Remote sensing of invasive plant species:

Optimization of Sentinel-2 and Landsat 8 imagery for enhanced mapping of the invasive *Parthenium hysterophorus* in South Africa

By

Zolo Zime Zinu Serge Kiala

Submitted in fulfilment of the academic requirements for the degree of Doctor of Philosophy in the School of Agricultural, Earth and Environmental Sciences, University of KwaZulu-Natal

Pietermaritzburg

South Africa

March 2020

ABSTRACT

Invasive Plant Species are rapidly spreading worldwide, causing irreversible damage to ecosystem functioning by accentuating the occurrence and severity of fires, altering the dynamics of nutrients, carbon storage, the microclimate and vegetate succession. Due the proliferation of IPSs, global biodiversity has been lost through the homogenization of flora and fauna. Parthenium weed (*Parthenium hysterophorus*) is considered as one of the most noxious IPSs in the world because its adverse impacts on not only, crop, animal and human health, but also on the economy and the environment. Parthenium weed is an upright annual and herbaceous weed of the *Asteracea* family (tribe: *Heliantheae*). Although it is native to neo-tropical regions in central Argentina and the Gulf of Mexico, Parthenium weed has spread to pan-tropical regions. Parthenium weed was first registered at Inanda in the province of KwaZulu Natal in South Africa in 1880. Since then, it has spread to the other provinces, such Mpumalanga, North West and Limpopo.

To optimize Parthenium mitigation, it is necessary to accurately monitor its spread by using cost-effective solutions, such as remote sensing technologies. In this regard, Sentinel-2 and Landsat 8 imagery, which are freely available, were implemented to accurately map Parthenium infestations. However, several challenges related to the mapping of landscapes infested by Parthenium weed, using conventional classifiers, in combination with Sentinel-2 and Landsat 8 imagery, have been overlooked in past studies. For instance, the application of a classifier, which is independent of data characteristics, has not been explored. Meanwhile, it is still not known, which dimension reduction algorithm is appropriate for discarding redundant features from the large volume of Sentinel-2 image data that can be acquired or derived in mapping Parthenium weed. Furthermore, the determination of the temporal window(s) within which the variability of phenological characteristics of Parthenium weed and associated species is the most prominent, and subsequently from which, most accurate maps can be derived, has been overlooked. Therefore, this study endevoured to tackle these issues in order to optimize a Sentinel-2 and Landsat 8 image for more accurate spatial detection of Parthenium weed.

In the first part of this study, the potential of an automated machine learning approach, the Tree-based Pipeline Optimization Tool (TPOT), was explored in mapping Parthenium weed infestations. It was established that the TPOT is an efficient method for automatically selecting and tuning algorithms for Parthenium weed discrimination and monitoring, regardless the data

characteristics. The TPOT model yielded higher overall classification accuracies (88.15% and 74%) than the most robust classifier after manual optimization (84.45% and 68.3%), using a Sentinel-2 and Landsat 8 images, respectively.

Secondly, ten feature selection algorithms, which belong to five groups, namely, sparse learning-based, statistical-based, information theoretical-based, similarity-based and wrappers methods, were compared on Sentinel-2 wavebands and their derived vegetation indices in mapping Parthenium weed, using specific class-based accuracy metrics. The results showed that the investigated feature selection algorithms could increase the classification accuracies of Parthenium weed, in addition to reducing the number of variables or features. The svm-b, a wrapper method, produced the highest classification accuracies, and ReliefF, a similarity-based method, could select the smallest size of the optimal features.

The third part of the study endeavoured to find the temporal window within which variability in the phenological characteristics of Parthenium weed and its associated species is the most asynchronous, and subsequently an accurate map of Parthenium weed can be derived using a Sentinel-2 image. The results showed that most accurate maps of Parthenium weed could be obtained at the beginning of February. Bands such as Blue (490 nm), NIR (835 nm), Red-edge (704 nm) and Green (560 nm) were the most contributing features in the developed models.

In the fourth part of the study, a hybrid feature algorithm was proposed for handling the correlated variables in a multi-date Sentinel-2 image. The proposed approach, which combines ReliefF, svm-b and RF, was compared against its constituent feature selection methods. The multi-date and the single-date images acquired at the beginning of February were also compared. The results showed that the proposed feature selection algorithm selects fewer features than the single feature selection methods, in addition to producing higher classification accuracies (e.g. Overall Accuracy, Producer's and User's Accuracies) than the single-date image. The Overall Accuracy was 86.6%, with 22 optimal features using the proposed approach, whereas it was 84.7% with 35 optimal features using svm-b, 84% with 31 optimal features using ReliefF, 85% with 38 optimal features using RF and 77.6% using the single-date image.

Finally, a hybrid feature selection algorithm and the TPOT were combined in a new algorithm system to explore the capability of the TPOT for handling high dimensional geo-datasets, such as the multi-date Sentinel-2 image. The results showed that the TPOT can be applied on high

dimensional datasets without affecting the classification accuracies. The highest Producer's and User's accuracies of Parthenium weed were achieved, using a multi-date image in combination with the TPOT (90% and 93%). Coupling feature selection with the TPOT reduces the computational costs (17%) at the expense of the classification accuracies.

Overall, this study has proved that, by overcoming some previously overlooked challenges related to weed mapping, a Sentinel-2 image can be optimized and hence, significant improvement of the spatial representation of Parthenium weed in infested landscapes can be achieved. Information on the accurate extent of Parthenium weed is crucial for enhancing decision-making in the management plans.

PREFACE

The work described in this thesis was carried out in the School of Agricultural, Earth and Environmental Sciences, University of KwaZulu-Natal, Pietermaritzburg. The research was undertaken from January 2016 to December 2019, under the supervision of Professor John Odindi and Professor Onisimo Mutanga.

The work in this thesis represents the original work of the author and has never been submitted in any form to any other tertiary institution. Where use has been made of the work of others, it is duly acknowledged in both the text and reference sections of this thesis.

Zolo Zime Zinu Serge Kiala:  Date: _____

As the candidate's supervisors, we certify the above statement and have approved this thesis for submission.

1. Prof. Onisimo Mutanga Signed: _____ Date: _____

2. Prof. John Odindi Signed: _____ Date: _____

DECLARATION 1: PLAGIARISM

I, Zolo Zime Zinu Serge Kiala, declare that:

1. the research reported in this thesis, except where otherwise indicated, is my original work;

2. this thesis has not been submitted for any degree or examination at any other university;

3. this thesis does not contain other persons' data, pictures, graphs, or other information, unless specifically acknowledged as being sourced from other persons;

4. this thesis does not contain any other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

    - their words have been re-written, but the general information attributed to them has been referenced; and

    - where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.

5. this thesis does not contain text, graphics, or tables copied and pasted from the Internet, unless specifically acknowledged and the source being detailed in the thesis and in the references section.

Signed: 

DECLARATION 2: PUBLICATIONS

1. **Zolo Kiala**, Onisimo Mutanga, John Odindi, Kabir Y Peerbhay and Rob Slotow. "The Automated classification of a tropical landscape infested by Parthenium weed (*Parthenium hysterophorus*)", International Journal of Remote Sensing, 41:22, 8497-8519, DOI:10.1080/01431161.2020.1779375

2. **Zolo Kiala**, Mutanga, O., Odindi, J., and Peerbhay, K. (2019). "Feature Selection on Sentinel-2 Multispectral Imagery for Mapping a Landscape Infested by the Parthenium Weed", Remote Sensing, 11(16), 1892.

3. **Zolo Kiala**, Onisimo Mutanga, John Odindi and Cecilia Masemola. "Optimal window period for mapping Parthenium weed in South Africa using high temporal resolution imagery and the ExtraTrees classifier", Biological Invasions, Under Review.

4. **Zolo Kiala**, Onisimo Mutanga, John Odindi, Serestina Viriri and Mbulisi Sibanda. "A hybrid feature method for handling redundant features in a Sentinel-2 multi-date image for mapping Parthenium weed", Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13:3644-3655, DOI:10.1109/JSTARS.2020.3001564.

5. **Zolo Kiala,** Onisimo Mutanga, John Odindi and Romano Lottering. "Exploring the capability of the Tree-based Pipeline Optimization Tool (TPOT) in handling high dimensional multi-date Sentinel-2 image data for mapping Parthenium weed", In preparation.

DEDICATION

This thesis is dedicated to my late twin babies, Joshua and Caleb Kiala.

ACKNOWLEDGEMENTS

The completion of this study was made possible through the contributions of many helping hands whom I would like to acknowledge.

I would like to thank the Southern African Science Service Centre for Climate Change and Adaptive Land Management (SASSCAL) and the Big Data for Science Society (BDSS) program for funding this research study.

To my supervisors, Prof Onisimo Mutanga and Prof John Odindi, thank you for your mentorship and for your trust. The road to the completion of this thesis was very rough, with many unforeseen events, but you kept on supporting me, both morally and financially. You were honest with me when I was not performing. You also taught me to be an independent scientist and a critical thinker. I will never forget your support, which sometimes went beyond the relationship between a student and supervisor during a very difficult time of my life.

I am grateful to my cousin, Ognelet Marie Claude, for your steadfast financial support throughout my studies. Your dream of having at least one person in the family with a PhD degree is now fulfilled, at last. To my wife and my daughter, you have made home such a beautiful place to live in. Your patience and understanding have greatly encouraged me. Certainly, you will be the first to eat the fruits of my labor. Many thanks to my parents and relatives for your patience and moral support. ***Papa and mama, I made it!***

I deeply acknowledge the contribution of precious friends and colleagues during the field work and the revision of the manuscripts emanating from this thesis. To Dr Mbulisi, Dr. Cecilia Masemola and Dr. Kabir Peerbhay, Mr. Brice Gijsbertsen and Mr. Donavan deVos, you have all been awesome! You were always willing to help, despite of your busy timetable.

Finally, I would like to thank my boss, Craig Saunders, and all my colleagues from Quartex Technologies, for your patience. It was not easy to combine work and school, but you understood that it was all worth it, in order to prepare for a possible future career.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

CHAPTER 1. GENERAL INTRODUCTION

1.1 Background

Invasive Plant Species (IPSs) are speedily disseminating across the globe, meanwhile causing irreversible changes in ecosystem structures. Some of the adverse impacts of IPSs include: accentuation of fire occurrence and severity; and alteration of the dynamics of nutrients, carbon storage, microclimate and vegetation succession (Joshi et al., 2004; Huang and Asner, 2009). In contrast to pollution, logging or wildfires, eradication on evaded sites may not be complete, even after spotting and controlling the source of infestation (Huang and Asner, 2009). The proliferation of IPSs has been recognized as the principal driver of biodiversity loss in the world through the homogenization of the fauna and flora (Ustin et al., 2002).

Parthenium weed *(Parthenium hysterophorus*) has the reputation of being one of the most noxious IPSs across the globe, mainly due to detrimental effects to animal and human health, agricultural production, the economy and the environment, among others (Swati et al., 2013). For instance, in the beef and pasture industry, Parthenium weed reduces production by lowering the carrying capacity and stocking rate by about 90% and 80%, respectively, and it increases the controlling costs. This impact is particularly felt in Africa, where land-owners commonly rely on hand-weeding as weed control practices and other weed control methods, such as chemical control, are practically unfeasible, due to financial constraints (Evans, 1997; Strathie et al., 2011). Long exposure to Parthenium weed or its pollen can cause severe allergic reactions in humans. Some of those allergies include hay fever, dermatitis and asthma. When consumed in large amounts by livestock, it becomes toxic and taints the milk and meat, thereby dropping their quality (Strathie et al., 2011). In infested cultivated lands, it may lead to the poor fruiting of legumes, such as Crotalaria and Desmodium. Evans (1997) reported that Parthenium pollen could reduce the chlorophyll content and serve as an alternate host for crop pests. Parthenium weed is aggressive towards co-existing species in invaded ecosystems because of its ecological attributes, such as its fast growth rate, its high reproductive potential, the longevity of buried seeds and interference by allelopathy (Javaid and Anjum, 2005; Strathie et al., 2011; Swati et al., 2013).

Parthenium weed is an upright annual and herbaceous weed of the *Asteracea* family (tribe: *Heliantheae*) (Figure 1.1). It is native to neo-tropical regions in the central Argentina and Gulf of Mexico. To date, it has spread to pan-tropical regions (Javaid and Anjum, 2005; Belz et al., 2007).

Parthenium weed niches are mainly disturbed lands characterized by poor ground cover, including roadsides, railway tracks, construction sites, wastelands, overgrazed pastures and cleared lands (Adkins and Shabbir, 2014). It may germinate, grow and flower at any time of the year under ideal climatic conditions, completing its cycle in four weeks. However, summer is the principal season of growth when temperatures range from 10 to 25^{o}C and rainfall is more than 500 mm per annum (McConnachie et al., 2011; Dinwiddie, 2014). Its aerial parts do not resist frost well, and the majority of plants perish under low winter temperatures, although after a mild winter, regrowth from their old stem bases may place for some plants (Adkins and Shabbir, 2014). Parthenium weed tolerates a wide diversity of soils, ranging from sandy loams to clay; however, it prefers alkaline, black and high fertility cracking clay soils (Swati et al., 2013; Kaur et al., 2014). It is extremely prolific in seed production (about 20 000 seeds per plant). Its seed bank is estimated to more than 340 million seeds per ha in abandoned fields, while 958 million seeds per ha have been found in some sites of South Africa (Goodall et al., 2010). Parthenium seeds can be dispersed by animals (domestic and wild), vehicles, farm machinery and river or flood water (Strathie et al., 2011; Swati et al., 2013). The aforementioned ecological attributes make it very challenging to control in infested areas.

Figure 1.1 Parthenium weed (*Parthenium hysterophorus*) (BioNET-EAFRINET, No_date)

Parthenium weed was first registered at Inanda in the province of KZN in South Africa in 1880, but it spread widely to other provinces during the 1980s after the passage of the tropical cyclone,

dubbed “Demoina” (McConnachie et al., 2011). In the Southern African Plant Invaders Atlas (SAPIA), in which the occurrence of IPSs is recorded in quarter degree grid cells (about 625 km^2 per cell), Parthenium weed has increased from only three cells in 1980 to 76 cells in 2014 (Terblanche et al., 2016). Currently, dense Parthenium weed stands can be found in the Limpopo, KwaZulu Natal (KZN), Mpumalanga, North West and provinces (Goodall et al., 2010; McConnachie et al., 2011; Strathie et al., 2011). Climatic suitability modelling showed that it could potentially infest many more provinces (McConnachie et al., 2011).

Being declared as a Category One weed (CARA, 1983), Parthenium weed constitutes a major threat to croplands and grazing in the northern parts of the KZN province, causing economic losses to the ecosystems (Belz et al., 2007). It has been reported to increase the costs of planting crops and depreciate livestock (meat and milk), crops and seeds due to contamination. When consumed in large quantities by cattle, it becomes toxic, resulting in their death (Lalla et al., 2013: Truter et al., 2014). Furthermore, the province of KZN has globally-recognized biodiversity hotspot areas (e.g. the Maputaland-Pondoland-Albany hotspot and the Isimangaliso Wetland Park). Such areas are known to have high concentrations of the biodiversity under threat. The goods and services provided by hotspot areas are crucial in supporting the country’s development. Parthenium weed has already invaded large areas of the KZN hotspot areas, thereby threatening ecosystems and their functionality. The loss of biodiversity caused by Parthenium invasion will have an adverse impact on many people who have a direct or indirect dependency on natural resources for livelihood (Lalla et al., 2013).

To effectively control and eradicate invasive plant species, managers require accurate spatial information at a relatively low cost (Lawrence et al., 2006). This spatial information, which can be acquired by traditional survey methods or remote sensing techniques, may also assist managers to assess the efficacy of an implemented control practice, to monitor possible future invasions and to assist in identifying species of interest and areas for clearance (Underwood et al., 2003).

In the field of biological invasion, remote sensing technology has gained considerable interest in recent years. For small management areas, survey methods (e.g. hand-mapping and Global Positioning System (GPS)) have been traditionally used. However, they are time-consuming, as well as financially, logistically and technically impractical for large areas (Erinjery et al., 2018).

Unlike survey methods, remote sensing technology offers several well-documented advantages, such as a broad and multispectral data, multi-date, synoptic view and cost effectiveness (Joshi et al., 2004; Lawrence et al., 2006; Resasco et al., 2007). The number of publications, dealing with GIS and remote sensing applications on collecting and analyzing data on invasive alien plants, their abundance, distribution, modelling and factors that influence their distribution, has considerably increased over the last years (Joshi et al., 2004; Tsai et al., 2005; McConnachie et al., 2011; Malahlela et al., 2015; Peerbhay et al., 2016).

As a result of these attributes, remote sensing technology has become the tool of choice in land use and natural resource management, particularly in detecting, mapping and monitoring weeds and their severity (Joshi et al., 2004; Chatziantoniou et al., 2017). It has also been considered as a practical approach for the study of complex geographic terrains and inaccessible environments. Furthermore, it has been proved to unravel spatial heterogeneity through its broad view and detect land cover changes and quantify the rate of changes over time through its multi-temporal nature (Kokaly et al., 2003; Joshi et al., 2004).

The latest advancement of space-sensor technologies has allowed investigations on the potential of new multispectral sensors, such as WorldView-2, GeoEye and RapidEye, in the field of biology invasion (Mullerova et al. 2005; Peerbhay et al., 2016). These multispectral sensors are characterized by high spatial resolution data and additional narrow features or bands (e.g. Red-Edge bands) and have considerably improved the accuracy of traditional multispectral scanners for detecting and discriminating weeds. For example, Lantz and Wang (2013) discriminated the common reed (*Phragmites australis)* from the native wetland plant species, using Worldview-2 image data, with an overall classification accuracy of 94.0%. Peerbhay et al. (2016) mapped the spatial distribution of bugweed (*Solanum mauritianum*) in forest margins, at an overall classification accuracy of 91.33%, using Worldview-2 image data. Although new generation of multispectral sensors, provides the higher classification accuracies than traditional multispectral scanners for mapping weeds, they are often costly for operational applications and do not cover a large area. Therefore, there is a need for a cheaper sensor system with image data that combines the characteristics of the two types of aforementioned multispectral sensors.

Sentinel-2 is an innovative multispectral scanner with a wide-swath (290 km) and fairly high spatial (up to 10 m) and spectral resolution (13 spectral bands). In addition to being freely accessible, Sentinel-2 data have a global revisit time of five days and contain novel spectral bands (e.g. three bands in the red-edge and two bands in the SWIR regions) with discriminatory potential (Immitzer et al., 2016). Sentinel-2 offers unprecedented opportunities for spatially detecting and monitoring invasive species, and its data have been used for estimating biophysical variables (Clevers and Gitelson, 2013; Ramoelo et al., 2015; Sibanda et al., 2015; Majasalmi and Rautiainen, 2016; Vincini et al., 2016), geological applications (van der Meer et al., 2014; Chen et al., 2019; Salehi et al., 2019), urban studies (Pesaresi et al., 2016; Nie et al., 2019; Sun et al., 2019; Tzelidi et al., 2019), tree species mapping (Immitzer et al., 2016, Laurin et al., 2016; Hościło and Lewandowska, 2019; dos Santos et al., 2020), fire management (Fernandez-Manso et al., 2016; Gargiulo et al., 2019; Picos et al., 2019; Roteta et al., 2019) and water body mapping (Du et al., 2016; Wang et al., 2018; Yang et al., 2018; Giuliani et al., 2019). However, in the field of the biological invasion of herbaceous plants, such as Parthenium weed, the application of Sentinel-2 has received less attention.

Over the past few decades, several challenges have been overlooked in studies that attempted to map land-covers, in general, and weeds, in particular, using conventional classifiers and new sensors, such as Sentinel-2 and Landsat 8. For example, studies have shown inconsistencies in the performance of compared algorithms (Nitze et al., 2012; Cracknell and Reading, 2014). A classifier may be suitable for some sites and not for others, hence resulting in poor performance. Generally, the identification of a data-independent algorithm remains a challenge, while the determination of an optimal classification algorithm is often time-consuming and laborious, as it involves a comparison of several manually-generated and complex trials. Therefore, an automated statistical algorithm selection and hyperparameter optimization approach, with a similar or superior classification performance to optimise manually-selected algorithms would be a desirable milestone in Parthenium weed and landscape mapping. Meanwhile, it is still not known, which dimension reduction algorithm is appropriate for discarding noisy data from the great number of Sentinel-2 image data that can be acquired or derived in mapping Parthenium weed. Generally, little attention has been given to multispectral images in dimension reduction related studies, because of the limited number of bands they contain. However, with the proliferation of high

temporal resolution sensors (e.g. Sentinel-2), the volume of image data that can be acquired within a short period has greatly increased (Aires et al., 2016). Although images with large sets of variables are crucial for mapping vegetation, they can misguide classification algorithms (Thejas et al., 2019). Hence, there is a need to evaluate existing feature selection algorithms and develop new ones for Sentinel-2 spectral bands and their derivatives. Furthermore, the spectral signature of herbaceous weeds, such as Parthenium weed is often similar to that of the surrounding herbaceous plant species (Matongera et al., 2017; Wang et al., 2018), resulting in low overall classification accuracies. The determination of the temporal window(s) within which the variability of phenological characteristics of Parthenium weed and associated species is the most prominent, is crucial for deriving accurate maps of Parthenium weed. Therefore, it is necessary to address these concerns for optimizing the Sentinel-2 and Landsat 8 images in delineating landscapes infested by Parthenium weed

1.2 Aim and Objectives

With regard to the discussion above, this study aims at optimizing the Sentinel-2 and Landsat 8 image data for spatially detecting Parthenium weed in KZN, South Africa.

The objectives of the study are:

a) to investigate the potential of a newly developed automated machine learning technique, the TPOT, for mapping a heterogeneous landscape characterized by significant Parthenium weed invasions, using Sentinel-2 and Landsat images. The TPOT automatically searches and finds optimized tree-based pipelines, using a genetic algorithm;

b) to provide a thorough comparison of ten feature selection algorithms (Trace ratio, ReliefF, Gini index, the F-score, Mutual Information, Mutual Information Maximization, Mutual Information Maximization (MIM), $\ell 1$-norm Regularizer , $\ell 2,1$-norm Regularizer , Support Vector Machines Backward) on Sentinel-2 wavebands, coupled with vegetation indices, using specific class-related accuracies metrics, such as f-score;

c) to determine the optimal window period for mapping Parthenium weed within a growing season, based on the most contributing band of a Sentinel-2 image. Mapping of Parthenium weed was carried out for different phenological stages to find that window;

d) to identify a novel hybrid feature method, made of ReliefF, Support Vector Machines Backward and Random Forest, for handling correlated variables in a multi-date Sentinel-2 image in mapping a landscape infested by the common Parthenium weed invasive plant species; and

e) to explore the capability of the TPOT to handle a high dimensional multi-date Sentinel-2 image for mapping Parthenium weed.

1.3 Description of Research area

This study was conducted at two sites (Figure 1.2). The first site was located within the Mtubatuba municipality on the North-East Coast of the KZN province, South Africa (Figure 1.2). In addition to covering an area of 129 km^2, this first research site is characterized by heavy Parthenium infestations. Geological formations, such as basalt, sand and mudstone, underlie the area (Norman and Whitfield, 2006). The annual average rainfall varies from 600 mm to 1250 mm. Temperatures vary in the vicinity of 21°C. Summers are typically warm to hot, while winters are cool to mild (Municipality, 2002). The research area is also characterized by a mosaic of several land use/land covers, including commercial agriculture (e.g. mining, forestry plantations and sugarcane farming), subsistence farming (beans, potatoes, bananas, and cattle), and high- and low-density residential areas (Municipality, 2002).

The second study area was located in the northern part of the KZN province (Figure 1.2). It covers an approximate area of 1660 km^2. The study site is characterized by dry winters and wet summers, with the annual average rainfall and temperatures that vary between 500 mm and 2000 mm, and 13.89^{o}C and 21.7^{o}C, respectively (Kganyago et al., 2018). Geological types of the area include Rhyolite, Siltstone, Sand, Basalt and Arenite. Bushveld, grassland and forest are the dominant vegetation types (Cooley et al., 2002). In contrast to the first study, the majority of land is used for conservation purposes. The choice of this study was determined by the high prevalence of increasingly prolific Parthenium weed infestations.

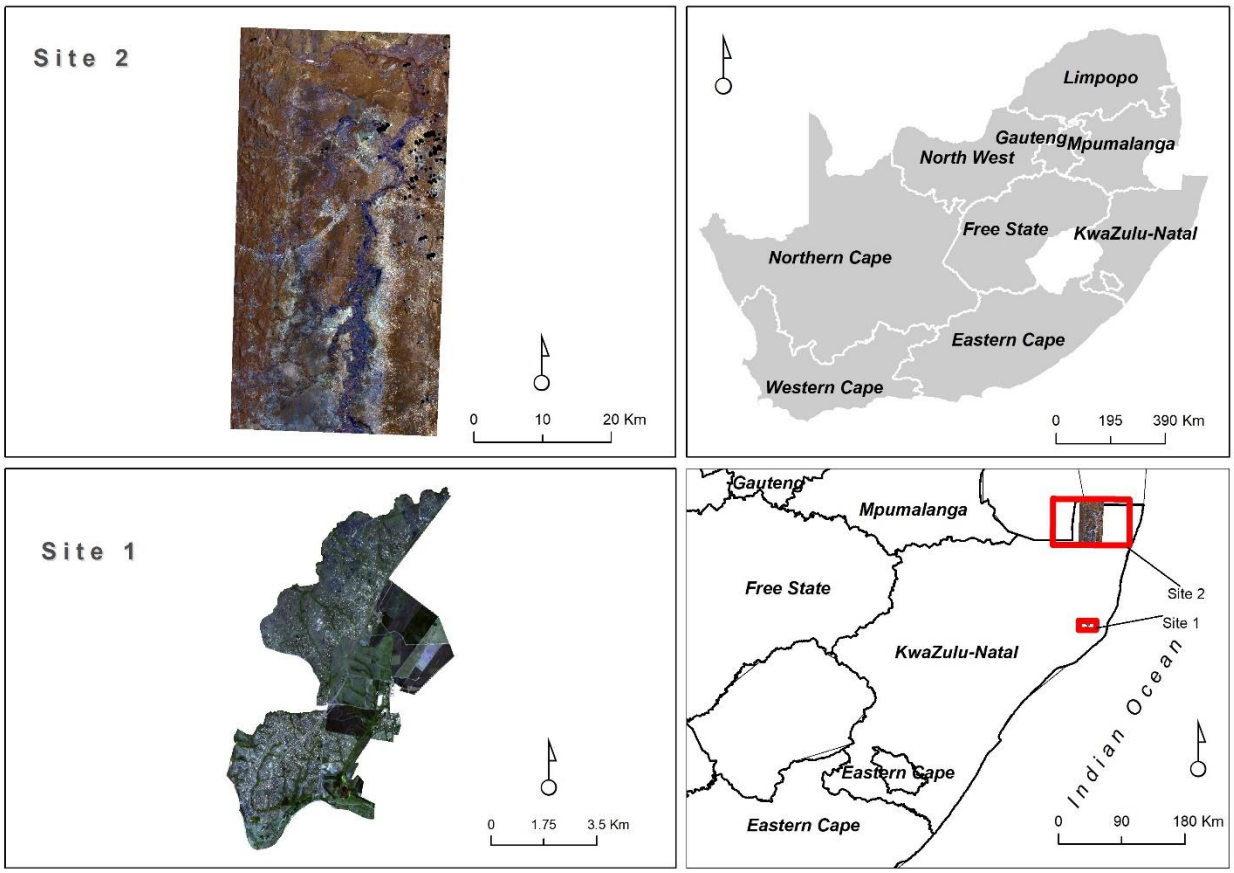


Figure 1.2 Location of the two research sites in KZN, South Africa

1.4 Reference data, image acquisition and pre-processing

Using a high resolution (50 cm) colour orthophotograph (NGI, 2008) of the study area, a broad map of land cover classes was created. Then, conspicuous patches of Parthenium infestation and investigated land cover types (i.e. water bodies, forests, grassland and settlements) were randomly selected across the study area. The selected sites were then located and surveyed in the field, using a differentially corrected Trimble GeoXT hand-held GPS receiver, with an accuracy of about 50 cm when it was placed at the approximate center of a patch. The field work was conducted during summer, between 12th of January and 2nd of February 2017, in the first study area, and throughout February 2014, in the second study area. In every patch, at least 10 m x 10 m quadrats were demarcated (Carter, 1994) using a tape measure and then GPS points were collected for Parthenium invasion and the investigated land cover types. Additional GPS points of investigated land covers were also extracted from the colour ortho-photograph in order to increase the number

of samples. In total, 447 and 326 GPS points were collected in the first and second study areas, respectively, for mapping Parthenium and its co-existing land covers.

Concerning image processing, at the first study area, a Level 1C Sentinel-2A image was acquired on February 19, 2017 under cloudless conditions. Sen2Cor Version 2.3.1 (Louis et al., 2016) software was applied on the image to correct for atmospheric effects. Sen2Cor is a processor that performs the atmospheric correction of Top-Of-Atmosphere (TOA) Level 1C input data, and generates a Level 2A product. The latter product has a Bottom-Of Atmosphere (BOA) corrected reflectance, with three-bands spatial resolution (60, 20, and 10 m). Bands (1, 9 and 10), with a 60 m-resolution, were omitted for this study. At the second study area, an orthorectified Landsat 8 Surface Reflectance Tier 1 scene was acquired on February 12, 2014, using the Google Earth Engine. Subsequent image processing, such as cloud masking, clipping to the extent of the study area boundary and spectral reflectance extraction for model development were performed by using scripts written in Google Earth Engine and Python. The thermal bands of Landsat 8 were not considered in the analyses, due to their low spatial resolution (100 m) and inappropriateness for vegetation mapping.

1.5 Outline of Thesis

The thesis is presented as a collection of research papers addressing each of the objectives listed in Section 1.2. Two research papers have been published and accepted in a peer-reviewed scientific journal, respectively, two other papers are under review and one paper is in preparation. This thesis consists of seven chapters, including the introduction and synthesis chapters.

Chapter Two assesses the TPOT on geospatial data as an alternative to human intervention in creating model pipelines for mapping the Parthenium weed. The TPOT was implemented on Sentinel-2 and Landsat 8 images that were collected at the two different sites.

Chapter Three compares ten feature selection methods on the basis of the f-score of Parthenium weed and the data size for mapping a landscape infested by the Parthenium weed. Random Forest (RF) was implemented to evaluate the different subsets of features selected by the investigated feature selection methods on vegetation indices derived from the Sentinel-2 image data.

Chapter Four determines the temporal window within which the variability of the phenological characteristics of Parthenium weed and associated species is the most prominent for optimizing the discrimination and subsequent mapping of the Parthenium weed. The ExtraTrees classifier (EXT) was compared to RF, in order to find that window.

Chapter Five introduces a new hybrid feature selection algorithm built from the most robust feature selection algorithms that have the previously been compared in mapping a multi-date Sentinel-2 image. The proposed approach was compared against its constituent based on the size of optimal feature subsets and classification accuracies.

Chapter Six explores the capability of The TPOT for handling a large set of multi-date Sentinel-2 image data in mapping Parthenium weed infestations and its co-existing land use/covers. The models created from the multi-date image and optimal features selected from the multi-date image, using an algorithm system that combines feature selection and the TPOT, were compared with the model created from a single-date image.

Chapter Seven summarizes the findings of the preceding chapters and indicates the new directions for future studies.

CHAPTER 2. THE AUTOMATED CLASSIFICATION OF A TROPICAL LANDSCAPE INFESTED BY PARTHENIUM WEED

This chapter is based on: **Zolo Kiala**, Onisimo Mutanga, John Odindi, Kabir Y Peerbhay and Rob Slotow. "The Automated classification of a tropical landscape infested by Parthenium weed (*Parthenium hysterophorus*)", International Journal of Remote Sensing, 41:22, 8497-8519, DOI: 10.1080/01431161.2020.1779375

Abstract

The invasive Parthenium weed (*Parthenium hysterophorus*) adversely affects agricultural productivity, animal and human health, rural livelihoods, local and national economies, as well as the environment. Its fast-spreading capability requires consistent monitoring for adoption of relevant mitigation approaches, potentially through remote sensing. To date, studies that have endeavoured to map Parthenium weed have commonly used popular classification algorithms that include Support Vector Machines and RF classifiers, which do not capture the complex structural characteristics of the weed. Furthermore, the determination of site- or data-specific algorithms, often achieved through the intensive comparison of algorithms, is often laborious and time-consuming. Selected algorithms may also not be optimal on datasets that are collected from other sites. Hence, this study adopted the TPOT, an automated machine learning approach that can be used to overcome high data variability during the classification process. Using Sentinel-2 and Landsat 8 images to map the Parthenium weed, we compared the outcome of the TPOT to the best-performing and most optimized algorithm selected from sixteen classifiers on different training datasets. The results showed that the results showed that the TPOT models, derived from Sentinel-2 and Landsat 8, yielded higher overall classification accuracies (88.15% and 74%, respectively) than models developed with the commonly used classifiers. This study is the first to demonstrate the value of the TPOT in mapping Parthenium weed infestations using satellite imagery. Its adoption would therefore be useful in limiting human intervention, while optimising the classification accuracies for mapping invasive plants. Based on these findings, we propose that the TPOT is an efficient method for selecting and tuning algorithms for Parthenium discrimination and monitoring, and indeed, general vegetation mapping.

Keywords: TPOT, Parthenium weed, Sentinel-2, Landsat 8, image classification

2.1 Introduction

Parthenium weed (*Parthenium hysterophorus*) is an alien plant that is reputed to adversely impact animal and human health, crop production, the environment and local and national economies (Swati et al., 2013). Its negative impacts are particularly damaging in Africa, where hand-weeding is commonly relied upon by farmers, while other weed-control methods, such as the chemical control, are often unaffordable. Examples of impacts include lowering of the stocking rate and the carrying capacity by about 80% and 90%, respectively, in the pasture industry (Adkins and Shabbir, 2014). Large consumption by livestock can become harmful, and taint milk and meat, thereby dropping their quality (Strathie et al., 2011). Prolonged exposure to Parthenium weed or its pollen may cause severe allergic reactions, dermatitis, asthma and hay fever to humans. Parthenium weed may lead to the poor fruiting of legumes, such as Crotalaria and Desmodium in infested cropland. Pollens of Parthenium weed can reduce the chlorophyll content of crops and act as an alternate host for pests (Evans, 1997). The competitive nature of Parthenium weed over surrounding plant species in invaded ecosystems can be attributed to its ecological attributes, such as its fast growth rate, high reproductive potential, longevity of buried seeds and allelopathic suppression (Strathie et al., 2011; Swati et al., 2013). Hence, it is necessary to regularly monitor the weed, so as to adopt optimal mitigation measures.

Remote sensing technology has become a tool of choice by managers of natural resource and land use, particularly in detecting, mapping and monitoring weeds (Joshi et al., 2016). Remote sensing, together with ancillary technologies, such as Geographic Information System (GIS) and GPS, can inform sound planning decisions (Rogan and Chen, 2004). Over the last decades, weed mapping and vegetation species discrimination have significantly improved due to the development of new generation multispectral sensors (e.g. Worldview and GeoEye (Peerbhay et al., 2015) and sophisticated classification algorithms (e.g. RF), Support Vector Machines (SVM) and Artificial Neural Networks (ANN)) (Lass et al., 2005; Rodriguez-Galiano et al., 2012). However, mapping and monitoring Parthenium weed have received less attention, with only a few documented studies in the literature. For instance, Kganyago et al. (2017) applied a hybrid feature selection approach on in-situ hyperspectral data for the determination of a subset of hyperspectral bands that are relevant for discriminating Parthenium weed from its co-occurring plant species. With a subset of 10 wavebands, they achieved an Overall Accuracy (OA) of 80.19% by using SVM on in-situ

hyperspectral data. Kganyago et al. (2018) implemented SPOT 6 and Landsat Operational Land Imager (OLI) image data for mapping the spatial distribution of Parthenium weed patches in the savannah landscapes of the KZN province. They found that SPOT 6 yielded a higher OA (86%) than OLI (83%). Royimani et al. (2018) used SPOT 6 and the RF classifier to detect the spatial and temporal distribution of Parthenium weed. The SPOT 6 image data yielded OA of 68%, 75%, 68% and 73% in 2006, 2009, 2012 and 2016, respectively. Based on the above, it can be concluded that very few sophisticated classification algorithms have been investigated. Hence, approaches that may yield higher classification accuracies are still necessary.

A major problem in the classification process of Parthenium weed and surrounding grass species is their erectophile canopy structure, which is known to absorb more spectral reflectance than planophile plants (Miphokasap et al., 2012; Adkins and Shabbir, 2014). Species-specific radiation, such as Near-Infrared (NIR), is scattered into the lower layers, hence affecting model's precision (Gitelson et al., 2002; Miphokasap et al., 2012). Hence, to accurately map Parthenium weed, it is necessary to test newly-developed or more robust algorithms that have the potential to overcome this problem. However, the performance of most recently-developed algorithms is restricted to hyperparameter tuning and data characteristics (Kiala et al., 2016; Luo, 2016). For example, Cracknell and Reading (2014) compared NB, K-Nearest Neighbours (KNN), RF, SVM and ANN in mapping 13 lithological classes in Australia, using Landsat ETM+. The RF generated the highest classification accuracy. Nitze et al. (2012) used RF, ANN, SVM and Maximum Likelihood (ML) on multi-temporal RapidEye images for crop classification in Canada, SVM outperformed all other tested algorithms. Among others, the above studies have shown inconsistencies in the performance of tested algorithms. Generally, the identification of a data-independent algorithm remains a challenge and the determination of an optimal classification algorithm is often time-consuming and laborious, as it involves a comparison of several manually-generated and complex trials. The process would also require a high skillset in machine learning (Luo, 2016). Therefore, an automated statistical algorithm selection and hyperparameter optimization approach, with a similar or superior classification performance to optimal manually-selected algorithms would be a desirable milestone in Parthenium and landscape mapping.

The TPOT is a new automated machine learning (AutoML) approach developed by Olson and Moore (Olson and Moore, 2016). The approach automatically generates and optimizes tree-based

pipelines, using a genetic algorithm. These pipelines are mainly made of three types of operators namely, supervised classification operators (e.g. Decision Tree), feature processing operators (e.g. RandomizedPCA) and feature selection operators (Recursive Feature Elimination) (Olson and Moore, 2016). Although the TPOT has been used in a wide range of application domains, such as image classification (Olson and Moore, 2016) and disease detection (Olson et al., 2016), it has not been tested against rigorous machine learning analyses where operations, such as feature selection, model optimization, and model selection, are simultaneously implemented. Such investigation would be useful in assessing the efficacy of the approach as an alternative to human intervention in creating model pipelines. In addition, the TPOT is an open-source tool that has never been applied to a geospatial dataset. Therefore, we investigate hereby the potential of the TPOT as an efficient approach for the satellite image classification of a heterogeneous landscape characterized by significant Parthenium weed invasions. Specifically, we test the value of the TPOT on image data from two new generation multispectral sensors, Sentinel-2 and Landsat-8, by comparing it to the best- performing algorithm of sixteen advanced classifiers.

2.2 Materials and Methods

2.2.1 Reference data

In total, 447 and 326 GPS points were collected in the first and second study areas, respectively, for mapping Parthenium and its co-existing land covers. The large sample size, as well as testing on different study sites, improved the reliability and representivity of the tested models. In addition to make the comparison and assessment of the developed models in this study more valuable, these ground reference data were randomly split into three sets of training and test datasets (Tables 2.1 and 2.2) (Pal and Foody, 2010).

Table 2.1 Calibration and test dataset of the land use and land cover classes in Site 1

Land-cover classes	Training set 1 (70%)	Test set 1 (30%)	Training set 2 (50%)	Test set 2 (50%)	Training set 3 (50%)	Test set 3 (70%)	Total
Forest	70	30	50	50	30	70	100
Water body	49	21	35	35	21	49	70
Parthenium Weed	63	27	45	45	27	63	90
Grassland	64	28	46	46	28	64	92
Settlement	66	29	48	48	29	66	95

Table 2.2 Calibration and test dataset for the land use and land cover classes in Site 2

Land-cover class	Training set 1 (70%)	Test set 1 (30%)	Training set 2 (50%)	Test set 2 (50%)	Training set 3 (50%)	Test set 3 (70%)	Total
Bare soil	10	29	19	20	29	10	39
Settlement	12	36	24	24	36	12	48
Dense vegetation	7	23	15	15	23	7	30
Grassland	8	23	15	16	23	8	31
Parthenium weed	19	57	38	38	57	19	76
Shrub	7	22	15	14	22	7	29
Swamp	10	30	20	20	30	10	40
Water	8	26	17	17	26	8	34

2.2.2 Statistical analysis

2.2.2.1 Brief description of used ML techniques

a) *AdaBoost.* The AdaBoost (AD) classifier (Yoav and Robert, 1996) is one of the ensemble boosting algorithms that combines multiple less-performing classifiers in an interactive manner to generate a better-performing classifier. The basic rule of AD is to set the weights of classifiers and the calibration data sample in each iteration, such so that it ensures the accurate prediction of unusual instances.

b) *Naïve Bayes*. The Naïve Bayes (NB) classifier is a set of supervised learning algorithms that apply the Bayes' theorem by assuming that all attributes are independent, given the value class (Zhang, 2004). Three variants of the Naïve Bayes classifier were implemented in this study:

- Bernoulli Naïve Bayes, which is best-suited for data that follow a Bernoulli distribution (Christopher et al., 2008);

- Multinomial Naïve Bayes is used for multinomial distributed data (Christopher et al., 2008); and

- Gaussian Naïve Bayes: the data here are assumed to likely follow a Gaussian distribution (Lou et al., 2014).

c) *Classification and Regression Trees.* Classification and Regression Trees (CART) is a non-parametric supervised learning method that utilizes simple decision rules, inferred from the data features, to predict the value of a target feature (Hastie et al., 2009).

d) *Random Forest*. RF is a combination of decision tree learning, where each learning casts a single vote for the most frequent class to classify an input vector (Breiman, 1996).

e) *Isolation Forest.* Isolation Forest (iF) algorithm is a variant of RF that is used for detecting anomalies in a dataset. Observations are separated by the length of a path in a forest of random trees. Instances with shorter path lengths are likely to be considered as anomalies (Liu et al., 2008).

f) *Extremely random trees.* Extremely randomized trees or EXT classifier is a variant of RF. Unlike RF, it computes an average of votes from randomized decision trees fitted on various sub-samples of the dataset (Geurts et al., 2006).

g) *Support Vector Machines.* First introduced by (Smola and Vapnik, 1997), the principle of SVM is to construct a hyperplane in a high or infinite-dimensional space, which can separate the classes of a dataset in such a way that the distance between the hyperplane and the closest instances from each class is maximized. Once the hyperplane is constructed, SVM classify new data instances by determining on which side of the hyperplane they land. Two variants of SVM were tested in this study: the Linear Support Vector Machines (LSVM) and KSVM.

h) *Discriminant Analysis.* Two variants of Discriminant Analysis (DA) were used in this study: Linear DA (LDA) and Quadratic DA (QDA). The LDA is a classifier that reduces the dimensionality in data by projecting its input into a linear subspace consisting of the directions, which maximizes the separation between classes (Hastie et al., 2009). QDA uses a quadratic decision boundary, obtained by fitting class conditional densities to data points and applying the Bayes' rule (Pedregosa et al., 2011).

i) *Stochastic Gradient Boosting.* The basic principle behind Stochastic Gradient Boosting (SGB) is to sequentially compute simple trees, where each successive tree is built for the

prediction of residuals from the preceding tree, hence yielding an incremental improvement in the model (Sankaran et al., 2008).

j) *K-Nearest Neighbour.* The K-Nearest Neighbor (KNN) works by finding a predefined number of training instances that are closest in distance to the new sample, and it predicts the class from the training instances. When the number of samples is a user-defined constant, the nearest neighbor is known as the k-nearest neighbor (Pedregosa et al., 2011).

k) *Passive Aggressive*. The Passive Aggressive (PA) classifier is an algorithm that is suited to process massive streams of online or live datasets. It is closely related to the perceptron algorithm (or SGB) in that it attempts to solve some of the issues of the perceptron algorithm. For example, given a linear classifier, the loss of function of the perceptron algorithm is difficult to move towards a local optimum (Crammer et al., 2006).

l) *Artificial Neural Network.* The ANN algorithm works as a simulation of human learning processes by establishing and reinforcing linkages between the input and output data. Then, in the absence of training data, those linkages connect the input and output data (Barrett et al., 2014).

2.2.2.2 Automated Machine Learning and TPOT

Thornton et al. (2013) first introduced the Automated Machine Learning (AutoML) system by using classifiers that were implemented in WEKA, and it was later dubbed as the Auto-Weka. According to Thornton et al. (2013), the autoML system is expressed as:

a) a set of classifiers: $\{A^{(1)},\ldots, A^{(R)}\}$ with $\Lambda(j)$ as the domain space of the hyperparameters (μ) for each algorithm $A(j)$;

b) a training dataset $D_{train} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ (Equation 2.1) which is split into $D_{test}^{(1)}, \ldots, D_{test}^{(k)}$ and $D_{train}^{(1)}, \ldots, D_{train}^{(k)}$ such that $D_{train}^{(i)} = {D_{train}}/{D_{test}^{(i)}}$ (Equation 2.2) for $i = 1, \ldots, k$; where k is the number of folds of cross-validation; and

c) the loss of the algorithm $A(j)$ that achieves on $D_{test}^{(1)}$ when it is trained on $D_{train}^{(1)}$ is denoted by $\mathcal{L}(A_{\mu}^{(j)}, D_{train}^{(i)}, D_{test}^{(i)})$.

An autoML system finds the algorithm that reduce the value of this loss $A(j)$: $A^*, \mu_* \epsilon \underset{A^{(j)} \epsilon \mathrm{A},\ \mu \epsilon \Lambda(j)}{argmin} \frac{1}{k} \sum_{i=1}^{k} \mathcal{L}\left(A_{\mu}^{(j)},\ D_{train}^{(i)}, D_{test}^{(i)}\right)$. (Equation 2.3)

The TPOT (Olson and Moore, 2016) is a novel AutoML that applies GP to optimize machine learning pipelines of the sklearn Python library for classification and regression problems. The following pipeline operators are implemented in the TPOT: pre-processors, decomposition, feature selection and prediction. During the optimization process, a subset of the ML algorithms is defined as GP primitives, which are organized as in a tree structure to form individuals. To obtain the optimal combination of processes, GP optimizes the number and order of pipeline operators, addition to each operator's parameters (Sohn et al., 2017). More details on the tool can be found in Olson and Moore (2016). An example of the TPOT workflow is illustrated in Figure 2.1.

In this study, the choice of TPOT parameters was premised on the assumption that better results are achieved with more allocated Central Processing Unit (CPU) time (Hutter et al., 2019). Therefore, parameters such as "generations", "population_size" and "verbosity" were set to 500, 100 and 2, respectively. Furthermore, a "random_state" parameter was added to the Python code containing the best pipeline generated by TPOT, to allow for replication.

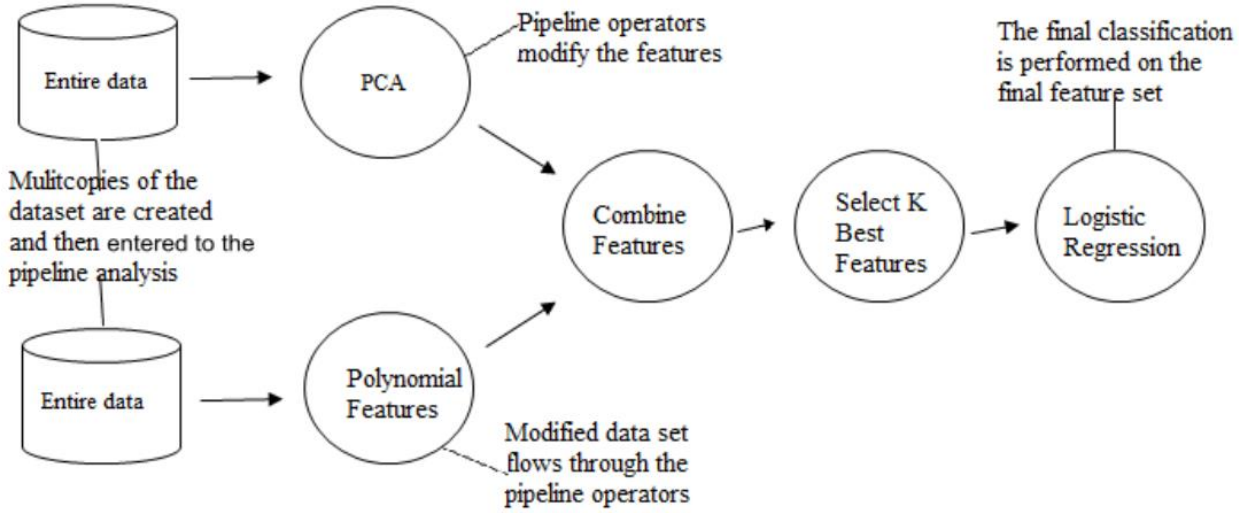


Figure 2.1 The TPOT flowchart (Source: Olson and Moore, 2016)

2.2.3 Data analysis

2.2.3.1 Model comparison

Sixteen classifiers were applied on the reflectance values extracted from Sentinel-2 and Landsat 8 band images to model the spatial distribution of Parthenium weed and other investigated land cover types. The developed models were evaluated on the test dataset, using the overall classification accuracy as a metric. Thereafter, the model that yielded the highest overall classification accuracy was further optimized by first applying feature selection and then tuning its hyper-parameters. Feature selection and hyperparameter tuning are typical preprocessing methods that are for improving the accuracy of a predictive model (Brownlee, 2013). A tree-based feature selection was implemented to generate a spectral band importance ranking (Saeys et al., 2007). Thereafter, a for-loop was used to iterate over all the band subsets. Starting with all the bands and ending with the least important band, models were trained and evaluated on the test dataset, using the overall classification accuracy to find the optimal subset of bands (Brownlee, 2013). The following optimization methods were implemented: a grid search (Hsu et al., 2003), a random search (Bergstra and Bengio, 2012) and Bayesian optimization (Snoek et al., 2012). Finally, the classification accuracies of different optimized models were compared to the accuracies of the model created by using the TPOT. In addition to the classification accuracies, computational complexity was evaluated to determine the time needed to run each optimization method.

2.2.3.2 Model assessment

To evaluate the classification accuracies of optimized models and the TPOT models on the test dataset, estimated classes were cross-tabulated against the ground-sampled classes for corresponding pixels in a confusion matrix. From the confusion matrix, metrics such as the OA, the User's Accuracies (UA), and the Producer's Accuracies (PA) were mathematically derived (Lunetta and Lyon, 2004). The OA indicates the accuracy of the entire classified map. The User's Accuracies refers to the probability that a pixel labeled as a certain class on the map represents that class on the ground. The Producer's Accuracies represents the pixels that belong to a ground-sampled class which fail to be classified in the correct class. Furthermore, the difference in accuracies between the best-performing classifier and the TPOT models was statistically compared, using the Wilcoxon test (Hogg and Craig, 1995). A program that compares algorithms as implemented in the sklearn (Version 0.20) with their default hyperparameters was written in Python (Version 3.6). According to Ahmad et al. (2017), default hyperparameters often yield excellent results. The Python libraries that were used are as follows: sklearn for machine learning implementation and model optimizations (grid search and random search), gdal for producing maps, TPOT and OptML for Bayesian optimization.

2.3 Results

2.3.1 Comparison of algorithms

Figure 2.2 illustrates the overall classification accuracies of different classifiers (with their acronyms in Table 2.3) on the three training sets for the two sites. In the first site, SGB was the best-performing classifier on the first training set (70% of the dataset), with an overall classification accuracy of 82.96%. SGB also outperformed other classifiers on the second training set (50% of the dataset), with an overall classification accuracy of 80.08%. On the third training set (30% of the dataset), the model that was produced by using the EXT classifier was the most accurate (79.76%). iF, Bernoulli naïve Bayes (BE) and Kernel Support Vector Machines (KSVM) produced the least accurate models on all the training sets. In the second site, RF and KNN outperformed the remaining classifiers. On the second and third datasets, RF yielded an overall classification accuracy of 67.7% and 65. 8%, respectively, while on the third dataset, KNN outperformed the rest of the classifiers with an OA of 63.8%. As in the first site, iF, BE and KSVM

were among the least-performing classifiers. Based on their superior performance, the SGB and EXT classifiers in the first site, and RF and KNN in the second site were selected for further analyses.

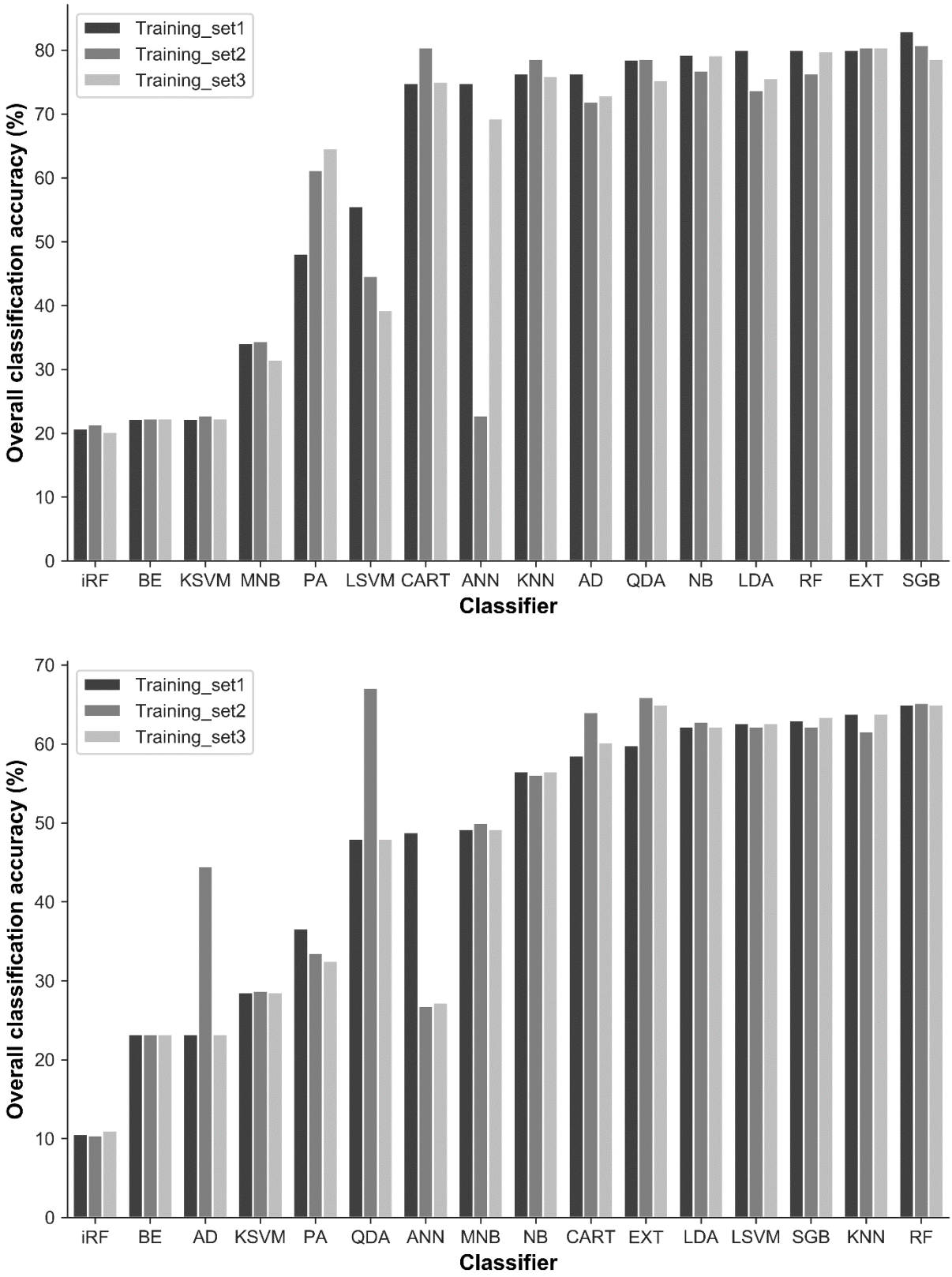


Figure 2.2 Overall accuracy for compared algorithms in Sites 1(a) and (b)

Table 2.3 Acronyms of the compared algorithms

No	Acronym	Algorithm
1	iF	Isolation forest
2	BE	Bernoulli naïve Bayes
3	AD	AdaBoost
4	KSVM	kernel support vector machine
5	PA	Passive aggressive
6	QDA	Quadratic discriminant analysis
7	ANN	Artificial neural network
8	MNB	Multinomial naïve Bayes
9	NB	Naïve Bayes
10	CART	Classification and Regression Trees
11	EXT	Extremely random trees
12	LDA	Linear discriminant analysis
13	LSVM	linear support vector machine
14	SGB	Stochastic gradient boosting
15	KNN	k-nearest neighbour
16	RF	Random forest

2.3.2 SGD and EXT optimization and TPOT

2.3.2.1 The feature selection

Figure 2.3 illustrates the overall accuracy for each band subset. Feature selection in this study did not improve the classification model for mapping Parthenium weed and surrounding land cover types on the three training sets. All ten and seven bands of Sentinel-2 and Landsat 8, respectively, could be used for model development.

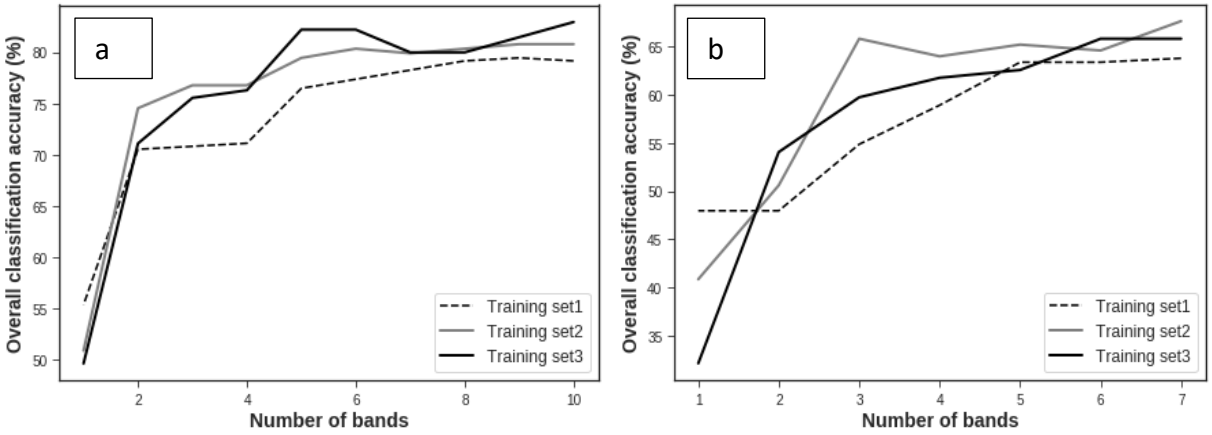


Figure 2.3 Relationship between overall classification accuracy and subset of bands in Sites 1 (a) and 2 (b)

Figure 2.4 shows the relative importance of each band in the SGB and EXT models for the first site, using the Sentinel-2 image, and RF for the second site, using the Landsat 8 image. Overall, Bands 4 (665 nm), 2 (490 nm), 11 (1610 nm), 8a (865 nm), 8 (842 nm), 6 (740 nm) and 3 (560 nm) were the significant contributors to the models created from SGB and EXT classifiers, with the feature importance measures above 0.05, using Sentinel-2 image. Bands 7 (186.66 nm), 6 (84.72 nm), 4 (37.47 nm) and 5 (28.25 nm) contributed the most in the model created by RF, using the Landsat image.

Using the Sentinel-2 image, all the bands had feature importance measures above 0.05 on the first training set. Bands 7 (783 nm), 12 (2190 nm), 5 (705 nm) showed little feature importance measures on the third training set. Band 11 (1610 nm) had high feature importance measures, regardless of the training size, followed by Bands 4, 2 and 8a. Using the Landsat image, the feature importance measure of all the bands was above 0.05 on all the datasets. Bands 1 and 2 contributed the least to the RF and KNN models.

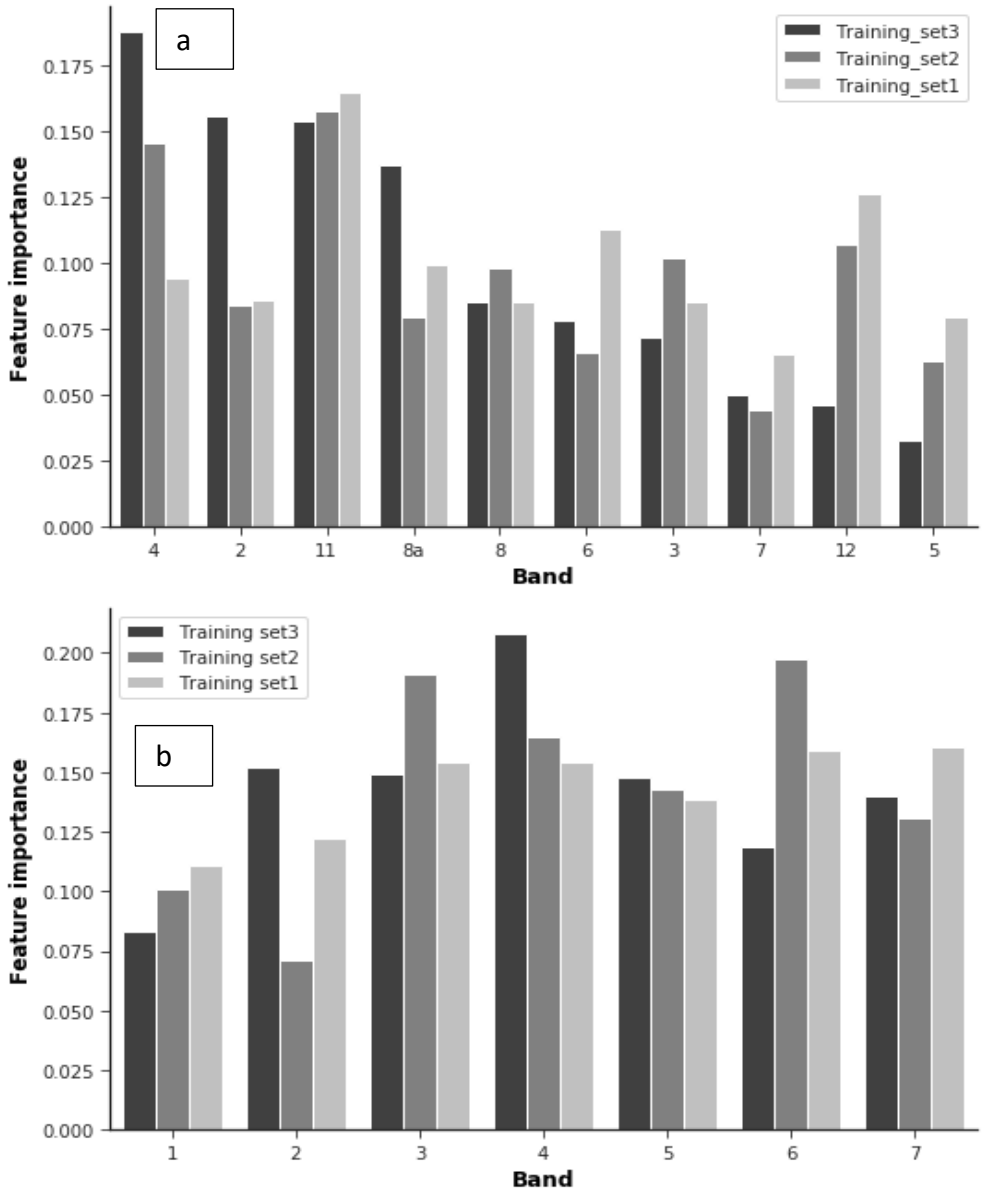


Figure 2.4 Feature importance of spectral bands for Sentinel-2 (a) (Band 4 = Red; Band2 = Blue; Band 11 = SWIR1; Band 8a = NIR2; Band 8 = NIR1; Band 6 = Red edge2; Band 3 = Green; Band 7 = Red Edge3; Band 12 = SWIR2; Band 5 = Red edge1) and for Landsat-8 (b) Band 1 = Coastal; Band 2 = Blue; Band 3 = Green; Band 4 = Red; Band 5 = NIR; Band 6 = SWIR1; Band 7 = SWIR2)

2.3.2.2 Hyperparameter tuning

Table 2.3 illustrates the classification accuracies of SGB, EXT (without optimization and optimized with the grid search, random search and Bayesian methods) and TPOT models, using the Sentinel-2 image. The SGB and EXT's hyperparameters, such as the number of estimators (or number of trees), the maximum number of features ("max_feature"), the minimum number of samples required to split an internal node **("**min_samples_split"), the minimum number of samples needed to be at a leaf node ("min_samples_leaf"), as implemented in the sklearn library, were selected after several trials, and then used for model optimization. Table 2.4 displays the classification accuracies of RF and KNN (without optimization and optimized with the grid search, random search and Bayesian methods) and TPOT models, using the Landsat 8 image. The RF was optimized the same way as the aforementioned classifiers. The optimized parameters of KNN included the number of neighbors ('n_neighbors'), the weight function ('weights') and the weight function ('metric').

Overall, there was an increase in the overall classification accuracy when an optimization method was implemented on most of the training sets. Of all the optimization methods, a random search produced the highest increase, reaching up to 2.1 % on all the datasets. For example, using Landsat 8, RF yielded an overall classification accuracy of 65.9% on the third training dataset without optimization, while it was 68.3% with the random search. However, there were also cases where model optimization did not improve the classification accuracies. The TPOT models (Wilcoxon test, $p < 0.05$) significantly outperformed models of the investigated classifiers without optimization on all the training sets. When compared to the optimization methods, the TPOT models significantly outperformed ($p < 0.05$) models of the best-performing optimization method, which was a random search, on the first and second training sets of the first site, and on the first and third dataset of the second site. On these training sets, the increase in overall classification accuracy ranged between 3.7% and 4.47% respectively, using Sentinel-2, and from 3.4 to 5.7%, using Landsat 8. Although the TPOT was superior to the random search, the difference was not significant ($p > 0.05$) on the third and the second training sets of the first and second sites, respectively.

With regard to specific class accuracy, Parthenium weed was most accurately mapped with the TPOT. For instance, using the Sentinel-2 image, the PA and UA values of TPOT models ranged from 66% to 80% and, from 75% to 81%, respectively, while random search models yielded PA and UA values ranging from 63% to 74% and 69% to 74%, respectively. Using the Landsat 8 image, the TPOT models yielded PA and UA values ranging from 66% to 74% and from 70% to 78%, respectively, while random search models produced PA and UA values ranging from 66% to 68% and from 68% to 79%. Parthenium weed was the most accurately modelled on the first training set, using the TPOT in both sites. The maps of different models are displayed in Figures 2.5 and 2.6.

Table 2.3 Error matrix of the classified map of Parthenium weed and coexistent land cover classes for the first (a), second (b) and third training set (c) in Site 1

a

	SGD		BayesOpt		Grid search		Random search		TPOT	
	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA
Forest	97	97	88	97	94	97	88	97	97	97
Water	100	90	100	90	100	95	100	90	100	95
Parthenium	61	70	68	70	69	67	74	74	73	81
Grassland	86	68	87	71	87	71	87	71	95	71
Settlement	79	90	81	90	74	90	79	90	82	97
OA (%)	82.97		83.7		83.7		84.45		88.15	

b

	SGD		BayesOpt		Grid_search		Random-search		TPOT	
	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA
Forest	84	94	84	92	84	92	82	92	82	94
Water	100	91	100	94	100	94	100	94	100	94
Parthenium	62	76	63	71	63	73	63	69	80	80
Grassland	78	61	80	61	80	61	79	65	91	67
Settlement	89	83	82	85	84	85	85	85	80	92
OA (%)	80.8		80.36		80.8		80.8		85.27	

c

	EXT		BayesOpt		Grid_search		Random-search		TPOT	
	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA
Forest	77	95	60	93	75	92	78	96	91	92
Water	100	98	96	96	100	98	100	98	100	98
Parthenium	64	65	63	32	68	68	70	71	66	75
Grassland	79	61	75	62	80	65	85	67	84	68
Settlement	84	83	70	73	87	83	87	85	84	89
OA (%)	79.76		70.83		80.65		82.74		83.93	

Table 2.4 Error matrix of the classified map of Parthenium weed and coexistent land cover classes for the first (a), second (b) and third training set (c) for Site 2

a

	RF		BayesOpt		Grid search		Random search		TPOT	
	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA
Bare soil	41	38	58	38	60	41	59	34	64	58
Settlement	79	72	82	75	85	64	82	78	86	80
Dense vegetation	59	70	71	87	63	74	71	87	64	100
Grassland	39	30	39	39	42	48	39	39	86	67
Parthenium	77	75	70	77	74	79	69	79	69	78
Shrub	53	73	56	68	57	77	56	68	86	67
Swamp	77	90	72	87	78	93	74	87	80	67
Water	80	62	88	58	79	58	88	58	80	80
OA (%)	65.9		67.9		68.3		68.3		74	

b

	KNN		BayesOpt		Grid search		Random search		TPOT	
	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA
Bare soil	43	41	43	34	44	41	50	38	57	41
Settlement	79	72	81	69	81	69	89	69	85	78
Dense vegetation	71	65	75	65	77	74	77	74	67	70
Grassland	33	35	36	39	35	30	34	43	38	39
Parthenium	68	68	63	65	65	70	66	72	74	70
Shrub	53	82	46	77	50	77	55	77	55	77
Swamp	73	73	70	77	72	77	74	83	79	90
Water	89	65	89	62	89	62	89	62	89	62
OA (%)	63.8		61.8		63.8		65.9		69.2	

c

	RF		BayesOpt		Grid search		Random search		TPOT	
	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA
Bare soil	53	40	67	30	62	25	46	30	64	45
Settlement	70	79	64	75	67	58	68	62	69	75
Dense vegetation	76	87	82	93	71	67	75	80	59	87
Grassland	43	38	60	38	25	25	42	50	54	44
Parthenium	68	68	59	71	59	68	68	68	70	74
Shrub	79	79	75	86	67	86	75	86	85	79
Swamp	72	65	68	75	77	85	78	90	82	70
Water	71	88	88	82	71	88	88	88	79	88
OA (%)	67.7		68.3		62.8		68.3		70	

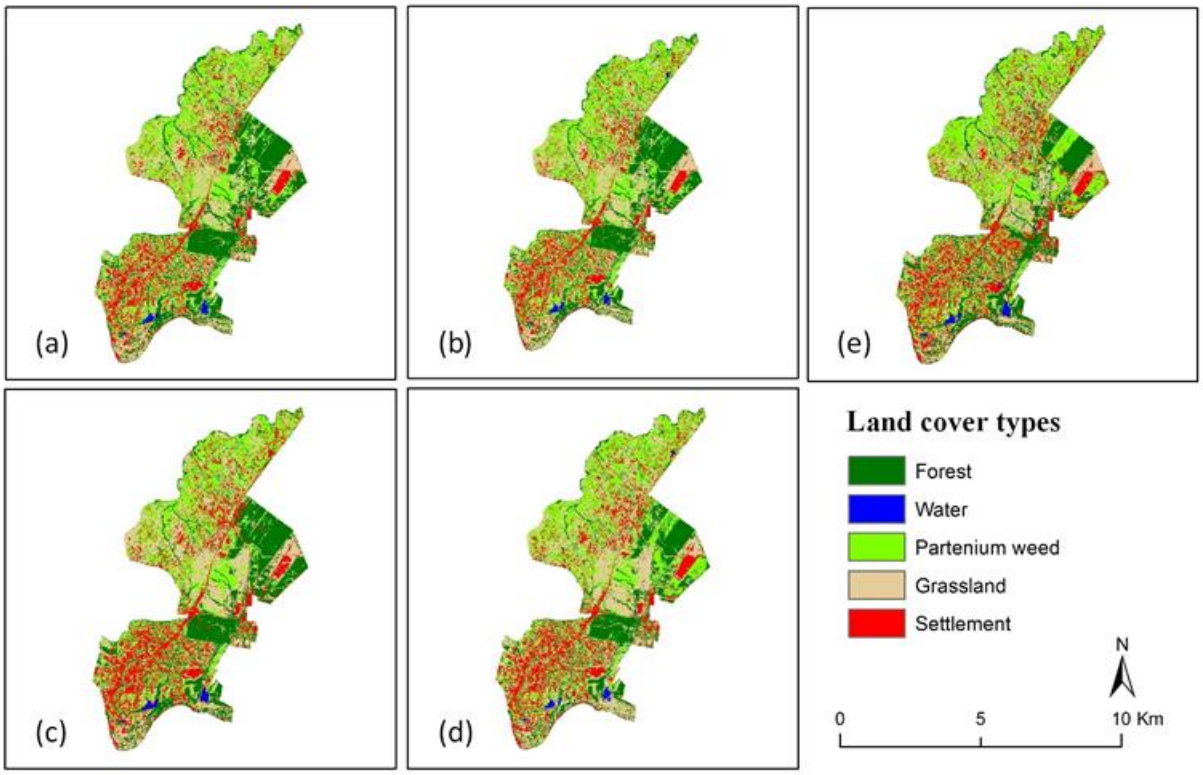


Figure 2.5 Classified map of Parthenium weed and surrounding land cover classes produced from SGB (a), SGB tuned with random search (b) on the first training set, and the TPOT on the first (c), second (d) and third training set (e) in Site 1

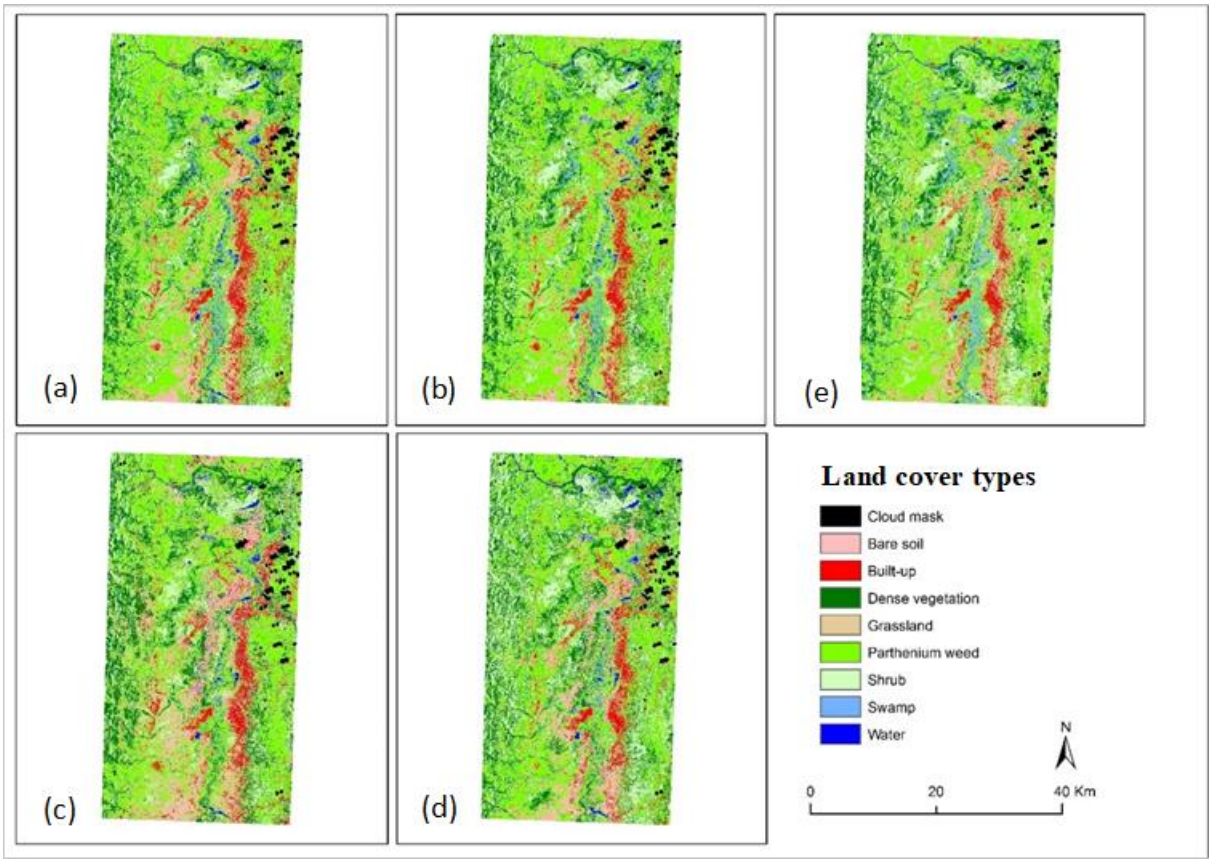


Figure 2.6 Classified map of Parthenium weed and surrounding land cover classes produced from RF (a), RF tuned with grid search (b) on the first training set, and the TPOT on the first (c), second (d) and third training set (e) in Site 2

2.3.3 Computational complexity analysis

Figure 2.7 illustrates the required time for running investigated classifiers and optimization methods. The Bayesian, grid search and random search optimization and investigated classifiers did not require a lot of time, with the number of hyper-parameters involved during the tuning process. However, the TPOT required over 40000 seconds (i.e. 11 hours) to find a suitable pipeline for our datasets. The sample size was directly linked to the computational cost of the TPOT.

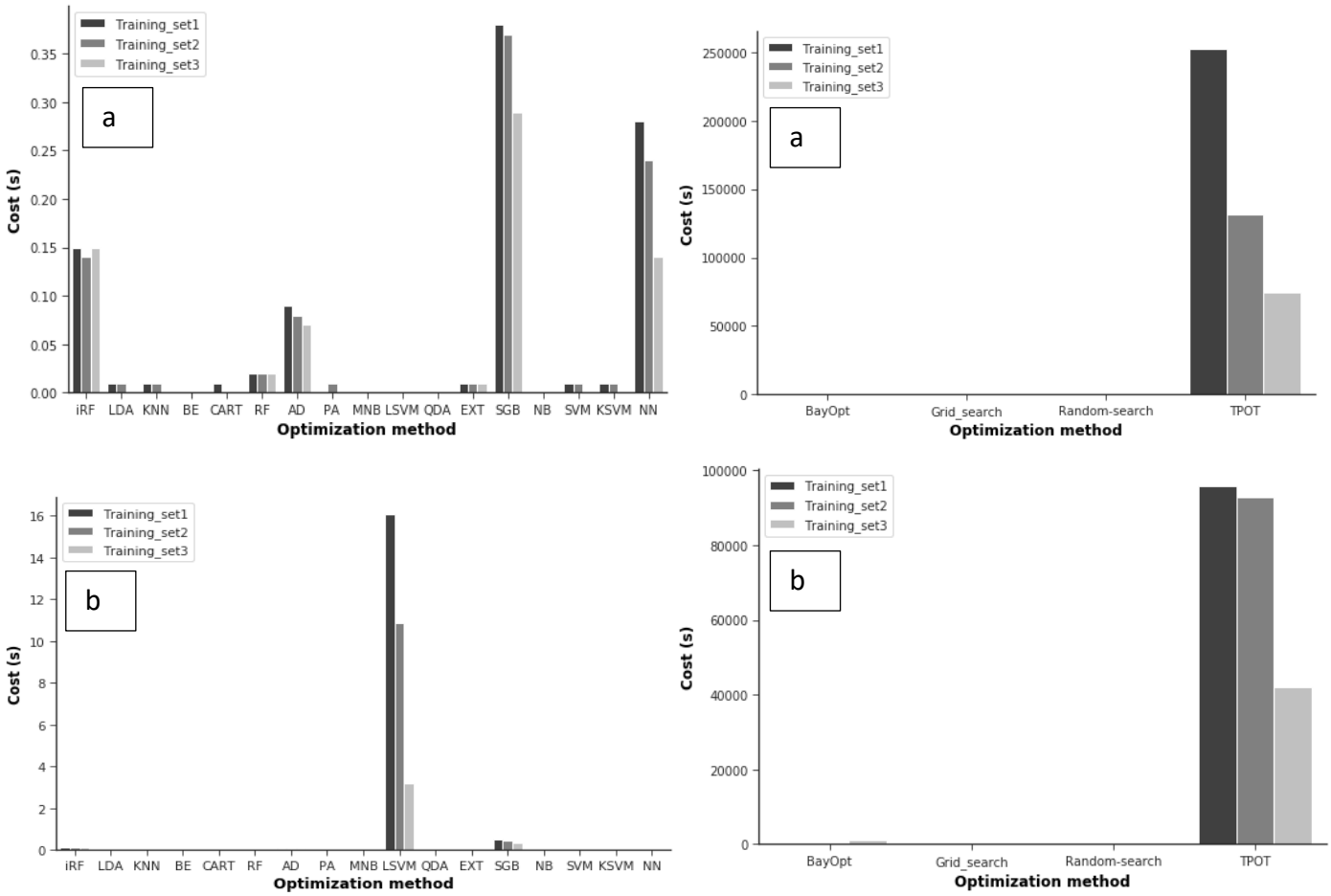


Figure 2.7 Computational cost (in seconds) of investigated classifiers and optimization methods in Site 1 (a) and Site (b)

2.4 Discussion

This study investigated the potential of Sentinel-2 and Landsat 8 images combined with the TPOT, open-source dataset and open-source methods, respectively, in spatially detecting Parthenium weed infestation in relation to existing land covers. To evaluate the efficacy of the TPOT as an automated method, it was compared with a benchmark algorithm derived from the comparison of seventeen algorithms, and then it was tuned, using random-search, grid search and Bayesian optimization. The dataset was split into three different training sets to assess the performance of the TPOT under different scenarios.

2.4.1 Comparison of individual classifiers

Among the compared classifiers, in the first site, SGB produced the highest overall classification on the first and second training sets (70% and 50% of the dataset). EXT classifier yielded the highest overall classification on the third training dataset (30% of the dataset). In the second site, RF and KNN outperformed other classifiers. Nevertheless, the SGB and EXT models were still among the most accurate in classifying a landscape infested by the Parthenium weed. These findings concur with existing literature; for example, Lawrence et al. (2004) found that SGB improved the overall accuracy of the IKONOS and Probe-1 image data from 84% to 95% and from 83% to 93%, respectively, when compared to a standard classification tree analysis. Freeman et al. (2015) found that SGB and RF generated similar performances in predicting tree canopy cover over four sites, using remotely sensed and ancillary data. Other studies (Chirici et al., 2013; Hitziger and Ließ, 2014) have noted its superior performance over the RF classifier. However, its adoption in land use/cover classification using satellite imagery, especially vegetation mapping and discrimination, has been limited (Baker et al., 2006). Regarding EXT, Barrett et al. (2014) reported that EXT, though under-used, out-performed SVM and RF in discriminating grassland types, using ancillary, multi-temporal, multi-sensor radar spatial datasets. As mentioned earlier, these findings confirm the strict dependency of algorithm performance on data characteristics (Kiala et al., 2017).

2.4.2 Contribution of spectral bands in the SGB and RF models

Bands 4 (Red), 2 (Blue), 11 (SWIR), 8a (NIR), 8 (NIR), 6 (Red-Edge 2) and 3 (Green) of models using Sentinel-2, and Bands 7 (SWIR-2), 6 (SWIR-2), 4 (Red) and 5 (NIR) of models using Landsat 8 were the most optimal in discriminating Parthenium weed from the surrounding land cover types on all the training sets. It is worth noting that Bands 2, 3, 4 and 8 of Sentinel-2 imagery have a higher spatial resolution. This highlights the contribution of spatial resolution in mapping weed-infested landscapes. Immitzer et al. (2016) reported a similar finding when mapping trees and crops. A high spatial resolution has also been advocated as a requirement for detecting weed infestation (Huang and Asner, 2009). Bands in the SWIR region have been proven to possess a high discriminatory power in the classification process (Adam et al., 2012; Atkinson et al., 2014; Immitzer et al., 2016). Immitzer et al. (2016), for instance, found that Band 11 of Sentinel-2 (1610 nm) was one of the most important bands in classifying tree species and crops. Kganyago et al.

(2017) applied a feature selection on hyperspectral data, and found that useful spectral bands in discriminating Parthenium weed from coexistent species were situated in the NIR and SWIR. This study has also confirmed the significance of Red-Edge bands of Sentinel-2 imagery (e.g. Band 6) in discriminating land cover types, as was reported in other studies (Zhu et al. 2007; Shafri and Hamdan, 2009). Previous studies have confirmed its strong correlation with plant biochemical (e.g. foliar chlorophyll content) and biophysical parameters (Mutanga and Skidmore, 2007). As these parameters differ according to plant species (Kanke et al., 2012), Red-Edge has been a useful variable for, amongst others, vegetation discrimination (Kim and Yeom, 2014).

2.4.3 Comparison between investigated algorithm and TPOT models

Overall, we demonstrated that the TPOT models developed from the Sentinel-2 and Landsat 8 image data could yield high accuracies for mapping Parthenium weed and coexistent land cover types (Tables 2.3 and 2.4). For instance, the overall classification accuracy reached 88.15% and 74% where the training set was made of 70% of the dataset using both satellite image data, respectively. Accuracies from Sentinel-2 image data were notably superior to those of other studies that endeavored to map Parthenium weed. For example, by reducing the dimensions of hyperspectral data, Kganyago et al. (2017) achieved an overall accuracy of 80.19% in the discrimination Parthenium weed from coexistent plant species (Kganyago et al., 2018). Royimani et al. (2018) achieved an overall classification accuracy of 68%, 75%, 68% and 73% for mapping Parthenium weed-infested landscapes in 2006, 2009, 2012 and 2016, respectively.

The TPOT models outperformed all the tuned models using investigated classifiers, by automatically discovering better-performing pipelines (Tables 2.3 and 2.4). Differences in accuracy were also statistically significant ($p < 0.05$). The superiority of the TPOT, in comparison to machine learning techniques, was reported in other studies. For instance, Olson and Moore (2016) compared the TPOT to RF on 149 benchmarks. They found that the TPOT outperformed RF on 25 benchmarks and performed equally and worse on 125 and 21 benchmarks, respectively. Sohn et al. (2017) found that an improved version of the TPOT, The Tree-based Pipeline Optimization Tool Multifactor Dimensionality Reduction (TPOT-MDR), outperformed a tuned logistic regression and XGBoost classifiers. However, the TPOT is computationally more complex than the tuning methods used in this study (Figure 2.7). According to Feurer et al. (2015),

automated machine learning methods are not intended to run for a few minutes. As a rule of thumb, the more CPU time is allocated for a data analysis, the better the result.

The findings of this study prove that Parthenium weed maps can be automatically generated with a high accuracies and limited human intervention. This eases the process of choosing the machine learning and preprocessing techniques to use for a given landscape that is infested by the Parthenium weed. This task is particularly challenging for novice and non-specialist machine learning practitioners (Feurer et al., 2015). Moreover, the classification accuracies of Parthenium weed maps can be compared more efficiently. For instance, given a task of classifying Parthenium weed and coexistent species in different sites using a classifier. Due to the site characteristics related to soil, or surrounding vegetation complexity (van der Walt and Barnard, 2006), the classifier may be suitable for some sites and not for others, hence resulting in a poor performance. However, by using the TPOT, it would likely be possible to find the best-performing classifier for every site. This may be very useful in mapping Parthenium weed; for example, in tropical regions at regional scale, particularly where the mapping process is undertaken on different tiles or sensor imagery collected from different years (Franklin and Wulder, 2002). The use of one classifier may not be efficient in such diverse regions. With the TPOT, accurate spatial extent of Parthenium weed across different tiles can be achieved. Therefore, planning control and mitigation on a regional scale would also improve accordingly. Furthermore, the application of the TPOT can be extended to mapping other invasive species.

2.5 Conclusions

This study has demonstrated the potential of Sentinel-2 and Landsat 8, in conjunction with the TPOT, in mapping Parthenium weed infestation in a heterogeneous landscape. The following conclusions can be drawn:

a) the TPOT is a reliable method for the classification of geospatial data, as it can automatically find a suitable pipeline for a dataset, with results that are superior to manually-selected and parameter-tuned algorithms;

b) Sentinel-2 image data can be used to map Parthenium weed infested landscapes with reliable classification accuracies; and

c) SWIR, NIR, red-edge bands and bands with the finer spatial resolution are the most valuable in modeling the spatial distribution of Parthenium weed invasion, with high ranking scores.

This study is valuable as it spares remote sensing practitioners laborious time, when searching and optimizing algorithms, as it can be done automatically and efficiently using the TPOT. We recommend that future studies should consider implementing the TPOT in regression problems and that it should be expanded by incorporating more remote sensing applications. Moreover, they should look at the relationship between the CPU time and classification accuracies to find the cut-off time at which a TPOT model can achieve acceptable classification accuracies.

2.6 Acknowledgements

The authors would like to thank the SASSCAL and Ognelet Marie Claude for funding the project. We are also grateful to Carlos de la Torre and Jason Brownlee for sharing their codes online. Many thanks to the anonymous reviewers for the time they spent in polishing up this work.

CHAPTER 3. FEATURE SELECTION ON SENTINEL-2 MULTISPECTRAL IMAGERY FOR MAPPING A LANDSCAPE INFESTED BY PARTHENIUM WEED

This chapter is based on: **Zolo Kiala**, Mutanga, O., Odindi, J., and Peerbhay, K. (2019). "Feature Selection on Sentinel-2 Multispectral Imagery for Mapping a Landscape Infested by Parthenium Weed", Remote Sensing, 11(16), 1892.

Abstract

In the recent past, the volume of spatial datasets has significantly increased. This is attributed to, among others, the higher sensor temporal resolutions of the recently-launched satellites. The increased data, combined with the computation and possible derivation of a large number of indices, may lead to high multi-collinearity and redundant features that compromise the performance of classifiers. By using dimension reduction algorithms, a subset of these features can be selected, hence increasing their predictive potential. In this regard, an investigation into the application of feature selection techniques on multi-temporal multispectral datasets, such as Sentinel-2, is valuable in vegetation mapping. In this study, ten feature selection methods belonging to five groups, namely, the similarity-based, statistical-based, sparse learning-based, information theoretical-based and wrappers methods, were compared, based on their f-score and data size, for mapping a landscape infested by Parthenium weed (*Parthenium hysterophorus*). Overall, results showed that ReliefF (a similarity-based approach) was the best-performing feature selection method as demonstrated by the high f-score values of Parthenium weed and the small size of optimal features selected. Although svm-b (a wrapper method) yielded the highest accuracies, the size of the optimal subset of selected features was quite large. The results also showed that data size affects the performance of feature selection algorithms, except for statistically-based methods such as Gini-index and F-score and svm-b. The findings in this study provide guidance on the application of feature selection methods for the accurate mapping of invasive plant species, in general, and Parthenium weed, in particular, using new multispectral imagery with a high temporal resolution.

Keywords: feature selection; Parthenium weed; Sentinel-2

3.1 Introduction

The dimension space of the variables that are given as input to a classifier can be reduced without an important loss of information, while decreasing its processing time and improving the quality of its output (Gitelson et al., 2001). To date, studies on dimension reduction in remote sensing have mostly focused on hyperspectral datasets (Adam and Mutanga, 2009; Xie et al. 2017; Zheng et al., 2017) and high spatial resolution multispectral imagery using Object-Based Image Analysis (OBIA) (Yu et al., 2006; Waser et al., 2014; Ma et al., 2017). Generally, multispectral images have received less attention likely due to the limited number of bands that do not require dimension reduction. However, with the launch of high temporal resolution sensors, such as Sentinel-2, the amount of image data that can be acquired within a short period has considerably increased (Aires et al., 2016). This is due to the sensor's improved spectral resolution (13 bands) and a five-day temporal resolution (Immitzer et al., 2016).

Generally, high-dimensional remotely-sensed datasets contain irrelevant information and highly-redundant features. Such dimensionality deteriorates the quantitative (e.g. leaf area index and biomass) and qualitative (e.g. land cover) performance of statistical algorithms by overfitting the data (Gnana et al., 2016). High-dimensional data are often associated with the Hughes effects, or the curse of dimensionality, a phenomenon that occurs when the number of features in a dataset is greater than the number of samples (Kavzoglu and Mather, 2002; Taşkın et al., 2017). The Hughes effects affect the performance of algorithms that were previously designed for low-dimensional data. Whereas high dimensionality can lead to the poor generalization of learning algorithms during the classification process (Taşkın et al., 2017), it can also embed features that are crucial for classification enhancement. Hence, when using dimension reduction algorithms, a subset of those features can be selected from the high-dimensional data, increasing their predictive potential (Lagrange et al., 2017).

There are two main components of dimension reduction strategies namely, feature extraction or construction and feature selection or feature ranking. Feature extraction (e.g. Principle Component Analysis (PCA)), constructs a new and low dimensional feature space by using linear or non-linear combinations of the original high-dimensional feature space (Li et al., 2017), while feature selection (e.g. Fisher Score and Information Gain) extracts subsets from existing features (Gnana

et al., 2016). Although feature extraction methods produce higher classification accuracies, the interpretation of the generated results is often challenging (Zheng et al., 2017). However, feature selection methods do not change the original information of the features, thus giving models better interpretability and readability. Feature selection techniques have been applied in text mining and genetic analysis (Li et al., 2017). Hence, in this study, they were preferred over the feature extraction methods.

Traditional feature selection techniques are typically grouped into three approaches, namely the filter, embedded and wrapper methods (Cao et al., 2017). In earth observation related studies, feature selection algorithms have generally been compared, based on this grouping (Novack et al., 2011; Ma et al., 2017). However, in the advent of big data, this grouping can be regarded to be very broad, necessitating the development of new feature selection algorithms. For instance, within filter feature selection methods, there are some that evaluate the importance of features that are based on the ability to preserve data similarity (e.g. Fisher Score, ReliefF), while others use a heuristic filter criterion (e.g. Mutual Information Maximization). Therefore, it is crucial to re-evaluate the comparison of feature selection algorithms in a data-specific perspective.

Li et al. (2017) reclassified traditional feature algorithms for generic data into five groups, namely the similarity-based feature selection, information theoretical-based feature selection, statistical-based feature selection, sparse learning-based feature selection and wrappers. They devised an open-source feature selection repository, named scikit-feature, that provides 40 feature selection algorithms (including unsupervised feature selection approaches). Some selections, such as Joint Mutual Information and dt-f, are relatively new in earth observation applications (Chen et al., 2003).

Over the last past few decades, the number of vegetation indices has significantly increased. For instance, Henrich et al. (2009) gathered 250 vegetation indices that were derivable from Sentinel-2. The computation of vegetation indices from higher temporal resolution imagery, like Sentinel-2, would lead to data with increased multi-collinearity, a higher number of derived variables, and increased dimensions. Hence, the need for accurate and efficient feature selection techniques when dealing with new generation multispectral imagers, such as Sentinel-2, is becoming increasingly valuable in vegetation mapping (Aires et al., 2016). To the best of our knowledge, no earth

observation-related study has endeavored to undertake an empirical evaluation of feature selection methods, as provided in the scikit-feature repository for generic data.

In this study, feature selection algorithms, based on the classification of Li et al. (2017), were compared for mapping a landscape infested by Parthenium weed, using Sentinel-2. Parthenium weed is an alien invasive herb of tropical American origin that has infested over thirty countries. It has been identified as one of the seven most devastating and hazardous weeds worldwide (Adkins and Shabbir, 2014). A number of studies (Dhileepan, 2007; McConnachie et al., 2011) have reported its adverse impacts on ecosystem functioning, biodiversity, agricultural productivity and human health. A detailed comparison of feature selection algorithms on Sentinel-2 spectral bands combined with vegetation indices, with respect to the classification of Li et al. (2017), would therefore: (a) improve mapping accuracy and be valuable for designing mitigation approaches; (b) shed light on the most valuable feature selection group; and (c) identify the most suitable feature selection method for accurately mapping a Parthenium weed-infested landscape. Unlike previous studies that evaluated the features selection methods on the basis of overall classification accuracy (Ao et al., 2017; Kganyago et al., 2017; Ma et al., 2017), this study investigated their performance on the mapping accuracy of a specific landscape phenomenon (Parthenium weed) as a high overall classification accuracy does not always mean a reliable accuracy for a specific class (Ao et al., 2017). In this study, we sought to provide a detailed comparison of feature selection algorithms on higher temporal resolution satellite images with a high data volume. More specifically, we looked at (a) their performance on Parthenium weed, using specific class-related accuracies as an evaluation criterion; and (b) the impact of data size on their accuracy.

3.2 Materials and Methods

3.2.1 Reference data

In total, 447 reference points for mapping Parthenium weed and its surrounding land cover classes were obtained. To determine the optimal feature selection methods, and to test the effect of the data sizes on the classification accuracies, these ground reference data were randomly split into training and test sets in three different ratios: 1:3; 1:1; 3:1, as shown in Table 3.1 (Pal and Foody, 2010). The random split was undertaken using the function “train_test_split” of the sklearn Python library. “Random-state” and “stratify” parameters were included in the function to, respectively,

allow reproducibility and obtain the same proportions of class labels as the input dataset. The data design also allowed evaluating the investigated feature selection methods with respect to the Hughes effect.

Table 3.1 Training and test dataset combinations for land cover classes

Land-Cover Classes	Training Set 3 (70%)	Test Set 3 (30%)	Training Set 2 (50%)	Test Set 2 (50%)	Training Set 1 (30%)	Test Set 1 (70%)	Total
Forest	70	30	50	50	30	70	100
Water Body	49	21	35	35	21	49	70
Parthenium Weed	63	27	45	45	27	63	90
Grassland	64	28	46	46	28	64	92
Settlements	66	29	48	48	29	66	95

3.2.2 Feature selection methods

In this section, the five groups of feature selection methods are briefly discussed. Two representative methods were randomly chosen from each group to achieve the comparison.

3.2.2.1 Similarity-based feature selection methods

Similarity-based feature selection methods evaluate the importance of features by determining their ability to preserve data similarity, using some performance criteria. The two selected feature selection algorithms in this study were Trace ratio and ReliefF. Trace ratio (Bradley, 2014) maximizes the data similarity for samples of the same class or those that are close to each other, while minimizing the data similarity for the sample of different classes or those that are far away from each other. More important features also have a larger score. ReliefF (Farrell et al., 2019) assigns a weight to each feature of a dataset, and the feature values that are above a predefined threshold are then selected. The rationale behind ReliefF is to select features randomly and, based on the nearest neighbors, the quality of the features is estimated according to how well their values distinguish among the instances of the same and different classes close to each other. The larger the weight value of a feature, the higher the relevance (Colkesen and Kavzoglu, 2018).

3.2.2.2 Statistical-based feature selection methods

Statistical-based feature selection methods rely on statistical measures in order to estimate the relevance of features. Some examples of statistical-based feature selection methods include the Gini index and the F-score. The Gini index (Gini, 1912) is a statistical measure that quantitatively

evaluates the ability of a feature to separate instances from different classes (Li et al., 2017). It was used earlier in decision tree for splitting attributes. The rationale behind the Gini index is as follows: Suppose S is the set of s samples with m different classes ($C_i, i = 1, \ldots, \mathrm{m}$). According to the differences of classes, S can be divided into m subset ($S_i, i = 1, \ldots, \mathrm{m}$). Given that S_i is the sample set which belongs to class C_i, s_i is the sample number of set Si, the Gini index of set S can be computed according to the equation below (Shang et al., 2007):

$$Gini(S) = 1 - \sum_{i=1}^{m} P_i^2 \qquad \text{(Equation 3.1)}$$

where P_i denotes the probability for any sample to belong to C_i and to estimate with s_i/s.

The F-score (Wright, 1965) is calculated as follows: Given that f_i, n_j, μ, u_j, σ_j represent the number of instances from class j, the mean feature value, the mean feature value on class j, the standard deviation of feature value on class j, respectively, the F-score of a feature f_i can be determined as follows:

$$F_{-score}(f_i) = \frac{\sum_j \frac{n_j}{c-1}(u_j - u)^2}{\frac{1}{n-c}\sum_j (n_j - 1)\,\sigma_j^2} \qquad \text{(Equation 3.2)}$$

3.2.2.3 Sparse learning-based methods

The Sparse learning-based methods constitute a group of embedded approaches. They aim at reducing the fitting errors, along with some sparse regularization terms, which make feature coefficients small or equal to zero. To make a selection, the corresponding features are discarded (Li et al., 2017). Feature selection algorithms belonging to this group have been recognized to produce good performance and interpretability. In this study, sparse learning-based methods with the following sparse regularization terms, namely ℓ1-norm Regularizer (LS-121) (Hastie et al., 2015) and ℓ2,1-norm Regularizer (LL-121) (Liu al., 2009), were implemented.

3.2.2.4 Information theoretical-based methods

Information theoretical-based methods apply some heuristic filter criteria in order to estimate the relevance of features. Some feature selection algorithms that belong to this family include Joint

Mutual Information (JMI) and Mutual Information Maximization (MIM) or Information Gain. The JMI seeks to incorporate new unselected features that are complementary to the existing features and are given the class labels in the feature selection process (Li et al., 2017), while the MIM measures the importance of a feature by its correlation with the class labels. MIM assumes that features with strong correlations would achieve a good classification performance (Li et al., 2017).

3.2.2.5 Wrappers

Wrapper methods use a predefined learning algorithm, which acts like a black box, to assess the importance measures of selected features, based on their predictive performance. Two steps are involved in selecting features. First, a subset of features is searched and then the selected features are evaluated repeatedly until the highest learning performance is reached. The features that are regarded as being relevant are the ones that yield the highest learning performance (Li et al., 2017). These two steps are implemented by using the forward or backward selection strategies. In the forward selection strategy, the search for relevant features starts with an empty set of features and then features are progressively added into larger subsets, whereas in backward elimination, it starts with the full set of features and then progressively eliminates the least relevant ones (Kohavi and John, 1997).

However, the implementation of wrapper methods is limited, in practice, for high-dimensional data, due to the large size of the search space. Some examples of wrapper methods that were used in this study include Decision Tree Forward (dt-f) (Guyon and Elisseeff, 2003) and Support Vector Machines Backward (svm-b) (Guyon and Elisseeff, 2003).

3.2.3 Vegetation indices computation

In total, 75 Vegetation Indices (VI) derived from the Sentinel-2 wavebands were computed, using the online Indices-Database (IDB) developed by Henrich et al. (2009). The IDB provides over 261 parametric and non-parametric indices that can be used for over 99 sensors and it allows the viewing of all available VI for specific sensors and applications (Henrich et al., 2009). In this study, VI were selected because of their usefulness for vegetation mapping and in order to increase the dimensionality of Sentinel-2 data.

3.2.4 Classification algorithm: RF

The RF classifier was used to infer models from different selected features, using the investigated feature selection methods. RF is a combination of decision tree classifiers, where each classifier casts a single vote for the most frequent class to classify an input vector (Breiman, 1996). RF grows trees from random subsets drawn from the input dataset, using methods such as bagging or bootstrap aggregation. A split of the input dataset is typically performed using attribute selection measures (Information Gain, Gini-Index), which are useful in maximizing the dissimilarity between classes and they therefore determine the best split selection in creating subsets (Rodriguez-Galianon et al., 2012). In the process of RF model training, the user defines the number of features at each node, in order to generate a tree and the number of trees to be grown. The classification of a new dataset is done by passing down each case of the datasets to each of the grown trees, and then the forest chooses a class that has the most votes of the trees for that case (Pal, 2005). More details on RF can be found in Breiman (2001). RF was chosen for this study, as it can efficiently handle large and highly dimensional datasets (Díaz-Uriarte and de Andres, 2006; Archer and Kimes, 2008).

3.2.5 Model assessment

To assess the classification accuracy of models on test datasets, the estimated classes in different models developed in this study were cross-tabulated against the ground-sampled classes for the corresponding pixels in a confusion matrix. The performance of the developed models was assessed on test data sets, using performance measures, such as the UA and the PA and the f-score of Parthenium weed class. Supplementary information, including the PA, the UA of other classes and the Kappa coefficient was added. The UA refers to the probability that a pixel labeled as a certain class on the map represents that class on the ground. The PA represents pixels that belong to a ground-sampled class, which fail to be classified into the correct class. The f-score is the harmonic mean of UA and PA and it is typically used to assess accuracy per-class, as it represents a true outcome for specific classes (Ao et al., 2017). The Kappa coefficient represents the extent to which the classes on the ground are correct representations of the classes on the map. Their formulae are as follows:

$$PA = \frac{TP}{TP + FN} \qquad \text{(Equation 3.3)}$$

$$UA = \frac{TP}{TP + FP} \quad \text{(Equation 3.4)}$$

$$F = \frac{2 \times PA \times UA}{PA + UA} \quad \text{(Equation 3.5)}$$

$$Kappa\ coefficient = \frac{P_o - P_e}{1 - P_e} \quad \text{(Equation 3.6)}$$

where:

True Positive (TP) represents the number of correctly labeled positive samples;

False Positive (FP) represents to the number of negative samples that were incorrectly labeled as positive samples;

False Negative (*FN*) represents the number of positive samples that were incorrectly labeled as negative samples;

P_o represents the relative observed agreement among classes;

P_e represents hypothetical probability of chance agreement.

3.2.6 Software and feature selection

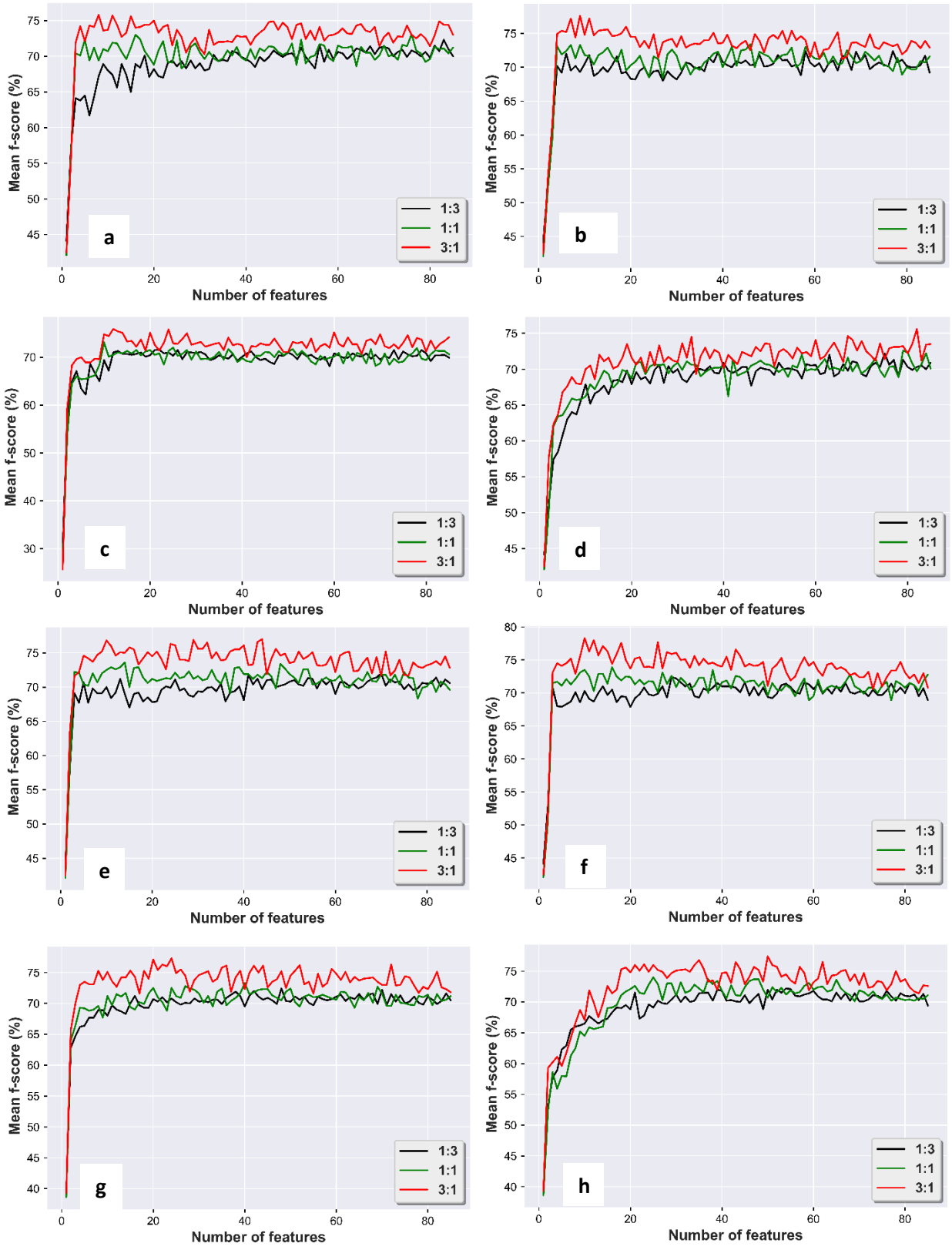
All investigated feature selection methods were applied on the three previously-mentioned training datasets using the scikit-feature library, a package of the Python (Version 3.6) programming language. This library was developed by Li et al. (2017) and provides more than 40 feature selection algorithms. To obtain the f-score accuracy of selected features for each feature selection method, we first created a range of numbers that started from 1 to 85 (which is the total number of variables) instead of specifying the number of selected features, as prescribed by Li et al. (2017). Each number in the range corresponded to the size of selected feature subsets. Then, for each dataset, the RF model was trained and evaluated on the test dataset, using the selected feature subset through a loop iteration. RF was run in Python (Version 3.6) using the sklearn library (Pedregosa et al., 2011). As RF was only used in this study for evaluating the performance of optimal variables selected by different feature selection methods, its hyperparameters were kept in default (e.g. the number of trees in the forest is equal to 10, and the criterion set to “gini”). Default hyperparameters often yield excellent results (Ahmad et al., 2017). In addition, according to Du et

al. (2015), the larger number of trees does not influence classification results. This procedure was repeated ten times by reshuffling the samples of training and test sets, and the mean of f-score was computed for each select feature subset. This was to ensure the reliability of the results of the investigated feature selection methods. The feature subset with the highest mean f-score was considered to be the most optimal. A code that automates the whole procedure, including deriving VI, was written in Python (Version 3.6).

3.3 Results

3.3.1 Comparison among of the investigated features algorithms

Figure 3.1 shows that the size of the feature subset and the training and test sets, using RF, determine the f-score accuracy of Parthenium weed. In general, the f-score increases with an increase in the size of the feature subset until it reaches a plateau at around 10 features, showing the insensitivity of RF in the face of the noisy or redundant variables. However, it is noticeable that the f-score accuracy of some feature selection methods, such as Gini-index and ReliefF, which belong to the statistical-based feature selection methods, and LL-121, were found at smaller feature subsets. With respect to the size of the training set, the f-score of Parthenium weed increased as the ratio between the training and test got larger. As a rule of thumb, this shows that when the ratio between the training and test datasets is large (for example, 3:1), a learning curve of a higher F-score would be produced. Nevertheless, some feature selection methods, such as svm-b, Gini-index and F-score, seemed to yield similar f-score accuracies, regardless of the size of the dataset.


Mean f-score (%)
Number of features
1:3
1:1
3:1
a
b
c
d
e
f
g
h

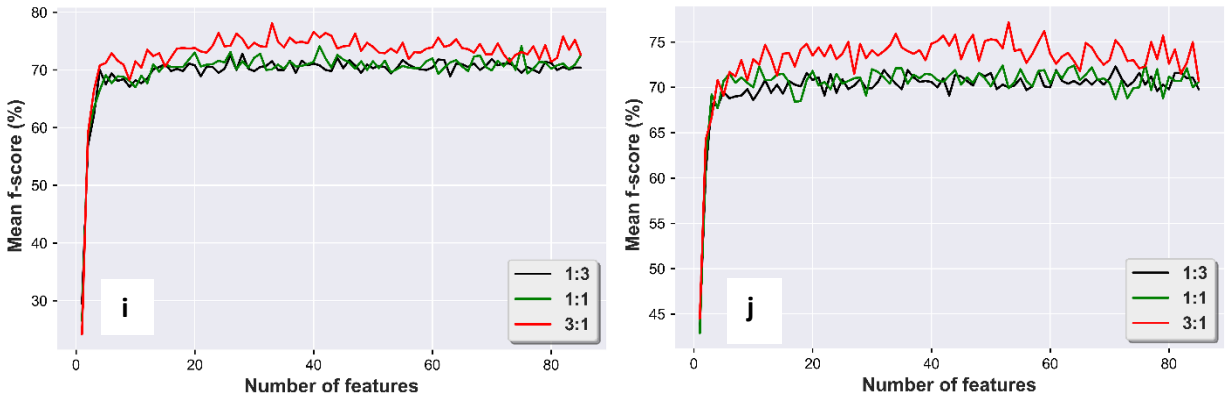


Figure 3.1 Mean f-score learning curve of trace ratio (a), ReliefF (b), Gini-index (c), F-score (d), LS_121(e), LL_121 (f), JMI (g), MIM (h), svm-b (i), dt-f (j) for different feature subsets (Features are made of 75 VIs and 10 Sentinel-2 bands)

3.3.2 Comparison of performance between peak accuracy and accuracy derived from full feature subsets

3.3.2.1 First training set

As per Table 3.2, all the investigated feature selection methods yielded a similar f-score of Parthenium weed for the optimal feature subset. F-score accuracies varied from 71% to 72%. Svm-b produced the highest f-score accuracy (72.5%). However, in terms of the size of optimal feature subset, ReliefF and Gini-index were the best in reducing the dimensionality of the full-dataset. The size of feature subsets was 6 and 13, respectively. The f-score method was the lowest at this point of view. For this dataset, the ReliefF method can be recommended, as its f-score is among the highest and the size of optimal feature subset among the smallest. The computational time and accuracies of other classes were also low and high, respectively, for the ReliefF method (Table 3.3).

Table 3.2 F-score, PA and UA of Parthenium weed using optimal feature subsets yielded by investigated feature selection methods for first training set

Feature Selection Method	Peak Accuracy			Number of Features	*Comput. Time
	PA (%)	UA (%)	F-score (%)		
Trace Ratio	74	69.8	71.6	56	0.58
ReliefF	74.5	70.1	72	6	1.05
Gini-Index	74.2	69.3	71.3	13	18.54
F-Score	74.9	69.5	72	63	0.25
LS_121	73.4	71.2	71.8	41	20.01
LL_121	75.2	70.2	72.4	29	0.37
JMI	74.8	69.6	71.9	38	38.58
MIM	74.4	70.1	72	39	39.45
SVM-B	74.7	71.1	72.5	26	0.5
DT-F	73.2	68.5	70.3	38	573.44
None	72.7	69.1	70.4	85	-

*Comput. Time (s): Computational time in seconds.

Table 3.3 Classification accuracies of other classes using optimal feature subsets yielded by investigated feature selection methods for first training set

	Forest		Water Body		Grassland		Settlements		
*F.S.M	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)	**Kappa Coef.
Trace Ratio	89.4	88.8	99	97	58.2	64.5	85.8	81.3	0.78
ReliefF	91.1	90	100	97.6	58.5	67.6	85.6	79	0.75
Gini-Index	90.8	90.4	99.2	95.4	57.9	64.5	85.4	82.4	0.74
F-Score	90.3	88.5	100	96.8	58.6	66.6	85.9	82.5	0.76
LS_121	91.8	88.8	100	96	55.8	62.8	85	81.7	0.75
LL_121	90.7	90.8	100	97.8	58.3	64.2	83.8	80.3	0.76
JMI	91.8	90.7	99.4	96.8	59.5	67.2	85.6	81.9	0.77
MIM	93.3	90.8	99.2	95	59.7	66.6	84.6	83.1	0.78
SVM-b	90.4	90.3	97.4	100	60	64.4	84.4	82.3	0.75
DT-F	91.8	89.1	100	96.8	58.7	64.8	83.3	80.8	0.73
None	91.4	89.7	100	97.4	58.8	66.2	87.1	81.8	0.78

*F.S.M.: Feature selection method; **Kappa Coef.: Kappa coefficient.

3.3.2.2 Second training set

As shown in Table 3.4, apart from LS_121 and the F-score, all feature selection methods could reduce the number of features with a higher f-score accuracy than the full dataset. As for the first dataset, svm-b was the best-performing feature selection method. Its PA, UA and f-score of Parthenium weed and Kappa coefficient (Table 3.5) were the highest. ReliefF selected the least number (4) of optimal features and was among the highest top-performing feature selection methods after svm_b, in terms of f-score and PA of Parthenium weed. It was followed by Gini-index, LL_121, with respect to f-score and size of feature subsets. Once more, the F-score method turned out to perform poorly because there was of a large number of selected features and no improvement of the f-score accuracy of Parthenium weed.

Table 3.4 F-score, PA and UA of Parthenium weed using optimal feature subsets yielded by investigated feature selection methods for the second training set

Feature Selection Method	Peak Accuracy			Number of Features	*Comput. Time
	PA (%)	UA (%)	F-score (%)		
Trace Ratio	77	70.3	73	16	0.6
ReliefF	77.1	69.6	73.1	4	2.14
Gini-Index	75.6	71.1	73.1	10	31.29
F-score	75.7	68.6	72	57	0.26
LS_121	76	68.7	71.9	41	20.12
LL_121	76.6	70	73.1	9	0.43
JMI	75.5	71.5	73.3	40	59.46
MIM	75.3	72.7	74	25	61.07
SVM-b	77.1	72	74.1	41	0.54
DT-F	78.3	69.6	73.3	39	934.08
None	75.9	69.7	72.6	85	-

Table 3.5 Classification accuracies of other classes using optimal feature subsets yielded by investigated feature selection methods for the second training set

	Forest		Water Body		Grassland		Settlements		
F.S.M	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)	Kappa Coef.
Trace Ratio	91.6	90.2	99.2	95.2	61.3	72.8	86.7	81	0.76
ReliefF	89	91.2	98.5	97.6	61.6	68.9	85.6	81.4	0.78
Gini-Index	91.5	91.6	98.9	95.5	60.3	66.8	86.8	83.1	0.79
F-Score	90.4	92.2	100	96.4	63.5	71.8	87.7	84.1	0.79
LS_121	90.7	91.4	98.8	95	59.9	69	86.5	81.7	0.78
LL_121	90.2	91.4	98.8	96.7	60.6	68.8	85.1	80.7	0.8
JMI	91.9	91	98.8	96.1	62.3	70.3	88	82.1	0.77
MIM	90	91.8	98.9	95.5	64.8	69.1	86.6	82.9	0.82
SVM-b	91.4	92.8	100	97	64	70.7	86.9	82.7	0.83
DT-F	91	86	100	96	60	68	83.3	80.8	78.1
None	87.2	82.5	99.7	97.3	62.2	68.6	87.2	82.5	0.79

3.3.2.3 Third training set

Table 3.6 shows that LL_121 and ReliefF were among the feature selection methods that selected a small subset of features with a high f-score of Parthenium weed, and with PA and UA accuracies above 3% difference from the full dataset. ReliefF, for example, yielded 77.2% of the f-score, 80% of the PA and 75% of the UA, with only seven optimal features. The full dataset yielded 72.6% of the f-score, 75.2% of the PA and 71.4% of the UA without any feature selection method being applied. Svm_b outperformed all the feature selection methods, with the highest PA (82.3%) and f-score of Parthenium weed (78.1%), and kappa coefficient (0.83) (Table 3.7). However, the number of optimal features that were selected was quite large (33). As for the previous datasets, the performance of the F-score method was the worst, with the lowest PA (78.5%), UA (73.6%) and f-score (75.6%) of Parthenium weed and the largest feature subset (82).

Figure 3.2 illustrates the spatial distribution of Parthenium weed and surrounding land cover, using the full dataset and optimal features from ReliefF.

Table 3.6 F-score, PA and UA of Parthenium weed using optimal feature subsets yielded by investigated feature selection methods for third training set

Feature Selection method	Peak Accuracy			Number of Features	*Comput. Time
	PA (%)	UA (%)	f-score (%)		
Trace Ratio	77.2	74.5	75.7	11	0.7
Relief	80	75	77.2	7	3.56
Gini-Index	78.5	74.2	75.9	12	43.97
F-Score	78.5	73.6	75.6	82	0.3
LS_121	78.6	76.1	76.8	10	20.64
LL_12 1	79.2	78.2	78.3	10	2.2
MIM	76.9	72.6	74.1	50	87.71
SVM-b	82.3	75	78.1	33	0.54
DT-F	76.6	72.4	74.1	26	1346.43
None	75.2	71.4	72.6	85	-

Table 3.7 Classification accuracies of other classes using optimal feature subsets yielded by investigated feature selection methods for third training set

	Forest		Water Body		Grassland		Settlements		
F.S.M	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)	Kappa Coef.
Trace Ratio	93.5	81	98.5	81	61.8	81	88.3	81	0.75
ReliefF	92	81.4	100	81.4	66.2	81.4	89.9	81.4	0.77
Gini-Index	92.1	83.1	99.5	83.1	64	83.1	86.3	83.1	0.78
F-Score	92.2	95.2	99.5	98	64.4	71.2	91.4	84.4	0.79
LS_121	92.5	81.7	99.5	81.7	62.9	81.7	89	81.7	0.82
LL_121	92.3	68.8	98.5	68.8	63.4	68.8	90	68.8	0.77
JMI	90.6	82.1	99.5	82.1	68.4	82.1	90.9	82.1	0.8
MIM	76.9	72.6	99	82.9	63.7	82.9	84.8	82.9	0.75
SVM-b	91.6	82.7	99.5	82.7	67.2	82.7	89.7	82.7	0.83
DT-F	93.1	73	100	73	63.7	73	87.9	73	0.82
None	90.8	82.5	100	82.5	64.3	82.5	89.1	82.5	0.74

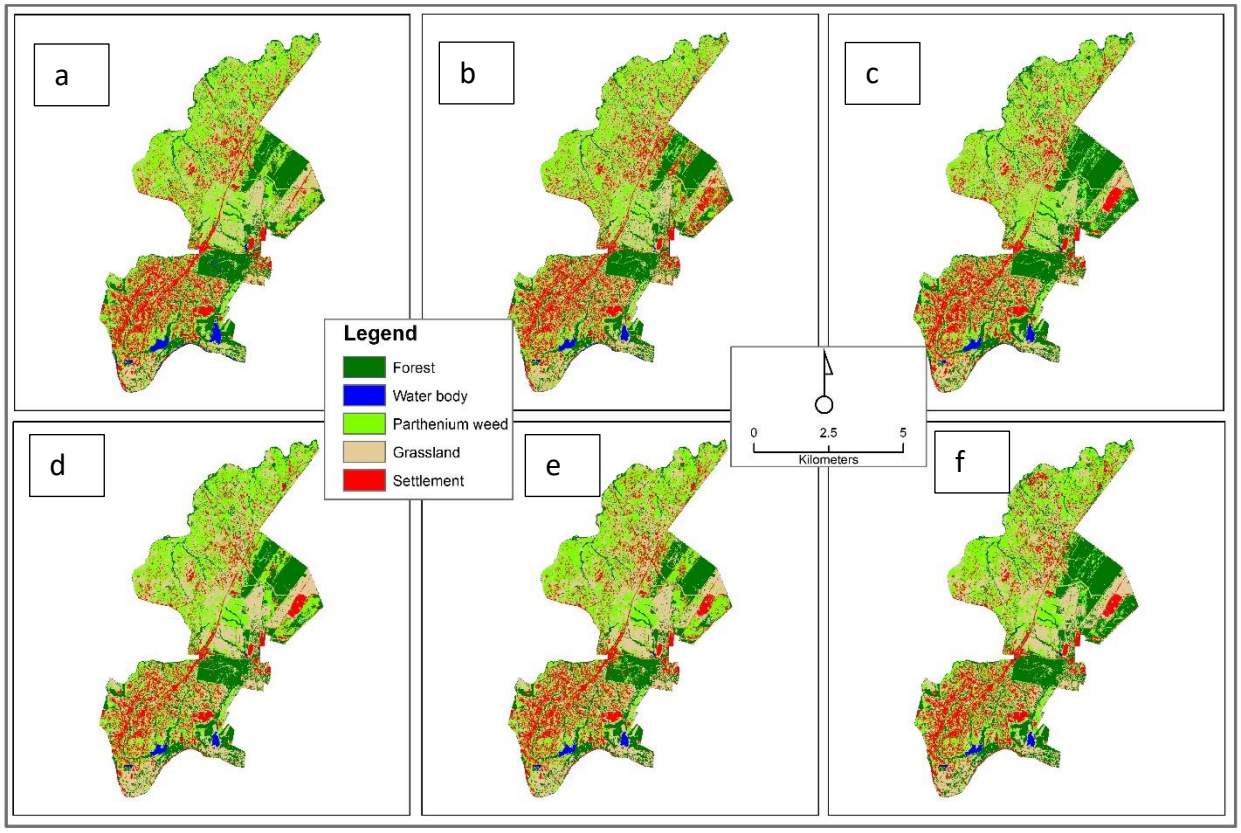


Figure 3.2 Spatial distribution of Parthenium weed and surrounding land cover with optimal features from ReliefF on first (a), second (b) and third (c) training set and from full dataset on the first (d), second (e) and third (f) training set

3.4 Discussion

This study sought to compare ten feature selection algorithms with two of those belonging to the following feature selection method groups, namely, similarity-based, statistical-based, sparse learning-based, information theoretical-based and wrapper feature selection methods. These feature selection algorithms were applied on Sentinel-2 spectral bands combined with 75 vegetation indices for mapping a landscape infested by Parthenium weed. The comparison was based on the f-score of Parthenium weed, using RF. We also tested the effect of training and test set sizes on the performance of the investigated feature selection algorithm.

3.4.1 Comparison of the feature selection methods

The results showed that the feature selection algorithms could reduce the dimensionality of Sentinel-2 spectral bands combined with vegetation indices. The algorithms could increase the classification accuracies of Parthenium weed by up to 4%, by using the RF classifier, depending on the adopted feature selection method and the size of the dataset. Previous studies found similar results (Martin et al., 2018; Vuolo et al., 2018). For example, Vuolo et al. (2018), who applied filter-based feature selection algorithms and three machine learning techniques on WorldView-2 image for determining the most effective object features in object-based image analysis, achieved a significant improvement (about 4%) by applying the feature selection methods. Overall, ReliefF was the best-performing feature selection method because it could bring down the number of features from 85 to 6, 4 and 7 on the first, second and third datasets, respectively. Its f-score, PA and UA accuracies for Parthenium weed were also among the highest (Table 3.2, for example). According to Vergara and Estévez (2014), the purpose of feature selection is to determine the smallest feature subset that can produce the minimum classification error. This finding concurs with studies that attempted to compare ReliefF with other feature selection methods. For example, Robnik-Šikonja and Kononenko (2003) found that the subset of features selected by ReliefF tend to be small, compared to other feature selection methods, as only statistically-relevant features are retained during the selection process. Studies that compared ReliefF with other feature selection methods reported that the mapping accuracy with selected features from ReliefF was similar to the best feature selection methods (Martin et al., 2018; Vuolo et al., 2018). However, our findings are opposed to Vuolo et al. (2018), who found that ReliefF selected the highest number of features in comparison to Chi-square and Information Gain algorithms, but slightly lower classification accuracies than Information Gain, using RF, SVM and ANN. We suggest that repeated classifications on reshuffled training and test data should be investigated to confirm their findings. In terms of the f-score, PA and UA of Parthenium weed, svm-b, which belongs to the wrapper group, outperformed all the feature selection methods for the three datasets (Tables 3.2, 3.4 and 3.6). Although computationally expensive, several authors have noted that wrappers outperform the filter methods (Hall and Smith, 1999; Talavera, 2005; Chrysostomou, 2009). However, in this study, svm-b was found to be less computationally intense than some of the filter methods (Tables 3.2, 3.4 and 3.6). The F-score method did not perform well for mapping Parthenium weed because of the low accuracies and large subset of the selected features. To the best of our knowledge, its

use is very limited in earth observation related studies. Further investigations are therefore necessary. Concerning the investigated feature selection groups, not a single group performed well on all the datasets. This supports the recommendation that there is no universal 'best' method for all the learning tasks (Ghioca-Robrecht et al., 2008).

3.4.2 Impact of training sizes on feature selection performance

The results showed that the performance of feature selection algorithm depends on the ratio between the training and test dataset. It can be noticed that smaller difference of f-score accuracies of Parthenium weed between optimal features and full dataset were obtained when the ratio between training and test was 1:3 or 1:1 (Tables 3.2 and 3.3). According to Jain and Zongker (1997), a small sample size and a large number of features impair the performance of feature selection methods, due to the curse of dimensionality. All the investigated feature selection algorithms positively influenced the f-score accuracies of the dataset when the ratio between training and test was large (1:3 or 70% training and 30% test) (Table 3.6). This concurs with (Chu et al., 2012), who demonstrated that the higher the training size, the better the classification accuracy. They also highlighted the necessity of selecting the appropriate feature selection for improving classification accuracies. In this study, we found that some feature selection algorithms, such as Gini-index and F-score, which belong to the Statistical-based Feature selection methods, and svm-b did not seem to be affected by the curse of dimensionality (Figure 3.1). We did not come across studies showing this finding; hence, further investigations should be carried out to corroborate our results.

3.4.3 Implications of findings in Parthenium weed management

In invaded landscapes, Parthenium weed expands more rapidly than the native plants do (Terblanche et al., 2016). Spectral bands alone are insufficient to achieve reliable mapping accuracies (Casady et al., 2005). Increasing data dimensionality by combining, among others, the Sentinel-2 image bands, vegetation indices and other variables and applying an appropriate feature selection approach, a higher Parthenium mapping accuracy can be achieved. This study provides guidance on how newly-developed feature selection methods, based on the classification of Li et al. (2017), should be used to reduce the dimension of high temporal resolution imagery, such as

Sentinel-2, in mapping Parthenium weed. The accurate spatial distribution of Parthenium weed would enhance the decision-making process relating to appropriate mitigation measures.

3.5 Conclusions

The following conclusions can be drawn from the findings:

a) wrapper methods, such as svm-b, yield higher accuracies for classifying Parthenium weed when using the RF classifier;

b) ReliefF was the best-performing feature selection method in terms of the f-score and the size of the optimal features selected;

c) to achieve a better performance of the feature selection methods, the ratio of 3:1 between the training and test set size turned out to be better than the ratios of 1.1 and 1:3;

d) Gini-index, F-score and svm-b were slightly affected by the curse of dimensionality; and

e) none of feature selection method groups seemed to perform at their best for all the datasets.

The findings of this study are critical for reducing the computational complexity of processing a large volume of Sentinel-2 image data. With the advent of Sentinel-2A and B, an increased volume of data is available, necessitating feature selection. This offers the possibility of deriving useful information, and hence, the accurate classification of Parthenium weed maps. Further research should look at comparing other feature selection methods with different classifiers. Moreover, a combination of feature selection methods, such as ReliefF and svm-b, should be considered, as they select a small number of features and yield a high f-score accuracy, respectively.

3.6 Acknowledgments

We would like to thank the two anonymous reviewers for providing constructive comments, which have greatly improved the manuscript.

CHAPTER 4. OPTIMAL WINDOW PERIOD FOR MAPPING PARTHENIUM WEED IN SOUTH AFRICA, USING HIGH TEMPORAL RESOLUTION IMAGERY AND THE EXTRATREES CLASSIFIER

This chapter is based on: **Zolo Kiala**, Onisimo Mutanga, John Odindi and Cecilia Masemola. "Optimal window period for mapping Parthenium weed in South Africa using high temporal resolution imagery and the EXT classifier", Biological Invasions, Under Review.

Abstract

Parthenium weed (*Parthenium hysterophorus*) is one of the most noxious herbaceous weeds in the world, having an adverse impact on human and animal health, crop production, the environment, local, as well as national, economies, among other things. To optimize Parthenium mitigation, it is necessary to accurately monitor its spread by using earth observation data. However, one of the challenges of mapping Parthenium weed is that its spectral response is similar to that of surrounding herbaceous plant species, resulting in low classification accuracies. Due to a variability in the phenological characteristics of Parthenium weed and associated species, an exploration of that variability within the growing season may optimize the discrimination and subsequent mapping of Parthenium weed. However, determining the window(s) of where that variability is the most prominent has been overlooked in previous studies. Furthermore, no specific algorithm has been identified as being efficient enough to find such (a) window(s). EXT, an under-used classifier in earth observation studies, possesses interesting properties for satellite image processing, such as high speed and performance. In this regard, this study attempted to: (a) determine the optimal window period for discriminating Parthenium weed from co-existing plant species; and (b) to compare EXT and the RF algorithms. The results showed that the beginning of February was the optimal period for mapping Parthenium weed, with an OA of 88.1%. EXT outperformed RF for most of the dates. This study lays the foundation for optimizing earth observation data-derived models for characterizing invasive species and leveraging on the high temporal resolution of the new generation sensors.

Keywords: Parthenium weed, EXT, Sentinel-2

4.1 Introduction

IPSs increasingly disseminate across the world with the immediate consequence of irreversibly damaging ecosystems. They are known to impair ecosystems functioning, through the accentuation of fire occurrence and severity and the alteration of the dynamics of nutrients, carbon storage, the microclimate and vegetation succession (Huang and Asner, 2009). Parthenium weed (*Parthenium hysterophorus*) is one of the most harmful herbaceous weeds in the world as it is known to negatively affect animal, crop and human health, the environment, as well as local and national economies (Swati et al., 2013). By infesting cultivated lands, for example, Parthenium weed can impact livelihoods by reducing incomes and increasing the cost of weeding. In crop production, Parthenium weed can reduce yields by about 40% (Adkins and Shabbir, 2014). Direct contact with Parthenium weed by humans may cause hay fever, dermatitis, asthma and bronchitis (Tottrup, 2004).

The competitive advantage of Parthenium weed over co-existing species in invaded ecosystems can be ascribed to its ecological attributes, such as its fast growth, high reproductive potential, longevity of buried seeds and allelopathic suppression (Strathie et al., 2011; Swati et al., 2013). According to the SAPIA, in which the occurrence of IPSs is captured within quarter degree grid cells (about 25 x 25 km), Parthenium weed has increased from three cells in 1980 to 76 cells in 2014 in South Africa (Terblanche et al., 2016). Due to the enormous damages caused by the Parthenium weed and its ability to spread quickly, it is crucial to regularly and accurately monitor it. This will shed some light on its spatial dynamics, hence enhancing control efforts (Qing et al., 2018).

Monitoring invasive species, in particular Parthenium weed, through field surveys poses technical difficulties, particularly in remote and large areas. Moreover, ground monitoring is time-consuming and tedious (Erinjery et al., 2018). Meanwhile, earth observation technologies could offer a relatively accurate, faster and plausible technique to comprehensively monitor changes in the spatial distribution of highly invasive plants, such as Parthenium weed (Erinjery et al., 2018). For instance, Royimani et al (2018) used multi-year SPOT satellite remotely sensed data to characterise the areal extent covered by Parthenium weed in the Mtubatuba local municipality in the north-east coast of the KZN province of South Africa, with optimal kappa statistic accuracies

ranging between 65% and 73%. In a related study, Arogoundade et al (2019) modelled the spatial distribution of Parthenium weed in the KZN province using remotely sensed data combined with environmental variables to a superior Area Under the Curve (AUC) value of 0.976. However, these studies were undertaken using a single-date image within the growing season. A growing body of literature that illustrates that mapping herbaceous weeds is a major challenge, especially during the post-emerging stage (Matongera et al., 2017; Wang et al., 2018). This is because the spectral signature of herbaceous weeds is similar to that of the surrounding herbaceous plant species, resulting in low overall classification accuracies. It is hypothesised that, due to the variability in the phenological characteristics of Parthenium weed and associated species, an exploration of that variability within the growing season may optimise the discrimination and subsequent mapping of the invasive species, using remotely sensed data. Therefore, there is a need for comprehensive and robust techniques that could be used to identify a specific temporal window to accurately detect Parthenium within the growing season, using remotely sensed data. Raczko and Zagajewski (2017) suggested that the rigorous selection of remotely sensed image data scenes from suitable seasons that are characterized by optimal atmospheric conditions across the growing season, together with new and robust algorithms, are valuable for improving mapping accuracies. In this regard, leveraging on the high temporal resolution of new generation sensors, in conjunction with newly developed algorithms, offers opportunities to accurately determine those urgently-required optimal window periods. Subsequently, relevant spectral bands or features can be derived for use in developing models during those periods.

The recent deployment of Sentinel-2 multispectral imager (MSI) by the European Space Agency (ESA) has set up a new paradigm for the mapping of herbaceous invasive plants that were indiscriminable from other land cover types, based on the relatively limited spectral resolution of its predecessors (i.e. Landsat). Specifically, Sentinel-2 is an innovative optical multispectral scanner with a relatively wide-swath (290 km), as well as fairly high spatial (up to 10 m) and spectral (13 spectral bands) resolutions. In addition to its free availability, Sentinel-2 data have a global revisit time of five days, a characteristic that is very useful for monitoring the phenological growth stages of plants and that optimizes the classification accuracies. Among others, Sentinel-2 also contains a novel spectral band setup that is characterized by three bands in the red-edge region

and two bands in the SWIR, which are known to possess a high discriminatory potential for vegetation mapping.

Furthermore, it is hereby hypothesized that applying a relatively new algorithm, such as EXT classifier or Extremely Randomised Trees, on Sentinel 2 data could enhance the process of discerning the optimal temporal window(s) and improve the accuracy of characterizing the spatial advance of the invasive Parthenium weed. The EXT classifier is a modified version of RF that has been developed primarily for computer vision and bio-medical imaging applications (Moosmann et al., 2008; Marée et al., 2013; Myrans et al., 2016). The classifier has interesting properties for satellite image processing, such as its high speed and performance (Kiala et al., 2019). EXT has the potential to outperform widely-used and robust algorithms, such as RF, because of the higher grade of the randomness of trees during the training process. This decreases the variance of the generated trees and discriminates classes better discriminate (Kiala et al., 2019). Several studies have proved that EXT outperforms some of the robust machine learning techniques, such as SGB, SVM and RF (Barrett et al., 2014; Li et al., 2017; Lu et al., 2017). However, its application has been limited in earth observation-related studies. In this regard, it is not yet known if EXT could be more robust and efficient than commonly-used algorithms in detecting Parthenium weed, using remotely sensed data.

Therefore, this study aimed at determining the optimal window period for mapping Parthenium weed within a growing season, based on the most influential band images of Sentinel-2. It also sought to assess the robustness and computational efficiency of the relatively new EXT classifier in relation to RF, a renowned algorithm in the remote sensing community.

4.2 Materials and Methods

4.2.1 Reference data

A field work took place between January 12 and February 2, 2017. According to Kushwaha and Maurya (2012), the phenological cycle of Parthenium weeds can be completed within five months. Typically, germination takes place between September and December, and senescence occurs between March and May (Henry, 2008). Subsequently, our field campaign was conducted between

January and February, following the phenological transitions of Parthenium. It was also conducted only once because the study area is not subject to Parthenium weed control throughout the year.

In total, 447 reference points for mapping Parthenium weed and major land cover classes were obtained (Table 4.1). For model development and assessment, these field-based data were randomly split into 70% and 30% to make up the training and test datasets (Table 4.1).

Table 4.1 Training and test dataset for Parthenium weed and surrounding land cover classes

Land-cover classes	Training dataset (70%)	Test dataset (30%)	Total
Forest	70	30	100
Water body	49	21	70
Parthenium weed	63	27	90
Grassland	64	28	92
Built-up	66	29	95

4.2.2 Acquisition of multi-date Sentinel-2 images and pre-processing

Five Level 1C Sentinel-2A satellite images that were acquired on 19th January, 8th February, 28th February, 7th March and 27th March were selected and downloaded from the ESA website. The band images acquired on 7th March were the only ones affected by cloud cover, hence a clouds mask was created and applied. The remaining images were acquired under cloudless conditions. These acquisition dates spanned across dominant phenological events (i.e. rosette growth, flowering, senescence). Images that were acquisitions in December and the beginning of January, which coincided with early growth of Parthenium weed (Henry, 2008), were not considered in this study due to high percentage of cloud cover over the study area. As the phenological growth stages of Parthenium weed patches vary tremendously both in space and time (Giuliani et al., 2019), Enhanced Vegetation Index (EVI), a proxy reflecting phenology, was used to provide an insight into the phenological growth stages of Parthenium weed at the different image acquisitions. The EVI was calculated according to Huete et al. (1999), using the reflectance values extracted from Sentinel images at the GPS points of Parthenium weed.

Sentinel-2 imagery consists of four spectral bands at a 10 m spatial resolution in Blue (Band 2: 497 nm), Green (Band 3: 560 nm), Red (Band 4: 664 nm), and the Near infrared (Band 8: 835 nm) spectra. It also includes six bands at a 20 m spatial resolution in Red-edge (Band 5: 704 nm, Band 6: 740 nm, Band 7: 782 nm), NIR (Band 8a: 865 nm), Short-wave infrared (SWIR) (Band 11: 1614 nm, Band 12: 2202 nm) spectra and atmospheric bands at a 60 m spatial resolution in coastal aerosol (Band 1: 442.7 nm), Water vapour (Band 9: 945.1 nm) and SWIR – cirrus (Band 10: 1373.5).

4.2.3 Data analysis

4.2.3.1 Classification based on RF and EXT classifiers

RF is created from the combination of several decision tree classifiers, where each classifier casts a single vote for the most frequent class to predict an input vector (Breiman, 1996). RF grows trees from random subsets drawn from the input dataset, using bagging or bootstrap aggregation methods. Typically, the split of input dataset is performed using attribute selection measures, such as Gini-Index, Information Gain. The usefulness of the attribute selection measures resides in the maximization of the dissimilarities between classes and therefore, they determine the best split selection in creating subsets (Rodriguez-Galiano et al., 2012). During the process of RF model calibration, the number of features at each node are defined by the user, in order to generate a tree and the number of trees to be grown. By passing down each case of the datasets to each of the grown trees, a new dataset is classified and then the forest chooses a class that obtains the majority of votes of the trees for that case (Pal, 2005). Breiman (2001) provides more details on RF.

The EXT classifier, or Extremely randomised tree, was first introduced by Geurts et al. (2006). Like RF, it constructs independent decision trees to perform classification and regression problems. In addition, EXT includes stronger randomisation techniques, in order to further reduce the variance of the prediction model. The two main differences between RF and EXT are: a) EXT uses the whole input training set for constructing each tree, while RF applies a bagging procedure; and b) the node split is randomly chosen by EXT (both variable splitting values and variable index are randomly selected), whereas the best split in RF is optimised by a feature index and a feature splitting value, among a random subset of features (Li et al., 2017).

4.2.3.2 Feature importance ranking, model calibration and validation

EXT and RF algorithms were also used to generate a variable importance ranking of spectral bands, using the Mean Decrease Accuracy (MDA) (Nembrini et al., 2018) at the optimal temporal window (s) for mapping Parthenium weed. To ensure that samples are classified more than once, the number of trees was set at 500 (Oon et al., 2019). Other hyperparameters of EXT and RF (e.g. the oob_score and the maximum depth of the tree) were kept in default (Pedregosa et al., 2011).

During the model assessment, an error (or confusion) matrix was derived and used. From the confusion matrix, performance measures (e.g. the OA, f-score, UA, and PA were computed (Lunetta and Lyon, 2004). The OA refers to the proportion of all the classes that were mapped correctly. The UA refers to the probability that a pixel labelled as a certain class on the map represents that class on the ground. The PA refers to the probability that a certain class on the ground is classified as such. The f-score is the harmonic mean of the UA and the PA. More focus was given to f-score, UA and PA of Parthenium weed class, as this study purposed to map its spatial distribution. Furthermore, the difference in the accuracies, among of all the created models was statistically compared, using the Wilcoxon test in Microsoft Excel (Hogg and Craig, 1995). The rest of the analyses and map creation were implemented in Python (Version 3.6).

4.3 Results

4.3.1 Spectral profile of investigated classes

Figure 4.1 illustrates the spectral signature of investigated land cover classes, using the single-date Sentinel-2 image. Except for settlements, a fair amount of overlap between the classes was evident for spectral bands in the visible region of the electromagnetic spectrum (Bands 2, 3 and 4). From Band 5 onward, there was more separability among the classes. Grassland and Parthenium weed were the classes with the most similar spectral signatures. However, the two classes were the most spectrally separable between Bands 6 and 11.

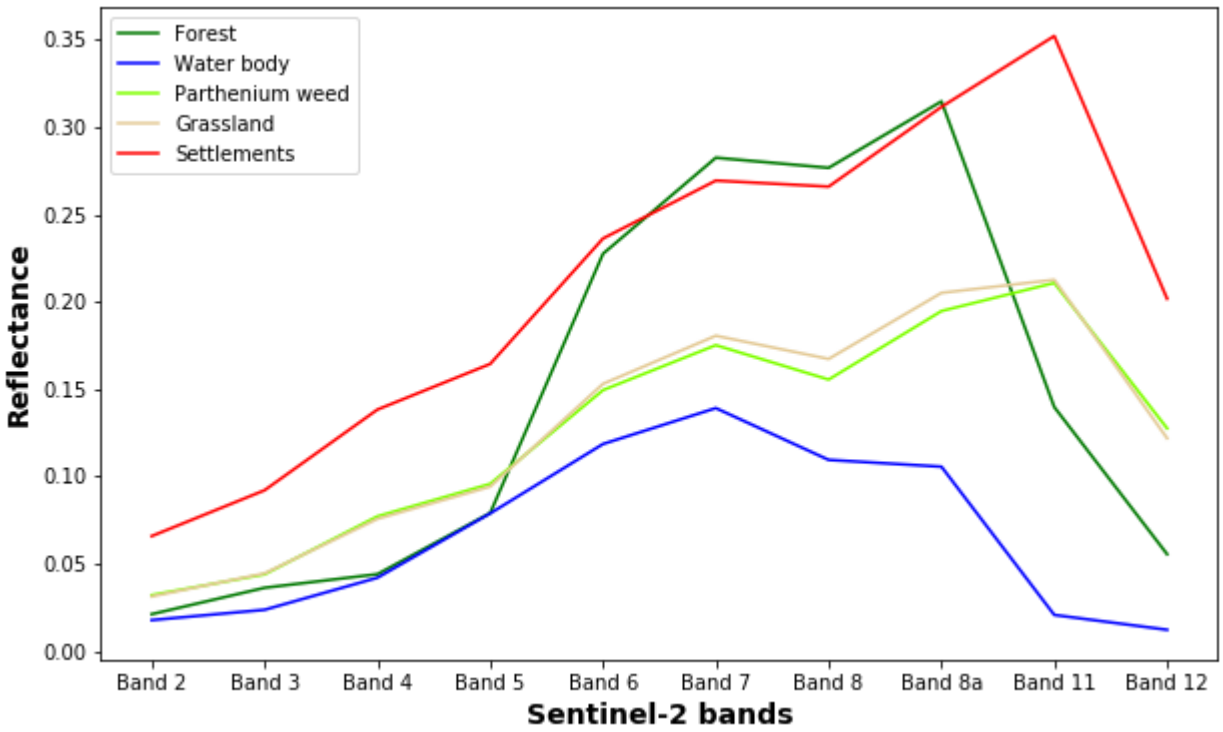


Figure 4.1 Spectral signature of investigated land cover types

4.3.2 Finding optimal temporal window for mapping Parthenium weed

4.3.2.1 EVI profile of Parthenium weed over time

Figure 4.2 illustrates the mean, maximum and minimum of the EVI values of Parthenium weed over time. Overall, the lowest and highest values of EVI were found on 8th February and on 28th February, respectively. A decrease in EVI was observed from 19th January to 8th February and from 28th February to 27th March. From 8th February to 28th February, there was a sharp increase in EVI.

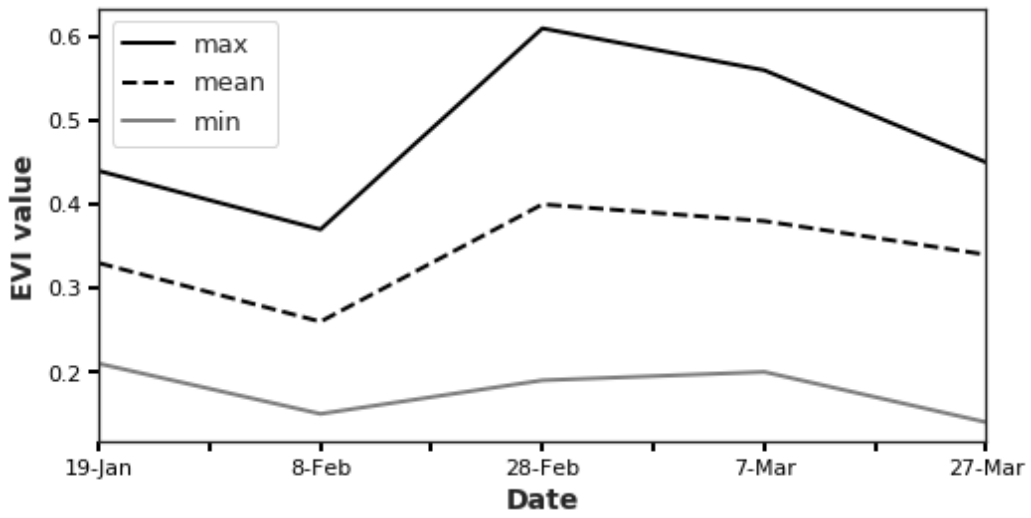


Figure 4.1 Maximum, mean and minimum EVI values of Parthenium weed over time

4.3.2.2 Classification accuracies of the RF and EXT models over time

Figure 4.3 shows the variation of the overall classification accuracy and the accuracy of Parthenium weed over time, using RF and EXT. The highest f-score and overall accuracies were observed on 8th February, while the lowest accuracies were observed on 28th February and on 7th March, using the RF and EXT models, respectively. With respect to the f-score of Parthenium weed, there was a marked decrease in accuracy from 8th February to 7th March for both classifiers. The EXT models exhibited a noticeable increase in the classification accuracies from 7th March to 27th March.

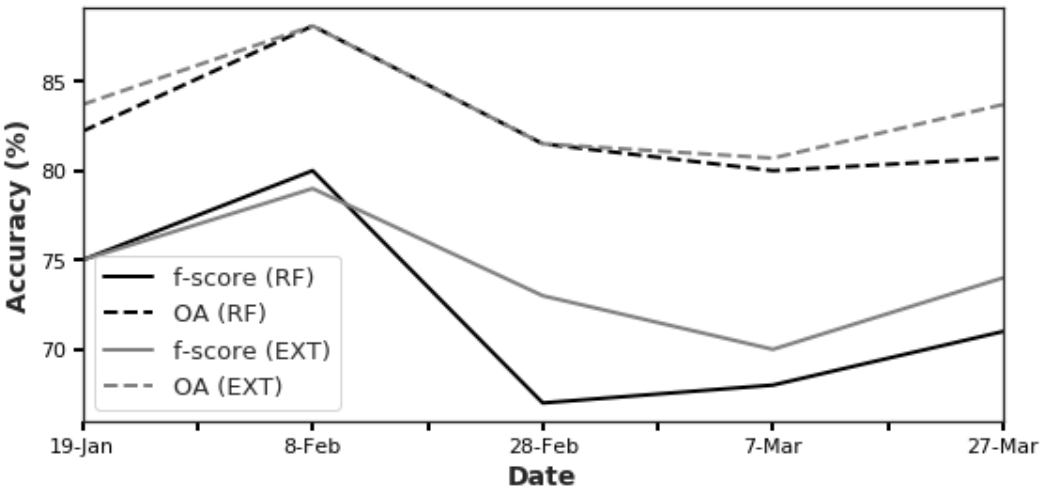


Figure 4.2 Variation of overall classification accuracy and f-score of Parthenium weed over time, using RF and EXT

4.3.3 Comparison between EXT and RF

4.3.3.1 Accuracies of EXT and RF

Tables 4.2 and 4.3 display the accuracies derived in discriminating Parthenium weed from other land covers, using Sentinel-2 images acquired on different dates with EXT and RF algorithms. An agreement on the identification of the optimal window period for discriminating Parthenium weed between the RF and EXT algorithms was observed to be exhibited by the image acquired on 8th February. Specifically, the image acquired on this date exhibited the highest overall accuracy as well as the f-score in this study. When comparing the two algorithms in terms of the OA, and the PA and the UA in discriminating Parthenium weed, the EXT models outperformed the RF models for most of the dates (Tables 4.2 and 4.3). For example, the PA accuracies exhibited by the EXT were 84%, 86%, 73% and 94%, whereas those exhibited by RF were 81%, 74%, 68%, and 85% on 8th February, 28th February, 7th March and 27th March, respectively, for Parthenium weed class. The lowest accuracies were observed on 7th March.

Table 4.2 Sentinel-2 accuracies (%) for investigated land cover classes on different dates using RF

	Parthenium weed		Forest		Water body		Grassland		Settlements		
Image date	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	OA
19-Jan	83	68	83	97	100	95	68	63	81	90	82.2
08-Feb	81	79	94	100	100	100	73	70	93	93	88.1
28-Feb	74	61	88	97	100	100	69	67	78	86	81.5
07-Mar	68	68	88	93	100	100	62	59	86	83	80
27-Mar	85	61	88	97	100	100	62	74	76	76	80.7

Table 4.3 Sentinel-2 accuracies (%) for investigated land cover classes at different dates using EXT

	Parthenium weed		Forest		Water body		Grassland		Settlements		
Image date	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	OA
19-Jan	83	68	94	97	100	95	69	67	77	93	83.7
08-Feb	84	75	86	100	100	100	83	74	90	93	88.1
28-Feb	86	64	91	97	100	100	64	67	76	86	82.2
07-Mar	73	68	90	93	100	100	59	63	86	83	80.7
27-Mar	94	61	91	97	100	100	63	81	83	83	83.7

When testing whether there were significant differences in the overall accuracies derived from the various image dates, the Wilcoxon test (Table 4.4) confirmed that the image acquired on 8th February yielded significantly ($p = 0.05$) higher accuracies than the other dates, based on both the RF and EXT algorithms. Consequently, the images acquired on 8th February were identified as the most optimal window date for mapping Parthenium weed.

Table 4.4 Pairwise comparisons of classification accuracies between optimal date and remaining dates (Significance at $p < 0.05$ with critical value at 29)

Pairwise comparison	t-test	Result	t-test	Result
19 Jan to 8 Feb	1	Significant	13	Significant
28 Feb to 8 Feb	0	Significant	5	Significant
7 Mar to 8 Feb	0	Significant	1	Significant
27 Mar to 8 Feb	4	Significant	11	Significant

4.3.3.2 Comparing the computational efficiency of RF and EXT algorithms

Figure 4.4 shows the time taken by each algorithm to discriminate Parthenium weed from other land cover types, using the Sentinel-2 satellite images acquired on different dates. It can be observed that EXT was computationally faster than RF in discriminating Parthenium weed from other land cover types. For example, it took 0.7 seconds, on image data acquired on 19th January, to create the model by using RF, whereas using EXT, it took only 0.48 seconds.

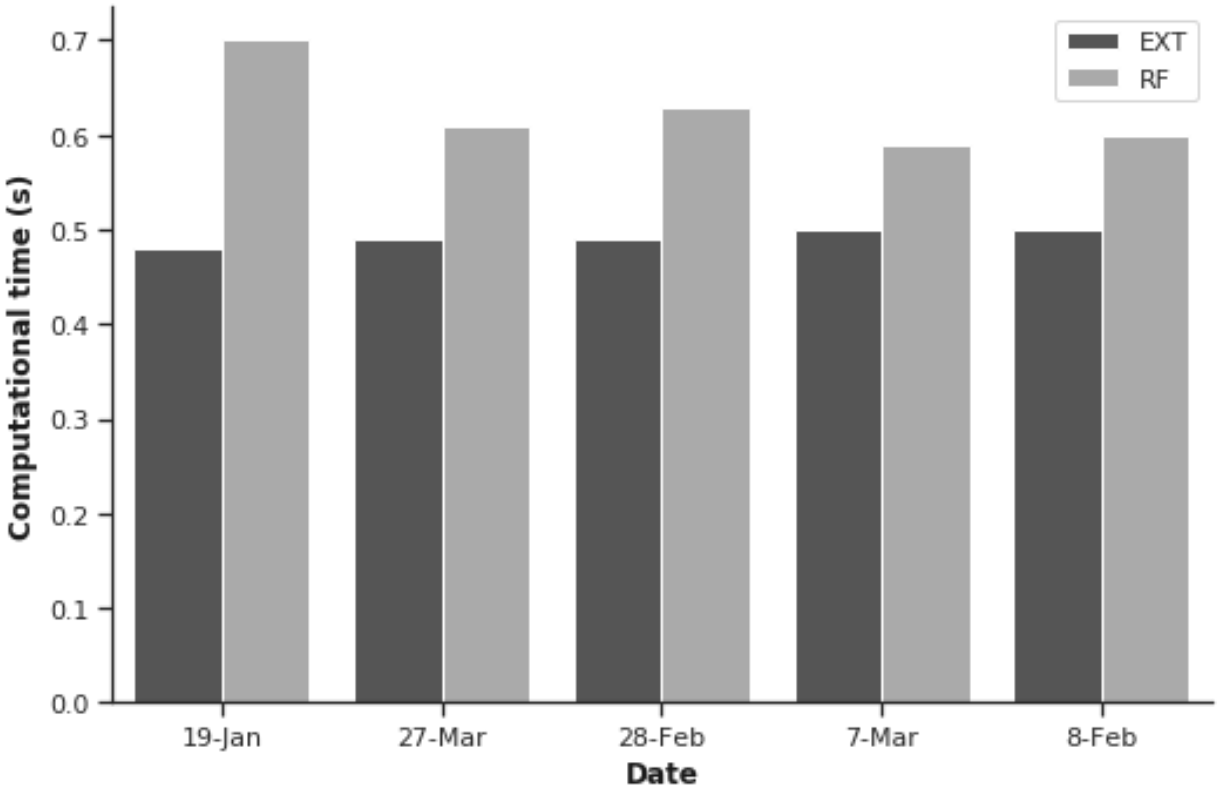


Figure 4.3 Computational time of RF and EXT models at investigated dates

4.3.3.3 Sentinel-2 spectral bands ranking on the optimal window and classified maps using EXT and RF

Figure 4.6 illustrates the importance of the weight of spectral bands of the RF and EXT models at the optimal date. RF and EXT did not rank their variables in the same way, but Bands 4 (664 nm), 2 (497 nm), 8 (835 nm), 5 (704 nm) and 3 (560 nm) were repeatedly ranked as the highest in both models. Bands 6 (740 nm) and 7 (782 nm) were the least important.

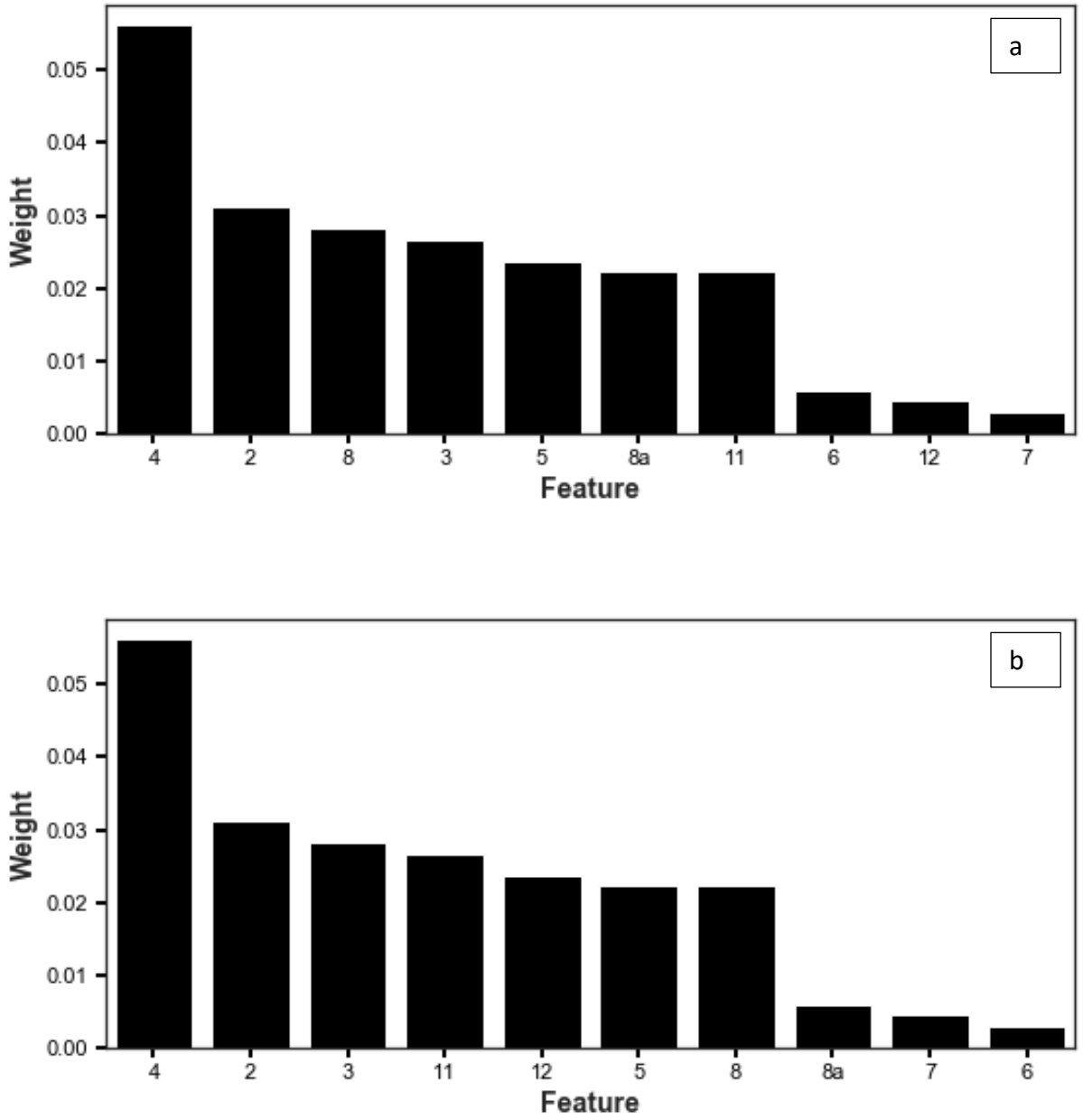


Figure 4.4 Variable importance generated by RF (a) and EXT (b)

Figure 4.6 shows the maps of Parthenium weed infestations and surrounding land cover types on images collected on 8th February, using RF and EXT. As the accuracies of their models were similar, the investigated classes were also similarly mapped. Overfitting was pronounced between grasses and Parthenium weed in low-density areas. With respect to the level of infestations in different land cover types, forested areas (in the lower center and North-East) seemed to be less infested by Parthenium weed. Low-density areas (in the North) were the most infested.

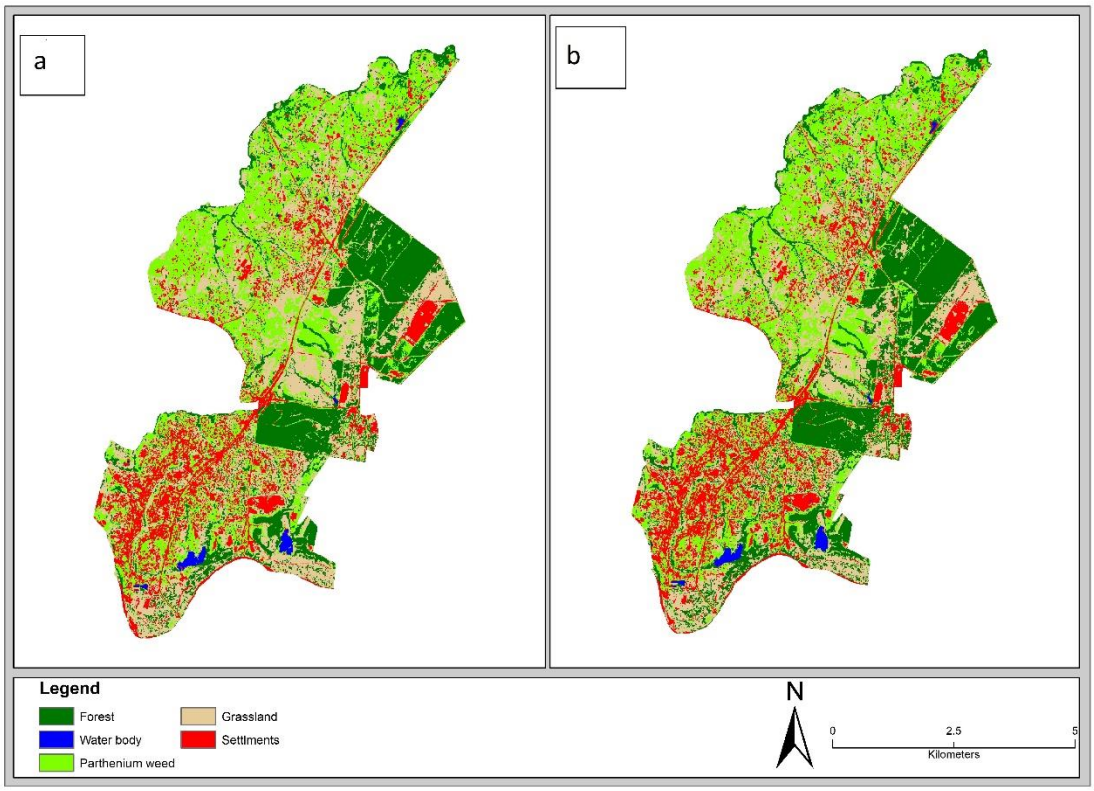


Figure 4.5 Maps of Parthenium weed and surrounding land cover types at the optimal temporal window, using RF (a) and EXT (b)

4.4 Discussion

This study sought to find the optimal temporal window for mapping Parthenium weed, using multi-date Sentinel-2 images. The study also sought to evaluate the efficiency of EXT compared to RF, based on their derived classification accuracies and computing time. Furthermore, both classifiers were also implemented for ranking spectral bands within that window period.

4.4.1 Finding the optimal temporal window for mapping Parthenium weed

Figures 4.2 and 4.3 showed that the temporal variation of overall classification accuracies and the f-score of Parthenium weed could be related to the EVI values of Parthenium. When the EVI values of Parthenium weed were high, the overall classification accuracies and f-score were low. According to Zhang et al. (2003) and Sepúlveda et al. (2018), the EVI and temporal dynamics of vegetation productivity are directly correlated, specifically at the onset and the peak of photosynthetic activity. This is a clear indication that the amount of Above Ground Biomass

(AGB), due to full canopy cover of Parthenium weed, may have influenced the classification accuracies in the mapping on Parthenium weed. For example, when using EXT during the peak of vegetation productivity (Figure 4.2) on 28th February, the acquired image yielded an overall classification of 81.5%, whereas during the least stage of vegetation productivity (Figure 4.2) on 8th February, the acquired image produced a higher overall classification (88.1%).

The least performance of EXT and RF during the period of full canopy cover can be attributed to a saturation problem that is associated with multispectral sensors (Mutanga and Skidmore, 2004; Adjorlolo et al., 2012; Shoko et al., 2018). The literature states that optical satellite images tend to fail to detect the inherent variation in vegetation in areas of great biomass density, due to the high saturation (Morandeira et al., 2016; Guo et al., 2017; de souza Mendes et al., 2019). In this instance, the high density of both Parthenium and the co-existing plant species could have resulted in the saturation of the Sentinel-2-derived signals, hence hindering their optimal discrimination and resulting in low classification accuracies. Moreover, the canopy of Parthenium weed and surrounding grasses have an erectophile architecture, which could have interacted with the incident radiation in a similar manner, resulting in impartial discriminations and low classification accuracies (Vaiphasa et al., 2007; Miphokasap et al. 2012; Adkins and Shabbir, 2014). Therefore, the high performance of images acquired on 8th February could be attributed to the fact that the density and architecture of the canopy cover of Parthenium weed and surrounding vegetation species, such as grasses, at that period were relatively lower and less complex, attenuating the saturation problem and interaction of incident radiation.

4.4.2 Comparison between EXT and RF

RF is one of the popular and most commonly-used classifiers in land cover classifications using hyper- and multi-spectral data. However, to date, EXT has been under-utilized for satellite image processing. The results showed that EXT outperformed RF in mapping the Parthenium weed. The higher performance of EXT was particularly noticeable at non-optimal periods for mapping Parthenium weed. According to Kiala et al. (2019), EXT tends to give better results than RF because of its higher grade of randomness during the training process. This yields more trees that are independent, and thus decreasing any further variance. In addition, EXT model was developed faster than the RF models in all the time periods that were tested (Figure 4.4). The higher speed of

EXT can be explained by the fact that the algorithm does not search exhaustively for the optimal split RF does. This simplifies the computational load during the training stage (Barrett et al., 2014). Our findings are similar with previous studies that highlighted the superiority of EXT over RF (Barrett et al., 2014; Li et al., 2017; Lu et al., 2017). For example, Barrett et al. (2014) compared EXT with SVM and RF for grassland type discrimination using ancillary, multi-temporal and multi-sensor radar spatial datasets. They found that the EXT classifier outperformed SVM and RF for most of the datasets. Geurts et al. (2006) and Kiala et al. (2019) also found that EXT models were faster than RF for training models. We can therefore recommend EXT for the satellite image classification of landscapes infested by Parthenium weed.

4.4.3 Spectral band ranking using EXT and RF at the optimal window period

Ranking spectral bands at the optimal window period may give an indication as to which vegetation indices to use for improving mapping accuracy. The two classifiers ranked the spectral bands with a high spatial resolution (e.g. the Red, Green, Blue), as well as NIR and red-edge (at 704 nm) bands of Sentinel-2, as the most optimal features for mapping Parthenium weed. The large contribution of NIR and red-edge (704 nm) bands was expected, particularly when phenological stages of Parthenium weed and the surrounding plant species are the most asynchronous, (Laba et al., 2005). The NIR and red-edge bands were reported to be species-specific, because of their sensitivity to different internal leaf structure, leaf pigments and water (Vaiphasa et al., 2007). However, as also established by Shoko et al. (2018), it is worth noting that not all the NIR and red-edge bands were the most relevant in developed models. Overall, our findings agree with previous studies. For instance, Arogoundade et al (2019) found that Band 5 (red-edge at 704 nm) of Sentinel-2 was one of the most important variables in predicting a habitat that is susceptible to famine weed invasion. Kganyago et al. (2017) applied a hybrid method to the hyperspectral data to find useful subsets of the spectral bands in discriminating Parthenium weed from coexistent species. They concluded that most of the suitable bands were situated in the NIR and SWIR. The selection of spectral bands with a high spatial resolution may be attributed to the fact that they suffer less from mixed pixels (Kganyago et al., 2018). Previous studies also highlighted the importance of high spatial resolution wavebands in detecting weed infestation (Lass et al., 2005; Huang and Asner, 2009).

4.5 Conclusions

Grounded on the main findings of this study, it can be concluded that:

a) the window period in which Parthenium weed can be discriminated from the surrounding plant species is located at the beginning of February, based on spectral bands such as Red (664 nm), Blue, NIR (835 nm), Red-edge (704 nm) and Green (560 nm); and

b) the EXT classifier outperforms RF for most of the images acquired at different phenological growth stages.

The findings of this study are valuable for choosing the optimal period for mapping Parthenium weed, which is very crucial for achieving an accurate representation of its spatial distribution, hence improving decision-making for the weed invasion control and mitigation. Further research should look at implementing the results of this study in other areas for validation in a multi-year scenario.

CHAPTER 5. A HYBRID FEATURE METHOD FOR HANDLING REDUNDANT IN A SENTINEL-2 MULTI-DATE IMAGE FOR MAPPING PARTHENIUM WEED

This chapter is based on: **Zolo Kiala**, Onisimo Mutanga, John Odindi, Serestina Viriri and Mbulisi Sibanda. "A hybrid feature method for handling redundant features in a Sentinel-2 multi-date image for mapping the Parthenium weed", Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13:3644-3655, DOI: 10.1109/JSTARS.2020.3001564.

Abstract

Multi-date images provide additional spectral information that is valuable for mapping plant species. However, correlated or redundant variables created from multiple image data and a large set of variables impede accurate and efficient landscape classification. Nevertheless, with the implementation of an appropriate feature selection method, the full potential of multi-date images and subsequent increased classification accuracies may be achieved. Feature selection is a process of automatically selecting features in a dataset that contribute the most to the prediction of a target variable. It improves the classification process in terms of the computation cost and predictive accuracy. Feature selection algorithms are typically classified into three groups, namely, filters, wrappers and those that are embedded. Due to the inherent tradeoffs provided by different feature selection approaches, we hypothesize that a hybrid approach could optimize landscape delineation by leveraging their complementary strengths. In this regard, a new feature selection method that combines a filter ReliefF, a wrapper svm-b, and the embedded RF is proposed hereby in mapping the noxious Parthenium weed, using a Sentinel-2 multi-date image. The new approach was compared to its three constituents, based on the size of the optimal feature subsets and classification accuracies across the three datasets and divided into varying training and test data set ratios. The new approach was also evaluated against the multi-date image without feature selection and a single-date image. The results showed that the new approach yielded the highest overall accuracies with the smallest optimal feature subsets. For instance, on Dataset 3, the overall accuracy was 86.6% with 22 optimal features, whereas it was 84.7% with 35 optimal features when using svm-b, which was the second-most performing feature selection algorithm. The new approach yielded higher classification accuracies than the multi-date image, without feature selection, and the single-date image. The findings of this study under score the capability of hybrid methods to select fewer features from multi-date images, with higher predictive accuracies than individual feature selection methods.

Keywords: Parthenium weed, multi-date image, single-date, hybrid feature selection method

5.1 Introduction

The recent advancements in space-sensor technologies have facilitated investigations into the potential of new multispectral sensors, such as GeoEye, WorldView-2 and RapidEye, in understanding the biology of invasive alien species (Mullerova et al., 2005; Peerbhay et al., 2016). Specifically, new multispectral sensors, together with sophisticated machine learning techniques, can provide the accurate spatial distribution of invasive alien species, which is critical for informing appropriate mitigation strategies (Matongera et al., 2017). Sentinel-2 multispectral imagery is one of the new sensors that offer an innovative optical multispectral scanner with a wide-swath of 290 km, a fairly high spatial of up to 10 m and a spectral resolution of 13 spectral bands. In addition to being freely accessible, Sentinel-2 data have a five-day global revisit time, which is a valuable characteristic for improving vegetation mapping. Vuolo et al. (2018) for instance, showed that the additional multi-temporal information of Sentinel-2 image data increases the classification accuracy of nine crop types for 2016 and 2017 in an agricultural region of Austria. The improvement can be explained by embedded information, such as phenological and canopy structural properties (Tottrup, 2004). However, multi-date image data are not always superior to single-date images. Some studies (MacLean and Congalton, 2013; Meddens et al., 2013; Hościło and Lewandowska, 2019) have proved that multi-date images do not improve the classification accuracies, when compared to single-date images. This is probably due to the correlated or redundant variables created from closely-acquired image data. Such variables are known to decrease the performance of learning algorithms during the classification process (Zhu et al., 2007; Venkatesh and Anuradha, 2019). Moreover, multi-date images contain a large set of variables that misguide the commonly-used machine learning techniques (Thejas et al., 2019). Therefore, by implementing an appropriate variable selection, the full potential of multi-date image classification and optimal accuracies can be achieved.

Variable or feature selection methods are typically subdivided into three groups namely, filter methods, wrapper methods and embedded methods (Chuang et al., 2011; Cao et al., 2017). Filter methods (e.g. ReliefF, gini-index) rely on the internal characteristics of data and are implemented as a preprocessing step, which is independent of the induction algorithm (Hsu et al., 2002). Since they do not use a learning algorithm to select an optimal subset of features, they are generally fast. Meanwhile, wrapper methods (e.g. Support Vector Machines Recursive Feature Elimination

(SVM_RFE) and svm-b evaluate a subset of features by using a specific induction algorithm (Cao et al., 2017). They take advantage of robust classification algorithms to achieve better predictive accuracy (Peralta and Soto, 2014; Li et al., 2017). On the other hand, embedded methods (e.g. RF search for optimal subset of features that is built in the induction algorithm. They are a trade-off between the filter and wrapper methods (Li et al., 2017). Embedded methods are relatively faster than wrapper methods and better than filter methods in predictive accuracy (Gnana et al., 2016).

However, the three aforementioned groups of feature selection methods have their respective disadvantages. For instance, since the correlation between features and classifiers is not considered in the filter methods, a subset of the selected features may not be optimal to the target induction algorithm (Zhu et al., 2007). Wrapper methods are computationally intractable when the feature dimension becomes very high. Embedded methods are typically based on a greedy mechanism; for instance, only top-ranked attributes are used to perform sample learning (Chuang et al., 2011). Moreover, embedded methods, such as RF, have difficulty in discarding redundant features from the datasets (Zhou et al., 2018). In this regard, a combination of the advantages of the filter, wrapper and embedded methods can achieve both a high predictive accuracy and a reasonable computational cost.

Hybrid feature selection methods generally use the strength of the filter and wrapper feature selection methods. Typically, the first feature dimension of data is reduced by using a filter method, then wrapper method is implemented, for the selection of optimal feature subset (Venkatesh and Anuradha, 2019). These approaches are usually faster than wrapper-based methods and yield better accuracies than the filter methods (Rouhi and Nezamabadi-pour, 2017). Previous studies have proposed different hybrid feature selection methods (Xie and Wang, 2011; Lin et al., 2012; Lu et al., 2017; Venkatesh and Anuradha, 2019). For example, Lin et al. (2012) proposed a hybrid method MI-SVM-RFE, which combines Mutual Information (MI) and SVM-RFE, for selecting the discriminative metabolites from high dimension metabolome data. They found that MI-SVM-RFE yielded a higher overall accuracy (74.3%) than SVM-RFE (72%) in distinguishing three liver diseases on one of the datasets. However, only filter and wrapper methods were combined in those studies. In addition, applications of hybrid methods on multi-temporal earth observation datasets, such as Sentinel-2, have not yet been explored. The use of hybrid feature methods is critical for accurately mapping invasive species, using multi-date images, since they can select fewer spectral

bands or features with high accuracies. In their research, Hall and Holmes (2003); Robnik-Šikonja and Kononenko (2003) and Kiala et al. (2019) found that ReliefF, a filter method, and svm-b, a wrapper method, could select small subset of optimal features and yield high classification accuracies, respectively. Therefore, we propose hereby a novel hybrid feature method that combines ReliefF, svm-b and RF for handling correlated variables in a multi-date Sentinel-2 image when mapping a landscape infested by the common Parthenium weed invasive plant species. Specifically, the new approach was first compared with its individual feature selection method component on three datasets, in terms of its predictive accuracy and the size of optimal feature subset. A comparison of the classification accuracies was then performed between the optimal feature subset selected by the new approach and a single-date image and a multi-date image, without any feature selection applied.

5.2 Materials and Methods

5.2.1 Reference data

As shown in Table 5.1, 447 sampling points were created to map Parthenium weed and major land cover classes in this study. To assess the efficiency of our proposed method and its constituent feature selection methods, these field-based data were randomly split into training and test sets in three different ratios: 3:1 (Dataset 1); 1:1 (Dataset 2); 1:3 (Dataset 3) (Table 5.1).

Table 5.1 Description of the datasets

	Dataset 1		**Dataset 2**		**Dataset 3**	
	Training set 1 (70 %)	Test set 1 (30 %)	Training set 2 (50 %)	Test set 2 (50 %)	Training set 3 (30 %)	Test set 3 (70 %)
Forest	70	30	50	50	30	70
Water body	49	21	35	35	21	49
Parthenium	63	27	45	45	27	63
Grassland	64	28	46	46	28	64
Settlement	66	29	48	48	29	66

5.2.2 Acquisition of multi-date Sentinel-2 images

According to Vanangamudi et al. (2013), Parthenium weed plants typically germinate between September and December and they senesce between March and May. Hence, four Level 1C Sentinel-2A satellite images, which spanned across the dominant phenological events (i.e. rosette

growth, flowering, senescence) were acquired on 19th January, 8th February, 28th February and 27th March, and they were downloaded from the website of the ESA. The multi-date image was then created by layer stacking these four single-date images (MacLean and Congalton, 2013). In this study, only the single-date image with the highest classification accuracies was retained for comparison with the multi-date image. Table 5.2 provides the characteristics of the Sentinel-2 imagery.

Table 5.2 Spectral band configuration of Sentinel-2A

Band Number	Central wavelength (nm)	Bandwidth (nm)	Pixel size (m)
1	442.7	44	60
2	492.4	94	10
3	559.8	45	10
4	664.6	38	10
5	704.1	19	20
6	740.5	18	20
7	782.8	28	20
8	832.8	147	10
8a	864.7	44	20
9	945.1	26	60
10	1373.5	75	60
11	1613.7	143	20
12	2202.4	242	20

5.2.3 Ecology of Parthenium weed

Parthenium weed (*Parthenium hysterophorus*) is an upright annual and herbaceous weed of the *Asteracea* family (Figure 5.1). It is native to neo-tropical regions around the Gulf of Mexico and central Argentina. To date, it has spread to pan-tropical regions (Javaid and Anjum, 2005; Belz et al., 2007). Parthenium weed niches are mainly disturbed areas of land that are characterized by poor ground cover, such as roadsides, railway tracks, construction sites, wastelands, overgrazed pastures and cleared lands (Adkins and Shabbir, 2014). Once established, Parthenium weed can fulfill its life cycle in four weeks. However, under ideal climatic conditions, it may germinate, grow and flower at any time of the year (Masum et al., 2013). Summer is the principal growth season, when temperatures range from 10 to 25^{o}C and rainfall is greater than 500mm per annum

(McConnachie, 2015). Its aerial parts cannot effectively resist frost; hence the majority of plants perish under low temperatures in winter although after a mild winter, regrowth from old stem bases may take place for some plants (Adkins and Shabbir, 2014). Parthenium weed tolerates several types of soil, ranging from sandy loams to clay loams; however, it prefers cracking clay soils of high fertility, as well as black and alkaline soils (Swati et al., 2013; Kaur et al., 2014). Furthermore, Parthenium weed is extremely prolific in seed production (about 20,000 seeds per plant). Its seed bank is estimated to be more than 340 million seeds per ha in abandoned fields (Goodall et al., 2010). In some sites of South Africa, an estimated 958 million seeds per ha have been recorded (Strathie et al., 2011). Parthenium weed seeds may be dispersed by animals (domestic and wild), vehicles, farm machinery, and river or flood water (Strathie et al., 2011; Swati et al., 2013). The above ecological attributes make the control of Parthenium weed very challenging.

Figure 5.1 Photograph of a Parthenium weed plant

5.2.4 Feature selection methods

5.2.4.1 ReliefF

ReliefF is a multi-class version of the Relief algorithm family (Farrell et al., 2019). The principle of ReliefF is to estimate the importance of features based on how well their values are different among the instances that are close to each other (Zhu et al., 2007). Let us assume that S, is a sample

set, R, is a selected sample instance from S, K is found near the neighbors of samples R, NH, ('near-hit') is the closest instance of sample R within the same class, NM, ('near-miss') is the closest instance of sample R among the different classes, and w_t, is the weight of feature t, which is updated after m times of the feature evaluation. The formula of the final weight of t (w_t^i) is calculated as follows (Zhou et al., 2018):

$$w_t^i = w_t^{i-1} + \frac{\sum c \neq class(x) \frac{p(x)}{1-p(class(x))} \sum_{j=1}^{k} diff(x, M(x))}{m * k} - \frac{\sum_{i=1}^{k} diff(x, H(x))}{m * k} \quad \text{(Equation 5.1)}$$

Where:

$M(x)$ and $H(x)$ stand for the closet sample in a same class and in different classes of sample x, respectively;

$diff(\,)$ indicates the distance of the sample of feature t;

$p(\,)$ denotes the ratio of the whole samples in class c_i to all heterogeneous samples in S; and

m and k represent the number of iterations and nearest neighbors, respectively.

5.2.4.2 Support Vector Machines Backward

svm-b and Support Vector Machines Forward (svm-f) rank features according to their predictive power, using the classical SVM in a backward or forward selection strategy, respectively (Guyon and Elisseeff 2003; Deng et al., 2013). In the forward selection strategy, the search for relevant features starts with an empty set of features, then the features are progressively added into larger subsets, whereas in the backward elimination, it starts with the full set of features and then progressively eliminates the least relevant ones (Kohavi and John, 1997). According to Kiala et al. (2019), svm-b is faster and more efficient in its predictive accuracy than svm-f. The code behind svm-b can be found in the skfeature Python package (Li et al., 2017).

5.2.4.3 RF

The RF classifier provides a self-contained importance measure for each feature, when calculating the mean decrease ($\overrightarrow{D_J}$) in the classification accuracy for the Out Of Bag (OOB) data from the bootstrap sampling. The following steps are used to compute the variable importance measure. First, given bootstrap samples $b = 1, \dots, B$, the $\overrightarrow{D_J}$, the mean decrease for the variable x_j is computed as follows (Ma et al., 2017):

$$\overrightarrow{D_J} = \frac{1}{B}\sum_{b=1}^{B}(R_b^{oob} - R_{bj}^{oob}) \quad \text{(Equation 5.2)}$$

Where R_b^{oob} stands for the classification accuracy for OOB data l_b^{oob} using Tb as the classification model; R_{bj}^{oob} is the classification accuracy for OOB data l_{bj}^{oob} by permuting the values of Variable x_j in l_b^{oob} $(j = 1, \dots, N)$. Second, the variable importance (z-score) of Variable x_j can be calculated as follows:

$$z_j = \frac{\overrightarrow{D_J}}{s_j/\sqrt{B}} \quad \text{(Equation 5.3)}$$

Where s_j represents the standard deviation of the classification accuracy decrease.

5.2.4.4 Proposed feature selection method

Figure 5.2 displays the pseudo-code of the proposed hybrid feature selection method. The first stage of the new approach consisted of creating a range of numbers starting from 1 to N, which is the number of bands of the multi-date image. Each number in the range corresponded to the size of the feature subsets to be selected through ReliefF. Then, the RF model was trained and evaluated on test dataset, using a selected feature subset through an iteration. The subset of selected features with the highest overall accuracy was considered as the output of the first stage. In the second stage, the steps of the previous stage were repeated using the optimal features selected by ReliefF as input and svm-b as feature selection. In the third stage, the resulting optimal features through svm-b were ranked by the RF algorithm, using the Mean Decrease in Impurity (MDI). Another

interaction was implemented on different subsets of the ranked features generated, using the "SelectFromModel" function of the sklearn package (Pedregosa et al., 2011). The subset with the highest predictive accuracy was the final output of the hybrid method.

```
Input:  original feature set (X) with N as number of features and Y predictor
Begin
        Feature selection with ReliefF
        for i = 1 to N:
              apply ReliefF on i
              evaluate selected features with Random forest on test dataset
          output subset of features with highest accuracy (Nsvm-b)
        Feature selection with svm-b
        For i = 1 to Nsvm-b:
              apply svm-b on i
              evaluate selected features with Random forest on test dataset
              output subset of features with highest accuracy (Nsvm-b)
        Feature selection with Random forest (Rf)
        Rank features in Rf
        Create different subsets of ranked features
        For i = 1 to T:
             apply Rf on i
             evaluate selected features with Random forest on test dataset
          output optimal feature subset
end
```

Figure 5.2 Pseudocode of the proposed method

5.2.5 Model assessment metrics

To assess the built models in this study, the estimated classes were cross-tabulated against the ground-sampled classes for corresponding pixels in an error matrix. Then, from the error matrix, performance metrics (e.g. the OA, the UA, and the PA) were computed (Lunetta and Lyon, 2004). The OA refers to the proportion of all the classes that were mapped correctly. The UA refers to the probability that a pixel labeled as a certain class on the map represents that class on the ground. The PA refers to the probability of real features on the ground and are classified as such. Since this study sought to map Parthenium weed, more focus was given to its UA and PA for model assessment. Furthermore, the difference in accuracies among all the created models was

statistically compared, using the Wilcoxon test (Hogg and Craig, 1995). All the analyses and map-making were performed in Python (Version 2).

5.3 Results

5.3.1 Comparison between the new approach and its constituent feature selection methods (Hybrid, ReliefF, svm-b and RF)

Tables 5.3 (a, b, c) show the classification accuracies of Parthenium weed and surrounding land cover types based on the optimal number of selected features, using different feature selection methods on Datasets 1, 2 and 3. The highest overall accuracies and the lowest optimal number of selected features were achieved using the new approach. For instance, on Dataset 1, the overall accuracy was 94,1% with 20 optimal selected features, using the new approach, whereas svm-b achieved 92.6% with 20 optimal selected features. Furthermore, on Datasets 1 and 2, the optimal subset of the selected features was the smallest (11 and 22, respectively) with the highest overall accuracy (90.6% and 86.6%). In terms of the classification accuracy of Parthenium weed, the optimal subset of the selected features, using the new approach, yielded the highest accuracy for Parthenium weed on the three datasets. Overall, the accuracies (PA and UA) of Parthenium weed were above 70%.

Table 5.3 Classification accuracies on Dataset 1 (a), Dataset 2 (b) and Dataset 3 (c) using ReliefF, svm-b, RF and the new approach

a)

	Parthenium weed		Forest		Water body		Grassland		Settlements			
Feature selection	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	OA	*Number of sel.
ReliefF	73	79	96	87	100	100	79	81	93	93	87.4	32
Svm-b	92	82	97	100	100	100	82	85	93	97	92.6	20
RF	83	86	97	93	100	100	85	85	100	100	92.6	32
Hybrid	86	89	97	93	100	100	89	89	100	100	94.1	20

b)

Feature selection	Parthenium weed		Forest		Water body		Grassland		Settlements			
	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	OA	Number of sel.
ReliefF	77	78	98	96	100	100	76	76	94	94	88.4	31
Svm-b	78	87	98	88	100	100	83	78	92	98	89.7	37
RF	80	89	98	98	100	100	85	76	92	92	90.6	38
Hybrid	82	87	98	94	100	100	81	78	94	96	90.6	11

c)

Feature selection	Parthenium weed		Forest		Water body		Grassland		Settlements			
	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	OA	Number of sel.
ReliefF	70	80	92	86	100	100	75	78	89	81	84	31
Svm-b	63	66	97	83	100	100	66	79	89	81	84.7	35
RF	74	66	96	91	100	98	70	90	93	82	85	38
Hybrid	75	75	94	93	100	100	75	81	92	87	86.6	22

*Number of sel.: number of selected features

5.3.2 Comparison between the image with optimal features selected with the new approach and the single-date and multi-date images

Tables 5.4 (a, b, c) display the classification accuracies of the model derived from the single-date image with the highest predictive accuracies (8th February), the stacked multi-date images and the new approach-selected band image for mapping the investigated land cover classes. The results show that the band image selected by the new approach yielded the highest predictive accuracies for the investigated classes. On Dataset 2, the difference in overall accuracy between the band images selected by the new approach and other images was approximately 6%. It is worth noting that the single-date image produced higher accuracies than the multi-date image on Dataset 1. Based on the Wilcoxon test, the difference in classification accuracies between the single-date and multi-date images were not significant ($p > 0.05$) on Dataset 2. Furthermore, the image with selected bands, using the new approach, yielded a higher PA and UA for Parthenium weed than the single-date and multi-date images. On Dataset 2, for instance, they were 82% and 87%, respectively, using the new approach, whereas they were 75% and 78%, respectively, using the single-date image, and 76% and 83%, using the multi-date image.

Table 5.4 Classification accuracies of the single-date and the multi-date image and the image with optimal features selected by the new approach

a)

	Parthenium weed		Forest		Water body		Grassland		Settlements		
Image date	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	OA
08-Feb	88	75	97	93	100	100	72	78	88	97	88.1
Hybrid	86	89	97	93	100	100	89	89	100	100	94.1
Multi-date	71	86	96	90	100	100	83	74	93	90	87.4

b)

	Parthenium weed		Forest		Water body		Grassland		Settlements		
Image date	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	OA
08-Feb	75	78	96	90	100	100	76	76	90	92	86.6
Hybrid	82	87	98	94	100	100	81	78	94	96	90.6
Multi-date	76	83	96	90	100	100	79	73	76	83	87.5

c)

	Parthenium weed		Forest		Water body		Grassland		Settlements		
Image date	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	OA
08-Feb	65	55	92	83	100	100	58	78	84	78	77.6
Hybrid	75	75	94	93	100	100	75	81	92	87	86.6
Multi-date	66	69	94	93	100	100	69	79	91	76	82.7

Figure 5.3 shows the map of Parthenium weed infestations using a single-date image, optimal-band image through the new approach and multi-date images on Dataset 1. Some differences in the spatial distribution of the investigated land cover types can be noted. For example, water bodies were not accurately mapped in the southern part of the study area, using the single-date image. Furthermore, grassland and Parthenium weed were the classes with the most pixel confusion. Forested areas (in the lower center and North-East) were less infested by Parthenium weed. Low-density residential areas (in the North) were the most vulnerable.

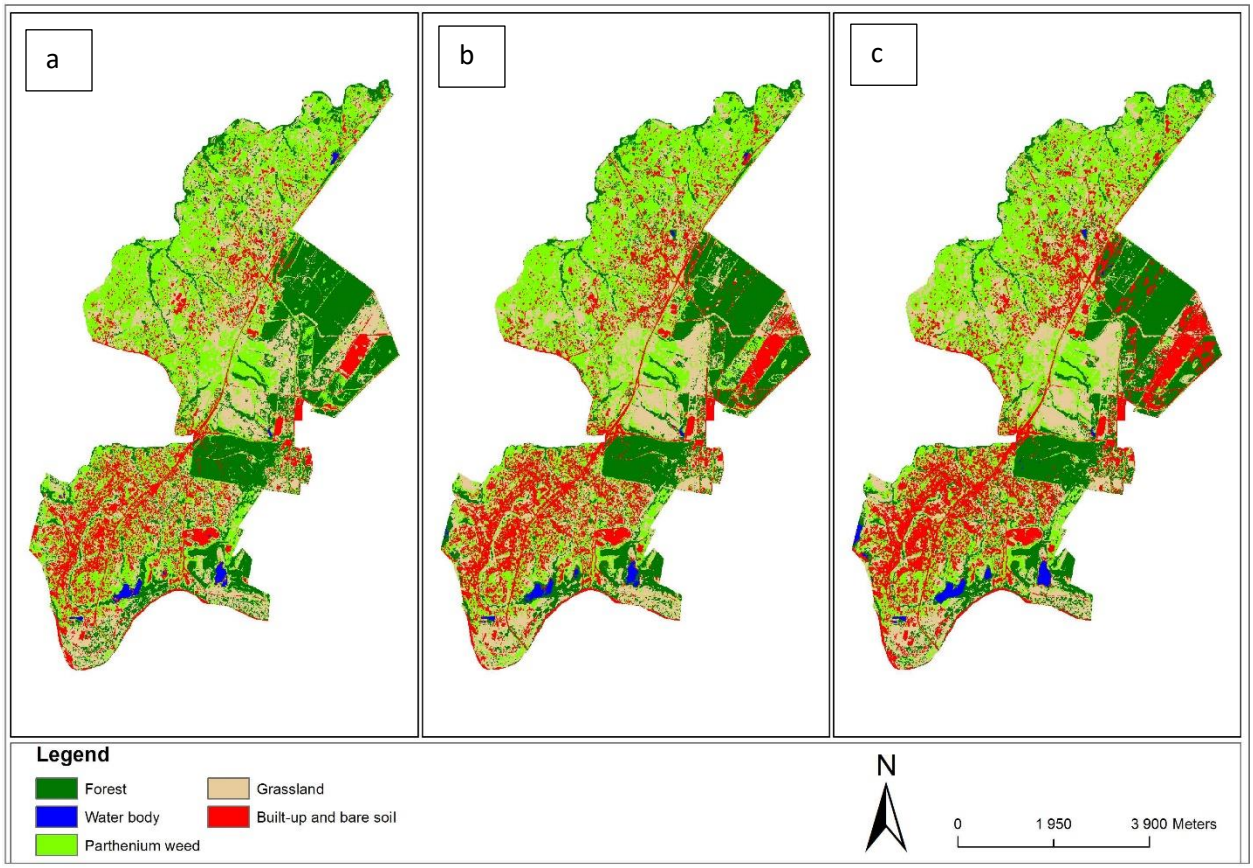


Figure 5.3 Maps of Parthenium weed and surrounding land cover types, using the single-date image (a), optimal band images, using the new approach (b) and multi-date image (c) on Dataset 1

5.4 Discussion

This study proposed a new hybrid feature selection method for reducing the dimension of a multi-date image for mapping a landscape infested by Parthenium weed. In addition, we compared the predictive accuracy of the optimal features selected by the new approach with the predictive accuracy of the multi-date and the best single-date images acquired on 8th February. All the comparisons were implemented across the three sub-datasets obtained by randomly splitting the sample dataset into training and the test datasets in three ratios (1:3, 1:1, 3:1). The performance of the hybrid feature selection algorithm, in relation to ReliefF, svm-b and RF, was evaluated using the PA, the UA and the OA.

The results showed that the optimal features selected by the new approach generated more accurate models than those developed through ReliefF, svm-b and RF. More specifically, the subset of optimal features selected by the new approach was the smallest and the highest, in terms of classification accuracies across the three datasets. For instance, on Dataset 1 (ratio of 3:1 between the training and test data set), the overall classification accuracy was 94.1% with 20 optimal features, whereas it was 87.4% with 32 features using ReliefF, 92.6% with 20 features using svm-b and 92.6% with 32 features using RF. The superiority of the new approach may be explained by the fact that it combines the strengths of its constituent feature selection methods. For instance, ReliefF and svm-b were found to select a small subset of optimal features and to yield high classification accuracies (Kira and Rendell, 1992; Kiala et al., 2019). RF further makes the selection of the most relevant variables from the output of ReliefF and svm-b (Belgiu and Drăguț, 2016). The finding concurs with that of Venkatesh and Anuradha (2019) which showed that a hybrid method made up of MI and Recursive Feature Elimination (RFE), were superior to its constituent feature methods, in terms of classification accuracies and the subset of optimal features. Kganyago et al. (2017) found that a hybrid feature selection method, which combined statistical analyses and SVM-RFE could select fewer bands (10) with higher accuracies (80.19%) than SVM-RFE (76.19%) and RF (66.67%) from field hyperspectral data in discriminating Parthenium weed and co-occurring plant species.

The results showed that the predictive accuracy of the new approach is superior to that of single-date and multi-date images. The difference in the OA ranged from 3% to 9%. It is worth noting that the classification accuracies of multi- and single-date images were not statistically significant on Datasets 1 and 2. These results are consistent with previous studies (Casady et al., 2005; Henry, 2008; Meddens et al., 2013). For instance, Casady et al. (2005) found that the improvement in accuracies of multi-date Ikonos imagery over the single-date images was not significant in detecting a leafy spurge (*Euphorbia esula*). Henry (2008) found that the single-date image approach yielded better classification accuracies than the multitemporal approach for mapping burned areas, using Landsat and CART. MacLean and Congalton (2013) argued that, in cases where the single-date classification was relatively good, the multi-date approach might not be necessary for the accurate creation of the land cover map. Furthermore, multi-date images contain a large set of variables that may be correlated, noisy or redundant, hence impairing the

classification methods (Thejas et al., 2019). However, this study has shown that the new hybrid method has the ability to enhance the potential of multi-date imagery, where an even better single-date classification is achieved.

The results produced by the new approach underscore the fact that hybrid feature selection methods select fewer features and higher classification accuracies than single feature selection methods. This is particularly useful for implementation on multi-date images. The latter contain added spectral information from the additional date for distinguishing plant species, when compared to single-date images (MacLean and Congalton, 2013). This information is particularly important given that plant species have a similar spectral reflectance at their peak biomass season (Ghioca-Robrecht et al., 2008).

5.5 Conclusions

The main findings of this study are as follows:

a) the new hybrid feature selection method outperformed its constituent feature selection methods (ReliefF, svm-b and RF), in terms of classification accuracies and the subset size of optimal features; and

b) optimal bands selected by the new hybrid method yielded higher classification accuracies than single-date and multi-date images.

This study demonstrated the potential of a new hybrid selection method in dealing with redundant features in the multi-date Sentinel-2 image for mapping Parthenium weed-infested landscapes. The removal of redundant variables has hereby been proved to enhance the potential of a multi-date image. The new approach can also be implemented on other high dimensional datasets such as hyperspectral data, where redundant or correlated variables are frequently found. Future work should look at testing other combinations of filters, wrappers and embedded methods. The development of a package would be beneficial for a fast combination.

5.6 Acknowledgments

The authors would like to thank the UKZN-funded BDSS program and the DST/NRF-funded

SARChI Chair in Land use Planning and Management (Grant Number: 84157) for funding this study. Many thanks to the anonymous reviewers for the time they spent on polishing this work.

CHAPTER 6. EXPLORING THE CAPABILITY OF THE TREE-BASED PIPELINE OPTIMIZATION TOOL (TPOT) IN HANDLING A HIGH DIMENSIONAL MULTI-DATE SENTINEL-2 IMAGE DATA FOR MAPPING PARTHENIUM WEED

This chapter is based on: **Zolo Kiala**, Onisimo Mutanga, John Odindi and Romano Lottering. "Exploring the capability of the Tree-based Pipeline Optimization Tool (TPOT) in handling high dimensional multi-date Sentinel-2 image data for mapping Parthenium weed", Under preparation.

Abstract

The TPOT is a new AutoML approach that automatically generates and optimizes tree-based pipelines using a genetic algorithm. The approach is new to the remote sensing community and unknown whether the TPOT can handle high dimensional datasets without significantly impacting both the classification accuracies and computational costs. Multi-date images are generally high dimensional datasets that contain embedded information such as phenological and canopy structural properties, which are known to enhance vegetation mapping. However, without the implementation of a powerful classification algorithm and feature selection, large sets of variables and the presence of redundant variables in multi-date images can impede accurate and efficient landscape classification. In this regard, the overarching goal of this study was to test the efficiency of the TPOT on the multi-date Sentinel-2 image for optimizing the classification accuracies and computational costs of a landscape infested by a noxious invasive plant species, Parthenium weed. Specifically, the models created from the multi-date image and optimal features selected from the multi-date image, using an algorithm system that combines feature selection and the TPOT, were compared with the model created from a single-date image. The results showed that the TPOT can perform well on data with large feature sets. The overall accuracies were 94.1%, 91.1%, 89.6% using the TPOT on the multi-date image, the selected features using the developed system and the single-date image, respectively. The results also showed that feature selection, as a separate pre-processing step to the TPOT, reduced the classification accuracies and computional costs. The study findings are crucial for the automatic and accurate mapping of Parthenium weed, with limited human intervention on high dimensial geospatial datasets.

Keywords: Parthenium weed, multi-date image, single-date, hybrid feature selection method, TPOT

6.1 Introduction

IPSs are rapidly spreading around the world, causing irreversible damage to ecosystems. Parthenium weed (*Parthenium hysterophorus*) is one of the prolific IPSs that adversely impact agricultural production, human and animal health, rural livelihoods, local and national economies, as well as the environment (Swati et al., 2013). Furthermore, in South Africa, it constitutes a threat to globally-recognized biodiversity hotspots like the Maputaland-Pondoland-Albany hotspot and the Isimangaliso Wetland Park in KZN and Eastern Cape provinces. Hence, to mitigate spread, it is necessary to determine its spatial distribution at a relatively low cost (Lawrence et al., 2006).

The recently launched Sentinel-2 sensor provides open-source image data with a wide-swath (290 km) and a relatively higher spatial resolution (up to 10 m) and spectral resolution (13 bands). Moreover, Sentinel-2 data has a five days global revisit time, which is a valuable characteristic for improved vegetation mapping. Vuolo et al. (2018), for instance, showed that additional multi-temporal Sentinel-2 image data increased the classification accuracies of nine crop types in 2016 and 2017 in an agricultural region of Austria; an improvement attributed to embedded information such as the phenological and canopy structural properties (Tottrup, 2004). Nevertheless, without a powerful classifier and/or a decent feature selection method, the correlated or redundant and a large set of variables created from multi-date image can impede accurate and efficient classification of a landscape. In this regard, we hypothesis that an AutoML approach, such as The TPOT together with feature selection, can considerably enhance the quality of the multi-date Sentinel-2 image for generating the weed's accurate spatial representation.

The TPOT is a novel AutoML that was developed by Olson and Moore (Olson and Moore, 2016). The adoption of TPOT limits human intervention by automating algorithm search and optimization. Previous studies have proven that the TPOT could create more accurate models than conventional machine learning techniques (Luo 2016; Olson and Moore, 2016). For example, Sohn and Moore (2017) found that an improved version of the TPOT, the TPOT-MDR, outperformed a tuned logistic regression and XGBoost classifiers. However, AutoML, such as the TPOT, is practically new to the remote sensing community. Furthermore, although TPOT seems promising for geospatial image processing, it requires a lot of time to determine an optimized pipeline (Elshawi et al., 2019). For instance, with its default parameters (i.e. 100 generations with a

population size of 100), the TPOT evaluates 10 000 pipeline configurations to find the recommended pipeline. This makes the TPOT impracticable on high dimensional datasets. The combination of feature selection, as a preprocessing step, and the TPOT would be crucial to overcome this limitation.

Feature selection algorithms are typically classified into three groups namely, filters, wrappers and embedded. Hybrid feature selection methods generally use the strength of the filter and wrapper feature selection methods. Typically, the first feature dimension of data is reduced by using a filter method, then a wrapper method is implemented for the selection of the optimal feature subset (Venkatesh and Anuradha, 2019). These approaches are usually faster than wrapper-based methods, yield better accuracies than filter methods and select fewer features (Rouhi and Nezamabadi-pour, 2017; Kganyago et al., 2017). Kiala et al. (2019), Robnik-Šikonja and Kononenko (2003), Hall and Holmes (2003) for instance, found that ReliefF, a filter method, and svm-b, a wrapper method, could select small subset of optimal features and yield high classification accuracies, respectively. An embedded EXT, a modified version of the RF classifier, was proved to be faster and more accurate than RF (Geurts et al., 2006). Therefore, in this study, a hybrid feature selection method dubbed "ReliefF-Svm-b-EXT", was developed to serve as a preprocessing step to the TPOT.

Hence, the main goal of this study was to evaluate the efficiency of the TPOT on the multi-date Sentinel-2 image for optimizing the classification accuracies and computational costs of a landscape infested by Parthenium weed. Specifically, the models created from the multi-date image and optimal features selected from the multi-date image, using an algorithm system that combines ReliefF-Svm-b-EXT and the TPOT, were compared with the model created from a single-date image.

6.2 Materials and Methods

6.2.1 Reference data

In total, 447 reference points for mapping Parthenium weed and major land use/cover classes were created (Table 6.1).

Table 6.1 Description of the dataset

Class	Number of GPS points
Forest	100
Water body	70
Parthenium weed	90
Grassland	92
Settlement	95

6.2.2 Acquisition of multi-date Sentinel-2 images

Parthenium weed typically germinates between September and December and senesces between March and May (Henry, 2008). Hence, four Level 1C Sentinel-2A satellite images, which spanned across the dominant phenological events (i.e. rosette growth, flowering, senescence), acquired on 19th January, 8th February, 28th February and 27th March, were downloaded from the website of ESA. The multi-date image was created by layer stacking the four single-date images. In this study, only the single-date image, with the highest classification accuracies was retained for comparison against the multi-date image.

Sentinel-2 imagery consists of four spectral bands at a 10 m spatial resolution in blue (Band 2: 497 nm), green (Band 3: 560 nm), red (Band 4: 0.664 um), and the near infrared (Band 8: 0.835 um) spectra. It also includes six bands at a 20 m spatial resolution in red-edge (Band 5: 704 nm, Band 6: 0.740 um, Band 7: 0.782 um), NIR (Band 8a: 0.865 um), SWIR (Band 11: 1.614 um, Band 12: 2.202 um) spectra and atmospheric bands at a 60 m spatial resolution in Coastal aerosol (Band 1: 0.442.7 um), Water vapor (Band 9: 945.1 nm) and SWIR – Cirrus (Band 10: 1373.5).

6.2.3 Feature selection methods

6.2.3.1 ReliefF

ReliefF is a multi-class version of the Relief algorithm family (Farrell et al., 2019). The principle of ReliefF is to estimate the importance of features based on how well their values are different among instances that are close to each other (Zhu et al., 2007). Assuming that S, is a sample set, R, is a selected sample instance from S, K is found near the nearest neighbors of samples R, NH

('near-hit') is the closest instance of sample R within the same class, NM ('near-miss') is the closest instance of sample R among the different classes, and w_t, is the weight of feature t, which is updated after m times of the feature evaluation. The formula of the final weight of t (w_t^i) is calculated as follows (Zhou et al., 2018):

$$w_t^i = w_t^{i-1} + \frac{\sum c \neq class(x) \frac{p(x)}{1-p(class(x))} \sum_{j=1}^{k} diff(x, M(x))}{m * k} - \frac{\sum_{i=1}^{k} diff(x, H(x))}{m * k} \quad \text{(Equation 6.1)}$$

Where:

$M(x)$ and $H(x)$ stand for the closet sample in a same class and in different classes of sample x, respectively;

$diff(\,)$ indicates the distance of the sample of feature t;

$p(\,)$ denotes the ratio of the whole samples in class c_i to all heterogeneous samples in S; and

m and k represent the number of iterations and nearest neighbors, respectively.

6.2.3.2 Svm-b

Svm-b ranks features according to their predictive power, using the classical SVM in backward selection strategy (Guyon and Elisseeff, 2003; Deng et al., 2013). In the backward elimination, it starts with the full set of features and then progressively eliminates the least relevant ones (Kohavi and John, 1997). According to Kiala et al. (2019), svm-b is faster and more efficient in predictive accuracy than svm-f. The svm-b code can be found in the skfeature Python package (Li et al., 2017).

6.2.3.3 EXT classifier

EXT classifier or Extremely randomized tree is a modified version of the RF that was first introduced by Geurts et al. (2006). It is similar to RF in the sense that it constructs independent decision trees to perform classification and regression analyses. However, EXT includes stronger randomization techniques to further reduce the variance of the prediction model. Like RF, EXT provides a self-contained importance measure for each feature when calculating the mean decrease ($\overrightarrow{D_J}$) in the classification accuracy for the OOB data from the bootstrap sampling. The following

steps are used to compute the variable importance measure. First, given the bootstrap samples $b = 1, \ldots, B$, the $\overrightarrow{D_J}$, the mean decrease for Variable x_j is computed as follows (Ma et al., 2017):

$$\overrightarrow{D_J} = \frac{1}{B}\sum_{b=1}^{B}(R_b^{oob} - R_{bj}^{oob}) \quad \text{(Equation 6.2)}$$

Where R_b^{oob} stands for the classification accuracy for OOB data l_b^{oob} using Tb as classification model; R_{bj}^{oob} is the classification accuracy for OOB data l_{bj}^{oob} by permuting the values of variable x_j in l_b^{oob} $(j = 1, \ldots, N)$. Second, the variable importance (z-score) of Variable x_j can be calculated as follows:

$$z_j = \frac{\overrightarrow{D_J}}{s_j/\sqrt{B}} \quad \text{(Equation 6.3)}$$

Where s_j represents the standard deviation of the classification accuracy decrease.

6.2.3.4 The TPOT

The TPOT (Elshawi et al., 2019) is a state-of-the-art AutoML that applies Genetic Programming (GP), using the Python package "Distributed Evolutionary Algorithms in Python" (DEAP) (Fortin et al., 2012) to optimize machine learning pipelines. The TPOT finds the optimized pipeline from a combination of the three types of machine learning pipeline operators, namely, feature preprocessing (StandardScaler, MinMaxScaler, etc.), feature selection (Variance Threshold, SelectKBest, etc.) and classification (DecisionTree, RF, etc.). Most of these machine learning pipeline operators are from the scikit-learn package (Pedregosa et al. 2011). More details on the TPOT can be found in Olson and Moore, (2016). In this study, the TPOT was run on the datasets using its default parameters, which are 100 generations with 100 population size.

6.2.3.5 Proposed system

Figure 6.1 displays the pseudo-code of the proposed system. The proposed system consists of two sections. The first one consists of using a hybrid feature selection method for reducing the dimension of the datasets. The second section consists of applying the TPOT on the features selected, using the hybrid method. The following steps were followed to construct the hybrid

method: first, a range of numbers that starts from 1 to N, which is the number of bands of the multi-date image, was created. Each number in the range corresponded to the size of the feature subsets to be selected through ReliefF. The EXT model was then trained and evaluated on the test dataset, using the selected feature subset through an iteration. The subset of selected features with the highest overall accuracy was considered as the output of the first stage. In the second stage, the steps of the previous stage were repeated, using the optimal features selected by ReliefF as input and svm-b as feature selection. In the third stage, the resulting optimal features through svm-b were ranked by the EXT algorithm, using the MDI. Another interaction was implemented on different subsets of the ranked features generated, using the "SelectFromModel" function of the sklearn package (Pedregosa et al., 2011). The subset with the highest predictive accuracy was the final output of the hybrid method and it serves as the final input of the TPOT.

```
Input: original feature set (X) with N as number of features and Y predictor
Begin
        Feature selection with ReliefF
        for i = 1 to N:
              apply ReliefF on i
              evaluate selected features with EXT on test dataset
         output subset of features with highest accuracy (ReliefF)
        Feature selection with RF
        For i = 1 to Nsvm-b:
              apply ReliefF on i
              evaluate selected features with EXT on test dataset
          output subset of features with highest accuracy (Nsvm-b)
        Feature selection with EXT
        Rank features in RF
        Compute thresholds (T)
        For i = 1 to T:
             apply EXT on i
              evaluate selected features with EXT on test dataset
         output optimal feature subset
        Apply TPOT on optimal feature subset
end
```

Figure 6.1 Pseudocode of the proposed system

6.2.4 Model assessment metrics

Estimated classes were cross-tabulated against the ground-sampled classes for corresponding pixels in a confusion matrix during the model assessment. From the confusion matrix, conventional performance metrics such as the OA, the UA, and the PA, were computed (Lunetta and Lyon 2004). The OA refers to the proportion of all the classes that were mapped correctly. The UA refers to the probability that a pixel labeled as a certain class on the map represents that class on the ground. The PA refers to the probability of real features on the ground and are classified as such. In this study, the focus was on the UA and PA of Parthenium weed class, as it endeavored to map its spatial distribution. All the analyses and map generation were performed using scripts written in Python (Version 2).

6.3 Results

6.3.1 Comparing TPOT models from the multi-date and single-date images

Table 6.2 shows the classification accuracies of TPOT models created from the multi-date image (with and without feature selection) and from the single-date image. The results showed that the highest classification accuracies were achieved with the multi-date image using the TPOT without feature selection. A 94.1%, 91.1% and 89.6% overall accuracies were achieved using the TPOT on the multi-date image, selected features using the developed system and on the single-date image, respectively. Based on the classification accuracies of the individual land cover, the TPOT, when used on the multi-date image, produced the most accurate spatial delineation of Parthenium weed, with PA and UA of 90% and 93%, respectively. The classification accuracies of other land cover types were also the highest. Of all the classes, the water bodies and settlements were the most accurately mapped.

Table 6.2 Classification accuracies of TPOT models from the single-date image, the proposed system and full multi-date Sentinel-2 image

	Parthenium weed		Forest		Water body		Grassland		Settlements		
Method	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	OA
TPOT-8feb	79	82	93	90	100	100	81	78	97	100	89.6
TPOT-hybrid	81	89	97	100	100	100	86	67	94	100	91.1
TPOT-alone	90	93	93	90	100	100	89	89	100	100	94.1

6.3.2 Computation costs of created models

Figure 6.1 shows that the lowest computational costs were achieved on the single-date image and on the optimal features from the multi-date image. The highest computation costs were achieved with the multi-date image using the TPOT alone. The difference in time between the TPOT models from the multi-date image and optimal features selected by the developed system was 11815 seconds (3 hours 18 minutes), which represent a reduction of 17% of the computational costs. The time was reduced by 18% between the TPOT models created from the single-date image and the multi-date image with feature selection. Meanwhile, the reduction in time was 35% between the TPOT model created from the single-date and multi-date images without feature selection.

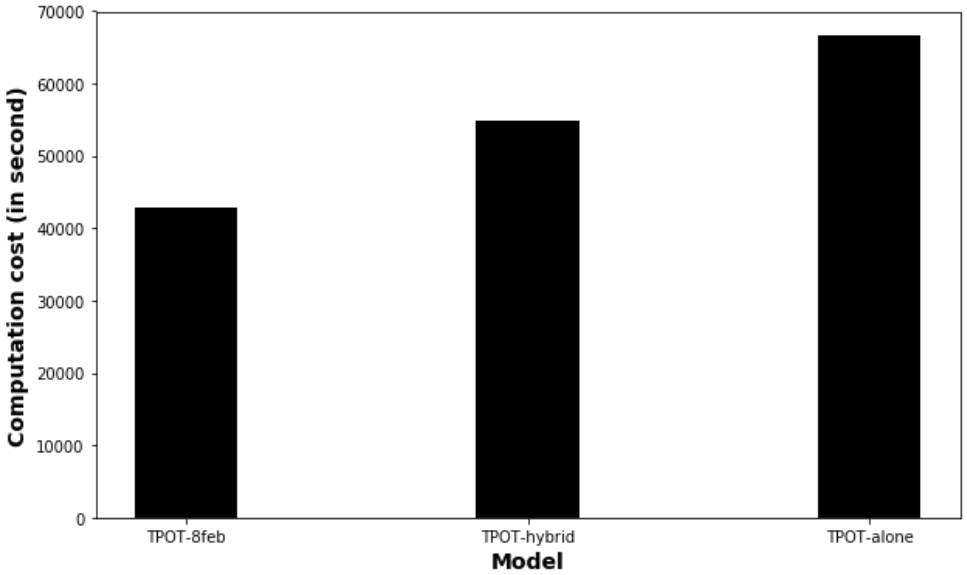


Figure 6.2 Computation cost of the TPOT from single-date image and multi-date image, with and without feature selection

Figure 6.3 shows the map of Parthenium weed infestations using TPOT models from single-date image (a), from full multi-date image and from multi-date with feature selection. It can be noticed that the single-date image did not accurately detect Parthenium weed infestations. The most accurate maps of Parthenium weed were achieved with the multi-date image together with the TPOT. For instance, Parthenium weed infestations was almost non-existent in forested areas. Grassland and Parthenium weed were also less confused in the two maps, even in disturbed areas, such as low-density residential areas (in the North).

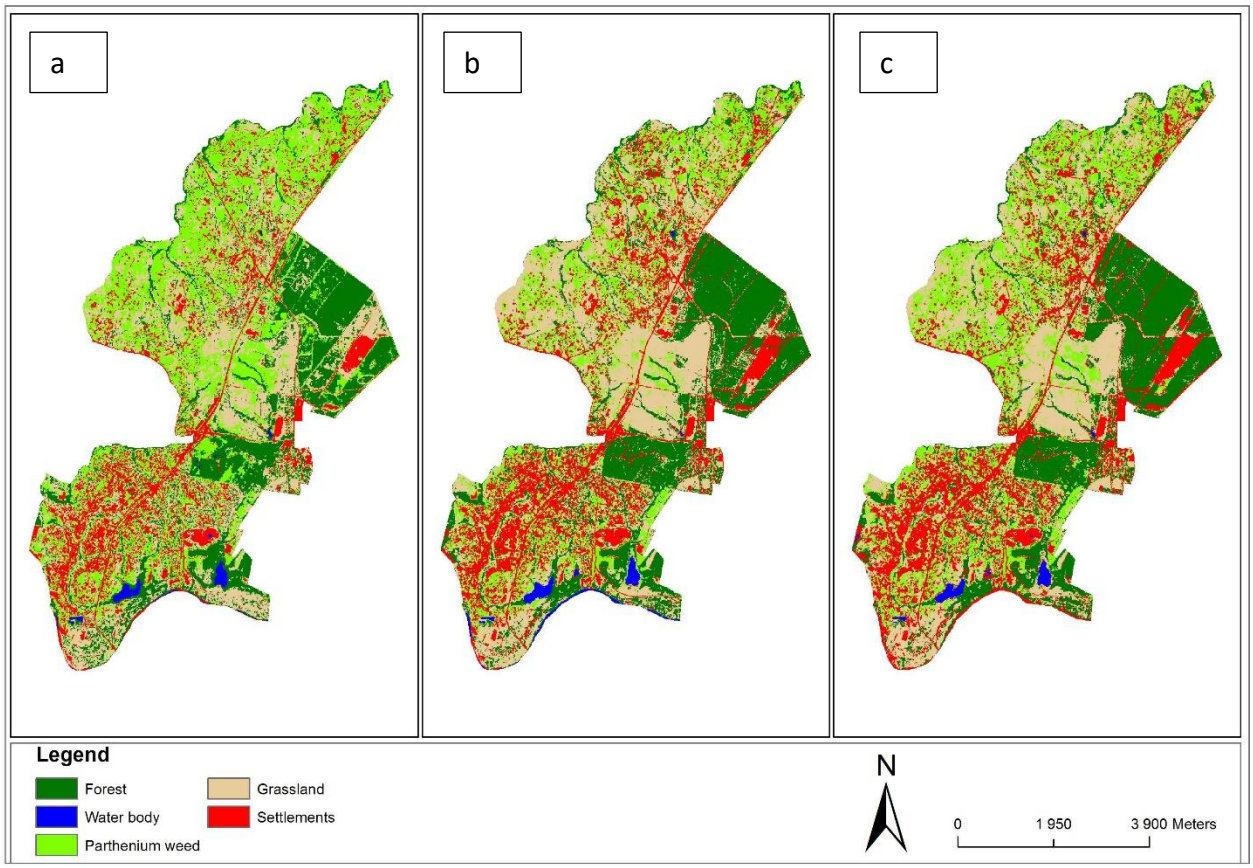


Figure 6.3 Parthenium weed infestations within coexistent land use/cover types using TPOT models created from a single-date image (a), a multi-date image with (b) and feature selection (c)

6.4 Discussion

This study explored the capability of The TPOT for handling a large set of multi-date Sentinel-2 imagery in mapping Parthenium weed infestations and its co-existing land use/covers. To achieve

this goal, an algorithm system, made up of a hybrid feature selection method, ReliefF-Svm-b-EXT, and the TPOT, was developed to determine the impact of feature selection on classification accuracies and computational costs using TPOT. Classification metrics such as OA, PA, UA and were used to assess the different models created in this study.

The results demonstrated that the highest classification accuracies were achieved with the model created from the multi-date Sentinel-2 image. The performance of the TPOT model created from the multi-date image was also superior to the models developed in other studies that mapped the Parthenium weed. For instance, Royimani et al. (2018) mapped Parthenium weed infestations with an overall classification accuracy, PA and UA of 73%, 60% and 61%, respectively using SPOT 6 and RF. Kganyago et al. (2018) found that SPOT 6 yielded an overall accuracy, and a PA and UA of 86%, 72.22% and 93.24%, respectively, using the Support Vector Machines SVM classifier.

The above results show that the TPOT can perform well on high dimensional data such as multi-date Sentinel-2 image, without prior application of a feature selection method. The finding underscores the fact that multi-date images are a good alternative over single-date images for mapping vegetation, particularly when using an appropriate classifier. For example, Casady et al. (2005) found similar results when they compared the IKONOS multi-date and single-date images in mapping a deep-rooted perennial weed, Leafy Spurge (*Euphorbia esula* L.), using the maximum likelihood classifier. Thejas et al. (2019) argued that multi-date images contain a large set of variables, which misguide the commonly-used machine learning techniques. The TPOT's better performance on the multi-date image may be explained by the fact that it intelligently selects algorithms in the recommended pipeline that can handle noisy or redundant features. For example, in this study, the optimized pipeline (Table 6.3) for classifying the multi-date image contained PCA as preprocessor, RFE and EXT classifier in the feature selection process and Gradient Boosting classifier as the classification method. These operators are known to efficiently deal with high dimensional data (Lusa 2017; Samat et al., 2018; Sencaki et al., 2018; Weisenthal et al., 2018). For example, Samat et al. (2018) found that EXT and their proposed method, Extremely Randomized Rotation Forest (ERRF), are capable of producing better classification accuracies than RF in handling high-dimensional data. Weisenthal et al. (2018) stated that Gradient boosting is an ensemble classifier that performs well in high-dimensional data.

On the other hand, in terms of classification accuracies and computational costs, the results demonstrated that the developed system provides a trade-off between the TPOT models from the single-date and multi-date images. When compared to the multi-date image, the computational costs and the overall accuracy of the model created from the optimal features selected by the developed system were reduced by 17% and decreased by 3%, respectively (Figure 6.2 and Table 6.1). This reduction can be attributed to the sequential and complementary use of the three feature selection methods, namely svm-b, ReliefF and EXT, that made up the hybrid feature selection methods. Svm-b is one of the wrapper methods that are known to yield a high predictive accuracy, as they use robust classification algorithms (Peralta and Soto, 2014). The ReliefF algorithm belongs to the filter methods that have a fast runtime (Hsu et al., 2003). EXT is one of embedded methods that are generally a trade-off between the wrapper and filter methods. Svm-b and EXT are also faster and yield higher predictive accuracies than their counterparts, such as svm-f and RF (Geurts et al., 2006; Kiala et al., 2019). Furthermore, previous studies show that hybrid feature selection methods select fewer features and higher predictive accuracies than single feature selection methods (Lin et al., 2012; Kganyago et al., 2017). In this regard, it was expected that the developed system could also increase the TPOT classification accuracies on the multi-date image. However, this was not achieved, probably due to the time allocated to run different TPOT models. Further studies should therefore investigate the time required for the developed system to find a pipeline with similar classification accuracies in the same way as the TPOT model was created from the multi-date Sentinel-2 image.

6.5 Conclusions

Based on the findings, the following conclusions can be drawn:

a) the TPOT can work well on a high dimensional dataset, such as the multi-date Sentinel-2 imagery;

b) the TPOT models from the multi-date image are more accurate than the TPOT model from the single-date image; and

c) combining a hybrid feature selection method with TPOT decreases the computational costs of the TPOT on a high dimensional dataset.

This study was the first to investigate the capability of TPOT to handle high dimensionality in multi-date Sentinel-2 imagery. In the advent of the big data, this study is valuable as it provides a basis for improved landscape delineation by selecting useful features from highly dimensional datasets. Furthermore, the study findings demonstrate the possibility for automatic and accurate Parthenium weed mapping, and indeed other plant species invaded landscapes, with limited human intervention.

Appendix

Table 6.3 Recommended pipelines of TPOT models from the proposed system, full multi-date image and single-date image

Models	Recommended pipelines
TPOT-8feb	Make_pipeline (RBFSampler (gamma=0.85), RandomForestClassifier(bootstrap=False, criterion="entropy", max_features=0.2, min_samples_leaf=2, min_samples_split=4, n_estimators=100))
TPOT-hybrid	Make_pipeline (PCA(iterated_power=8, svd_solver="randomized"), RFE(estimator = ExtraTreesClassifier (criterion = 'gini', max_features = 0.3, n_estimators = 100), step=0.4), GradientBoostingClassifier (learning_rate = 0.1, max_depth = 7, max_features = 0.2, min_samples_leaf = 10, mini_samples_split = 13, n_estimators = 100, subsample = 0.75)
TPOT-alone	Make_pipeline (Normalizer (norm="l1), PolynomialFeatures (degree = 2, include_bias = False, interaction_only = False), StackingEstimator (estimator = DecisionTreeClassifier (criterion = "gini", max_depth = 10, min_samples_leaf = 20, min_samples_split = 13)), GradientBoostingClassifier (learning_rate=0.1, max_depth=7, max_features=0.05, min_samples_leaf=16, min_samples_split=4, n_estimators=100, subsample=0.65))

CHAPTER 7. OPTIMIZING SENTINEL-2 IMAGE FOR MAPPING PARTHENIUM WEED IN SOUTH AFRICA: A SYNTHESIS

7.1 Synthesis

Over the last decades, accurate mapping and monitoring of Parthenium weed in the world, and especially in South Africa have been challenging, probably because of the complex structure of its canopy. This requires more robust classifiers and their performance should also be independent of site characteristics. Moreover, with the recent launch of satellites, such as Sentinel-2, which have high temporal resolution, it is imperative to investigate on suitable feature selection methods to use in combination with these classifiers. This area of study has received less attention for multispectral image data. Large volume of image data with large feature sets can compromise the performance of classifiers. At the same time, they can contain relevant features for accurate mapping of Parthenium weed. Furthermore, it is still unknown when it is suitable to discriminate Parthenium weed from similarly spectral plant species such as grasses, using robust classifiers and appropriate feature selection methods. This temporal window is worth finding out as mapping of Parthenium weed can be focused, hence saving time and resources. Therefore, this study endevoured to tackle these issues in order to advance knowledge on weed mapping in the remote sensing community.

To do so, a Sentinel-2 and Landsat 8 images were optimized for accurate mapping of Parthenium weed in infested landscapes in South Africa. Four possible avenues were implemented to achieve this goal, namely, the application of an automated machine learning technique, the TPOT; the determination of appropriate feature selection methods; the exploration of the optimal temporal window within which it is suitable to map Parthenium weed; and the use a multi-date image instead of a single-date image.

The findings showed that the TPOT increases the classification accuracies of a Sentinel-2 image in mapping Parthenium weed by automatically finding the optimized pipeline, with limited human intervention. Although its computational costs are high, the TPOT achieves higher classification accuracies than manually-selected and parameter-tuned algorithms. The use of the TPOT on earth observation data, as presented in this study, has solved the issue of site-dependent classifiers that have been long faced in the remote sensing community. To optimize the discrimination and subsequent mapping of Parthenium weed, using Sentinel-2, the beginning of February would be the best period for applying the TPOT or any robust classifier, such as RF or EXT classifiers.

Nevertheless, the classification accuracies obtained using a multi-date Sentinel-2 image are superior to those yielded by using a single-date image acquired during this period, if only an appropriate classifier or/and feature selection method is used.

With regard to feature selection, svm-b, a wrapper method, and ReliefF, a similarity-based approach, respectively, yielded the highest accuracies in classifying Parthenium weed and selected the smallest size of optimal features. However, the complementary use of the two feature selection methods, together with an embedded method such as RF or EXT in a hybrid approach, yielded better results. The hybrid feature selection algorithms proposed in this study selected fewer features and produced higher classification accuracies than the single feature selection methods. It is worth noting that the TPOT can still perform well on the high dimensional geo-datasets, such as the multi-date Sentinel-2 image data. It selects operators in the optimized pipeline that can handle the redundant or noisy features.

7.2 Conclusions

This thesis aimed at optimizing the freely available Sentinel-2 and Landsat 8 imagery for the accurate mapping of a landscape that is infested by Parthenium weed (*Parthenium hysterophorus*) in South Africa. This was achieved by addressing some of the challenges that were overlooked in past studies in mapping Parthenium weed. The findings derived from this study have proved that by implementing Sentinel-2 image optimization, it is possible to significantly improve the spatial representation of Parthenium weed in infested landscapes. The main conclusions are as follows:

a) the TPOT is an efficient method for automatically finding the optimized pipeline for Parthenium weed discrimination and monitoring, using the Sentinel-2 image particularly. Its performance is not linked to data characteristics;

b) Svm-b and ReliefF feature selection methods yield the highest classification accuracies and select the smallest subset of features, respectively, on classifying Parthenium weed and co-existent land cover types, using a large Sentinel-2 feature set image;

c) the beginning of February should be the best period for mapping Parthenium weed, using Sentinel-2 image with the Blue, NIR (835 nm), Red-edge (704 nm) and Green (560 nm) bands as the most contributing bands in the developed models;

d) The combination of Svm-b, ReliefF and RF or EXT in a hybrid feature method selects fewer features a with higher predictive accuracies than individual feature selection methods. A model created from a multi-date Sentinel-2 image, using the developed hybrid feature method, is more accurate than a model developed from a single-date Sentinel-2 image; and

e) the TPOT can be applied on high dimensional datasets, such as a multi-date Sentinel-2 image, without affecting the classification accuracies.

Overall, using an automated machine learning approach, feature selection and a determination of the temporal window approaches, this study highlights the significance of optimizing Sentinel-2 image in developing accurate Parthenium weed and surrounding land cover maps. The accurate spatial representation of Parthenium weed is crucial for informing decision-making on mitigation. This study advances the frontiers of knowledge in the remote sensing community by developing innovative ways for detecting and mapping Parthenium weed.

7.3 Recommendations

The retrieval of the accurate extent of Parthenium weed infestations lies in understanding its spatial distribution, in relation to the surrounding environment. In addition to spectral bands, this could be achieved by using ancillary data (e.g. environmental, topographical and edaphic variables, vegetation indices, texture measures) in combination with feature selection methods and classification algorithms for the detection and mapping of Parthenium weed. This study has proved that only the efficient and intelligent implementation of these variables and algorithms can improve the classification accuracies of Parthenium weed and surrounding land cover, using freely-available satellite imagery, such as Sentinel-2. However, this study does not claim to have explored all the possible avenues for optimizing Sentinel-2 and Landsat 8 imagery in mapping Parthenium weed. In this regard, the following recommendations should be considered for future research:

a) the significance of an optimized Sentinel-2 image should be assessed by comparing it to commercial satellite imagery, such as WorldView and SPOT. This would indicate to what extent an optimized Sentinel-2 image can be used as an alternative to commercial image data;

b) although the TPOT was successfully used in this study, the cut-off time at which a TPOT model can achieve acceptable classification accuracies for vegetation mapping, in general, remains unknown. In addition, a modified version of TPOT should be created for spatial data and maybe dubbed the "Spatial Tree-based Pipeline Optimization Tool" (STPOT);

c) feature selection methods have proved in this study to derive useful information from increased Sentinel image data volume. An assessment of the investigated feature selection algorithms was performed using class-based metrics (f-score of Parthenium weed). It would be interesting to use the same approach for evaluating feature extraction methods on Sentinel-2 spectral bands and their derivatives (vegetation indices, texture measures, topography, etc.);

d) this study found that the optimal temporal window for mapping Parthenium weed was at the beginning of February. This needs to be confirmed by implementing the findings in other areas of South Africa, in a multi-year scenario; and

e) the complementary use of the filter, wrapper and embedded methods have been proved to select fewer features with higher classifications accuracies. In this study, Relielf, svm-b and, RF and EXT were used in the developed hybrid methods. The testing of other combinations should be investigated. The development of a package for the hybrid feature selection methods would be beneficial for a fast combination;

REFERENCES

Adam, E. and O. Mutanga (2009). "Spectral discrimination of papyrus vegetation (*Cyperus papyrus* L.) in swamp wetlands using field spectrometry." ISPRS Journal of Photogrammetry and Remote Sensing **64**(6): 612-620.

Adam, E., O. Mutanga, D. Rugege and R. Ismail (2012). "Discriminating the papyrus vegetation (*Cyperus papyrus* L.) and its co-existent species using random forest and hyperspectral data resampled to HYMAP." International Journal of Remote Sensing **33**(2): 552-569.

Adjorlolo, C., O. Mutanga, M. Cho and R. Ismail (2012). "Challenges and opportunities in the use of remote sensing for C3 and C4 grass species discrimination and mapping." *African* Journal of Range and Forage Science **29**(2): 47-61.

Adkins, S. and A. Shabbir (2014). "Biology, ecology and management of the invasive parthenium weed (*Parthenium hysterophorus* L.)." Pest Management Science **70**(7): 1023-1029.

Ahmad, M. W., M. Mourshed and Y. Rezgui (2017). "Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption." Energy and Buildings **147**: 77-89.

Aires, F., V. Pellet, C. Prigent and J. L. Moncet (2016). "Dimension reduction of satellite observations for remote sensing. Part 1: A comparison of compression, channel selection and bottleneck channel approaches." Quarterly Journal of the Royal Meteorological Society **142**(700): 2658-2669.

Anderson, G. L., J. H. Everitt, A. J. Richardson and D. E. Escobar (1993). "Using satellite data to map false broomweed (*ericameria-austrotexana*) infestations on south texas rangelands." Weed Technology **7**(4): 865-871.

Ao, Z., Y. Su, W. Li, Q. Guo and J. Zhang (2017). "One-class classification of airborne LiDAR data in urban areas using a presence and background learning algorithm." Remote Sensing **9**(10): 1001.

Archer, K. J. and R. V. Kimes (2008). "Empirical characterization of random forest variable importance measures." Computational Statistics and Data Analysis **52**(4): 2249-2260.

Arogoundade, A. M., J. Odindi and O. Mutanga (2019). "Modelling Parthenium hysterophorus invasion in KwaZulu-Natal province using remotely sensed data and environmental variables." Geocarto International: 1-16.

Atkinson, J. T., R. Ismail and M. Robertson (2014). "Mapping bugweed (*Solanum mauritianum*) infestations in *Pinus patula* plantations using hyperspectral imagery and support vector machines." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **7**(1): 17-28.

Baker, C., R. Lawrence, C. Montagne and D. Patten (2006). "Mapping wetlands and riparian areas using Landsat ETM+ imagery and decision-tree-based models." Wetlands **26**(2): 465.

Barrett, B., I. Nitze, S. Green and F. Cawkwell (2014). "Assessment of multi-temporal, multi-sensor radar and ancillary spatial data for grasslands monitoring in Ireland using machine learning approaches." Remote Sensing of Environment **152**: 109-124.

Belgiu, M. and L. Drăguţ (2016). "Random forest in remote sensing: A review of applications and future directions." ISPRS Journal of Photogrammetry and Remote Sensing **114**: 24-31.

Belz, R. G., C. F. Reinhardt, L. C. Foxcroft and K. Hurle (2007). "Residue allelopathy in *Parthenium hysterophorus* L.—Does parthenin play a leading role?" Crop Protection **26**(3): 237-245.

Bergstra, J. and Y. Bengio (2012). "Random search for hyper-parameter optimization." Journal of Machine Learning Research **13**(Feb): 281-305.

Bradley, B. A. (2014). "Remote detection of invasive plants: a review of spectral, textural and phenological approaches." Biological Invasions **16**(7): 1411-1425.

Breiman, L. (1996). "Bagging predictors." Machine Learning **24**(2): 123-140.

Breiman, L. (2001). "Random forests." Machine Learning **45**(1): 5-32.

Brownlee, J. (2013). "A tour of machine learning algorithms." Journal of Intelligent Learning Systems and Applications, **9**(4), pp.50-62.

Cao, X., C. Wei, J. Han and L. Jiao (2017). "Hyperspectral Band Selection Using Improved Classification Map." IEEE Geoscience and Remote Sensing Letters **14**(11): 2147-2151.

CARA (1983). Conservation of agricultural resources act. D. O. AGRICULTURE.

Carter, G. A. (1994). "Ratios of leaf reflectances in narrow wavebands as indicators of plant stress." Remote Sensing **15**(3): 697-703.

Carter, G. A., K. L. Lucas, G. A. Blossom, C. L. Lassitter, D. M. Holiday, D. S. Mooneyhan, D. R. Fastring, T. R. Holcombe and J. A. Griffith (2009). "Remote sensing and mapping of tamarisk along the Colorado river, USA: a comparative use of summer-acquired Hyperion, Thematic Mapper and QuickBird data." Remote Sensing **1**(3): 318-329.

Casady, G. M., R. S. Hanley and S. K. Seelan (2005). "Detection of leafy spurge (*Euphorbia esula*) using multidate high-resolution satellite imagery." Weed Technology **19**(2): 462-467.

Chatziantoniou, A., E. Psomiadis and G. Petropoulos (2017). "Co-Orbital Sentinel 1 and 2 for LULC mapping with emphasis on wetlands in a mediterranean setting based on machine learning." Remote Sensing 9(12): 1259

Chen, H.-M., P. K. Varshney and M. K. Arora (2003). "Performance of mutual information similarity measure for registration of multitemporal remote sensing images." *IEEE* Transactions on Geoscience and Remote Sensing **41**(11): 2445-2454.

Chen, L., X. Yang and G. Zhen (2019). "Potential of Sentinel-2 data for alteration extraction in coal-bed methane reservoirs." Ore Geology Reviews **108**: 134-146.

Chirici, G., R. Scotti, A. Montaghi, A. Barbati, R. Cartisano, G. Lopez, M. Marchetti, R. E. McRoberts, H. Olsson and P. Corona (2013). "Stochastic gradient boosting classification trees for forest fuel types mapping through airborne laser scanning and IRS LISS-III imagery." International Journal of Applied Earth Observation and Geoinformation **25**: 87-97.

Christopher, D. M., R. Prabhakar and S. Hinrich (2008). "Introduction to information retrieval." An Introduction to Information Retrieval **151**(177): 5.

Chrysostomou, K. (2009). Wrapper feature selection. Encyclopedia of Data Warehousing and Mining, Second Edition, IGI Global**:** 2103-2108GI Global.

Chu, C., A.-L. Hsu, K.-H. Chou, P. Bandettini, C. Lin and A. s. D. N. Initiative (2012). "Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images." Neuroimage **60**(1): 59-70.

Chuang, L.-Y., C.-H. Yang, K.-C. Wu and C.-H. Yang (2011). "A hybrid feature selection method for DNA microarray data." Computers in Biology and Medicine **41**(4): 228-237.

Clevers, J. and A. A. Gitelson (2013). "Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on Sentinel-2 and-3." International Journal of Applied Earth Observation and Geoinformation **23**: 344-351.

Colkesen, I. and T. Kavzoglu (2018). "Selection of Optimal Object Features in Object-Based Image Analysis Using Filter-Based Algorithms." Journal of the Indian Society of Remote Sensing **46**(8): 1233-1242.

Congedo, L. (2016). "Semi-automatic classification plugin documentation." Release **4**(0.1): 29.

Cooley, T., G. P. Anderson, G. W. Felde, M. L. Hoke, A. J. Ratkowski, J. H. Chetwynd, J. A.Gardner, S. M. Adler-Golden, M. W. Matthew and A. Berk (2002). FLAASH, a MODTRAN4-based atmospheric correction algorithm, its application and validation. IEEE International Geoscience and Remote Sensing Symposium, IEEE.

Cracknell, M. J. and A. M. Reading (2014). "Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information." Computers and Geosciences **63**: 22-33.

Crammer, K., O. Dekel, J. Keshet, S. Shalev-Shwartz and Y. Singer (2006). "Online passive-aggressive algorithms." Journal of Machine Learning Research **7**(Mar): 551-585.

de Neergaard, A., C. Saarnak, T. Hill, M. Khanyile, A. M. Berzosa and T. Birch-Thomsen (2005). "Australian wattle species in the Drakensberg region of South Africa - An invasive alien or a natural resource?" Agricultural Systems **85**(3): 216-233.

de Souza Mendes, F., D. Baron, G. Gerold, V. Liesenberg and S. Erasmi (2019). "Optical and SAR Remote Sensing Synergism for Mapping Vegetation Types in the Endangered Cerrado/Amazon Ecotone of Nova Mutum—Mato Grosso." Remote Sensing **11**(10): 1161.

Deng, S., Y. Xu, L. Li, X. Li and Y. He (2013). "A feature-selection algorithm based on support vector machine-multiclass for hyperspectral visible spectral analysis." Journal of Food Engineering **119**(1): 159-166.

Dhileepan, K. (2007). "Biological control of parthenium (Parthenium hysterophorus) in Australian rangeland translates to improved grass production." Weed Science **55**(5): 497-501.

Díaz-Uriarte, R. and S. A. de Andres (2006). " Gene selection and classification of microarray data using random forest. BMC bioinformatics, **7**(1), p.3.

Dinwiddie, R. (2014). "Composting of an invasive weed species Parthenium hysterophorus L."

dos Santos, A., I. C. d. L. Santos, N. d. Silva, R. Zanetti, Z. Oumar, L. F. R. Guimarães, M. B. d. Camargo and J. C. Zanuncio (2020). "Mapping defoliation by leaf-cutting ants Atta species in Eucalyptus plantations using the Sentinel-2 sensor." International Journal of Remote Sensing **41**(4): 1542-1554.

Du, P., A. Samat, B. Waske, S. Liu and Z. Li (2015). "Random forest and rotation forest for fully polarized SAR image classification using polarimetric and spatial features." ISPRS Journal of Photogrammetry and Remote Sensing **105**: 38-53.

Elshawi, R., M. Maher and S. Sakr (2019). "Automated Machine Learning: State-of-The-Art and Open Challenges." arXiv preprint arXiv:1906.02287.

Erinjery, J. J., M. Singh and R. Kent (2018). "Mapping and assessment of vegetation types in the tropical rainforests of the Western Ghats using multispectral Sentinel-2 and SAR Sentinel-1 satellite imagery." Remote Sensing of Environment **216**: 345-354.

Evans, H. C. (1997). "Parthenium hysterophorus. a review of its weed status and the possibilities for biological control." Biocontrol News and Information: 89N-98N

Farrell, A., G. Wang, S. A. Rush, J. A. Martin, J. L. Belant, A. B. Butler and D. Godwin (2019). "Machine learning of large-scale spatial distributions of wild turkeys with high-dimensional environmental data." Ecology and evolution **9**(10), pp.5938-5949..

Fernandez-Manso, A., O. Fernandez-Manso and C. Quintano (2016). "SENTINEL-2A red-edge spectral indices suitability for discriminating burn severity." International Journal of Applied Earth Observation and Geoinformation **50**: 170-175.

Feurer, M., A. Klein, K. Eggensperger, J. Springenberg, M. Blum and F. Hutter (2015). Efficient and robust automated machine learning. In Advances in Neural Information Processing Systems (pp. 2962-2970).

Fortin, F.-A., F.-M. D. Rainville, M.-A. Gardner, M. Parizeau and C. Gagné (2012). "DEAP: Evolutionary algorithms made easy." Journal of Machine Learning Research **13**(Jul): 2171-2175.

Franklin, S. and M. Wulder (2002). "Remote sensing methods in medium spatial resolution satellite data land cover classification of large areas." Progress in Physical Geography **26**(2): 173-205.

Freeman, E. A., G. G. Moisen, J. W. Coulston and B. T. Wilson (2015). "Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance 1." Canadian Journal of Forest Research **46**(3): 323-339.

Gargiulo, M., D. A. G. Dell'Aglio, A. Iodice, D. Riccio and G. Ruello (2019). "A CNN-Based Super-Resolution Technique for Active Fire Detection on Sentinel-2 Data." arXiv preprint arXiv:1906.10413.

Geurts, P., D. Ernst and L. Wehenkel (2006). "Extremely randomized trees." Machine Learning **63**(1): 3-42.

Ghioca-Robrecht, D. M., C. A. Johnston and M. G. Tulbure (2008). "Assessing the use of multiseason QuickBird imagery for mapping invasive species in a Lake Erie coastal marsh." Wetlands **28**(4): 1028-1039.

Gini, C. (1912). "Variability and mutability, contribution to the study of statistical distribution and relaitons." Studi Economico-Giuricici della R.

Gitelson, A. A., Y. J. Kaufman, R. Stark and D. Rundquist (2002). "Novel algorithms for remote estimation of vegetation fraction." Remote Sensing of Environment **80**(1): 76-87.

Gitelson, A. A., M. N. Merzlyak, Y. Zur, R. Stark and U. Gritz (2001). Non-destructive and remote sensing techniques for estimation of vegetation status. Third European Conference on Precision Agriculture, Montpellier, France.

Giuliani, C., A. C. Veisz, M. Piccinno and F. Recanatesi (2019). "Estimating vulnerability of water body using Sentinel-2 images and environmental modelling: the study case of Bracciano Lake (Italy)." European Journal of Remote Sensing **52**(sup4): 64-73.

Gnana, D. A. A., S. A. A. Balamurugan and E. J. Leavline (2016). "Literature review on feature selection methods for high-dimensional data." International Journal of Computer Applications **136**(1).

Gnana, D. A. A., S. A. A. Balamurugan and E. J. Leavline (2016). "Literature review on feature selection methods for high-dimensional data." International Journal of Computer Applications **975**: 8887.

Goodall, J., M. Braack, J. de Klerk and C. Keen (2010). "Study on the early effects of several weed-control methods on Parthenium hysterophorus L." African Journal of Range and Forage Science **27**(2): 95-99.

Guo, M., J. Li, C. Sheng, J. Xu and L. Wu (2017). "A review of wetland remote sensing." Sensors **17**(4): 777.

Guyon, I. and A. Elisseeff (2003). "An introduction to variable and feature selection." Journal of Machine Learning Research **3**(Mar): 1157-1182.

Hall, M. A. and G. Holmes (2003). "Benchmarking attribute selection techniques for discrete class data mining." IEEE Transactions on Knowledge and Data engineering **15**(6): 1437-1447.

Hall, M. A. and L. A. Smith (1999). Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. In FLAIRS conference (Vol. 1999, pp. 235-239).

Harvey, K. R. and G. J. E. Hill (2001). "Vegetation mapping of a tropical freshwater swamp in the Northern Territory, Australia: a comparison of aerial photography, Landsat TM and SPOT satellite imagery." International Journal of Remote Sensing **22**(15): 2911-2925.

Hastie, T., R. Tibshirani and J. Friedman (2009). "The elements of statistical learning New York." NY: Springer.

Hastie, T., R. Tibshirani and M. Wainwright (2015). Statistical learning with sparsity: the lasso and generalizations, CRC press.

Henrich, V., E. Götze, A. Jung, C. Sandow, D. Thürkow and C. Gläßer (2009). "Development of an online indices database: Motivation, concept and implementation." EARSeL proceedings, EARSeL, Tel Aviv.

Henry, M. C. (2008). "Comparison of single-and multi-date Landsat data for mapping wildfire scars in Ocala National Forest, Florida." Photogrammetric Engineering and Remote Sensing **74**(7): 881-891.

Hitziger, M. and M. Ließ (2014). "Comparison of three supervised learning methods for digital soil mapping: application to a complex terrain in the Ecuadorian Andes." Applied and Environmental Soil Science **2014**.

Hogg, R. V. and A. T. Craig (1995). Introduction to Mathematical Statistics (5th edition), Upper Saddle River, New Jersey: Prentice Hall.

Hościło, A. and A. Lewandowska (2019). "Mapping Forest Type and Tree Species on a Regional Scale Using Multi-Temporal Sentinel-2 Data." Remote Sensing **11**(8): 929.

https://keys.lucidcentral.org/keys/v3/eafrinet/weeds/key/weeds/Media/Html/Parthenium_hysterophorus_(Parthenium_Weed).htm.

Hsu, C.-N., H.-J. Huang and S. Dietrich (2002). "The ANNIGMA-wrapper approach to fast feature selection for neural nets." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **32**(2): 207-212.

Hsu, C.-W., C.-C. Chang and C.-J. Lin (2003). "A practical guide to support vector classification."

Huang, C.-y. and G. P. Asner (2009). "Applications of Remote Sensing to Alien Invasive Plant Studies." Sensors **9**(6): 4869-4889.

Huete, A., C. Justice and W. van Leeuwen (1999). "MODIS vegetation index (MOD13)." Algorithm theoretical basis document **3**: 213.

Hutter, F., L. Kotthoff and J. Vanschoren (2019). "Automatic machine learning: methods, systems, challenges." Challenges in Machine Learning.

Immitzer, M., F. Vuolo and C. Atzberger (2016). "First Experience with Sentinel-2 Data for Crop and Tree Species Classifications in Central Europe." Remote Sensing **8**(3): 166.

Jain, A. and D. Zongker (1997). "Feature selection: Evaluation, application, and small sample performance." IEEE Transactions on Pattern Analysis and Machine Intelligence **19**(2): 153-158.

Javaid, A. and T. Anjum (2005). "*Parthenium hysterophorus* L.–a noxious alien weed." Pak. J. Weed Sci. Res **11**(3-4): 81-87.

Joshi, C., J. d. Leeuw and I. C. v. Duren (2004). Remote sensing and gis applications for mapping and spatial modelling of invasive species. Proceedings of ISPRS.

Joshi, N., M. Baumann, A. Ehammer, R. Fensholt, K. Grogan, P. Hostert, M. R. Jepsen, T. Kuemmerle, P. Meyfroidt, E. T. A. Mitchard, J. Reiche, C. M. Ryan and B. Waske (2016). "A Review of the Application of Optical and Radar Remote Sensing Data Fusion to Land Use Mapping and Monitoring." Remote Sensing **8**(1): 23.

Kanke, Y., W. Raun, J. Solie, M. Stone and R. Taylor (2012). "Red edge as a potential index for detecting differences in plant nitrogen status in winter wheat." Journal of Plant Nutrition **35**(10): 1526-1541.

Kaur, M., N. K. Aggarwal, V. Kumar and R. Dhiman (2014). "Effects and Management of Parthenium hysterophorus: A Weed of Global Significance." International Scholarly Research Notices **2014**: 368647-368647.

Kavzoglu, T. and P. Mather (2002). "The role of feature selection in artificial neural network applications." International Journal of Remote Sensing **23**(15): 2919-2937.

Kganyago, M., J. Odindi, C. Adjorlolo and P. Mhangara (2017). "Selecting a subset of spectral bands for mapping invasive alien plants: a case of discriminating *Parthenium hysterophorus* using field spectroscopy data." International Journal of Remote Sensing **38**(20): 5608-5625.

Kganyago, M., J. Odindi, C. Adjorlolo and P. Mhangara (2018). "Evaluating the capability of Landsat 8 OLI and SPOT 6 for discriminating invasive alien species in the African Savanna landscape." International Journal of Applied Earth Observation and Geoinformation **67**: 10-19.

Kiala, Z., O. Mutanga, J. Odindi and K. Peerbhay (2019). "Feature Selection on Sentinel-2 Multispectral Imagery for Mapping a Landscape Infested by the Parthenium Weed." Remote Sensing **11**(16): 1892.

Kiala, Z., J. Odindi and O. Mutanga (2017). "Potential of interval partial least square regression in estimating leaf area index." South African Journal of Science **113**(9-10): 1-9.

Kiala, Z., J. Odindi, O. Mutanga and K. Peerbhay (2016). "Comparison of partial least squares and support vector regressions for predicting leaf area index on a tropical grassland using hyperspectral data." Journal of Applied Remote Sensing **10**(3): 036015-036015.

Kim, H.-O. and J.-M. Yeom (2014). "Effect of red-edge and texture features for object-based paddy rice crop classification using RapidEye multi-spectral satellite image data." International Journal of Remote Sensing **35**(19): 7046-7068.

Kira, K. and L. A. Rendell (1992). A practical approach to feature selection. Machine Learning Proceedings 1992, Elsevier: 249-256.

Kohavi, R. and G. John (1997). "Wrappers for feature subset selection." Artificial Intelligence **97**(1-2): 273-324.

Kokaly, R. F., D. G. Despain, R. N. Clark and K. E. Livo (2003). "Mapping vegetation in Yellowstone National Park using spectral feature analysis of AVIRIS data." Remote Sensing of Environment **84**(3): 437-456.

Kushwaha, V. B. and S. Maurya (2012). "Biological utilities of *Parthenium hysterophorus*." Journal of Applied and Natural Science **4**(1): 137-143.

Laba, M., F. Tsai, D. Ogurcak, S. Smith and M. E. Richmond (2005). "Field determination of optimal dates for the discrimination of invasive wetland plant species using derivative spectral analysis." Photogrammetric Engineering and Remote Sensing **71**(5): 603-611.

Lagrange, A., M. Fauvel and M. Grizonnet (2017). "Large-scale feature selection with Gaussian mixture models for the classification of high dimensional remote sensing images." IEEE Transactions on Computational Imaging **3**(2): 230-242.

Lalla, R., N. Mthimkhulu, Ian and Rushworth (2013). Update of action taken against *Parthenium hysterophorus* in KwaZulu-Natal (KZN), South Africa during 2013. International parthenium news. Tropical and Sub-Tropical Weed Research Unit, The University of Queensland, Australia., Dr Asad Shabbir.

Lantz, N. J. and J. Wang (2013). "Object-based classification of Worldview-2 imagery for mapping invasive common reed, *Phragmites australis*." Canadian Journal of Remote Sensing **39**(4): 328-340.

Lass, L. W., T. S. Prather, N. F. Glenn, K. T. Weber, J. T. Mundt and J. Pettingill (2005). "A review of remote sensing of invasive weeds and example of the early detection of spotted knapweed (*Centaurea maculosa*) and babysbreath (*Gypsophila paniculata*) with a hyperspectral sensor." Weed Science **53**(2): 242-251.

Laurin, G. V., N. Puletti, W. Hawthorne, V. Liesenberg, P. Corona, D. Papale, Q. Chen and R. Valentini (2016). "Discrimination of tropical forest types, dominant species, and mapping of functional guilds by hyperspectral and simulated multispectral Sentinel-2 data." Remote Sensing of Environment **176**: 163-176.

Lawrence, R., A. Bunn, S. Powell and M. Zambon (2004). "Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis." Remote Sensing of Environment **90**(3): 331-336.

Lawrence, R. L., S. D. Wood and R. L. Sheley (2006). "Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForest)." Remote Sensing of Environment **100**(3): 356-362.

Li, J., K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang and H. Liu (2017). "Feature selection: A data perspective." ACM Computing Surveys (CSUR) **50**(6): 94.

Li, J., J. Tang and H. Liu (2017). Reconstruction-based unsupervised feature selection: an embedded approach. Proceedings of the 26th International Joint Conference on Artificial Intelligence. IJCAI/AAAI.

Lin, X., F. Yang, L. Zhou, P. Yin, H. Kong, W. Xing, X. Lu, L. Jia, Q. Wang and G. Xu (2012). "A support vector machine-recursive feature elimination feature selection method based

on artificial contrast variables and mutual information." Journal of Chromatography B **910**: 149-155.

Liu, F. T., K. M. Ting and Z.-H. Zhou (2008). Isolation forest. 2008 Eighth IEEE International Conference on Data Mining, IEEE.

Liu, J., S. Ji and J. Ye (2009). Multi-task feature learning via efficient l 2, 1-norm minimization. Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence, AUAI Press.

Lou, W., X. Wang, F. Chen, Y. Chen, B. Jiang and H. Zhang (2014). "Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes." PLoS One **9**(1): e86703.

Louis, J., V. Debaecker, B. Pflug, M. Main-Knorn, J. Bieniarz, U. Mueller-Wilm, E. Cadau and F. Gascon (2016). Sentinel-2 Sen2Cor: L2A Processor for Users. Proceedings Living Planet Symposium 2016, Spacebooks Online.

Lu, H., J. Chen, K. Yan, Q. Jin, Y. Xue and Z. Gao (2017). "A hybrid feature selection algorithm for gene expression data classification." Neurocomputing **256**: 56-62.

Lunetta, R. S. and J. G. Lyon (2004). Remote sensing and GIS accuracy assessment, CRC press.

Luo, G. (2016). "A review of automatic selection methods for machine learning algorithms and hyper-parameter values." Network Modeling Analysis in Health Informatics and Bioinformatics **5**(1): 1-16.

Lusa, L. (2017). "Gradient boosting for high-dimensional prediction of rare events." Computational Statistics and Data Analysis **113**: 19-37.

Ma, L., T. Fu, T. Blaschke, M. Li, D. Tiede, Z. Zhou, X. Ma and D. Chen (2017). "Evaluation of feature selection methods for object-based land cover mapping of unmanned aerial vehicle imagery using random forest and support vector machine classifiers." ISPRS International Journal of Geo-Information **6**(2): 51.

MacLean, M. G. and R. G. Congalton (2013). "Applicability of Multi-date Land Cover Mapping using Landsat-5 TM Imagery in the Northeastern US." Photogrammetric Engineering and Remote Sensing **79**(4): 359-368.

Majasalmi, T. and M. Rautiainen (2016). "The potential of Sentinel-2 data for estimating biophysical variables in a boreal forest: a simulation study." Remote Sensing Letters **7**(5): 427-436.

Malahlela, O. E., M. A. Cho and O. Mutanga (2015). "Mapping the occurrence of Chromolaena odorata (L.) in subtropical forest gaps using environmental and remote sensing data." Biological Invasions **17**(7): 2027-2042.

Marée, R., L. Wehenkel and P. Geurts (2013). Extremely randomized trees and random subwindows for image classification, annotation, and retrieval. Decision Forests for Computer Vision and Medical Image Analysis, Springer: 125-141.

Martin, F.-M., J. Müllerová, L. Borgniet, F. Dommanget, V. Breton and A. Evette (2018). "Using Single-and Multi-Date UAV and Satellite Imagery to Accurately Monitor Invasive Knotweed Species." Remote Sensing **10**(10): 1662.

Masum, S., M. Hasanuzzaman and M. Ali (2013). "Threats of Parthenium hysterophorus on agro-ecosystems and its management: a review." International Journal of Agriculture and Crop Sciences **6**(11): 684.

Matongera, T. N., O. Mutanga, T. Dube and M. Sibanda (2017). "Detection and mapping the spatial distribution of bracken fern weeds using the Landsat 8 OLI new generation sensor." International Journal of Applied Earth Observation and Geoinformation **57**: 93-103.

McConnachie, A. J. (2015). "Host range and risk assessment of Zygogramma bicolorata, a defoliating agent released in South Africa for the biological control of Parthenium hysterophorus." Biocontrol Science and Technology **25**(9): 975-991.

McConnachie, A. J., L. W. Strathie, W. Mersie, L. Gebrehiwot, K. Zewdie, A. Abdurehim, B. Abrha, T. Araya, F. Asaregew, F. Assefa, R. Gebre-Tsadik, L. Nigatu, B. Tadesse and T. Tana (2011). "Current and potential geographical distribution of the invasive plant Parthenium hysterophorus (Asteraceae) in eastern and southern Africa." Weed Research **51**(1): 71-84.

Meddens, A. J., J. A. Hicke, L. A. Vierling and A. T. Hudak (2013). "Evaluating methods to detect bark beetle-caused tree mortality using single-date and multi-date Landsat imagery." Remote Sensing of Environment **132**: 49-58.

Miphokasap, P., K. Honda, C. Vaiphasa, M. Souris and M. Nagai (2012). "Estimating canopy nitrogen concentration in sugarcane using field imaging spectroscopy." Remote Sensing **4**(6): 1651-1670.

Moosmann, F., E. Nowak and F. Jurie (2008). "Randomized clustering forests for image classification." IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(9): 1632-1646.

Morandeira, N., F. Grings, C. Facchinetti and P. Kandus (2016). "Mapping plant functional types in floodplain wetlands: an analysis of C-band polarimetric SAR data from RADARSAT-2." Remote Sensing **8**(3): 174.

Mullerova, J., P. Pysek, V. Jarosik and J. Pergl (2005). "Aerial photographs as a tool for assessing the regional dynamics of the invasive plant species *Heracleum mantegazzianum*." Journal of Applied Ecology **42**(6): 1042-1053.

Municipality, M. L. (2002). "Integrated development plan." Prepared by the Councillors and Officials of the Msunduzi Municipality.

Mutanga, O. and A. K. Skidmore (2004). "Narrow band vegetation indices overcome the saturation problem in biomass estimation." International journal of remote sensing **25**(19): 3999-4014.

Mutanga, O. and A. K. Skidmore (2007). "Red edge shift and biochemical content in grass canopies." ISPRS Journal of Photogrammetry and Remote Sensing **62**(1): 34-42.

Myrans, J., Z. Kapelan and R. Everson (2016). "Automated detection of faults in wastewater pipes from CCTV footage by using random forests." Procedia Engineering **154**: 36-41.

Nembrini, S., I. R. König and M. N. Wright (2018). "The revival of the Gini importance?" Bioinformatics **34**(21): 3711-3718.

NGI (2008). Pietermaritzburg (Air Photo). Mowbray, Cape Town, National Geo-spatial Information.

Nie, Z., K. K. Y. Chan and B. Xu (2019). "Preliminary Evaluation of the Consistency of Landsat 8 and Sentinel-2 Time Series Products in An Urban Area—An Example in Beijing, China." Remote Sensing **11**(24): 2957.

Nitze, I., U. Schulthess and H. Asche (2012). "Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification." Proc. of the 4th GEOBIA: 7-9.

Norman, N. and G. Whitfield (2006). Geological journeys: A traveller's guide to South Africa's rocks and landforms, Struik.

Novack, T., T. Esch, H. Kux and U. Stilla (2011). "Machine learning comparison between WorldView-2 and QuickBird-2-simulated imagery regarding object-based urban land cover classification." Remote Sensing **3**(10): 2263-2282.

Olson, R. S. and J. H. Moore (2016). TPOT: A tree-based pipeline optimization tool for automating machine learning. Workshop on Automatic Machine Learning.

Olson, R. S., R. J. Urbanowicz, P. C. Andrews, N. A. Lavender, L. C. Kidd and J. H. Moore (2016). "Automating biomedical data science through tree-based pipeline optimization." arXiv preprint arXiv:1601.07925.

Oon, A., H. Z. Mohd Shafri, A. M. Lechner and B. Azhar (2019). "Discriminating between large-scale oil palm plantations and smallholdings on tropical peatlands using vegetation indices and supervised classification of LANDSAT-8." International Journal of Remote Sensing **40**(19): 1-17.

Pal, M. (2005). "Random forest classifier for remote sensing classification." International Journal of Remote Sensing **26**(1): 217-222.

Pal, M. and G. M. Foody (2010). "Feature selection for classification of hyperspectral data by SVM." IEEE Transactions on Geoscience and Remote Sensing **48**(5): 2297-2307.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg (2011). "Scikit-learn: Machine learning in Python." Journal of Machine Learning Research **12**(Oct): 2825-2830.

Peerbhay, K., O. Mutanga, R. Lottering and R. Ismail (2016). "*Mapping Solanum mauritianum* plant invasions using WorldView-2 imagery and unsupervised random forests." Remote Sensing of Environment **182**: 39-48.

Peerbhay, K. Y., O. Mutanga and R. Ismail (2015). "Random Forests Unsupervised Classification: The Detection and Mapping of *Solanum mauritianum* Infestations in Plantation Forestry Using Hyperspectral Data." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **8**(6): 3107-3122.

Peralta, B. and A. Soto (2014). "Embedded local feature selection within mixture of experts." Information Sciences **269**: 176-187.

Pesaresi, M., C. Corbane, A. Julea, A. Florczyk, V. Syrris and P. Soille (2016). "Assessment of the Added-Value of Sentinel-2 for Detecting Built-up Areas." Remote Sensing **8**(4): 299.

Picos, J., L. Alonso, G. Bastos and J. Armesto (2019). "Event-Based Integrated Assessment of Environmental Variables and Wildfire Severity through Sentinel-2 Data." Forests **10**(11): 1021.

Qing, C., X. Xiao and W. Jiao (2018). Characterizing spring phenology of snow-covered forests by vegetation indices, primary productivity and solar-induced chlorophyll fluorescence. AGU Fall Meeting Abstracts.

Raczko, E. and B. Zagajewski (2017). "Comparison of support vector machine, random forest and neural network classifiers for tree species classification on airborne hyperspectral APEX images." European Journal of Remote Sensing **50**(1): 144-154.

Ramoelo, A., M. Cho, R. Mathieu and A. K. Skidmore (2015). "Potential of Sentinel-2 spectral configuration to assess rangeland quality." Journal of Applied Remote Sensing **9**: 11.

Resasco, J., A. N. Hale, M. C. Henry and D. L. Gorchov (2007). "Detecting an invasive shrub in a deciduous forest understory using late-fall Landsat sensor imagery." International Journal of Remote Sensing **28**(16): 3739-3745.

Robnik-Šikonja, M. and I. Kononenko (2003). "Theoretical and empirical analysis of ReliefF and RReliefF." Machine Learning **53**(1-2): 23-69.

Rodriguez-Galiano, V., M. Chica-Olmo, F. Abarca-Hernandez, P. M. Atkinson and C. Jeganathan (2012). "Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture." Remote Sensing of Environment **121**: 93-107.

Rodriguez-Galiano, V. F., B. Ghimire, J. Rogan, M. Chica-Olmo and J. P. Rigol-Sanchez (2012). "An assessment of the effectiveness of a random forest classifier for land cover classification." ISPRS Journal of Photogrammetry and Remote Sensing **67**: 93-104.

Rogan, J. and D. Chen (2004). "Remote sensing technology for mapping and monitoring land cover and land use change." Progress in Planning **61**(4): 301-325.

Roteta, E., A. Bastarrika, M. Padilla, T. Storm and E. Chuvieco (2019). "Development of a Sentinel-2 burned area algorithm: Generation of a small fire database for sub-Saharan Africa." Remote Sensing of Environment **222**: 1-17.

Rouhi, A. and H. Nezamabadi-pour (2017). A hybrid feature selection approach based on ensemble method for high-dimensional data. 2017 2nd Conference on Swarm Intelligence and Evolutionary Computation (CSIEC), IEEE.

Royimani, L., O. Mutanga, J. Odindi, K. S. Zolo, M. Sibanda and T. Dube (2018). "Distribution of Parthenium hysterophoru L. with variation in rainfall using multi-year SPOT data and random forest classification." Remote Sensing Applications: Society and Environment **13**: 215-223.

Saeys, Y., I. Inza and P. Larrañaga (2007). "A review of feature selection techniques in bioinformatics." Bioinformatics **23**(19): 2507-2517.

Salehi, S., C. Mielke, C. B. Pedersen and S. D. Olsen (2019). "Comparison of ASTER and Sentinel-2 spaceborne datasets for geological mapping: a case study from North-East Greenland." Geological Survey of Denmark & Greenland Bulletin **43**: 1-6.

Samat, A., C. Persello, S. Liu, E. Li, Z. Miao and J. Abuduwaili (2018). "Classification of VHR multispectral images using extratrees and maximally stable extremal region-guided morphological profile." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **11**(9): 3179-3195.

Sankaran, M., J. Ratnam and N. Hanan (2008). "Woody cover in African savannas: the role of resources, fire and herbivory." Global Ecology and Biogeography **17**(2): 236-245.

Sencaki, D. B., D. J. Muhammad, L. Sumargana and L. Gandharum (2018). Peatland Delineation Using Remote Sensing Data in Sumatera Island. 2018 IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS), IEEE.

Sepúlveda, M., H. E. Bown, M. D. Miranda and B. Fernández (2018). "Impact of rainfall frequency and intensity on inter-and intra-annual satellite-derived EVI vegetation productivity of an Acacia caven shrubland community in Central Chile." Plant Ecology **219**(10): 1209-1223.

Shafri, H. Z. and N. Hamdan (2009). "Hyperspectral imagery for mapping disease infection in oil palm plantation using vegetation indices and red edge techniques." American Journal of Applied Sciences **6**(6): 1031.

Shang, W., H. Huang, H. Zhu, Y. Lin, Y. Qu and Z. Wang (2007). "A novel feature selection algorithm for text categorization." Expert Systems with Applications **33**(1): 1-5.

Shoko, C., O. Mutanga, T. Dube and R. Slotow (2018). "Characterizing the spatio-temporal variations of C3 and C4 dominated grasslands aboveground biomass in the Drakensberg, South Africa." International Journal of Applied Earth Observation and Geoinformation **68**: 51-60.

Sibanda, M., O. Mutanga and M. Rouget (2015). "Examining the potential of Sentinel-2 MSI spectral resolution in quantifying above ground biomass across different fertilizer treatments." Isprs Journal of Photogrammetry and Remote Sensing **110**: 55-65.

Smola, A. and V. Vapnik (1997). "Support vector regression machines." Advances in Neural Information Processing Systems **9**: 155-161.

Snoek, J., H. Larochelle and R. P. Adams (2012). Practical bayesian optimization of machine learning algorithms. Advances in Neural Information Processing Systems.

Sohn, A., R. S. Olson and J. H. Moore (2017). Toward the automated analysis of complex diseases in genome-wide association studies using genetic programming. Proceedings of the Genetic and Evolutionary Computation Conference, ACM.

Strathie, L. W., A. J. McConnachie and E. Retief (2011). "Initiation of biological control against *Parthenium hysterophorus* L. (*Asteraceae*) in South Africa." African Entomology **19**(2): 378-392.

Sun, G., H. Huang, Q. Weng, A. Zhang, X. Jia, J. Ren, L. Sun and X. Chen (2019). "Combinational shadow index for building shadow extraction in urban areas from sentinel-2a msi imagery." International Journal of Applied Earth Observation and Geoinformation **78**: 53-65.

Swati, G., S. Haldar, A. Ganguly and P. K. Chatterjee (2013). "Review on Parthenium hysterphorus as a potential energy source." Renewable and Sustainable Energy Reviews **20**: 420-429.

Talavera, L. (2005). An evaluation of filter and wrapper methods for feature selection in categorical clustering. International Symposium on Intelligent Data Analysis, Springer.

Taşkın, G., H. Kaya and L. Bruzzone (2017). "Feature selection based on high dimensional model representation for hyperspectral images." IEEE Transactions on Image Processing **26**(6): 2918-2928.

Terblanche, C., I. Nänni, H. Kaplan, L. W. Strathie, A. J. McConnachie and J. Goodall (2016). "An approach to the development of a national strategy for controlling invasive alien plant species: The case of *Parthenium hysterophorus* in South Africa." Bothalia **46**(1): 11 pages.

Thejas, G., S. R. Joshi, S. Iyengar, N. Sunitha and P. Badrinath (2019). "Mini-Batch Normalized Mutual Information: A Hybrid Feature Selection Method." IEEE Access 7: 116875-116885.

Thornton, C., F. Hutter, H. H. Hoos and K. Leyton-Brown (2013). Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.

Tottrup, C. (2004). "Improving tropical forest mapping using multi-date Landsat TM data and pre-classification image smoothing." International Journal of Remote sensing **25**(4): 717-730.

Truter, M., A. Dippenaar-Schoeman, A. v. d. Berg, I. Millar, M. v. d. Merwe, A. d. Klerk, P. Marais, E. v. Niekerk and L. Besaans (2014). Integrated management and soilborne plant diseases. Plant Protection News. Queenswood, South Africa, ARC-Plant Protection Research Institute.

Tsai, F., E.-K. Lin and H.-H. Wang (2005). Detecting invasive plant species using hyperspectral satellite imagery. Proceedings. 2005 IEEE International Geoscience and Remote Sensing Symposium, 2005. IGARSS'05., IEEE.

Tzelidi, D., S. Stagakis, Z. Mitraka and N. Chrysoulakis (2019). "Detailed urban surface characterization using spectra from enhanced spatial resolution Sentinel-2 imagery and a hierarchical multiple endmember spectral mixture analysis approach." Journal of Applied Remote Sensing **13**(1): 016514.

Underwood, E., S. Ustin and D. DiPietro (2003). "Mapping nonnative plants using hyperspectral imagery." Remote Sensing of Environment **86**(2): 150-161.

Ustin, S. L., D. DiPietro, K. Olmstead, E. Underwood and G. J. Scheer (2002). Hyperspectral Remote Sensing for Invasive Species Detection and Mapping. Geoscience and Remote Sensing Symposium, 2002. IGARSS'02. 2002 IEEE International

Vaiphasa, C., A. Skidmore, W. de Boer and T. Vaiphasa (2007). "A hyperspectral band selector for plant species discrimination." ISPRS Journal of Photogrammetry and Remote Sensing **62**: 225-235.

van der Meer, F. D., H. M. A. van der Werff and F. J. A. van Ruitenbeek (2014). "Potential of ESA's Sentinel-2 for geological applications." Remote Sensing of Environment **148**: 124-133.

Van der Walt, C. M. and E. Barnard (2006). "Data characteristics that determine classifier performance."

van der Werff, H. and F. van der Meer (2015). "Sentinel-2 for Mapping Iron Absorption Feature Parameters." Remote Sensing **7**(10): 12635-12653.

Venkatesh, B. and J. Anuradha (2019). A Hybrid Feature Selection Approach for Handling a High-Dimensional Data. Innovations in Computer Science and Engineering, Springer: 365-373.

Vergara, J. R. and P. A. Estévez (2014). "A review of feature selection methods based on mutual information." Neural Computing and Applications **24**(1): 175-186.

Vincini, M., F. Calegari and R. Casa (2016). "Sensitivity of leaf chlorophyll empirical estimators obtained at Sentinel-2 spectral resolution for different canopy structures." Precision Agriculture **17**(3): 313-331.

Vuolo, F., M. Neuwirth, M. Immitzer, C. Atzberger and W.-T. Ng (2018). "How much does multi-temporal Sentinel-2 data improve crop type classification?" International Journal of Applied Earth Observation and Geoinformation **72**: 122-130.

Wang, Z., J. Liu, J. Li and D. Zhang (2018). "Multi-Spectral Water Index (MuWI): A Native 10-m Multi-Spectral Water Index for Accurate Water Mapping on Sentinel-2." Remote Sensing **10**(10): 1643.

Waser, L. T., M. Küchler, K. Jütte and T. Stampfer (2014). "Evaluating the potential of WorldView-2 data to classify tree species and different levels of ash mortality." Remote Sensing **6**(5): 4515-4545.

Weisenthal, S. J., C. Quill, S. Farooq, H. Kautz and M. S. Zand (2018). "Predicting acute kidney injury at hospital re-entry using high-dimensional electronic health record data." PloS one **13**(11): e0204920.

Wright, S. (1965). "The interpretation of population structure by F-statistics with special regard to systems of mating." Evolution **19**(3): 395-420.

Xie, J. and C. Wang (2011). "Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases." Expert Systems with Applications **38**(5): 5809-5815.

Xie, L., G. Y. Li, L. Peng, Q. C. Chen, Y. L. Tan and M. Xiao (2017). "Band selection algorithm based on information entropy for hyperspectral image classification." Journal of Applied Remote Sensing **11**: 17.

Yu, Q., P. Gong, N. Clinton, G. Biging, M. Kelly and D. Schirokauer (2006). "Object-based detailed vegetation classification with airborne high spatial resolution remote sensing imagery." Photogrammetric Engineering and Remote Sensing **72**(7): 799-811.

Zhang, H. (2004). "The optimality of naive Bayes." AA **1**(2): 3.

Zhang, X., M. A. Friedl, C. B. Schaaf, A. H. Strahler, J. C. Hodges, F. Gao, B. C. Reed and A. Huete (2003). "Monitoring vegetation phenology using MODIS." Remote Sensing of Environment **84**(3): 471-475.

Zheng, X., Y. Yuan and X. Lu (2017). "Dimensionality reduction by spatial–spectral preservation in selected bands." IEEE Transactions on Geoscience and Remote Sensing **55**(9): 5185-5197.

Zhou, Y., R. Zhang, S. Wang and F. Wang (2018). "Feature selection method based on high-resolution remote sensing images and the effect of sensitive features on classification accuracy." Sensors **18**(7): 2013.

Zhu, Z., Y.-S. Ong and M. Dash (2007). "Wrapper–filter feature selection algorithm using a memetic framework." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **37**(1): 70-76.