MASTERS THESIS

Remote sensing and prediction of invasive species

Author:
Gareth W. Hillston

Supervisor:
Prof. Rene Breton

Co-Supervisor:
Dr. Angela Harris

A thesis submitted to the University of Manchester
for the degree of Master of Philosophy

in the

Department of Physics and Astronomy in the School of Natural Sciences
Faculty of Science and Engineering

2023

Contents

List of Figures

List of Tables

Popular Abstract

Gareth W. Hillston

Remote sensing and prediction of invasive species

Globally, invasive plants present a major challenge to environments, public health and economies. They are often very difficult, if not impossible, to eradicate once they had gained a foothold in a region. If it is possible, containment or removal are usually very costly and time consuming. it is therefore of great advantage to understand these plants well and if possible, to be able to preempt their propagation.

Computer models are often used to simulate the behaviours of invasive plants, however they are often complex require a variety and abundance of data. In this thesis we examine the use of epidemiological disease models as a more simple alternative, requiring less data be gathered. We experimented with different implementations in order to achieve an accurate simulation of the population dynamics of the plant Parthenium in Pakistan. Models such as this may offer an easy way to examine the behaviours and influencing factors of other invasive plants, in order to better manage and contain them.

In our experimentation we were able to develop an epidemiological model which was capable of accurately simulating the population fluctuations of Parthenium throughout the year. The model requires relatively fewer sets of data than a species distribution model which might typically be used, and it able to offer insights into the characteristics and behaviours of Parthenium.

Declaration of Authorship

I, Gareth W. Hillston, declare that this thesis titled, “Remote sensing and prediction of invasive species” and the work presented in it are my own. I confirm that:

- This work was done wholly while in candidature for a research degree at this University.
- No part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Copyright Statement

Acknowledgements

This thesis was something of an ordeal, made infinitely more manageable by an entire village of wonderful people who helped me in a myriad of ways. My parents, Lynn and Stuart, and my brother Cameron have been supportive in every, and have given me every advantage they could to get me here. My supervisors Rene Breton and Angela Harris have been endlessly patient and helpful throughout the entire process and lead me to fill in the numerous prior gaps in my scientific knowledge. My advisor Mike Garrett was kind and approachable when I needed help. There are a whole host of fellow postdocs in JBCA who offered me support, reassurance and advice throughout. The sense of community during the pandemic lockdown was incredibly helpful to me, as I started an alien, new process, entirely online. Special thanks go to Joshua Hayes and Emma Alexander who helped me a great deal both directly and indirectly with my research and writing, as well as Laura Driessen for creating the handsome template used for the styling of this thesis. A great many other friends have offered me help and support in too many ways to count. I have to thank Elena Cavallero and Chris Pedge for daring their reputations to vouch for me as references, to get me onto the course in the first place. Thanks also to all the friends who offered me their advice, encouragement, homes, and most importantly, food. Many already mentioned, but importantly also Saboor Qureshi, Charlotte Mason, Eleanor Farrant-Jayawant, James Farrant-Jayawant, Andy Gorton, Claudia Ivan and Matt Taylor. Finally I have to offer a special thanks to those friends who offered me important emotional support; Adora, Edith, The Cat, Charlie, Fish, Ferret and Grey Cat. This work wouldn't have been possible without everyone above, and for that I'm immensely grateful.

Chapter 1

Introduction

Invasive species are a global issue of growing concern, negatively affecting ecosystems, public health and economies (Büyüktahtakın & Haight, 2018). They have been an issue for as long as humans have been travelling between disparate regions, and have only grown more problematic with each leap in global inter-connectivity (Vitousek et al., 1997). Generally, "invasive species" are defined as those which are non-native to a region and are considered to have negative consequences for the ecosystem or human population (Rejmánek, 2000). Communities will seek to curtail the spread of invasive species by monitoring them, preventing their migration, and attempting to eradicate any populations which have gained a foothold in a new environment (Richards et al., 2012). These efforts are often difficult, expensive and ineffective, as large areas of terrain must be observed and small quantities of an invasive species can escape detection and rapidly proliferate. This then gives rise to the use of behavioural and environmental models to attempt to predict and preempt the movements of particular alien species (Dana et al., 2014).

Computer models are often used in invasive species prediction, with most falling under the umbrella of species distribution models (SDMs). These can be either correlative or mechanistic. Correlative SDMs model the species as a product of environmental factors, determining where it is or is not likely to be, based on its currently inhabited environment. Mechanistic SDMs directly model the behaviour of the species in question, to predict its preferred conditions and proliferation (Srivastava, 2019). While widely used, both have their advantages and disadvantages and both require a variety of data, such as plant behaviours or environmental data such as precipitation, temperature and humidity. It would be useful to be able to construct models with fewer additional data sets required. Looking to alternative forms of modelling may offer some solutions. It has been remarked that invasive species behave in many ways similar to disease epidemics (Strickland, 2015). The spread of diseases, largely based on proximity, but with some spontaneous emergence in new regions, and the fluctuations in population, bare striking resemblances to the behaviours of invasive plant populations (Strickland, 2015). Investigating the models used to simulate and predict the spread of infectious diseases could provide new insights into the spread of invasive species. Epidemiological models often

require fewer sets of environmental and behavioural data than invasive species models, while remaining useful in behavioural prediction. Epidemiological models can be generated using the rates at which infections occur and the period of infectivity, and therefore do not necessarily require other data sets regarding external factors or the mechanics of the disease in question.

To this end, we selected a case study on which to experiment with, to determine whether epidemiological modelling can applied to invasive plants, and yield results which can inform management decisions. We experimented using the data of the invasive plant Parthenium in Pakistan. Originating in the Americas, Parthenium has spread to 46 countries around the world, and presents a major ecological problem (Steve Adkins & Asad Shabbir, 2014). Each plant can spread tens of thousands of seeds and propagate quickly, with a new plant reaching maturity in as little as 30 days (Nguyen et al., 2017). It can also chemically inhibit the growth of other plants and out-compete them for resources (Belgeri et al., 2011). Parthenium seeds can cause serious respiratory problems in humans, and the plant's leaves, flowers and stem hairs produce a dry powder which can cause dermatitis (Sharma et al., 2013). The plant causes stomach irritation when ingested by humans and can spoil meat and milk of animals which ingest it. Parthenium has also been known to cause famines due to its suppression of crops (Towers & Mitchell, 1983). It is a major problem in Pakistan, having a detrimental affect on public health and agriculture, and has now spread to three of the country's provinces (Steve Adkins & Asad Shabbir, 2014). The problem is also amplified by other factors such as climate change, which will benefit the especially hardy and fast-growing Parthenium and worsen the famines it already causes with drought (Steve Adkins & Asad Shabbir, 2014). Eradication of an invasive species once it has gained a foothold is often highly difficult, if not impossible. The problem is worsened as the communities largely affected are rural farming communities which often do not have the requisite time or resources to properly tackle the issue, and the government does not have sufficient resources to attempt a full eradication.

This thesis aims to leverage satellite imagery to create a model based on epidemiological models, which can simulate the spread of Parthenium in Pakistan, in order to assist in its containment and eradication. Our work builds upon previous efforts in Fennel & Breton (2023) to predict the presence of Parthenium in Pakistan using the multi-spectral imaging data from the Sentinel-2 satellite (Fennel & Breton, 2023). In this paper, Fennel et al. built a machine learning classifier for data from the European Space Agency satellite Sentinel-2 data, validated with a large-scale field campaign to gather ground truth data. In this work, we use archival data dating back several years in conjunction with the already-trained Parthenium classifier to create temporal maps of how Parthenium has spread as the basis to inform an epidemiological-type model simulating where it might spread next. We experiment with different models, specifically the potential use of epidemiological population models, as applied to invasive species.

Chapter 2

Literature Review

2.1 Invasive species

In the context of ecology and biology, the term "invasive", as applied to plants or other organisms, is not used in a consistent way. Generally, "invasive species" are defined as those which are non-native to a region and are considered to have negative consequences for the ecosystem or human population (Rejmánek, 2000). Often, non-native species have little, if any, positive effects on their new habitats. In some cases, even native species may be referred to as invasive if other factors, such as the eradication of a natural predator, cause them to reproduce or spread at such rates that they upset balances previously held in the biological or geological systems they interacted with (Rejmánek, 2000).

Invasive species are a global issue, affecting environments, cultures and economies all over the world (Büyüktahtakın & Haight, 2018). Whilst they are pervasive, each invasive species is different in the way it disrupts, unique to the ecosystem it is invading. Caused largely by human intervention, invasive species have been transported or introduced almost everywhere humans have been, intentionally and accidentally (Vitousek et al., 1997). However, invasive species did not spread at their current alarming rates for most of the thousands of years of human global travel and trade. It is only more recently, since the industrial revolution, that the rate of spread has significantly increased. This is owing to the increased ease and frequency of global travel and trade, and modification of the environment (Hulme, 2009). One infamous example is that of Japanese knotweed. A species native to Japan, Korea and China, it has since been reported across Europe and in the USA (Richards et al., 2012). It was introduced into the UK in the 1850s as an ornamental plant due to its ease of growth and exotic appearance (Hollingsworth, 2000). However, this ease of growth meant it was able to spread rapidly across most of the country and remains a widespread problem, difficult to remove and often out-competing native plants.

Invasive plants specifically, are often able to out-perform native species for resources, spread more easily or actively attack them, which can profoundly disrupt local ecosystems (Ehrenfeld, 2010). Invasive plants can alter soil composition and nutrients, or lead to the use of herbicides harmful for native species (Weidenhamer & Callaway, 2010). They

can also have significant economic costs ranging into the billions, by reducing yields from fishing, livestock or crops, or due to the high costs of attempts at management or eradication (Marbuah et al., 2014; Pimentel et al., 2001; Lovell et al., 2006). These reduced yields can also lead to food shortages in the areas affected.

There are multiple frameworks to describe the progression of an alien species' invasion, however a useful one for our purposes is that of describing plants overcoming a series of barriers, as detailed in (Richards et al., 2012, and Fig. 2.1). The process is described as follows:

A) The initial barrier is geographical. Overcoming this represents the plant entering the area in question, but being unable to maintain this new population.

B) This represents the plant overcoming environmental factors, i.e. being able to survive in the local environment.

C) The barriers to regular reproduction, having passed this threshold the plant is now able to sustain its population in the new environment and is considered 'naturalised'.

D) The obstacle of dispersal beyond the initial area of invasion.

E) After this point, the plant will be able to resist or adapt to biotic and abiotic features of its new environment; this allows for the invasion of already disturbed ecosystems. After this point the plant is considered invasive, as it can leave the initial point of ingress and take hold, in order the spread further.

F) Passing the final barrier, the plant will be able to overcome the defenses mounted by undisturbed, mature ecosystems.

Management of invasive species may largely be divided into three parts: prevention, surveillance and control. Prevention strategies focus on preventing the introduction of a new invasive species to an area, using methods such as quarantine and inspection of organic products on entry. Surveillance aims to detect new incursions at an early stage, such that they can be eradicated before they gain a foothold. Control encompasses all efforts at mitigation or eradication of an invasive species after it has established itself (Büyüktahtakın & Haight, 2018).

Prevention methods may include the blacklisting of certain species for import (Hulme et al., 2018), inspections and analysis of imported products which carry a risk of contamination with alien species, i.e. grain shipments carrying seeds or plant specimens carrying pathogens. Imports may also be treated with heat, cold, radiation or pesticides in order to neutralise any contamination without need for time-consuming inspections. Surveillance methods may include 'sentinel plots', sample plots of land which are monitored regularly for the introduction of foreign species. Visual surveys may also be conducted from the air or ground of a more general area (Büyüktahtakın & Haight, 2018). Finally, control measures may include the use of chemical agents (Simberloff, 2001), biological

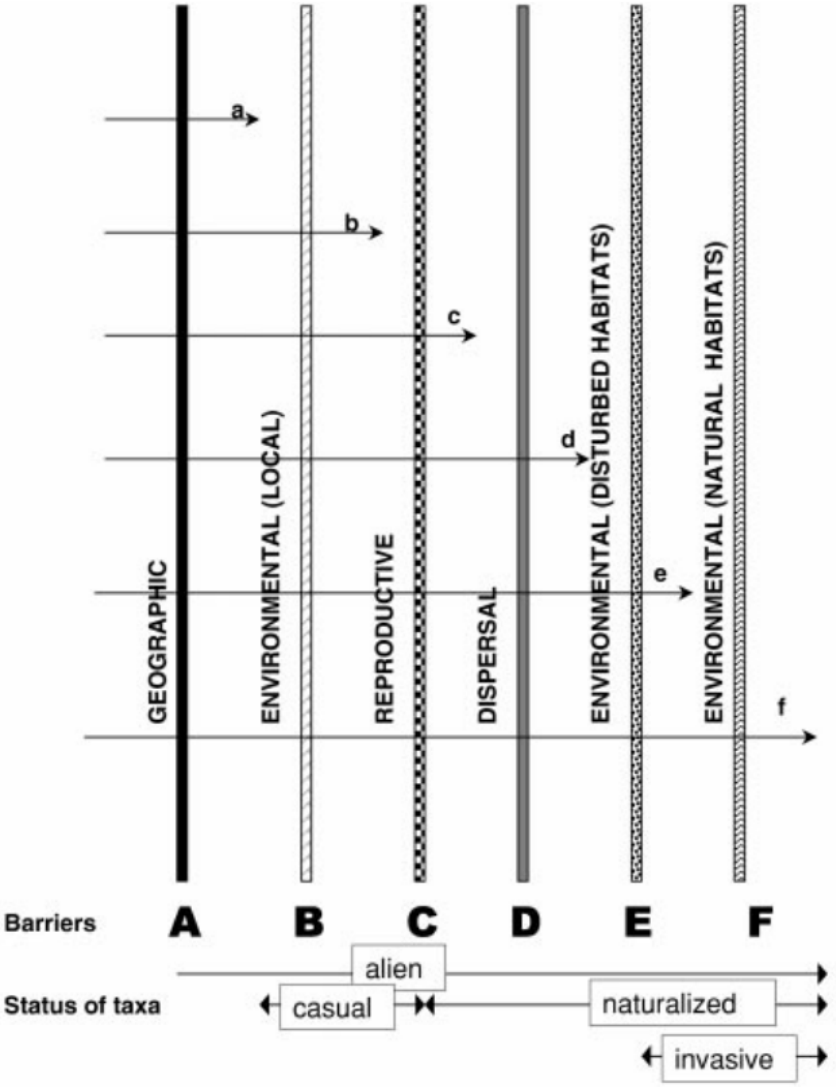


FIGURE 2.1: The invasive spread model, as described by Richards et al. (2012). The figure displays the 6 progressive barriers to be overcome for an invasive species. A - geographic - moving into the area in question. B - local environmental - being able to adapt to the local environmental challenges. C - reproductive - being able to reproduce consistently in this new environment. D - dispersal - the species can propagate outside of it's initial invasion point. E - disturbed environmental - disturbed or weakened ecosystems can be invaded. E - natural environmental - mature, natural environments can be invaded. Credit: Richards et al. (2012).

controls like herbivorous insects (Shabbir et al., 2013), or mechanical and manual methods of physically removing the plants (Simberloff, 2003).

These methods all come with limitations, however. Blacklists may be easily circumvented and can be difficult to enforce (Hulme et al., 2018). Inspections are time-consuming and imperfect, especially regarding contaminants which are difficult to visually identify or differentiate, such as small seeds (Steve Adkins & Asad Shabbir, 2014). The blanket treatment of imports will have varying efficacy depending on the qualities of any individual species, and may not catch contaminants outside the import products themselves, such as in or on the vehicles themselves (Hulme et al., 2008). Additionally, such screening methods will not cover any vectors of transport which are not directly human, such as plant movement due to disrupted or changing environments, or animal hosts.

Surveillance is dependant on key areas being under observation, or the observed areas being well representative. This is not always possible, as invasions are often highly localised initially, until surpassing the dispersal stage in the invasion model. Additionally, new invasive species may behave differently to previous ones, and therefore the observation areas which proved effective for previous species may not work for future ones (Hauser & McCarthy, 2009).

Control generally suffers from the most issues, as at this point the plant has usually established itself in the local environment and becomes difficult to disentangle from the local flora. Chemical methods often adversely affect native plants and animals, (Myers et al., 2000). Biological methods are difficult to direct and can present risks to native wildlife, as with chemical methods. Use of non-native species can also result in yet further invasive species (Myers & Cory, 2017). Both chemical and biological methods tend to be expensive. Mechanical and manual removal are often very time-consuming, and manual removal is often required where invasive plants are mixed with desirable plants, especially crops. Additionally, many plants are able to regrow from roots, so the removal has to be very thorough and any small amount missed can quickly re-establish a population (Steve Adkins & Asad Shabbir, 2014). These methods also potentially negatively affect the health of the workers involved, if the plant is in some way toxic or allergenic (Rao et al., 1977).

Deciding where to allocate resources between prevention, surveillance and control, and all the various strategies therein, is a complex and difficult problem. Therefore, "operations research" (OR) models are developed to decide the most efficient and cost-effective combination of management methods (Born et al., 2005; Buhle et al., 2005; Dana et al., 2014). Incursions of alien species are often not recognised until they are already well established and therefore expensive to manage. Often the greatest hurdle in tackling outbreaks of invasive plants is finding an effective form of management, and gaining the required policy, coordination and resources to tackle what are often national or international problems (Head, 2017). It is therefore important to understand, and if possible predict the behaviours of invasive plants as best as possible (Pluess et al., 2012) in order to make the best use of limited resources.

2.2 Modelling of invasive plant species

Distribution modelling of invasive plants can be useful to better understand their behaviour, and therefore be able to effectively manage them. It can also be useful in order to create predictive models, which allow for the spread of a plant to be predicted. This is either to preempt the plant's spread, or to prepare to manage future outbreaks. In this thesis, when we refer to species distribution models, we are discussing models specifically pertaining to the plant's behaviour, and not OR models, which help determine the best combination of management techniques to tackle an invasive species. The simulation models considered in this thesis often inform OR models, in terms of the efficacy of different management techniques. Adding to the methods of devising management strategies is part of the aims for our research.

The most common form of model used to simulate invasive species behaviour is a species distribution model (SDM) Srivastava (2019). SDMs use environmental data to predict where a species will be found across time and space. They fall into two main categories: correlative SDMs, which model the species' distribution as a product of environmental factors, and mechanistic SDMs, which model the species physiology to predict the conditions they will thrive in. Correlative SDMs need existing presence and absence data for the species under observation. While presence data may be available, absence data often are not, which can pose problems for the accuracy of the model. As these models orient themselves around where the species is and is not at present, they can only model its "realised niche", i.e. the types of environments it is in now. They are not capable of projecting what other environments it may be able to spread to. Conversely, mechanistic SDMs need exhaustive experimental data on the traits of the species in order to make accurate predictions. However, because mechanistic models model the plants behaviour, and not its current environment, they are able to project its "fundamental niche", i.e. everywhere it could possibly grow.

There has comparatively less research in modelling the fluctuations in population size and dispersal over time, using spatial data to generate diffusion models. One area of interest is the similarities between the spread of infectious diseases and invasive species. The models used to simulate infectious disease populations can be of use in monitoring invasive species, especially as they tend to focus more on populations and individuals, rather than broad the likelihood of presence in environments (Strickland, 2015). For an example we can look at the work of Jones (2017) in modelling the populations of red and grey squirrels, and the transmission of disease in Anglesey, Wales. They use a series of differential equations to model the change in portions of the population, i.e. red squirrel, grey squirrel, infected, uninfected. This method was able to accurately model the population dynamics which had been observed in the past, and could be used to suggest optimal methods of containment and eradication. It was also expanded with geospatial data about land types and correlations with squirrel populations to be able to generate a 2D density map and model the diffusion of populations in the space. These disease models have not been applied extensively to invasive species, and especially not widely

to invasive plants.

2.3 Parthenium

In this thesis we focus on an invasive plant called Parthenium, and analyse its presence in Pakistan as a test case for using epidemiological models to model invasive species. Parthenium is a particularly harmful plant as it is damaging to other plants, people and animals, and is considered a serious issue for agriculture and public health. Parthenium (*Parthenium hysterophorus L. (Asteraceae; Heliantheae)*) is a herbaceous plant which is highly specialised in out-competing other plants for space and resources. It has proven toxic to both humans and animals, detrimental to crop yields and exceedingly difficult to remove.

Typical adult specimens reach a height usually not exceeding 1.5 m, although they have been known to grow to 2.5 m in height. It typically grows annually, germinating in the spring, reaching maturity in as little as 30 days and continuing to flower and produce seeds for 6-8 months until late autumn. Each plant will produce hundreds of flowers in groups of 5 (rarely 6-8) (Fig. 2.2b), each containing 4 or 5 achenes (seeds) (Fig. 2.2d) enclosed in a straw-coloured fruit (Fig. 2.2c), with two laterally attached sterile florets (Steve Adkins & Asad Shabbir, 2014).

One Parthenium plant can produce up to 12,000–28,000 seeds per year (Nguyen et al., 2017). The seed florets are very light, and as such can act as buoyancy aids to help seeds spread both on the wind and down water courses (Mao et al., 2019). Additionally the seeds, once removed from their fruit, are very small, $\sim$ 2 mm, and as such are easily caught and spread by vehicles, farm machinery and animals (Parsons et al., 2001). They also often contaminate shipments of grain or other produce. It is thought that Parthenium has mainly spread internationally via contaminated grain shipments or vehicles. Soil seed banks can range from 1000s to 100,000s per m^2 (Nguyen et al., 2017).

The spread of Parthenium is in part dictated by its preferred growing conditions. It generally grows best in warm, wet conditions. Its aerial parts do not tolerate frost well, and most plants die during cold winters, although after mild winters some may regrow from their stem bases.

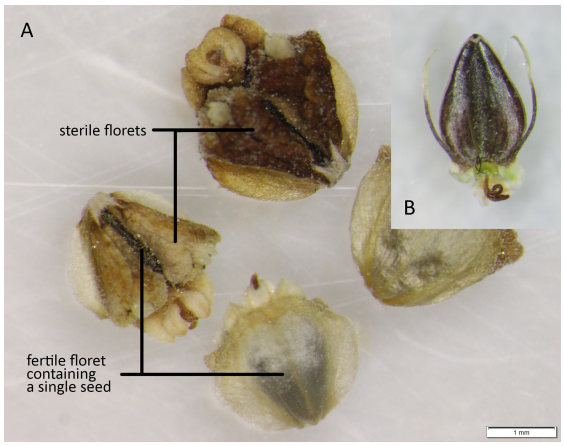
It is most commonly found in areas with poor ground cover, so naturally distributed areas, wastelands, or cleared ground (Steve Adkins & Asad Shabbir, 2014). It also thrives in maintained and irrigated areas, such as croplands and orchards, generally at the fringes (Finch, 2023). Parthenium is able to germinate and grow rapidly, which in combination with the production of allelopathic secretions, means it is able to hinder and surpass the rate of growth of neighbouring plants (Belgeri et al., 2011).

2.3.1 Negative effects

A high proportion of people develop severe allergic reactions to the plant or pollen. It can cause dermatitis, asthma and hay fever in people with no prior history of these conditions. People can be affected either by direct contact or indirectly through airborne

(A) Mature *Parthenium Hysterophorus* on a road-side, credit: CABI Pakistan

(B) *Parthenium Hysterophorus* seed pods, credit: CABI Pakistan



(C) Close up of *Parthenium Hysterophorus* seed pods, credit: (Mao et al., 2019)

(D) Close up of *Parthenium Hysterophorus* seeds, credit: CABI Pakistan

FIGURE 2.2: Images showing Parthenium as a whole, and its seed pods and seeds.

particles (Towers & Mitchell, 1983). There have been some reports in India of deaths due to respiratory conditions. There is no known treatment or desensitising therapies. Symptoms can be managed, but as it is often found in poorer areas, remedies are not always readily available (Steve Adkins & Asad Shabbir, 2014). Additionally, strong medications cannot be used in conjunction with heavy machinery, which is frequently necessary in the daily lives of the farmers often affected, or which is necessary to clear the Parthenium.

Ingestion by livestock is known to give their milk a bitter taste and their meat an unpleasant flavour (Tudor et al., 1982; Nigatu et al., 2010). Parthenium has been shown to seriously reduce the yield of affected crops, and reduce the carrying capacity of pasture land (Tamado et al., 2002; Shabbir et al., 2012). Time and effort are also wasted in trying to contain it. Parthenium can be host to other diseases and pests, and interfere with natural ecosystems (Sharman et al., 2009; Remadevi & Sivaramakrishnan, 1996).

Parthenium therefore poses serious risks to the health, food supply and livelihoods of communities affected, making it a major cause for concern where ever it is found and a priority for eradication.

2.3.2 Management

Unfortunately, Parthenium proves very difficult and costly to remove in practice, and often requires a combination of methods to be fully eradicated. Removal at the surface level is often not enough, as Parthenium is able to regrow from cut stalks or roots, and its seeds can lay dormant for 7 or more years before germinating (Adkins, 2010).

Some countries have attempted to create habits and practices which prevent the spread of the plant in the first place. Parthenium has been given serious designations in some countries, with some restrictions on moving plants between regions. The strongest measure thus far have come from Australia, where it is a serious problem. In Australia, farmers are required to report Parthenium to a special service which monitors it. Machines and vehicles coming from infested areas must be thoroughly cleaned at designated cleaning stations. Some have found it useful to adjust stocking rates and rotational timings between grazing events (Parsons et al., 2001).

In areas where labour is cheap and people are not aware of the adverse health effects, hand pulling and ploughing are used to remove the plant (Rao et al., 1977). Ploughing may be effective, if done in the week before replanting. Some countries do not consider physical management to be cost-effective due to the size of Parthenium infestations, its effect on workers and ability to regrow from cut stalks or roots. Fire has been studied in Australia as a management technique, but was found to just clear the ground for Parthenium to grow more easily (Vogler et al., 2006).

Due to costs, chemical management is often only useful in protecting expensive crops. It is usually only feasible for easily accessible areas, and therefore less effective in wastelands and rangelands, where Parthenium is often most prevalent. It is also necessary to catch it in the pre-flowering stage, and to revisit the site for as long as seeds in the soil may re-emerge, i.e. at least 7 years (Steve Adkins & Asad Shabbir, 2014). Some

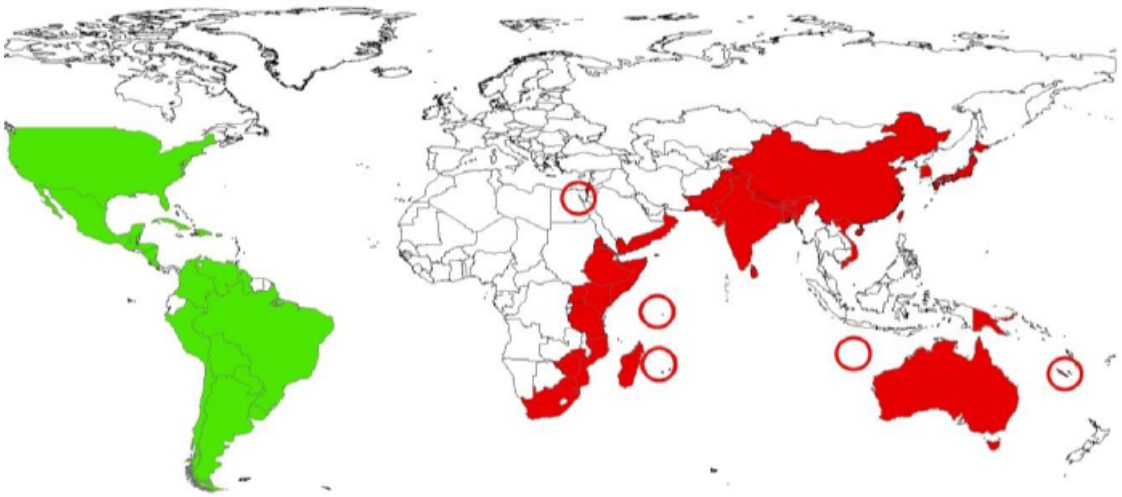


FIGURE 2.3: A map of the world, showing the spread of Parthenium. It is considered native in the countries highlighted in green, it is considered invasive in the countries highlighted in red. Credit: (Shabbir et al., 2012).

herbicides have been effective, but it has already garnered resistance to some forms of herbicides (Vila-Aiub et al., 2008).

Some plants and insects have proven effective in suppressing the growth of Parthenium. A combination of plants, insects, pathogens and other biological agents have shown to be very effective in suppressing growth and even seed production. Using native insects and plant pathogens has the added advantage of not further disrupting the local environment with foreign species. The approach considered best is a mix of all the aforementioned management techniques, as Parthenium can often survive one method alone (Steve Adkins & Asad Shabbir, 2014).

2.4 The Spread of Parthenium Globally and in Pakistan

Pakistan has a major problem with Parthenium, and difficulty in quantifying the issue and managing it, due to the country's mountainous terrain and limited resources on a governmental and individual level. Parthenium is native to regions surrounding the Gulf of Mexico. Parthenium is originally thought to have spread throughout 11 different countries, from the Southern USA and further South to Bolivia, Brazil and Northern Argentina (Steve Adkins & Asad Shabbir, 2014). It is believed that Parthenium was first carried outside the Americas in shipments of grain, which is a major vector of spread for it. It often grows in and around crops, its seeds are small enough to avoid detection and numerous enough to have a high chance of being transported on humans, animals or vehicles (Mao et al., 2021). Currently Parthenium is known to be found in 46 countries and territories worldwide (see Fig. 2.3) and is not effectively contained in most of them. Parthenium is especially prevalent in warm, wet countries (Mao et al., 2021).

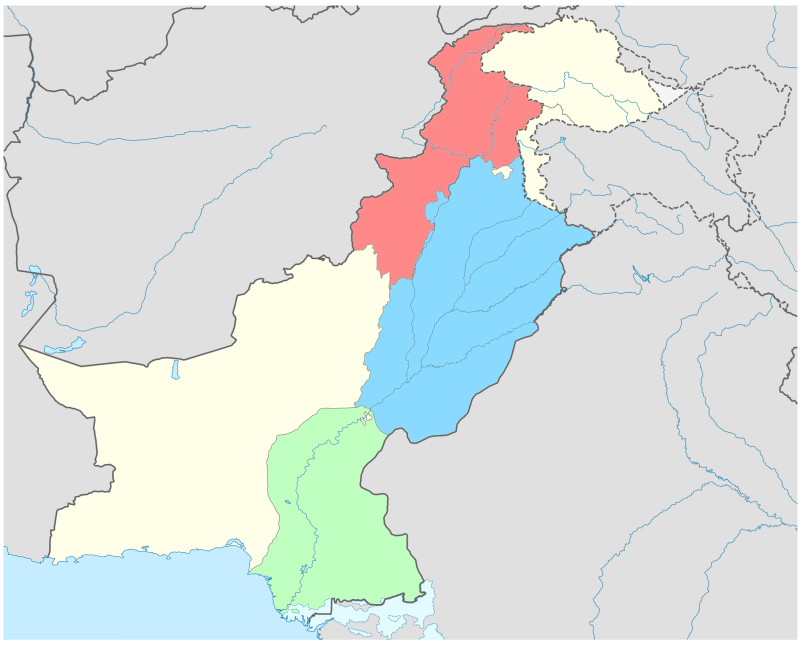


FIGURE 2.4: Map of Pakistan with Khyber Pakhtunkhwa in red, Punjab in blue and Sindh in green. Image adapted from: Wikipedia - Administrative units of Pakistan - NordNordWest - own work, using United States National Imagery and Mapping Agency data World Data Base II data.

Pakistan is located in South Asia, within the latitudes 24°N - 37°N and longitudes 61°E - 76°E. It ranges from mountain ranges in the North, through the Indus River valley in the North and East, through more arid land in the South, the Balochistan Plateau in the West and the Arabian Sea in the South. While Pakistan consists of several large provinces, and several smaller territories, in this thesis we are mainly concerned with 3 provinces: Punjab in the North East, bearing the majority of the plant, Sindh in the South East, and Khyber Pakhtunkhwa in the North West (see Fig. 2.4). The varied topography lends itself to a tropical to sub-tropical climate. The temperature ranges from 12°C to 20°C in the winter and 19°C to 35°C in the summer. The monsoons in the summer and Western disturbances in the winter deliver most of the country's annual rain; 45% and 31%, respectively (Adnan et al., 2018).

Parthenium was first reported in Pakistan in the Gujarat district of Punjab Province in 1980 (Shabbir et al., 2012). Parthenium is suspected to have been introduced to Pakistan in grain shipments contaminated with seeds, driven in from India, which has suffered the infestation since at least the 1950s and possibly ever since the late 1800s (Steve Adkins & Asad Shabbir, 2014). Parthenium was originally only reported in Northern Punjab Province in the 1990s, as of 2012 it had spread into Southern Punjab and threatened other regions. Parthenium has since spread across Pakistan, especially the Punjab and Khyber Pukhtunkhwa Provinces, where it is reported in many districts, and possibly to Sindh Province. It is often found having spread between urban areas and along roadsides, far from other known populations of Parthenium. Given its small and numerous seeds, and its known propensity for becoming lodged in vehicles and invading in grain shipments,

it seems likely that it is often spread by vehicles travelling from invaded regions. There are also reports of higher densities of Parthenium near plant nurseries, which implies its seeds may also spread in the soil or seeds of plant products. It is largely found in wasteland, but has caused significant disruption to agriculture, causing significantly reduced yields in some crops (Shabbir et al., 2012). It is likely to have mainly spread in Punjab due to the region's warm, wet climate and fertile soil, which provide ideal growing conditions. These conditions also lead to an abundance of agriculture in the area, which leads to irrigation and fertilisation which often benefits Parthenium at field margins, or sometimes directly competing with the crops themselves.

2.5 Remote Sensing of Parthenium

Here we will look at some existing work in observing Parthenium, using remote sensing imagery from the Sentinel-2 satellite. Sentinel-2 is an ESA-operated mission comprised of 2 polar-orbiting satellites offering multi-spectral band imaging across most of the Earth's land surface[1]. Each satellite has a re-visit period of 10 days, meaning a 5-day re-visit period overall, in cloud-free conditions. It captures images in 13 bands of the electromagnetic spectrum at either 10, 20 or 60 m per pixel resolution, in the visible light and infra-red wavelengths.

Arogoundade et al. (2020) investigated the use of multi-spectral imaging (MSI) in supplementing an SDM - MaxEnt - in identifying the presence of Parthenium in the KwaZulu-Natal province of South Africa. They gathered data using the Sentinel-2 satellite and added this to the environmental data used by the MaxEnt algorithm to predict Parthenium presence. Ground truth data was gathered at 274 locations, as 10 m x 10 m quadrants, measuring the presence of Parthenium. This was then used to train and test the MaxEnt model. Distance to roads, precipitation and elevation were all used as environmental factors, in addition to the multi-spectral data from Sentinel-2. The resulting model was found to have a high degree of accuracy, higher than without the Sentinel-2 data. Elevation was found to be the most important environmental factor for presence of Parthenium. It also found that the most important indicators for the model were Sentinel-2 red edge band 705 nm and the red edge index, which indicates the presence of healthy vegetation.

Kganyago et al. (2017) addresses the issue of the "curse of dimensionality" in machine learning classification, as applied to remote sensing data on Parthenium in KwaZulu-Natal. The "curse of dimensionality" is where the number of observations is less than the number of variables being measured, which reduces the resulting classification's accuracy. Kganyago et al. (2017) sought to compare different methods for reducing the number of variables by reducing the number of different bands needed to accurately identify Parthenium from multi-spectral data. They found that a hybrid approach yielded small subsets with higher classification accuracies than other techniques. They also found that

[1]Technical documentations here: https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/overview

the most important bands for distinguishing Parthenium from other plants commonly found with it, were the red edge bands at 685 and 707 nm, as well as a near infrared band at 1115 nm and short wave infrared bands at 1971, 1982, 1990, 1966, 2003, 2005 and 2013 nm.

In a study based on the same dataset, Kiala et al. (2020) compared different methods of automatic identification of the presence of Parthenium using remote sensing imagery of KwaZulu-Natal. Several common methods of presence predictions – stochastic gradient boosting, extra-trees classifier, random forest and k-nearest neighbour – were evaluated against a relatively new method, Tree-based Pipeline Optimization Tool (Olson et al., 2016). It was found that this new method yielded more accurate predictions than the other methods at an overall rate of 88.15%. The most important bands for Parthenium identification were visible blue, green and red (bands 4, 3 and 2), short wave infrared (band 11), near infrared (bands 8 and 8a) and red edge (band 6).

In a study in Pakistan, Finch (2023), data on the presence of Parthenium was collected via roadside surveys of 10m square areas across a random selection of 10 km x 10 km grid cells, across the districts of Punjab, Khyber Pukhtunkhwa and Sindh. Crop type, road type, land type and proximity to other Parthenium were found to be significant in determining the abundance and likelihood of the presence of Parthenium. The authors found proximity to be the most important factor in the presence of Parthenium, which correlates with the current understanding of Parthenium and other invasive plants.

Building upon the previous study, Fennel & Breton (2023) trained a machine learning algorithm to predict the presence of Parthenium based on its spectral signature, using imaging data from the Sentinel-2 satellite and using the previous study's data as ground truth. Remote imaging data was acquired from the Sentinel-2 satellite via online archives. The machine learning classifier selected was an Extremely-randomised Tree classifier (ETC), due to its resistance to over-fitting, computational scalability and clear trade-offs between bias and variance in training the model. The variance of accuracy of predictions can be mitigated by averaging over large groups of predictions, and high levels of bias away from accurate average predictions can be tolerated in classification without high error rates (Geurts et al., 2006). The trained model was run against a reserved test set of images, and was found to have a high degree of accuracy. The model was then tested with new ground truth data from a region South of Islamabad with a mixture of land cover types. All but 2 of 39 points predicted to contain Parthenium did contain it, although these 2 points were recently ploughed. The model was then applied generally. The Southern Khyber Pukhtunkhwa and Northern and Western Punjab were shown to have the most Parthenium, with decreasing prevalence through central Punjab and little to none in Sindh and the South and East of Punjab.

Chapter 3

Gathering Data

In this chapter we describe the structure and facets of the data used in the course of our research, and the methods used to attain it. Firstly we address our main source of data, the ESA archives of Sentinel-2 imaging data. Additionally we detail the features of this data; the multiple wavelengths of imaging, and classifications included with images. Next, we explain the work done in Finch (2023) and Fennel & Breton (2023), upon which we have built much of this thesis. Specifically, we used the survey work done in Finch (2023) to inform our understanding of Parthenium and its preferred habitats, and the classifier built in Fennel & Breton (2023) as our base truth for the presence of Parthenium. Finally we explain the use of K-means classification and conduct our own classification of the Sentinel-2 data, in order to judge the spatial distribution of land types in the areas being observed. This all provided us with both data, and context for that data, in order to draw conclusions about the behaviour of Parthenium and create our simulations of that behaviour.

3.1 Pre-processing Sentinel-2 data

To gather our multi-spectral data we used the archives of the Sentinel-2 satellite (see Section 2.5). Data were searched and downloaded from Google Earth Engine [1], as individual *.tif* files for each date, with one file representing the 10 bands for that image area and date. We used images processed to the 2A level, which is corrected to match cartographic geometry, and remove atmospheric distortions, with associated classification data. These data are available from December 2018.

Using Sentinel-2 allows us to survey the entire country with ease, at regular intervals (i.e. 1 satellite pass per 5 or 10 days), at a resolution of 10, 20 or 60 m, depending on the band[2], although we do not use the 60 m resolution data. Note that when using the 20-m resolution data we used bi-linear interpolation to match it to the image size of the 10-m resolution data. This is useful as it enables us to gather data on all affected areas without

[1] https://code.earthengine.google.com/
[2] https://sentinel.esa.int/web/sentinel/user-guides/sentinel-2-msi/resolutions/spatial

Sentinel-2 Spectral Bands			
Band Number	Area of EM Spectrum	Wavelength (nm)	Resolution (m/pixel)
1	Blue	443	60
2	Blue	490	10
3	Green	560	10
4	Red	665	10
5	VNIR	705	20
6	VNIR	740	20
7	VNIR	783	20
8	Low SWIR	842	10
8a	Low SWIR	865	20
9	Low SWIR	940	60
10	Mid SWIR	1375	60
11	Mid SWIR	1610	20
12	High SWIR	2190	20

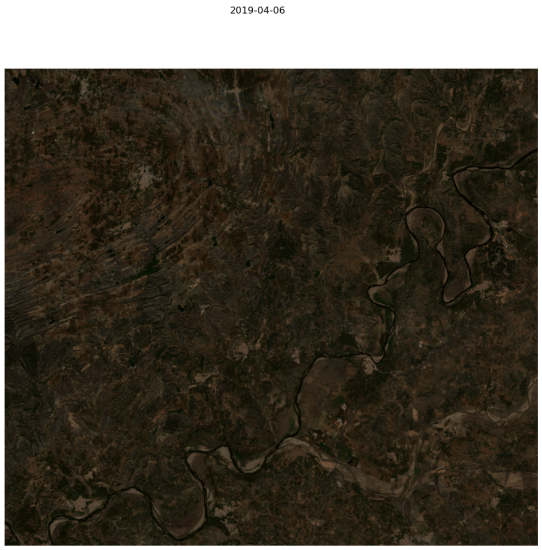
TABLE 3.1: The various spectral bands at which Sentinel-2 can capture images. VNIR stands for Very Near Infra-Red, SWIR stands for Short-Wave Infra-Red.

the expense and time entailed in an on-the-ground data gathering operation. Parthenium resembles many other ground cover plants and its effects may not be immediately apparent to those unfamiliar with it. Therefore, reporting of Parthenium presence from individuals in affected areas is unreliable. Satellite imaging allows us to get a clearer idea of the total spread of the plant.

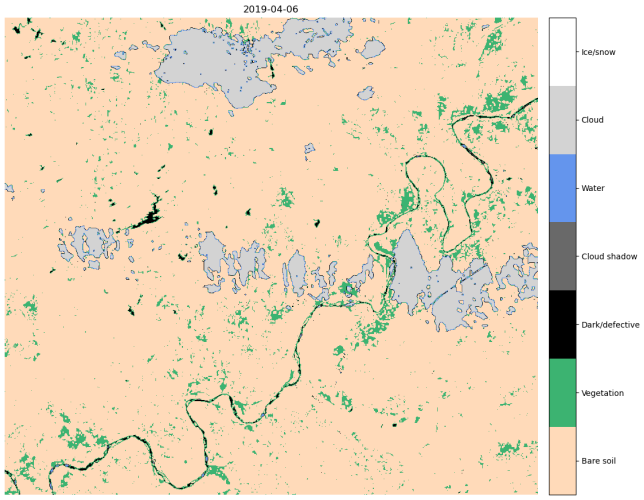
This multitude of imaging at different wavelengths is very useful, as different land types can be identified by their "spectral signatures". Different types of land, such as bare soil, water or foliage may reflect light in the same way at certain wavelengths but not at others. For example, soil, vegetation and rocks all reflect in band 10 (1375 nm) at around 35-45%, but look different in other bands. Telling apart various land types using a single band would be difficult if not impossible, however we can make use of the multiple bands provided by Sentinel-2 in order to differentiate them.

We gathered spectral reflectance measurements from December 2018 to January 2022. The Sentinel-2 reflectance data are provided with a Scene Classification Layer (SCL), which provides classifications of pixels at 20 m spatial resolution. We used nearest-neighbour interpolation to spatially re-sample the classification layer from 20 m to 10 m resolution, to match the spatial resolution of our surface reflectance data. The SCL contains eleven classes including four different classes for clouds and six classes for shadows, cloud shadows, vegetation, soils/deserts, water and snow. These classifications are designated by the ESA's scene classification algorithm and can be one of the following 11 types, as shown in Table 3.2.

In some instances, heavy cloud cover may render images less useful, due to the low proportion of surface detail visible, therefore it is important to know the level of cloud



(A) True colour image of the area under observation.



(B) A representation of the Scene Classification Layer for the observation area. This representation is combining all cloud layers together.

FIGURE 3.1: A comparison of a true colour image of the observed area and a visualisation of the Scene Classification Layer of the same area. Both are from the same data set, gathered on 06/04/2019.

cover, to filter out potentially unrepresentative images. Sentinel-2 uses a custom algorithm for classifying cloud pixels[3], described below. In each step after 1a, if a high chance of cloud cover is indicated, the probability is unchanged. For a middling chance, the existing probability for that pixel is multiplied by the probability calculated by the current index being tested. If the chance is low, then the probability is set to 0.0, the pixel is considered cloud-free and any remaining steps are skipped.

1. (a) Band 4 (red) is checked first to classify cloud-free pixels. With less than 0.07 reflectance, the pixel is considered cloud-free, otherwise it is considered cloudy. These are represented as a probability of clouds of 0.0 or 1.0, respectively.

   (b) Next is the Normalised Difference Snow Index (NDSI), which compares bands 3 (VNIR) and 11 (SWIR). Cloud-free is considered to be a score below -0.1, potentially cloudy is between -0.1 and $+0.2$, and a high probability of cloud is above 0.2.

2. Any pixel which has recorded snow in the last 10 years is tested for snow in 4 consecutive tests on bands 2, 3, 8 11. Band 12 is then checked to remove false reports of clouds at the borders of snowy regions. If, after these checks, the snow probability is above a given threshold, then the pixel is marked as snow. Otherwise, the algorithm proceeds to the next step.

3. The Normalised Difference Vegetation Index (NDVI) is determined for each pixel, by comparing bands 8 and 4. This gives an indication of the presence of green, leafy vegetation.

4. Bands 8 and 3 are then tested for plant senescence (signs of aging), which can result in higher reflectivity.

[3]See the description at https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-2-msi/level-2a/algorithm

Sentinel-2 Scene Classification Layers	
Number	Layer
1	Saturated / Defective
2	Dark Area Pixels
3	Cloud Shadows
4	Vegetation
5	Bare Soils
6	Water
7	Clouds low probability / Unclassified
8	Clouds medium probability
9	Clouds high probability
10	Cirrus clouds
11	Snow / Ice

TABLE 3.2: The Scene Classification Layers available with Sentinel-2 data. The Number indicates the order of inclusion in the SCL list, the Layer describes the classification of that layer.

5. Bands 2 and 11 are checked in 2 separate indexes for bare soil or bright water.
6. The ratio of bands 8 and 11 are compared to check for desert landscapes composed of rocks and sand.
7. An optional filtering can be applied on cloud layers to clear up misclassifications at high contrast class boundaries.

After general areas of cloud are determined, some specific tests are run. Cirrus cloud detection is then run on band 10, checking the level of water vapour absorption at that wavelength. If the reflectance is greater than the clear sky threshold (0.012) and less than the thick cloud threshold (0.035), then it is classified as thin cirrus clouds. Finally, cloud shadows are identified by determining the likely areas of shadow from cloud position and elevation, and Sun orientation, as well as detection of darker areas via neural network.

Unfortunately, the classifications are not perfect and there are some obvious areas where it lacks. For instance water is often identified at cloud edges, where this is often only temporary due to rain. Additionally, areas marked as low or medium probability of cloud are often clear, and masking them out would remove large swathes of data; see 3.2b This may cause issues when we are using the SCL to mask the image. It is necessary to mask out cloud cover, so that it does not interfere with the surface reflectance data and skew our analysis. However, as the SCL has the aforementioned issues this sometimes results in water being masked out as well.

We tried different combinations of masking, drawing from different types of clouds, as listed 7-10 (Fig. 3.2), and masking out water altogether. However, as we are committed to using the existing trained Parthenium classifier, we are best served by using the same masking criteria used by that classifier. From the layers listed in Table 3.2, the vegetation (4), bare soils (5) and low cloud probability (7) layers will be kept, and the rest will be masked out.

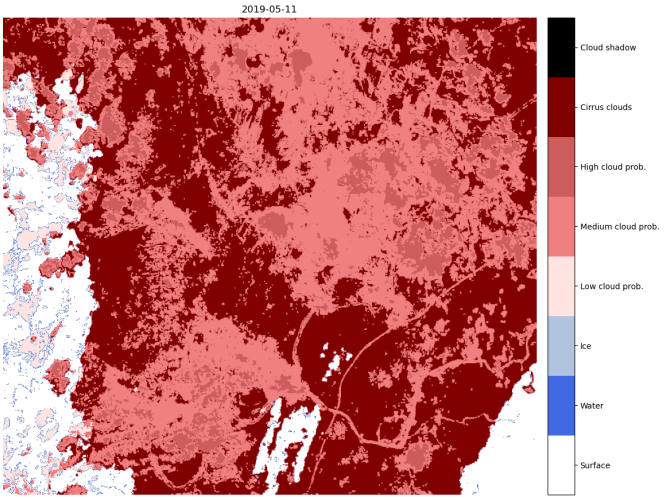
3.2 Predicting the presence of Parthenium

In our research, we made extensive use of the model developed in Fennel & Breton (2023) to predict the presence of Parthenium across Pakistan. In this project, a machine learning algorithm was trained to predict the presence of Parthenium based on its reflectance spectral signature, using imaging data from the Sentinel-2 satellite. For the classifier presented in Fennel & Breton (2023), the ground truth data was used from another study, Finch (2023), wherein they surveyed various locations in Pakistan for the presence of Parthenium.

3.2.1 Surveying Parthenium

In Finch (2023), Parthenium presence data was collected via roadside surveys across a random selection of 10 km x 10 km grid cells, across the provinces of Punjab, Khyber

(A) True colour image of the observation area, with what appears to be light cloud cover across the whole image.



(B) A visualisation of the different levels of cloud probability, from the SCL. Note that although in the true colour image, it appears to mostly not be obscured by cloud, in the SCL the image is mostly covered by low and medium probability cloud layers.

FIGURE 3.2: A comparison of a true colour image of a cloudy day and a visualisation of the cloud layers of the SCL. Both are from the same data set, gathered on 11/05/2019.

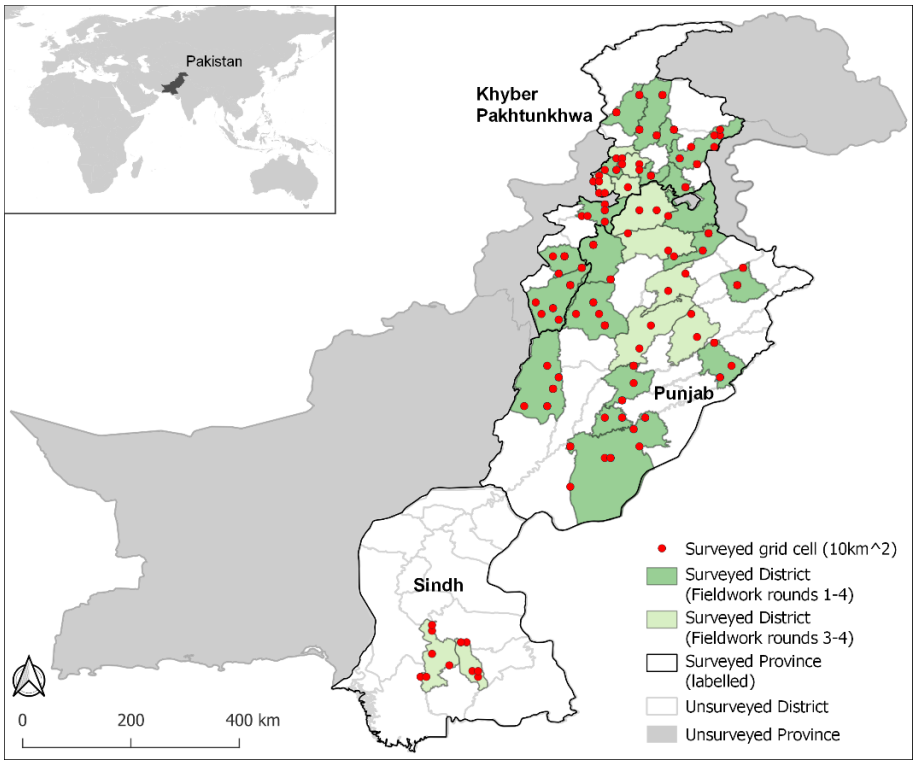


FIGURE 3.3: Map showing fieldwork survey locations across Pakistan. Red dots indicate surveyed locations. Dark and light green areas indicate districts which were surveyed in rounds 1-4 and 3-4, respectively. Credit: Finch (2023)

Pukhtunkhwa and Sindh. Twenty such cells were selected for each province, with cells re-selected if the original selection comprised >80% urban area, bare area or water. If a cell was not accessible by road, the nearest accessible cell was selected in its place. Data was collected in 4 field campaigns: December 2018, March 2019, June 2019 and October 2019. The first two rounds were conducted in Punjab and Khyber Pukhtunkhwa, the third and fourth rounds were extended to include additional districts in these two, and two districts in the Sindh Province. With each round of fieldwork, 50% of previously visited cells were chosen for revisit, with the remaining 50% being newly chosen. In each selected cell, surveys were conducted every 2 km along the road, where it was safe to stop. Approximately 20 locations were surveyed per cell, with locations where fewer measurements were taken being supplemented with additional data from the nearest possible cell(s). Each survey location observed an area of 20 m x 20 m, noting the road and habitat types, presence and abundance of Parthenium, and presence, type, growth stage and irrigation of crops. To increase the amount of available ground truth data containing positive identification of Parthenium, extra land plots from all 3 provinces were manually selected where there was a higher than 80% coverage of Parthenium.

Road types were found to have a strong effect on the presence of Parthenium, with larger district and national roads having a higher chance, local roads middling and unpaved roads and paths the lowest chance. Habitat type also had a significant effect; it was most common in grasslands, parks, gardens and cemeteries and least common in

forests. In crop lands Parthenium was most common in fallow or bare fields and crop margins, it was also much more likely to be found in irrigated fields than non-irrigated ones. Parthenium was less likely to be found in some crops like wheat and rapeseed, than maize, potatoes, sugarcane and others, and was never found in cotton and rarely in rice. Proximity also played a big role, with Parthenium being more likely to be found in grid cells adjacent to others which also contained Parthenium. Road type was not found to significantly affect the abundance of Parthenium, however crop type and habitat were. Abundances for crop types were lower in planted and bare soil fields, middling in fallow fields and highest in crop margins. Abundance was statistically lower in wheat, middling in maize and sugarcane and highest in potatoes or other vegetables. In other habitats, abundance was highest in grasslands and orchards and lowest near water.

The authors found proximity to be the most important factor in the presence of Parthenium, which correlates with the current understanding of Parthenium and other invasive plants. It was found to mostly spread by road, with bigger, busier roads carrying more of it, likely due to higher vehicle traffic. Transport via waterways was not found to be as significant, however this may be due to the surveys being limited to road-sides. Whereas previous studies have found Parthenium to be most common in "bare areas", this study found that there was a range of presence and abundance in different types of such areas. It was highest in crop margins, intermediate around buildings, roads and railways, and near the lowest in bare crop land. Bare areas had a middling probability of presence, but some of the highest abundances, which indicates that "bare areas" is a wide classifier and other factors in the environment are at play.

Cash crops, such as tobacco, sugarcane and potatoes, have been perceived by farmers as having a higher populations of Parthenium. It was found that there was a higher probability of Parthenium presence in these crops, therefore these perceptions were not purely due to the crops' increased value. There were no consistent trends between different methods of planted crops, and so planting methods were ruled out. Differences in the times when crops are planted also showed no consistent pattern in levels of Parthenium. Irrigated crops did show higher levels of presence and abundance, however almost all crops in the study were both rain-fed and irrigated, therefore this does not explain the discrepancies between crops. It is likely a combination of qualities, both of the crops themselves and their preferred environments, that dictate these differing issues with Parthenium.

The findings supported general management of the plant, rather than specifically focusing on areas of monetary value like crops. Farmers in the area are generally not tackling Parthenium in non-crop areas, despite this then providing it a foothold to spread from. It was suggested that resistant crops, such as wheat, rice and cotton, are grown along vectors of infection like main roads, as a buffer for the more susceptible crops. It was also recommended to look more into the economic thresholds of different crops; the level of Parthenium at which it starts to significantly affect the crop's yield. As it was found that crops vary widely in their susceptibility to Parthenium, they will likely have very different thresholds, and so crops with higher thresholds could be used to shield

those without.

3.2.2 Predicting Parthenium Presence

In Fennel & Breton (2023), Fennel et al. built upon the previous work to create a machine learning algorithm to predict the presence of Parthenium, based on its unique pattern of reflectance of different bands of light. We used this classifier to predict the presence of Parthenium in the areas under observation. While the machine learning model is not expected to produce 100% accurate classification of Parthenium's presence or absence (i.e. accuracy, precision and recall are not perfect), we detail below the authors' validation of the model, showing a high level of efficacy. Therefore we have assumed, for the purpose of this thesis, that the classifications produced by the model are always accurate. In reality, the inaccuracy of the classifier will introduce a source of systematic uncertainties which could be investigated at a latter time.

Remote imaging data was acquired from the Sentinel-2 satellite via online archives. This provides images in 10 wavelength bands of the visible and infra red spectrum, in 10 or 20 m per pixel resolution (see Sentinel-2 2.5). 20 m images were up-sampled to 10 m using nearest-neighbour interpolation and for each ground truth data point, the corresponding image was selected for that area, with dates ranging from $t - 35$ to $t + 7$, where t is the sampling date. The machine learning classifier selected was an Extremely-randomised Tree classifier (ETC), due to its resistance to over-fitting, computational scalability and clear trade-offs between bias and variance in training the model. The variance of accuracy of predictions can be mitigated by averaging over large groups of predictions, and high levels of bias away from accurate average predictions can be tolerated in classification without high error rates (Geurts et al., 2006). To allow for varying weather conditions and Sentinel-2's 5-10 day revisit period, a mosaic of images was created, from images in a 1 month period around the target date. Pixels not necessary to be categorised were masked out using the Scene Classification Layer for the image (see Sentinel-2 image processing 3.1).

The trained model was then run against a reserved set of images to test its efficacy. The model's performance was measured using its accuracy and F1 score. Accuracy being the percentage of total classifications which are correct. The F1 score is calculated as follows:

$$\text{F1 score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.1}$$

Where precision is the proportion of true positive classifications, amongst all true and false positives, and recall is the proportion of true positives amongst all true positives and false negatives, i.e. all actual positives. This gives us an average of precision and recall, which can give a better representation of the model's usefulness, especially in unbalanced data. For instance, where there are far more positive or negative observations.

In a single image test between classifications in the training and test sets, the F1 score was 0.90 and the accuracy was 0.95. Testing against the mean classification over a 7 week

window of 6 weeks before the target and 1 week after, the F1 score increased to 0.93 and the accuracy increased to 0.96.

A test region was selected 5 km South of Islamabad with a mixture of land cover types. The model was applied at 3 different times from November 2019 to March 2020. This showed a generally high level of Parthenium across the area. To validate these results, 39 randomly selected points were selected where the probability of Parthenium in March 2020 was given as between 0.10 and 1.0. All but 2 points contained Parthenium, although these 2 points were recently ploughed. For the validation points, the F1 score was similar to the test at 0.90, but the accuracy was lower at 0.82, due to the absence of control points without Parthenium.

The classifier was then applied to Khyber Pukhtunkhwa, Punjab and Sindh provinces between 25.72N and 72.41N, with Northern regions of Khyber Pukhtunkhwa excluded due to the interference of steep topography and high cloud cover, shown in fig 3.4. The data was aggregated to a 1 km per pixel scale, with values estimated for any pixels unable to be classified. Southern Khyber Pukhtunkhwa and Northern and Western Punjab were shown to have the most Parthenium, with decreasing prevalence through central Punjab and little to none in Sindh and the South and East of Punjab. Regions with higher abundance also registered more double and triple detections across the 3 classification times, indicating a longer presence throughout the year, compared to areas with lower abundance.

3.3 Land classification via K-means clustering

The SCL provided with the Sentinel-2 data includes 4 classes of surface measurement: vegetation, bare soil, water and snow/ice. As Parthenium cannot flourish on water, snow or ice, this leaves only two land cover classes upon which Parthenium could potentially be found. We therefore decided to create our own classifications, using a K-means clustering algorithm in order to identify the types of land cover present with greater variety of categorisations. This in turn will give use better information about the types of land cover preferred by Parthenium and enable us to separate out analysis or modelling by land type.

K-means clustering Cam & Neyman (1967); Hartigan & Wong (1979) is a method of grouping n data points together into k groups, or "clusters", based on proximity in a given feature space. In our case each data point will be a pixel in the reflectance data from a given date. The feature space is the reflectance value for each of the 10 bands for that pixel. K-means clustering is considered an unsupervised machine learning method, as it is given no supervision in the form of what clusters should be, merely a starting point. The algorithm itself determines the criteria for each cluster that minimise the difference between members of the same cluster, and maximise the difference between clusters. A data point is considered a member of a cluster if it is closer to the centre point, or "centroid", of a given cluster than of any other cluster. A point can therefore only belong to one cluster, and clusters do not overlap.

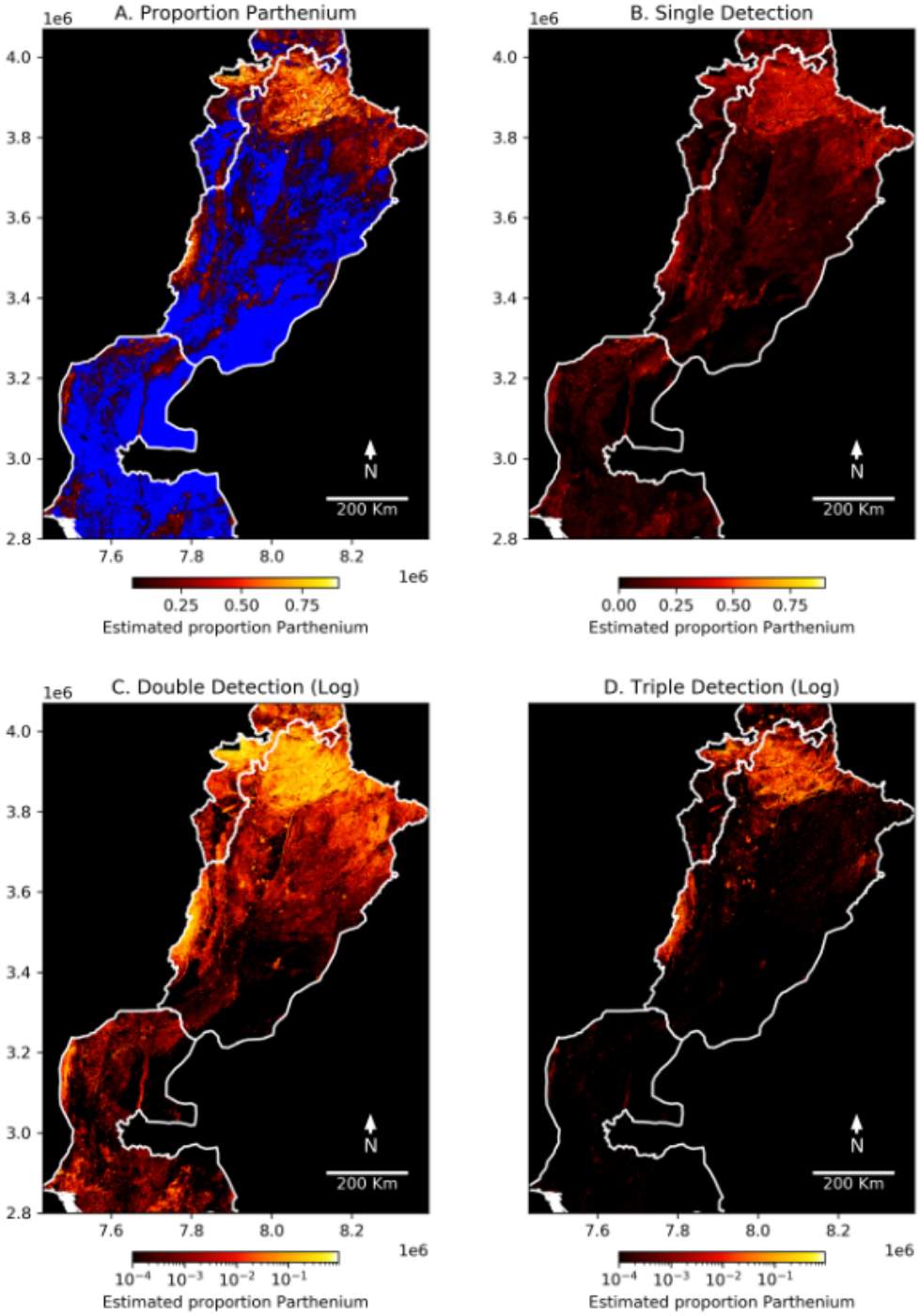


FIGURE 3.4: Aggregated Parthenium probability at 1 km resolution from remote sensing classification. Subplots show (A) the probability of Parthenium. Areas with aggregated probability below the classifier accuracy (p = 0.95) are coloured blue. Subplots B-D show probability of single (B), double (C) or triple (D) detection over the 3 time points classified. Credit: Fennel & Breton (2023)

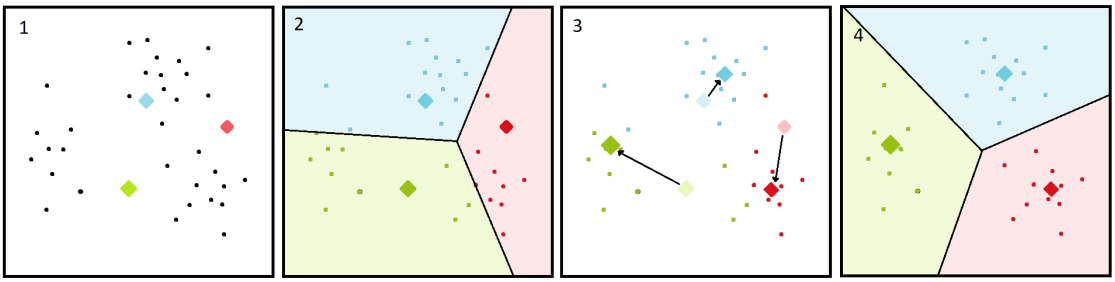


FIGURE 3.5: A representation of the process of K-means classification, as applied with 3 clusters. (1): Centroids are initialised in random positions. (2): Each point is assigned a centroid according to which is the shortest distance away. (3): The centroids are moved to the mean position of the points assigned to them. (4): Points are then reassigned centroids, based on the new centroid positions. Steps 2-3 are repeated until the centroid positions do not meaningfully change.)

To run, the algorithm is initially given a starting set of centroids, equal to the number of intended clusters, k. The algorithm will group all points according to which centroid they have the smallest squared distance from (inertia), then the centroids are updated to be the average of their new data points. This is done repeatedly, until no meaningful improvements can be made and it is considered to have reached the optimum values for the centroids. The data points are then classified into one of the k clusters, depending on which one they are closest to. See Fig. 3.5

The minimisation of inertia of data points can be described as follows:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} \left(\|x_i - \mu_j\|^2 \right) . \tag{3.2}$$

It is worth noting that this method will only cluster pixels based on their quantitative values, in this case their reflectance data. The algorithm ascribes no qualitative meaning, and so we will have to determine what shared characteristics each cluster represents, if any.

We trained an unsupervised k-means classifier on all 10 bands of reflectance data to identify 8 different classes of land type. 8 classes were arrived by observing that after 8, extra classes were duplicates or unused. Classes are demonstrated at 2 different times of year, with tentative labels assigned in figures 3.6 and 3.7.

3.4 Conclusion

In this chapter, we aimed to detail the data used in this thesis, and its significance. We can gather Sentinel-2 data starting from December 2018, at a resolution of 10m per pixel, in 10 wavelengths from visible light to short-wave infra red. Additionally, the data is provided with pre-generated classifications, mainly distinguishing between land, water and clouds. These classifications will be useful later in filtering out areas covered by water or cloud which are not relevant. The research conducted in Finch (2023) found that Parthenium is most common along road ways, with the chance of its presence increasing

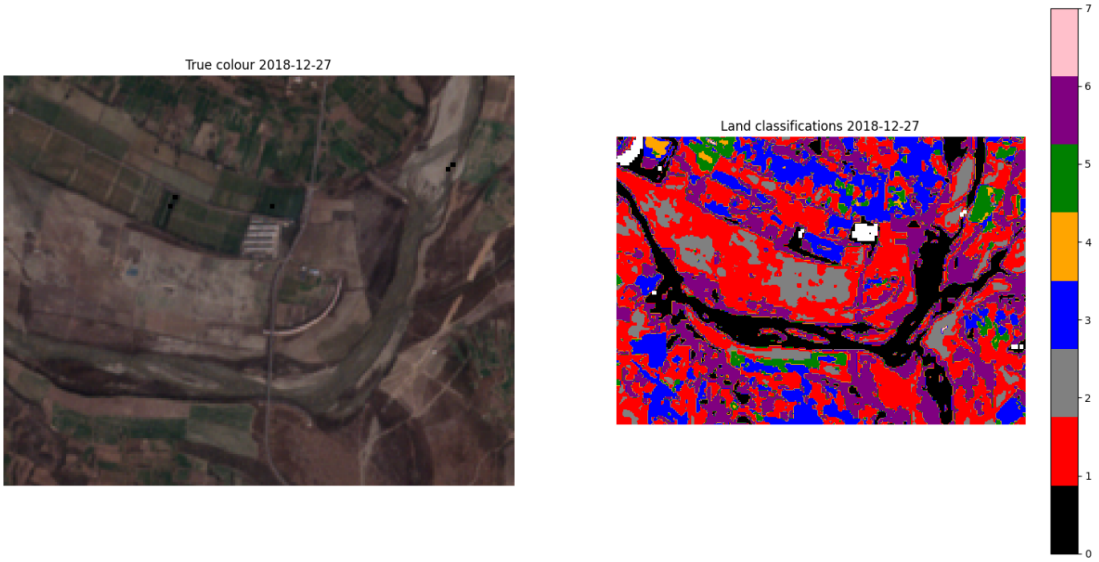


FIGURE 3.6: True colour (left) vs AI classification (right) December 2018. We can see class 0 correlating with the water of the river. 1 correlates with dryer, bare soil. 2 is found in bare soil patches and in the sand of the dry riverbed in the top-right. 3 is variable, but seems to correlate best with green vegetation. 4 and 5 are sparse, but mostly match vegetation. 6 is found along the riverside and vegetation, potentially wet soil. 7 is not present in this image.

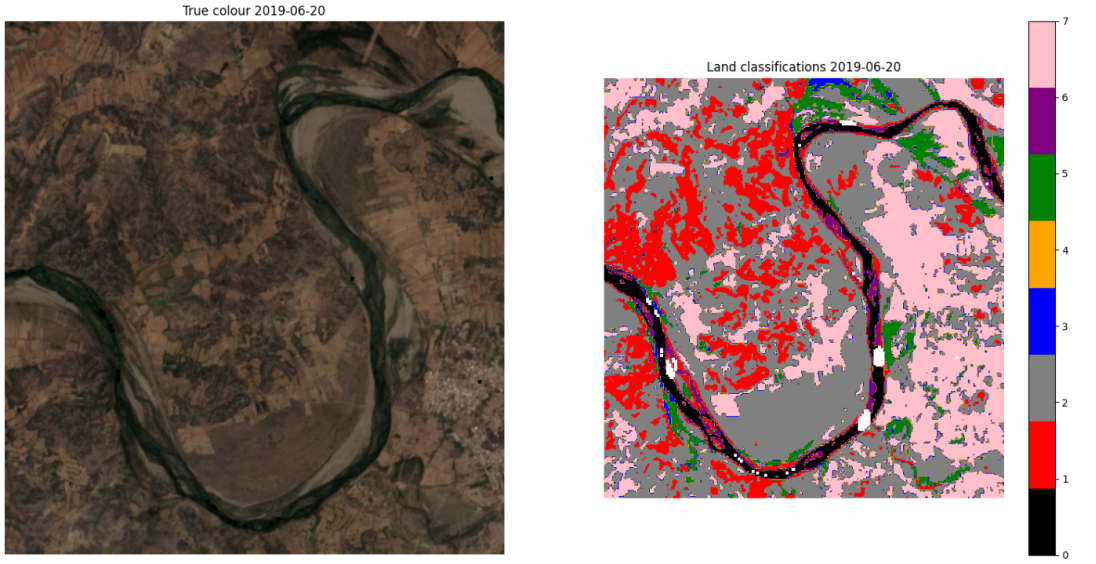


FIGURE 3.7: True colour (left) vs AI classification (right) June 2019. Class 0 again lining up with the water of the river. 1 and 2 correlating with dry soil and sand, now more prevalent in the summer. 3, 4 and 5, previously matching vegetation, are now sparse in the drier climate, mostly found near the water. 6 again seems to relate to wet soil, being found only on the banks and sand bars of the river. 7 is much more prevalent in June than December, mostly on sites that seem to be delineated as fields. It therefore likely represents dried out vegetation.

with road size. Parthenium was found to be most common in areas with little competition, such as crop margins or, wastelands. It was least common in areas with competition from other plants, such as crop fields and forested areas. Overall, the biggest factor in its presence was found to be proximity to existing patches of Parthenium.

In addition to the gathering of data, we detailed some evaluation of the data, to help us in our research. In Fennel & Breton (2023), the authors outlined a classifier they developed to predict the presence of Parthenium, using Sentinel-2 multi-spectral data. The resulting predictions were found to have a high degree of accuracy and precision. As we have access to the data used to train and validate this model, and similar data for surrounding regions, we can perform our own predictions using the model, as well as other analyses on the same areas. We will detail in Chapter 5 how we use this classifier to inform our simulations of Parthenium. Finally, we used a K-means classifier to categorise the different types of surface land in the observation area, in order to give better context to the spatial predictions generated by the Parthenium classifier. We categorised 8 classes, roughly correlated with water, dry soil, sand, green vegetation, two of agricultural vegetation, wet soil and desiccated vegetation. All together, these sources gave us the data we needed in order to make predictions about the presence and behaviour of Parthenium, and to be about to simulate its behaviour in the future.

Chapter 4

Data analysis and exploration

In this chapter, we expand on the data from the previous chapter, looking for context for our research and means by which to understand the overall patterns of plant life and Parthenium in the region. We outline the data used in our study, and the various methods used to analyse it. This includes the use of NDVI to get a sense of the patterns of leafy vegetation in the area, the seasonal trends of the land classifications we made previously, trends in Parthenium predictions, and comparisons between these different factors.

4.1 Data Set

We initially chose a sample area of approximately 20km by 20km for analysis, see Fig. 4.1. It is in the Punjab area of Pakistan, the area most affected by Parthenium. The area comprises a mix of urban, agricultural and wild areas; it also features a river running North East to South West and hills to the North West. This gives us a good mix of elevation, terrain and land use to observe Parthenium's behaviours in.

The data set runs from 17/12/2018 to 25/01/2022, with intervals between observation dates of usually 5 days, sometimes 10. This is with the exception of images which had to be removed from the data set. 26 images had to be discarded due to corrupted data, presenting as entirely blank images.

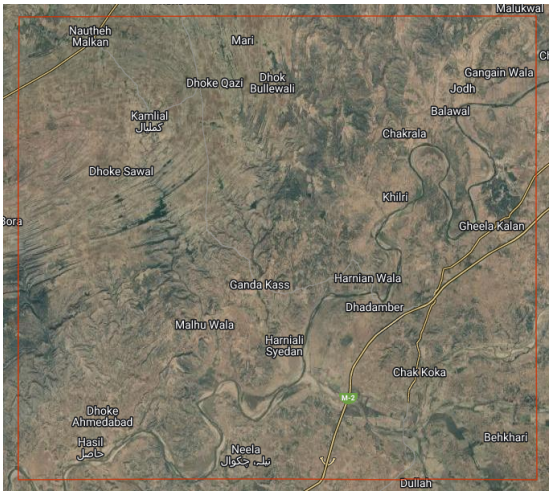
4.2 NDVI

Normalised Difference Vegetation Index (NDVI) gives us an indication of which areas are covered in leafy vegetation and which aren't. We can use this to gain an understanding of the general trends of plant growth over the year. Vegetation strongly reflects near infrared light, and absorbs red light. By measuring the difference between them, we attain a rough measure of the prevalence of green, leafy vegetation. NDVI is calculated as

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}}, \tag{4.1}$$

(A) A view of the study area, in red, within Pakistan. The area is within the Punjab province, which is heavily farmed and most affected by Parthenium.

(B) Map of the area under observation, bounded in red. Black names denote settlements, beige areas are land, blue areas are water, white and yellow lines are roads.



(C) True colour image of the area under observation, bounded in red. We can see more clearly here the striations of the hills running from West to North. We can also see that the map's depiction of the river is much wider than that which is visible, likely due seasonal variation and the river's flood plains being included.

FIGURE 4.1: Location of the study area within Pakistan, marked in red, as described by the following coordinates - [72.521,33.159],[72.742,33.159],[72.742,33.324],[72.521,33.324]. Credit: Google Earth Engine.

where NIR stands for reflectance measurements captured in the near-infrared spectral regions, and Red is reflectance measurements in the visible red region.

As the difference between NIR and Red is divided by their sum, this gives us a proportion of the total of both frequencies measured which is NIR, as an index between -1 and 1. A result close to 1 is a strong indicator of green vegetation, a result close to -1 likely indicates water, with a result around 0 suggesting little green vegetation and is indicative of bare soil, rock or urban areas. NDVI can give us an idea of the areas and times of year in which plants are flourishing. It thus can act as an approximation of active months for plant life in general.

Initially we want to observe the same area at regular intervals, at the same time each year, so that we can measure like-for-like. The first simulations will be run using yearly observations to give a consistent ecological state and time step, such that seasonal effects do not need to be taken into consideration. The Parthenium classifier we are using was trained largely on data from October, therefore it would be preferable to use data from October observations to train our model. However, as we only have access to data from 3 Octobers, using December may be preferable to gain more data.

We analysed the NDVI and Parthenium scores over the year and specifically in October and December to see if they offered comparable states in terms of plant activity and of Parthenium specifically. The hope was to use the October data for 2019-2021, being the likely most reliable, and December for 2018, being the closest available. As we can see from Figs. 4.2, 4.3 and 4.4, December has a peak just above 0, as with October, but is far more concentrated around that peak, as well as having a generally smaller spread. While, unfortunately, the two months are not exactly comparable, they were close enough that using entirely December data would suffice for our modelling, given the advantage gained from an extra data point.

Fig. 4.5 shows how the average NDVI values change over the course of each observed year, as well as the standard deviations represented as the blue area. It shows that there is a variation of 0.2-0.3 between the maximum, around September/October and minimum, around June/July. There is an increase in the standard deviation size around the higher values, and a marked decrease around the lower values. This indicates that during these months, one is more likely to find higher NVDI values and a greater range of values both higher and lower. This would correlate with with the increased likelihood of plant matter at the higher end of NDVI values, but also with an increased prevalence of surface water in cooler months. The months with a lower NDVI variance correlate with hotter months, when plants are less likely to be green and leafy, and more likely to be dried out and register a lower NDVI value. Additionally there will be less surface water at this time, so the breadth of possible scores will be reduced on both ends of the scale. All together, this suggests that Parthenium will be harder to recognise during there months, using the remote sensing methods we are employing. It also indicates that plant life and growth is hampered in these hotter, drier months, and that the plants spread may be more pronounced in the cooler, wetter months.

In order to better understand the spatial distribution of plant presence and growth

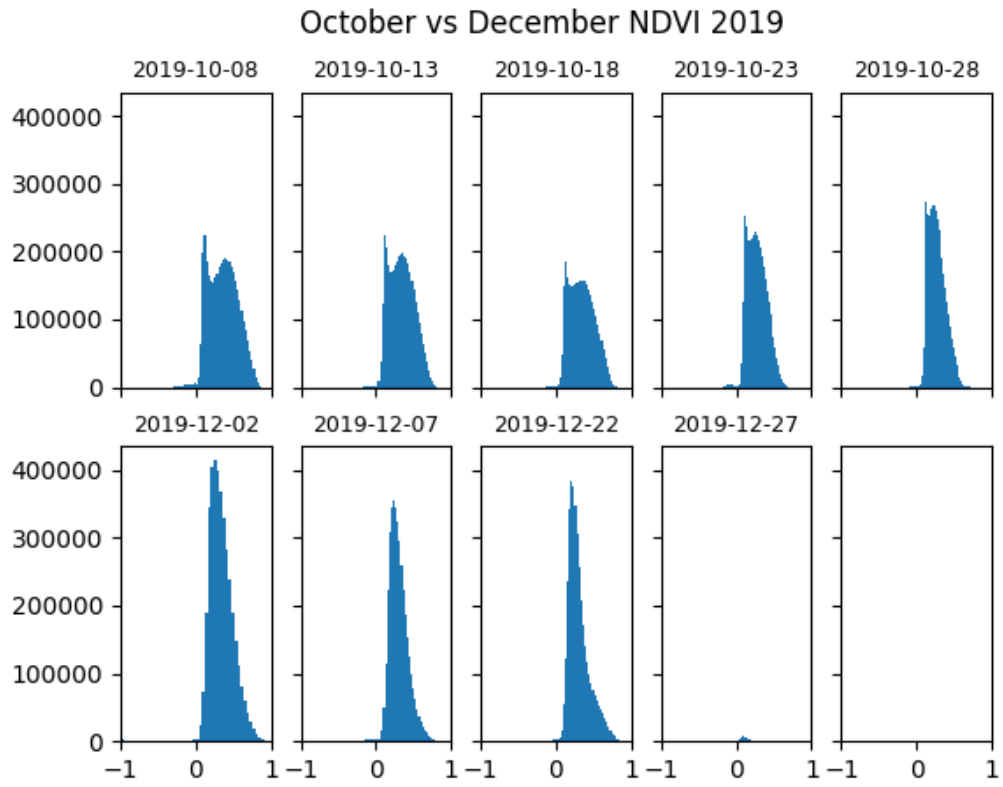


FIGURE 4.2: Distributions of NDVI values for October (above) and December (below) 2019, for the dates available in each month.

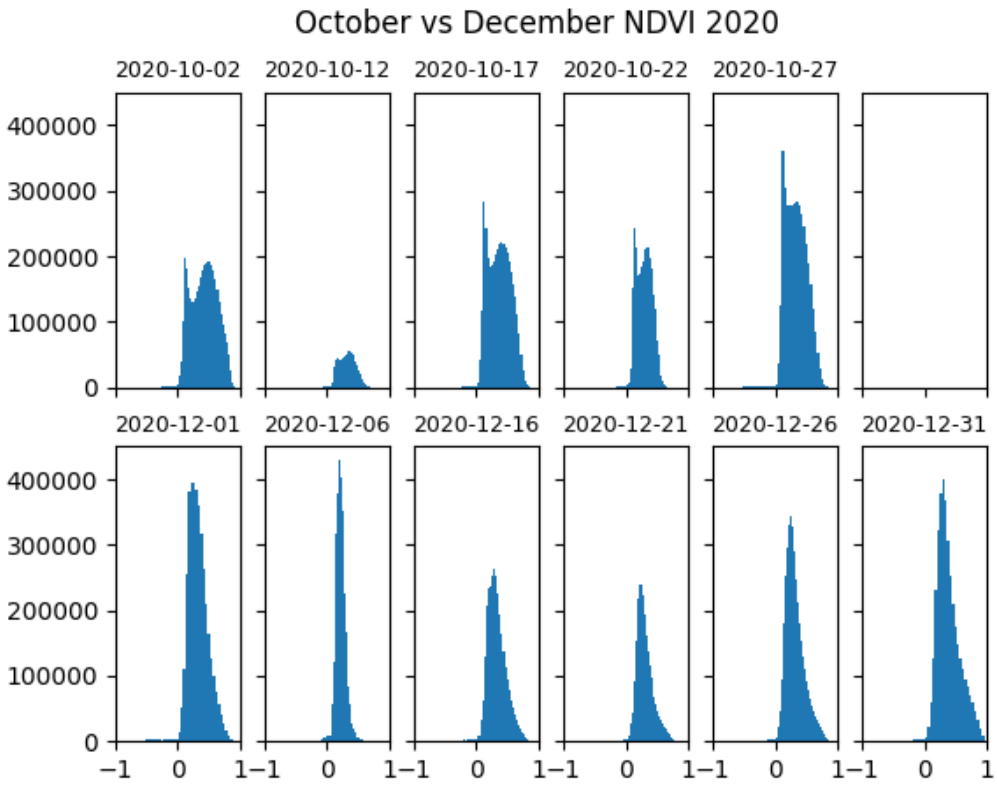


FIGURE 4.3: Distributions of NDVI values for October (above) and December (below) 2020, for the dates available in each month.

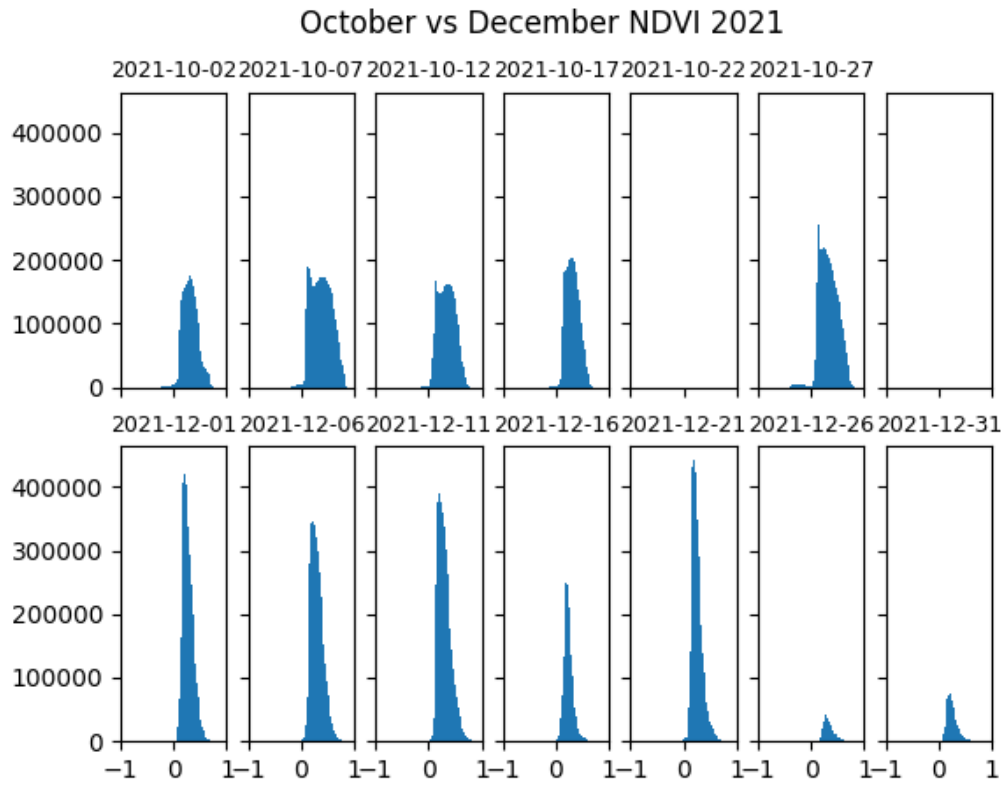


FIGURE 4.4: Distributions of NDVI values for October (above) and December (below) 2021, for the dates available in each month.

FIGURE 4.5: Displays the spread of average NDVI values per date, over the observation period, with mean and median values per date in red and blue, respectively. The blue highlighted area represents one standard deviation above and below the mean, to give an idea of the spread of the values. Each chart displays a year of data, from December to January, from 2018/2019 at the top to 2021/2022 at the bottom, with the turn of the year marked by the grey line.

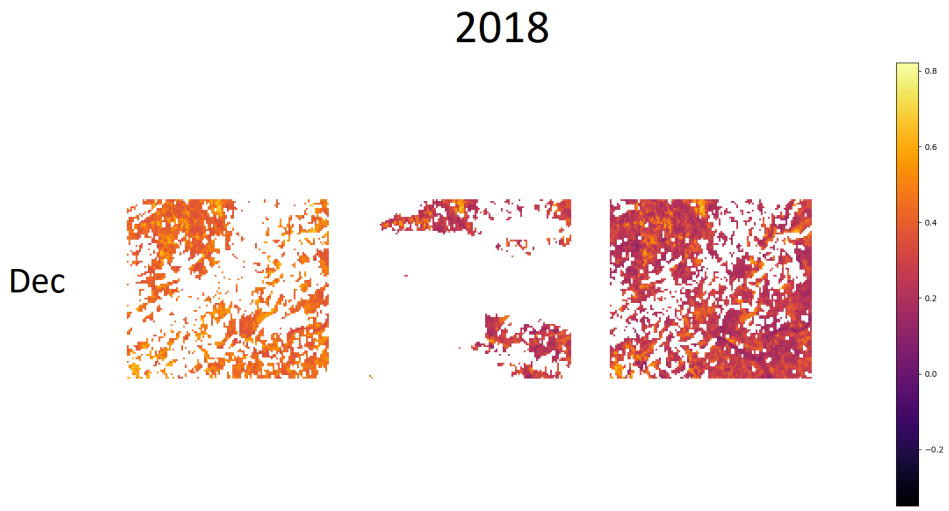


FIGURE 4.6: Maps of NDVI scores for available dates in December 2018. White pixels represent data masked out due to cloud or water cover.

over the course of each year, we rendered the NDVI values for the study area over the entire observation period. In Figs. 4.6, 4.7, 4.8, 4.9, 4.10 we can see how the NDVI values of the area changes over each year. This illustrates the trends seen in Fig. 4.5. The intensity increases around the winter months and decreases in the summer. This shows the prevalence of leafy vegetation in those winter months, and likely the desiccation of much plant matter in the summer. We can also see that there are far more removed pixels in the winter, this is due to increased cloud cover. It will be important to bear this in mind moving forward, as it may affect the accuracy of any analysis of proportions of different types of land cover if the surface is irregularly obscured.

4.3 Land cover

We can also gain more insight into what the different land classifications represent, by looking at the relative proportion of each over the course of the year. In Fig. 4.11, we show the percentage of total pixels represented by each classification, for each date, over the entire period. We screened out any dates with 10% or greater cloud cover, as this was found to produce the least noise, while also leaving a representative amount of data in any given month.

Fig. 4.11 shows some notable seasonal characteristics in the proportions of different land covers. Classifications used here are established in section 3.3. Class 1 is more prevalent between October and May, correlating with the cooler months and corroborating a classification as water. Classes 2 and 3, are present year-round, but more prevalent between April and September, especially class 3. This correlates with the hotter months, and overall supports their designation as dry or sandy soils. Class 4 is best represented

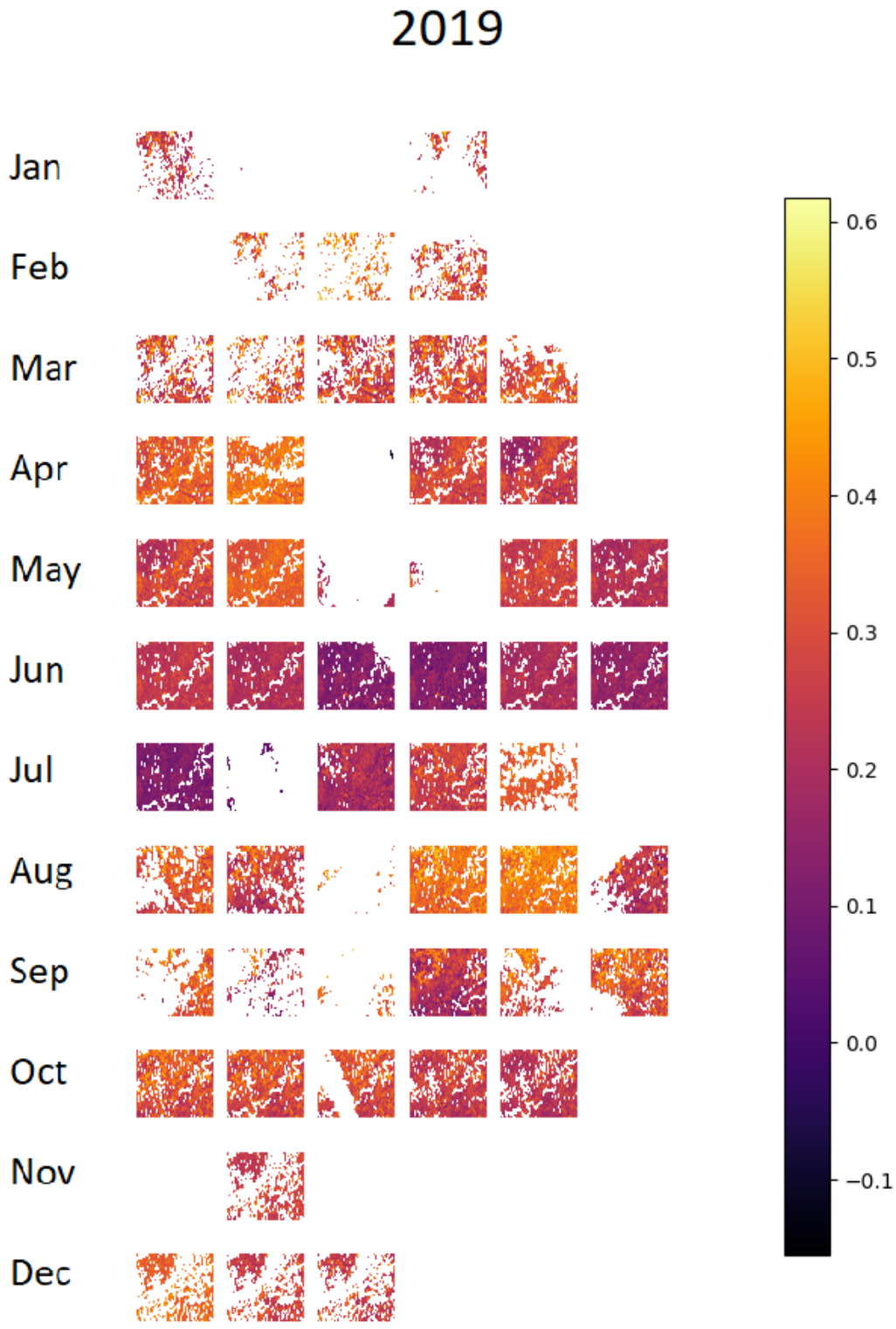


FIGURE 4.7: Maps of NDVI scores for available dates in 2019. Each row contains all the available observation dates for that month, from January at the top to December at the bottom. White pixels represent data masked out due to cloud or water cover.

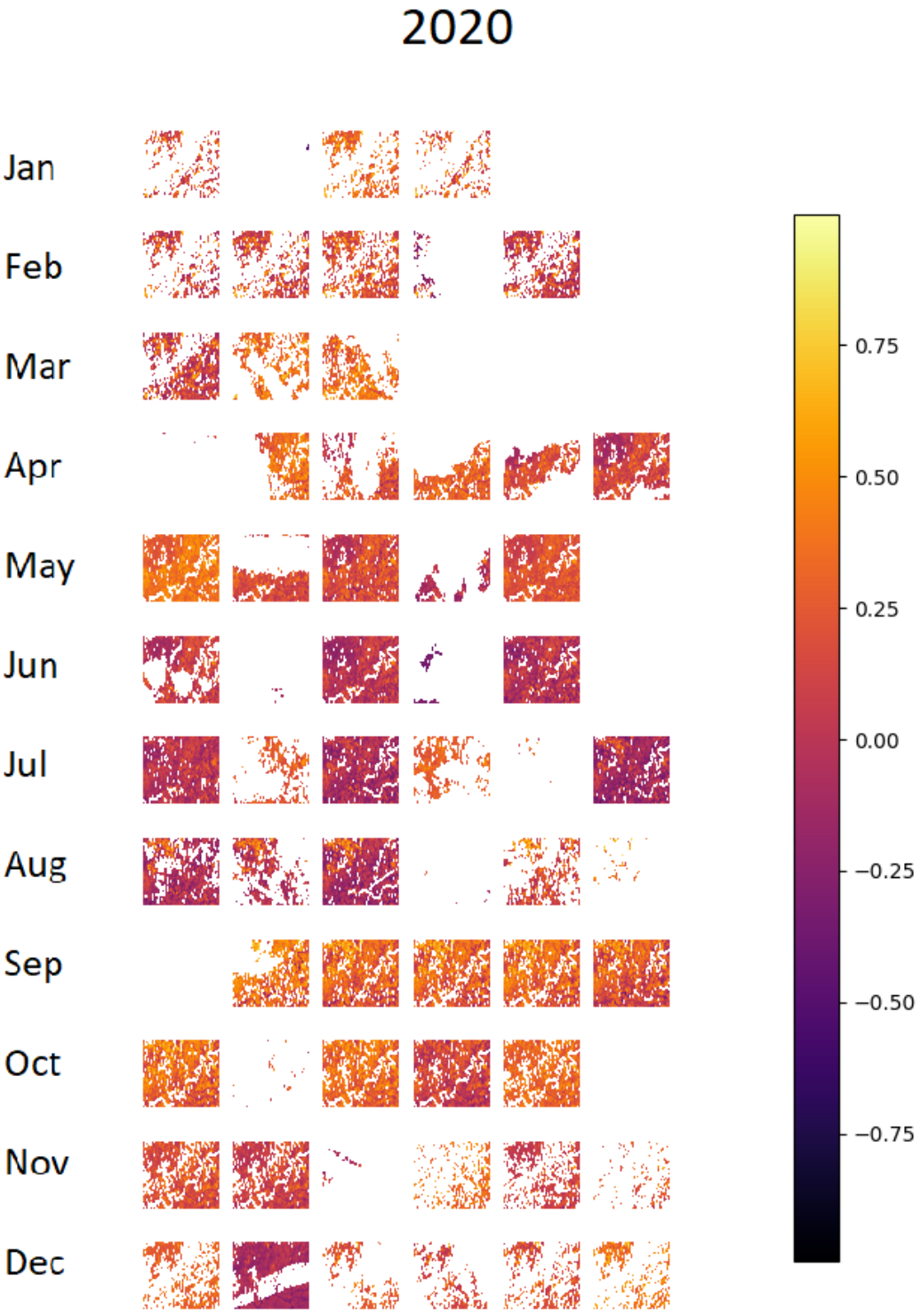


FIGURE 4.8: Maps of NDVI scores for available dates in 2020. Each row contains all the available observation dates for that month, from January at the top to December at the bottom. White pixels represent data masked out due to cloud or water cover.

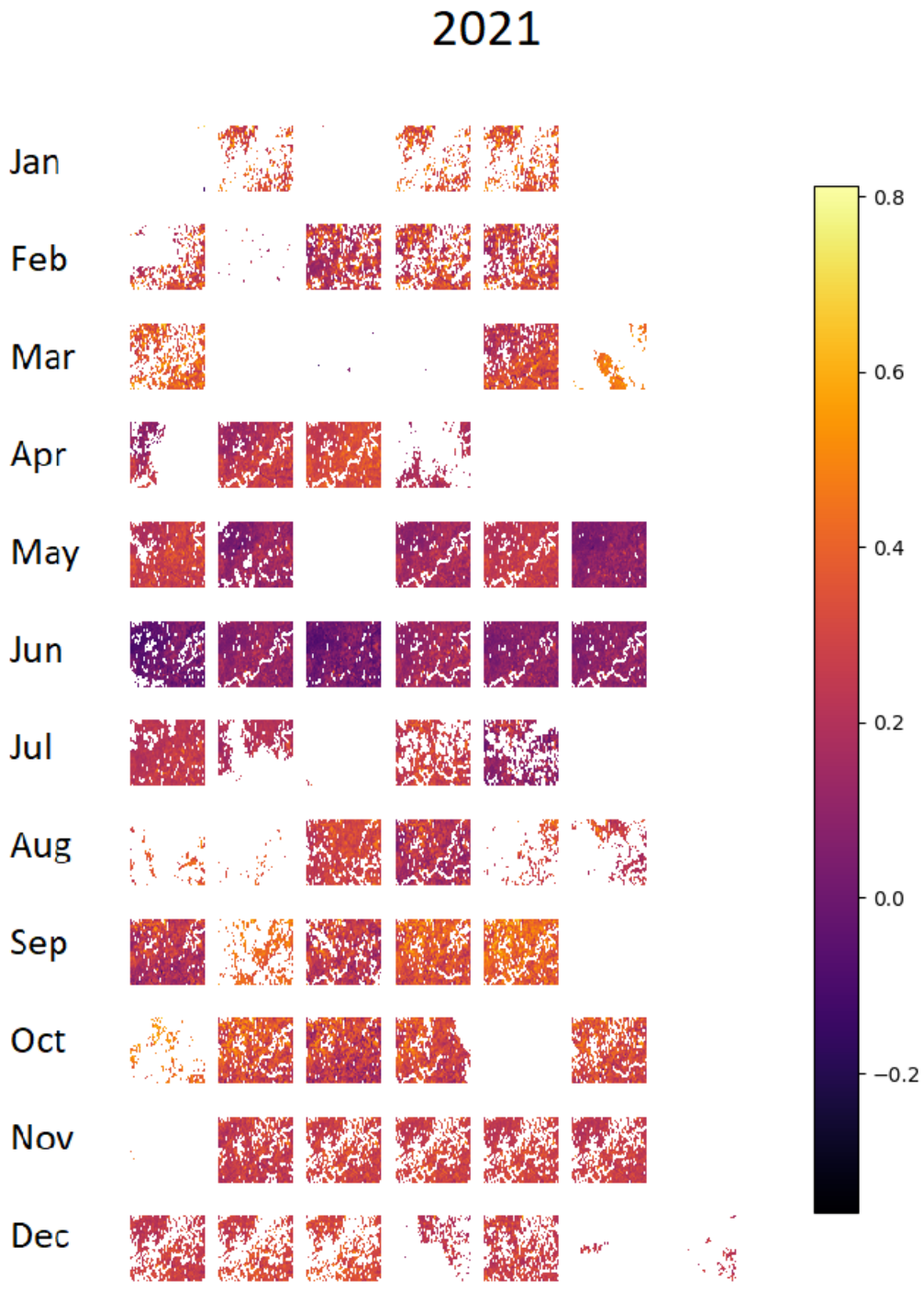


FIGURE 4.9: Maps of NDVI scores for available dates in 2021. Each row contains all the available observation dates for that month, from January at the top to December at the bottom. White pixels represent data masked out due to cloud or water cover.

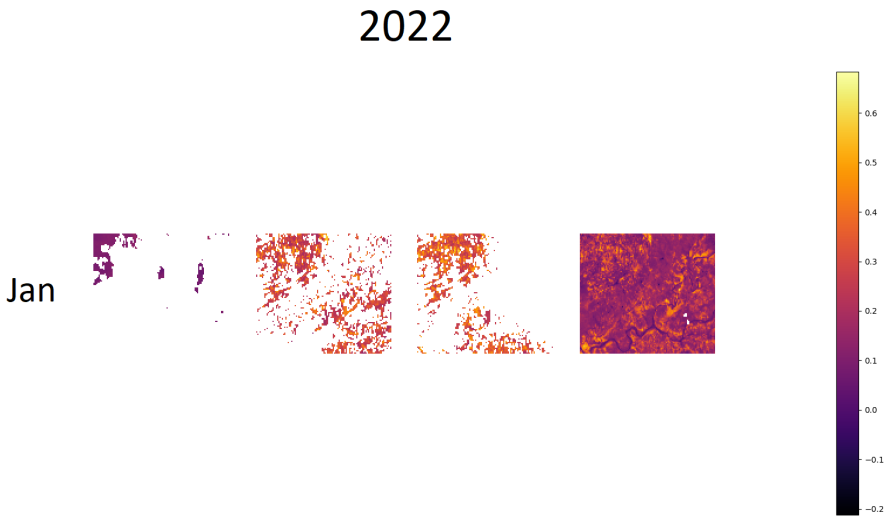


FIGURE 4.10: Maps of NDVI scores for available dates in January 2022. White pixels represent data masked out due to cloud or water cover.

between September and May. It increases and decreases in opposition to class 8, which is most common between April and October and decreases to 0 in the winter. This supports the classification of class 4 as green, leafy vegetation and class 8 as dried out vegetation. Classes 5 and 6 are most prevalent from July to November. Class 7 is found most commonly from October to May, this correlates with the cooler, wetter months and supports its designation as wet soil.

Observing the NDVI values of the different land type classifications, as seen in Fig. 4.12, helps us to understand the types of land cover they represent by their coverage of plant matter. Here, for each date, we calculated the average NDVI score per land cover class. We only calculated scores for classes which were represented by at least 100 points, to ensure there weren't too few data points to be meaningful. Fig. 4.12 shows that the highest scores of around 0.5, 0.7 and 0.4 belong to classes 4, 5 and 6 respectively. This corroborates their designation as leafy vegetation. The rest of the classes then have scores between 0.3 and 0, representing an indeterminate NDVI score. Class 8 is consistently the lowest scoring, which further suggests it being dried out vegetation.

4.4 Parthenium Presence

In Fig. 4.13 we can see the average and spread of Parthenium scores across each year. This score is an average of the Parthenium prediction score across all pixels for a given date, along with the standard deviation. The trend shows that the averages are more consistent around lower values between June and October, along with a reduction in the standard deviation and less variance in the scores. This is likely because the decrease in foliage around these months, as demonstrated previously, will also affect the Parthenium.

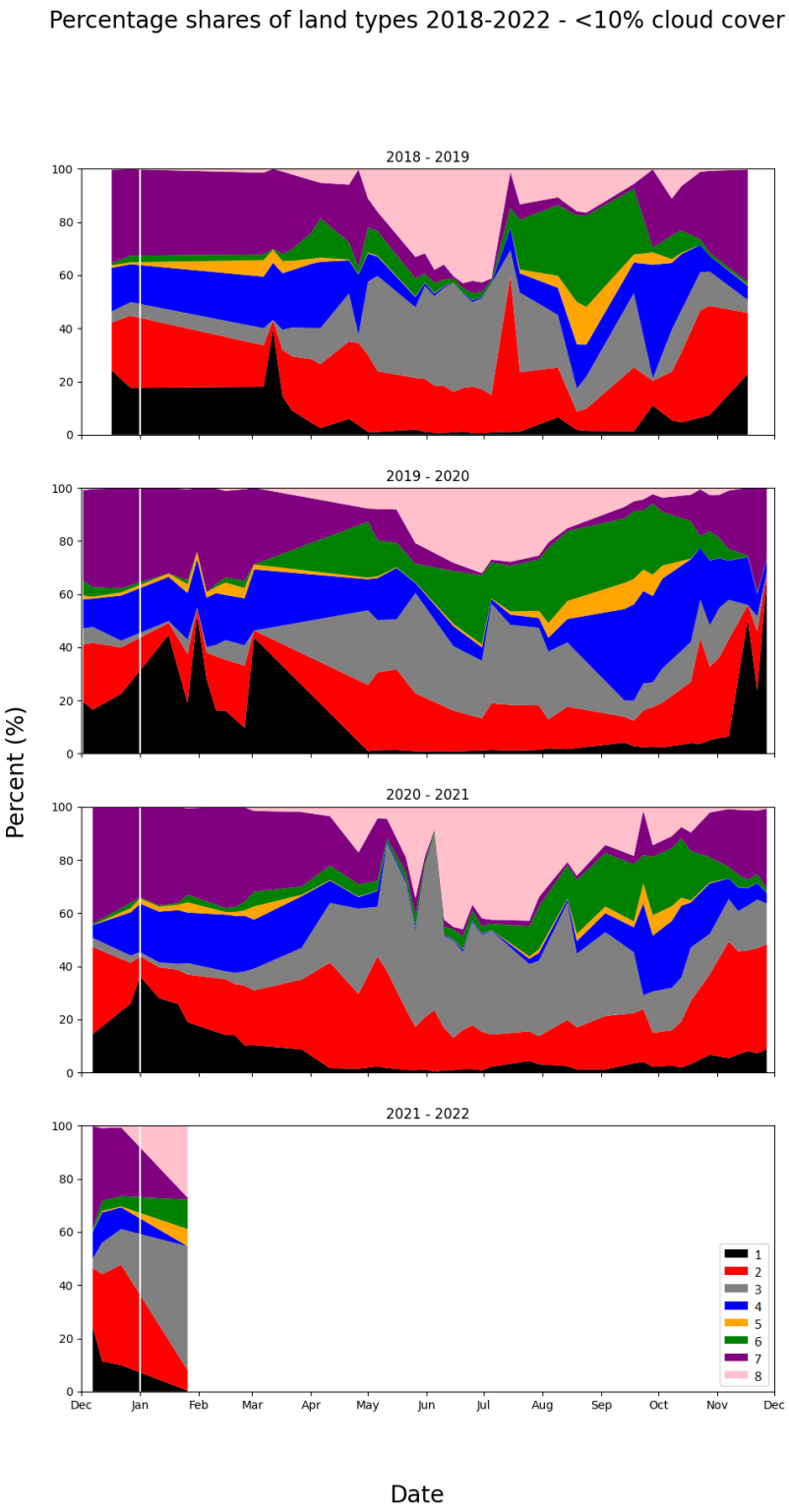


FIGURE 4.11: Shows the percentage share of each classified land type, as defined in section 3.3. Each chart displays a year of data, from December to January, from 2018/2019 at the top to 2021/2022 at the bottom, with the turn of the year marked by the white line. We are displaying only data from dates which had less than 10% cloud cover, as cloud cover can obscure the relative proportions of land types and cause excessive noise in the data.

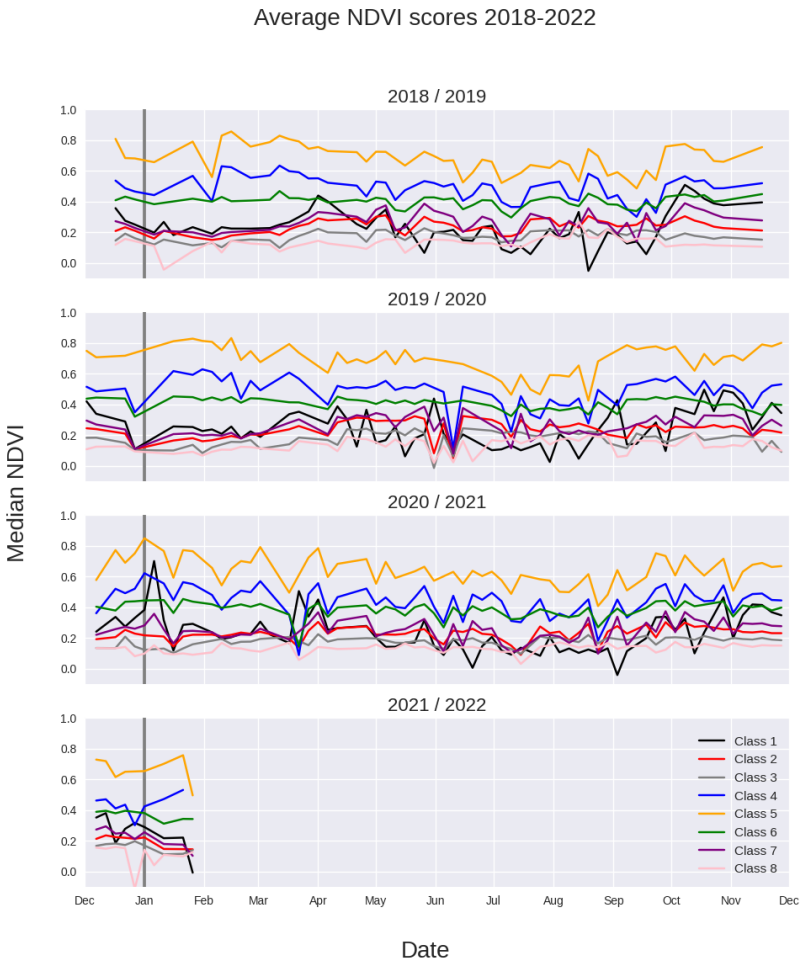


FIGURE 4.12: Shows the mean NDVI value, averaged over every pixel per date, for each land type, as classified in section 3.3. Each chart displays a year of data, from December to January, from 2018/2019 at the top to 2021/2022 at the bottom, with the turn of the year marked by the grey line. We only display data points where there are at least 100 pixels on that date to be averaged. This is to ensure a representative sample size.

During cooler weather, there is a range of areas with likely Parthenium and areas without. During the hotter months most foliage dries out or dies off, and so we do not have the variance of likely and unlikely presences of Parthenium, only a cluster around and very low (0.1) likelihood.

4.5 Comparisons

We have an idea of what the different land classes represent, and how Parthenium presence changes over time. We can now compare the two to see if we can observe any preferred habitats and how these might compare with observations from field data. Fig. 4.14 shows the distributions of average Parthenium scores per date for each land class, over the entire observation period.

We can see all the classes have an average score for Parthenium of around 0.05 to 0.1. The first standard deviation is between 0.125 and 0.175, and two standard deviations between 0.275 and 0.375. Outliers extend up to around 0.525 for all except class 7. Classes 2 and 7, dry, light soil and wet, dark soil, have a larger spread of values of likelihood of Parthenium, skewing higher than other classes. These classes would correspond well with wasteland or wet areas, which are preferred habitats for Parthenium. However the larger spread is not significant enough to indicate a correlation, merely a slightly better chance of finding higher Parthenium prediction values as per the classification model.

Observing the NDVI values against Parthenium values helps us evaluate if there is any correlation between the presence of plants generally, and Parthenium. In Fig. 4.15, we can see that Parthenium scores are highest around NDVI of 0.1-0.2, with the proportion of higher values falling off after that, up to an NDVI of 0.8. Below an NDVI of 0.1, the Parthenium likelihood drops off precipitously, down to about -0.3. This makes sense, given that an NDVI of 0 indicates an uncertainty of plant matter, with 1 being certainty of green, leafy vegetation, and -1 being a certainty of no plant matter. Past 0.1 NDVI, the predicted Parthenium chances are evenly distributed between 0 and 1. Overall this indicates no correlation between NDVI and Parthenium where plants are likely, but with the distribution of Parthenium predictions skewed lower for lower NDVI values. It is worth noting that this graph only covers the data for one day. However this is in December, around the peak of Parthenium prevalence, giving us predicted chances at the higher possible end of the spectrum, and the variety of cover within the observation area gives us a good cross-section of different levels of Parthenium coverage and NDVI values.

4.6 Conclusion

In this chapter we investigated the significance of the previously gathered data, in providing context for the simulations we would be conducting later. We found from the NDVI that there was a seasonality to the data, with a narrower spread of data and possibly lower values around June to September. We also noted an increase in cloud cover outside this summer period, which would affect the quantity of data points available for

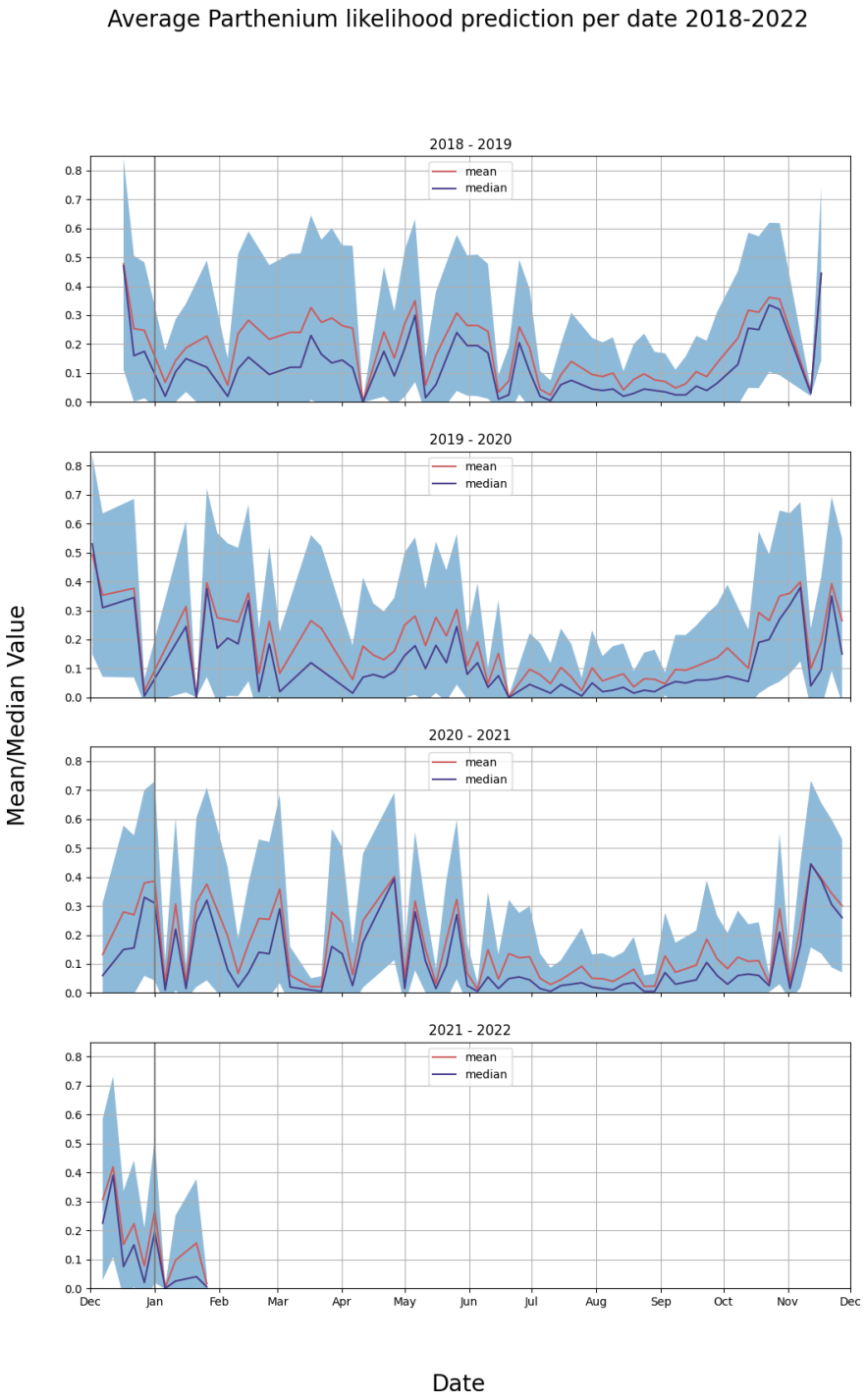


FIGURE 4.13: Displays the spread of Parthenium likelihood over the observation period, with mean and median values per date in red and blue, respectively. The blue highlighted area represents one standard deviation above and below the mean, to give an idea of the spread of the values. Each chart displays a year of data, from December to January, from 2018/2019 at the top to 2021/2022 at the bottom, with the turn of the year marked by the grey line.

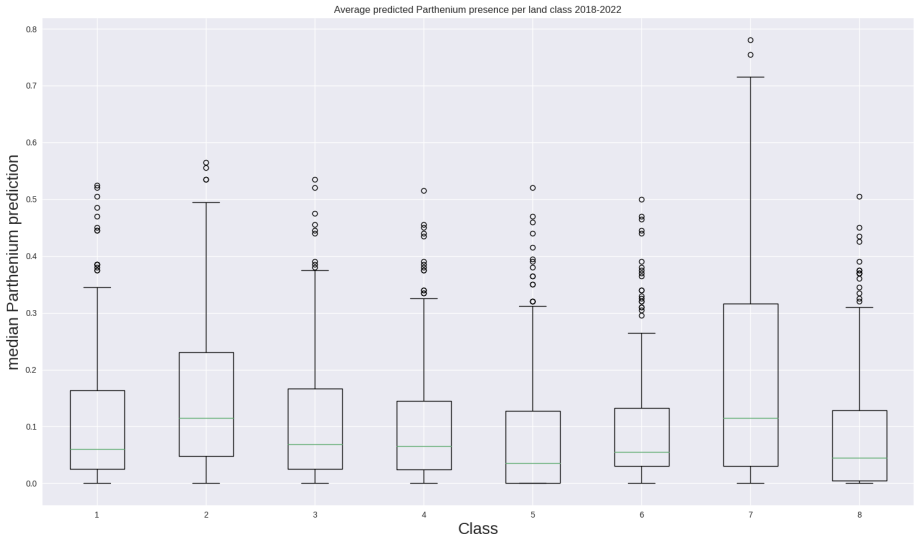


FIGURE 4.14: Distribution of Parthenium scores by classified land type. The green bar represents the median likelihood value of that land type, with the upper and lower ends of the box representing the third quartiles and first quartiles of the data, respectively. The top and bottom whiskers represent 1.5x the inter-quartile range, limited to 0 in the case of the lower whiskers. Circles represent individual data points outside these bounds. This data was calculated for the entire observation period.

those times. Observation of trends in land cover further supported this seasonality, with most classes of land type showing large differences in presence between the two periods. The seasonal patterns in land cover additionally supported the designations of these classifications made in Section 3.3. Analysis of Parthenium values furthered this seasonal trend, with lower variance and slightly lower values in the summer period. Comparing Parthenium values against land classes indicates no strong correlation in Parthenium distribution between land classes. Comparing against NDVI only indicates that Parthenium is only present along with NDVI values above 0, but shows no correlation between NDVI and Parthenium prediction values themselves. Overall we gained a validation of the land types we have previously classified, and a trend of seasonality among the various data, which would inform our understanding of Parthenium for our simulations.

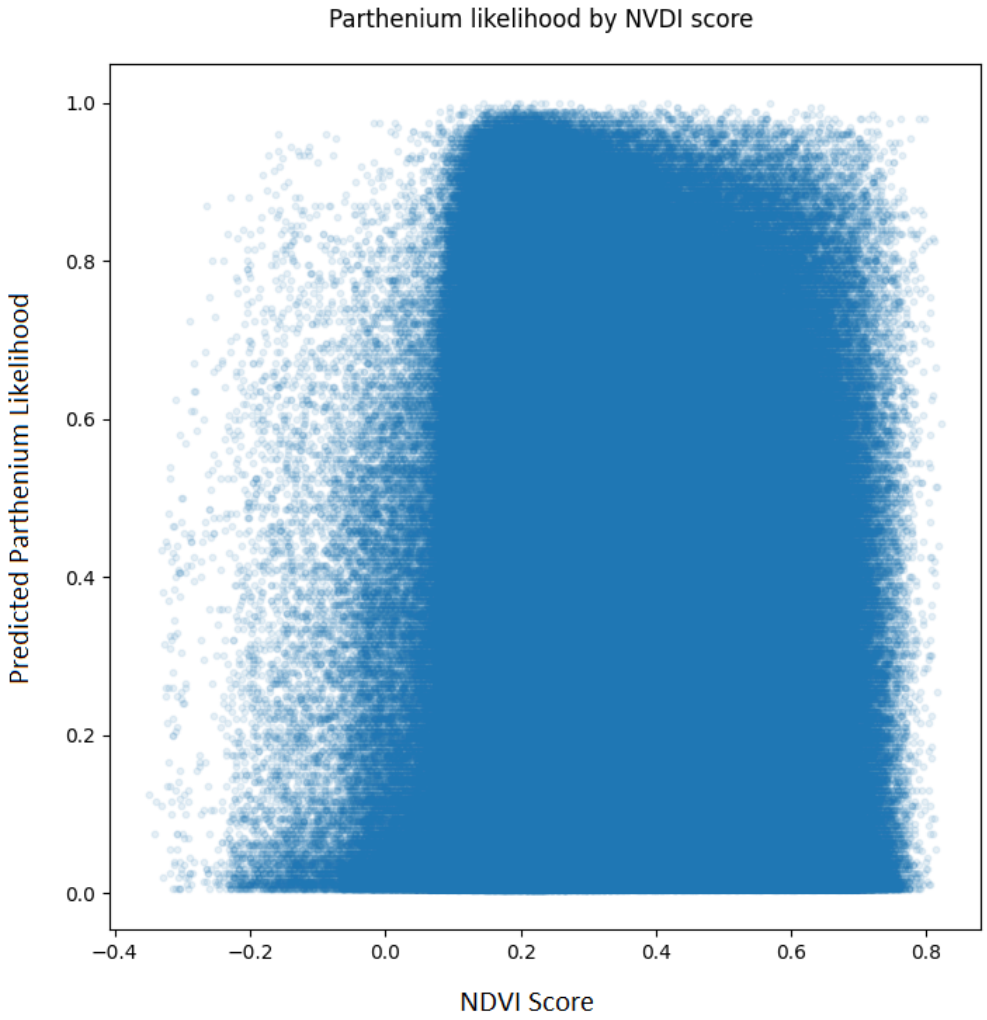


FIGURE 4.15: Displays the NDVI and Parthenium score for each pixel, for the data of 27/12/2018. Data points have partial transparency, such that areas of denser concentration with be darker in hue.

Chapter 5

Modelling

In this chapter, we bring together the data and insight gathered in previous chapters, in order to construct a simulation model capable of replicating the trends in Parthenium behaviour. First we look at epidemiological models, as we had previously established their use regarding invasive species in Section 2.2, and show how they can be used to model population trends. We then explain the function of cellular automata as a form of simulation with a spatial component. Finally we create our own simulations, using epidemiological models, to replicate Parthenium presence patterns as closely as possible, and use these models to gain insight into the factors affecting Parthenium growth.

5.1 Compartmental epidemiological models

A compartmental model is a type of epidemiological model used to simulate the changes in a population, so called because it splits the population into different compartments based on their status regarding the disease in question (Brauer et al., 2008). For instance, the most common type of compartmental model is the SIR model. In this, every individual in the population is defined as being either susceptible to the disease (S), infectious to others (I), or recovered (R).

This simplifies the modelling, in that one only has to model the transitions of individuals from one compartment to another. These transitions can be modelled stochastically, simulating the change for each individual, with assigned probabilities for the transitions between states.

Alternatively it can be simulated deterministically, wherein individuals are not simulated, but compartments as a whole. This has the advantage of being quicker and easier to calculate, as you do not need to calculate for each individual, but each compartment. This comes at the cost of the granularity of the data as one only sees the sizes of the different compartments change and not the effects on individuals. The transitions in this case are calculated using differential equations, for example (Brauer et al., 2008):

$$\mathrm{d}S/\mathrm{d}t = \frac{-\beta IS}{N}, \tag{5.1}$$

$$\mathrm{d}I/\mathrm{d}t = \frac{\beta IS}{N} - \gamma I\,, \tag{5.2}$$

$$\mathrm{d}R/\mathrm{d}t = \gamma I\,. \tag{5.3}$$

Where S is the susceptible population, those who can be infected, I is the infected population, R is the recovered population, those who have contracted the disease and recovered, now having immunity to it and N is the total population. β represents the rate of infections, γ represents the number of people an infected individual will infect, and t represents the time value.

The SIR model is the most common and one of the simplest, however there are a variety of compartmental models which we considered for our purposes. We needed a model which could best map onto the behaviours of an invasive plant species. SIR was therefore inappropriate, as there is no analogue to the recovered state in this model. Similarly, other additional states such as in the SIRD and SIRV models, which add deceased (D) and vaccinated (V) states respectively, do not correlate to the behaviours of areas of land in an invasive plant scenario.

Ultimately we settled on SIS as an appropriate first-order representation, wherein one can transition from susceptible, to infected, and back to susceptible. There is the chance of removing the infection, but no long term cure or immunity. This correlates to the ability to remove Parthenium from an area, but not to completely prevent its reintroduction. We also identified the SEIS model as potentially useful, wherein the E stands for exposed, denoting an individual which has been exposed to the infection but is not yet infectious to others. This helps simulate the nuance of an area having been seeded with Parthenium, or bearing maturing Parthenium plants, but none that can yet produce seeds themselves.

5.2 Cellular automaton

Initial modelling was attempted using a cellular automata. A cellular automata is a model, whereby the system to be modelled is represented by a 2-dimensional grid which evolves over time, with each cell representing a unit of area, or agent in the system. Each cell may exist in a multitude of states, and the state of all cells is updated at regular time intervals, as per a set of criteria defined in the creation of the model. A common example is Conway's Game of Life (Schulman & Seiden, 1978), which takes place in an infinite grid, wherein each cell represents an organism, either alive or dead. With each update or "tick" the state of each cell is re-evaluated, based on the state of each its 8 orthogonally and diagonally adjacent neighbours. The criteria for state changes are as follows:

- **Underpopulation** Any live cell with 0 or 1 live neighbours dies.
- **Sustain** Any live cell with 2 or 3 live neighbours stays alive.
- **Overpopulation** Any live cell with 3 or more live neighbours dies.

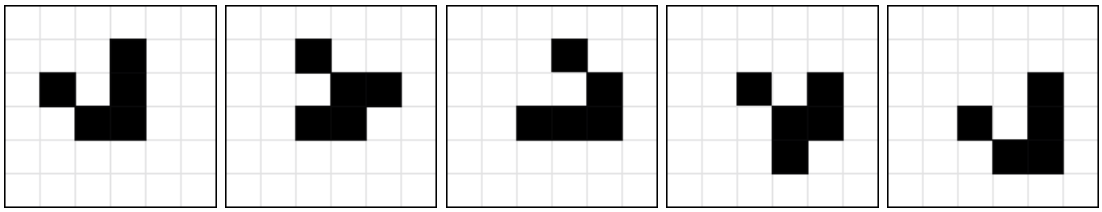


FIGURE 5.1: Illustration of the "glider" pattern from Conway's Game of Life. After the initial state, the pattern moves through 3 interstitial states to arrive at the same arrangement as the initial state, but translated 1 cell down and to the right.

- **Reproduction** Any dead cell with exactly 3 living neighbours becomes a live cell.

From this relatively simple setup, complex interactions can emerge, with repeating patterns and objects that are able to "move" across the grid, as demonstrated in Fig. 5.1 with a "glider", a common pattern.

While Game of Life is intended purely as an amusement, cellular automata may take many forms and can be useful in modelling real systems. Cells may have many states, and be used to represent portions of an area under observation, or individuals in a population. In our case, a cellular automata can be used to represent 10m x 10m portions of an area imaged with Sentinel-2. The state of each cell can represent the presence of Parthenium. An initial model was created, using a constrained grid of cells, where each cell could either contain Parthenium or not. An empty cell became a Parthenium cell if a neighbour contained Parthenium. Otherwise cells remained the same in updates. Several "seed" cells were randomly assigned to start containing Parthenium. However, this simulation was strongly shaped by the constraints of its own rules. First, the square shape of the expansion is a result of the square grid of the simulation, it does not consider a diagonal neighbour to be further away than an orthogonal one. In order to make the spread better reflect proximity to the infection source, we wanted to emulate a radial spread pattern[1]. An easy way to achieve this is by introducing randomness to the infection process. Demonstrated in Fig. 5.2, we can overlay a circle over the neighbourhood of an infected cell, with a diameter equal to the width of the neighbourhood (3 cells wide). If we then consider each neighbour to contain a circle, whose diameter is equal to the width of its cell, we can calculate how much of each neighbour circle is occupied by the area of the infected cell's infection radius. The neighbours fall into two categories; orthogonals with 100% covered by the circle, and diagonals with 57.3%. If we use these values to determine the chance of a neighbouring cell being infected, then we see a much more natural, radial looking spread pattern emerge. These values can now be applied on top of any future infection calculations, to imitate a radial infection spread.

Of course, a 100% guaranteed infection is not accurate to real world conditions, and we needed more nuanced criteria for infection. We used a Markov chain to structure the possible transitions between cell states. Markov chains represent possibility spaces as a series of states, with transitions that can occur between different states, or from a

[1]We will be using a method outlined at: https://scipython.com/blog/the-forest-fire-model/

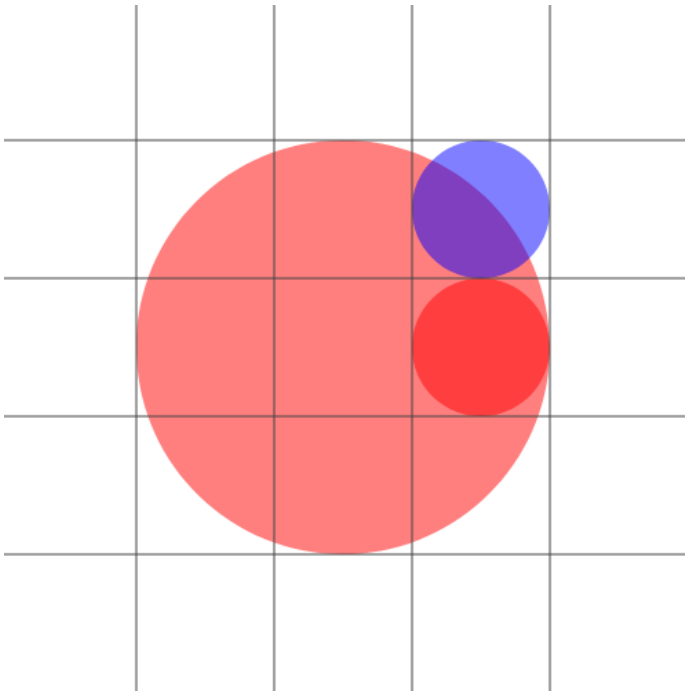


FIGURE 5.2: A demonstration of the relative areas of neighbours' circles which are overlapped by the infection radius of the infected cell, in the centre. The large red circle is the infected cell's infection radius, the small blue circle is a diagonal neighbour and the small red circle is an orthogonal neighbour. Credit: https://scipython.com/blog/the-forest-fire-model/

state to itself. In our case, the criteria for transitions were random chance, as dictated by the transition probabilities collected from the data. We used observations for all dates, binarised to hold either a value or 1 or 0 for Parthenium present or absent. We then recorded, for each pixel, what state it was in between each pair of consecutive dates. We then used these data to construct a normalised transition matrix, containing the transition probabilities from each state to each other state.

In order to get a better idea of how the model thus far compares to real life, and where it needs improving, we next used real data as a starting point. We used Parthenium prediction data from the area identified in 4.1 as a starting state for the model, correlating a pixel in the imaging data to a cell in the model. We used an observation from the beginning of the available period - 17/12/2018. For the sake of removing variables in initial tests, we binarised the data, such that any pixel with greater than 50% predicted chance of Parthenium was assigned a Parthenium cell, and any pixel below the threshold was assigned an empty cell. The probability of state changes was determined using the transition matrix previously developed. Each iteration was to represent the passing of a day, with the total simulation time being roughly equivalent to the amount of observed data we have access to, 3 years. This was so we could start at the first date we have data for and finish after 3 years, where we can validate the simulation results against the real data.

By comparing the end result of the simulation with the Parthenium predictions for 3

years later, we saw that they are drastically different. Our model seems to lead to a rapid expansion in the population, whereas the real life population seems to be largely stable. Additionally, the existing population is more prevalent in some areas than others and does not spread at all to some areas, like water. The simulation needed to be able to take into account these differing interactions with different types of land. It is also apparent that using the averaged transitions rates between states for the transition matrix is not a good approximation. It is also worth noting that this model does not feature spontaneous emergence of Parthenium, i.e. Parthenium growing in areas which are not adjacent to existing populations. As seeds are often carried great distances by humans and vehicles, and this is the main method of spread to new locales, this seems to be a crucial element of the plant's spread to be simulated. However, this would prove difficult to accurately represent, being as random as it is.

With a multitude of issues to be resolved with the model, we decided to simplify it further to reduce the variables that needed accounting for. We decided to do this by reducing the dimensionality from a 2-dimensional map to a 1-dimensional population count.

5.3 Population model

In this model we aimed to still use the basis of epidemiological modelling as we did previously. However, instead of modelling individual cells we would be modelling the overall population of Parthenium in the area. This removed the complication of adjacency, and the need to simulate individual parcels of land, instead being able to model the entire population in terms of a mass of fungible entities. In this way, rather than using the transition matrices of the previous model, we can directly use the differential equations of the SIS model.

Previously, we found that the Parthenium was not drastically increasing or decreasing in the observed area, and so seemed to have achieved a kind of equilibrium. In order to simulate this effectively, we needed data on the Parthenium population to compare against. We opted to continue using binarised Parthenium predictions, using the classifier from Fennel & Breton (2023), using a threshold of over 50% of the prediction score to determine if a pixel was Parthenium. We found the average Parthenium coverage across the year, which was approximately 20%. This would then be the target for the simulation, achieving a steady population of 20%.

In order to perform the simulation, we need the differential equations for our chosen model, SIS. The susceptible population's size, $S(t)$, will decrease by the number of infections per day, $\beta I(t)$, multiplied by the proportion of population still susceptible, $S(t)/N$. It will also increase by the number of people recovering per day, $\gamma I(t)$. The infected population changes by the number of new infections, $\beta I(t)S(t)/N$, minus the recoveries, $\gamma I(t)$. N represents the total population count. β represents the expected number of individuals an infected individual infects per day. γ represents the proportion of infected population which recovers each day $(1/D)$. D represents the number of days an infected

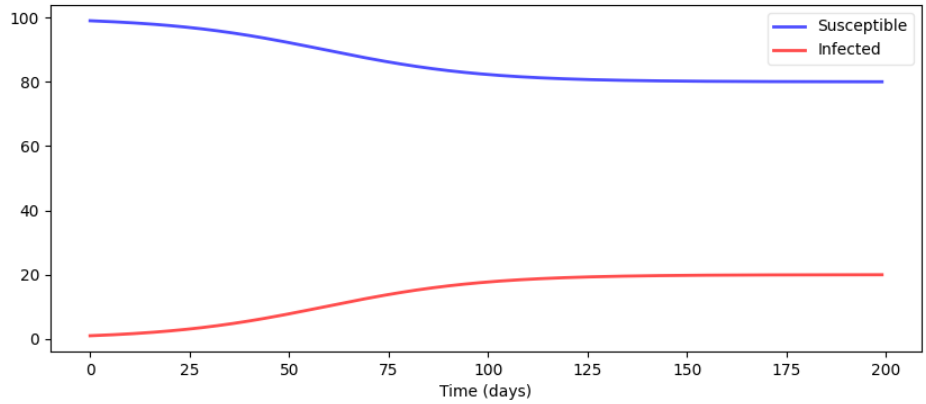


FIGURE 5.3: A test simulation, with the starting parameters of N = 100, I(0) = 1, S(0) = 99, β = 0.25, D = 5, γ = 0.2, simulation length = 200.

person stays infectious. Useful for future reference will be R_0, which represents the total number of other individuals an infected individual will infect ($R_0 = \beta/\gamma$). All told we get:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \gamma I - \frac{\beta I S}{N}, \tag{5.4}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \frac{\beta I S}{N} - \gamma I. \tag{5.5}$$

Following from these equations, we can calculate where the equilibrium of the model will resolve to:

$$\frac{\beta}{\gamma} \leq 1 \Rightarrow \lim_{t \to +\infty} I(t) = 0, \tag{5.6}$$

$$\frac{\beta}{\gamma} > 1 \Rightarrow \lim_{t \to +\infty} I(t) = \left(1 - \frac{\gamma}{\beta}\right) N. \tag{5.7}$$

i.e. if $\beta/\gamma \leq 1$, the disease cannot sustain itself and will die out. If $\beta/\gamma > 1$, the fraction of infectious population will reach an equilibrium at $1 - \gamma/\beta$. As immunity cannot be reached in this model, all individuals will eventually become infected, however the infected population is constantly changing, as new individuals recover and get infected.

We were aiming for the population to reach a stable state with 20% infected, which must mean that $1 - \gamma/\beta = 0.2$. This can be simplified to $\gamma/\beta = 0.8$. This means that we only need consider the interaction of γ and β to achieve a steady equilibrium. In Fig. 5.3 we can see the steady state achieved, from a starting infected population of 1.

Whilst we need to keep the ratio between β and γ the same, we can modify their values, keeping the same relationship between them. This allows us to change how quickly the population changes, and how fast the rate of change accelerates with new infections. For example, see Fig. 5.4. The higher infection rate (β) and shorter symptomatic period

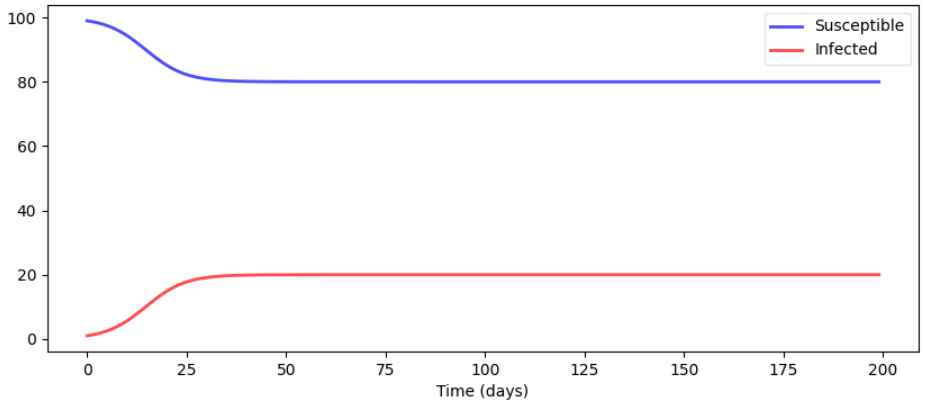


FIGURE 5.4: Modified from the previous simulation, with the starting parameters of N = 100, I(0) = 1, S(0) = 99, β = 1, D = 1.25, γ = 0.8, simulation length = 200.

(D, with $\gamma = 1/D$), results in a much higher rate of infection, leading to the same equilibrium in a much shorter time period. The same equilibrium is reached as with a lower infectivity rate, due to the shorter symptomatic period, meaning that individuals have less time to infect others, despite being more infectious.

Observing the population change of Parthenium over time however, we can see that while the population is consistent year-on-year, it fluctuates throughout the year. Fig. 5.5 shows the percentage of the total observed area covered by Parthenium, averaged over each month, for the duration of the observation period. As observed in previous graphs of Parthenium (see section 4.4), we can see that there is a larger variance and higher values from around October to May, and lower variance and values from May to October. Amongst the higher scoring periods, the incidence of Parthenium seems to be higher at the end of the year, about 2.7 to 4.1 in November and December, compared to 1.6 to 3.0 from January to April, with a possible outlier in March of 2020. This period from February to April could also represent a second, smaller depression, but it is difficult to say with only 3 years of data and the noise involved. It is clear, however, that there is a large degree of variation in the population over the year, which correlates with what we found in section 4.4. This poses an issue for our model, as currently it is assumed that the qualities of the system being modelled are consistent, however we can see a distinct seasonality in the population trends of Parthenium.

The issue is that our model started from the estimated average Parthenium population and expanded from there, with no limit. However, in actuality Parthenium appears to have reached a seasonal equilibrium in the region, with the population being approximately the same in any given month between years, but varying greatly over the year. Therefore the model needed to be modified to reflect this reality. This entailed both simulating the population dynamics in such a way that the population did not increase or decrease from year to year, but also to simulate the seasonal variation in population over the year. The primary ways we had of affecting the infected population levels in the simulation were by modifying the β and γ variables.

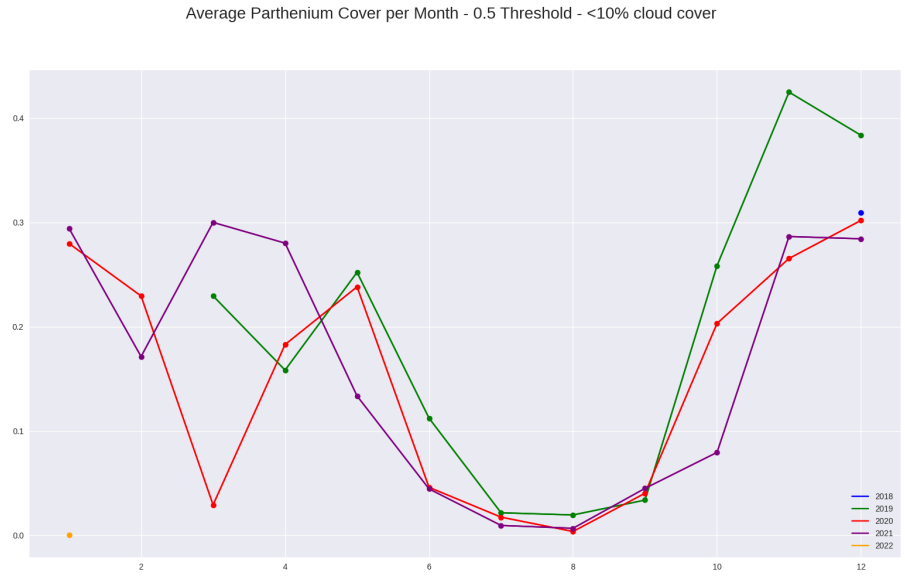


FIGURE 5.5: Percentage of the observed area covered by Parthenium, using a threshold for Parthenium presence of 0.5 and omitting any dates with more than 10% cloud cover. Each series represents a different year of data.

The dip in population was in the summer months, likely correlated with the lack of water and dying back of green, leafy vegetation. The best corollary in the model would be β, the infectivity, standing in for the plant's ability to produce and spread seeds. γ, representing how many other individuals are infected, or how much seeds are spread, is less likely to fluctuate with the seasons.

In order to modulate the β value over the course of the year, we needed to change it from a fixed value, to a function of time. Given that the variation was regular, and entailed a peak and a trough, the simplest matching pattern would be a sine wave. Therefore we can model the value of β as a sinusoidal function over time. To ensure that the result is consistent over more than one year, we ran the simulation over multiple years. We can describe this new sinusoidal β with the following formula:

$$\beta(t) = \beta_1 \sin(\omega t + \phi) + \beta_0 \tag{5.8}$$

where β_1 is the amplitude, ω is the frequency, ϕ is the phase offset, and β_0 is the y offset. The function's starting parameters are set such that the infectivity does not dip below 0, a maximum amplitude congruent with previous data, and the cyclical period equal to one year. This gives us an oscillation between 0 and 1, which we then multiply by the desired maximum β, in this case 0.11, to achieve an oscillation between the desired maximum and 0. In Fig. 5.6 we can see the effect of this new β function on the population simulation. We achieved a yearly modulation in the population, however the trough occurs later than it should. It does not coincide with the trough of the sine wave, around the middle of the year, but rather is offset by about a quarter cycle later. This will be due to the

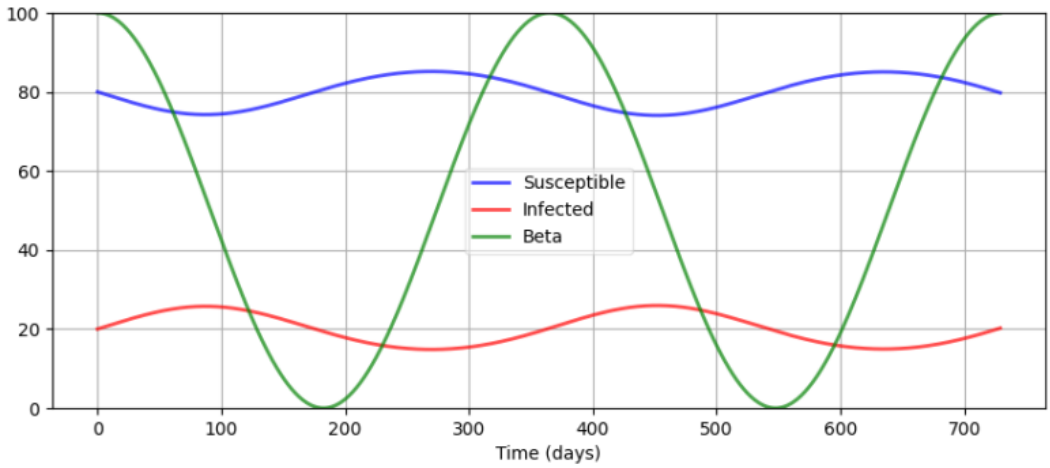


FIGURE 5.6: Example simulation with parameters of duration = 730 days, N = 100, S(0) = 80, I(0) = 20, β = 0.011, D = 210, γ = 0.00476. β function parameters of β_1 = 0.5, ω = 1, ϕ = 0, β_0 = 0.5

delayed effect of any change in infectivity, taking some time to resolve throughout the population. Offsetting the β function to be a quarter cycle earlier, moved the population trough to the right period. However, we could not make such manual tweaks for all the relevant variables; we needed a more programmatic way of optimising the function.

Least Squares Fitting (LSF) is a technique which can be employed in order to arrive at an optimal set of parameters for the simulation, without lengthy manual adjustments and guesswork. The method works, given a set of data points, by finding a curve of best fit through the points. To do this, it is assumed that there is a model describing such a curve, which can be optimised to best describe the data by minimizing the sum of the squares of the normalised difference between the model and the data. In application, this is often a simple linear or parabolic relationship. In our case, we will be assuming the optimal formula is that of the population differential equations, with a sine wave function for the γ variable. The aim is to minimise the squared of the distance between the best-fit curve at each data point. This will be achieved by modifying each of the free variables in the formula in tandem, until the best result is obtained, i.e. the lowest sum of squares of differences between the resulting formula and actual data points. We will be optimising for 4 variables: the symptomatic period (D) of the simulation model itself, and the amplitude (β_1), phase (ϕ) and y offset (β_0) of the β sine wave function. Additionally, upper and lower bounds were placed on some variables, as it would be unnecessary for the algorithm to experiment with values outside these bounds. Namely, the phase was bound to between $-\pi$ and π, as this represents half a phase in either direction– any more would be redundant. The value of β at any given time was also limited to 0 or above, as a negative β value would not fit in the confines of our definition of it as the rate of infection. A negative infection rate is not physically possible.

In our case we used the *least_squares* function from the *scipy.optimize* python package. This allowed us to provide a set of example real-world data to emulate the function

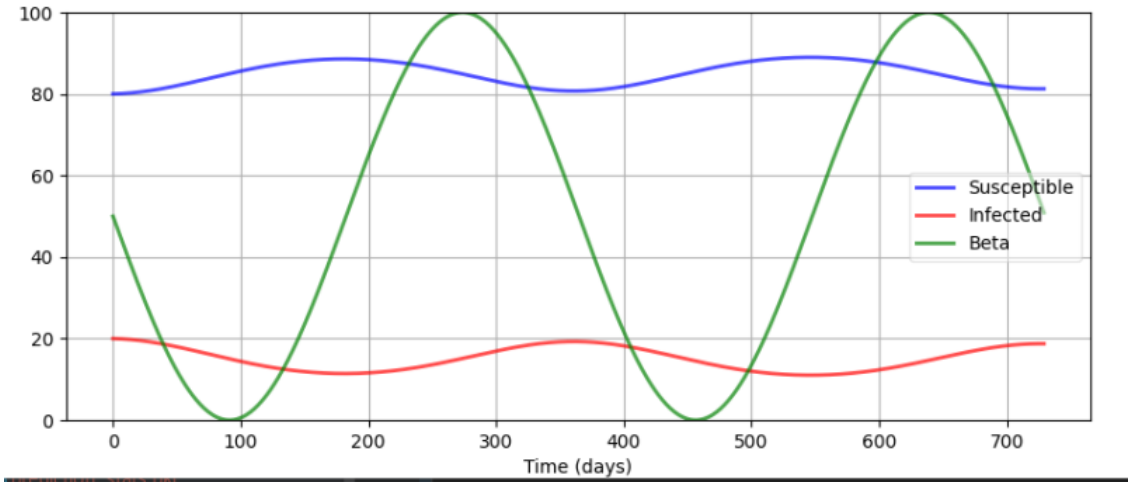


FIGURE 5.7: Example simulation with parameters of duration = 730 days, N = 100, S(0) = 80, I(0) = 20, β = 0.011, D = 210, γ = 0.00476. β function parameters of β_1 = 0.5, ω = 1, $\phi = \frac{-\pi}{2}$, β_0 = 0.5

Model Free Parameters		
Parameter	Initial Value	Optimised Value
D	210	209.863
β_1	0.5	1.028
ϕ	$\frac{-\pi}{2}$	-3.718
β_0	0.5	0.508

TABLE 5.1: Initial and optimised parameter values for initial model.

which parameters need to be optimised, and the set of parameters to be optimised. The *least_squares* function will then run a least squares fitting, to find the optimal parameters values to attain a function output as close as possible to the example data given. The fitting model needs criteria to aim for, and for this purpose we will be using the monthly average Parthenium cover, as shown in Fig. 5.8.

To simplify the model we opted to approximate the year to be 360 days, composed of 12 30-day months. The difference of a day or two between months would make minimal difference to the model, as the data we are comparing against are averaged over the course of each month. Our initial attempt at fitting yielded a promising result: see Fig. 5.9, with values shown in Table. 5.1. The optimised parameter values produce a curve which mirrors the shape of the data well, with a dip around the summer months and a peak in the winter. However, the curve is too broad, smoothing out the plateau from January to May and reducing the drop in the summer. Additionally, the generated curve does not level off between November and December, resulting in an upward trend which would not lead to a reasonable decrease leading into January, as seen in the real data. We saw the latter issue as more pressing, as this indicated that the model was only being optimised to a single year, and not to seasonality over multiple years, as was the case in actual fact. Therefore, our next priority was to optimise the model over the course of several years.

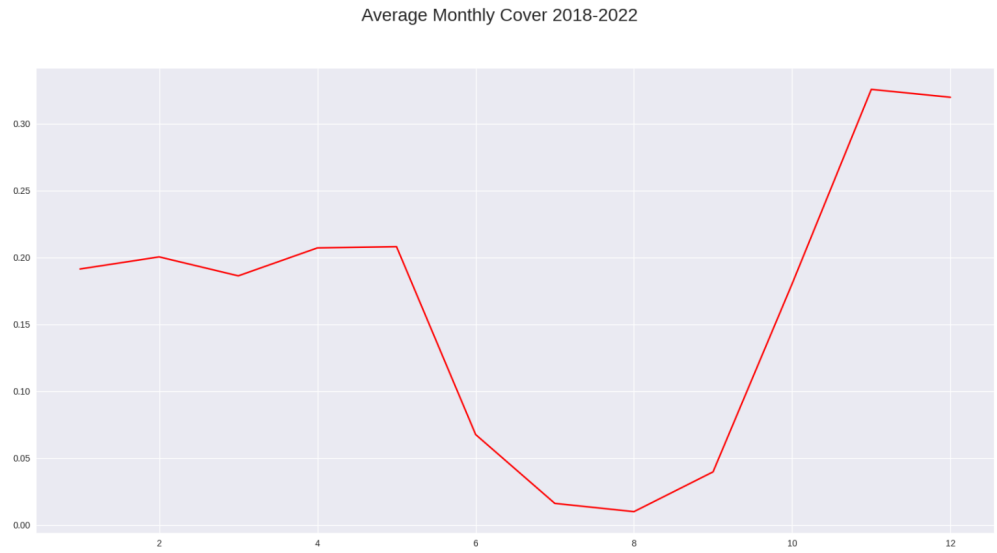


FIGURE 5.8: Average Parthenium cover by month.

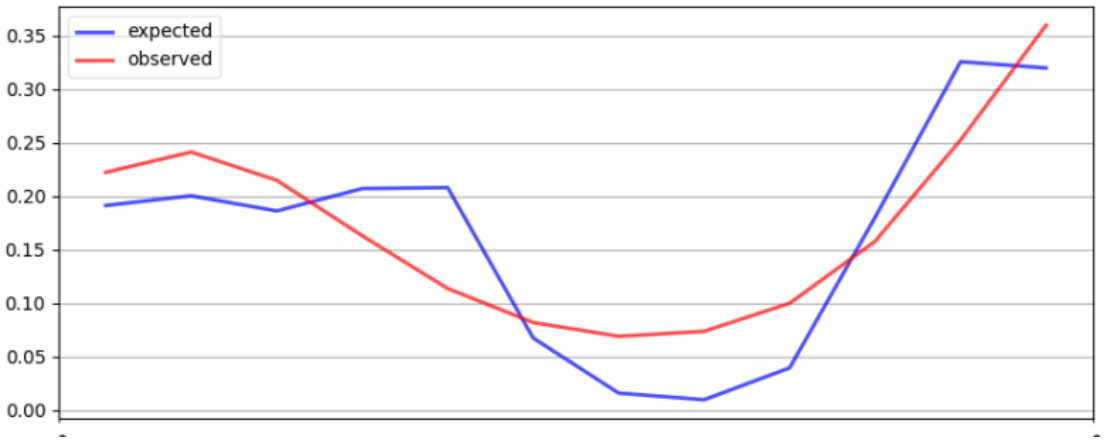


FIGURE 5.9: Initial Least Squares Fitting attempt.

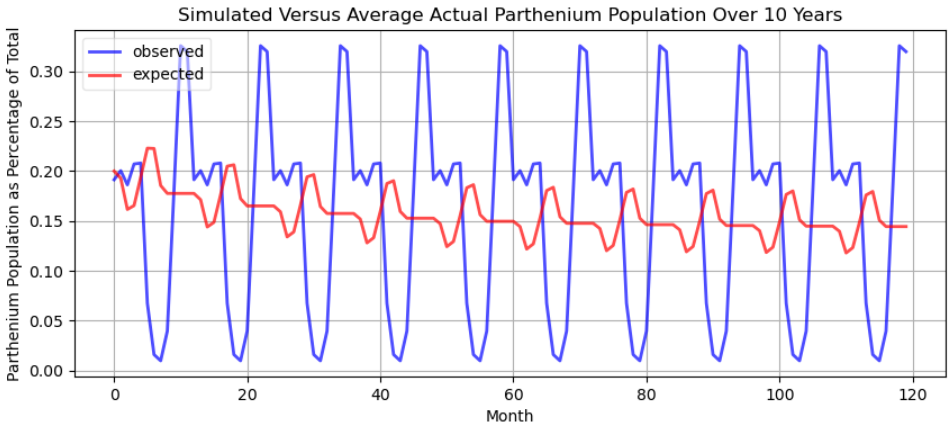


FIGURE 5.10: Initial test of SIS model using LSF optimised parameters, compared against averaged Parthenium population data duplicated over 10 years.

Model Free Parameters		
Parameter	Initial Value	Optimised Value
D	210	81.319
β_1	0.05	0.023
ϕ	$\frac{-\pi}{4}$	-0.776
β_0	0.05	0.0000735

TABLE 5.2: Initial and optimised parameter values for initial 10 year fitted model.

We opted to simulate next over 10 years, using the averaged monthly data, duplicated 10 times in sequence. This was not intended to be representative of the inter-year fluctuations of the Parthenium population over those 10 years, but rather to ensure that the model could generate a stable pattern of population changes that is consistent between years and independent from the initial conditions. Fig. 5.10, with parameters in Table. 5.2, shows that this approach is approximating the right shape, but much too small and out of phase. Fig. 5.11 shows the output of the β function for this simulation. This may be due to the LSF algorithm having settled on a local minima. A local minima is a point on a curve whose value is the lowest of those in its immediate area, but not necessarily the lowest in the whole series. If the algorithm's starting value for a given parameter is within the bounds of the "valley" surrounding this local minima, it may arrive at that value, erroneously believing it to be the absolute minimum value. In order to check, we can modify the starting values of the free parameters given to the LSF algorithm to see if we can achieve a better optimised value.

Indeed, by changing the starting values, we are able to achieve the result seen in Fig. 5.12, with parameters as in Fig. 5.3, with β function output seen in Fig. 5.13, which fits the observed data far better. It fits the overall shape of the data better, reaching the same peak, approaching the trough and imitating the plateau in the middle. However this was the best that could be achieved using a sinusoidal β. We wanted to generate a beta pattern with a double peak, similar to that of the population graph itself, however this is not possible with a sine function. When multiple trigonometric functions are added

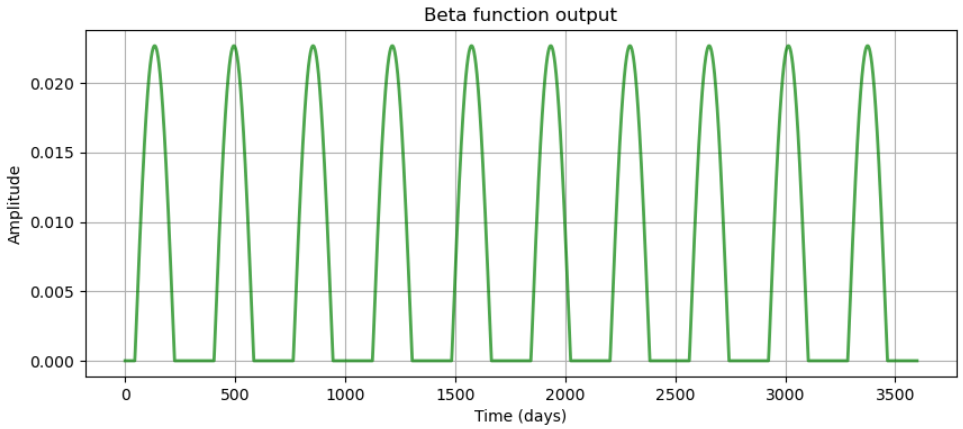


FIGURE 5.11: Sinusoidal β function output for initial test of SIS model using LSF optimised parameters over 10 years.

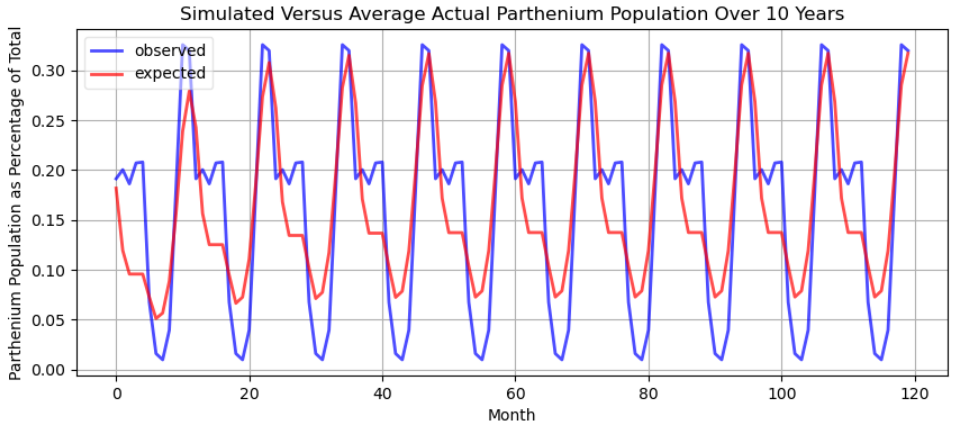


FIGURE 5.12: Improved model fitting for 10 year SIS simulation.

together, the result is always one consistent waveform. Multiple peaks with one cycle are not possible. We opted to use Gaussian functions to calculate the β value instead, as this would allow us to layer together multiple functions.

A Gaussian function, if truncated at either end and repeated, has roughly the same shape as a truncated sine wave, and so can serve the same function of imitating seasonal highs and lows, but allows us to achieve more irregular profiles by adding layering multiple function outputs. This can also more adequately model an underlying physical process responsible for triggering Parthenium's growth which is more limited in time, such as a rainy season. Equation 5.9 shows the definition of a Gaussian function for our purposes, where a is an arbitrary constant, t is the current time value, t_0 is the median value, denoting the peak of the curve, and σ is the standard deviation of the distribution. In order to have the pattern repeat, rather than tend towards an asymptote at either end, we applied the modulo function, with the value 360, to the time value t. This caused the function to repeat its values every 360 days, forming a wave with a cycle of 1 year. We initially tested with just one Gaussian function output as the β value, see Fig. 5.14,

Model Free Parameters		
Parameter	Initial Value	Optimised Value
D	210	41.588
β_1	0	0.03
ϕ	$\frac{\pi}{2}$	2.94
β_0	0.05	0.02

TABLE 5.3: Initial and optimised parameter values for improved model.

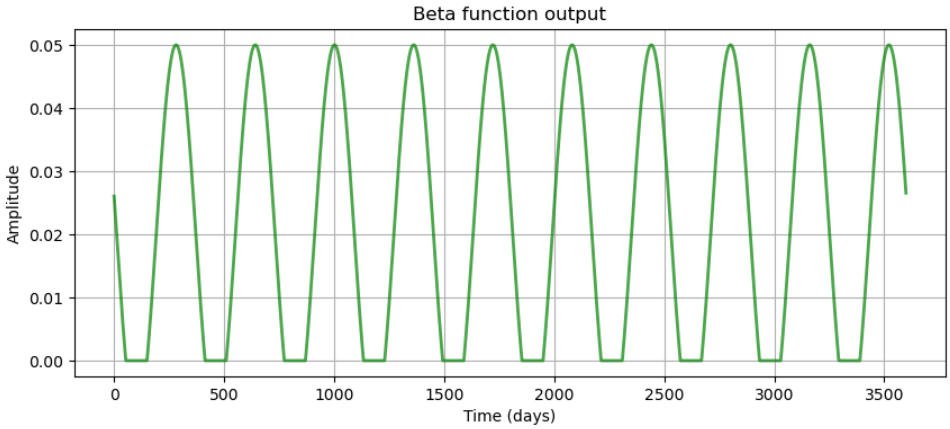


FIGURE 5.13: Sinusoidal β function for the improved model.

Table. 5.4 and Fig. 5.15. This resulted in only a rough approximation of the data and so we moved on to layering two Gaussian functions together, simply by calculating them separately and adding together the results.

$$f(x) = a \exp\left(\frac{-(t-t_0)^2}{2\sigma^2}\right) \tag{5.9}$$

Fig. 5.17, with parameters in Table. 5.5, shows the two Gaussian functions layered together, resulting in the simulation seen in Fig. 5.16. This provided a much better match than we had achieved previously, with both the high and low points matched closely, the plateau in the spring months represented, and the curves in between closely followed. It is also stable over the simulated 10 year time span.

Model Free Parameters		
Parameter	Initial Value	Optimised Value
D	231	232.22
a	0.07	0.0598
t_0	284	284.679
σ	13.5	12.803

TABLE 5.4: Initial and optimised parameter values for initial Gaussian model.

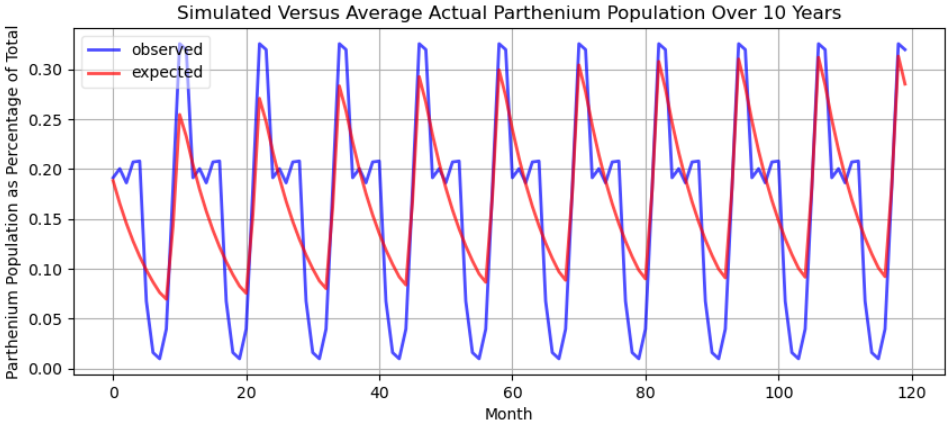


FIGURE 5.14: Initial test of the single Gaussian function applied as the β function for a 10 year SIS simulation.

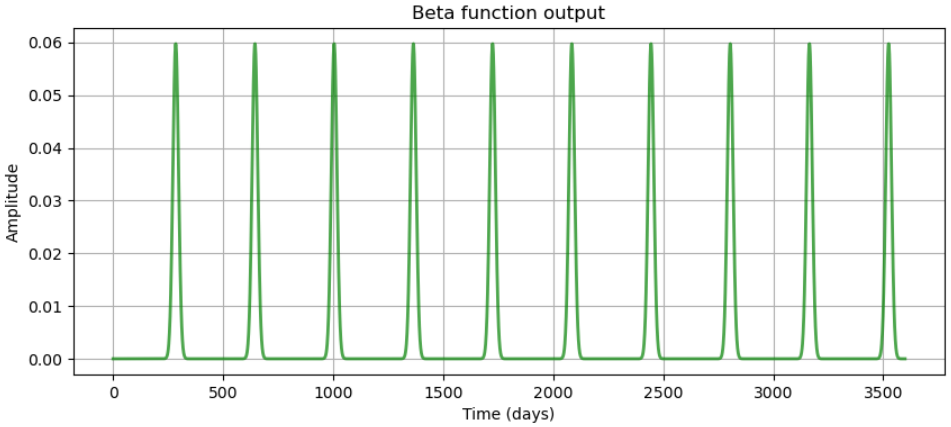


FIGURE 5.15: Gaussian β function for the initial Gaussian test.

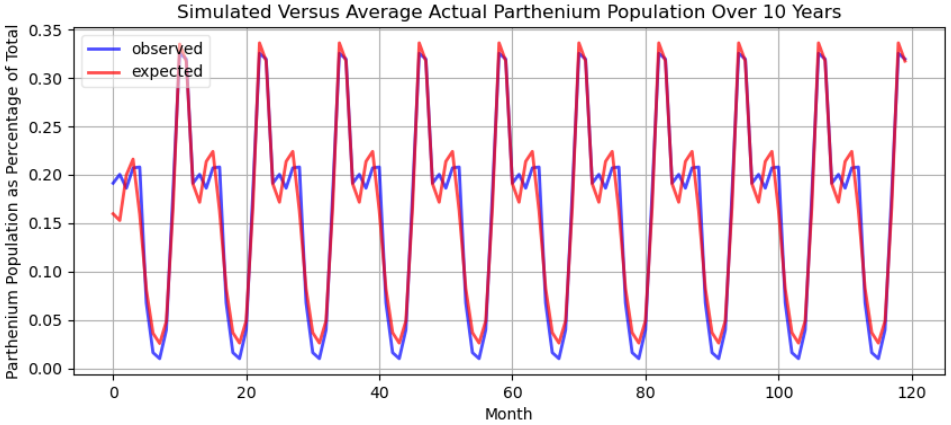


FIGURE 5.16: Optimised fit for the double Gaussian function model, modelled over 10 years and compared against averaged Parthenium population data, duplicated over the same period.

Model Free Parameters		
Parameter	Initial Value	Optimised Value
D	231	13.644
a	1	0.102
t_0	180	68.647
σ	90	65.684
a_2	1	0.136
t_02	180	290.628
σ_2	90	58.847

TABLE 5.5: Initial and optimised parameter values for double Gaussian model.

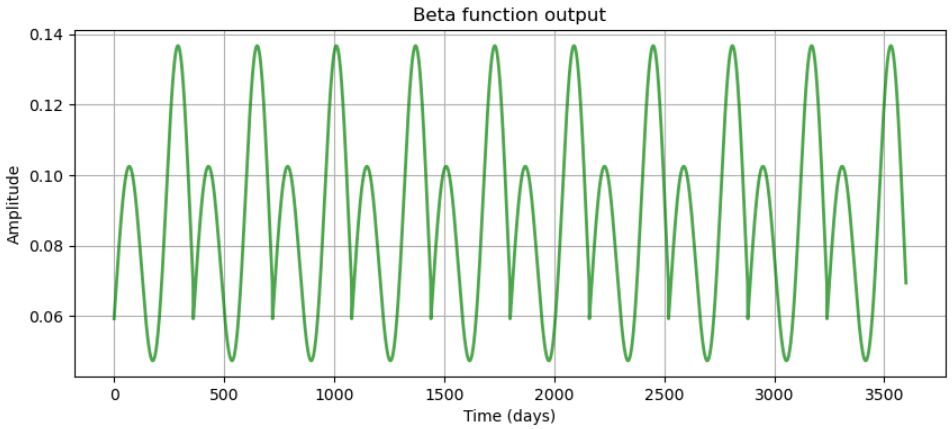


FIGURE 5.17: The combined output of the 2 Gaussian functions used as the β function.

5.4 Analysis

In this experimentation, we have found that we can generate a reasonable approximation of the actual population fluctuations of Parthenium using a simple epidemiological model. The relationship between the β value, representing "infectiousness", and γ, the number of other individuals "infected", is the important factor here. This relationship dictated the stable levels of the population, and so was the determining factor in the fluctuations of that population throughout the year. This then indicates a factor which influences the prevalence of Parthenium throughout the year. It is possible that this factor is external, an environmental factor like precipitation, temperature or the interference of other plants or animals. The deciding factor could also be internal, simply the regular pattern of seed dispersal of Parthenium throughout the year.

We can draw parallels between the parameters of the model and real world factors, which may be useful in further research. β, the "infectiousness", represents in disease modelling the ease of spread to other individuals. In our application of the model, the analogue would be the dispersal of seeds to spread Parthenium to new areas. D, the symptomatic period, may represent the period of time during which seeds are spread. Its inverse γ, originally representing the average number of other individuals infected by an infected individual, can correlate here to the number of other areas to which an area infested with Parthenium will spread to. If β represents the seed dispersal of Parthenium, then t_0 will represent the peak rate of dispersal, and σ will dictate the spread of time over which this dispersal accelerates and decelerates.

Taking the values from the best model fitting we can draw some inferences from the optimised values. The ultimate value arrived at for D was 13.64 days, which would indicate a period of seeding of around 2 weeks. However, the σ values for the first and second Gaussian functions were 65.68 and 58.84, respectively. With σ representing the standard deviation of a Gaussian distribution, this would indicate that the 95% of the seeding occurred within 2 standard deviations, or around 90 days in each case, much longer than 1 week. This may suggest a disconnect; that these parameters do not correlate exactly with real world factors, as we had assumed. It may also indicate that there are more complicated interactions between the parameters, potentially with currently unknown factors coming into play.

The values for the t_0 for the first and second Gaussian functions were 68.65 and 290.62, respectively. These correlate with the peaks in the spring and winter, however not exactly. The spring plateau occurs at around January-May, but peaks April-May, so with our approximated 30-day months, the middle of the plateau is around day 75 of the year and the peak at day 120. In the winter, the peak is between November and December, putting the peak around day 330. The winter population peak being preceded by 40 days by the peak in β supports the theory of β representing intensity of seeding, given the time to grow to maturity for Parthenium is around 30 days, with some more time for seeds to take hold and germinate. The spring population aligns less well, being either exactly at the same time, or approximately 50 days later. However, this may be influenced by the

winter peak supporting some growth in the early months of the year, with the spring β peak "topping up" the rate of spread, leading to the steady plateau and slight peak in April-May.

The values of t_0 seem to support the assumptions made about their significance in terms of real life factors, while the correlation between D and σ seem not to. This, coupled with the relative accuracy of the model to actual population trends, suggests that elements of the model, such as the variations in β value, hold some significance. However, some elements, such as the D and σ values may require further scrutiny and be affected by other, external factors. Regardless, this model poses a promising application of epidemiological models to invasive plants, and a validation of existing knowledge about Parthenium and its behaviours.

5.5 Conclusion

In this chapter we aimed to create simulations of Parthenium, both to test if epidemiological models could be used to model invasive species, and to gain insight into the behaviors of Parthenium and external factors affecting it. We concluded that an SIS type compartmental model best suited our purposes, with potential of using an SEIS model in future. We initially attempted to create a cellular automata model, but found that the complexity of such a model would be beyond the scope of this thesis. We therefore moved to a less complex population model, still using the basis of an epidemiological model. We experimented with different parameters and the use of Least Square Fitting to optimise these, as well as adapting the model from a stable population to match the seasonal fluctuations noted in Parthenium values.

The model we generated had some limitations, in that we were not able to validate it against other areas of the country and it is constructed of just over 3 years worth of data, as well as the Parthenium data being based of the predictions of a classifier. The end result was a model that was able to closely replicate the real life trends of Parthenium presence with a relatively simple epidemiological model, as we had aimed for. The parameters it arrived at offer some insight into the behaviours of Parthenium, for instance there appear to be external factors affecting its propagation, which warrants further investigation. Additionally, some parameters which we assumed would match more closely to real-world values did not. This may indicate inaccuracies in the model, or that the relationship between parameters and real world values is more complex than we assumed, the investigation of which may yield useful insight into the behaviours of Parthenium.

Chapter 6

Conclusion

6.1 The problem

Invasive species are a global issue of growing concern, negatively affecting ecosystems, public health and economies (Büyüktahtakın & Haight, 2018). Communities will seek to curtail the spread of invasive species by monitoring them, preventing their migration, and attempting to eradicate any populations which have gained a foothold in a new environment (Richards et al., 2012). These efforts are often difficult, expensive and ineffective, as large areas of terrain must be observed and small quantities of an invasive species can escape detection and rapidly proliferate. This then gives rise to the use of behavioural and environmental models to attempt to predict and preempt the movements of particular alien species (Dana et al., 2014).

Computer models are often used in the prediction of invasive plant behaviours, with most falling under the umbrella of species distribution models (SDMs). While widely used, these models have their advantages and disadvantages and require a large variety of data to be collected and processed. It would be useful to be able to construct models with a simpler set of data. Invasive species behave in many ways like infectious diseases. Their spread, largely based on proximity, but with some far reaching vectors, and the fluctuations in population, bear striking resemblances (Strickland, 2015). Epidemiological models therefore offer a possible alternative, often requiring fewer forms of data than invasive species models, while remaining useful in behavioural prediction.

We experimented using data of the invasive plant Parthenium in Pakistan. Originating in the Americas, it has spread to now occupy 46 countries around the world, and presents a major ecological problem (Steve Adkins & Asad Shabbir, 2014). Each plant can propagate quickly (Nguyen et al., 2017) and chemically inhibit the growth of other plants (Belgeri et al., 2011). Parthenium can cause serious respiratory problems and dermatitis (Sharma et al., 2013). Upon ingestion, the plant causes stomach irritation in humans and the spoiling of meat and milk in animals. Parthenium has also been known to cause famines due to its suppression of crops (Towers & Mitchell, 1983). It is a major problem in Pakistan, having a detrimental affect on public health and agriculture, and has now spread to three of the country's provinces (Steve Adkins & Asad Shabbir, 2014). Being

as difficult as Parthenium is to tackle, further insight into its reaction to environmental stimuli and possible future behaviours would be valuable in directing efforts to combat it.

6.2 Our work

We first explored the data obtained from Sentinel-2, observing the variance in predicted Parthenium presence and NDVI values across time and land type, and the distribution and change in coverage of various land types over the year. We investigated the use of epidemiological models, settling on an SIS compartmental model to represent the spread of Parthenium through an area. Initially we investigated the use of cellular automata, as a representation of the physical distribution of Parthenium over time. However, it proved to be too complicated to construct an accurate model in this way, in the time we had. Therefore, we reduced the dimensionality of the problem to only model the size of the Parthenium population over time, and not its physical distribution. Using the classifier developed in Fennel & Breton (2023) on data from Sentinel-2, we were able to generate predictions for the presence of Parthenium across the observed area. We set a threshold of 50% confidence of Parthenium, above which we would presume the presence of Parthenium, and below its absence. This gave us a binary map of the presence of Parthenium for each image.

As the population of Parthenium did not appreciably change from year to year, we initially sought to create a model of a system at equilibrium. However, this did not accurately represent the actual state of the population, as it fluctuated throughout the year. As the compartmental model assumed constant values for its parameters throughout, we had to adapt it to induce a seasonal fluctuation in population. We achieved this by modifying the "infectivity" parameter over the year. Initially we created this seasonal change using a sinusoidal function, as a sine wave roughly matched the peak and trough of the Parthenium population. We used Least Squares Fitting to optimise the parameters of the model, however even with optimised values the model still fell short of the real data. We attributed this to the actual population not following a simple rise and fall, but rather a major peak in the winter and a lower "plateau" in the spring, before the trough in the summer. In order to represent this, we needed a more irregular model, which was achieved using two Gaussian functions in tandem. This approached yielded a model which closely matched the averaged data for the actual Parthenium population in the area.

In our analysis of the NDVI values over the course of each year, we found that there was a pronounced seasonal variance, with higher variances in values in the winter/spring and much lower in the summer. This was in-keeping with Pakistan's climate being generally warm and wet, with very hot, dry summers. This was supported by our land type classifications, which showed a marked change in the summer months to a prevalence of more dry and bare land cover. The variance of Parthenium over the year matched these trends. Parthenium was more prevalent, with higher variance in cover, in

the winter and spring, becoming less so, with much less variance, in the summer. The Parthenium population always made a quick recovery after the summer, which is consistent with what has previously been found about its abundant distribution of seeds, those seeds' hardiness, and the speed with which Parthenium is able to progress from germination to a mature, seeding plant. Comparing the prevalence of Parthenium in the various land types also revealed patterns consistent with previous findings, with the plant showing a preference for bare and wet soils, which would commonly represent the marginal and wasteland spaces that Parthenium is known to thrive in.

We were also able to gain some insight from our modelling, which showed that the population of Parthenium could be simulated accurately with a relatively simple model. The relationship between the β value, representing "infectiousness", and γ, the number of other individuals "infected", proved to be crucial. This relationship dictated the stable levels of the population, and so was the determining factor in the fluctuations of that population throughout the year. This then indicates one or more factors which influence the prevalence of Parthenium throughout the year, either environmental or inherent to the plant.

6.3 Potential improvements

Our approach yielded some interesting results, however it was not without its drawbacks. Due to time constraints a number of assumptions had to be made, which we would have liked to further investigate. The assumptions around our thresholding of Parthenium presence could use further inquiry, to ascertain which level of threshold provides the best reduction of false positives and false negatives. The levels of clouds cover to be masked out were determined based on the classifier we used to predict the presence of Parthenium. However, there were many instances where the level of cloud cover rendered an observation date unusable. It may prove useful to investigate altering the acceptable levels of cloud cover, for both the predictive classifier and our own data processing. Access to only a few years' worth of data was limiting for our experimentation. It would be valuable to investigate if Sentinel-2 data from 2018 and earlier could be processed to a state suitable for our experimentation. Other satellites could also provide extra data. This would be invaluable in providing a more representative sample size for validating the model's output. It could also enable us to observe trends over longer periods of time than individual years. We would have liked to make better use of the land classifications done in Section 3.3. It is clear from other work, such as that of (Finch, 2023), that the presence of Parthenium is highly dependant on the type of land cover in an area. Certainly the land type would have played a large role in any cellular automata model, which was the cause for carrying out a classification in the first instance. It could also provide useful insight for our simpler population models, had they been separated into different land type populations.

There were also several areas of further study that we would have liked to pursue, had we have more time. Experimenting with the SEIS compartmental model may have

produced more accurate or insightful results, as the latency period of the exposed status may have correlated well with the period between seed dispersal and new plant growth. Additionally, other types of disease models may provide more nuanced or accurate simulation and would be a good avenue for experimentation. It would doubtless be insightful to apply our findings to another observation area, in order to validate our model and investigate for similarities or differences in behaviour. This would help us to establish which patterns were universal, and which unique to the area being observed. We would have liked to take the insights from our population model and apply them to a cellular automata. While we did not have time to pursue this form of model, it would have provided information more immediately useful to those affected by Parthenium in terms of its geographic behaviours and movements. Finally, it would be very helpful to those tackling Parthenium to have some idea of its future dispersal. A model which could predict the behaviours of a population not at equilibrium, but newly introduced to a region would be highly informative.

6.4 Summary

Invasive plants represent a major global challenge to environment, economy and health, and the pursuit of tackling them will only become more relevant in the coming years. Parthenium is an especially troublesome example and one which is likely not done propagating around the world. This makes it crucially important to study such invasive plants and develop methods by which they can be monitored, contained and predicted. We have contributed in small part to this effort, both in validating previous research into the preferred habitats and growth patterns of Parthenium, and in offering a simple yet effective way of modelling its population. We hope that this research can be of use in future work to model and predict the behaviours of Parthenium and other invasive plants, facilitating a healthier environment and peoples, especially for those often worst affected and least able to mitigate the effects.

Bibliography

Adkins S. W., 2010, Parthenium weed (Parthenium hysterophorus L.) research in Australia: new management possibilities., https://www.cabi.org/ISC/abstract/20123079451

Adnan S., Ullah K., Shuanglin L., Gao S., Khan A. H., Mahmood R., 2018, Climate Dynamics, 51, 1885

Arogoundade A. M., Odindi J., Mutanga O., 2020, Geocarto International, 35, 1450

Belgeri A. M., Navie S. C., Adkins S. W., 2011, in 23rd Asian-Pacific Weed Science Society Conference. p. 5

Born W., Rauschmayer F., Bräuer I., 2005, Ecological Economics, 55, 321

Brauer F., van den Driessche P., Wu J., Morel J. M., Takens F., Teissier B., eds, 2008, Mathematical Epidemiology. Lecture Notes in Mathematics Vol. 1945, Springer Berlin Heidelberg, Berlin, Heidelberg, doi:10.1007/978-3-540-78911-6, http://link.springer.com/10.1007/978-3-540-78911-6

Buhle E. R., Margolis M., Ruesink J. L., 2005, Ecological Economics, 52, 355

Büyüktahtakın E., Haight R. G., 2018, Annals of Operations Research, 271, 357

Cam L. M. L., Neyman J., 1967, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Weather modification. University of California Press

Dana E. D., Jeschke J. M., García-de Lomas J., 2014, Oryx, 48, 56

Ehrenfeld J. G., 2010, Annual Review of Ecology, Evolution, and Systematics, 4, 59

Fennel J., Breton R., 2023, Monitoring of an invasive weed at high spatial resolution

Finch E., 2023, Factors affecting invasion and persistence of an invasive weed species: a case study of Parthenium weed, Parthenium hysterophorus L., in Pakistan

Geurts P., Ernst D., Wehenkel L., 2006, Machine Learning, 63, 3

Hartigan J. A., Wong M. A., 1979, Journal of the Royal Statistical Society. Series C (Applied Statistics), 28, 100

Hauser C. E., McCarthy M. A., 2009, Ecology Letters, 12, 683

Head L., 2017, Nature Plants, 3, 1

Hollingsworth M., 2000, Botanical Journal of the Linnean Society, 133, 463

Hulme P. E., 2009, Journal of Applied Ecology, 46, 10

Hulme P. E., et al., 2008, Journal of Applied Ecology, 45, 403

Hulme P. E., et al., 2018, Journal of Applied Ecology, 55, 92

Jones H., 2017, Ecological Modelling, 359, 276

Kganyago M., Odindi J., Adjorlolo C., Mhangara P., 2017, International Journal of Remote Sensing, 38, 5608

Kiala Z., Mutanga O., Odindi J., Peerbhay K. Y., Slotow R., 2020, International Journal of Remote Sensing, 41, 8497

Lovell S. J., Stone S. F., Fernandez L., 2006, Agricultural and Resource Economics Review, 35, 195

Mao R., Nguyen T. L. T., Osunkoya O. O., Adkins S. W., 2019, Austral Ecology, 44, 1111

Mao R., Shabbir A., Adkins S., 2021, Journal of Environmental Management - Volume 279

Marbuah G., Gren I.-M., McKie B., 2014, Diversity, 6, 500

Myers J. H., Cory J. S., 2017, in Vilà M., Hulme P. E., eds, Invading Nature - Springer Series in Invasion Ecology, Impact of Biological Invasions on Ecosystem Services. Springer International Publishing, Cham, pp 191–202, doi:10.1007/978-3-319-45121-3_12, https://doi.org/10.1007/978-3-319-45121-3_12

Myers J. H., Simberloff D., Kuris A. M., Carey J. R., 2000, Trends in Ecology & Evolution, 15, 316

Nguyen T. L. T., Bajwa A. A., Navie S. C., O'Donnell C., Adkins S. W., 2017, Rangeland Ecology & Management, 70, 244

Nigatu L., Hassen A., Sharma J., Adkins S. W., 2010, Weed Biology and Management, 10, 143

Olson R. S., Urbanowicz R. J., Andrews P. C., Lavender N. A., Kidd L. C., Moore J. H., 2016, in Squillero G., Burelli P., eds, Applications of Evolutionary Computation. Lecture Notes in Computer Science. Springer International Publishing, Cham, pp 123–137, doi:10.1007/978-3-319-31204-0_9

Parsons W. T., Parsons W. T., Cuthbertson E. G., 2001, Noxious Weeds of Australia. Csiro Publishing

Pimentel D., et al., 2001, Agriculture, Ecosystems & Environment, 84, 1

Pluess T., Cannon R., Jarošík V., Pergl J., Pyšek P., Bacher S., 2012, Biological Invasions, 14, 1365

Rao P. V. S., Mangala A., Rao B. S. S., Prakash K. M., 1977, Experientia, 33, 1387

Rejmánek M., 2000, Austral Ecology, 25, 497

Remadevi O. K., Sivaramakrishnan V. R., 1996, Impact of diseases and insect pests in tropical forests Proceedings of the IUFRO Symposium, Peechi, India, 23-26 November, 1993, 441

Richards C. L., Schrey A. W., Pigliucci M., 2012, Ecology Letters, 15, 1016

Schulman L. S., Seiden P. E., 1978, Journal of Statistical Physics, 19, 293

Shabbir A., Dhileepan K., Adkins S. W., 2012, Pak. J. Weed Sci. Res., 18: 581-588, Special Issue, October, 2012, p. 9

Shabbir A., Dhileepan K., O'Donnell C., Adkins S. W., 2013, Biological Control, 64, 270

Sharma V. K., Verma P., Maharaja K., 2013, Photochemical & Photobiological Sciences: Official Journal of the European Photochemistry Association and the European Society for Photobiology, 12, 85

Sharman M., Persley D. M., Thomas J. E., 2009, Plant Disease, 93, 708

Simberloff D., 2001, Trends in Ecology & Evolution, 16, 273

Simberloff D., 2003, Conservation Biology, 17, 83

Srivastava V., 2019, CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources, 14

Steve Adkins Asad Shabbir 2014, Pest Management Science, p. 7

Strickland C., 2015, Ecological Modelling, 309-310, 1

Tamado T., Ohlander L., Milberg P., 2002, International Journal of Pest Management, 48, 183

Towers G. H. N., Mitchell J. C., 1983, Contact Dermatitis, 9, 465

Tudor G. D., Ford A. L., Armstrong T. R., Bromage E. K., 1982, Australian Journal of Experimental Agriculture, 22, 43

Vila-Aiub M. M., Vidal R. A., Balbi M. C., Gundel P. E., Trucco F., Ghersa C. M., 2008, Pest Management Science, 64, 366

Vitousek P. M., D'Antonio C. M., Loope L. L., Rejmánek M., Westbrooks R., 1997, New Zealand Journal of Ecology, 21, 1

Vogler W., Navie S., Adkins S., Setter C., 2006, A report for the Rural Industries Research and Development Corporation - RIRDC Publication No 06/130 - RIRDC Project No QDN-7A, p. 51

Weidenhamer J. D., Callaway R. M., 2010, Journal of Chemical Ecology, 36, 59