}t}(t) &= N(t)(b - dN(t)), \quad t > 0 \\ N(0) &= \alpha. \end{aligned}\right\} \tag{2.8}$$

The solution to the differential equation in (2.8) may be determined by following a similar approach to solving (2.6), by evaluating

$$\int_{\alpha}^{N} \frac{\mathrm{d}N}{N(b - dN)} = \int_{0}^{t} \mathrm{d}t,$$

which requires the left-hand side integral to be expanded by applying the method of partial fraction decomposition. This yields the integral equation

$$\frac{1}{b}\int_{\alpha}^{N} \frac{\mathrm{d}N}{N} - \frac{1}{b}\int_{\alpha}^{N} \frac{-d\ \mathrm{d}N}{b - dN} = \int_{0}^{t} \mathrm{d}t. \tag{2.9}$$

Solving for the equation in (2.9) at the initial condition of (2.8) yields the solution

$$\log \frac{N(b - d\alpha)}{\alpha(b - dN)} = bt,$$

which after exponentiating and making N(t) the subject gives

$$N(t) = \frac{\alpha b}{\alpha d + (b - d\alpha)e^{-bt}}.$$

The relationship between the species birth and death rate can be expressed as the ratio $K = \frac{b}{d}$, otherwise known as the carrying capacity of the species. By incorporating the carrying capacity K into (2.9), the population size at time t for $t \geq 0$ resulting from the logistic growth model may be expressed as

$$N(t) = \frac{K\alpha}{\alpha + (K - \alpha)\, e^{-bt}}. \tag{2.10}$$

The expression in (2.10) necessitates the elucidation of four possible behaviours [64] that may arise:

1) for $0 < \alpha < K$, the population size increases asymptotically to K according to the logistic function, and $\frac{\mathrm{d}N}{\mathrm{d}t}(t) > 0$,

2) for $\alpha > K$, the population size decreases asymptotically to K, and $\frac{\mathrm{d}N}{\mathrm{d}t}(t) < 0$,

3) for $\alpha = K$, the population remains constant at a size of $N(t) = K$ for $t \geq 0$, and $\frac{\mathrm{d}N}{\mathrm{d}t}(t) = 0$, and

4) for $\alpha = 0$, the population remains constant at a size of $N(t) = 0$ for $t \geq 0$, and $\frac{\mathrm{d}N}{\mathrm{d}t}(t) = 0$.

The solution in 2.10 is displayed in Figure 2.8 and illustrates behaviours (i) and (ii). From Figure 2.8, it is evident that an initial population size significantly less than K, yet greater than zero (*i.e.* α_1), may exhibit exponential growth in the beginning due to the abundance of resources. As the population size asymptotically approaches its carrying capacity K, however, the growth rate begins to decrease as of the point of inflection. The growth rate continues to decrease as a result of resource scarcity until it has reached a value of zero. In the theoretical case of a population size exceeding K (*i.e.* α_2), the population experiences a strictly negative growth rate due to overcrowding and competition for resources, thereby resulting in a decline of the population size until the carrying capacity K is reached [64]. The logistic growth model, unlike the Malthusian model, may exhibit two steady states and occurs in the cases where the initial population size is either $\alpha = 0$ or $\alpha = K$. Trivially, the first steady state remains in a population size of $N(t) \equiv 0$, whereas the second maintains a population size of $N(t) \equiv \frac{b}{d}$.

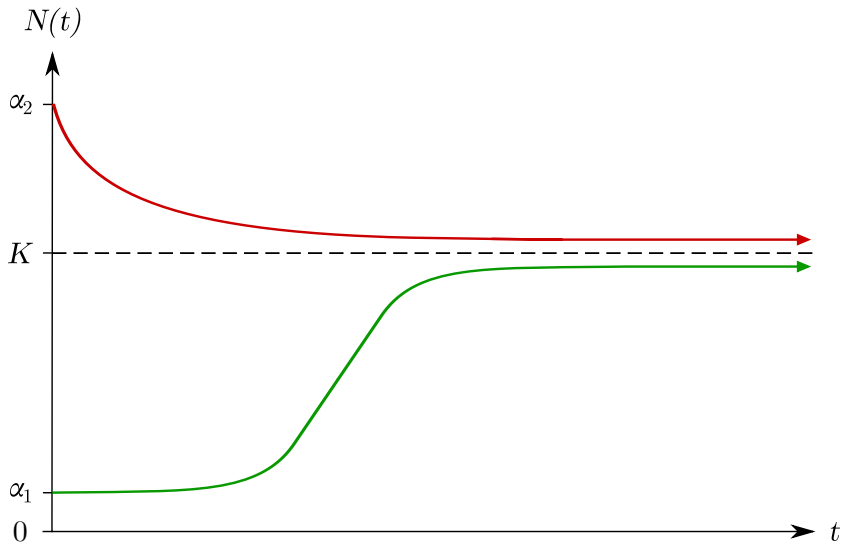


FIGURE 2.8: *Logistic growth model displaying initial populations α_1 and α_2 as well as their corresponding behaviours, (1) and (2) respectively.*

2.4.2 Modelling population growth over space and time

The *mean-field approximation* [60] is an idealised assumption employed by a host of population growth models and it assumes that the landscape of an environment, as well as the population density of the individuals contained within that environment, is homogeneous. It follows that

the probability of interaction between a randomly selected individual and any other individual belonging to the same, or a different species population, is independent of the individual selected. As a result, individuals have an equal probability of interaction, irrespective of the distance separating them. When spatial models are developed under this assumption, they are said to predict the dynamics — with respect to the size of the population — of the global model by upscaling the local interactions between individuals to the scale of the entire environment. In doing so, these mathematical representations unrealistically ignore the terms governing the dispersal of a population, resulting in comparatively poor model results. Consequently, models should ensure that they accurately represent population growth by allowing for spatially varying population densities.

The earliest mathematical representations attempting to model reaction-diffusion equations for biological populations were pursued in 1951 by John G. Skellam [83]. The model proposed by Skellam was based on the spread of muskrat populations in Europe according to Malthusian growth over time, but stating that an area occupied by an invading species will linearly also increase in size over time. This model has since been adopted to successfully model the spread of other biological species, such as the grey squirrel and the Californian sea otter [32].

The diffusion component of a reaction-diffusion model is derived from Fick's first law, which describes the flux of a substance's diffusing particles [13]. In 1855 Fick published an article titled *On liquid diffusion* [22] in which this approach of diffusion was mathematically formulated, analogous to the diffusion of heat through a conducting medium. In the case of a one-dimensional medium, Fick stated that the diffusion of particles occurs in the direction from a high density to a lower density. As such, the movement of particles between two spatial points per unit of time is directly proportional to the spatial derivative of the particle density, and inversely proportional to the distance between the two spatial points [22]. The flux of a population N at position x and at time $t \geq 0$, $F(N, x, t)$, can thus be given by

$$F(N, x, t) = D\frac{\partial N}{\partial x}(x, t), \tag{2.11}$$

where D is the rate of diffusion experienced by the population. In order to combine the logistic growth function and the diffusion of species population members into a single reaction-diffusion system, however, the population diffusion represented by the flux in (2.11) should be presented as the spatial change in the population density over time. This is achieved by finding the partial derivative of (2.11) over time as

$$\frac{\partial F}{\partial t}(N, x, t) = D\frac{\partial^2 N}{\partial x^2}(x, t). \tag{2.12}$$

The diffusion in a single spatial dimension of x may be exemplified by considering the arbitrary interval $I \in [v, w]$ defining the opposite boundaries of a spatial region on x, where $w > v$. The change in population density on x then equates to the difference in flux of the populations at the opposite boundary points v and w, expressed mathematically as

$$F(N, w, t) - F(N, v, t) = D\int_v^w \frac{\partial^2 N}{\partial x^2}(x, t)\mathrm{d}x. \tag{2.13}$$

The final reaction-diffusion model aims to find the change in population size over the temporal and spatial dimension by combining the logistic growth model and the partial derivative in (2.12) into a single expression, yielding

$$\frac{\partial N}{\partial t}(x, t) = N(x, t)(b - dN(x, t)) + D\frac{\partial^2 N}{\partial x^2}, \tag{2.14}$$

which may be solved over a given time period and for the interval I, provided some initial population size $N(x,0) = f(x)$ at time $t = 0$, where $x \in [v, w]$.

In the context of discrete modelling of population growth, the *finite difference method* [85] is often employed in order to approximate derivative values over a discretised spatial domain. As such, the finite difference method requires a discretisation procedure for the spatio-temporal domain $[v, w] \times [0, T]$ where $T > 0$ is a constant, on a set of equidistant grid points

$$x_i = i\Delta x, \quad i = 0, \ldots, \frac{(w-v)}{\Delta x}$$

and

$$t_m = m\Delta t, \quad m = 0, \ldots, \frac{T}{\Delta t}$$

for the selected spatial and temporal step lengths Δx and Δt, respectively. In the discretised spatial and temporal domains, the finite difference method may be employed to discretise the second-order derivative of the reaction-diffusion equation in (2.14) at the grid point (x_i, t_m) [82] as

$$\frac{N(x_i, t_{m+1}) - N(x_i, t_m)}{\Delta t} = D\frac{N(x_{i+1}, t_m) - 2N(x_i, t_m) + N(x_{i-1}, t_m)}{\Delta x^2}. \tag{2.15}$$

As such, solving for $N(x_i, t_{m+1})$ yields

$$N(x_i, t_{m+1}) = N(x_i, t_m) + Z\left(N(x_{i+1}, t_m) - 2N(x_i, t_m) + N(x_{i-1}, t_m)\right), \tag{2.16}$$

where Z is a dimensionless parameter representing the species diffusion rate D from equation (2.15), as well as the step lengths Δx and Δt, into a single parameter. As such, the properties of the finite difference method depends greatly on the parameter Z.

2.5 Machine learning

ML models seek to extract knowledge from data [56], allowing for the prediction of new data. ML paradigms are burgeoning in the field of conservation planning and, in particular, the modelling and prediction of static species distributions [19]. This can be attributed to the fact that ML models exhibit high accuracy and are not as constrained as many traditional, parametric species distribution modelling techniques, employing mathematical functions to predict distributions [6].

The use of prediction models is twofold: First, these models may be employed in imputing or interpolating areas for which species data has not yet been recorded, or does not exist. This is achieved by evaluating regions for which species data exists, and predicting theoretical species distributions in areas exhibiting similar environmental conditions over the same time period. Second, these models may be utilised by predictive models in order to forecast species distributions for future time periods by employing past and current distribution data [55]. Both aforementioned applications are achieved by extracting the species-environment relationships of variables in historical species and environmental data [38].

The mechanism of extracting knowledge, and as a result predicting outputs [6] is illustrated graphically in Figure 2.9. The process begins by obtaining a relevant pre-existing data set, or building one's own data set, using the tools reviewed in §2.3.1. The data is then *cleaned* and *pre-processed* so as to obtain a transformed data set from the original data set. Thereafter, the ML algorithm accepts the transformed data as input. Feature relationships and patterns are extracted during the *training* stage of the ML implementation and predictions are made based on the extracted patterns. Finally, once the output of the ML model is *interpreted* and *validated*

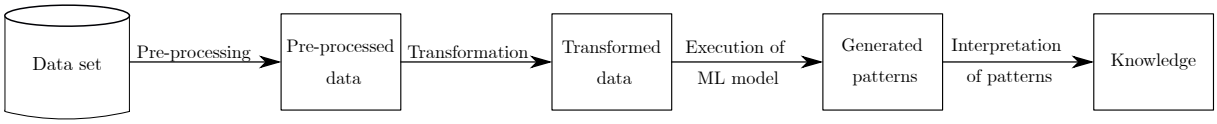


FIGURE 2.9: *Steps involved in the ML process, adapted from [6].*

by a human expert, it may be employed as a predictive tool for obtaining knowledge and insight for stakeholders, thereby aiding the decision-making process.

It is important to note that no single ML algorithm is universally better than another in the prediction of species distribution [55], and the success (or failure) of a model is highly dependent on the particular application of the study. As a result, the behaviour of the ML algorithm selected should be thoroughly understood prior to implementation. Moreover, the powerful paradigm of ML is not intended to replace humans as the domain expert, but rather to act as a tool to aid in the prediction, explanation, and interpretation of species distribution and other environmental phenomena [6].

There are four dominant paradigms of ML, namely *supervised*, *unsupervised*, *semi-supervised*, and *reinforcement* learning. Supervised learning works on the principle of extracting the underlying relationships between descriptive features and a target variable in order to make accurate predictions on unseen data [12]. The algorithm compares the output of the model with the actual or expected results in order to improve the model, thereby reducing the prediction error [78]. Conversely, data sets employed in unsupervised learning instances contain unlabelled instances and these models seek to extract the hidden structure of the underlying patterns in the data [1]. Data sets employed in semi-supervised learning instances contain both labelled and unlabelled instances, however, the data set typically consists of more unlabelled instances. The aim of this learning method is the same as supervised learning and the inclusion of the few labelled instances is to assist the unlabelled instances in training. Finally, reinforcement learning assesses the actions of agents in an environment and offers different rewards for different actions. The method aims to learn a *policy*, that is, a function that accepts a feature vector as input and outputs the optimal action for that instance. An action is termed optimal if it maximises the expected reward.

2.5.1 The supervised learning paradigm

Suppose an input variable x and a corresponding output variable y exists, then a supervised ML algorithm will attempt to learn a functional mapping, $f(x)$, from input x to output y. Two classes of supervised ML algorithms exist, namely *regression* and *classification*. Regression algorithms map the input variable x to a *continuous*, real-valued output variable y. An example of a regression problem is predicting the price of a house, given input variables, such as the 'number of rooms,' 'size,' and 'neighbourhood safety.' Classification algorithms, however, map the input variable x to a *categorical* output variable y. An example of a classification problem is predicting whether or not a client will default on their loan, since the client can be classified as belonging to one of the two classes: *Default* or *Non-default*. Supervised learning focused on classifying the occurrence of a species is typically employed in species distribution modelling [27]. The algorithm is trained using environmental data of a study region as input (*e.g. Mean Temperature, Annual precipitation, and Distance to water*), together with a binary encoded target variable indicating whether a species is *Present* (1) or *Absent* (0), as exemplified in Table 2.3.

TABLE 2.3: *Binary encoding of the target variable for classification in species distribution modelling.*

ID	Mean temp	Annual prec.	Dist to water	⋯	Presence
1	23	49	3 300	⋯	1
2	24	51	3 000	⋯	1
3	19	46	800	⋯	0
4	21	48	2 100	⋯	1
⋮	⋮	⋮	⋮	⋱	⋮

2.5.2 Preprocessing

The representation and quality of the data presented to an ML algorithm is an indispensable consideration that needs to be made as it directly affects the performance and accuracy of the model's predictions [43]. If the data presented to the ML algorithm exhibits quality issues, then the model is likely to struggle with knowledge discovery during the training phase of the ML process. As such, the step of preprocessing data seeks to clean and manipulate the data so that it is in the best state for the selected ML algorithm. Data quality issues may arise either due to invalid, or erroneously recorded instances within the data, or they may arise due to valid data which is incorrectly formatted that will cause an ML algorithm to struggle [41]. Some of the common data quality issues include missing values of features, imbalanced target classes, varying scales within the data, outliers, and irregular cardinality. For the sake of brevity, only the data quality issues relevant to the data for this project and the treatment thereof will be considered further.

Missing values

Incomplete data sets are inevitable in real world problems, and result from information being unavailable for an instance of one or more features. Due to the unique nature of each data set, the treatment of missing values should be carefully considered by evaluating the causes of missing information, which may be summarised as follows: (1) An entry is missing due to human error in capturing the data, (2) the value of a specific feature is missing because there is no information for that instance, or (3) for a specific instance, the value of a feature is not of concern or relevancy to the expert and is conveniently marked as missing [43]. Kelleher [41] proposes that features containing missing values in excess of 60% should be omitted from the data. Another approach of handling instances containing missing values is to select a suitable *imputation* method capable of replacing missing values with appropriate estimated values. Common imputation techniques include substituting the missing values with the mean or median of the available instances for continuous features, and the modal class for categorical features. Suitable feature values may also be calculated by considering the values associated with other features describing the entries in the data set.

Imbalanced data

An imbalanced data set consists of significantly more instances of one target class, than any other, leading to misrepresentation of the classes for the ML model to learn from [41]. As a result, the imbalance will cause some ML algorithms to be biased towards one of the classes. Contributing to this bias is a defect known as *overfitting*, which occurs when the algorithm induces a model that is too complex, thereby fitting the data too closely due to the presence of

under-represented classes exhibiting noisy patterns. Consequently, the model performs well on the training data, but poorly on unseen data [43].

While several techniques exist as solutions for treating imbalanced data, the most common approaches include the methods of *oversampling*, *undersampling*, and *synthetic minority oversampling technique* (SMOTE). Oversampling entails resampling or duplicating instances from the under-represented class in order to increase its significance in the data set. Conversely, undersampling involves randomly removing instances from the over-represented class in order to reduce its dominance in the data set. However, since data instances are removed in this technique, it may result in the model *underfitting*, that is, the algorithm induces a model that is too simple to learn the complex mappings between the descriptive and target variable [41]. Finally, SMOTE balances the data by synthesising new instances of the under-represented class, in addition to the existing instances. Suppose a set $\boldsymbol{A}$ contains all instances of the minority class, then for each instance $x_i \in \boldsymbol{A}$, k similar data entries are evaluated and stored in the subset set $\boldsymbol{S}_k$. A new instance, x_{new} is then synthesised from the subset $\boldsymbol{S}_k$ according to: $x_{new} = x_i + \lambda(x_k - x_i)$, where x_k is a randomly sampled instance from the subset, and λ is a random number in the interval $[0, 1]$ [12].

2.5.3 Supervised ML algorithms

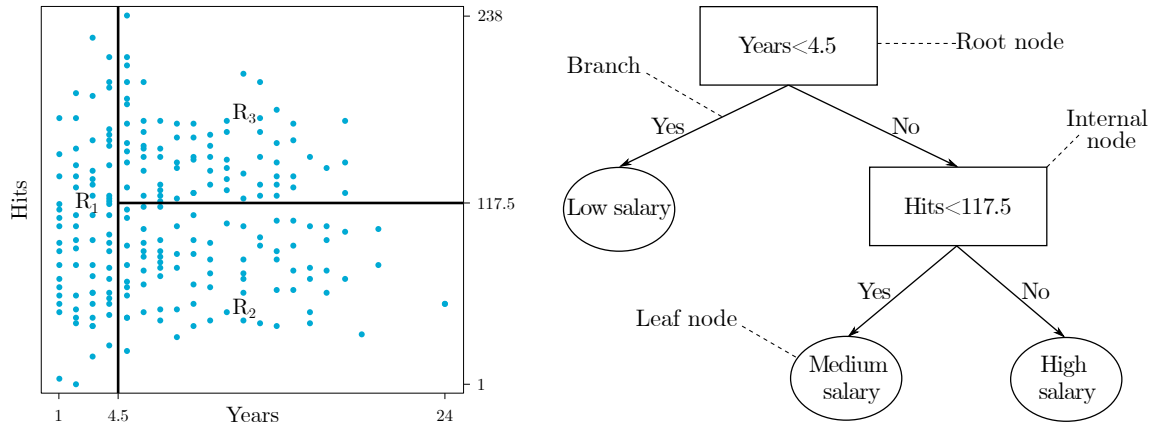
Many ML algorithms employed in the the field of ecological modelling fall within the paradigm of supervised ML. This is largely due to the fact that most data sets comprising ecological data contain some variable representing the presence or population size of a biological species and is often employed as a target variable alongside numerous environmental descriptive features. With this in mind, the remainder of this section provides an in-depth review of four supervised ML algorithms which are often employed within the field of ecological modelling, namely the *decision tree*, *random forest*, *k–nearest neighbours* (KNN), and *logistic regression* algorithms.

Decision trees algorithm

Tree-inspired models have been fondly utilised by researchers in many fields as a means of expressing knowledge and guiding decision making. In the 1980s, decision tree algorithms, commonly referred to as *Classification and Regression Trees* (CART), became popularised for the first time in the field of ML [95]. CART models are widely employed in solving both classification and regression problems due to their ease of interpretability, robustness to noisy data patterns, and relatively low computational cost [57]. Decision trees follow a hierarchical structure, consisting of *internal nodes*, *branches*, and *leaves* or *terminal nodes*. Following the logic of 'if-else-then' conditions, the tree starts at the top-most internal node, known as the *root node*, and branches out to one or more internal node(s). Each internal node represents a condition associated with evaluating a feature variable in the data set so as to split the data into its different target classes. The branches stemming from the internal nodes represent the different outcomes resulting from the internal node's conditional test and follows a path to the leaf node which represents the corresponding predicted classification label [5].

The workings of a decision tree algorithm is illustrated by considering a classification problem using the *Hitters* data set[5]. The problem is concerned with predicting a baseball player's salary, given some information about the number of years he has played in the major leagues, as well as the number of hits he made in the previous year [36]. Historical data is plotted in Figure 2.10(a), where the two feature variables, *Years* and *Hits*, are denoted as x_1 and x_2, respectively.

[5] A data set containing records and salaries for baseball players, available in the ISLR library, adapted from [36].



(a) *Two dimensional feature space partitioned into three regions*

(b) *Classification tree induced from the data set*

FIGURE 2.10: *Relating the salary of baseball players' to their years played and hits made. The partitioning of the feature space is shown in (a), and the resulting classification tree is shown in (b), adapted from [36].*

The decision tree algorithm partitions all of the observations in the training data set into several regions, using decision boundaries that are strictly parallel to the axes in a process known as *recursive binary splitting*. The process entails starting at the root node (*Years*) and successively splitting the feature space, indicated by two new branches on the tree, illustrated in Figure 2.10(b). The partitioned regions formed on the plotted data as a result of splitting the feature space can be denoted by R_i, where represents the number of leaf nodes on the tree. R_1 represents the region in which instances are predicted to have the class label *Low salary*, R_2 represents players predicted to earn a *Medium salary*, and R_3 represents players predicted to earn a *High salary*. In the case of classification trees, the algorithm predicts that a test instance belongs to the modal class of the region within which the test instance is partitioned. For regression trees, a test instance is predicted as being the mean value of all the training instances belonging to the same region as the test instance [36]. The overall structure of the example in Figure 2.10 can thus be understood as: $R_1 = \{\text{Low salary} \mid \text{Years} < 4.5\}$, $R_2 = \{\text{Medium salary} \mid \text{Years} \geq 4.5, \text{Hits} \leq 117.5\}$, and $R_3 = \{\text{High salary} \mid \text{Years} \geq 4.5, \text{Hits} \geq 117.5\}$.

A fundamental consideration for the analyst lies in deciding how the nodes of a decision tree are split. This is achieved by evaluating the *entropy*, a measure of homogeneity, uncertainty or randomness, with respect to each stratified region of the data [76]. If all instances within a region of the data set share the same class label, the entropy of the data in that region is 0, indicating no randomness in the data of that specific region. Conversely, if an equal frequency of different class labels exist within a region of the data set, the entropy of the data in that region is 1, indicating the highest degree of randomness possible in the data of that specific region [36]. The concept of entropy can be visually understood by Figure 2.11. Each container exhibits a different degree of entropy due to the composition of two classes denoted by circles and squares in each container being different. The first container contains a low entropy because only one class exists in the container, resulting in no uncertainty of what the contents of the container is. The last container exhibits the highest possible entropy, which is equal to 1 due to the equal frequency of the present classes in the container [58].

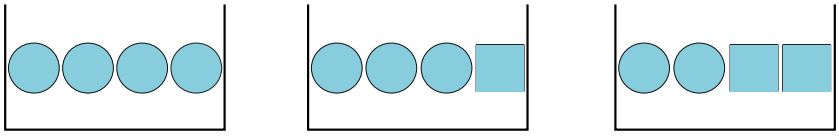


FIGURE 2.11: *Visual interpretation of entropy, adapted from [58].*

The entropy of a data set D may be expressed as

$$H(D) = -\sum_{i=1}^{I} p(y_i) \log_2 p(y_i), \tag{2.17}$$

where I denotes the number of classes present in D, and $p(y_i)$ denotes the proportion of instances within D that class i occurs. The algorithm will seek to split on the variable whose subsets exhibit the lowest possible entropy. This is because subsets exhibiting lower entropy values contribute more information about the variable, making them more desirable for the algorithm. As such, a statistical theory known as *information gain* is evaluated as the minimisation of the average entropy in 2.17 across all partitioned regions R [42]. Information gain can be expressed as

$$Gain(D, R) = H(D) - \sum_{o \in R} \frac{|D_o|}{|D|} H(D_o), \tag{2.18}$$

where D_o denotes the subset of instances corresponding to outcome o, after the split has been induced on a variable. Finally, after comparing all possible splits at a particular node, the split resulting in a maximised information gain is selected in order to grow the decision tree further [93].

Random forest

Ensemble learning methods have gained much interest due to their considerable improvement in classification accuracy [11]. This comes as result of combining a host of separately developed prediction models and aggregating their resulting predictions. Two celebrated ensemble learning methods relevant to tree-based models are *boosting* and *bagging*. Boosting methods assign additional weight to successive trees exhibiting a poor prediction accuracy. By combining the poor predictions made by the trees, a weighted *vote* or average is calculated across all of the trees and is used for prediction [48]. Bagging methods also entail combining trees, but ensures that successive trees are independent of earlier trees. Furthermore, the trees are constructed through a process of randomly sampling from the training data set with replacement, commonly known as *bootstrap sampling*. Once all of the trees are developed, a standard majority vote is used as a prediction [93].

Breiman [11] introduced the random forest algorithm to improve upon the method of bagging. The algorithm grows an ensemble of tree-based prediction models in such a way that each tree in the forest is subject to independent random sampling and is identically sized by sampling the data set with replacement. In the aforementioned stand-alone decision trees, nodes of the tree were *greedily* split, meaning that the split yielding the best information gain across all variables was selected. This technique, however, often results in a poor prediction accuracy of the model due to the high *predictive variance* that exists when the model is tested using other data sets, making it a potentially unreliable model [93]. Random forests, on the other hand, splits each node using the best split among a subset of variables which are randomly selected

at each node [48]. In doing so, the ensemble of trees improves upon the standard tree model by *decorrelating* it, thereby reducing the average variance of the results yielded by the trees and producing a more reliable predictive model [36].

k–Nearest neighbours

The KNN algorithm is a similarity-based learning technique which is considered to be simple and effective for classifying test instances [28]. The aim of the algorithm is to correctly predict the class of a test instance q by evaluating the distance between q and its k-nearest neighbours, where k is a specified non-zero integer value [14]. In the case of regression problems, the algorithm will classify the test instance with the average values of all training instances within the neighbourhood specified by k. However, for classification problems, the algorithm will classify the test instance with the modal class of the k neighbours [30].

Let a represent a new test instance and b represent a neighbouring training instance in a data set defined by n features. The simplest similarity measure between two data instances a and b is to evaluate the distances between the instances within the n-dimensional feature space. For continuous values, the Euclidean distance is perhaps the most popular measure and is given by

$$d(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}. \tag{2.19}$$

For discrete values, however, the *Hamming* distance is commonly employed, which is expressed mathematically as

$$d_{\text{Hamming}}\,(a,b) = \begin{cases} 0, & \text{if } a = b \\ 1, & \text{otherwise.} \end{cases} \tag{2.20}$$

The workings of the KNN algorithm can be easily understood through a simple example, illustrated in Figure 2.12. The algorithm is required to estimate the class of a new observation denoted by the star, given historical data containing two distinct classes denoted by circles (Class 1) and triangles (Class 2). As such, for a set of $k = 4$ nearest neighbours, the dominant class in the set is Class 1, and the class of the new instance is, therefore, classified as being Class 1. In the case where a set of $k = 15$ nearest neighbours are employed, the dominant class in the set is Class 2, and the class of the new instance is now classified as being Class 2.

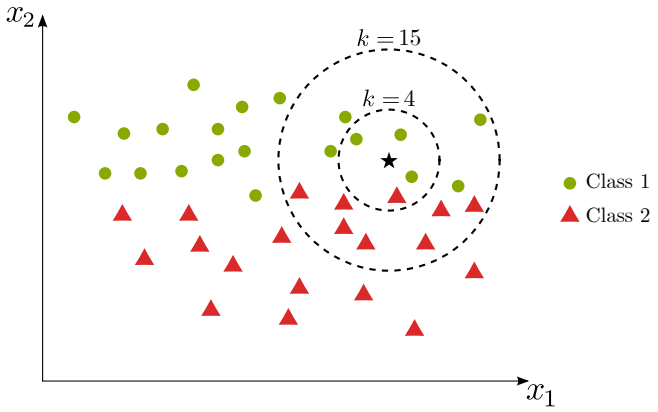


FIGURE 2.12: *Illustration of the KNN algorithm.*

One of the criticisms regarding the KNN algorithm is the selection of an optimal value for k, as this decision strongly affects the quality of the model's predictions. In particular, if the value

of k is selected as being too low, the algorithm becomes sensitive to noisy patterns in the data, and consequently susceptible to overfitting. Conversely, if the value of k is selected as being too high, the algorithm becomes susceptible to underfitting [41]. A simple technique to overcome the adverse effects of inductive biases, such as overfitting and underfitting, is to select an optimal value for k by iteratively running the algorithm, employing a different value for k each time. Upon inspecting the results of each iteration, the value for k should be selected based on the model exhibiting the best performance [28].

Logistic regression

Logistic regression is a classical learning approach deeply rooted in statistics and predicts the likelihood of a categorical variable belonging to one of two classes [36]. As such, this method is often employed in classification problems by passing the output of a basic *linear regression* model through the *sigmoid* function given by

$$\sigma(z) = \frac{1}{1 + \mathrm{e}^{-z}}, \tag{2.21}$$

where $\sigma(z)$ is the binary response prediction in the interval $\{0, 1\}$, and z is a scalar score which is representative of the characteristic traits present within the data set and can be expressed mathematically as

$$z = w_0 + \sum_{i=1}^{n} w_i x_i. \tag{2.22}$$

Furthermore, $w_1, w_2, ..., w_n$ denotes the numerical weights assigned to each of n descriptive features of the data set $x_1, x_2, ..., x_n$, and w_0 denotes a bias term utilised by the algorithm. The sigmoid function employs a threshold on the input instance, thereby producing the aforementioned binary response. Consider a problem in which the status of a generator is required to be classified as being either *faulty* (0) or *good* (1), given a set of descriptive features, $\boldsymbol{x}$ [41]. If the sigmoid function in (2.21) evaluates to $\sigma(z) < 0.5$, the binary response will be predicted as 0. Conversely, if the sigmoid function evaluates to $\sigma(z) \geq 0.5$, then the binary response will be predicted as 1.

Moreover, an advantage of the output of a logistic regression model may be interpreted as the probability associated with each target level occurring when the model is presented with a test instance t to classify. Mathematically, this can be expressed as,

$$P(t = faulty \mid \boldsymbol{x}) = \sigma(z) \tag{2.23}$$

and

$$P(t = good \mid \boldsymbol{x}) = 1 - \sigma(z). \tag{2.24}$$

Logistic regression models can be visualised by their *decision surfaces* which maps all of the possible values that the decision variables may assume, and their response according to (2.21). An example of a decision surface containing two arbitrary descriptive features, which have been normalised in the range $[-1, 1]$, and a binary target variable is displayed in Figure 2.13. The decision boundary of the model is determined by the weights $w_1, w_2, ..., w_n$, described earlier.

This discussion leads on to a fundamental component in the training of logistic regression models which seeks to determine the optimal decision boundary of the model, facilitated by an algorithm commonly known as *gradient descent*. The algorithm seeks to minimise the sum of squared errors

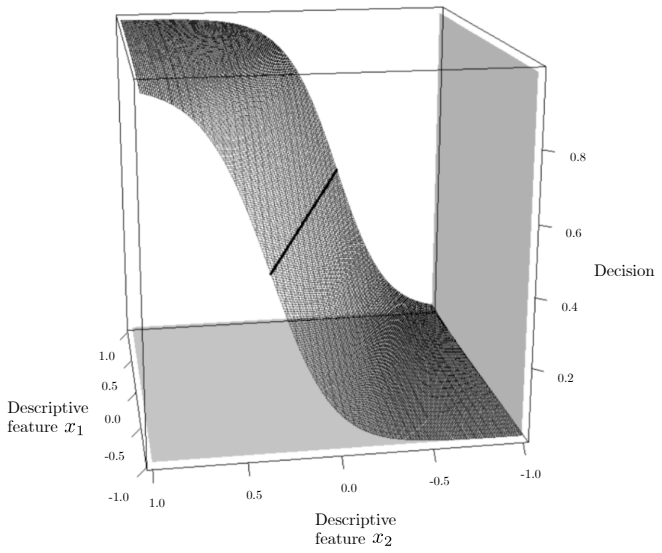


FIGURE 2.13: *Logistic decision surface, adapted from [41].*

of the training set by adjusting the weights of each descriptive variable in the model, w_i, also referred to as the model parameters. The update rule is given by

$$w_i \leftarrow w_i + \eta \times \sum_{j=1}^{m} (y_j - \sigma(z)) \times \sigma(z) \times (1 - \sigma(z)) \times x_{i,j}, \tag{2.25}$$

where η is a specified learning rate, m is the number of instances in the data set, and y_j is the binary target value of the data set instance j.

Thus far, regression models containing only linear relationships between descriptive features and a binary target feature has been considered. In some cases, however, the data exhibits non-linear relationships that are required to be captured by the model. As such, the approach to capture the underlying non-linear relationships in the data is to transform the input data, rather than the entire model itself by implementing a set of *basis functions* in the model. The sigmoid function described in (2.21) is therefore adjusted to

$$\sigma(z) = \sum_{i=0}^{n} (w_i \times \phi_i(d)), \tag{2.26}$$

where ϕ_0 to ϕ_n is a set of n basis functions that transforms each feature to the sigmoid function. The process of deploying basis functions can be visualised according to Figure 2.14 in which the feature space of a two-dimensional data set is transformed to a higher dimensional space where linear relationships can be captured.

A common application of basis functions is to capture the underlying polynomial behaviour that exists between descriptive features in the training set. Consider a simple example of a model having a single numeric descriptive feature, *Rain*, and seeks to predict the value of the target variable *Grass growth*. The relationship between the two features may be captured by the well-known second order polynomial of the form $a = bx^2 + cx$, according to the following model

$$Grass\,growth = w[0] \times \phi_0(Rain) + w[1] \times \phi_1(Rain) + w[3] \times \phi_3(Rain) \tag{2.27}$$

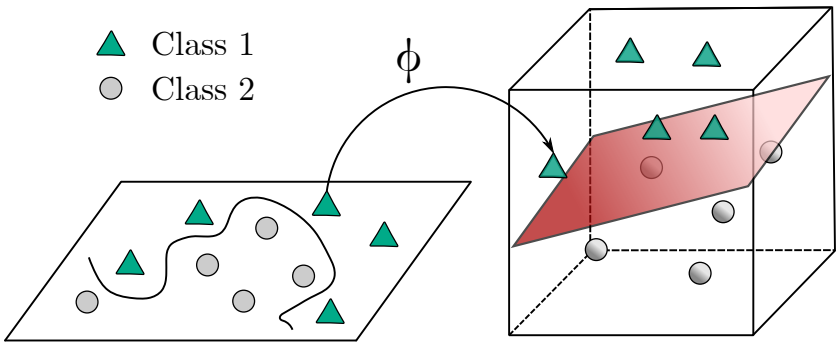


FIGURE 2.14: *Visualisation of employing basis functions.*

where the basis functions, ϕ_0 to ϕ_3 can be given as

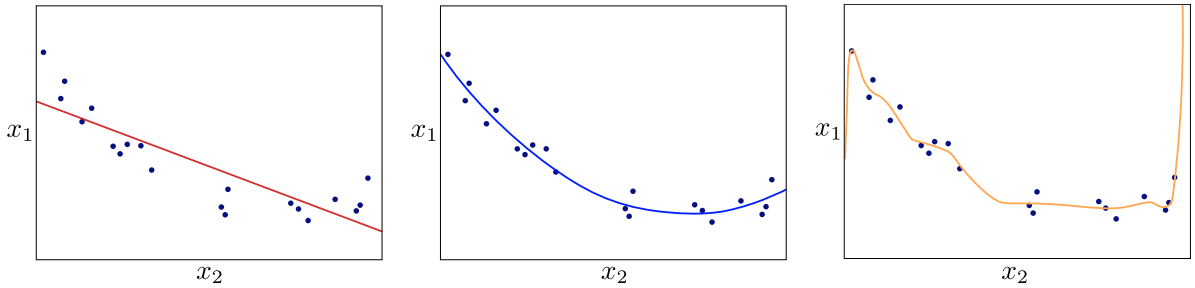
$$\begin{aligned} \phi_0(Rain) &= 1 \\ \phi_1(Rain) &= Rain \\ \phi_2(Rain) &= Rain^2. \end{aligned}$$

Finally, it is important to note that it is acceptable and common for there to be more basis functions than there are descriptive features, as illustrated in the aforementioned model [41].

2.5.4 Validation of the ML models

One of the main considerations in the construction of ML models involves selecting the best model for a specific application [36]. As such, data sets are typically divided into three subsets, namely *training*, *validation*, and *test* sets and are strategically employed to address the aforementioned issue. The training set alluded to throughout this section is typically the largest of the three subsets and is employed to build the model. The remaining two subsets are approximately equally sized, smaller than the training set, and do not contain instances of the training set. The model built from the aforementioned training set that performs most optimally on the validation set is selected as the model for the application. Thus, the validation set facilitates the selection of the best algorithm, as well as the best parameters for that algorithm with which the model is built [12]. The reason for all three sets is to ensure that the model does not only perform well on seen data as this would render a trivial model that has memorised the instances in the training set, and thereafter uses its memory to make 'predictions' with no mistakes.

This discussion leads to the notion of *inductive bias*, as well as the briefly described concepts of underfitting and overfitting in §2.5.2, which can be be avoided by applying model validation techniques [41]. A model that predicts poorly on the training set is said to have a *high bias* or underfits the data as a result of the model being too simple to capture the complex relationships within the data. An underfitting model exhibits a *low variance*, that is, the model is insensitive to large fluctuations in the training set and thus resampling the training set would result in a very similar model. Conversely, a model that predicts very well on the training set is said to have a *low bias* or overfits the data as a result of the model being too complex, thereby learning the noisy patterns present in the data. An overfitting model exhibits a *high variance*, that is, the model is sensitive to small fluctuations in the training set and thus resampling the training set would result in a very different model [12]. An illustration of an underfit, good fit, and overfit model is displayed in Figure 2.15.



(a) *Underfitting* (b) *Good fit* (c) *Overfitting*

FIGURE 2.15: *The three fittings of predictive models to data, adapted from [12].*

Holdout validation

Holdout validation is one of the simplest methods for training and validating predictive models. The data set is split into two sets by randomly sampling instances to form the training set and the validation set. A rule of thumb for the proportion of the split is 70% for training, and 30% for testing [12], before further splitting the training data subset into a smaller training data subset and validation data subset. The training set is used to train the model, while the remaining set is exclusively used to test the model to evaluate its performance on unseen data, as illustrated in Figure 2.16. Holdout validation is most useful when dealing with large data sets as it ensures that the size of the training and test sets are large enough to accurately train and evaluate the model [41].

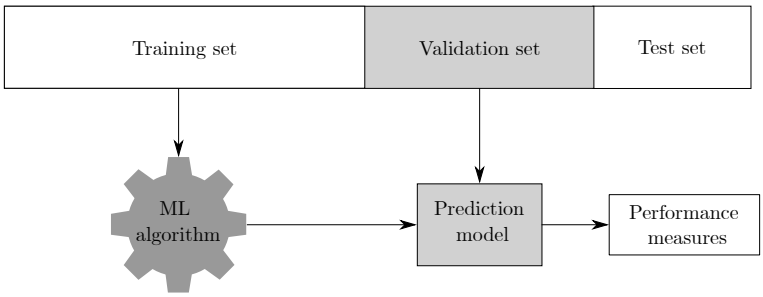


FIGURE 2.16: *Building and evaluating a predictive model using hold-out validation with the data set split into training and test sets, adapted from [41].*

k-fold cross-validation

Cross-validation is the process of repeatedly using the same data, split differently for each iteration during validation [41]. In the k-fold cross-validation method, the data set is partitioned into k equally-sized subsets or folds. During each of the k iterations, one of the subsets are kept out and used as the validation set, whilst the remaining $k-1$ subsets are combined to form the training set. As such, this method may be understood as simply repeating the aforementioned holdout validation method k times. After k experiments are performed, the performance measures across all folds are aggregated to serve as a single performance measure for the model. The value of k may be set to any value, however, 10-fold cross-validation is the most widely used variant [41]. As the value for k increases, the proportion of training instances increases, resulting in more robust estimators, however, the size of the validation set decreases [1]. The workings of this validation method may be explained via a simple example. Consider a data set

containing 1 000 data observations. If the performance of the model is to be evaluated using 5-fold cross-validation, then each fold will contain 200 data observations. The splitting of the data set into the training and validation sets for each fold is illustrated graphically in Figure 2.17.

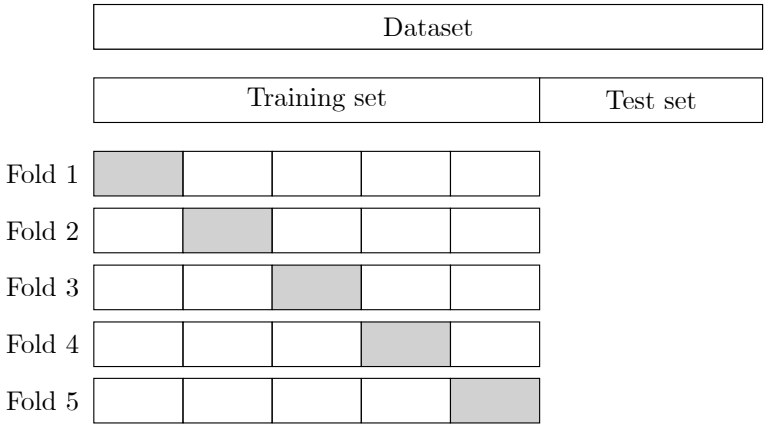


FIGURE 2.17: *5-fold cross validation displaying a data set split into training sets (white cells) and validation sets (grey cells).*

Hyperparameter tuning

Hyperparameters are the parameters of the model which are defined by the user prior to training and validation of the model such, as the value of k for the KNN algorithm, or η for the gradient descent algorithm. The purpose of this step is to obtain the optimal combination of values for the hyperparameters, such that a predefined loss function is minimised, thereby yielding a better performing model. The *Gridsearch* tuning technique is an approach that is widely utilised and involves testing every possible combination of hyperparameters, which are required to be tuned on a model, and evaluating the performance of the models until an optimal combination of parameters have been found that maximises the model performance. Thereafter, the best performing model is selected to be assessed using the test data set [12]. The total number of combinations assessed during a Gridsearch is the combinatoric product of each parameter of the model. As such, a model comprising 2 combinations for each of three parameters and employing 10-fold cross-validation would assess a total of $2 \times 3 \times 10 = 60$ combinations before reporting the best performing model.

2.5.5 Feature selection

The addition of more descriptive features to a data set counter-intuitively does not lead to more accurate models. At some point in the process of adding more descriptive features to the data set, the predictive power of the model begins to decrease. This comes as a result of many descriptive features being either redundant or irrelevant features and adversely contributes to the model predicting a target value [41]. Four types of descriptive features are important to define prior to considering the process of feature selection: (1) *Predictive* features provide beneficial information to the model in order to accurately predict the value of a target feature, (2) *interacting* features alone are uninformative about the value of the target feature, however, when used in combination with one or more features, they can be informative, (3) *redundant* features are strongly correlated with other descriptive features and therefore do not provide any

new or additional information to the model that may assist in accurately predicting the target variable, and finally (4) *irrelevant* features do not provide any information to the model that may assist in accurately predicting the target variable.

As a result, feature selection is the process of identifying and removing as many redundant and irrelevant features from the data set so that the data set mainly contains predictive and interacting features [41]. The general structure of the feature selection process is displayed in Figure 2.18. Feature selection methods typically comprise of two components: (1) A selection or generation algorithm responsible for determining the optimal subset of features, and (2) an algorithm that evaluates the subset of features from (1) and returns some measure of the subset quality. A stopping criterion is declared to ensure that the feature selection process is not computationally exhausted and can either be when the addition or removal of a feature to a subset does not produce a higher quality subset, or if an optimal subset condition is met that was predefined according to a measure of subset quality [43].

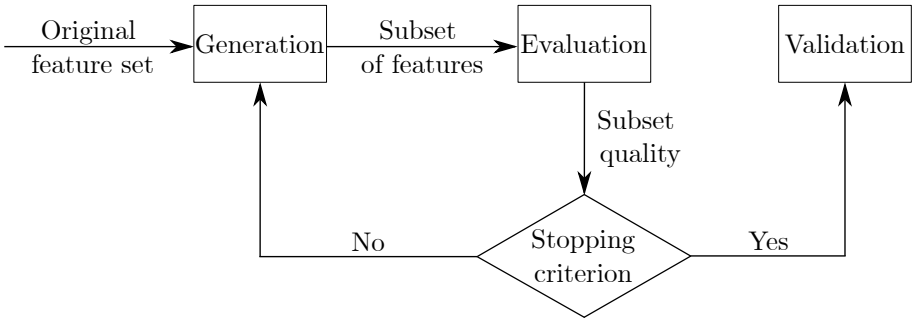


FIGURE 2.18: *General structure of the feature selection process.*

The *rank and prune* method is perhaps the simplest and most widely used approach to conduct feature selection. In this method, features are evaluated according to some measure of predictiveness and ranked accordingly. Thereafter, features that fall outside of a threshold predictive score are pruned from the data set. The measures of predictiveness, commonly known as *filters* that filter out uninformative features, are heuristics that evaluate the extent to which a feature is predictive. Moreover, these heuristics are independent of the selected model and makes use of the intrinsic characteristics within the data. An example of a filter is information gain, as discussed in §2.5.3. A disadvantage of the rank and prune approach is that interacting features may be excluded, whilst redundant features may be included in the final data set due to each feature being evaluated independently from each other [41].

2.5.6 Model evaluation

After training and validating an ML model, the test set is utilised to evaluate how well the model performs on unseen data. A model is said to *generalise* well if it accurately predicts the class labels of the data in the test set. Many formal metrics have been developed to quantify the performance of a classification model such as the *area under the ROC*[6] *curve* (AUC). The ROC represents a summary of the classification performance by combining the proportion of instances correctly predicted from the positive class (true-positive labels) and the proportion of instances incorrectly predicted from the negative class (false-negative) into a single rate curve. Moreover, the AUC score is the preferred performance metric for ML models comprising imbalanced data sets [41]. Since this metric may only be employed to assess classifiers that make predictions

[6] *Receiver operating characteristic curve* (ROC) [12].

with an associated degree of confidence, the logistic regression, decision tree, and random forest algorithms discussed in §2.5.3 are suited to adopting this performance measure [12]

2.6 Chapter summary

This chapter reviewed the relevant literature for the concepts required to understand the topic of this project. The chapter opened in §2.1 and discussed the introduction and dynamics relating to Prosopis invasions, as well as the state of Prosopis invasions in South Africa. In §2.2, the conventional control strategies, namely mechanical, chemical, and biological were briefly defined and evaluated for Prosopis in South Africa. The GIS and CA components of this project were explored and explained in §2.3. Thereafter, §2.4 considered the mathematical modelling of population growth over time, as well as over space and time. Finally, §2.5 extensively reviewed the necessary concepts and models surrounding the paradigms of ML.

CHAPTER 3

Modelling components

Contents

This chapter is devoted to describing and implementing each of the three modelling components of this project. The chapter is opened in §3.1.1 with a focus on the spatial analysis component in which the importance of visualising and exploring spatial data, discretising the study area, as well as the application of GIS software in the analysis is explored. The implementation of the ML component is detailed in §3.1.2, explaining each phase of the model construction as well as the expectations regarding the output of the model. An in-depth description of the CA model is provided in §3.1.3, defining the hexagonal neighbourhood structure, the cell states of the model, as well as the transition rule required to update the state of the cells in the study area. Finally, the chapter concludes in §3.2 with a summary of the work covered in the aforementioned sections.

3.1 Model description and implementation

The symbols displayed in Figure 3.1 are employed to graphically illustrate the workflow of the modelling components implemented in this project in the form of a *data flow diagram* (DFD). A DFD represents the inputs, processes, and outputs of a system, and the symbols in the figure represent a *process*, a *data store*, and a *data flow*. The model developed for this project is

FIGURE 3.1: *Symbols employed in DFDs.*

illustrated by a high-level DFD in Figure 3.2 comprises three primary modelling components: (1) A *spatial analysis* component facilitating the construction of the spatial data set, (2) an ML

component responsible for extracting species-environment relationships in order to predict the suitable habitats of the invasive species, and (3) a CA component which facilitates the spatio-temporal modelling of the species population growth in a region. As per the high-level DFD in Figure 3.2, the spatial analysis component is facilitated by a GIS software in order to identify the underlying spatial patterns as well as build a data set containing the relevant environmental variables. Thereafter, the data set is employed by the ML component in order to build an ML model capable of predicting the likelihood of species presence in a particular area, thereby indicating the habitat suitability and preferences of the species. Finally, the CA component employs the output data from the ML component in order to model the intrinsic growth and dispersal of the species over space and time.

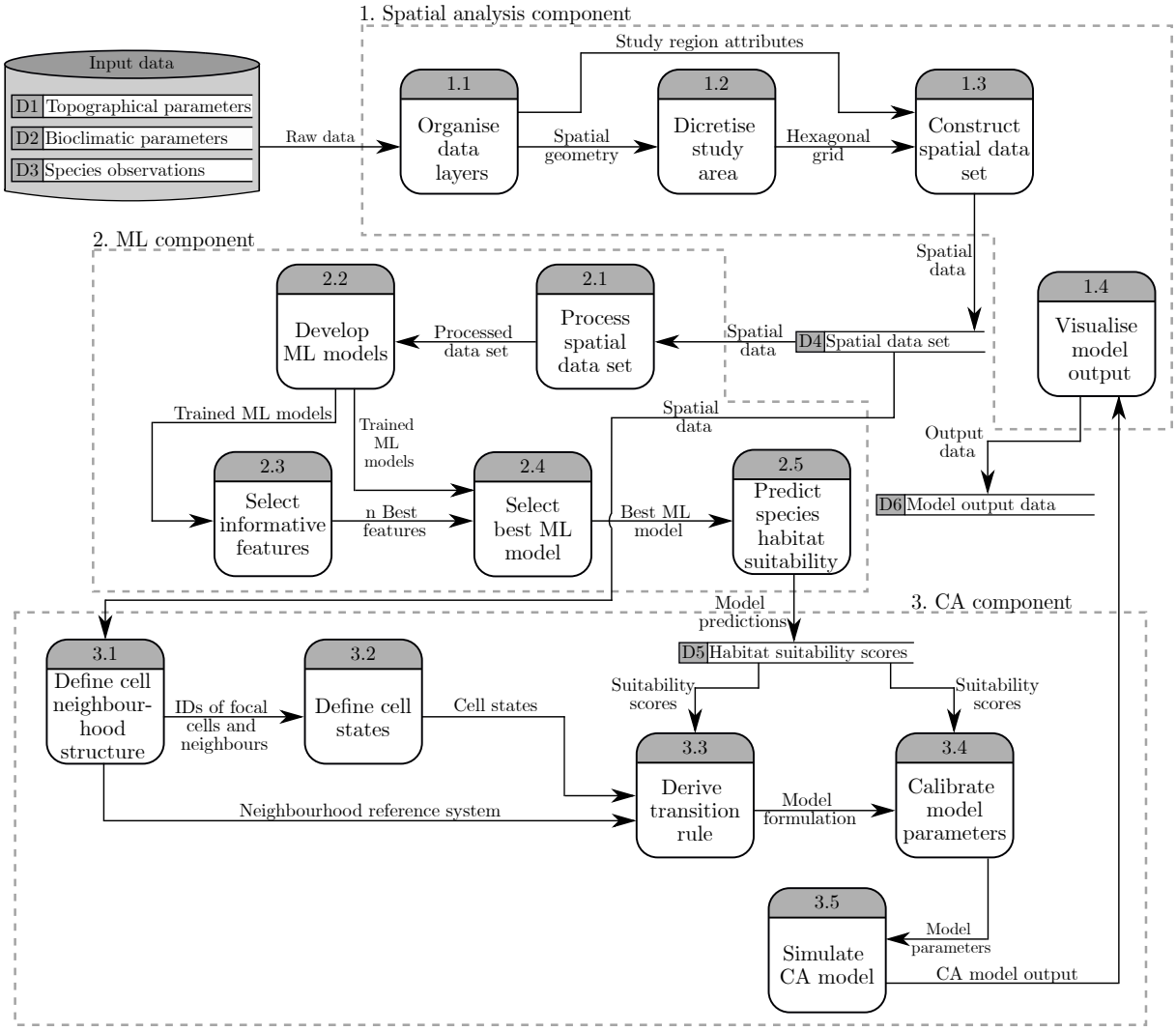


FIGURE 3.2: *A high-level DFD of the three modelling components.*

3.1.1 The spatial analysis component

The purpose of the spatial analysis component, described by Modules 1.1 to 1.4 in Figure 3.2, is to visualise and explore spatial data and to perform data manipulations in order to gain a better understanding of the underlying spatial patterns that may exist between the environment and a species population, before capturing all of the spatial data and relationships in a spatial data set

to be utilised throughout the remaining components of the model. Fundamental to the analysis is the discretisation of the study area into a spatial grid comprising smaller regions, as described in §2.3.2. The spatial domain of the model is discretised into a hexagonal lattice structure, as exemplified in Figure 3.3, with each hexagon cell specified to a side-to-opposite-side diameter of 1 km. The left-hand side of Figure 3.3 illustrates the initial display resulting from adding the hexagonal lattice layer to the modelling canvas. Thereafter, the hexagons intersecting with the extent of the study region is extracted to form the discretised study region, as illustrated on the right-hand side of Figure 3.3.

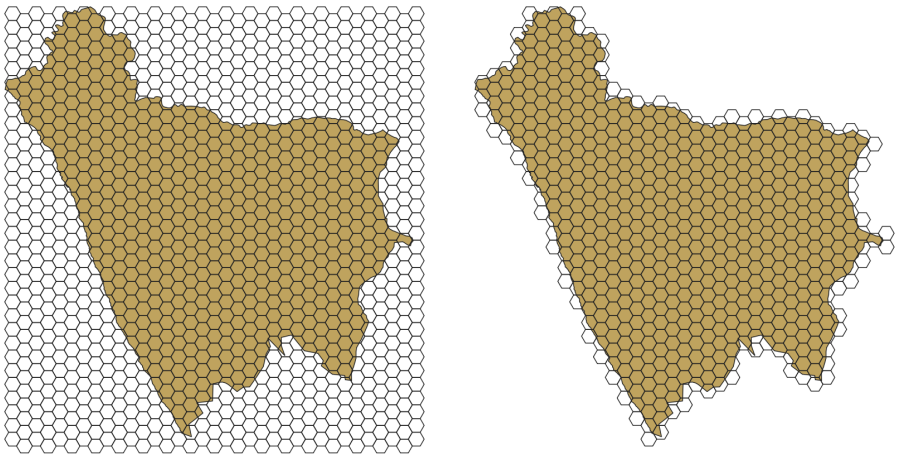


FIGURE 3.3: *A demonstration of the process of spatially discretising an area.*

The hexagonal lattice is employed as opposed to the square lattice for two main reasons. Firstly, hexagons are the roundest polygons that are able to tessellate uniformly over a grid. As such, when considering a large study area, an hexagonal grid will minimise the distortion which results from the Earth's curvature. Secondly, the centre-to-centre distance between a focal cell and any of its neighbours is the same, making it more convenient to obtain the neighbours of each cell. As a result, hexagonal grids are better suited for applications consisting of movement paths such as species dispersal [8]. The discretising process is achieved in the GIS software by employing the *Create grid* tool and selecting an hexagonal grid to be fitted to the extent of a study region according to the specified hexagon dimensions, as illustrated on the left-hand side of Figure 3.3. Thereafter, the *Extract by location* tool is used to refine the grid structure, so that it extracts only the hexagons that intersect with the study area and discards the hexagons that do not intersect. The final result of the discretised area is visualised in Figure 3.3. The overarching purpose of discretising the study area is to establish a convenient reference system that will be used in the subsequent modelling components.

Moreover, the spatial analysis component is utilised in order to populate each hexagon in the spatial grid with topographical and bioclimatic environmental descriptive features, and a target variable representing presence or absence of a species. The GIS software combines various vector and raster layers representing the geographical attributes of the Earth's surface into each of the hexagons comprising the spatial region in order to create a spatial data set that serves as input for the ML component. Each observation in the spatial data set represents a unique cell in the discretisation of the study region. The concept of data layering reviewed, in §2.3.1, is utilised by the GIS software to combine the environmental feature values with the hexagonal grid by employing the *Join attributes by location* tool. Since each environmental feature is represented by a layer, the process of constructing a data set containing n variables therefore requires $n-1$ iterations of applying the aforementioned tool.

The target variable value is derived from a species density data layer joined with the spatial data set. In particular, a cell containing a species density value greater than 0 indicates that the species is present and that the cell's target variable should be assigned a value of 1. Conversely, a cell containing no species density indicates that the species is absent and that the cell's target variable should be assigned a value of 0. The process of mapping species density values to a binary target variable may be visualised according to Figure 3.4. The left-hand side of Figure 3.4 is annotated with species density values for all cells within the study region containing non-zero species densities. The right-hand side of Figure 3.4 illustrates the study region after being shaded according to the classes of the target variable. As such, the shaded cells indicate species presence (1), whereas the unshaded cells indicate species absence (0).

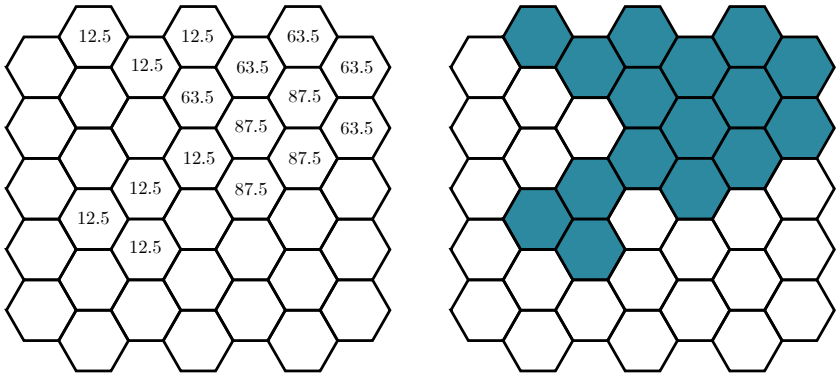


FIGURE 3.4: *Visualisation of converting species density to presence or absence classes.*

The result of joining all of the relevant layers is exemplified in Table 3.1, where F_1, F_2,...,F_n represents the descriptive features of the data set and the binary target variable in the last column represents species presence or absence within each hexagonal cell of the data set.

TABLE 3.1: *An extract from the data set constructed in the GIS software.*

ID	Latitude	Longitude	Features				Species
			F_1	F_2	F_3	...	presence
1	−30.209	21.065	89.982	21.192	6	...	1
2	−30.210	21.067	89.675	22.994	5	...	1
3	−30.213	21.067	88.699	23.083	5	...	0
4	−30.214	21.072	79.598	20.781	3	...	1
5	−30.216	21.072	82.710	19.562	2	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮

3.1.2 The ML component

The purpose of the ML component, described by Modules 2.1 to 2.5 in Figure 3.2, is to develop an ML model capable of predicting the likelihood of species presence within a study region, effectively extracting the habitat suitability of that area for the species under consideration. The expected output from the model is a prediction score in the range $[0, 1]$ that represents the habitat suitability for each hexagonal cell in the study region, with 0 translating to a low habitat suitability and 1 translating to a high habitat suitability. The ML component workflow

is illustrated graphically in the ML component section of Figure 3.2, where the spatial data set constructed in the spatial analysis component is provided as input. The data then undergoes preprocessing which involves imputing the missing values, as well as balancing the classes of the target variable using the appropriate techniques. Thereafter, the relevant ML models are trained and evaluated in order to assess their predictive performance, after which the process of feature selection is conducted in order to determine the n best features within the data set. Finally, the best suited ML model for the particular problem that achieved the highest performance score when evaluated is selected to predict the habitat suitability throughout the study region.

For the purpose of this project, *Iterative* imputation is employed to handle the missing values since it is a powerful method for imputation on large data sets. Furthermore, since the method performs multiple predictions for each missing value, as opposed to single imputations, it accounts for the standard error resulting from imputing a missing value, known as *statistical uncertainty* [3]. Briefly, this method models each of the descriptive features as a function of every other feature in the data set and thereafter imputes the missing values in ascending order, allowing for prior imputed values to be included as part of the model for subsequent imputations. Thereafter, the distribution of the target classes in the data set are investigated as they are often heavily skewed. In order to address this, random undersampling and SMOTE are employed to correct the imbalance.

The transformed data is then employed as input for the relevant ML algorithms which predicts the species presence or absence in each of the cells within the discretised study region. As such, the supervised ML paradigm is employed as it is highly effective at mapping a given set of input variables to output labels in the context of classification problems. The ML component of this project is concerned with a binary classification problem, which further necessitates the utilisation of this paradigm of algorithms. This project compares the predictive performance of the four supervised ML algorithms discussed at length in §2.5.3, namely decision trees, random forests, logistic regression, and the KNN algorithm. The decision tree algorithm requires minimal effort when preparing the data and trains relatively quickly, making it suitable for large data sets. Despite being a very intuitive algorithm, decision trees are prone to overfitting. As such, the random forest algorithm proves to be a more robust classifier as it mitigates the inefficiencies of the decision tree algorithm by aggregating the result of multiple decision trees. The logistic regression algorithm is simple to implement and its output may conveniently be interpreted as the probability of each target class occurring. However, the model is prone to overfitting the data when presented with a high dimensional data set. Finally, the KNN algorithm is a popular choice for solving classification problems, however, in the context of this project it exhibits many drawbacks. Most importantly is the fact that KNN is not suited to large data sets since the computational cost of calculating the distance between a test instance and each instance in the data set is exceptionally high and adversely affects the performance of the algorithm. Furthermore, as was reviewed in §2.5.3, the user's selection of k strongly affects the quality of the model's prediction and may lead to the model overfitting or underfitting the data [41].

The classification performance of the constructed ML models may be evaluated according to their AUC scores, and if their performances are deemed unsatisfactory, hyperparameter tuning may be ensued. This is achieved by employing the Gridsearch optimisation algorithm which determines the optimal combination of hyperparameters for the ML model. A final step to the validation of the ML models is to evaluate and rank the descriptive features according to the importance of the models. The most informative features are then selected and employed as input to re-compute the relevant ML algorithms. Finally, the validated ML model exhibiting the best predictive performance should be selected for predicting the species habitat suitability of each hexagonal cell in the study region. This is achieved by outputting the predicted probability

that a classification algorithm produces in order to identify the class to which an observation belongs, rather than the identified class. The ML probability scores of a species being present in the hexagon, is interpreted as the habitat suitability, and may be visualised as illustrated in Figure 3.5. The hexagonal cells on the grid are shaded according to the ML habitat suitability scores, ranging between 0 and 1, where the dark red cells indicate a high habitat suitability for the species, and the light yellow cells indicate a low habitat suitability for the species. The overlaid green regions illustrate the actual species distributions that may be employed to validate the output of the ML models.

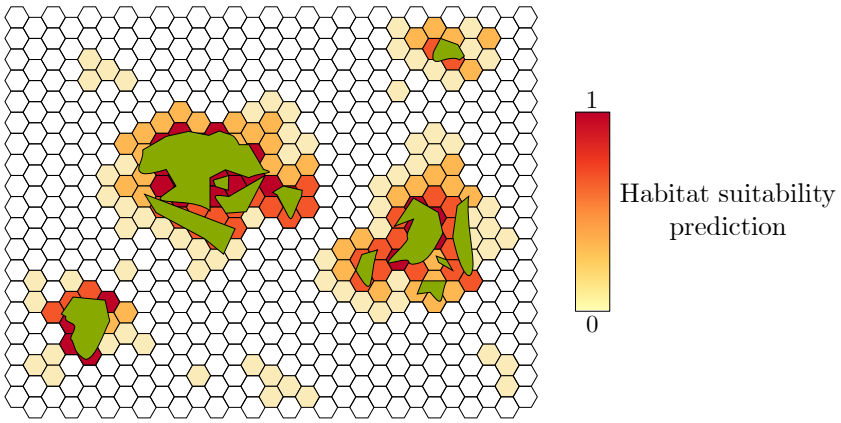


FIGURE 3.5: *A visualisation of ML habitat suitability scores, overlaid with the actual species distribution.*

3.1.3 The CA component

The CA component is the final modelling component and seeks to construct a spatio-temporal model by employing the spatial data set generated from the GIS software in the spatial analysis component and the habitat suitability scores of the ML component. The purpose of the spatio-temporal model facilitated by a CA is to simulate the future growth and dispersal of the species over space and time. The processes responsible for the development of the CA are illustrated by Modules 3.1 to 3.5 in Figure 3.2. As discussed in §3.1.1, a two-dimensional hexagonal grid structure is employed for the purpose of this project since the centre-to-centre distance between a focal cell and any of its neighbours are the same. The hexagonal neighbourhood structure and relevant axis system around a shaded focal cell is illustrated graphically in Figure 3.6, similar to how it was reviewed in §2.3.

The structure of the hexagonal neighbourhood allows a focal cell $C_{(i,j,k)}$ to interact directly with all six of its surrounding cells. As such, the set of neighbours for the focal cell $C_{(i,j,k)}$ is defined as

$$\Omega_{(i,j,k)} = \{C_{(i,j+1,k-1)}, C_{(i+1,j,k-1)}, C_{(i+1,j-1,k)}, C_{(i,j-1,k+1)}, C_{(i-1,j,k+1)}, C_{(i-1,j+1,k)}\}. \quad (3.1)$$

In the case of boundary cells observed on the perimeter of the study region, the set of neighbours are the cells that are adjacent to the boundary cell. As such, the neighbours of the boundary cell which are not contained within the study region are truncated accordingly, as illustrated in Figure 3.7. The shaded cells in the figure represent the boundary cells, and the cells containing a dot represents the cells within the neighbourhood of focal cell $C_{(i,j,k)}$. Boundary cells are, therefore, identified as those cells that have strictly less than six neighbouring cells in its neighbourhood set Ω.

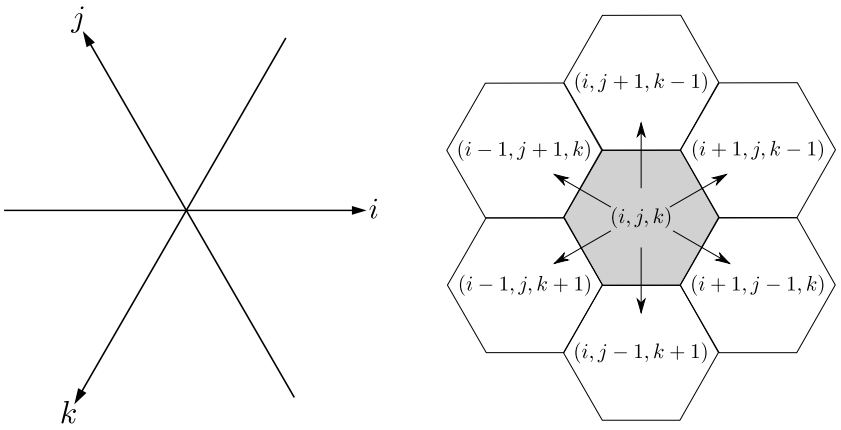


FIGURE 3.6: *The hexagonal neighbourhood structure employed by the CA model.*

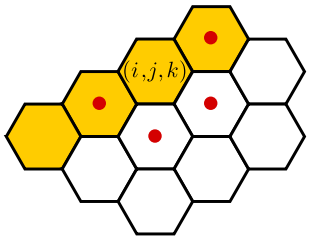


FIGURE 3.7: *An illustration of the set of neighbours for a focal cell on the perimeter of the study area.*

Cell states

The state $S^t_{(i,j,k)}$ of each cell $C_{(i,j,k)}$ at time t is equivalent to the density of the species population in that cell and its value can exhibit one of the three following states:

Vacant. A focal cell $C_{(i,j,k)}$ exhibiting a vacant state $S^t_{(i,j,k)} = 0$ indicates that the cell does not contain any presence of the species (*i.e.* species density = 0).

Inhabited. A focal cell $C_{(i,j,k)}$ exhibiting an inhabited state $0 < S^t_{(i,j,k)} < 1$ indicates that the cell contains a moderate degree of species presence (*i.e.* $0 <$ species density < 1).

Saturated. A focal cell $C_{(i,j,k)}$ exhibiting a saturated state $S^t_{(i,j,k)} = 1$ indicates that the cell is fully populated by the species (*i.e.* species density = 1). As such, the species population size has reached its carrying capacity.

Restricted. A focal cell $C_{(i,j,k)}$ exhibiting a restricted state indicates that the majority of the area of the cell intersects with features, such as buildings, roads, or bodies of water, which do not permit the presence of the species, irrespective of the time at which the cell is evaluated. As such, restricted cells remain in this state throughout the duration for which the model runs.

Transition rule

During the execution of the CA model, the state of each cell contained in the spatial data set comprising the hexagonal lattice structure is updated iteratively over m discrete time steps $t \in [0, 1, ..., m]$ according to a set of transition rules. In order to model the spatio-temporal dynamics of the invasive species population, the transition rule is composed of the logistic

growth population model which governs the species population growth, and the second rule which governs the dispersal of the species population members. The set of transition rules governing the state of cell $C_{(i,j,k)}$ during the discrete time step $[t-1, t)$ are, therefore, functions of the population growth occurring within a cell, as well as the diffusion of the population across its cell boundary to and from suitable neighbouring cells during the specified interval. More succinctly, the updating of cell states can be given by the transition rule $S^t_{(i,j,k)} = f\big(\Delta S^{\text{growth}}_{(i,j,k)}(t), \Delta S^{\text{diff}}_{(i,j,k)}(t)\big)$.

Growth: The reaction term of the model represented by logistic growth, as derived in §2.4.1, is employed to define the transition rule for growth of the species within cell $C_{(i,j,k)}$ at time t, at a rate of r which is proportional to the ML habitat suitability score $M_{(i,j,k)}$ of the cell [64], and can be given as

$$S^{\text{growth}}_{(i,j,k)}(t) = \frac{S^{t-1}_{(i,j,k)}}{S^{t-1}_{(i,j,k)} + \left(1 - S^{t-1}_{(i,j,k)}\right) e^{-rM_{(i,j,k)}\Delta t}}. \tag{3.2}$$

Dispersal: To account for the diffusion of the species across its cell boundaries, into its hexagonal neighbouring cells, the change in species density in cell $C_{(i,j,k)}$ between time $t-1$ and t may be modelled by Fickian diffusion [22, 15], and can be given by

$$\begin{aligned}\Delta S^{\text{diff}}_{(i,j,k)}(t) =& K\left[S^{t-1}_{(i-1,j+1,k)} - 2S^{t-1}_{(i,j,k)} + S^{t-1}_{(i+1,j-1,k)}\right] + \\ & K\left[S^{t-1}_{(i-1,j,k+1)} - 2S^{t-1}_{(i,j,k)} + S^{t-1}_{(i+1,j,k-1)}\right] + \\ & K\left[S^{t-1}_{(i,j-1,k+1)} - 2S^{t-1}_{(i,j,k)} + S^{t-1}_{(i,j+1,k-1)}\right], \\ =& K \sum_{C_{(p,q,r)} \in \Omega_{(i,j,k)}} \left[S^{t-1}_{(p,q,r)} - S^{t-1}_{(i,j,k)}\right],\end{aligned} \tag{3.3}$$

where $S^{t-1}_{(i,j,k)}$ and $S^{t-1}_{(p,q,r)}$ are the cell states representing the respective population densities of cell $C_{(i,j,k)}$ and its neighbouring cells denoted by the set $\Omega_{(i,j,k)}$ at time $t-1$. Moreover, $K = D/\ell$ represents a constant which combines the species diffusion rate across its cell boundary and the distance between the centroids of the hexagonal cells.

Since the habitat suitability for the species across the study area is non-uniform due to varying environmental conditions, each hexagonal cell has its own degree of habitat suitability. As such, the output of the ML model described in §3.1.2 is employed to satisfy this requirement in order to realistically model the diffusion of the species among cells with suitable habitats. In particular, the representation of the change in species density during a time interval described by (3.3) can be adjusted by incorporating the habitat suitability score predicted by the ML component and denoted as $M_{(i,j,k)}$, as a multiplier for the diffusion from a focal cell $C_{(i,j,k)}$ to its neighbouring cells $C_{(p,q,r)} \in \Omega_{(i,j,k)}$ and vice versa [15], yielding the diffusion component of the transition rule as

$$\begin{aligned}\Delta S^{\text{diff}}_{(i,j,k)}(t) = K \sum_{C_{(p,q,r)} \in \Omega_{(i,j,k)}} & \Big[M_{(i,j,k)} \cdot \max\left(S^{t-1}_{(p,q,r)} - S^{t-1}_{(i,j,k)}, 0\right) - \\ & M_{(p,q,r)} \cdot \max\left(S^{t-1}_{(i,j,k)} - S^{t-1}_{(p,q,r)}, 0\right)\Big].\end{aligned} \tag{3.4}$$

From (3.4), it is important to note that the diffusion of the species always occurs in the direction from a cell having a high density to a cell having a lower density in its neighbourhood, and the proportion of the population diffusing determined by the ML

habitat suitability score for the cell accepting the species. The combined transition rule for updating the cell states at time t may be calculated as the sum of the growth and dispersal terms, and can be expressed as

$$S^{t}_{(i,j,k)} = S^{\text{growth}}_{(i,j,k)}(t) + \Delta S^{\text{diff}}_{(i,j,k)}(t). \tag{3.5}$$

Eradication: The aforementioned growth and dispersal rules governed the transition between the inhabited and saturated states, however, it is necessary to develop another transition rule that accounts for the transition from the inhabited state to the vacant state. This comes as a result of the consideration of various control methods being implemented in hexagonal cells experiencing rapid growth within a time step. As such, the transition rule which accounts for the control method may be employed to calculate the new state of cell $C_{(i,j,k)}$ after control is implemented as

$$S^{t}_{(i,j,k)} = \begin{cases} \alpha\left(S^{\text{growth}}_{(i,j,k)}(t) + \Delta S^{\text{diff}}_{(i,j,k)}(t)\right), & \text{if } \frac{S^{t}_{(i,j,k)} - S^{t-1}_{(i,j,k)}}{S^{t-1}_{(i,j,k)}} \geq \beta \\ S^{t}_{(i,j,k)}, & \text{otherwise.} \end{cases} \tag{3.6}$$

The value β denotes a threshold value which needs to be exceeded in order for the control method to be implemented, and $\alpha \in [0, 1]$ denotes the effectiveness of the control method. Moreover, the control method is assumed to be implemented at the end of the time step at which the threshold β was exceeded so as to ensure that the states of the eradicated cells in time step t are correctly reflected as the initial states at time $t + 1$.

Inactivity: In the case where a cell's state remains unchanged between two time steps such as when the cell is restricted or vacant and no population density diffuses into it during the time period, the transition rule assigns the value of the cell's state at time $t - 1$ to the state of the corresponding cell at time t and can be expressed as $S^{t}_{(i,j,k)} = S^{t-1}_{(i,j,k)}$.

Once the CA model is executed, the updated distribution of species density is iteratively stored at each time step for the duration of the simulation. This enables for the visualisation of the spatio-temporal growth and dispersal of the species population after each time step and is achieved by storing and displaying each iteration as its own layer in the GIS software. Figure 3.8 illustrates the typical output of a CA model between two time steps where the shading of each cell indicates the relative population density of the species within the cell, making it convenient to observe the regions experiencing a large increase in population density between time steps.

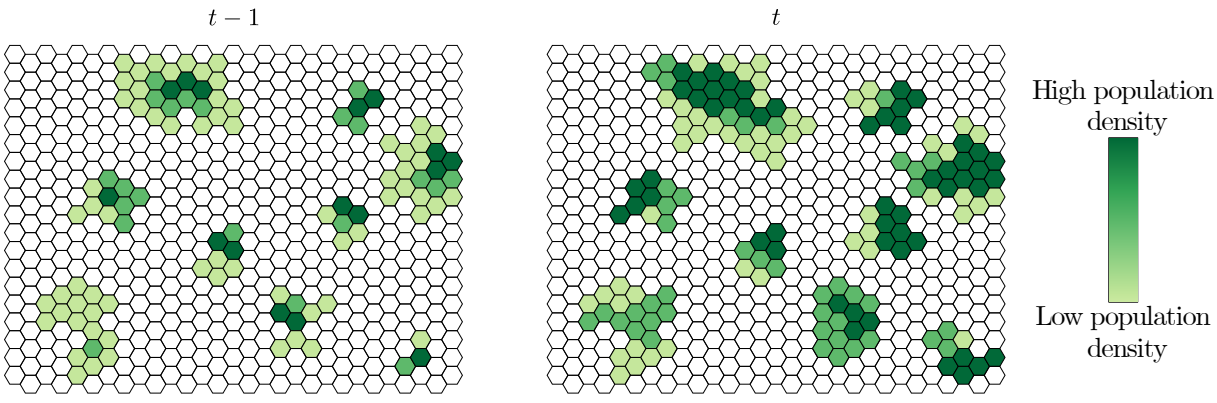


FIGURE 3.8: *A visualisation of the expected CA output between two time steps.*

3.2 Chapter summary

This chapter was devoted to reviewing the three modelling components required for the execution of this project. The chapter opened in §3.1.1 with a discussion on the spatial analysis component and the analysis completed using GIS software. Thereafter, the implementation of the ML component was described in §3.1.2, facilitating and producing the species habitat suitability scores. The chapter concluded with a detailed description of the CA component in §3.1.3, including the derivations necessary to develop and execute the CA model, facilitating the spatio-temporal modelling of the species population growth and dispersal, as well as the implementation of controlling species populations experiencing rapid growth.

CHAPTER 4

Model verification and validation

Contents

This chapter focuses on the verification and validation considerations for the ML and CA models employed in predicting the habitat suitability and modelling the spread and control of the invasive species Prosopis in this project. First, in §4.1, the verification of the ML models is discussed in great detail and the performance results of the considered models are tabulated in support thereof. Thereafter, the process of feature importance and feature selection is conducted in order to identify the minimum number of features from those deemed most important to the prediction of Prosopis habitat suitability. A subset of these features are then employed to validate the feature selection process. This is achieved by comparing the range of values of the subset of features of the hexagonal cells in which Prosopis was predicted present, with the preferred ranges of values for those features found within the literature. Secondly, the validation and calibration considerations of the spatio-temporal model are discussed in §4.2 by comparing various approaches which have been adopted in studies similar to this project. The chapter finally concludes in §4.3 with a brief summary of the aforementioned sections.

4.1 ML model verification and validation

The purpose of verifying and validating the predictive model is to firstly ensure that the model has been correctly built, and secondly, to ensure that the correct model was built. As such, all of the ML algorithms considered in §2.5.3, except for the KNN algorithm, were evaluated for their for their ability to correctly predict the presence or absence of Prosopis in each hexagonal cell within a spatial data set constructed from the entire Northern Cape, the study region selected for this project. The KNN algorithm was omitted from the analysis as it proved to be significantly more computationally expensive than the remaining three algorithms. This is due to the fact that it relies heavily on computing distance measures when classifying a new test instance, which is undesirable when confronted with a large data set — as is the case in this project.

The AUC scores achieved by each of the relevant ML models are tabulated in Table 4.1, and verifies that the random forest model performed best when compared to the other algorithms considered. The high performance score obtained by the random forest model further verifies

TABLE 4.1: *AUC scores for the relevant ML algorithms considered.*

ML algorithm	AUC score
Random forest	0.998
Decision tree	0.910
Logistic rgression	0.735
KNN (infeasible)	N/A

that the algorithm correctly predicted the absence or presence of Prosopis in almost all of the hexagonal cells within the data set. Furthermore, the various algorithms were evaluated and compared for each of the class balancing techniques, namely random undersampling and SMOTE. As a result, the model employed for the remainder of the ML modelling component was the random forest algorithm, which was constructed using the SMOTE balanced data set. In order to optimise the model, hyperparameter tuning was implemented for the random forest algorithm in order to obtain the combination of hyperparameters that yield the best performance of the algorithm. This was achieved using the Gridsearch algorithm and employing 3-fold cross-validation. Upon re-evaluating the model with the parameters obtained from the hyperparameter tuning process, the performance of the baseline model was not significantly improved, yielding an AUC score of 0.999. The lack of improvement may be attributed to the fact that the baseline model had already performed excellently, leaving minimal room for improvement of the model's performance. Moreover, since the decision tree model performed very well, it was expected that an ensemble of trees would perform even better.

A feature importance analysis was conducted in order to identify the features that were deemed most important by the random forest ML model in predicting the absence or presence of Prosopis throughout the study region. The process of ranking the features of a data set was be achieved by investigating the amount that each feature contributes to decreasing the average entropy, as reviewed in §2.5.3, across all trees within the forest [47]. As such, the 30 environmental features in the data set considered in this project were scored and ranked, as illustrated in Figure 4.1.

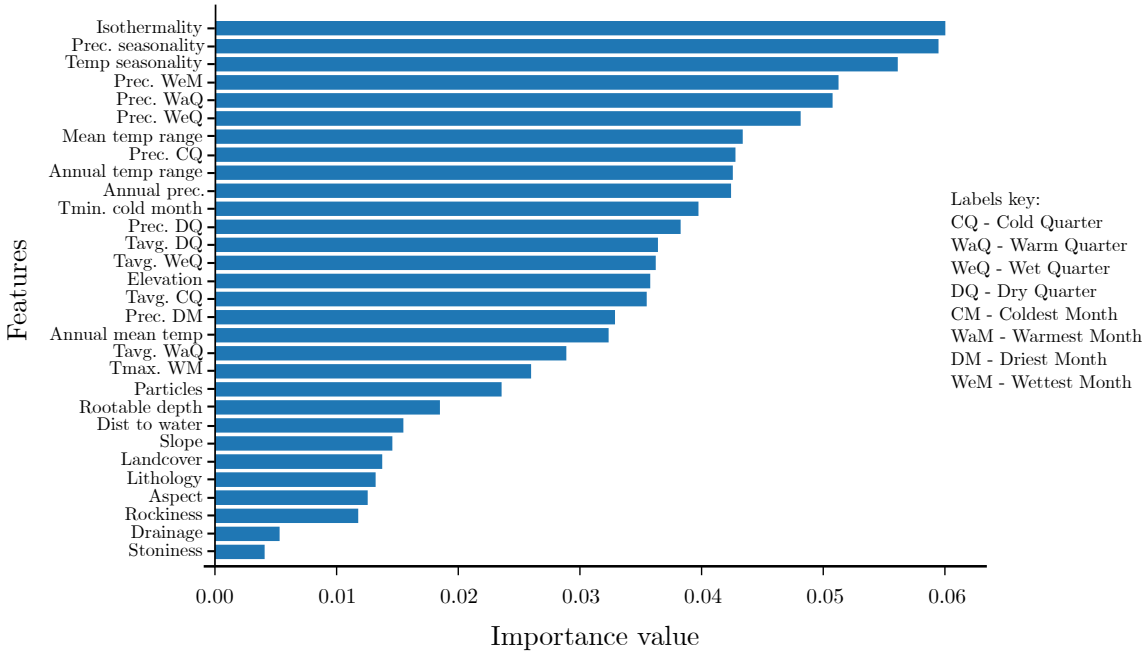


FIGURE 4.1: *The 30 environmental features in the data set ranked in order of decreasing importance.*

As a result of obtaining the features in terms of their importance, the process of feature selection was executed in order to reduce the dimensionality of the data set, whilst maintaining a desirable level of performance. In doing so, the computational cost expended by the model was reduced, leading to faster training of the model. The process entailed constructing a model for each configuration of the possible number of features and evaluating the corresponding AUC score of each of the models. With each model evaluated, the succeeding model was simply the same model, but constructed to include the next most important feature in the data set. The effect on the AUC score from adding additional features, in their ranked order, to the data set for the ML model to consider is graphically illustrated according to the curve in Figure 4.2.

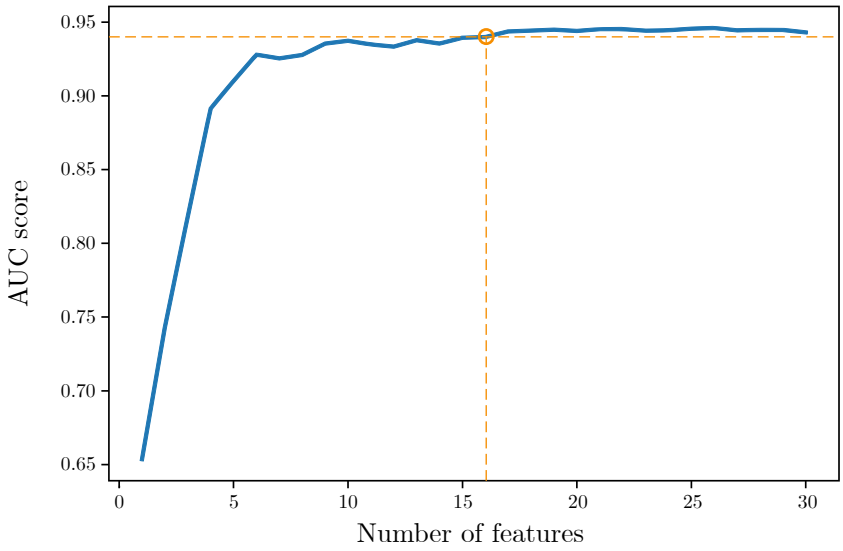


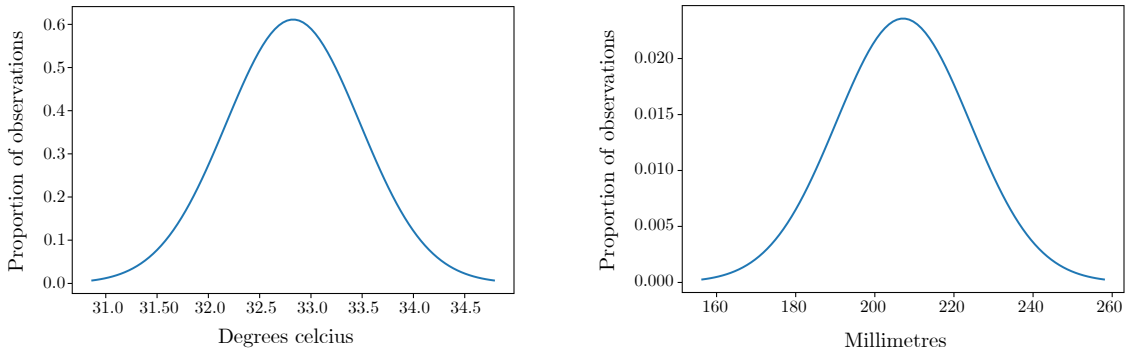
FIGURE 4.2: *Selecting a suitable number of features by comparing the number of features considered to the resulting AUC scores.*

The minimum number of features corresponding to a suitable AUC score is typically left to the discretion of the analyst. Upon inspecting Figure 4.2, it was found that the top 16 features corresponded with an AUC score of approximately 94% and was deemed an acceptable number of features to consider for the model, since adding more features to the data set did not yield significantly improved AUC scores. After developing the ML model only considering only the 16 most important features, the model was further validated by comparing the range of values that each feature assumes for all instances in which Prosopis was predicted as being present, with the known ranges of features extracted from the literature that are specific to Prosopis. It can be assumed that if the range of values observed from the model fall within the range of the values obtained from the literature, then the model was capable of accurately extracting the real-world environmental requirements with respect to Prosopis. For the sake of concision as well as the limited literature regarding the preferred habitat of Prosopis, three features were considered in this regard, namely *Annual temperature range*, *Annual precipitation*, and *Elevation*.

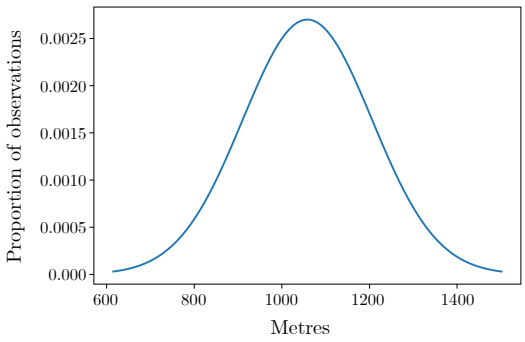
Annual temperature range. The annual range of temperatures feature was ranked as the ninth most important feature according to the ML model as per Figure 4.1, and hexagons with Prosopis present assumed values in the range 30.39 – 33.70°C within the data set, as illustrated by the normalised curve in Figure 4.3(a). The invasive species compendium, compiled by the *Centre for Agriculture and Bioscience International* (CABI) [65], confirms that this range of values is plausible, since the species is known to be observed in areas with an annual temperature range of 25 – 35°C.

Annual precipitation. The annual precipitation feature was ranked as the tenth most important feature according to the ML model as per Figure 4.1, and assumed values in the range 171 – 271 mm throughout the data set, as illustrated by the normalised curve in Figure 4.3(b). The CABI [65], confirms that this range of values is plausible, since Prosopis is known to be observed in regions where the annual precipitation is less than 860 mm.

Elevation. The elevation above sea level feature was ranked as the 15-th most important feature according to the ML model as per Figure 4.1, and hexagons containing the presence of Prosopis assumed values in the range 863 – 1 606 m, as illustrated by the normalised curve in Figure 4.3(c). The literature confirms that this range of values is plausible, since the species is known to be observed within a range of 300 – 1 900 m above sea level [7].



(a) *Annual temperature range* (b) *Annual precipitation*



(c) *Elevation*

FIGURE 4.3: *The normalised feature plots for the three validation features considered.*

4.2 Spatio-temporal model validation and calibration

Traditional science has been developed on the foundation of theories which focus on making predictions that are both accurate and testable. These theories are typically deterministic in nature as they do not account for randomness in the development of predicting the future state of the system being studied. Until recently, stochastic theories have focused on applications in which deterministic predictions for a system are made. As geographically-rooted models increase in popularity, many are focussed on representing a real-world process or system as realistically

as possible [20]. As such, these models often extend beyond the realm of purely theoretical considerations, and CA models are an example thereof.

Since the output values resulting from CA models are dynamic throughout the execution of the model, the parameters provided as input critically affects the quality of the model's ability to replicate the equivalent real-world system. As such, it was necessary to calibrate the parameters of the model and validate its output over the duration of the model execution. CA models are often adopted in very specific applications, resulting in a scarcity of generally applicable calibration and validation methods in the literature. Moreover, calibrating the model parameters and validating spatial models are time-consuming, resulting in additional challenges during the model development phase [31].

A suitable approach for addressing calibration and validation, and the approach followed for the purpose of this project, was to survey the literature within the field of ecological modelling for studies covering a variety of scenarios employing CA models [75]. The insights obtained from the studies in the literature was then be employed as a point of departure for deriving the calibration and validation methods for the project at hand. Furthermore, particular emphasis was placed on reviewing studies that incorporated the spatio-temporal growth and dispersal of single biological species populations.

In a study conducted by Cannas *et al.* [51], aerial photographs of the study region for three different years were observed and compared with the CA model predictions. In doing so, they were able to validate whether the results obtained from the model could be deemed as representative of the real-world system being studied. In a study titled *Validation of a spatial simulation model of a spreading alien plant population*, Higgins *et al.* [31] focused on various methods for validating spatial models. As such, they were able to employ known values to initialise some of the parameters in order to calibrate the remaining parameters. Briefly, the initialised parameters were obtained by utilising demographic data from aerial photographs. Thereafter, the level of spatial agreement between the output of the model and the existing data was obtained for each cell in the study region. The spatial agreement was determined by evaluating whether the existing data fell within the range of the model's predictions. Naturally, if the existing data is inconsistent with the model output data, a low confidence in the model predictions are assumed, while the converse holds true for the case in which the existing data is consistent with the model output data. This approach, however, is not suitable for studies in which the range of values predicted by the model is large.

While the methods addressed in the aforementioned studies focused primarily on comparing the model outputs with the existing data, an alternative approach for calibrating and validating the model parameters is to consult a so-called *subject-matter expert* (SME) [9]. Employing an SME for the calibration of the model parameters greatly reduces the challenge associated with the long calibration times observed when following the procedure of optimising the model through manual calculation of parameter values. SMEs are likely to estimate an appropriate parameter value based off a similar model, or draw upon personal experience gained from working with similar studies. As such, employing an SME from the relevant field of study to perform parameter calibration may yield favourable results that serve as a good representation of the real-world system being studied [44]. For the purpose of this project, some of the parameters that an SME could assist in calibrating included the intrinsic species population growth and dispersal rates, as well as the effectiveness of the relevant control strategies in inhibiting the spread of Prosopis. These values were estimated from similar studies in the literature that essentially functioned as recommendations from an SME, and the values were further tuned by comparing their effects on the several simulation instances before the final values were selected and the CA model was considered calibrated.

4.3 Chapter summary

This chapter reviewed the verification and validation considerations for the ML and CA models employed in this project. The chapter opened with a description of the verification and validation procedure of the ML model in §4.1. This included the process of evaluating the performance scores of the various models considered, the process of ranking features by importance and the selection of the most important features relevant to the study. The range of values assumed by the features of the model was subsequently compared with the known range of values assumed by those features within the literature in order to further validate the ML component. The chapter concluded in §4.2 with a discussion on the relevant validation and calibration techniques that have been adopted in order to calibrate and validate spatio-temporal models of studies similar to this project.

CHAPTER 5

A case study in the Northern Cape

Contents

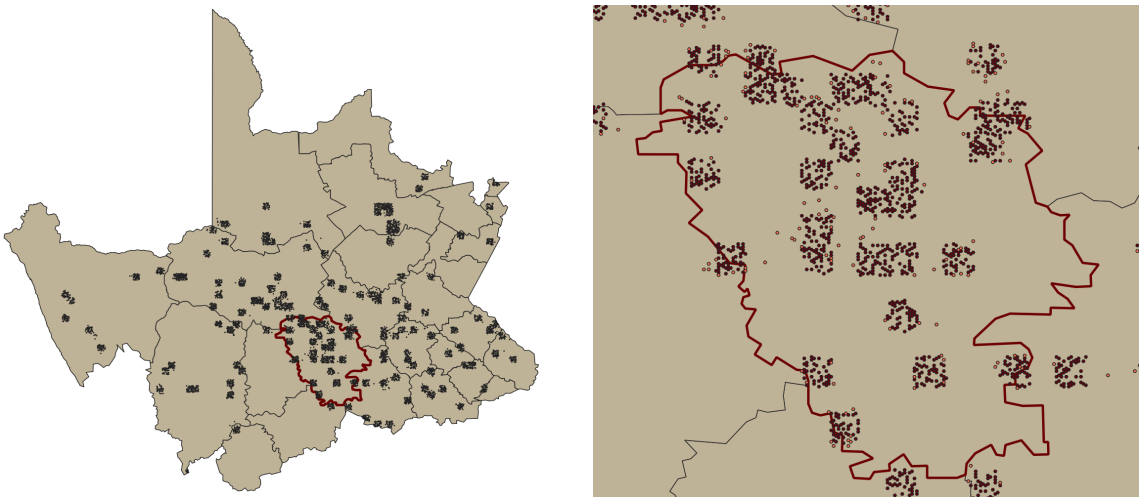
This chapter is devoted to the implementation of the model described in Chapter 3 and takes the form of a case study in the Northern Cape. The chapter opens in §5.1 with a description of the procedure followed in order to identify a suitable study region within the Northern Cape for the execution of the CA model. Thereafter, the process followed for obtaining the habitat suitability predictions for the identified study region is explained and visualised in §5.2. This is followed by §5.3 in which the combination of parameters employed during the execution of the model is defined. Moreover, the corresponding summarised results are visualised and briefly discussed. Finally, the chapter concludes in §5.4 with a brief summary of the work covered in this chapter.

5.1 Study region selection

This project specifically focusses on modelling the spread and control of Prosopis in the Northern Cape, however, due to the sheer size of the area, there is a need to select a specific study region with the aim of reducing the computational burden associated with executing the CA simulation on a large spatial grid. As such, the entire area of the province was partitioned into its 26 local municipalities and a suitable municipality was be identified and selected as the region considered for the implementation of the CA model developed in §3.1.3.

The identification of the study region or municipality considered for the case study was determined by employing an ML model capable of predicting the habitat suitability across the extent of the Northern Cape, as outlined in §3.1.2. Moreover, the predictions made were based on a 2007 density data set of Prosopis. The ML model was constructed and trained on a subset of the full data set (*i.e.* the entire Northern Cape area), after which the test data subset was provided to the model to visualise the habitat suitability distribution across the 26 municipalities. The test data subset are displayed in Figure 5.1. It is evident from Figure 5.1(a) that the outlined Kareeberg local municipality in which the town of Carnarvon is located is the most suitable region for Prosopis from a habitat suitability perspective, given that it was the region containing

the most test data subset predictions of Prosopis presence made by the ML model. An enlarged illustration of the identified Carnarvon region is displayed in Figure 5.1(b) and served as the region of interest discussed in the remainder of this chapter. The sporadic visual output of the ML model's predictions is attributed to the fact that the randomly sampled test data subset was employed to predict the habitat suitability within the 26 municipalities of the Northern Cape, as well as the fact that the predictions with a low habitat suitability score was omitted for visualisation purposes in order to identify the regions for which the habitat suitability was the highest.



(a) *Habitat suitability for Prosopis in the Northern Cape*

(b) *The most suitable region for Prosopis*

FIGURE 5.1: *Identification of the case study region by inspecting the habitat suitability within the Northern Cape.*

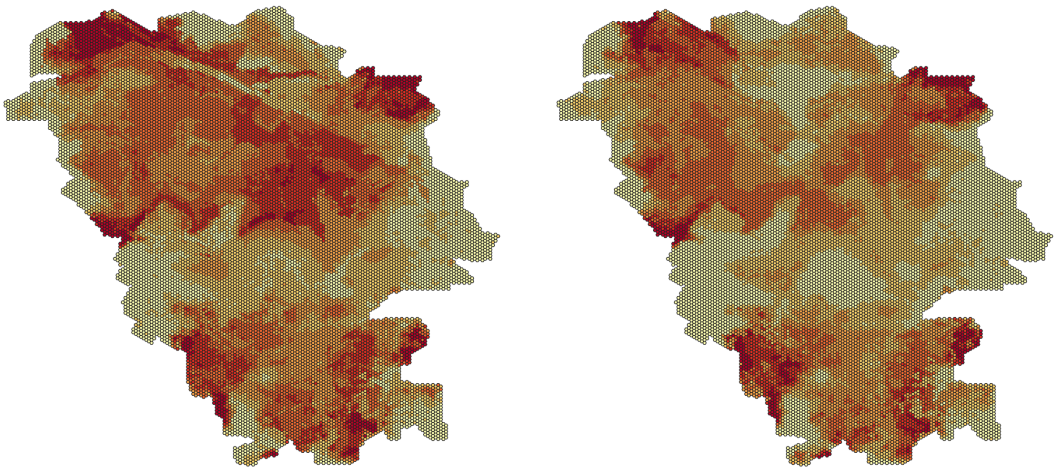
The Kareeberg local municipality occupies an area of 17 702 km^2, and accounts for 17% of the total area of the Pixley ka Seme District municipality. The main town, Carnarvon, is situated in the south of the municipal area and resides amongst the hills of the Kareeberg mountain range. Carnarvon is home to a large sheep and game farming community which contributes to Kareeberg's largest sector, agriculture, accounting for 33.8% of its economy. As such, Kareeberg is one of South Africa's largest producers of mutton and wool. Furthermore, the area is considered to be a *micro bioregion*, an area which is naturally defined by topographical and biological features, as opposed to man-made features. As such, the region comprises mountains, hills, plains, lowlands, and annual minimum and maximum temperatures ranging from -10℃ to 40℃ on average [40].

5.2 Predicting habitat suitability in the Carnarvon region

Fundamental to the execution of the CA model is the requirement of habitat suitability scores for the study region. In this regard, a ML model was constructed with a holdout test set defined to be the Kareeberg municipal region surrounding Carnarvon, comprising approximately 30% of the data set composed of hexagonal cells discretising the Northern Cape. Consequently, the remaining 70% of the data set was employed as the training data on which the random forest ML model was developed—the ML model deemed deemed the most suitable in §4.1. Two ML models were constructed for the prediction of the habitat suitability distribution around

Carnarvon: The first considered all 30 descriptive features, and the second considered the 16 most important features, as per the analysis conducted in §4.1. The results of this habitat suitability predictions made by each of the ML models are illustrated in Figure 5.2.

Upon comparing the distributions of the predictions, it is evident that the model which considered all 30 descriptive features, depicted by Figure 5.2(a), predicted a greater degree of habitat suitability, particularly in the upper half of the region, when compared to the model which considered the 16 most important features, depicted by Figure 5.2(b). The model considering 16 features was employed for the remainder of the project since its habitat suitability predictions were based on the most informative features which contributed meaningfully to the training and testing phases of the ML model, and so yielding higher quality predictions compared to the model employing all 30 descriptive features, for comparatively shorter training times.



(a) *Habitat suitability for Prosopis around Carnarvon employing all 30 descriptive features*

(b) *Habitat suitability for Prosopis around Carnarvon employing the 16 most important descriptive features from §4.1*

FIGURE 5.2: *The visualisation of the predictions made by two ML models.*

5.3 CA execution and results

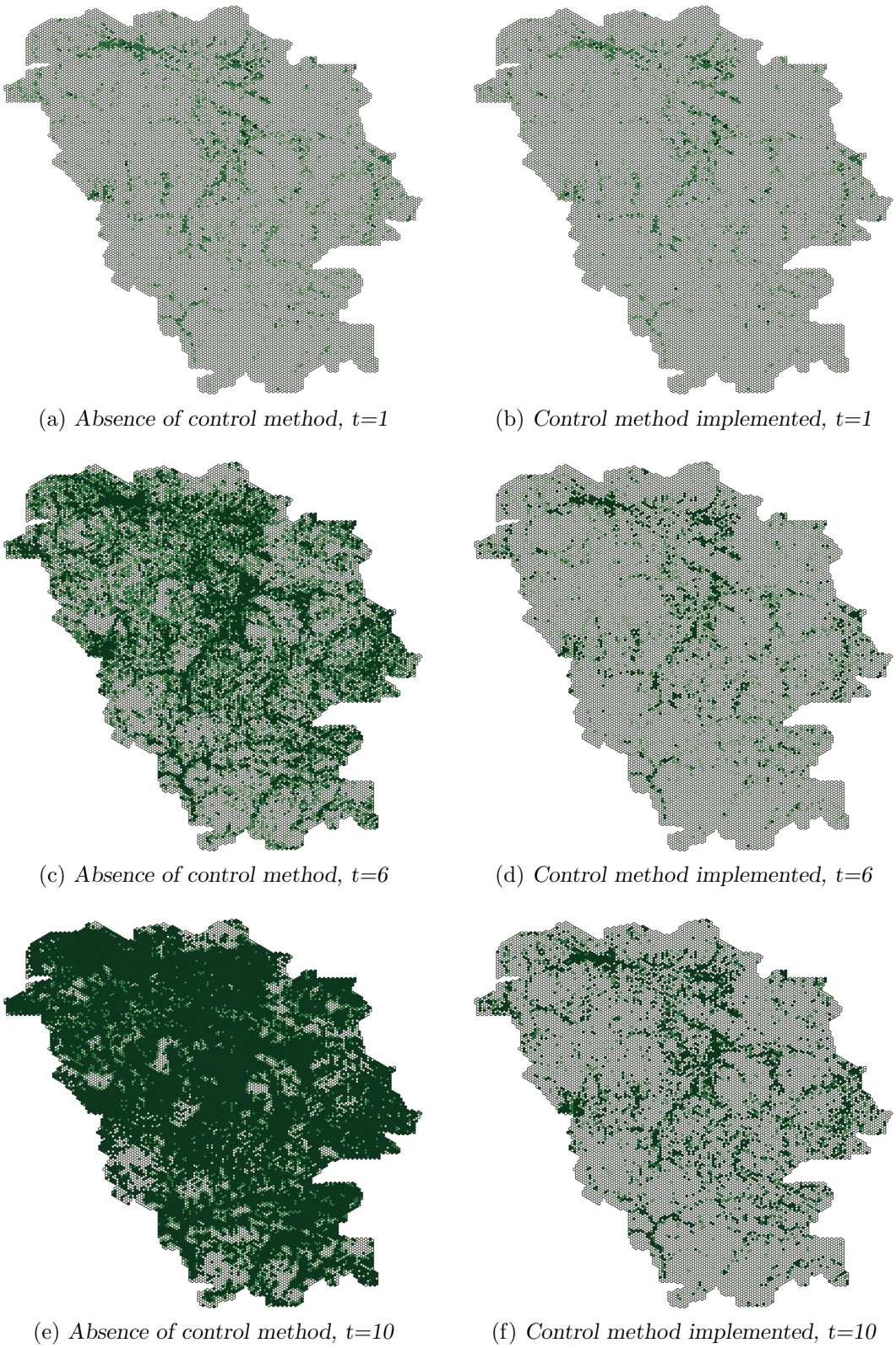
The execution of the CA model required the definition of five parameters defined in §3.1.3, namely the annual growth rate of the species r, the annual dispersal rate of the species which is included in the diffusion constant K, the threshold value indicating when to implement a control strategy β, the efficiency of the control strategy α, as well as the study period over which the simulation is required to be observed. The combination of parameters implemented were determined from the literature, as well as experimentally with the assistance of an SME to be: r=0.18, K=0.05, β=1.5. Finally, α was given a theoretical minimum and maximum effectiveness range of 40–60%. Moreover, the study period was observed over a period of ten years. The pseudocode for the implementation of the CA model is provided in Appendix B.

A consideration that greatly affects the results of the CA model is the effectiveness of the control method, α. As such, this parameter should be estimated with the aid of an SME so as to make a realistic assumption, especially since these values are not explicitly available in the literature. For the purpose of simulating the effectiveness of control methods in the real-world system as realistically as possible in this project, the value for the control method effectiveness

α was sampled randomly from the range 0.4–0.6 whenever a control method was required to be implemented. By adopting a range of effectiveness values as opposed to a single value, the varying effectiveness accounts for factors such as inefficient removal, human error, as well as the sheer size of each hexagonal cell and the effort required to clear such a large region. This consideration enables the CA model to be more representative of the real-world system, thus producing a simulation of a higher quality. While the complete results for the 10-year study period is given in Appendix C, a summary of three intervals during the execution of the model is illustrated in Figure 5.3. The time steps illustrated in the figure show the growth of Prosopis over ten years with no control implemented at year 1 in Figure 5.3(a), year 6 in Figure 5.3(c), and year 10 in Figure 5.3(e). By implementing the control method for the same intervals, the effectiveness of the control method significantly reduces the spread of Prosopis and is clearly visualised in Figures 5.3(b), 5.3(d), and 5.3(f).

5.4 Chapter summary

This chapter was dedicated to a case study by implementing the CA model to simulate the spatio-temporal spread and control of Prosopis in the Northern Cape. The procedure followed towards identifying and selecting a suitable study region within the Northern Cape was described in §5.1. Thereafter, the habitat suitability of the identified region was investigated and the results thereof was visually illustrated in §5.2. Finally, the summarised results of the CA model with, and without a control method implemented was discussed in §5.3.



(a) *Absence of control method, t=1*

(b) *Control method implemented, t=1*

(c) *Absence of control method, t=6*

(d) *Control method implemented, t=6*

(e) *Absence of control method, t=10*

(f) *Control method implemented, t=10*

FIGURE 5.3: *A summary of the CA model's results comparing the spread of Prosopis with, and without a control method implemented.*

CHAPTER 6

Conclusion

Contents

The final chapter of this project is devoted to a reflection on the work covered in this project and what the author has learnt throughout its execution. The chapter opens in §6.1 with a summary of each of the five chapters contained within the project. Thereafter, an appraisal of the project contributions is provided in §6.2. This is followed by suggestions for possible future work relating to the project in §6.3. Finally, a reflection by the author is provided in §6.4 about what he has learnt during the execution of the project.

6.1 Project summary

The introductory chapter of this project opened with a foundational background on invasive biology, a brief introduction of the invasive plant species considered in this project, as well as the various modelling components required to successfully execute the the problem addressed in this project. Thereafter, the problem statement was formally declared in §1.2 and the objectives were stated in §1.3. This was followed by §1.4 in which the scope of the project was outlined. Lastly, the report organisation and research methodology was described in §1.5.

Chapter 2 consisted of a thorough review of the relevant literature in fulfilment of Objectives I(a)–(f) as outlined in §1.3. The chapter opened with a detailed review on th characteristics of Prosopis invasions, as well as the state of Prosopis invasions in South Africa. Subsequently, the various control methods which are typically employed to inhibit the spread of a biological invasive species was addressed. This was followed by an in-depth discussion introducing the three modelling fields adopted for the successful execution of this project, including spatial analysis, spatio-temporal population modelling, and supervised ML. The spatial analysis section explained the use of GIS software as an important tool for visualising and analysing spatial data, as well as the paradigm of CA modelling and how it may be integrated with the GIS software. The spatio-temporal section was then discussed by elucidating the mathematical modelling of population growth over time as well as over space and time. Finally, the ML component went into great depth explaining the paradigm of supervised ML, the considerations for data pre-

processing, the models considered for this project, typical validation techniques, the process of hyperparameter tuning, the process of feature selection, as well as model evaluation.

The purpose of Chapter 3 was to provide a more detailed description for each of the modelling components and explain how they were implemented in the project, in fulfilment of Objectives II, III, and IV. A process model in the form of a high-level DFD was provided in order to explain the workflow proposed for the execution of the project and the communication and data flow between the spatial analysis, ML, and CA modelling components.

In pursuit of fulfilling Objective V, the verification and validation considerations of the model was discussed in Chapter 4. The chapter opened in §4.1 and addressed the verification and validation of the ML model. This entailed comparing the performance scores of the models considered, describing the implementation of conducting feature importance and feature selection, as well as comparing the range of values assumed by the model with the values identified by the literature as being most suitable for Prosopis. Finally, §4.2 compared the various approaches which have been adopted for validating and calibrating CA models similar to the one developed in this project, so as to aid in selecting an appropriate approach that was used in this project.

Chapter 5 was devoted to applying the developed CA model to a real-world case study in the Northern Cape in fulfilment of Objective VI. The identification process of the study region considered for the CA model, the Kareeberg municipality, was described in §5.1. This was followed by the process of predicting the habitat suitability for Prosopis in the Carnarvon region in §5.2. Finally in §5.3, the adopted parameters and input data were specified, and the corresponding output of the CA simulation was provided. The results of the CA model were then evaluated in an attempt to discover how accurately it was able to simulate the spread and control of the Prosopis species in the Northern Cape, as per Objective VII.

6.2 Project appraisal

The model proposed in this project is capable of predicting the extent at which a single invasive species spreads within a given study area. As such, the CA model employs the habitat suitability predictions made by an appropriate ML model in order to simulate the growth and dispersal of the species, as well as the effect that control strategies have on the spread of the species.

The CA model is governed by well-known mathematical models which are deeply rooted in the modelling of population growth. As such, the model requires species-specific data, as well as pre-defined parameters as input. The species-specific data refer to the habitat suitability predictions made by the ML model, as well as the species density within each cell in the study region. The pre-defined parameters refer to the species growth rate, dispersal rate, growth threshold for necessitating a control strategy to be implemented, and the effectiveness thereof. The output of the model is the simulated spread of the species at each of the discrete time steps for the duration over which the simulation is specified to run.

The intended goal is that the proposed CA model is utilised as an effective and efficient decision-making tool in the hands of management responsible for controlling the spread of an invasive species within a region. Furthermore, the CA model in this project serves as a mechanism for providing more insight into the complex nature of biological invasions. In particular, the model may be utilised by management to understand how and why a species behaves in a certain manner, given particular pre-defined parameter conditions. The insight gained from such a model is believed to guide decision-makers in understanding the spatio-temporal dynamics of the species and making better informed decisions regarding the management and control of the species.

The nature of the study conducted in this project may be an effective approach for introducing inexperienced practitioners to the realm of spatial analysis, ML, and CA modelling in the context of simulating population growth. This is attributed to the powerful visualisation and spatial analysis tools which are employed when working in GIS software, the insightful ML results obtained regarding the species habitat preferences, as well as the intuitive understanding of the model outputs as a result of the visual output produced by the CA model.

Finally, the input and output data of the case study conducted in this project may be published on public platforms such as *Soar*, a digital atlas of the world's maps and imagery. Through sharing of maps and environmental data, others may conduct similar studies and validate their constructed model with the results obtained in this study. As such, this form of collaboration may enrich the process of conducting thorough research, which in turn leads to greater insights being derived in the field of invasive biology and an improved understanding of how best to manage and control the spread of invasive species.

6.3 Suggestions for future work

The purpose of this section is to provide suggestions for future follow-up work that may be conducted under less time-pressured conditions in order to build upon or improve the work presented in this project.

Proposal I *Including a user-interface to automate execution of the CA model.*

The execution and visualisation of the outputs produced by the CA model in this project was conducted manually. This required the user to execute the CA simulation in the Python programming language, and thereafter export the data into the appropriate format in order to visualise the simulation in the GIS software. This may prove to be a tedious task, especially when the user would like to compare the outputs for various combinations of parameters. As such, a convenient computerised user-interface may be developed in which a user is able to upload the relevant data and specify the parameters required to execute and visualise the CA simulation. This proposal may be incorporated into a computerised *decision support system* (DSS) and employed by management as a tool for comparing the effect of varying parameters in order to determine which control method is likely to be most effective.

Proposal II *Improving the parameter calibration process.*

The parameters employed in the CA model were determined either from known values within the literature, or by estimating values and manually calibrating the model based on the output produced by the previous simulation run. While this method may not yield optimality in terms of obtaining the most accurate values for the parameters, it is an acceptable technique, given the time constraints of this project. As such, in the presence of more time, the analyst may estimate and calibrate the model parameters more accurately by employing, for example, the Gridsearch algorithm, consulting additional SMEs to validate the implemented parameters, or extracting parameters from surveillance data such as aerial photographs.

Proposal III *Approaching the ML component as a multinomial regression problem.*

The ML component in this project was treated as a binary classification problem. This meant that the target variable in the data set contained just two classes, namely *Presence* or *Absence*. An alternative approach may be to investigate the effects of addressing the

ML component as a regression problem, where the target variable consists of continuous values. In particular, the target variable may be selected as the density of the species within the study region.

Proposal IV *Introducing an additional species to the model.*

The model developed in this project considered only the invasive species Prosopis in the Northern Cape. Many population growth models are, however, developed as multi-species models, since no species live in complete isolation. As such, additional species may be introduced into the model to account for the interactions that typically occur between species. Additionally, the biological control method may be exploited in this regard by developing a well-known model known as the *predator-prey* model, where the predator is assumed to be the biological agent responsible for hindering the spread of Prosopis.

6.4 Reflections by the author

The author of this project thoroughly enjoyed the challenging task of completing a research project of this magnitude. From a technical perspective, the author utilised many of the technical problem solving tools and techniques learnt during his four-year undergraduate degree in Industrial Engineering. Most prominently, this project taught the author how to conduct thorough research which required him to acquire knowledge within a field of work which was previously foreign to him.

During the execution of the project, the author acquired many new skills which required him to gain proficiency in the appropriate software, all of which was not formally taught during the course of the undergraduate degree. In this regard, the author gained experience in working with the QGIS software during the construction of the data set, as well as the visualisation of spatial data. Furthermore, the author developed a proficiency in the PYTHON programming language which facilitated the development of the ML component, as well as the implementation of the population dynamics mathematical models required to execute the CA simulation.

With respect to producing a professionally written scientific report, the author had mastered the LaTeX typesetting environment in order to present the information in an appealing and concise manner. The author also gained valuable experience in the BEAMER environment in order to produce aesthetically pleasing presentations for sharing his research progress, as well as the INKSCAPE and IPE software for producing high quality and visually pleasing figures.

From a non-technical perspective, the project strengthened the author's time management skills, and taught him the importance of communication in the context of working alongside a supervisor. As such, this experience enabled the author to trust his abilities and work independently, with his supervisors providing feedback and guidance when required. Furthermore, conducting research within the SUnORE research group allowed the author to experience being a part of a structured group which prides itself in delivering research excellence. This provided the author with an invaluable opportunity to share, present, and engage in academic conversations with fellow peers and postgraduate students. Finally, the project taught the author that rigour, perseverance, and discipline are fundamental characteristics in pursuit of excellence.

References

[1] Alpaydin E, 2009, *Introduction to machine learning*, 2nd Edition, The MIT Press, Cambridge (MA.

[2] Andrews B, 2003, *Techniques for spatial analysis and visualization of benthic mapping data*, (Unpublished) Technical Report 623, Science Applications International Corporation, Newport (RI).

[3] Azur M, Stuart E, Frangakis C & Leaf P, 2007, *Multiple imputation by chained equations: What is it and how does it work?*, International journal of methods in psychiatric research, **20**, pp. 9–40.

[4] Berke EM, 2010, *Geographic information systems (GIS): Recognizing the importance of place in primary care research and practice*, The Journal of the American Board of Family Medicine, **23(1)**, pp. 9–12.

[5] Bhattacharya M, 2015, *Bioclimatic modelling: A machine learning perspective*, Innovations and advances in computing, informatics, systems sciences, networking and engineering, **313**, pp. 413–421.

[6] Bhattacharya M, 2013, *Machine learning for bioclimatic modelling*, Lecture Notes in Electrical Engineering, **313**.

[7] BioNET-international Secretariat, 2011, *Keys and fact sheets: Prosopis juliflora (Prosopis or Mesquite)*, 2021], Available from `https://keys.lucidcentral.org/keys/v3/eafrinet/weeds/key/weeds/Media/Html/Prosopis_juliflora_(Prosopis_or_Mesquite).html`.

[8] Birch CP, Oom SP & Beecham JA, 2007, *Rectangular and hexagonal grids used for observation, experiment and simulation in ecology*, Ecological Modelling, **206(3)**, pp. 347–359.

[9] Brath A, Montanari A & Moretti G, 2006, *Assessing the Effect on Flood Frequency of Land Use Change via Hydrological Simulation (With Uncertainty)*, Journal of Hydrology, **324**, pp. 141–153.

[10] Brauer F & Castillo-Chavez C, 2001, *Mathematical models in population biology and epidemiology*, 2nd Edition, Springer, New York (NY).

[11] Breiman L, 2001, *Random forests*, Machine learning, **45(1)**, pp. 5–32.

[12] Burkov A, 2019, *The hundred-page machine learning book*, Andriy Burkov.

[13] Cantrell R & Cosner C, 2003, *Spatial ecology via reaction-diffusion equations.*

[14] Christopher A, 2021, *K-nearest neighbor*, [Online], [Cited February 2021], Available from `https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4`.

[15] Cilliers P, Vuuren JV & Heerden QV, 2021, *A framework for modelling spatio-temporal informal settlement growth prediction*, Computers, Environment and Urban Systems, **90**, pp. 4–12.

[16] Cole V & Albrecht J, 1999, *Modeling the spread of invasive species—parameter estimation using cellular automata in GIS*, **32**, pp. 1–3.

[17] Dahl BE, 1982, *Mesquite as a rangeland plant*, pp. 8–27.

[18] Department of Environmental Affairs, 2014, *South Africa's national listed invasive species*, Department of Environmental Affairs, URL: `https://www.invasives.org.za/files/132/Books---Booklets/943/South-Africa---Listed-Invasive-Species-A5-Booklet.pdf`.

[19] Elith J & Leathwick J, 2007, *Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines*, Diversity and Distributions, **13(3)**, pp. 265–275.

[20] Engelen G & White R, 2008, *Validating and calibrating integrated cellular automata based models of land use change*, pp. 185–211 in Albeverio S, Denise A, Paolo G & Alberto V (Eds), *The dynamics of complex urban systems: An interdisciplinary approach*, Physica-Verlag HD.

[21] Febbraro MD, Sallustio L, Vizzarri M, Rosa DD, Lisio LD, Loy A, Eichelberger B & Marchetti M, 2018, *Expert-based and correlative models to map habitat quality: Which gives better support to conservation planning?*, Global Ecology and Conservation, **16**, pp. 2–10.

[22] Fick A, 1855, *On liquid diffusion*, Journal of Membrane Science, **100**, pp. 33–38.

[23] Foody G, 2007, *Map comparison in GIS*, Progress in Physical Geography, **31**, pp. 439–445.

[24] Fortin Marie-Josee MRTDale, 2007, *Spatial analysis: A guide for ecologists*, Cambridge University Press, Alberta.

[25] Fowler SV, Syrett P & Hill RL, 2000, *Success and safety in the biological control of environmental weeds in New Zealand*, Austral Ecology: A Journal of Ecology in the Southern Hemisphere, **25(5)**, pp. 553–562.

[26] Gamito S, 1998, *Growth models and their use in ecological modelling: An application to a fish population*, Ecological Modelling, **113**, pp. 83–94.

[27] Gobeyn S, Mouton AM, Cord AF, Kaim A, Volk M & Goethals PL, 2019, *Evolutionary algorithms for species distribution modelling: A review in the context of machine learning*, Ecological Modelling, **392**, pp. 179–195.

[28] Guo G, Wang H, Bell D, Bi Y & Greer K, 2003, *KNN model-based approach in classification*, Proceedings of the Lecture Notes in Computer Science, pp. 986–996.

[29] Hanley N & Roberts M, 2019, *Economic and environmental threats of alien plant, animal, and microbe invasions*, People and Nature, **1(2)**, pp. 124–137.

[30] Harrison O, 2018, *Machine learning basics with the K-nearest neighbors algorithm*, [Online], [Cited September 2021], Available from `https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761`.

[31] Higgins S & Richardson D, 2001, *Validation of a spatial simulation model of a spreading alien plant population*, Journal of Applied Ecology, **38**, pp. 571–584.

[32] Holmes E, Lewis M, Banks J & Veit D, 1994, *Partial differential equations in ecology: Spatial interactions and population dnamics*, Ecology, **75**, pp. 17–29.

[33] Hritonenko N, 2013, *Mathematical modeling in economics, ecology, and environment*, 2nd Edition, Springer, New York (NY).

[34] Hua H, Li Z, Yu JG & Dong W, 2008, *A cellular automata model for population expansion of Spartina alterniflora at Jiuduansha Shoals, Shanghai, China*, Estuarine, Coastal and Shelf Science, **77(1)**, pp. 47–55.

[35] IUCN, 2021, [Online], [Cited August 2021], Available from `https://www.iucn.org/resources/issues-briefs/invasive-alien-species-and-climate-change`.

[36] James G, Witten D, Hastie T & Tibshirani R, 2013, *An introduction to statistical learning*, Springer, Basking Ridge (NJ).

[37] Jeschke JM, Bacher S, Blackburn TM, Dick JTA, Essl F, Evans T, Gaertner M, Hulme PE, Kuhn I, Mrugala A, Pergl J, Pysek P, Rabitsch W, Ricciardi A, Richardson DM, Sendek A, Vila M, Winter M & Kumschick S, 2014, *Defining the impact of non-native species*, Conservation Biology, **28**, pp. 1188–1194.

[38] Jordan MI & Mitchell TM, 2015, *Machine learning: Trends, perspectives, and prospects*, International Journal of Engineering Science and Computing, **349(6245)**, pp. 255–260.

[39] Karafyllidis I & Thanailakis A, 1997, *A model for predicting forest fire spreading using cellular automata*, Ecological Modelling, **99(1)**, pp. 87–97.

[40] Kareeberg mayoral council, 2019, *Integrated Development Plan 2017-2022, 2nd review*, 2021], Available from `http://www.kareeberg.co.za/Docs/doc/IDP/KAR%5C%20IDP%5C%202017-2022%5C%20-%5C%20Review%5C%202%5C%20-%5C%202019-20%5C%20Final%5C%20to%5C%20Council%5C%20May_2019%5C%20V3.pdf`.

[41] Kelleher JD, Namee BM & D'Arcy A, 2015, *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies*, The MIT Press, Massachusetts (MA).

[42] Kleinhans R & Nel S, 2020, *A behavioural data science approach towards modelling Capitec clients' financial behaviour*, Final year thesis, Stellenbosch University, Stellenbosch.

[43] Kotsiantis S, Kanellopoulos D & Pintelas P, 2006, *Data preprocessing for supervised learning*, International Journal of Computer Science, **1**, pp. 111–117.

[44] Krueger T, Page T, Hubacek K, Smith L & Hiscock K, 2012, *The role of expert opinion in environmental modelling*, Environmental Modelling & Software, **36**, pp. 4–18.

[45] Le Maitre D, Pasiecznik N & Richardson D, 2014, *Prosopis: A global assessment of the biogeography, benefits, impacts and management of one of the world's worst woody invasive plant taxa*, Journal of the Annals of Botany, **1–15**.

[46] Le Maitre D & Richardson D, 2014, *Stakeholder perceptions and practices regarding Prosopis (mesquite) invasions and management in South Africa*, Ambio - A Journal of Environment and Society, **44**, pp. 569–578.

[47] Lewinson E, 2019, *Explaining feature importance by example of a random forest*, [Online], [Cited February 2021], Available from `https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e`.

[48] Liaw A & Wiener M, 2001, *Classification and regression by random forest*, The R Journal, **2(3)**, pp. 18–22.

[49] Malczewski J, 2004, *GIS-based land-use suitability analysis: A critical overview*, Progress in Planning, **62(1)**, pp. 3–65.

[50] Malthus T, 1798, *An essay on the principle of population (1798)*, Yale University Press, London.

[51] Marco D, Páez S & Cannas S, 2002, *Species invasiveness in biological invasions: A modelling approach*, Biological Invasions, **4**, pp. 193–205.

[52] Masocha M & Skidmore AK, 2011, *Integrating conventional classifiers with a GIS expert system to increase the accuracy of invasive species mapping*, International Journal of Applied Earth Observation and Geoinformation, **13(3)**, pp. 487–494.

[53] Mazibuko DM, 2012, *Phylogenetic relationships of Prosopis In South Africa : An assessment of the extent of hybridization, and the role of genome size and seed size In the invasion dynamics*, MSc Thesis, Stellenbosch University, Stellenbosch.

[54] Micalizio C.-S, 2017, *GIS (Geographic Information System)*, [Online], [Cited June 2021], Available from `https://www.nationalgeographic.org/encyclopedia/geographic-information-system-gis/`.

[55] Miller J, 2010, *Species distribution modeling*, Geography Compass, **4(6)**, pp. 490–509.

[56] Mitchell TM, 1997, *Machine learning*, 1st Edition, McGraw-Hill, Pittsburgh (PA).

[57] Mithrakumar M, 2019, *How to tune a decision tree?*, [Online], [Cited November 2021], Available from `https://towardsdatascience.com/how-to-tune-a-decision-tree-f03721801680`.

[58] MLMath.io, 2019, *Math behind decision tree algorithm*, [Online], [Cited February 2021], Available from `https://medium.com/@ankitnitjsr13/math-behind-decision-tree-algorithm-2aa398561d6d`.

[59] Mooney HA, 2005, *Invasive alien species: A new synthesis*, Island Press, Washington (DC).

[60] Morozov A & Poggiale J.-C, 2012, *From spatially explicit ecological models to mean-field dynamics: The state of the art and perspectives*, Ecological Complexity, **10**, pp. 1–11.

[61] Neumann JV & Burks AW, 1966, *Theory of self-reproducing automata*, University of Illinois Press, Champaign (IL).

[62] O'Kelly M, 1993, *Spatial analysis and GIS*, pp. 65–79 in , *Spatial analysis and GIS: A position paper*, Taylor and Francis.

[63] Olivier G, 2013, *Invasive* Alien Plants - NEMBA listSouth African Nursey Association (SANA), URL: `https://sana.co.za/2013/01/16/invasive-alien-plants-nemba-list/`.

[64] Panik M, 2017, *Stochastic differential equations*, Ecological Modelling, pp. 275–277.

[65] Pasiecznik N, 2017, *Invasive species compendium: Prosopis juliflora (Mesquite)*, [Online], [Cited February 2021], Available from `https://www.cabi.org/isc/datasheet/43942#toairTemperature`.

[66] Pasiecznik N, Felker P, Harris P, Harsh L, Cruz G, Tewari J, Cadoret K & Maldonado L, 2001, *Mathematical modeling in economics, ecology, and environment*, Forest Ecology and Management, **174**, pp. 3–102.

[67] Patnaik P, Abbasi T & Abbasi SA, 2017, *Prosopis (Prosopis juliflora): Blessing and bane*, Tropical Ecology, **58**, pp. 455–483.

[68] Pimentel D, 2011, *Review of biological invasions: Economic and environmental costs of alien plant, animal, and microbe species*, Journal of Agricultural & Food Information, **13**, pp. 1–14.

[69] Pimentel D, McNair S, Janecka J, Wightman J, Simmonds C, O'Connell C, Wong E, Russel L, Zern J, Aquino T & Tsomondo T, 2001, *Economic and environmental threats of alien plant, animal, and microbe invasions*, Agriculture, Ecosystems and Environment, **84(1)**, pp. 1–20.

[70] Pucha-Cofrep F, Fries A, Canovas-Garcia F & Onate-Valdivieso F, 2018, *Fundamentals of GIS*, Cottbus.

[71] Rejmanek M & Richardson D, 2013, *Tree invasions: Patterns, processes, challenges and opportunities*, Diversity and Distributions, **19**, pp. 1093–1094.

[72] Renshaw E, 1991, *Modelling biological populations in space and time*, Cambridge University Press, Cambridge.

[73] Richardson D, Hui C, Nunez M & Pauchard A, 2014, *Tree invasions: Patterns, processes, challenges and opportunities*, Biological Invasions, **16**, pp. 473–481.

[74] Richardson DM, Pysek P, Rejmanek M, Barbour MG, Panetta FD & West CJ, 2000, *Naturalization and invasion of alien pants: Concepts and definitions*, Diversity and Distributions, **6(2)**, pp. 93–107.

[75] Rykiel EJ, 1996, *Testing ecological models: The meaning of validation*, Ecological Modelling, **90(3)**, pp. 229–244.

[76] Sam T, 2019, *Entropy: How decision trees make decisions*, [Online], [Cited January 2021], Available from `https://towardsdatascience.com/entropy-how-decision-trees-make-decisions-2946b9c18c8`.

[77] Sante I, Garcia AM, Miranda D & Crecente R, 2010, *Cellular automata models for the simulation of real-world urban processes: A review and analysis*, Landscape and Urban Planning, **96(2)**, pp. 108–122.

[78] Saravanan R & Sujatha P, 2018, *A state of art techniques on machine learning algorithms: A perspective of supervised learning approaches in data classification*, Proceedings of the 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 945–949.

[79] Schroders, 2019, *Location, location, location: Why geospatial data science matters for investors*, [Online], [Cited May 2021], Available from `https://www.schroders.com/id/uk/realestate/insights/thought-leadership/location-location-location-why-geospatial-data-science-matters-for-investors/?t=true`.

[80] Shackleton RT, Maitre DCL, Wilgen BWV & Richardson DM, 2015, *The impact of invasive alien Prosopis species (mesquite) on native plants in different environments in South Africa*, South African Journal of Botany, **97**, pp. 25–31.

[81] Shackleton RT, Maitre DCL, Wilgen BWV & Richardson DM, 2017, *Towards a national strategy to optimise the management of a widespread invasive tree (Prosopis species; Mesquite) in South Africa*, Ecosystem Services, **27**, pp. 242–252.

[82] Shamey R & Zhao X, 2014, *Modelling, simulation and control of the dyeing process*, Woodhead Publishing, Sawston.

[83] Skellam J, 1991, *Random dispersal in theoretical populations*, Bulletin of Mathematical Biology, **53(1)**, pp. 135–165.

[84] Store R & Kangas J, 2001, *Integrating spatial multi-criteria evaluation and expert knowledge for GIS-based habitat suitability modelling*, Landscape and Urban Planning, **55(2)**, pp. 2–10.

[85] Strikwerda JC, 2004, *Finite Difference Schemes and Partial Differential Equations, Second Edition*, Society for Industrial and Applied Mathematics.

[86] Taillandier P, Banos A, Drogoul A, Gaudou B, Marilleau N & Quang TC, 2016, *Simulating urban growth with raster and vector models: A case study for the city of Can Tho, Vietnam*, Proceedings of the International conference Autonomous Agents and Multiagent Systems, pp. 154–171.

[87] Team EP, 2018, *How to perform spatial analysis*, [Online], [Cited February 2021], Available from `https://www.esri.com/arcgis-blog/products/product/analytics/how-to-perform-spatial-analysis/`.

[88] Terblanche C, Nanni I, Kaplan H, Strathie L, Mcconnachie A, Goodall J & van Wilgen B, 2016, *An approach to the development of a natonal strategy for controlling invasive alien plant species: The case of Parthenium hysterophorus in South Africa*, Bothalia, **46**, pp. 1–5.

[89] Turbelin A, Malamud B & Francis R, 2016, *Mapping the global state of invasive alien species: Patterns of invasion and policy responses*, Global Ecology and Biogeography, **26**, pp. 79–87.

[90] University of Massachusetts Department of Geosciences, 2008, *GIS Fundamentals*, [Online], [Cited July 2021], Available from `http://www.geo.umass.edu/courses/geo494a/Chapter2_GIS_Fundamentals.pdf`.

[91] Van den Berg E, Kotze I & Beukes H, 2013, *Detection, quantification and monitoring Prosopis in the Northern Cape province of South Africa using remote sensing and GIS*, South African Journal of Geomatics, **2**, pp. 68–81.

[92] Van Wilgen BW, Wilson JR, Wannenburgh A & Foxcroft LC, 2020, *The extent and effectiveness of alien plant control projects in South Africa*, pp. 597–628 in Van Wilgen BW, Measey J, Richardson DM, Wilson JR & Zengeya TA (Eds), *Biological Invasions in South Africa*, Springer International Publishing, Cham.

[93] Van Wyk D, 2020, *Algorithmic performance prediction based on fitness landscape features*, Final year thesis, Stellenbosch University, Stellenbosch.

[94] van Wilgen BW & Wannenburgh A, 2016, *Co-facilitating invasive species control, water conservation and poverty relief: achievements and challenges in South Africa's Working for Water programme*, Current Opinion in Environmental Sustainability, **19**, pp. 7–17.

[95] Venables B & Ripley B, 2002, *Modern applied statistics with S-Plus*, pp. 251–258 in.

[96] Versfeld DB, Le Maitre DC & Chapman RA, 1998, *The impact of invading alien plants on surface water resources in South Africa: A preliminary assessment*, (Unpublished) Technical Report TT99/98, CSIR, Stellenbosch.

[97] Wakie T, Evangelista P, Jarnevich C & Laituri M, 2014, *Mapping current and potential distribution of non-native Prosopis juliflora in the afar region of Ethiopia*, Public Library of Science, **9**, pp. 1–9.

[98] Wieczorek W & Delmerico A, 2009, *Geographic information systems*, Computational statistics, **1**, pp. 167–186.

[99] Wilgen BWV, Dyer C, Hoffmann JH, Ivey P, Maitre DCL, Moore JL, Richardson DM, Rouget M, Wannenburgh A & Wilson JRU, 2011, *National-scale strategic approaches for managing introduced plants: insights from Australian acacias in South Africa*, Diversity and Distributions, **17(5)**, pp. 1060–1075.

[100] Woodford DJ, Richardson DM, MacIsaac HJ, Mandrak NE, Wilgen BWV, Wilson JRU & Weyl OLF, 2016, *Confronting the wicked problem of managing biological invasions*, NeoBiota, **31**, pp. 63–86.

[101] Yoshimoto A, Asante P, Konoshima M & Surovy P, 2016, *Integer programming approach to control invasive species spread based on cellular automaton model*, Natural Resource Modeling, **30**, pp. 1–11.

[102] Zachariades C, 2021, Senior Researcher, (Email communication).

[103] Zachariades C, Hoffmann J & Roberts A, 2011, *Biological control of mesquite (Prosopis species) (Fabaceae) in South Africa*, African Entomology, **19(2)**, pp. 402–415.

[104] Zhao Y & Billings S, 2006, *Neighbourhood detection using mutual information for the Identification of cellular automata*, IEEE transactions on systems, man, and cybernetics, **36**, pp. 473–479.

APPENDIX A

Project Timeline

The expected timeline is given in Figure A.1 in Gantt-chart form.

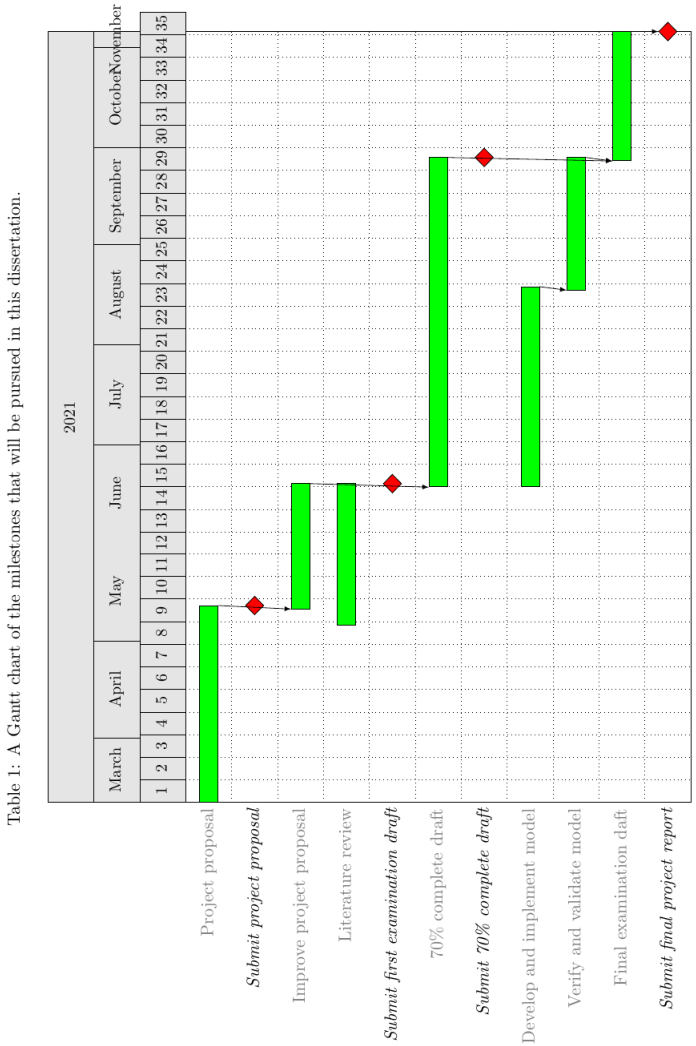


FIGURE A.1: *Expected timeline in Gannt-chart form.*

APPENDIX B

CA model implementation pseudocode

The updating of cell states according to the defined transition rules for the spatio temporal CA model may be given by the algorithm in this appendix.

Algorithm B.1: CA model implementation of transition rules during time t

Input : The state $S^{t-1}_{(i,j,k)}$ and ML habitat suitability score $M^{t-1}_{(i,j,k)}$ of cell $C_{(i,j,k)}$ and its set of neighbouring cells $C_{(p,q,r)} \in \Omega_{(i,j,k)}$ at time $t-1$, the species growth rate r, the species-specific diffusion constant K, and a control method threshold β.

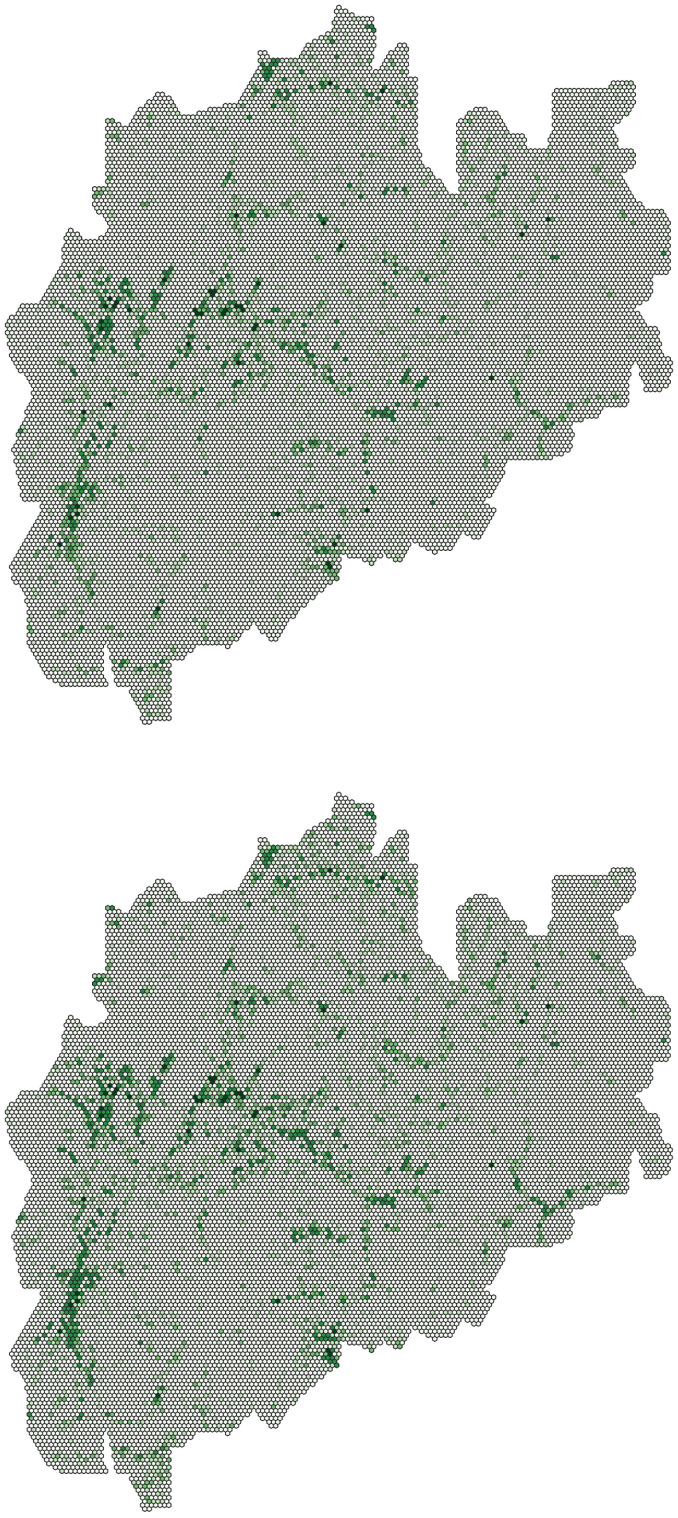
Output: The state $S^{t}_{(i,j,k)}$ of cell $C_{(i,j,k)}$ at time t.

for *each cell* $C_{(i,j,k)}$ *in the study region* **do**
  if $S^{t-1}_{(i,j,k)} \geq 0$ **then**
    $S^{t}_{(i,j,k)} \leftarrow S^{\text{growth (t-1)}}_{(i,j,k)} + \Delta S^{\text{diff (t-1)}}_{(i,j,k)}$
  if $\frac{S^{t}_{(i,j,k)} - S^{t-1}_{(i,j,k)}}{S^{t-1}_{(i,j,k)}} \geq \beta$ **then**
    $\alpha \leftarrow$ random number $\in [0.4, 0.6]$
    $S^{t}_{(i,j,k)} \leftarrow S^{t}_{(i,j,k)}\alpha$
  else
    $S^{t}_{(i,j,k)} \leftarrow S^{t}_{(i,j,k)}$

APPENDIX C

Complete results of the CA model

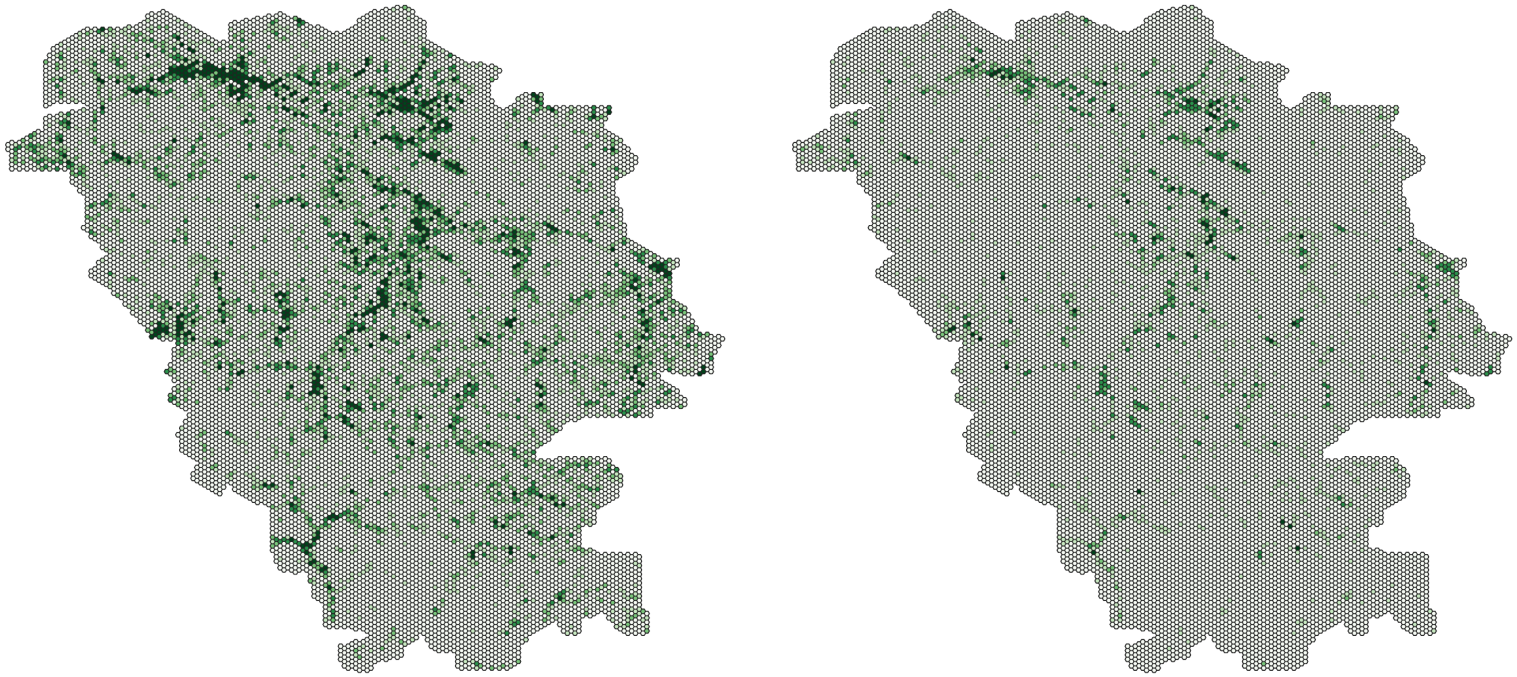
The complete results of the CA model for the 10-year study period is given by Figure C.1–C.10.



(a) *Absence of control method, t=1*

(b) *Control method implemented, t=1*

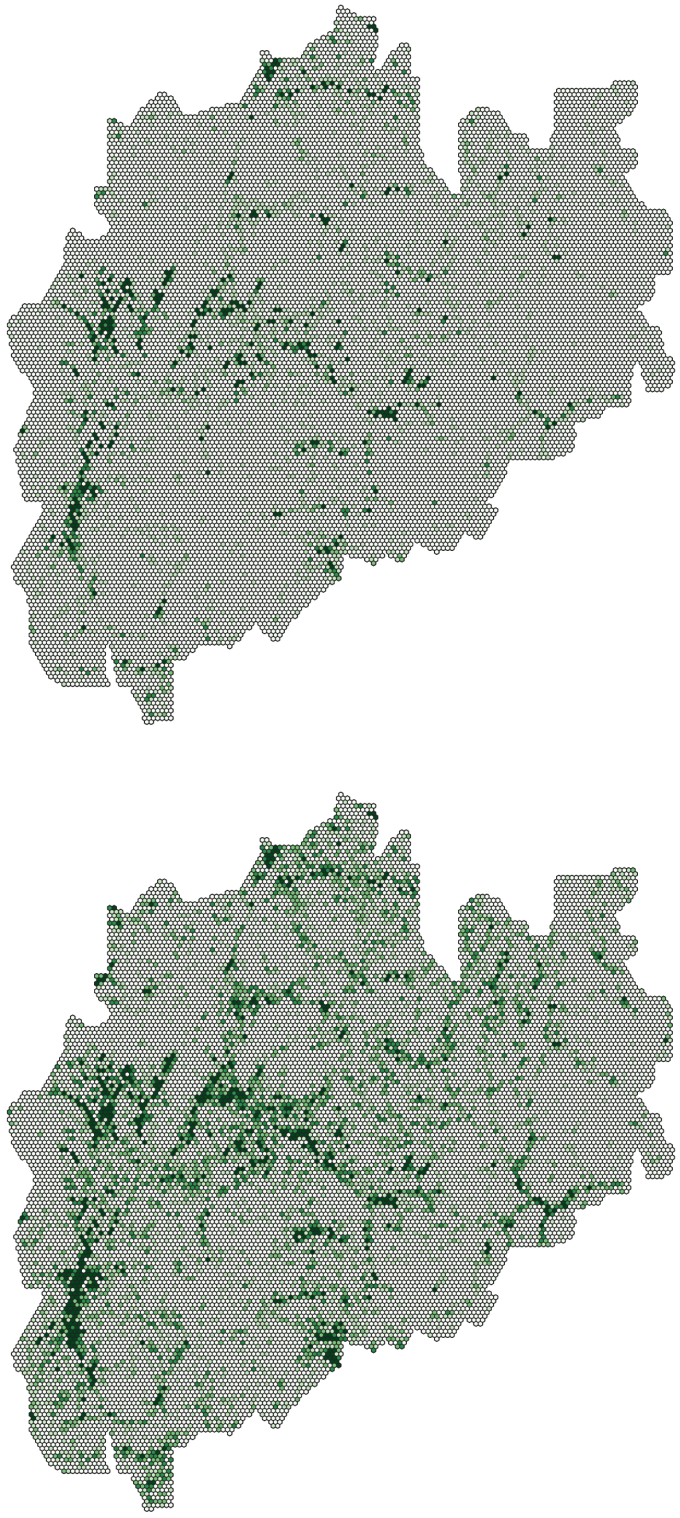
Figure C.1: *The spread and control of Prosopis at year 1.*



(a) *Absence of control method, t=2*

(b) *Control method implemented, t=2*

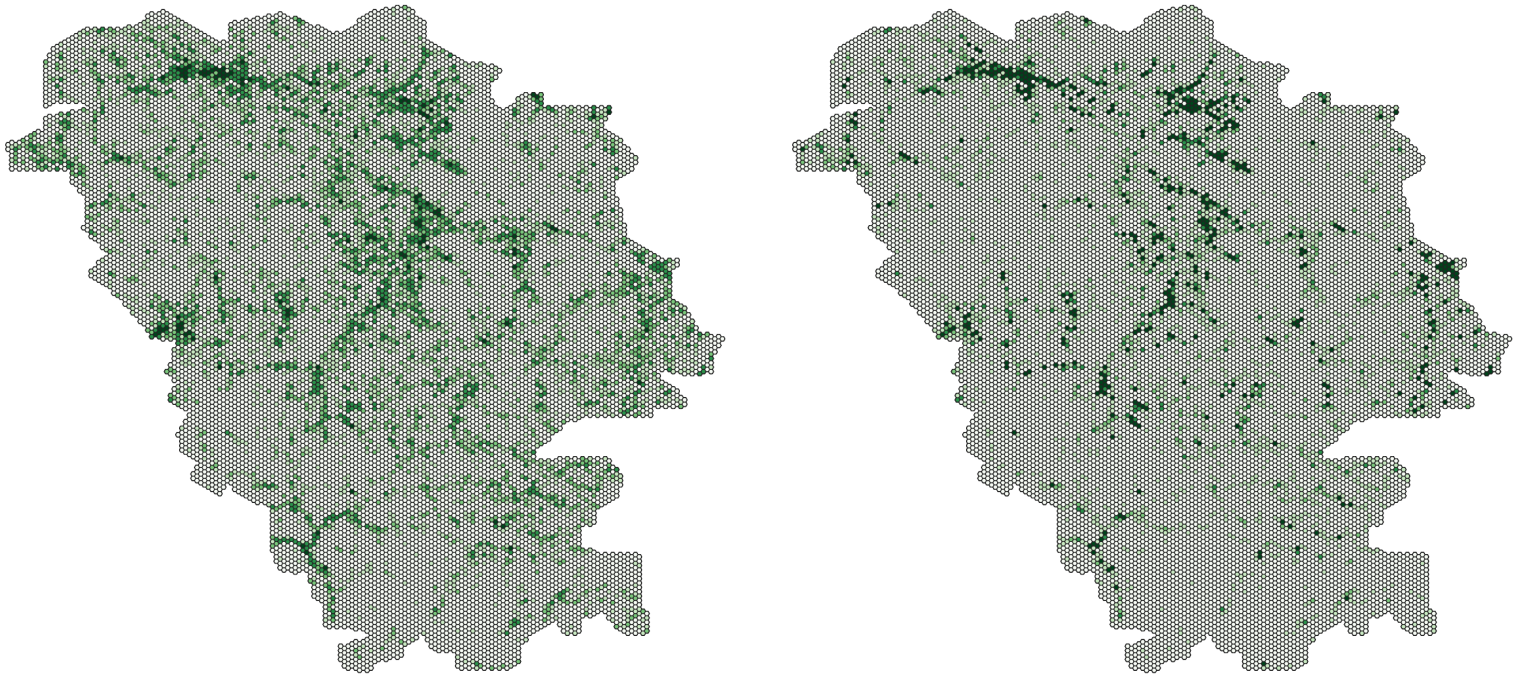
FIGURE C.2: *The spread and control of Prosopis at year 2.*



(a) *Absence of control method, t=3*

(b) *Control method implemented, t=3*

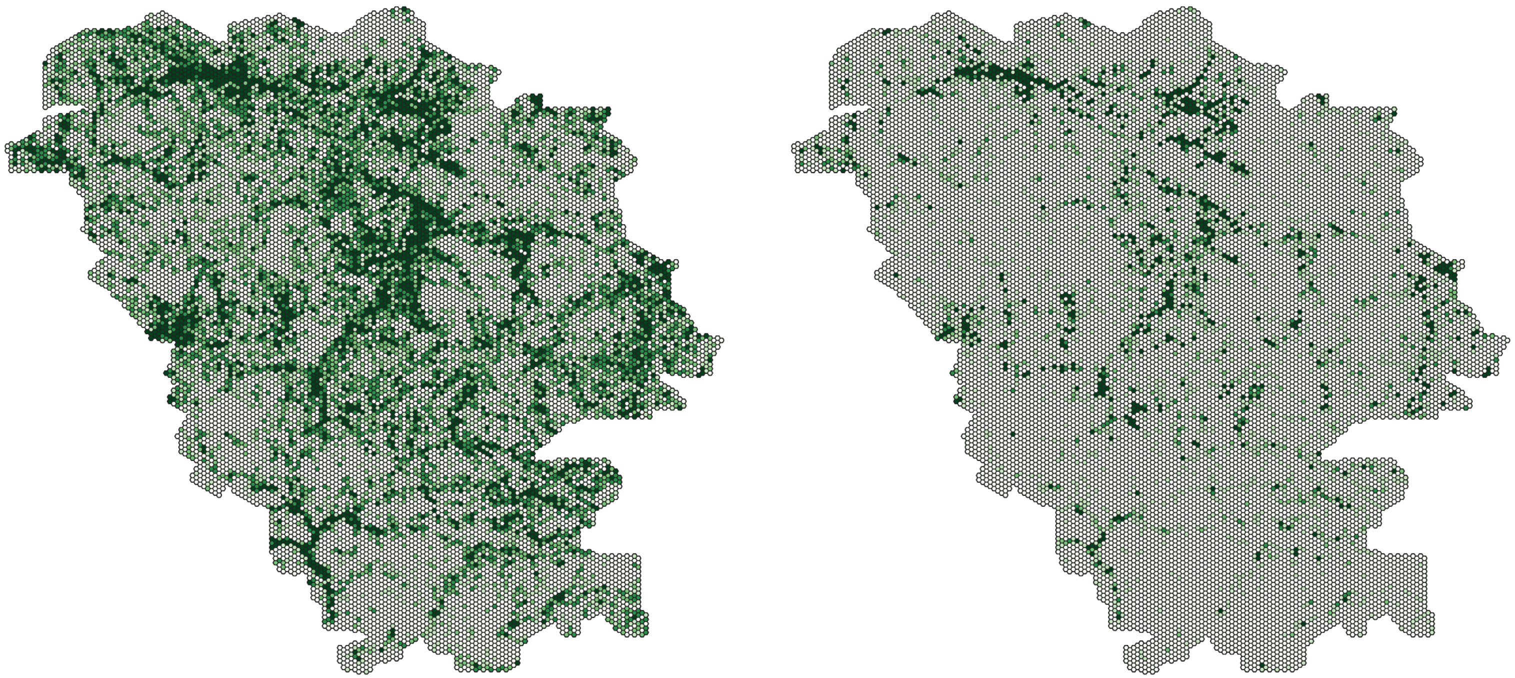
FIGURE C.3: *The spread and control of Prosopis at year 3.*



(a) *Absence of control method, t=4*

(b) *Control method implemented, t=4*

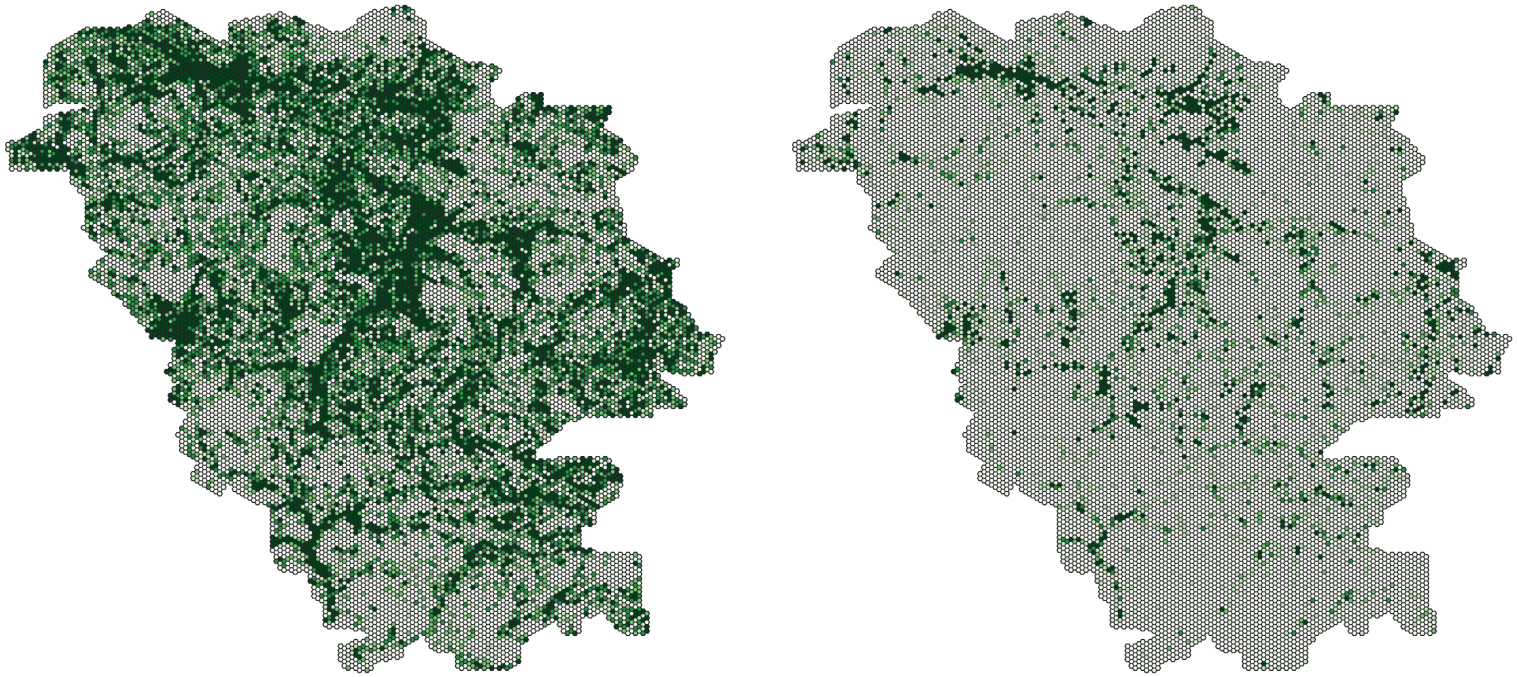
FIGURE C.4: *The spread and control of Prosopis at year 4.*



(a) *Absence of control method, t=5*

(b) *Control method implemented, t=5*

Figure C.5: *The spread and control of Prosopis at year 5.*



(a) *Absence of control method, t=6*

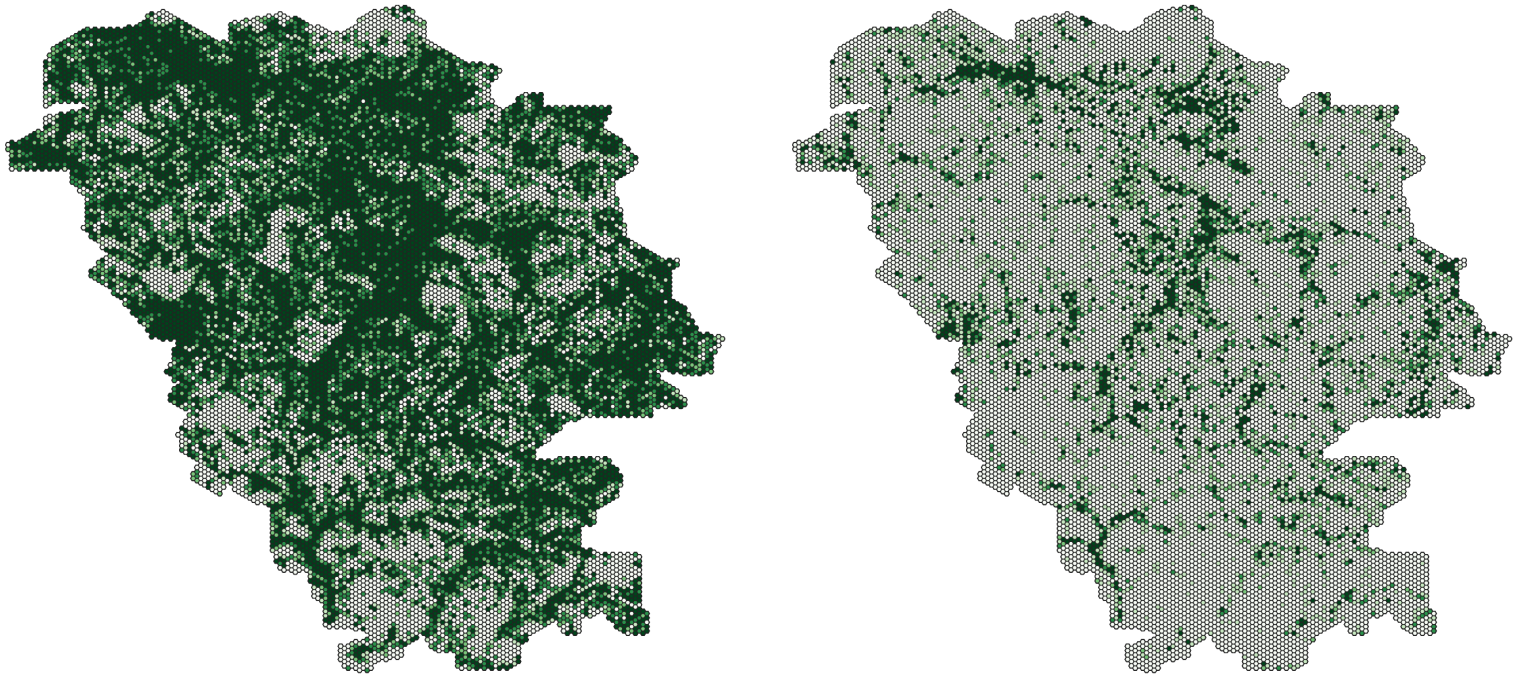
(b) *Control method implemented, t=6*

FIGURE C.6: *The spread and control of Prosopis at year 6.*

(a) *Absence of control method, t=7*

(b) *Control method implemented, t=7*

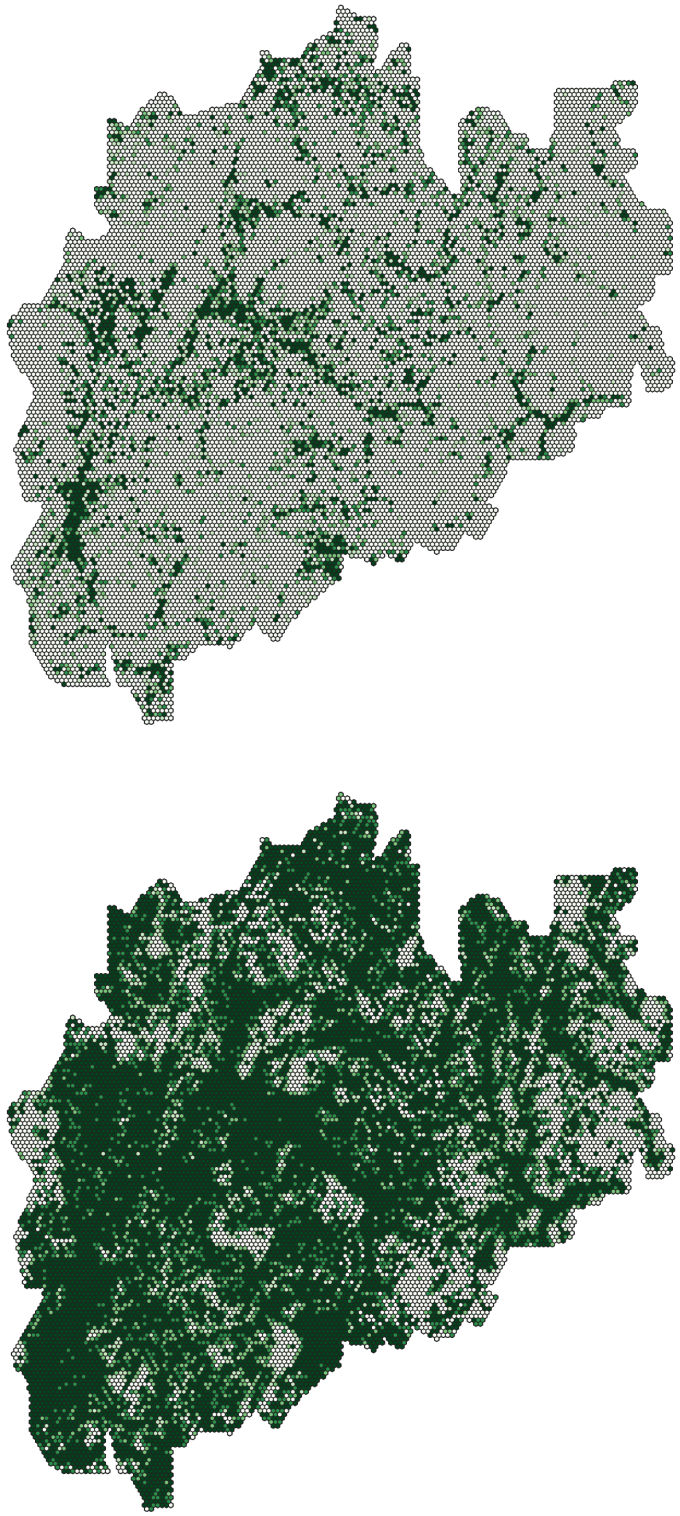
FIGURE C.7: *The spread and control of Prosopis at year 7.*



(a) *Absence of control method, t=8*

(b) *Control method implemented, t=8*

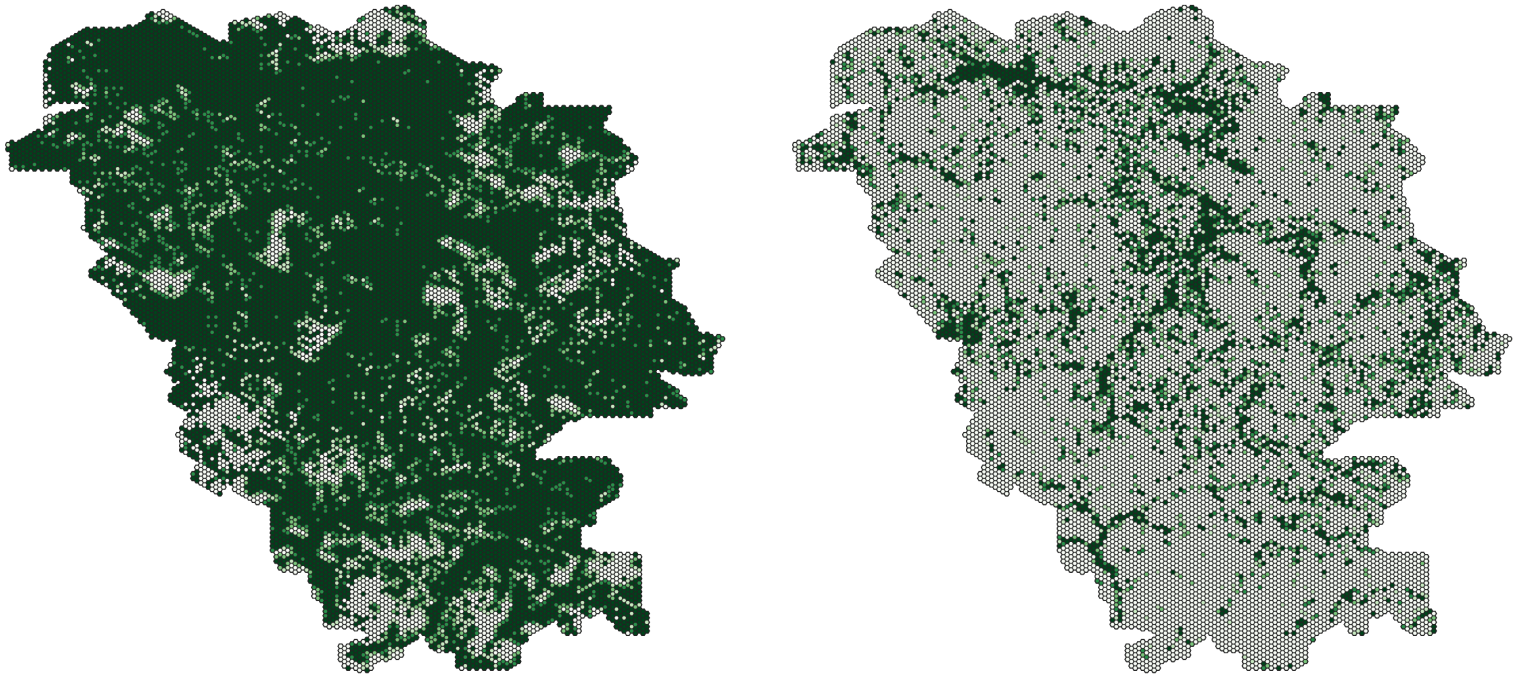
FIGURE C.8: *The spread and control of Prosopis at year 8.*



(a) *Absence of control method, t=9*

(b) *Control method implemented, t=9*

FIGURE C.9: *The spread and control of Prosopis at year 9.*



(a) *Absence of control method, t=10*

(b) *Control method implemented, t=10*

FIGURE C.10: *The spread and control of Prosopis at year 10.*