Understanding Striga occurrence and risk under changing climatic conditions across different agroecological farming systems at local and regional scales

Bester Tawona Mudereri

A thesis submitted to the Department of Earth Sciences, Faculty of Natural Sciences at the University of the Western Cape in fulfillment of the academic requirements for the degree of Doctor of Philosophy in Environmental and Water Science

South Africa

October 2020

Supervisor : Prof. Timothy Dube

UNIVERSITY *of the*
WESTERN CAPE

ABSTRACT

The invasion by Striga in most cereal crop fields in Africa has posed an acute threat to food security and socioeconomic integrity. Consequently, numerous technological and research developments have been made to minimize and even control the Striga impacts on crop production. So far, efforts to control Striga have primarily focused on the manipulation of the genetics of the host crops, as well as understanding the phenological and physiological traits, along with the chemical composition of the weed. These initiatives have immensely contributed to the management of Striga across the continents. However, on-farm Striga control technologies require spatial explicit locational information on farms experiencing Striga occurrence and potential risk. This information affords precise and accurate intervention mechanisms and allows for the prescription of site-specific and befitting control approaches. Unfortunately, the requisite baseline information on Striga occurrence, spatial configuration, infestation extent, and intensity remain rudimentary in sub-Saharan Africa. This study, therefore, aimed to examine Striga occurrence and the potential farming areas at risk within different agroecological regions and varying climatic scenarios in Kenya and Zimbabwe. To achieve this aim, the relatively new generation remotely sensed data coupled with biophysical variables, Striga occurrence, and cropping systems data were used. Specifically, the study sought to establish operational spatial methodologies that can help understand and empirically determine the prospective risk posed by two of the most economically detrimental Striga species in Africa (i.e. *Striga hermonthica and S. asiatica*) in agroecological farming systems. In addition, the likely impacts of climate change on Striga distribution and future spread by integrating climatic and cropland data were also examined. Further, different machine learning algorithms were used for data analysis at different mapping scales. Results from this study demonstrated that Striga's occurrence within agroecological systems can be characterized at reasonable accuracy, using relatively new generation sensor datasets across various scales of monitoring i.e. plot, field, and landscape. Comparatively, in-situ hyperspectral measurements and Sentinel-2 satellite data coupled with machine learning and subpixel classification approaches surpassed the traditional broadband sensor data in the detection and understanding of the spatial dispersion of the two Striga weed species across different agroecological farming systems. Further, the Striga flowering period was established as the most optimal period for its detection and monitoring. It was also observed that as the climatic conditions continue to change i.e. the

atmospheric CO_2 and the temperature increase, the suitable area for Striga propagation will also increase, making more farming areas to be susceptible to a higher risk of invasion. In particular, the use of the projected climate change scenarios showed that by the year 2050, the Striga suitable area propagation will increase and spread into new areas by approximately 0.73%. Also, it was established that the ecological niche and habitat suitability assessments using multi-source remotely sensed data are fundamental in characterizing and monitoring *S. asiatica* occurrence and risk areas. Therefore, immediate mitigation and adaptive actions such as awareness and advocacy for the adoption of Striga control methods in the current and the future risk areas is critical to contain and manage the spread and intensity of Striga under changing climatic conditions. Overall, the findings of this study underscore the relevance of using multi-source data and machine learning algorithms for Striga weed detection and monitoring across different agroecological farming systems.

Keywords: climate variability; food security; integrated modeling; invasive weeds; mapping scale; precision agriculture; remote sensing

PREFACE

The present study was undertaken to determine and assess Striga weed (i.e. *Striga hermonthica and S. asiatica*) occurrence and potential risk imposed on the agroecological farming systems. Further, the study investigated the likely impacts that climate change will have on Striga distribution and future spread by integrating spatial explicit data with climatic data. The approach used in this study was a succession of independent but related research papers that form different chapters of the thesis. The thesis comprises six chapters in total, with four chapters conceptualized as stand-alone research articles that address each of the objectives listed in section 1.5.

Four articles making up chapters 2 to 5 have already been published as research articles or as a conference paper i.e. chapter 2 (International Journal of Applied Earth Observation and Geoinformation); chapter 3 (ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences); chapter 4 (Science of the Total Environment) and chapter 5 (GIScience and Remote sensing). Each chapter can be read independently from the rest of the thesis but draws conclusions linked and relevant to the work as a whole. Although the document generally conforms to the University of the Western Cape style, some degree of repetition has been inevitable, given the shared thread of the papers.

- Chapter 1 is the general introduction and contextualization of the study.

- Chapter 2 examines the potential use of in-situ hyperspectral data at plot-scale in detecting *S. hermonthica* under field conditions. The hyperspectral data were also resampled to the Sentinel-2 band configurations to test the applicability of the multispectral Sentinel-2 sensor data in detecting and characterizing Striga.

- Chapter 3 follows up the work demonstrated in chapter 2 by upscaling the mapping scale to field level by comparing and evaluating the strength of using a single image of the Sentinel-2 multispectral sensor (10 m pixel size) and PlanetScope nanosatellite (3 m pixel size) to characterize *S. hermonthica*.

- Chapter 4, determines if using subpixel analysis i.e. multiple endmember spectral mixture analysis (MESMA) in an exclusive cropland class within a Sentinel-2 satellite

image, improves the detection and estimation of the level of infestation of *S. hermonthica* at both field and landscape scales. The chapter also tests the utility of the Google Earth Engine (GEE) in cropland mapping using Sentinel-2 derived bands, red-edge bands in integration with their respective vegetation indices and red-edge derived indices.

- Chapter 5 tests a wider landscape-scale i.e. country-level mapping of the *S. asiatica* using an integrated machine learning and ecological niche modeling approach using multi-source remotely sensed data (i.e. cropping system, edaphic, land surface temperature, and terrain) under current and future climate scenarios.

- Chapter 6 provides a synthesis of the research work.

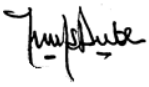
DECLARATION 1

The data described in this thesis were collected in the Rongo sub-county in Kenya, Masvingo, and Midlands provinces in Zimbabwe from December 2017 until December 2019. The study and fieldwork were carried out as part of the 'push-pull for sub-Saharan Africa' project with funding from the Biovision Foundation for ecological development (Switzerland) grant to the International Center of Insect Physiology and Ecology (*icipe*), Kenya. The candidate was supported by a German Academic Exchange Service (DAAD), In-Region Postgraduate Scholarship, through the Africa Regional Postgraduate Program in Insect Science (ARPPIS), while registered with the Department of Earth Sciences, Faculty of Natural Sciences, University of the Western Cape, under the supervision of Professor Timothy Dube.

This thesis, submitted for the degree of Doctor of Philosophy in the Department of Earth Sciences, Faculty of Natural Sciences, University of the Western Cape, represents original work by the candidate and has not otherwise been submitted in any form for any degree or diploma to any University. Where use has been made of the work of others, it is duly acknowledged in the text.

……...
Bester Tawona Mudereri
October 2020

We certify that the above statement is correct and as the candidate's supervisors we have approved this thesis for submission.



……..
Professor Timothy Dube
Supervisor
October 2020

…….
Dr. Elfatih. M. Abdel-Rahman
Co-supervisor
October 2020

……..
Dr. Saliou Niassy
Co-supervisor
October 2020

DECLARATION 2 – PLAGIARISM

I, Bester Tawona Mudereri, declare that:

1. The research reported in this thesis, except where otherwise indicated, is my original research.

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 other persons' writing unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

    a. Their words have been re-written, but the general information attributed to them has been referenced.

    b. 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 the References sections.

Signed: ……

Bester Tawona Mudereri

October 2020

DECLARATION 3 – PUBLICATIONS AND MANUSCRIPTS

Details of publications and manuscripts that form part and/ or include research presented in this thesis.

Publication 1:

Mudereri, B.T., Dube, T., Niassy, S., Kimathi, E., Landmann, T., Khan, Z., Abdel-Rahman, E.M., 2020. *Is it possible to discern Striga weed (Striga hermonthica) infestation levels in maize agroecological systems using in-situ spectroscopy?* International Journal of Applied Earth Observation and Geoinformation 85: 102008. https://doi.org/10.1016/j.jag.2019.102008

Publication 2:

Mudereri, B.T., Dube, T., Adel-Rahman, E.M., Niassy, S., Kimathi, E., Khan, Z., Landmann, T., 2019. *A comparative analysis of PlanetScope and Sentinel-2 space-borne sensors in mapping Striga weed using Guided Regularized Random Forest classification ensemble.* ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-2/W13, 701–708. https://doi.org/10.5194/isprs-archives-XLII-2-W13-701-2019

Publication 3:

Mudereri, B.T., Adel-Rahman, E.M., Dube, T., Niassy, S., Khan, Z., Tonnang H.E.Z., Landmann, T., (2020). *A two-step approach for detecting Striga in a complex agroecological system using Sentinel-2 data*. Science of the Total Environment. https://doi.org/10.1016/j.scitotenv.2020.143151

Publication 4:

Mudereri, B. T., Abdel-Rahman, E.M., Dube, T., Landmann, T., Khan, Z. R., Kimathi, E., Owino, R., Niassy, S. (2020). *Multi-source spatial data-based invasion risk modeling of Striga (Striga asiatica) in Zimbabwe*. GIScience & Remote Sensing 57: 553–571. https://doi.org/10.1080/15481603.2020.1744250.

All the conceptual and experimental work, analysis of data, and preparation of the above publications and manuscripts were accomplished by the candidate, Bester Tawona Mudereri, under the supervision of Professor Timothy Dube, Dr. Elfatih. M. Abdel-Rahman and Dr. Tobias Landmann. Dr. Saliou Niassy and Professor Zeyaur Khan sourced the funding in addition to giving logistical support for field data collection

Signed:
Bester Tawona Mudereri
October 2020

ACKNOWLEDGMENTS

I give thanks and praise to the Almighty, 'the author and finisher of my faith' for giving me insurance, health, strength, hope, and sustenance to do this work.

In a doctoral study, one is indebted to so many sources of assistance, hence it is practically impossible to list them all. However, at the risk of omission, I would like to particularly thank the following individuals and institutions for their support.

My research project was conducted with the collaboration of the International Center of Insect Physiology and Ecology (*icipe*) and the German Academic Exchange Program (DAAD) through the In-Region Africa Regional Postgraduate Program in Insect Science (ARPPIS). I am, therefore, sincerely grateful to *icipe* and the entire organizational fraternity, for hosting my research stay in Kenya and providing state-of-the-art facilities that facilitated top-notch research. I will forever be indebted to the taxpayers in Germany through their donation in the form of a scholarship that catered for all my living, health, and research expenses for the entire study period. I am also grateful to the Biovision Foundation for ecological development (Switzerland) for funding the 'push-pull for sub-Saharan Africa' project, which co-funded my field research work. I am also thankful to the University of the Western Cape (UWC) for granting me the opportunity to pursue my doctoral studies and the provision of an efficient and flawless registration during my entire study period.

I would like to give a special thank you to all my 'supervisors' whom I would like to call 'mentors' i.e. Professor Timothy Dube (UWC), Dr. Elfatih. M. Abdel-Rahman (*icipe*), Dr. Tobias Landmann (RSS-Germany), Dr. Saliou Niassy (*icipe*), and Professor Zeyaur. R. Khan (*icipe*) for their steadfast leadership and dedicated support throughout the progression of this mentorship. I can now self-confidently say, your shrewdness molded a researcher, a critical and independent thinker, and an innovator with sound scientific acumen. Again, my sincere gratitude to the Head of the *icipe* Data Management, Modeling, and Geoinformation (DMMG) unit i.e. Dr. Henri E. Z. Tonnang and all the unit members, for the inspiration, encouragement, and support through the various discussions and seminars that aided to expand the scope and quality of this work.

Many thanks are also due to the *icipe* Technology Transfer Unit (TTU crew) and administrators

for facilitating and organizing all my data collection trips and engagements with the local communities and local collaborators while I rode on their wings and names. I also would like to say nimeshukuru sana (Kiswahili for saying many thanks) to my two office mates and idea punching bags i.e. Emily Kimathi (mama Zi) and Gladys Mosomtai (Jebi) for permitting me to toss and pitch at you, scientific ideas that you would often frustrate, crush but ultimately shape and mold, which resulted in this lifetime piece of work. I wish you well in your careers as well.

To all my ARPPIS 2017 friends, thank you for all the fun, jokes, teasing, and above all enthusiasm on all our platforms that inculcated in me hope, that it can be done. We all meet for a purpose and I am grateful that we met at the right time and you were doors to so many of my life lessons and opportunities, cheers, and kindly do well in your careers and life after *icipe*. Much gratitude is extended to the entire *icipe* Scholars Association (IScA) fraternity for 'Science club' and resolute comradeship. It would not be enough if I do not thank my 'crazy' IScA executive committee members for the 2018–2019 year i.e. Yosef Hamba Tola, Gladys Mosomtai, Pascal Ayelo, Iman Brema Hassaballa, Naomi Riithi, and Getachew Eticha Bokore. All I can say is 'We came, We saw and We conquered'.

My very sincere gratitude goes to my family: My father (who called me Doc from the day I was accepted at *icipe* and my mother (for your prayers), brothers (for taking care of business while I was away), sisters (for continuously asking me about the day of graduation) and all relatives for the love and committed support.

Finally, and most importantly, I am grateful to my provocation, my stimulus, my incentive, my biggest fan, my motivator, and my cheerleader, my wife Caroline (Mkanyaress). I am sincerely grateful for your extraordinary altruism, the immense sacrifice, your serenity, your fortitude, and bearing the burden of looking after our kids while I spent my entire time working. It extremely aided me to focus on my Ph.D. work. I owe it all to you, thus, my triumphs, as well as the kids' achievements, are your own brilliance and distinction.

To Tawonashe (Tawo), Nenyasha (Nene), and Natswainashe (Nati) in the future you will understand.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

General introduction

Photo: courtesy of Bester Tawona Mudereri 2017

1.1 Agriculture in the African context

Most African households live in rural areas (~ 645 million people) and depend on agricultural produce for their food and source of income (Worldbank, 2018). Cereal crops mainly sorghum, wheat, millet, rice, and primarily maize constitute some of the most important staple food crops on the African continent that secure nutrients need to about 1.2 billion people (~ 13% of the global population) (Prăvălie et al., 2019). Of particular concern is the cereal crop productivity, which has been on the decline, increasing the undernourished population from 44 million to reach 218 million in the last decade in sub-Saharan Africa (OECD/FAO, 2016). Hunger is thus, on the rise in almost all African countries, making it one of the regions with the highest prevalence of undernourishment, at almost 20%, underscoring the immense challenge of achieving the Zero Hunger target by 2030 (FAO, IFAD, UNICEF, WFP, WHO, 2019). Several agroecological farming systems aimed at improving food production have been adopted and used extensively in Africa such as organic farming, diversified crop rotations, biological pest control, extensive agro-pastoral systems, and agroforestry, among others, with limited success (FAO, IFAD, UNICEF, WFP, WHO, 2018). This is greatly attributed to the diverse climatic conditions experienced in Africa, which are largely defined by the variations in rainfall patterns, altitude, temperatures, soil characteristics, emerging crop pests, and diseases that impede crop production and adaptability across different agroecological systems (FAO, IFAD, UNICEF, WFP, WHO, 2017). It is therefore prudent to identify tailor-made and across-the-board techniques that can increase crop production in sub-Saharan Africa and help to advance economic growth, combat poverty, and halt environmental degradation despite the various crop production and socioeconomic restraints (Khan et al., 2014).

1.2 Cereal crop production restraints in sub-Saharan Africa

Recent agricultural statistics point towards a steady improvement in total cereal production in sub-Saharan Africa, since 2005 (AGRA, 2017). However, the study conducted by FAO in the year 2019 shows that between the years 2017 and 2019, sub-Saharan Africa experienced approximately 5.3% decline in total cereal production, while Kenya and Zimbabwe went down by 20.4% and 52.7%, respectively. These changes have been attributed to climate change and variability induced droughts, emerging invasive pests, and diseases such as stem borers, fall armyworm, maize lethal necrosis, and invasive weeds like Striga (FAO, IFAD, UNICEF, WFP, WHO, 2018; Khan et al.,

2014). While the impacts of climate change and variability on food production are manifest across the various continents, they are irregularly distributed geographically and the magnitude of the impacts varies from place to place (Nhamo et al., 2019). So far, research conducted on the impacts of drought, climate change, and variability on the agricultural sector showed a general decline in crop productivity (Abegunde et al., 2019; Moore et al., 2017; Nhamo et al., 2019; Sultan et al., 2019). However, to this point, very few studies have widely considered these inherent variations that are largely influenced by weed occurrence and spread, in particular, the most economically important Striga weed species.

Striga weeds, which are obligate parasites that attach themselves to the roots of the cereal crops, after germination, out-compete their hosts for space, nutrients, water, and cause devastation in the agroecological farming systems, together with insects, birds, and plant diseases (Atera et al., 2013; Scholes and Press, 2008). The genus Striga (*Orobanchaceae*) is composed of more than 20 species of parasitic weeds of global economic importance, causing yearly socioeconomic losses of over US$ 1 billion in Africa (Ejeta and Gressel, 2007; Spallek et al., 2013). In some areas in Africa, the scourge of Striga has reached epidemic magnitudes, affecting mainly poor small-scale farmers who comprise ~ 80% of all the farmers in sub-Saharan Africa (Makurira, 2010).

Specifically, *Striga hermonthica* and *S. asiatica* are the most prevalent among the Striga species predating on all the major and most economically important cereal crops (Khan et al., 2014). Approximately, 80% of all Striga species are endemic to Africa (Rodenburg et al., 2010). In total, about 50 African countries are reported to have at least one species of Striga, with *S. hermonthica* being found in at least 32 countries and *S. asiatica* occurring in at least 44 of the 50 countries (Rodenburg et al., 2016). Consequently, these two species successfully thrive in climatic conditions, which are also favorable to most of the economically important cereal crops, prompting the need to understand their distribution, infestation, severity, and more importantly possible areas at risk (Ejeta and Gressel, 2007). To reverse these trends, there is a need for targeted investments in research that warrant support to small-holder farmers in Africa to battle the scourge of Striga with the necessary capacity, effective, sustainable, and affordable solutions.

1.3 Striga detection and control mechanisms

Several methods to control crop losses caused by Striga have been developed with partial or limited

success (Oswald, 2005). The majority of such methods mainly targeted manipulating the physical, biological, genetic, chemical, ecological, and phenological characteristics of the weed (Badu-Apraku et al., 2015; Ejeta & Butler, 1993). The most commonly used control strategies include practices such as fuming ethylene to initiate 'suicide germination' (Samejima et al., 2016), hand weeding, crop rotations (Oswald & Ransom, 2001). Planting Striga resistant cultivars (Ransom et al., 2012), trap cropping, improving soil fertility and the use of *Desmodium* species and *Napier* grass in the 'push-pull technology' has also been widely adopted (Khan et al., 2008). For the 'push-pull' strategy, specifically chosen companion plants are grown in between and around the main crop e.g. maize. These companion plants (*Desmodium* species and *Napier* grass) release semio-chemicals that (i) repel insect pests from the main crop, using an intercrop which is the 'push' component; and (ii) attract insect pests away from the main crop using a trap crop, which is the 'pull' component (Khan et al., 2008). However, most of these efforts and methods have been hindered by socioeconomic or environmental mismatches and improper prioritization of intervention areas.

Striga infestation is irregular in both space and time, making field assessments and surveys inadequate to systematically and explicitly understand the impacts caused by the weed at a landscape or regional scale. Spatial explicit information on inter- and intra-field variability of Striga weeds within the agroecological systems is imperative in developing site-specific adaptive management, precise resource allocation, labor efficiency for Striga control, and drafting of management decisions (Houborg and McCabe, 2018a). There is, therefore, a need to develop innovative spatial explicit methodologies that complement the established methodological advancements in Striga control mechanisms and its impact on crop production.

Remote sensing provides an invaluable toolset and operational framework for timely monitoring and modeling of crop weeds occurrence and spread (Phalke and Özdoğan, 2018). Thus, it presents unique opportunities that could aid in precision intervention and determining areas to further advance meaningful and localized Striga control interventions. The recent technological advancements in sensor design and sensing characteristics, as well as the recent trends in remote sensing and data analytics, provide new avenues and opportunities for continuous and near-real-time assessment of crop health, crop phenological development, and weed predictions in the face of climate change and variability (Mutanga et al., 2017). This is necessitated by relatively new

sensors that provide data at near-real-time temporal resolutions, with high-to-medium spatial resolutions and with new and strategically positioned bands (Shoko and Mutanga, 2017). This, therefore, permits for the plot-, farm- or landscape-scale detection and assessment of crop weeds such as Striga (Mandanici and Bitelli, 2016; Shoko and Mutanga, 2017).

1.4 Climate change impacts, Striga adaptation, and mitigation capacity

The risk inflicted by the Striga weed is likely to be worsened by climate change and the inadequate adaptive or mitigation capacity across Africa, in addition to the limited impact documentation leading to inadequate preparedness (Nhamo et al., 2019; Niang et al., 2014). The impacts of climate change on the agricultural sector and the spread of Striga weeds in sub-Saharan Africa can be catastrophic. This is because agriculture provides approximately 17% of regional gross domestic product (GDP) and contributes 13% of the total export value (AGRA, 2017). Additionally, about 60% of the region's population depends on agriculture for their livelihood (FAO, IFAD, UNICEF, WFP, WHO, 2019). Therefore, understanding the climate change dynamics and their impacts on this climate-sensitive sector is critical for the sustenance of the region's livelihoods.

Thus, identifying cost-effective tools that can help to monitor and control these invasive Striga weeds before they spread to new environments is imperative. Spatial characterization of Striga weeds, using remote sensing and geospatial methods at multiple mapping scales is likely to enhance localized early warning systems, national preparedness, development of informed and scientifically sound policies, and adaptive management strategies. This, in turn, will enhance strides toward reducing the impact and spread of Striga, thus, ultimately achieving food security in Africa on a long-lasting and sustainable basis.

1.5 Aim and objectives

To determine and assess Striga weed (i.e. *S. hermonthica and S. asiatica*) occurrence and potential farming areas at risk of invasion across different agroecological farming systems and varying climatic scenarios in Kenya and Zimbabwe.

1.5.1 Specific Objectives

The specific objectives of this study were to:

1. Assess the potential of in-situ spectroscopy in detecting Striga weed at plot level using various machine learning algorithms,
2. Compare the potential of PlanetScope and Sentinel-2 spaceborne sensors in detecting and characterizing Striga weed at the field level,
3. Evaluate field- and landscape-scales Striga detection using Sentinel-2 data coupled with hierarchical cropland classification and a subpixel technique, and
4. Evaluate the utility of multi-source data, and integrative ecological niche modeling approach to detect Striga weed at a landscape scale and under different climatic scenarios.

1.6 Scope of the study

This study examined multiple-scale, multi-resolution data coupled with robust, cross-cutting, and efficient remote sensing techniques and machine learning modeling routines to detect, predict, and monitor Striga weed infestation in cereal crop fields. The study targets to detect and predict key Striga infestation and susceptible areas in Kenya and Zimbabwe for the most economically important Striga species *S. hermonthica* and *S. asiatica,* respectively using in-situ spectroscopic measurements and satellite-based imagery. This study demonstrated the potential of using the handheld spectral device to detect Striga infestation; and the utility of satellite-based data, at both ground and satellite levels using advanced machine learning algorithms. These algorithms were further used to compare the efficacy of a high spatial resolution image (PlanetScope) to a medium resolution image (Sentinel-2) as proxies for Striga determination and mapping at a landscape scale. The study further demonstrated a two-step classification method using Sentinel-2 multi-temporal data in Google Earth Engine (GEE) to characterize croplands in conjunction with the multiple endmember spectral mixture analysis (MESMA) algorithm at field and landscape levels. The study further highlighted the potential and strength of ecological niche and machine learning models in integrating ancillary data from multiple sources such as bioclimatic, cropping system, edaphic and phenological characteristics in predicting and characterizing the potential risk invasion of *S. asiatica* within heterogenous maize fields at a landscape scale under different climate scenarios. Accordingly, all the mapping approaches were designed to enable the inter-annual and inter-region transfer of classification methods across all countries in sub-Saharan Africa.

1.7 Study area

The study was conducted in two separate countries (Kenya and Zimbabwe) known to be hubs of the two most economically important Striga species i.e. *S. hermonthica and S. asiatica*. These two species significantly vary in their morphological structure, hence demanded different remote sensing approaches, mapping scale, and study sites where they occur in abundance *viz.* Kenya for the *S. hermonthica* and *S. asiatica* in Zimbabwe. The investigations for the occurrence, distribution, and detection of the *S. hermonthica* were conducted in Rongo sub-county in Kenya while the study to understand the potential *S. asiatica* modeling and mapping approach at landscape scale was conducted in Zimbabwe. Rongo sub-county covers an area of 213 km^2 and lies within the Migori county in western Kenya, while Zimbabwe is a landlocked country in southern Africa covering a land area of ~ 390 753 km^2 (Figure 1.1). A detailed description of each of the study sites relative to the specific chapter objective is given in each chapter.

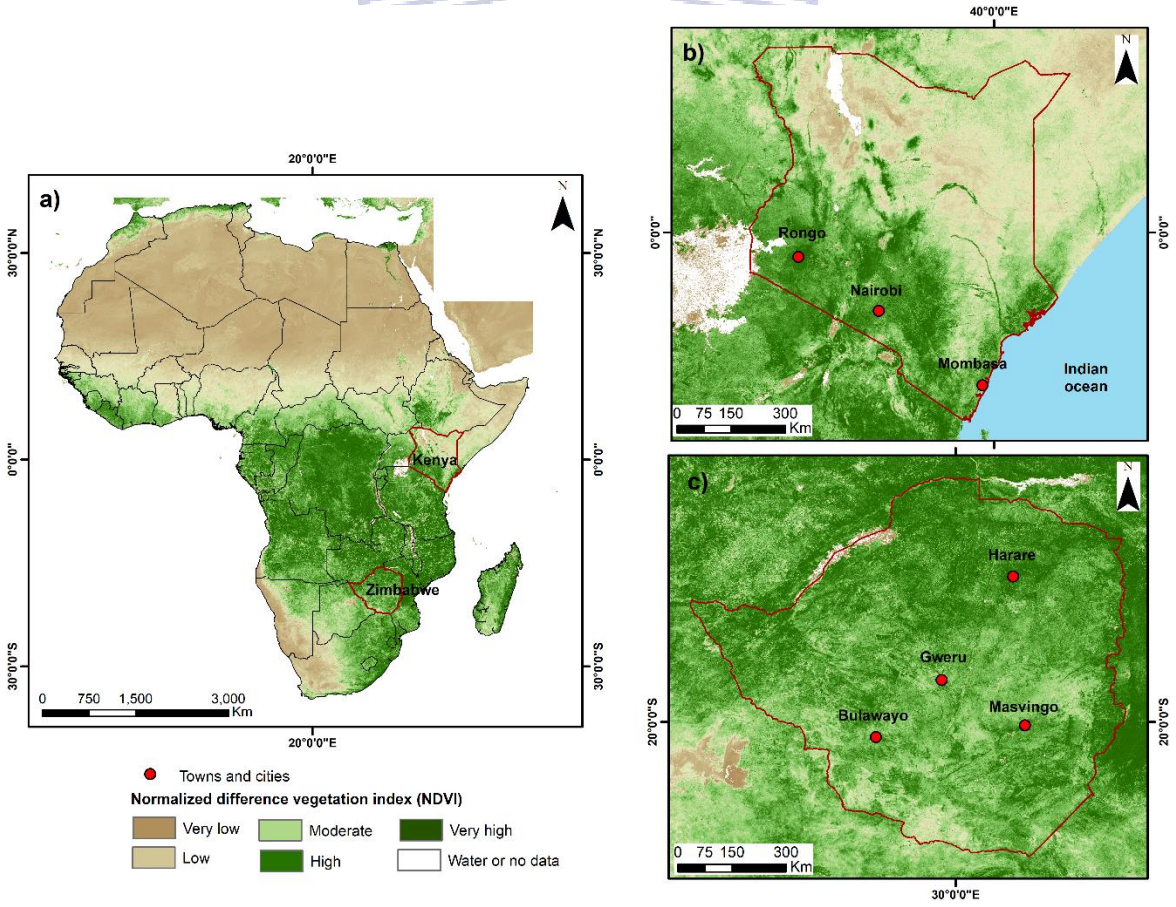


Figure 1.1: a) Location of (b) Kenya and (c) Zimbabwe in Africa. The background image is a December 2018 and 250 m moderate resolution imaging spectroradiometer (MODIS) normalized difference vegetation index (NDVI).

Chapter 2

Plot-scale Striga infestation detection using in-situ hyperspectral data

Photo: courtesy of Bester Tawona Mudereri 2017

This chapter is based on:

Mudereri, B.T., Dube, T., Niassy, S., Kimathi, E., Landmann, T., Khan, Z., Abdel-Rahman, E.M., 2020. *Is it possible to discern Striga weed (Striga hermonthica) infestation levels in maize agroecological systems using in-situ spectroscopy?* International Journal of Applied Earth Observation and Geoinformation 85: 102008. https://doi.org/10.1016/j.jag.2019.102008

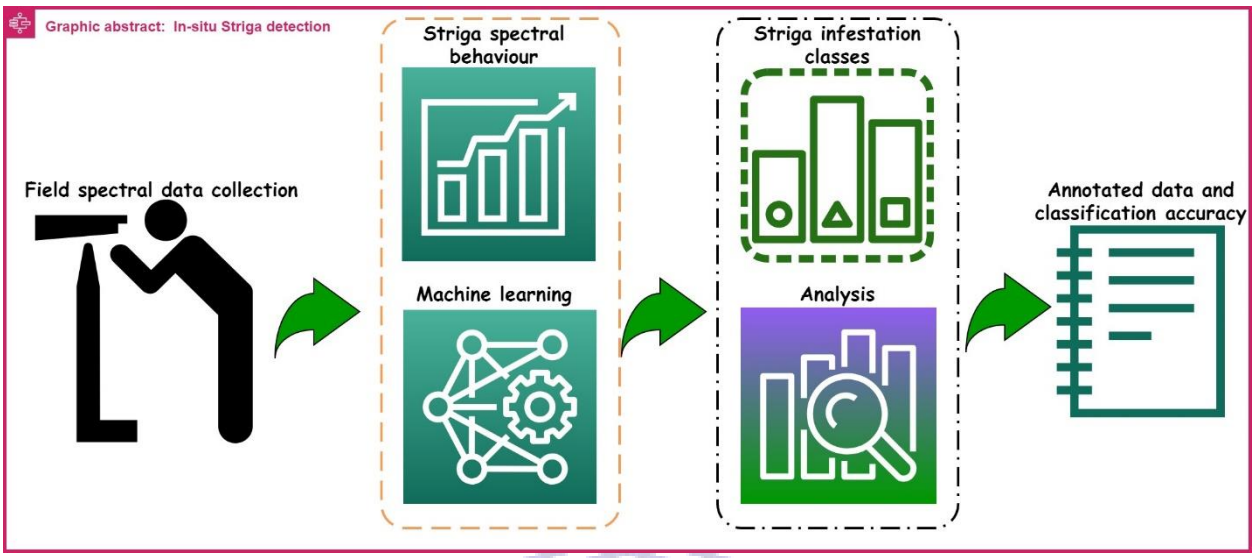
Abstract

The invasion by Striga in most cereal crop fields in Africa has posed a significant threat to food security and has caused substantial socioeconomic losses. Hyperspectral remote sensing is an effective means to discriminate plant species, providing possibilities to track such weed invasions and improve precision agriculture. However, essential baseline information using remotely sensed data is missing, specifically for the Striga weed in Africa. In this chapter, the spectral uniqueness of Striga compared to other co-occurring maize crops and weeds was investigated. In-situ FieldSpec® Handheld 2™ analytical spectral device (ASD), hyperspectral data and their respective narrow-waveband indices in the visible and near-infrared (VNIR) region of the electromagnetic spectrum (EMS) and four machine learning discriminant algorithms (i.e. random forest: RF, linear discriminant analysis: LDA, gradient boosting: GB and support vector machines: SVM) were used to discriminate among different levels of Striga (*Striga hermonthica*) infestations in maize fields in western Kenya. The Sentinel-2 waveband configurations were examined in their ability to map and discriminate Striga infestation in heterogeneous cereal crop fields. The in-situ hyperspectral reflectance data were resampled to the spectral waveband configurations of Sentinel-2 using published spectral response functions. Seven Striga infestation classes were sampled and detected based on three flowering Striga classes (low, moderate, and high) against two background endmembers (soil and a mixture of maize and other co-occurring weeds). A guided regularized random forest (GRRF) algorithm was used to select the most relevant hyperspectral wavebands and vegetation indices (VIs) as well as for the resampled Sentinel-2 multispectral wavebands for Striga infestation discrimination. The performance of the four discriminant algorithms was compared using classification accuracy assessment metrics. Striga from the two background endmembers i.e. soil and co-occurring vegetation (maize and co-occurring weeds) based on the few GRRF selected hyperspectral vegetation indices and the GRRF selected resampled Sentinel-2 multispectral bands was positively discriminated. RF outperformed all the other discriminant methods and produced the highest overall accuracy of 91% and 85%, using the hyperspectral and resampled Sentinel-2 multispectral wavebands, respectively, across the four different discriminant models tested in this chapter. The class with the highest detection accuracy across all the four discriminant algorithms was the “exclusively maize and other co-occurring weeds” (>70%). The GRRF reduced the dimensionality of the hyperspectral data and selected only 9 most relevant wavebands out of 750 wavebands, 6 VIs out of 15, and 6 out of 10 resampled Sentinel-2

multispectral wavebands for discriminating among the Striga and co-occurring classes. Resampled Sentinel-2 multispectral wavebands 3 (green) and 4 (red) were the most crucial for Striga detection. The use of the most relevant hyperspectral features (i.e. wavebands and VIs) significantly ($p \leq 0.05$) increased the overall classification accuracy and Kappa scores (±5% and ±0.2, respectively) in all the machine learning discriminant models. The results show the potential of hyperspectral, resampled Sentinel-2 multispectral datasets, and machine learning discriminant algorithms as a tool to accurately discern Striga in heterogeneous maize agroecological systems.

Keywords: invasive weeds; detection; maize; in-situ hyperspectral data; machine learning; resampled Sentinel-2.

Graphic abstract



2.1 Introduction

In Africa, food and nutrition insecurity due to crop losses is a chronic problem caused by insect pests, diseases, weeds, poor agronomic, and soil management practices (Sasson, 2012). This food insecurity is likely to be worsened by frequent unfavorable climatic conditions like droughts, climate change, and variability, among others (Rakotoarisoa et al., 2012). The most important staple crops on the African continent that secure food and nutrient to about 1.2 billion people are maize, sorghum, wheat, millet, and rice (Prăvălie et al., 2019). Among these economically important crops, maize plays a major role in the livelihood of people in sub-Saharan Africa. However, the productivity of maize has been on the decline in the last decade in sub-Saharan Africa, mainly due to the emerging of invasive pests and diseases such as stem borers, fall armyworm, maize lethal necrosis, and invasive weeds like Striga (FAO, IFAD, UNICEF, WFP, WHO, 2018).

Striga commonly referred to as the "witchweed" is considered to be the most economically important parasitic weed globally (Unachukwu et al., 2017). This parasitic weed attaches to the roots of the host plants after germination and causes considerable photosynthetic and productivity interference (Khan et al., 2002). Of the 23 Striga species predominant in Africa, *Striga hermonthica* is the most destructive, affecting a wide range of crops including maize, sorghum, millet, rice, and sugarcane (Ejeta and Gressel, 2007). Striga can reduce cereal production as much as 20–100% to more than 40 million households every year across Africa (Atera et al., 2013; Scholes and Press, 2008). Although these socioeconomic losses are difficult to quantify, it is estimated that in Africa alone, over US$ 1 billion is lost every year due to Striga infestation (De Groote, 2007; Spallek et al., 2013). Smallholder farmers are the most affected since they cannot afford the expensive Striga control mechanisms currently available on the market. These farmers often resort to inefficient hand weeding aimed at reducing the Striga seed bank within the soil, which is unsustainable. This problem is aggravated by the viability of Striga seeds in the soil for up to 20 years and their complex potential to spread via both mechanical and cultural processes (Khan et al., 2002).

Due to the destructive nature of Striga, numerous technological and research developments have been made to help control or minimize Striga's impacts on crop production. So far, efforts to control Striga have focused on the manipulation of genetics, chemical ecology, and phenology of

the weed (Midega et al., 2017; Oswald, 2005; Rispail et al., 2007; Samejima et al., 2016). However, on-farm Striga control technologies require spatiotemporal information on the weed to precisely prioritize sites for intervention and applications of such technologies. Usually, ground-based surveys and inspection methods are used to detect Striga-infested farms. This approach is often expensive, has a long-time lag, is laborious, and provides incomplete information on Striga hotspots. In contrast, remote sensing provides efficient, timely, synoptic, and inexpensive data that could effectively capture weeds spectral phenological responses at different spatiotemporal scales (Mutanga et al., 2017).

Studies have shown that weeds distribution and abundance can be estimated using diverse types of sensors and instruments such as field-based automated sensors (Smith and Blackshaw, 2003), unmanned aerial vehicles (de Castro et al., 2018; Peña et al., 2013), airborne multispectral and hyperspectral remote sensing (Mirik et al., 2013) among others. Yet, essential baseline information for the usage of such remote sensing information is absent for many high-impact invasive parasitic weeds like Striga (Große-Stoltenberg et al., 2016). In this chapter, the potential of using in-situ hyperspectral remotely sensed data at the plot level, to monitor Striga infestation in maize crops grown in an agroecological landscape in Kenya was explored. The candidate essentially tested whether canopy level in-situ hyperspectral data could discriminate among different Striga infestation levels and their co-occurring maize crop and other weeds. The potential capability of the Sentinel-2 multispectral band settings to detect and predict Striga infestation intensity was further tested, at the plot level, in heterogenous cereal crop fields.

Hyperspectral instruments acquire data in numerous quasi-contiguous spectral wavebands, allowing detection of the spectral features of plant biochemical and physical characteristics like pigments, nutrients, and water which are often masked when using the broadband multispectral data (Abdel-Rahman et al., 2013b; Landmann et al., 2015). Hence, hyperspectral data are efficient in discriminating weed species from their co-occurring crops based on their biochemical and physical characteristics providing the vast potential to precision farming for weed management (Große-Stoltenberg et al., 2016; Mureriwa et al., 2016). Additionally, in-situ hyperspectral data capture subtle spectral differences that are spectrally less distinct in airborne and spaceborne data (Sibanda et al., 2015b). Thus, in-situ hyperspectral platforms enable quick spectral measurements of targets on the ground and offer the opportunity for band-specific indices that breakdown

complexes concealed in biochemical and physical characteristics of plants (Huang et al., 2015).

These in-situ hyperspectral data are also operated under chosen appropriate atmospheric conditions unlike when operating satellite sensors (Chen et al., 2009). This enables the quality detection of unmixed energy captured from target objects without the influence of the bidirectional and diffuse scattering effects from other non-target features and the atmosphere (Jia et al., 2011). Also, the in-situ hyperspectral platforms can acquire spectral data at finer spatial resolution (up to a sub-meter), capturing the spectral vegetation signals at levels of a plant or an assemblage of plants. Such fine-scale remotely sensed data offer a deeper understanding of the interaction between parasitic weeds like Striga and the electromagnetic radiation at ground level before upscaling to airborne or spaceborne platforms such as Sentinel-2 (Kumar et al., 2001).

The relatively new generation of multispectral spaceborne sensors such as Sentinel-2 have assumed the use of relatively narrower wavebands (e.g. 15 nm spectral width), including those in the red-edge region of the electromagnetic spectrum (EMS), centered at 705, 740, and 783 nm that were not present in previous broadband sensors like Landsat 7, 8 and the advanced spaceborne thermal emission and reflection (ASTER: (Chemura et al., 2017a). Therefore, there has been a growing interest to test the Sentinel-2 data, regarding its potential to advance precision agriculture and other operational uses, particularly in low-income regions (Dhau et al., 2018; Mudereri et al., 2019b). This is mainly because Sentinel-2 data are freely available, with a relatively higher spatial resolution (10 m) and possess strategically placed bands at the red-edge region of the EMS, which makes the sensor versatile for many applications (Ochungo et al., 2019). Therefore, citing these positive characteristics, Sentinel-2 is hypothesized to be capable of providing timely data for the generation of critical products for Striga monitoring.

Despite the previously mentioned advantages posed by both hyperspectral and Sentinel-2 multispectral datasets, these datasets alone might not be adequate for detecting Striga infestation in complex and heterogeneous croplands. Merging the magnitude of the detail provided by hyperspectral data and the strength and capabilities of machine learning algorithms provides opportunities to reveal these complex structural and biophysical characteristics of weeds. However, one of the prominent problems in hyperspectral data processing and analysis is the dimensionality and multicollinearity inherent in the data (Adam et al., 2017). Multicollinearity is

associated with the limited number of training samples (n) in contrast to the abundance of hyperspectral wavebands (p), that often hinder the performance of the predictive models when they are validated using independent test dataset (i.e. overfitting) (Adam et al., 2017; Mureriwa et al., 2016). Studies have utilized robust machine learning classification algorithms like support vector machines (SVM, Vapnik, 1979), linear discriminant analysis (LDA, Fisher, 1936), gradient boosting (GB, Friedman, 1999) and random forest (RF, (Breiman, 2001) to deal with both the dimensionality and multicollinearity problems in the hyperspectral data.

All these mentioned machine learning classifiers are assumption-free methods that do not encounter variable overfitting challenges and yield a variable importance by-product which enables the selection of fewer, yet relevant input predictors (i.e. Striga weed). Specifically, GRRF and RF have shown to be successful methods in reducing the dimensionality of the hyperspectral data and simultaneously handle the multicollinearity in the data (Adam et al., 2017; Deng and Runger, 2013; Mureriwa et al., 2016). Nevertheless, previous studies demonstrated no consensus on the best machine learning classification algorithm and the best dimension reduction technique for invasive weeds discrimination (Große-Stoltenberg et al., 2016; Maxwell et al., 2018).

In this chapter, the innovation hinges on the hypothesis that flowering Striga is conspicuous from the rest of the photosynthetically green vegetation through their unique anthocyanins purple pigment in the flowers. To the best of the candidate's knowledge, the use of hyperspectral data and multivariate machine learning predictive models to separate different levels of Striga infestation using specifically the floral signal have not been examined. The present chapter was conducted with the following objectives:

a) To investigate the spectral uniqueness and behavior of flowering and non-flowering Striga owing to varying levels of infestation and co-occurring vegetation (i.e. maize crop and other chlorophyll-active materials), and

b) To discriminate among different levels of Striga infestations in maize crop using the most relevant hyperspectral and resampled Sentinel-2 multispectral features and machine learning classification algorithms

2.2 Methods

2.2.1 Study site

The study was conducted in the Rongo sub-county which lies within the Migori county in western Kenya. The study area is bound by the coordinates 0^0 39'12" S; 34^0 35'.40" E and 0^0 59'16" S; 34^0 37'21" E (Figure 2.1) at an altitude of 1 470 m above sea level. The climate in the study area is tropical and characterized by a yearly bimodal rainfall model with an average annual rainfall of 1 600 mm across the two rainy seasons i.e. during the "long rains" season occurring between March and June and a "short rains" season spanning November to January. The annual average temperature is 20.6^0 C and the relative humidity ranges between 50% and 70% while the soil type is loam, sandy, and clay.

The agro-natural ecosystem in the study area is dominated by scattered savanna grasslands in combination with deciduous and exotic forest vegetation, while the agricultural activities are mainly subsistence and small-scale farming. The crops grown in the Rongo sub-county include sugarcane as the main cash crop, maize, bean, groundnut, green gram, cassava, and some horticultural crops such as mango, banana, avocado, pawpaw, and indigenous vegetables. Maize in the Rongo sub-county area is grown as a mixed cropping system, with an average field size of 0.1 ha. The production of the crop in the study area is constrained mainly by rainfall variability and the invasive Striga weed. The yearly peak flowering period for Striga occurs between December and January during the short rains and again between May and June during the long rains season.

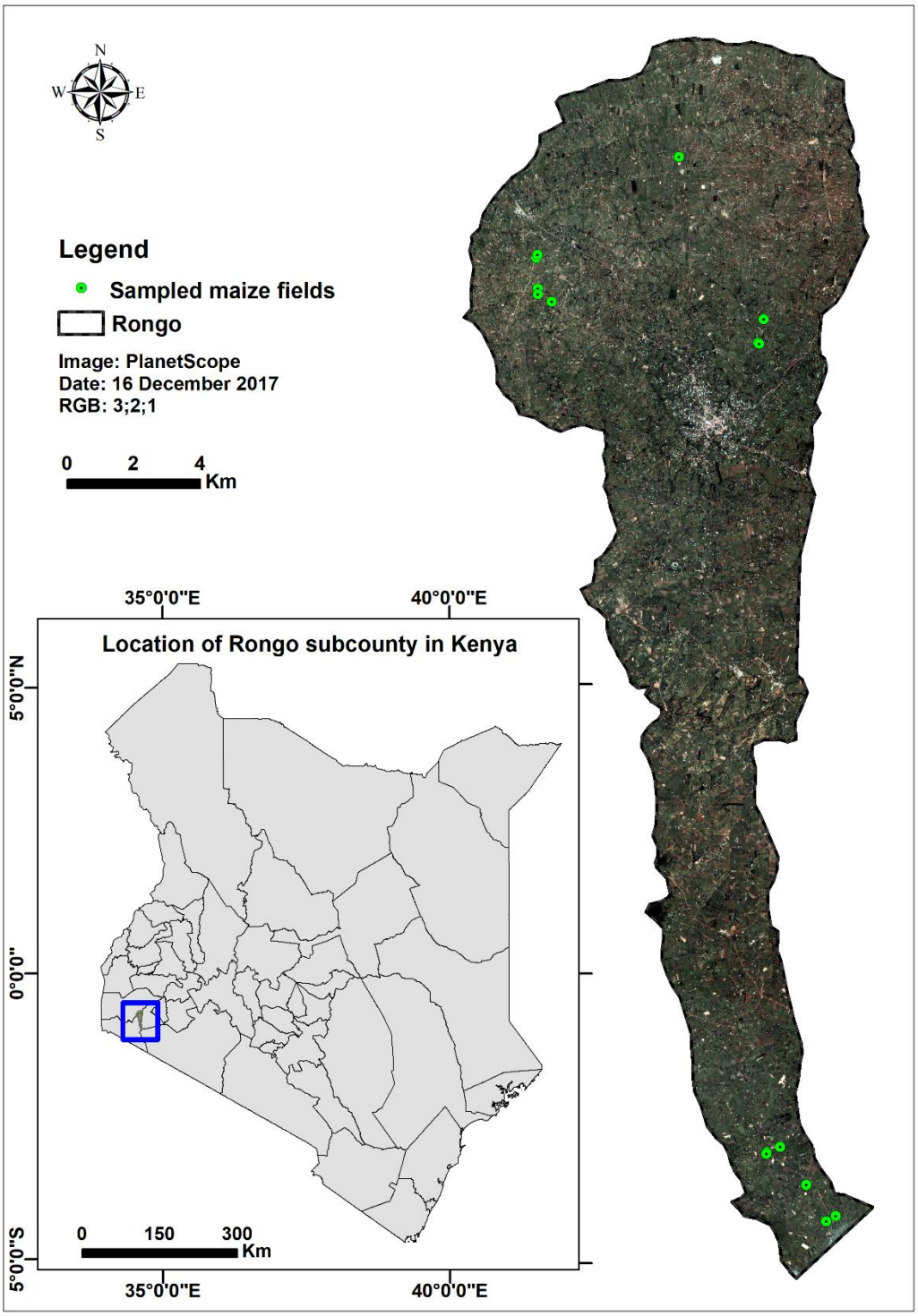


Figure 2.1: Location of Rongo sub-county in Migori county, Kenya, and the distribution of the sampled maize fields (n = 14). The image in the background is a PlanetScope image acquired on the 16th of December 2017 and displayed in RGB: red (band3), green (band2), and blue (band1).

2.2.2. Field sampling design

In total, 70 quadrats were sampled within all representative fields for Striga infestations. A total of 14 fields were purposively sampled during the period 12–16 December 2017 which coincided with the peak Striga flowering window in the study area. The purposive sampling procedure was guided by the presence and intensity of Striga infestation within each sampled maize field. In each sample maize field, a plot of 30 m x 30 m was selected, and within each plot, five quadrats measuring 1 m x 1 m each, were laid out along two crossing diagonal transects. Specifically, two quadrats were laid out across each of the two diagonal transects and 10 m away from the plot edges while one quadrat was laid in the center of the sample plot (Figure 2.2). In each quadrat, flowering and emerged Striga plants were counted. Infestation levels were categorized into three main classes; namely low (0–29 plants m^{-2}), moderate (30–90 plants m^{-2}), and high (> 90 plants m^{-2}) Striga infestation classes. Specifically, the Striga infestation classes were characterized according to the average Striga population in each quadrat and the damage it causes to the maize crop; following the procedure described in Ekeleme et al. (2014).

To test the influence of confounding features on the Striga spectral signal, spectral samples from soil background, and a combination of maize and other weeds in the sampling quadrat were also collected. This was necessary to test the influence of background spectral endmembers on the sensitivity of the Striga floral spectral signal. A total of seven classes of Striga infestation levels were assembled based on Striga floral signal sensitivity strength (number of Striga flowers per m^{-2}) and other EM abundances in the quadrat. These seven classes were derived from Striga infestation levels and corresponding combination background materials (soil or other non-Striga photosynthetically active vegetation) in the sample quadrats. Firstly, three flowering Striga infestation classes (i.e. low: LW, moderate: MW, and high: HW) with other green vegetation (maize and other weeds) were categorized. The second category comprised pre-flowering (PF) Striga infestation in combination with green vegetation. The third category included exclusively the green vegetation (GV) class (maize and other weeds) with no Striga infestation. Again, data for exclusively Striga within soil background (SB) with no green vegetation were enumerated, and finally bare soil (SO). The spectra of Striga classes that occurred with a soil background (SB and SO) were collected by manually removing all non-Striga photosynthetically active materials (for SB) and removing all vegetative material (for SO) from the sampling quadrat.

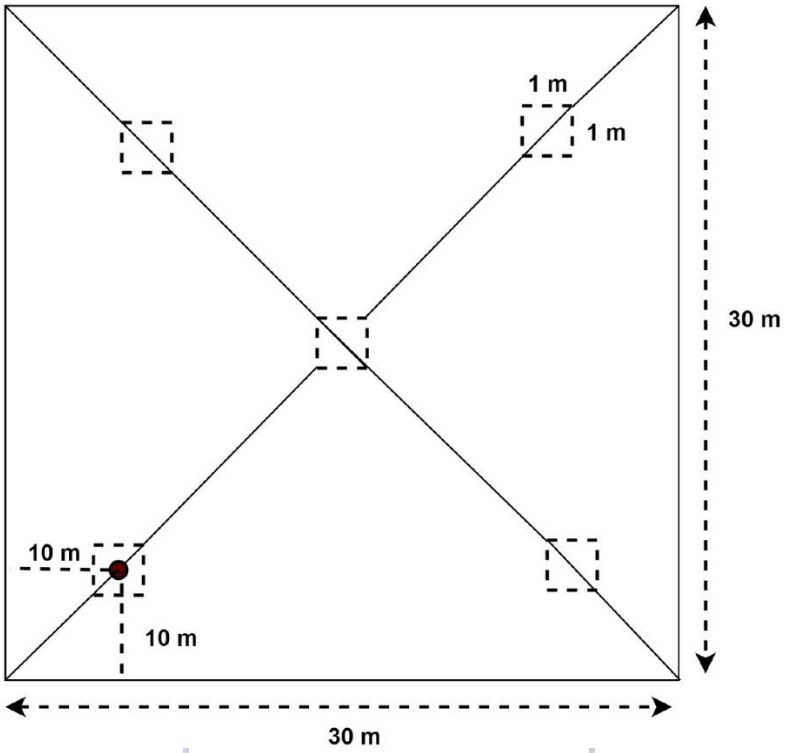


Figure 2.2: Example structure of the distribution of quadrats within a 30 m × 30 m maize plot used for Striga sampling data collection.

2.2.3. In-situ hyperspectral data acquisition

Canopy-level in-situ hyperspectral data were collected within the sample quadrats using the FieldSpec® Handheld 2™ spectroradiometer (HH2: ASD, 2010) under clear skies and stable wind conditions at between 10:00 hrs. and 14:00 hrs. local time (Greenwich Mean Time: GTM+3) as recommended by Sibanda et al., (2015b). The field spectroradiometer used for the spectral data collection captures reflected radiation in 325–1075 nm of the EMS with a built-in 2 nm sampling resolution (ASD, 2010). The device then resamples the spectral data to 1 nm spectral range. The hyperspectral measurements were collected from 1 m above the maize crop canopy using the bare optical input at the nadir field of view. This covers an area of ~0.5 m in diameter on the target, which was enough for capturing the spectral signal of a group of maize and Striga plants. It is a rule of thumb that the diameter of a spot of light that is covered by the HH2 when it is perpendicularly positioned to a target is approximately half the distance of the instrument to any specific target area (FieldSpec, 2017).

The instrument was held at arm's length (~0.9 m) from the observer to avoid scattered light from

surrounding objects including the instrument and the operator (Kumar et al., 2013). The spectroradiometer was set to internally and automatically collect and average 20 spectral readings for each sample spectrum. In each of the 70 quadrats, five spectra after optimizing and calibrating the measured radiance were measured using a Spectralon white reference of ~100% reflectance. The optimization and calibration were done before a first measurement and after collecting the spectra of each sampling unit (i.e. quadrat), or when the instrument saturated because of changing ambient weather conditions like sun irradiance (FieldSpec, 2017). The final total of averaged spectra for each respective class that was used in this chapter are summarized in Table 2.1

Table 2.1: Striga infestation level classes, with their respective class descriptions, class codes, sample sizes, training and testing samples used for employing the classification machine learning algorithms

Class description	Class code	Sampled spectra	Train	Test
Maize and other weeds (green vegetation) with no Striga infestation	GV	32	22	10
High Striga infestation level with other green vegetation	HW	101	70	31
Moderate Striga infestation level with other green vegetation	MW	71	50	21
Low Striga infestation level with other green vegetation	LW	56	40	16
Pre-flowering Striga with other green vegetation	PF	20	15	5
Exclusive Striga within a soil background with no other green vegetation	SB	21	15	6
Bare soil devoid of any photosynthetic material	SO	20	15	5

The spectral measurements acquired using the ASD were filtered using the "noiseFiltering" function and smoothened using the "Savitzky–Golay" filter in the "hsdar" package (Lukas et al., 2018) in R software (R Core Team, 2019). These filtered spectra were resampled to the spectral configuration of Sentinel-2 using the spectral response function, i.e. "SpectralResampling" of Sentinel-2 present in the "hsdar" package (Figure 2.3). The Sentinel-2 multispectral wavebands description, waveband centers, and their respective spectral wave ranges are shown in

Table 2.2: The wavebands, waveband centers, and their respective spectral width of the Sentinel-2 multispectral sensor. The wavebands that correspond to the in-situ hyperspectral data used in this chapter are shown in bold

Waveband	Waveband description	Waveband center (nm)	Wave range (nm)
1	**Coastal aerosol**	**443**	**433–453**
2	**Blue**	**490**	**458–523**
3	**Green**	**560**	**543–578**
4	**Red**	**665**	**650–680**
5	**Red-edge 1**	**705**	**698–713**
6	**Red-edge 2**	**740**	**733–748**
7	**Red-edge 3**	**783**	**773–793**
8	**Near-Infrared (NIR)**	**842**	**785–900**
8a	**Near-Infrared narrow (NIRn)**	**865**	**855–875**
9	**Water vapor**	**945**	**935–955**
10	Shortwave Infrared (cirrus)	1380	1360–1390
11	Shortwave Infrared 1 (SWIR1)	1610	1565–1655
12	Shortwave Infrared 2(SWIR2)	2190	2100–2280

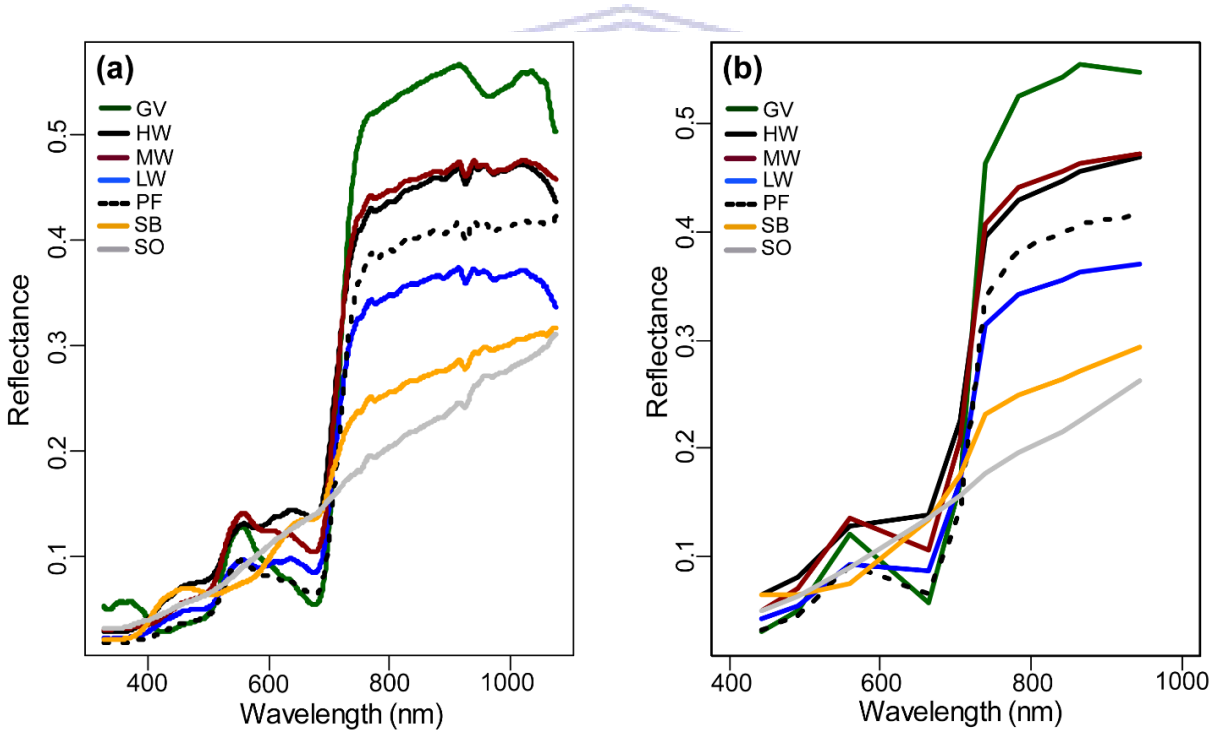


Figure 2.3: Mean canopy-level spectra of (a) in-situ hyperspectral reflectance and (b) resampled Sentinel-2 multispectral reflectance of the seven studied classes: maize and other weeds (green vegetation) with no Striga infestation (GV); high Striga infestation level with other green vegetation (HW); moderate Striga infestation level with other green vegetation (MW); low Striga infestation level with other green vegetation (LW); pre-flowering Striga with other green vegetation (PF); exclusive Striga stands within a soil background with no other green vegetation (SB); bare soil devoid of any photosynthetic material (SO) measured using Hand Held FieldSpec®2 (HH2) spectroradiometer in the 325–1075 nm wave range of the electromagnetic spectrum

2.2.4. Calculation of the narrow-waveband vegetation indices

In addition to the 750 wavebands, fifteen narrow-waveband vegetation indices (VIs) (Table 2.3) were calculated using the "hsdar" package (Lukas et al., 2018) in R software (R Core Team, 2019) and were used as predictor variables to discriminate among the seven Striga infestation classes. These indices were selected based on their relatedness to specific plant biophysical parameters (*e.g.* floral signal strength, plant health condition, plant pigments, and plant water content) and the availability of the narrow-wavebands used in their formulae in the hyperspectral data that ranged between 325 nm and 1075 nm.

Table 2.3: Hyperspectral narrow-waveband vegetation indices used in this chapter.

Vegetation index	Related to:	**Equation	Reference
Fluorescence ratio Blue/Red (SR7)	Fluorescence	R_{440}/R_{690}	(Große-Stoltenberg et al., 2016)
Water band index (WBI)	Water	R_{900}/R_{970}	(Ho, 2009)
Simple ratio pigment index (SRPI)	Pigments	R_{430}/R_{680}	(Große-Stoltenberg et al., 2016)
Double peak index (DPI)	Vegetation stress	$(R_{688} \times R_{710})/R^2_{697}$	(Große-Stoltenberg et al., 2016)
Anthocyanin reflectance index (ARI)	Anthocyanin	$(1/R_{500}) - (1/R_{700})$	(Ho, 2009)
Anthocyanin reflectance index2 (ARI_2)	Anthocyanin	$(1/R_{550}) - (1/R_{700})$	(Ho, 2009)
Datt4	Pigments	$R_{672}/ (R_{550} \times R_{708})$	(Große-Stoltenberg et al., 2016)
Plant Senescing reflectance index (PSRI)	Leaf senescence	$(R_{678} - R_{500}) / R_{750}$	(Große-Stoltenberg et al., 2016)
Double difference index (DDN)	Chlorophyll	$2 X (R_{710} - R_{660} - R_{760})$	(Große-Stoltenberg et al., 2016)
Modified Simple ratio (mSR)	Chlorophyll	$(R_{800}-R_{445})/ (R_{680} - R_{445})$	(Sims and Gamon, 2002)
Structure insensitive pigment index (SIPI)	Pigments	$(R_{800} - R_{445})/ (R_{800} - R_{680})$	(Ho, 2009)
Photochemical reflectance index (PRI)	Carotenoid	$(R_{531} - R_{570})/ (R_{531} + R_{570})$	(Sims and Gamon, 2002)
Photochemical Reflection Index x Chlorophyll content (PRI.CI2)	Carotenoid	$(R_{531} - R_{570})/ (R_{531} + R_{570})$ X $(R_{760}/R_{700} - 1)$	(Große-Stoltenberg et al., 2016)
Transformed Chlorophyll Absorption Ratio Index (TCARI2)	Chlorophyll	$3 \times ((R_{750}-R_{705}) - 0.2 (R_{750}-R_{550}) (R_{750}/R_{705})$	(Große-Stoltenberg et al., 2016)
Enhanced Vegetation Index (EVI)	Biomass/LAI	$2.5 \times ((R_{800} - R_{670})/ (R_{800} - (6 \times R_{670}) - (7.5 \times R_{475}) + 1)$	(Ho, 2009)

** R is reflectance at the respective hyperspectral narrow-waveband

2.2.5. Predictor variables' selection using guided regularized random forest (GRRF) algorithm

GRRF was used to select fewer, yet the most relevant narrow-wavebands, VIs, and resampled Sentinel-2 multispectral wavebands to discriminate the seven Striga infestation classes. The package "RRF" in R software (Deng, 2013; R Core Team, 2019) was used. The regularized framework considerably reduces the training time by building a single model (Deng and Runger, 2013). The GRRF uses the same concept of the RF model but uses the importance scores generated from RF to guide the variable selection process (Mudereri et al., 2019a; Mureriwa et al., 2016). The importance score of a variable in RF is obtained through the "Gini index" of all nodes across all RF decision trees obtained, and the variable is used to measure the purity of the feature at every

node to facilitate the voting process of RF trees (Breiman, 2002). Compared to the variable importance feature in the ordinary RF, GRRF provides the precise variables that are most suitable for predicting the feature from the multiple features dataset (Deng, 2013).

GRRF uses a gamma value to penalize the selection of new features over features already selected that possess similar gain (importance). The gamma value occurs between 0 and 1 with values closer to 1 executing higher penalties, hence selecting fewer relevant variables within GRRF. Comparatively, the values closer and equal to 0 increase the number of potentially relevant features selected, while the value of 0 yields similar variables to those produced when using an ordinary regularized random forest (RRF) (Deng and Runger, 2013).

In this chapter, a *gamma (γ)* value of 0.8 to limit the variables (i.e. narrow-wavebands or VIs) selection was used. The choice of $\gamma = 0.8$ was conservative as the highest gamma value of 1, extremely reduced the variables to too few (n = 3). The raw importance scores obtained from RF are normalized for each feature using Equations 2.1 to 2.3 to get the score used for variable selection in GRRF. For a detailed explanation of the theoretical and mathematical background of GRRF and how it functions, the readers are referred to Deng and Runger, (2013).

$$\text{gainG}(Xi) = \lambda i \, \text{gain}(Xi) \quad (2.1)$$

where: gain (Xi) denote the Gini information gain of using a feature Xi to split a tree node. And λi is calculated as:

$$\lambda i = 1 - \gamma + \gamma \, (\text{Impi} / \text{Imp}*) \quad (2.2)$$

where: Impi is the importance score of Xi from RF and Imp$*$ is the maximum importance score, Therefore:

$$\text{gainG}(Xi) = (\text{Impi} / \text{Imp}*) \text{ X gain}(Xi) \quad (2.3)$$

2.2.6. Machine learning discriminant algorithms

Four machine learning discriminant models namely; GB, LDA, RF, and SVM were used to discriminate the seven Striga infestation classes. The performance of these four models in discriminating the seven Striga infestation classes was compared using the in-situ hyperspectral data under the following five predictor variable criteria: (1) the clipped range (400–1075 nm) of

the collected narrow-wavebands (n = 676); (2) all the calculated narrow-waveband VIs are shown in Table 2.3 (n = 15); (3) The GRRF selected narrow-wavebands (n = 9); (4) The GRFF selected narrow-waveband VIs (n = 6), and (5) a combination of both GRRF selected narrow-wavebands and narrow-waveband VIs (n = 15).

Similarly, using the mentioned four machine learning discriminant models, the performance of (a) all the resampled Sentinel-2 multispectral wavebands (n = 10) and (b) the GRRF selected resampled Sentinel-2 multispectral wavebands (n = 6) were further compared. GB, LDA, RF, and SVM algorithms were selected because they have been widely used in classifying vegetation-related classes with reasonably high classification accuracies when hyperspectral datasets were utilized (Abdel-Rahman et al., 2014a; Dube and Mutanga, 2015; Ramoelo et al., 2015). Further, these four classification algorithms do not require the traditional regression assumptions which makes them useful in many cases (Holloway and Mengersen, 2018). They possess advantages such as: (1) easily identifying and adapting to inherent patterns and trends in data, (2) little to no human intervention in the running process, (3) versatile in handling ad-hoc multi-dimensional and multivariate types of data, and (4) mapping classes with complex characteristics (Maxwell et al., 2018).

The "Caret" package (Kuhn et al., 2018) in R software was used to run and validate all the four machine learning, discriminant models. The "Caret" package provides a standard syntax to execute a variety of machine learning discriminant approaches, thus simplifying the procedure of systematically comparing different algorithms and approaches (Maxwell et al., 2018). For consistency, the tune length parameter was set to 10, so that 10 values for each parameter were assessed. Also, all the variables were centered and rescaled for consistency before classification. Table 2.4 summarizes the "Caret" packages used to execute the four algorithms: RF, SVM, LDA, and GB. Again, for consistency purposes, the comparison of the performance of the four machine learning discriminant algorithms, were evaluated using the same dataset for all the models, split into a training set (70%) for model training and a test set (30%) for validation of the models (Dube and Mutanga, 2015; Qiao et al., 2017).

Model performance was presented using boxplots of overall accuracy and Kappa within the five different predictor variable criteria mentioned for the hyperspectral data and two-variable grouping

criteria of the resampled Sentinel-2 multispectral wavebands. The inter-class prediction performances of the different models were further assessed using confusion matrices derived from the best performance predictor variables category for each algorithm. The McNemar test for paired categorical data represented in contingency tables was performed at a 95% confidence interval (CI) to compare the performance among the four models in their ability to predict Striga severity classes using the GRRF selected variables.

Table 2.4: R software packages used by "Caret" that were used in this chapter and their respective caret syntax code.

Algorithm	Caret code	Package	Reference
Random forest	"rf"	Ranger	(Liaw et al., 2002)
Support vector machines	"svmRadial"	Kernlab	(Karatzoglou et al., 2004)
Linear discriminant analysis	"lda"	Mass	(Venables and Ripley, 2002)
Stochastic gradient boosting	"gbm"	gbm and plyr	(Greenwell et al., 2019)

2.3 Results

2.3.1. Spectral behavior of flowering and non-flowering Striga

Figure 2.4 demonstrates the variation in spectral responses of the different classes according to flowering compaction and color. The near-infrared (NIR: 750–1075 nm) revealed multiple scattering within the leaf structure, emanating from the different compositions of the classes. Nonetheless, the NIR displayed significant differences in magnitude with green vegetation producing a plateau of high reflectance compared to all other classes where pigments no longer absorb the radiation. The patterns of Striga infestation levels low, moderate, and high in the whole spectrum were similar but differed in the magnitude (> 5% difference between high and low). This is emphasized in Figure 2.4b in which the pattern for the classes is the same, but the magnitude of the reflectance differs as influenced by the Striga floral compaction and color. The higher the number of flowers in a plot, the more the reflectance magnitude increased within the region 500–700 nm. However, when compared to the other non-Striga classes, much variation was observed in the red section (620–680 nm).

The hyperspectral reflectance values for the "high", "moderate" and "low" Striga infestation, all peak around 550 nm and 670 nm which corresponded to the resampled Sentinel-2 wavebands 2 and 4, respectively. However, there is a slight depression in the reflectance values around 680 nm proceeding to increase again in the red-edge and NIR waveband regions. When Striga-infested plots were compared to GV, they all have peaks at 550 nm but there is a marked difference at 680 nm where there is a huge depression for the GV class. Similarly, within the blue region (400–500

nm), Striga infestation classes show higher values compared to the GV class. Furthermore, a similar trend is observed between flowering Striga and PF class (Figure 2.4c). Although visually the red-edge (680–750 nm) does not show any considerable variations in the spectra, green vegetation had the steepest gradient. The red-edge remains crucial in calculating vegetation indices as the gradient of the graphs reveals biochemical and ecophysiological vegetation parameters.

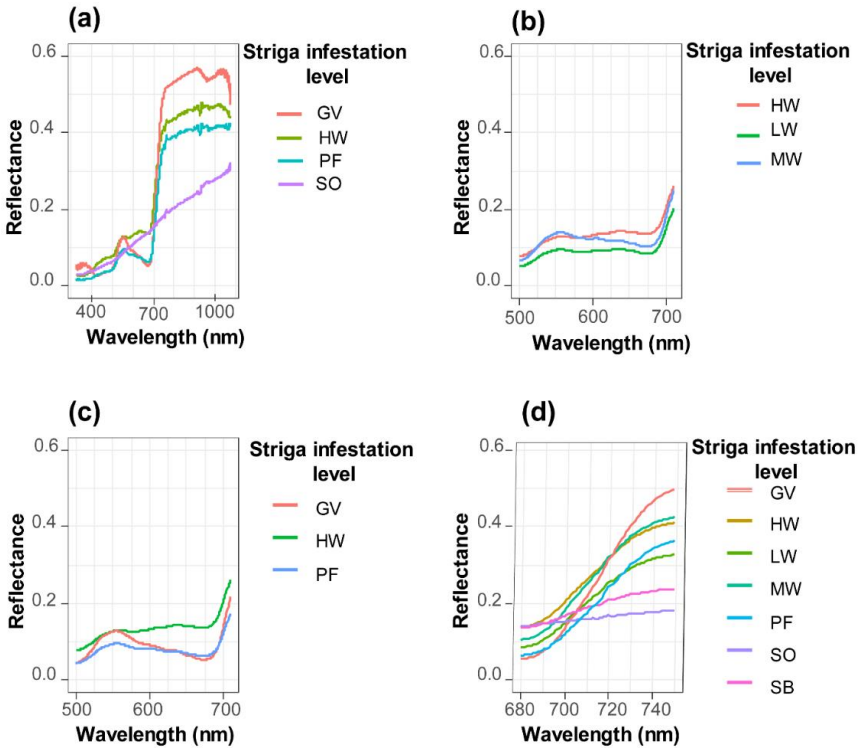


Figure 2.4: Comparison of spectral behavior for:(a) bare soil (SO), green vegetation (GV), high Striga infestation (HW) and pre-flowering (PF)Striga within the full spectral range(325–1075 nm); (b) high (HW), moderate(MW) and low (LW) Striga infestation levels within the visible range (500–700 nm); (c)green vegetation (GV), high Striga infestation(HW) and pre-flowering (PF) Striga within the visible range (500–700 nm); (d) all the seven classes used in this chapter within the red-edge spectral range (680–750 nm)

2.3.2. Predictor variables selection

The GRRF was able to determine only 6 narrow-waveband VIs, 9 narrow-wavebands, and 6 resampled Sentinel-2 multispectral wavebands to be of utmost relevance for discriminating the seven Striga infestation classes (Figure 2.5). The VIs that were selected by GRRF as the most relevant predictor variables were mainly related to pigments (Datt4, ARI, ARI2, and PRI.CI2). The most important narrow-wavebands are well distributed across the VNIR electromagnetic spectrum.

Many of the most relevant narrow-wavebands and resampled Sentinel-2 multispectral wavebands for discriminating among the infestation levels occurred within the green (band 3) and the red (band 4) regions of the EMS. Although most of these GRRF selected narrow-wavebands occurred within the visible range, the one with the highest variable importance value among them was identified within the red-edge region of the EMS (at 677 nm). The GRRF algorithm was able to determine uncorrelated variables for the VIs, however, most of the narrow-wavebands selected were correlated **(**Figure 2.6). The most correlated (> 80%) among the narrow-wavebands were those in the green region of the EMS, while Datt4 and DDN were also negatively correlated to most of the green, red, and NIR wavebands of the EMS.

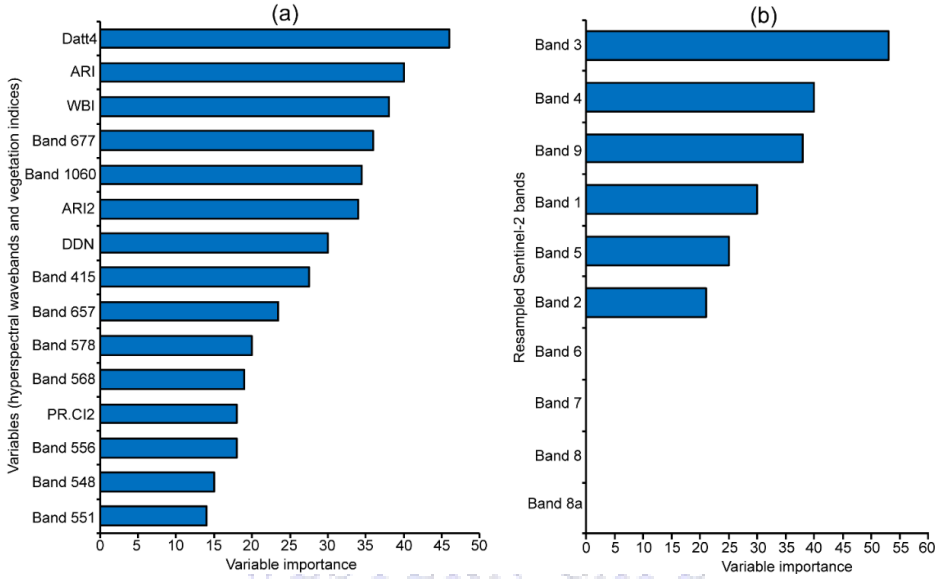


Figure 2.5: Predictor variables relevance for (a) both hyperspectral narrow-waveband vegetation indices and narrow-wavebands selected using the variable selection measure of the guided regularized random forest (GRRF) algorithm (b) Resampled Sentinel-2 multispectral wavebands selected using the variable selection measure of the GRRF algorithm. *See* Table 2.2 for the descriptions of Sentinel-2 multispectral wavebands

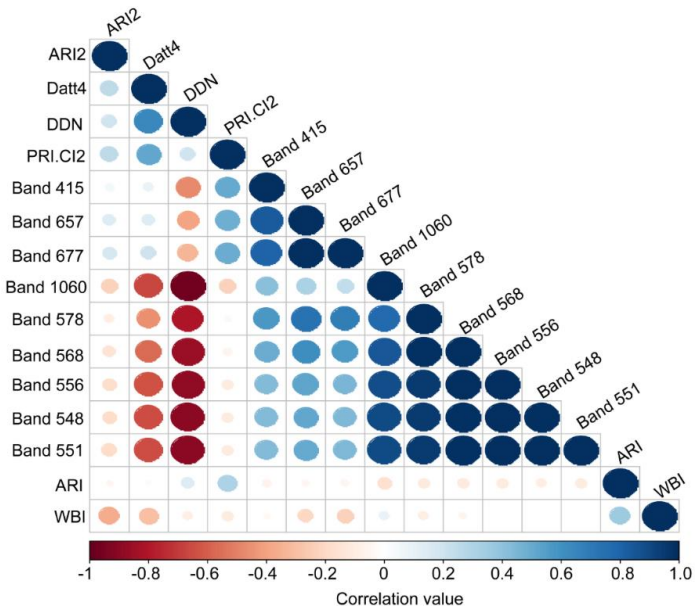


Figure 2.6: Correlation matrix for the guided regularized random forest (GRRF) selected hyperspectral wavebands and indices. Darker shades of blue and red colors indicate high variable correlation, while light shades indicate a low correlation between variables

2.3.3. Striga infestation levels discrimination using the four machine learning GB, LDA, RF, and SVM models

The results showed that the RF algorithm outperformed (overall accuracy of 91% and Kappa of 0.84, Figure 2.7) all the other three machine learning discriminant algorithms in discriminating among the seven Striga infestation classes using the hyperspectral data; and the resampled Sentinel-2 multispectral wavebands (overall accuracy 85% and Kappa of 0.80: Figure 2.8). This performance was followed by GB, LDA, and SVM, respectively. The use of only the selected VIs for the hyperspectral data and the GRRF selected wavebands for the resampled Sentinel-2, resulted in more accurate Striga infestation discrimination results compared with the use of other predictor variables across the machine learning discriminant algorithms, except for LDA (Figures 2.7 and 2.8).

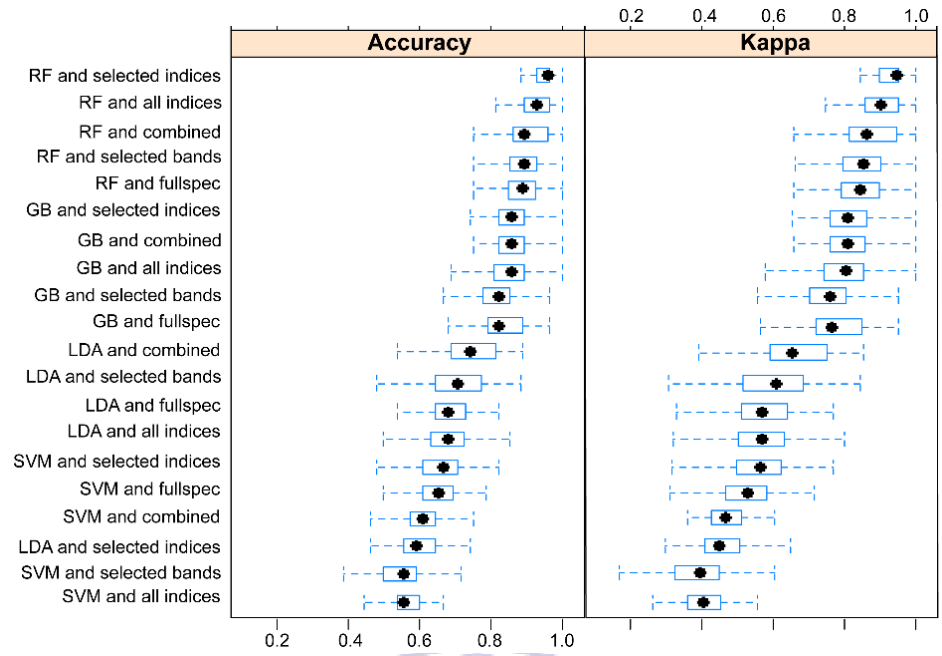


Figure 2.7: Striga infestation discrimination models performance as evaluated by overall accuracy and Kappa statistics using the clipped range (400–1075 nm) of spectral narrow-wavebands(*fullspec), narrow-waveband vegetation indices (*indices), all indices, and all narrow-wavebands (*combined), selected narrow-waveband indices (*selected indices) and selected narrow-wavebands (*selected bands). RF, GB, LDA, and SVM are random forest, stochastic gradient boosting, linear discriminant analysis, and support vector machines, respectively

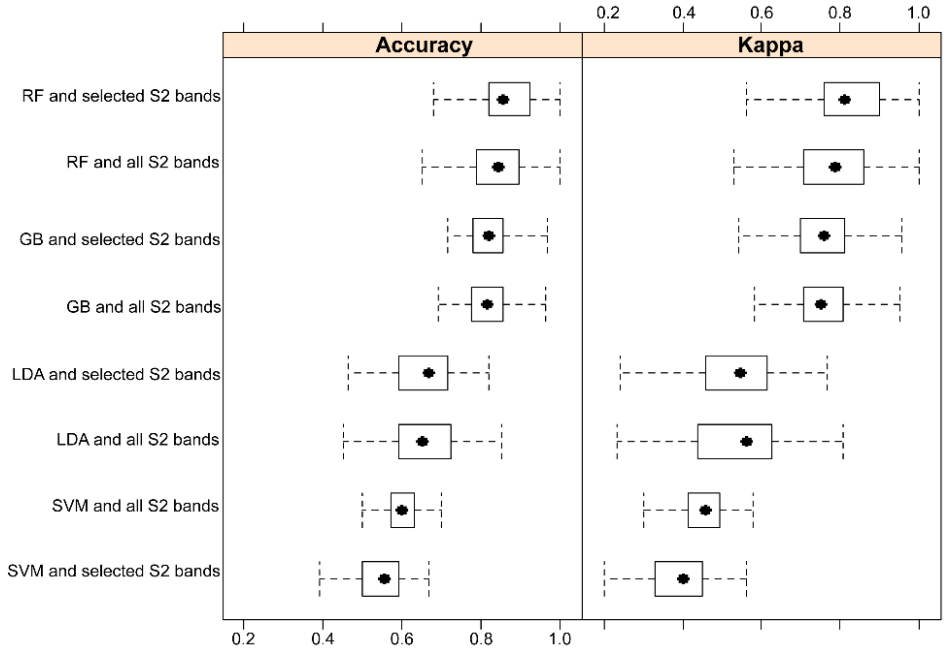


Figure 2.8: Striga infestation discrimination models performance as evaluated by overall accuracy and Kappa statistics using all the resampled Sentinel-2 multispectral wavebands or only the 6 guided regularized random forest (GRRF) selected re-sampled Sentinel-2 multispectral wavebands. RF, GB, LDA, and SVM are random forest, stochastic gradient boosting, linear discriminant analysis, and support vector machines, respectively

Although RF results showed superiority over all the other algorithms, the overlaps of the boxplots tests showed that there was no significant difference ($p \geq 0.05$) with the performance of the GB algorithm. The LDA and SVM algorithms achieved the least accurate Striga discrimination results to predict Striga infestation, using all the predictor variable criteria. Also, the Kappa statistic revealed that LDA and SVM models performed not significantly different ($p \geq 0.05$) from the model performance at random, whereas both the RF and the GB models produced high Kappa statistic values (Kappa > 0.75) for discriminating the seven classes. Considering the inter-class prediction accuracies (i.e. individual producer's accuracy (PA) and user's accuracy (UA) metrics), the RF model was superior over all the other models in predicting each of the seven classes using both the hyperspectral data or the resampled Sentinel-2 multispectral data (Table 2.5 and 2.6). Although all the four machine learning discriminant models were able to predict HW with a PA of at least 70%, MW and LW were not consistent among the models. On the contrary, all the machine learning discriminant models were relatively reliable in predicting GV, SO, and SB. Generally, the PA and UA metrics for LDA and SVM were relatively poor compared to the nonlinear decision tree-based algorithms (RF and GB) as shown in Table 2.5 and

Table **2.6**.

Table 2.5: Summarized confusion matrices and classification accuracies, overall accuracy (OA), producer's accuracy (PA) and user's accuracy (UA) of the random forest (RF), stochastic gradient boosting (GB), linear discriminant analysis (LDA) and support vector machines (SVM) discriminant models using the guided regularized random forest (GRRF) selected narrow-band indices

	Machine learning algorithm							
	RF		GB		SVM		LDA	
Class: Infestation level	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)
High (HW)	94	92	93	78	82	100	72	58
Moderate (MW)	85	92	78	84	100	10	41	44
Low (LW)	88	83	74	78	36	83	44	44
Green vegetation (GV)	100	100	70	100	100	100	100	100
Soil (SO)	100	100	100	100	0	0	100	100
Soil Background (SB)	100	100	86	100	0	0	71	83
Pre-flowering (PF)	100	100	75	100	33	33	30	67
OA (%)	91		83		60		55	

Table 2.6: Summarized confusion matrices and classification accuracies, overall accuracy (OA), producer's accuracy (PA) and user's accuracy (UA) of the random forest (RF), stochastic gradient boosting (GB), linear discriminant analysis (LDA) and support vector machines (SVM) discriminant models using the guided regularized random forest (GRRF) selected resampled Sentinel-2 multispectral wavebands

	Machine learning algorithm							
	RF		GB		SVM		LDA	
Class: Infestation level	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)
High (HW)	89	94	90	72	75	90	69	86
Moderate (MW)	87	76	70	85	50	10	76	37
Low (LW)	79	80	69	71	46	89	69	71
Green vegetation (GV)	92	100	80	90	50	100	40	33
Soil (SO)	100	100	100	100	100	100	100	100
Soil Background (SB)	95	100	94	100	100	30	83	100
Pre-flowering (PF)	75	75	78	80	0	0	40	67
OA (%)	85		81		65		55	

2.3.4. Pairwise model performance comparison using the McNemar test

The performances of the four machine learning discriminant models in predicting Striga infestation were significantly different ($p \leq 0.05$) from each other, except the comparison between LDA and SVM (hyperspectral wavebands) and GB and SVM (resampled Sentinel-2 multispectral wavebands) when using the pairwise McNemar test (Table 2.7). This further confirms the superiority of RF in the prediction of Striga infestation when compared to GB, LDA, and SVM.

Table 2.7: McNemar test for comparing the performance of the four machine learning discriminant models in predicting the seven studied Striga infestation classes using the hyperspectral wavebands and the resampled Sentinel-2 multispectral wavebands. RF, GB, LDA, and SVM are random forest, stochastic gradient boosting, linear discriminant analysis, and support vector machines models, respectively

	Hyperspectral wavebands		Resampled Sentinel-2 multispectral wavebands	
Comparison	Chi-square	*p*-value	Chi-square	*p*-value
RF vs GB	4.93	0.02*	5.14	0.02*
RF vs LDA	31.03	<0.001**	15.75	<0.001**
RF vs SVM	22.42	<0.001**	5.79	0.02*
GB vs LDA	19.86	<0.001**	7.84	<0.001**
GB vs SVM	10.26	0.001**	0.36	0.55
LDA vs SVM	0.21	0.64	6.05	0.01*

* Significant at 95% confidence interval (CI); ** Significant at 99% CI

2.4. Discussion

Several studies have demonstrated the importance of managing and controlling Striga infestation

and spatial spread (Atera et al., 2013; Khan et al., 2007; Spallek et al., 2013). This chapter examined the potential to use canopy-level in-situ hyperspectral data and resampled Sentinel-2 multispectral wavebands in distinguishing Striga from other co-occurring vegetated and non-vegetated materials within maize fields and to differentiate among different Striga infestation levels. Results from this current chapter show that Striga can accurately be discriminated from other vegetation and soil classes in maize fields using hyperspectral wavebands, narrow-band indices, resampled Sentinel-2 multispectral wavebands, and machine learning discrimination algorithms.

2.4.1. Spectral behavior of flowering and non-flowering Striga infestation classes and their co-occurring vegetation and soil.

In this chapter, the results showed that the spectral behavior of flowering Striga, non-flowering Striga, and other green vegetation differ. This can be attributed to the diversity of plant pigments occurring at different levels within the different co-occurring flora. Plant pigments are inherently associated with the biological function of leaves. Chlorophylls absorb light energy and allocate it to the photosynthetic system while, yellow pigments (carotenoids) also contribute energy to the photosynthetic apparatus and assist in resistance to environmental stress (Blackburn, 2007). Additionally, anthocyanins (red, pink, and purple pigments) may also serve as scavengers of reactive oxygen intermediates or as antifungal compounds (Litchenthaler and Buschmann, 2001; Sims and Gamon, 2002).

In light of the importance of pigments for leaf and petal function, dynamics in pigment quantities may provide details regarding their physiological state (Thenkabail et al., 2013). For instance, chlorophylls tend to decline more rapidly than carotenoids when plants are under stress or during leaf senescence (Sims and Gamon, 2002). Similarly, the reflectance response to incident radiation is influenced by the quantity and the interplay between the ratios of these pigments (Blackburn, 2007). These could have led to the differences observed among the flowering Striga, non-flowering Striga, and other vegetation classes. It was established that although the 530–570 nm are the portions that are mainly inclined to the green reflectance peak, all the plants had the highest peak at 550 nm within the visible region of the EMS, however, they differed significantly in the range 550–680 nm. This information is thus masked out when using spaceborne sensors such as Sentinel-2 that group these multiple narrow-wavebands into single broadbands i.e. in this case, into

wavebands 3 and 4.

Using the hyperspectral data revealed that the reflectance of the Striga classes within the 550–680 nm region of the EMS remained high, whilst the other green vegetation classes reflectance dropped considerably. This could be attributed to the presence and concentrations of anthocyanins within the petals of the Striga plants when flowering because chlorophylls and carotenoids absorbance increases whereas reflectance by anthocyanins increases significantly in this region of the EMS (Huang et al., 2015; Sims and Gamon, 2002). This concurs with the results of Blackburn (2007) who observed that an increase in reflectance in the red region of the EMS was linked with an increase in anthocyanins. In addition, in the analysis among the three Striga infestation levels (HW, MW, and LW), the magnitude of this reflectance increased with an increase in flower compaction (number of flowering Striga m^{-2}) which confirmed the importance of this region for discriminating Striga from other photosynthetic-active plants.

Leveraging on the presence of the purple color pigment in the flowers proved crucial in the separation of Striga occurring with other similarly co-occurring weeds and crops. *S. hermonthica* flowers are purple (Ejeta and Gressel, 2007; Khan et al., 2002; Spallek et al., 2013), hence the violet section of the EMS is very decisive to distinguish flowering Striga from green plants' material. However, the downside to the violet section is the short wavelength characteristic which is easily scattered by the atmosphere through Rayleigh scattering, thus most space-based satellites including Sentinel-2 do not consider wavebands within the violet section of the EMS (Campbell and Wynne, 2007).

Similarly, the spectral region between 413–420 nm region in the vegetation spectra are influenced by chlorophyll 'a' absorption, whereas the band 600 nm is influenced by chlorophyll 'b' absorption peak (Kumar et al., 2001; Thenkabail et al., 2013). There was a very sharp decline in the reflectance of green vegetation in these regions compared to the Striga classes. This decline is attributed to the absorption of chlorophyll by chlorophyll active plants. On the other hand, the region also demonstrates the low chlorophyll presence and the high influence of the pigments in the flowers (Litchenthaler and Buschmann, 2001).

The spectral differences observed in this chapter between flowers and green materials were anticipated in the hypothesis to positively produce reliable spectral differences between Striga and

co-occurring plants during the Striga flowering period (Ge et al., 2006). From other studies, it is known that the waveband 650 nm is the EMS region of vegetation pre-maxima spectral absorption, and 670 nm is sensitive to biomass and leaf canopy (Kumar et al., 2001; Sims and Gamon, 2002; Thenkabail et al., 2013). This was also observed in this chapter; however, for Striga detection, the significance of spectral absorption at these wavebands was not substantial enough to be useful.

Although there was multiple scattering within the NIR caused by water sensitive wavebands, the enormous variation in the magnitude of the reflectance is key in the separation of the classes used in this chapter. GV had the highest reflectance compared to all the other classes in the NIR spectral range. Apart from chemical composition, vegetation structure can also affect spectral features and influence spectral sensitivity and reflectance (He et al., 2011; Huang et al., 2015). The ability to separate among the studied classes could also be attributed to flower structure that influences the spectral features associated with the angle and arrangement of the petals. The flower structure coupled with lower water content could have similarly aided in the differences revealed in the prediction of the infestation classes as compared to the pre-flowering Striga which exhibited no difference with GV plots (Ge et al., 2006).

Therefore, the best period to predict Striga's presence within crop fields is during the peak Striga flowering period. These results concur with Best et al., (1981), who concluded that the best period to discriminate among eight plant species that they were studying was during the flowering and early seed development stages. It is critical to note that using spectral signatures for detecting weeds should be used with caution since different phenological stages of plants show significant variations in their spectral reflectance depending on the flowering stage of the species (Schmidt and Skidmore, 2003). Carvalho et al., (2013) suggested that further studies might be necessary to analyze what could cause such leaf and flower predictive spectral differences.

Additionally, although the results are valid at the plot scale, future studies using Sentinel-2 data, airborne, or UAV could be used for seamless wall-to-wall Striga mapping and upscaling from plot scale to field and landscape scales during the peak Striga flowering season.

2.4.2. Most relevant wavebands and indices using the GRRF approach

This chapter employed the robust GRRF for the resampled Sentinel-2 multispectral wavebands,

hyperspectral wavebands, and their derived VIs variable selection for multiclass classification. The traditional RF positively provides the variable importance parameters to lead the GRRF variable selection procedure (Deng and Runger, 2013). Because of the expected very high autocorrelation among the quasi-contiguous hyperspectral wavebands (1-nm interval), the variable importance values were also very similar among the different wavebands. Hence, some of the selected narrow-wavebands for detecting Striga were still autocorrelated. However, the GRRF algorithm decreased the multidimensionality of narrow-wavebands (9 out of 750 were selected) and their derived VIs (6 out of 15 were selected) as well as for the resampled Sentinel-2 multispectral wavebands (6 out of 10) without compromising key information relevant to the Striga and co-occurring vegetation classes. This was following the findings of Adam et al., (2017); Deng (2013), and Mureriwa et al. (2016) who reported a considerably reduced hyperspectral narrow-waveband dimensionality as a result of using the GRRF algorithm.

Specifically, the selected narrow-wavebands and VIs concurred and vindicated the importance of the already identified important EMS regions for separating the seven Striga infestation classes. Previously used variable selection methods like "varSelRF" and "Boruta" in R (R Core Team, 2019) are computationally expensive and may yield inexplicit variable importance outputs compared to GRRF (Mureriwa et al., 2016). In other words, the variable importance by-product of such RF-based variable importance procedures could remain dimensionally huge and redundant without identifying a few non-correlated and the most relevant variables.

Furthermore, several plant characteristics such as biochemistry, canopy structure, and soil parameters are combined within the canopy spectrum (Große-Stoltenberg et al., 2016). Thus, using the canopy-level spectrum rather than leaf-level data is key for further contrast with airborne or satellite remote sensing data such as Sentinel-2 used in the present chapter. Therefore, in this chapter, the candidate targeted to use VIs that incorporate these characteristics and could easily be upscaled to space-based satellite data.

Pigment indices cater to the problem of overlapping absorption characteristics of the different pigments rendering them more informative than the raw wavebands alone. The fact, this chapter showed that the most relevant VIs were related to pigments (chlorophylls, carotenoids, and anthocyanins) and water content, is attributable to the sampling season which was during the peak

Striga flowering and maize crop vegetative growth stage. In addition, the higher chlorophyll content detecting VIs (DDN and Datt4) that were selected among the relevant VIs for Striga detection, are known features that correlated to plant health, leaf area index (LAI), and light use efficiency.

The differences in chlorophylls, carotenoids, and anthocyanins among the seven Striga classes are vital to the physiological responses and resilience of plants to natural episodic events or seasonal fluctuations (Blackburn, 2007). These dynamics are captured very well in the indices calculated at the canopy level. Several studies offer credit to the two-band VIs, which correspond to the flanks of the main chlorophyll absorption feature in the red (530–630 nm) and the waveband located at 700 nm (which resembles waveband 5 of the Sentinel-2 sensor) to be the most sensitive to pigment concentration over the normal range (Kumar et al., 2001).

2.4.3. Performance of the machine learning classifiers for discriminating among the Striga infestation classes

In general, the determination of the best classifier for a particular application case when remotely sensed data are utilized depends on the accuracy measure selected and the intended objective of the analysis (Maxwell et al., 2018). For the case of this chapter, the RF and GB were identified as the best classifiers for Striga detection according to their overall accuracies. The resampled Sentinel-2 multispectral wavebands showed good potential to detect and map Striga at acceptable accuracies using the RF and GB algorithm with overall accuracies of 85% and 81%, respectively. However, the RF and GB classifiers experienced challenges by having some false positive cases when attempting to separate the Striga severity classes. These false-positive instances are important as detecting a damaging parasitic weed at the early stages of invasion is also fundamental for a real-time intervention (Große-Stoltenberg et al., 2016).

In both cases (i.e. using hyperspectral data or resampled Sentinel- 2 multispectral wavebands), it is expected the SVM to have performed better than its obtained results in this chapter since other studies have shown the high performance of SVM for detecting weed infestation using remotely sensed data (Brereton and Lloyd, 2010; Große-Stoltenberg et al., 2016; Pal and Mather, 2005). However, the relatively small sample size in some of the classes could have hindered the performance of the SVM algorithm as it is quite sensitive to imbalanced and small sample sizes

(Maxwell et al., 2018). One other reason for a relatively low SVM performance could be due to the use of a default linear hyperplane and SVM parameters; *viz.* gamma (γ) and sigma (C). Studies have shown that the optimization of these two SVM parameters would counter for the expected nonlinear relationship among the classes, hence enhancing the performance of the classifier (Abdel-Rahman et al., 2014b; Maxwell et al., 2018).

The performance of RF, GB, and SVM in this chapter is in agreement with some recent studies that have utilized leaf-level or canopy-level hyperspectral data and one of these classifiers to detect a plant trait (Große-Stoltenberg et al., 2016; Litchenthaler and Buschmann, 2001; Thenkabail et al., 2013). Specifically, the two non-linear classification algorithms (i.e. GB and RF) attained the best performance results when using the GRRF selected variables for both selected hyperspectral wavebands or when the selected resampled Sentinel-2 multispectral wavebands were used.

These results concur with Mureriwa et al. (2016) who used GRRF and RF to detect *Prosopis* using field spectral measurement data and found that reducing the number of redundant spectral variables increased the accuracy of the detection. In all cases, the RF classifier proved to be a very robust and reliable model for predicting subtle differences between classes and non-linear effects from spectral scattering between plant components. This is because RF is robust, yet can still accomplish high prediction accuracies even when the observation data are low or when the variables are highly correlated (Khadka and James, 2017; Thamaga and Dube, 2018). However, GRRF outperforms standard RF in relevant variable selection. It can, therefore, be concluded from the results that, if accurate detection of Striga infestation is to be conducted, combining the RF model and GRRF would provide the best model of choice regardless of the dimensions offered by the prediction variables and observations or the mapping scale.

2.5. Conclusions

In this chapter, the possibilities of using canopy-level in-situ hyperspectral data for predicting the presence and level of Striga infestation using their flowering characteristics are demonstrated. Prediction of seven classes of Striga infestation is possible with satisfactory overall accuracies (up to 94% overall accuracy), specifically during the peak flowering period including at Sentinel-2 spatial and spectral scales. However, due to the reliance on the flowering, it remains a challenge to pre-detect Striga before the damage is done. For more precise results on a global scale, remote

sensing could, therefore, be used to detect and model the condition of the infested maize rather than targeting Striga itself.

The GRRF model provided an easy and accurate variable selection platform that selects fewer and uncorrelated hyperspectral features relevant to the features of interest like Striga infestation. Specifically, this chapter shows that the selected narrow-waveband VIs; WBI, ARI, ARI2, Datt4, DDN, and PRI.CI2, narrow-wavebands (415 nm, 548 nm, 551 nm, 556 nm, 568 nm, 578 nm, 657 nm, 677 nm, and 1 060 nm) and Sentinel-2 multispectral wavebands (band 3, band 4, band 9, band 1, band 5 and band 2 in order of importance), are very relevant for Striga infestation prediction in maize fields in semi-arid agro-ecosystem.

The machine learning RF classification algorithm emerged as a very robust and reliable model for predicting differences among Striga occurring and other weeds and crop classes. However, there is a need to investigate the temporal and spatial variability of the flowering signal of Striga during the peak flowering season to explore upscaling options for the monitoring of the floral cycle using high spatial resolution multispectral data. Sentinel-2 data coupled with MESMA which separates spectra within image pixels by identifying the percentage contribution of each class with more than one endmember could also be explored to detect Striga infestation when large scale image data are utilized. This would bring an immense benefit to the landscape assessment of the floral cycle and infestation.

Findings from this chapter will be of utmost importance in understanding Striga infestation in heterogenous crop fields in sub-Saharan Africa. Although the use of the field hyperspectral data in vegetation studies is no longer new, the results indicate the capabilities and application of such remotely sensed data, as a tool for excellent detection of Striga infestation and other vegetation classes. These results provide opportunities to researchers, to apply a similar approach in precision agriculture using airborne or UAV data and platforms to detect the hotspots of Striga infestation at localized scales.

Chapter 3

Field-scale Striga characterization using multi-resolution satellite-based multispectral data and machine learning algorithms

Photo: courtesy of Bester Tawona Mudereri 2018

This chapter is based on:

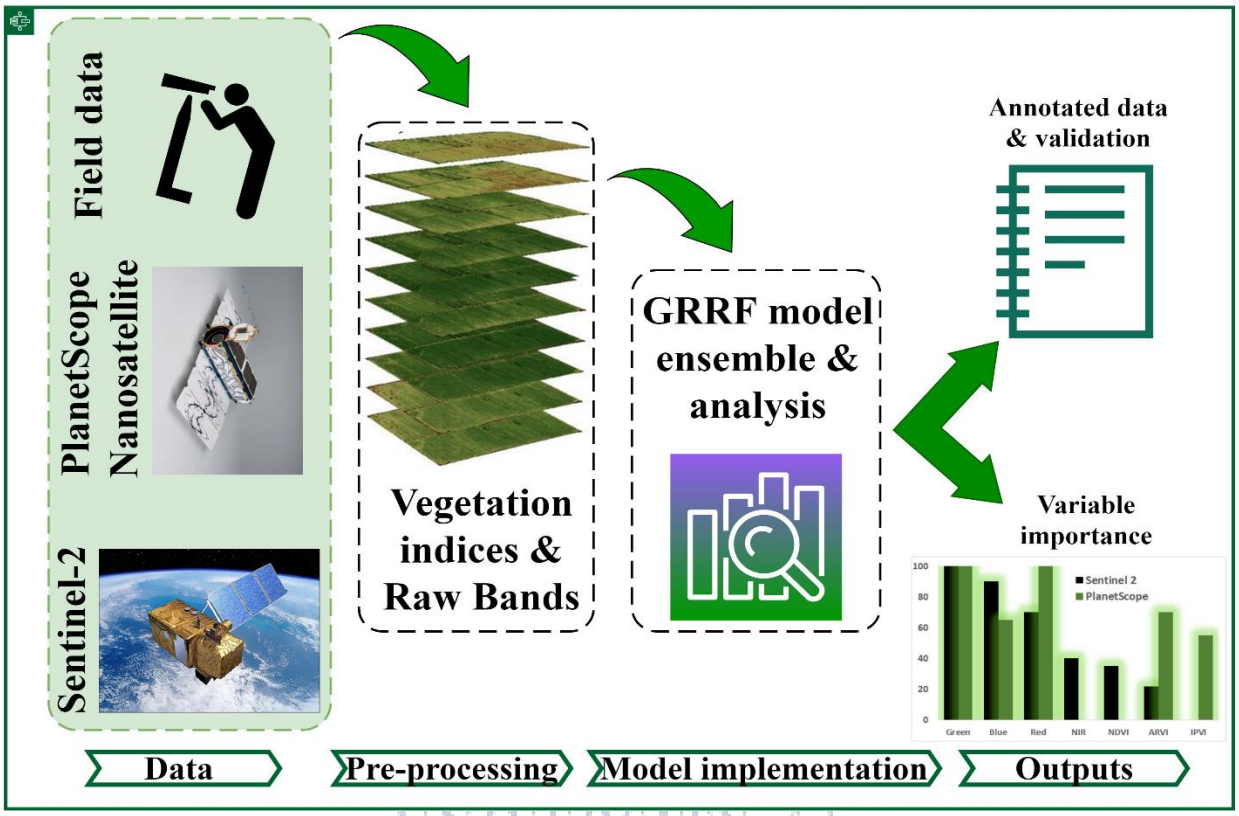
Mudereri, B.T., Dube, T., Adel-Rahman, E.M., Niassy, S., Kimathi, E., Khan, Z., Landmann, T., 2019. *A comparative analysis of PlanetScope and Sentinel-2 space-borne sensors in mapping Striga weed using Guided Regularized Random Forest classification ensemble.* ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-2/W13, 701–708. https://doi.org/10.5194/isprs-archives-XLII-2-W13-701-2019

Abstract

Weeds are one of the major restrictions to sustaining crop productivity. Weeds often out-compete crops for nutrients, soil moisture, solar radiation, space and provide platforms for the breeding of pests and diseases. The ever-growing global food insecurity triggers the need for spatially explicit innovative geospatial technologies that can deliver timely detection of weeds within agroecological systems. This, in turn, will aid to pinpoint maize fields that have to be prioritized for weed control. Satellite remote sensing offers incomparable opportunities for precision agriculture, ecological applications, and vegetation characterization, with vast socioeconomic benefits. This chapter compares and evaluates the strength of Sentinel-2 satellite with the constellation of Dove nanosatellites i.e. PlanetScope data in detecting and mapping Striga (*Striga hermonthica*) weed within intercropped maize fields in Rongo sub-county in western Kenya. The Sentinel-2 and PlanetScope derived spectral data and vegetation indices were applied for mapping the Striga occurrence. Data analysis was implemented, using the guided regularized random forest (GRRF) classifier. Comparatively, Sentinel-2 demonstrated slightly lower Striga detection capacity than PlanetScope, with an overall accuracy of 88% and 92%, respectively. The results further showed that the VNIR (blue, green, red, and NIR) and the ARVI were the most fundamental variables in detecting and mapping the Striga presence in maize fields. Findings from this work demonstrate that Sentinel-2 data can provide spatial explicit near real-time field-level Striga detection, a previously daunting task with broadband multispectral sensors.

Keywords: feature selection; food security; high spatial-temporal resolution; nanosatellites; new generation sensors; precision agriculture; weed detection; sub-Saharan Africa

Graphic abstract



3.1 Introduction

3.1.1 Background

Timely and repeatable spatial explicit information on inter- and intra-field variability of agroecological systems is key to devise adaptive and informed management decisions regarding crop productivity, resource allocation, and labor efficiency (Houborg and McCabe, 2018a). Advancements in remote sensing technologies, data analytics, research, and monitoring initiatives of invasive pests and weeds provide tremendous benefits to the continuous and near-real-time observations of crop health and crop phenological development (Mutanga et al., 2017). However, there is an inevitable unbalanced trade-off between the spatial resolution and frequency of observations that can be derived from the conventional satellites. Generally, most conventional satellites providing very frequent observations (daily and near-daily revisit time) lack high spatial resolution, such as the moderate-resolution imaging spectrometer (MODIS) (Masocha et al., 2018). High frequency of observation coupled with high spatial resolutions facilitates the rapid detection evolving from crop dynamics at spatial scales, high enough for timely interventions, and effective resource management (Houborg and McCabe, 2018a).

The advent of the CubeSat (~ 10 kg) in particular, data from Planet Labs Inc, provides satellite images collected by a constellation of nanosatellites (Doves) at very high spatial resolutions (VHSR) (Baloloy et al., 2018). Constellations of these 175 CubeSats are proving to be an innovative source of data, with vast potential to overcome the spatial-temporal limitations of conventional single-sensor satellite missions (Houborg and McCabe, 2018b).

These nanosatellites are equipped with unique multispectral sensors VNIR (420–900 nm) at 3 m spatial resolution and provide daily global data equivalent to a daily collection capacity of 350 million km²/day (Planet, 2018). However, since the first launch of the first flock of Doves on the 22nd of June 2016, research on the possible benefits afforded by PlanetScope data for landscape species characterization is still limited (Cooley et al., 2017). This could be attributed to the newness of CubeSats, the cost of imagery, concerns over cross-sensor calibration, location accuracy, image quality, data availability, and accessibility (Cooley et al., 2017). However, these nanosatellites can essentially improve spatial analysis methods for weed identification, detection, and mapping within agroecological systems by providing consistent information showing in-field variability in

weeds configuration and growth patterns (Yue et al., 2018).

Although the CubeSats are superior in their spatiotemporal resolution, the radiometric coverage is inferior to the conventional satellites such as Sentinel-2. This chapter evaluates and intercompares the capabilities and potential scientific utilities of Sentinel-2 data with the very high spatial resolution nanosatellite data of PlanetScope to detect and map Striga (*Striga hermonthica*) in intercropped African farm fields. Sentinel-2 is a combination of two sensors (2A-2B) developed by the European space agency (ESA) (Chemura et al., 2018). These sensors provide images at 10, 20 and 60 m spatial resolution and a spectral range of 440 to 2 280 nm at a combined global average revisit time of 5 days (Immitzer et al., 2016). Although several capabilities of Sentinel-2 have been successfully tested, such as evaluating rangeland quality (Ramoelo et al., 2015; Shoko and Mutanga, 2017), above biomass estimation (Sibanda et al., 2015a) water hyacinth mapping (Thamaga and Dube, 2018) and understanding crop behavior (Veloso et al., 2017; Vrieling et al., 2018) its utility in detecting and mapping Striga occurrence is not documented. The use of Sentinel-2 imagery is advantageous because of its spatiotemporal coverage, the presence of the strategically positioned red-edge bands, 13 multispectral bands, its global footprint, and free-availability.

Striga is one of the severe biological crop pests reducing food production in sub-Saharan Africa resulting in more than US$1 billion in losses every year (Atera et al., 2013; Scholes and Press, 2008). Striga is a parasitic weed whose existence relies on the availability of cereal crops in particular maize and sorghum (Khan et al., 2002; Midega et al., 2017). Since most optical remote sensing instruments capture the top of canopy reflection, the use of spaceborne sensors in mapping Striga occurrence poses huge mapping challenges as the weeds mostly occur underneath cereal plants. Crop health condition and phenology metrics entrenched within vegetation indices are, therefore, crucial to predict and inform on the existence of the understory weed.

In this chapter, the strength and usability of Sentinel-2 compared to the PlanetScope data in detecting and mapping Striga occurrence within maize fields in western Kenya were evaluated using the guided regularized random forest (GRRF) classification ensemble. GRRF algorithm is superior in its ability to identify key variables and robust in data processing capabilities (Mureriwa et al., 2016). GRRF is regarded as one of the most robust and effective classification methods for

agroecological systems, using very high spatial resolution satellites like PlanetScope, particularly within heterogeneous landscapes (Adam et al., 2017; Kyalo et al., 2017).

3.2 Materials and methods

3.2.1 Study area

The study was conducted in the Rongo sub-county, Migori county of Western Kenya. Rongo is administratively divided into four wards namely; North, East, Central, and South Kamagambo (Figure 3.1). The town is located at latitude -0.75786S and longitude 34.60901E at an altitude of 1 470 m. The study area is bound by the coordinates 0^0 39'12" S; 34^0 35'.40" E and 0^0 59'16" S; 34^0 37'21" E. It receives an average annual rainfall of 1 600 mm, with a bimodal yearly seasonal pattern. The average annual temperature is around 22^0 C, and humidity ranges between 50% and 70%.

The agroecological system is characterized by loam, sandy, and clay soils, with most of the agricultural activities being subsistence and small-scale agriculture. Crops grown in Rongo include maize, beans, groundnuts, green grams, cassava, and some horticultural crops such as mangoes, banana, avocado, pawpaw, and indigenous vegetables with sugarcane being the main cash crop in the area. The crop fields vary (±30 m^2) in dimension. Due to the limited farmlands allocated to individual farmers, intercropping of maize with either beans, cowpea, or groundnuts is the main farming practice in the area.

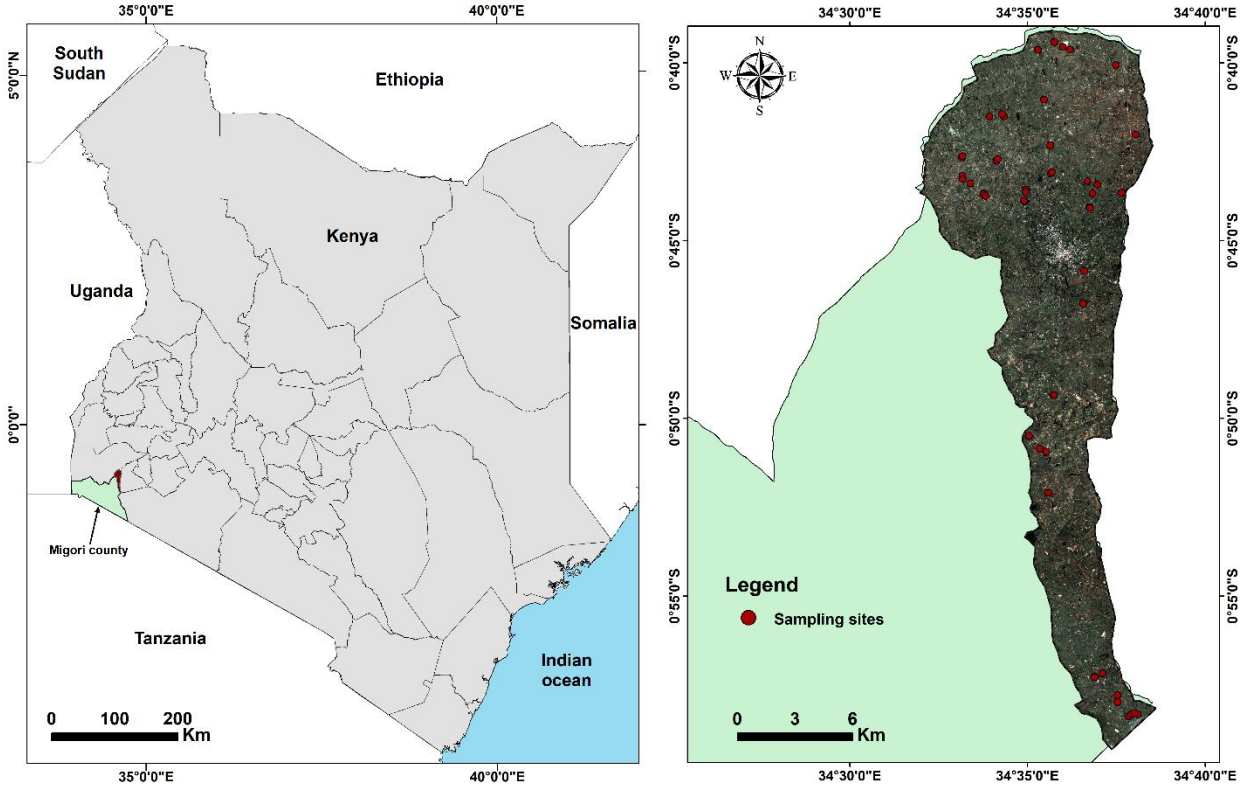


Figure 3.1: Domain and geographic location of the Rongo subcounty in Kenya and Migori county. The background is a true color RGB (3, 2, 1) Sentinel-2 image of the 13th of December 2017.

3.2.2 Data collection

Field surveys were conducted between the 12th and the 16th of December 2017. This period coincided with the peak Striga flowering phase and maize reproductive stages R1-R6. Reference data were gathered within croplands. In the study area, the cropland system was mainly mono and mixed maize cropping. A stratified random sampling approach, using the administrative ward boundaries as strata was followed to determine areas with Striga the presence or absence reference data. A handheld global positioning system (GPS) instrument with an accuracy margin of ±3 m was used to locate the reference data. The 'presence' or 'absence' Striga occurrence data at the center of the field were collected to avoid the influence of the edge effect. Each plot in the sampled fields was geo-tagged using photographs taken from the center and the four cardinal directions for supplementary analysis of the cropping systems and crop age. The other landcover components in the study area i.e. built-up area, bare patches, water, and natural vegetation reference data were collected by digitizing regions of interest (ROI) using visual interpretations of very high spatial resolution imagery within Google earth® as described by Chemura et al., (2017b) and summarized

in Table 3.1.

Table 3.1: Landcover classes used in the classification analysis for both PlanetScope and Sentinel-2 images

Class	Class ID	Description
Bare patches	1	Surfaces without vegetation
Built-up	2	Human-made constructions
Natural vegetation	3	Wood vegetation and grasslands
Non-Striga fields	4	Crop fields without Striga
Striga infested fields	5	Crop fields infested with Striga
Water	6	Water bodies

3.2.3 Image acquisition and pre-processing

Images of PlanetScope (16th December 2017) and Sentinel-2 (13th December 2017) were used in the analysis. PlanetScope was provided by Planet Labs Inc. https://www.planet.com and delivered as an analytic 4-band product of the visible and near-infrared (VNIR) (Planet, 2018). The data was received as an orthoscene product (level 3B), which relates to top of atmosphere (TOA) radiance suitable for an analytic and visual application. The scenes were provided already orthorectified to < 10 m root mean square error (RMSE) position accuracy and projected to UTM/WGS84 cartographic projection. The scaled radiance was converted to TOA reflectance using a Planet Labs python guide (https://developers.planet.com/tutorials/convert-planetscope-imagery-from-radiance-to-reflectance/) and the corresponding reflectance calibration coefficients associated with each band from the metadata.xml files.

The scenes were mosaiced to cover the entire study area. Similarly, Sentinel-2 data, processing level 1C was freely downloaded from the Copernicus data download platform https://scihub.copernicus.eu/dhus/#/home within the granule T36MXE. Level 1C data from Sentinels are provided as TOA reflectance, already orthorectified in cartographic geometry in tiles of 100 km^2, UTM/WGS84 projection. These were converted to atmospherically corrected, surface reflectance level 2A using the SEN2COR in SNAP v6.0 software run using the default parameter settings. The choice of images was informed by the alignment with the same period the field reference data was collected, the availability from the sensor's archive, and the corresponding cloud cover. All the images were acquired on days of low cloud cover (< 5%) and during the period when vegetation was in full vigor.

3.2.4 Broadband vegetation indices

Seven vegetation indices were computed from PlanetScope and Sentinel-2 images (Table 3.2). Visible and near-infrared (VNIR) bands were used to derive vegetation indices. The indices were computed, using the freely available SNAP v6.0 software. These indices were chosen founded on (1) their ability to reduce soil background effects, (2) their strength to separate between flowering and nonflowering plants, and (3) capability to enhance vegetation greenness (Xue and Su, 2017). The normalized difference vegetation index (NDVI) has been successfully used to estimate biomass and crop yields (Chemura et al., 2017c; Dube et al., 2015). Also, the usage of NDVI, as opposed to the classification of raw bands, is recommended for discrimination of vegetation from other non-photosynthetic classes (Bannari et al., 1995; Matongera et al., 2017). Ratio vegetation index (RVI) is based on the principle that leaves absorb red more than near infrared (NIR) (Bannari et al., 1995; Xue and Su, 2017).

Two indices that correct for the influence of the soil background were used namely; modified secondary soil adjusted vegetation index 2 (MSAVI2) and soil adjusted vegetation index (SAVI). MSAVI2 does not rely on the soil line to eliminate the soil influence but introduces a function rather than a constant 'L' value, which is the modification of SAVI and MSAVI. SAVI minimizes the influence of soil brightness by introducing the soil condition 'L'. The value of L ranges between 0 to 1, where values of L are close to zero, SAVI equals NDVI (Royimani et al., 2019). In this chapter, L = 0.5 was used since this is common practice for a wide range of environmental conditions (Xue and Su, 2017).

The atmospheric resistant vegetation index (ARVI) is based on the correction of the atmospheric influences through the elimination of the effects of atmospheric aerosols using the difference between the blue and red bands. The difference vegetation index (DVI), which subtracts the red band from the near-infrared band, has been useful for vegetation monitoring, however, it is very sensitive to changes in the soil background (Bannari et al., 1995). The infrared percentage vegetation index (IPVI) is recommended for vegetation mapping as it is sensitive to vegetation cover and vegetation biomass (Royimani et al., 2019).

Table 3.2: Selected vegetation indices for the discrimination of Striga infested crop fields from other land cover classes: where RB in the atmospheric resistance vegetation index is the difference between the red band and blue band

Vegetation index	Equation	Reference
Normalized difference vegetation index (NDVI)	$\frac{NIR - R}{NIR + R}$	(Abdel-Rahman et al., 2016)
Infrared percentage vegetation index (IPVI)	$\frac{NIR}{NIR + R}$	(Royimani et al., 2019)
Simple ratio/ ratio vegetation index (RVI)	$\frac{NIR}{R}$	(Evangelista et al., 2009)
Soil adjusted vegetation index (SAVI)	$\frac{NIR - R}{NIR + R + L}(1 + L)$	(Hadjimitsis et al., 2010)
Modified secondary soil adjusted vegetation index (MSAVI2)	0.5 X [2 NIR + 1 – $\sqrt{2(NIR + 1)2 - 8\,(NIR - R)}$]	(Xue and Su, 2017)
Atmospheric resistant vegetation index (ARVI)	$\frac{NIR - RB}{NIR + RB}$	(Xue and Su, 2017)
Difference vegetation index (DVI)	$NIR - R$	(Dube et al., 2015)

3.2.5 Guided regularized random forest

GRRF was used to select the most important bands and indices for the prediction of the six land cover classes (Table 3.1). Many researchers have used the random forest (RF) as a dimensionality reduction tool to reduce data redundancy within explanatory variables (Dube et al., 2014; Han et al., 2017). However, research has shown disadvantages for the use of the random forest as a tool to measure variable importance, as well as a variable selection method (Mureriwa et al., 2016). The package 'RRF' in R software was used to perform the variable importance analysis and classification (Deng, 2013; R Core Team, 2019). A gamma value of 0.6 was used in the 'CoefReg' function to determine the limit used to restrain the explanatory variables to the minimum without compromising on the capacity of the algorithm to predict accurately. Compared to the standard algorithms, the regularized framework significantly reduces the training time by building a single model (Deng and Runger, 2013). The GRRF uses a similar concept of the RF model but uses the importance scores generated from RF to guide the variable selection process (Adam et al., 2017; Mureriwa et al., 2016).

The importance value of a variable in RF is attained through the "Gini index" over all nodes, across all trees generated and the feature is used to split measure the pureness of the feature at every node to facilitate the voting process (Breiman, 2002). Hence, GRRF provides the precise variables that are most suitable for predicting the feature, out of the multiple features (Deng, 2013). Like the standard algorithm of the random forest for classification, the regularized random forest algorithm is an ensemble machine learning approach, which combines a large set of decision trees (Breiman, 2001).

The approach provides randomness by the bagging operation to fit numerous decision trees on random subsamples (Abdel-Rahman et al., 2013a; Adam et al., 2017). The multiple classification trees then vote by plurality on the correct classification (Royimani et al., 2019). Most studies rely on the out of bag (OOB) accuracy for the model evaluation, however, the current chapter used a 10 x 10-fold repeated cross-validation technique, set to train and validate the performance of the models (Immitzer et al., 2016). The three parameters, *mtry* and *ntree* = 500, *flagReg* = 1, were used as default using the *tune length* of 3. According to (Abdel-Rahman et al., 2013b) increasing *ntree* beyond 500 does not show a pronounced effect on RMSE.

3.2.6 Accuracy assessment

Data generated from the digitized training areas were split into two datasets 70% and 30%, which is 'training' and 'testing' respectively, as per the standard machine learning evaluation criteria (Dube et al., 2014; Kyalo et al., 2017). Model testing data was used in the construction of the confusion matrices for the validation of the performance of the GRRF model. The overall accuracy (OA), user's accuracy (UA), and the producer's accuracy (PA) were used together with Kappa statistics to evaluate the performance of the classifier and the sensors.

3.3 Results and discussion

3.3.1 Variable selection using guided regularized random forest

Results from both PlanetScope and Sentinel-2 imagery show that the visible bands (blue, green, and red) were critical for separating the two Striga classes (Striga and non-Striga infested fields) from the four land cover classes (bare patches, built-up, natural vegetation, and water). The GRRF was able to determine only six key variables from Sentinel-2 and five variables for prediction using

PlanetScope. Amongst the vegetation indices, only ARVI and NDVI for Sentinel-2 were selected by the GRRF whilst ARVI and IPVI were selected for the PlanetScope (Figure 3.2).

These results prove the strength and importance of the raw bands of these two sensors in discriminating vegetation from manmade features. However, the indices were also crucial for highlighting plant inherent properties associated with leaf properties, such as chlorophyll content and vigor. Vegetation indices have specific expressions and ratios, which can represent green vegetation properties better than using individual bands (Baloloy et al., 2018). NDVI has been tested in various related research and has proven a valuable index correlated to biophysical plant characteristics such as chlorophyll content and LAI (Baugh and Groeneveld, 2006). Similarly, ARVI is a modified version of NDVI that overcomes the influence of the atmosphere.

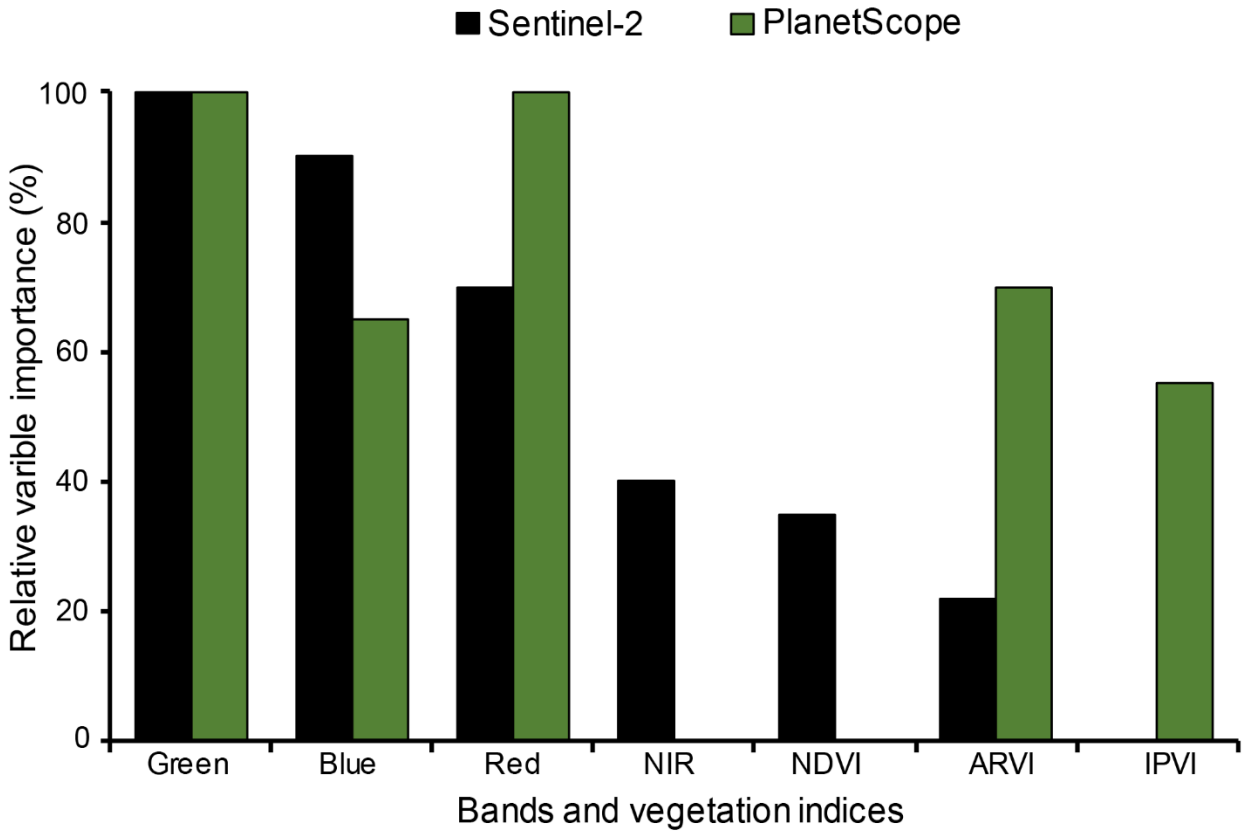


Figure 3.2: Variable importance percentage for both Sentinel-2 and PlanetScope derived metrics using the GRRF variable selection measure

3.3.2 Mapping accuracy assessment

The use of Sentinel-2 selected variables (red, green, blue, near-infrared, and ARVI) produced an overall accuracy of 87% and Kappa of 0.82 in detecting Striga occurrence in maize fields (Figure 3.3). The results were slightly lower (-5% deviation) compared to those obtained using PlanetScope. The results show that PlanetScope selected variables (red, green, blue, ARVI, and IPVI) produced the slightly higher land use/land cover (LULC) mapping results (92%) and Kappa of 0.89. The use of a few systematically selected number of effective bands has shown that it is possible to attain and exceed the classification accuracy of the entire waveband dataset (Cao et al., 2018). This also reduces the redundancy produced by correlated variables. The GRRF process decreases the multidimensionality of the variable data without compromising key information relevant to the features (Adam et al., 2017; Deng, 2013; Mureriwa et al., 2016)

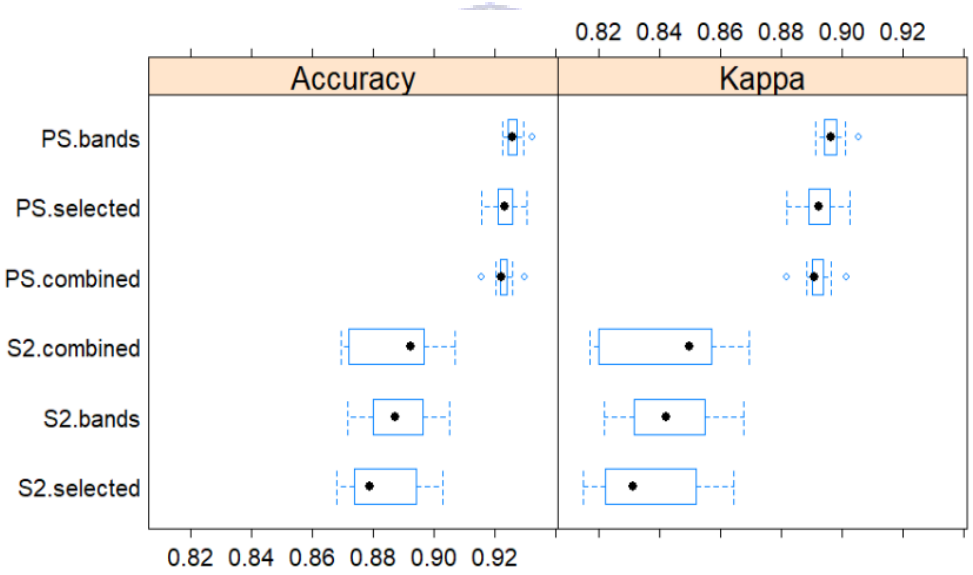


Figure 3.3: Model prediction performances using accuracy and Kappa for bands only variables (*.bands), combined indices and bands(*.combined) and guided regularized random forest (GRRF) selected variables (*.selected) using PlanetScope (PS) and Sentinel-2 (S2) imagery.

Although Sentinel-2 and PlanetScope both performed very well in predicting other classes, the accuracy results for Striga and non-Striga fields were low 66% and 72% respectively (Table 3.3). Based on the accuracy produced by the single scene of PlanetScope, these results demonstrate the ±175 CubeSats constellation is a relevant source of spatial data capable of consistently monitoring the presence of Striga and other land cover classes. However, due to the limiting costs of acquisition of the PlanetScope imagery, Sentinel-2 produced acceptable accuracies, which are

essential for monitoring Striga occurrence.

These classification errors are related to the split of cropland classes into Striga and non-Striga crop fields, which have similar spectral responses. The split was necessary to differentiate maize fields infested with Striga and non-infested fields. The classification error margins could also be attributed to the complexity produced by the miniature (< 30 m^2) heterogeneous fields mainly with intercropped plants that characterize the study area. Since other non-cropland classes could be classified with very high accuracy these results could potentially be improved by masking other classes except cropland and predicting Striga within the cropland class.

Table 3.3: Summarized confusion matrices and classification accuracies, overall accuracy (OA), producer's accuracy (PA), user's accuracy (UA) and Kappa statistics using the guided regularized random forest image classification of PlanetScope and Sentinel-2 comparing combined variables (all bands and all indices) and GRRF selected variables (fewer optimum bands and indices selected with the GRRF algorithm.

Class	PlanetScope				Sentinel-2			
	Combine variables		GRRF selected variables		Combine variables		GRRF selected variables	
	PA	UA	PA	UA	PA	UA	PA	UA
Bare	92	95	93	96	88	92	88	91
Built-up	93	87	93	88	92	85	90	85
Natural Vegetation	98	99	98	99	96	95	94	96
Non-Striga Fields	70	70	67	70	51	52	53	56
Striga Fields	65	72	67	68	53	66	54	52
Water	89	94	91	91	100	97	93	93
OA	92%		92%		88%		87%	
Kappa	0.89		0.89		0.83		0.82	

3.3.3 Striga mapping

The final thematic maps produced via the GRRF algorithm are shown in Figure 3.4. Both images show that much of the Striga infestation is in the north and north-west of Rongo whilst low on the south. The south is mainly characterized by sugarcane farms, which are less susceptible to Striga infestation whereas the north and northwest regions are predominantly maize fields. There are patches of Striga infested fields within the built-up area surrounding Rongo town. Most of the urban dwellers practice urban agriculture, hence the presence of Striga within the vicinity of the central business district.

Both images ably predicted for the built-up, water, natural vegetation, and bare areas. The ability of these two sensors and the capability of the GRRF model to separate these classes is commendable. Interestingly, there were able to classify the river quite distinctively from the rest

of all the surrounding classes. Apart from the capacity to map Striga these results confirm the potential of Sentinel-2 to monitor LCLU changes over large landscape scales.

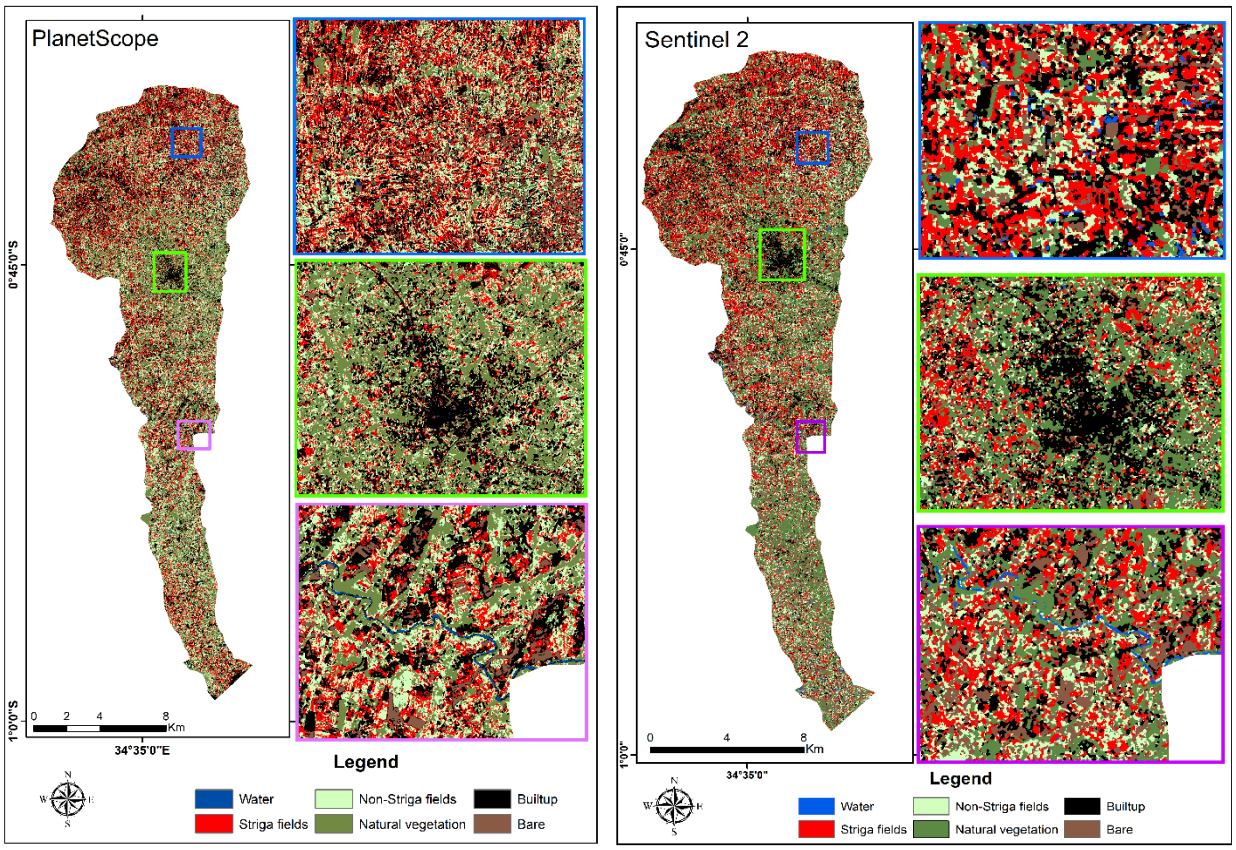


Figure 3.4: Land use and land cover map of Rongo showing the distribution of Striga infested fields from the GRRF classification of PlanetScope and Sentinel-2.

3.4 Conclusions

From the results of this chapter, it is concluded that:

• The Sentinel-2 sensor detected Striga infested maize fields with high accuracy, almost like those derived using PlanetScope metrics.

• The GRRF feature selection shortlisted the red, green, blue, NIR, and ARVI as the most important variables for predicting Striga occurrence, with an overall classification accuracy of 87% (-5% deviation from PlanetScope selected variable performance.

Overall, the findings of this work provide baseline information appropriate to devise adaptive weed management strategies in sub-Saharan Africa. However, since the results proved that there were errors of commission between Striga and non-Striga fields caused by similarities in the spectral behaviors of the crops, this chapter recommends that a hierarchical approach be tested by masking out croplands and using subpixel-unmixing technics to improve the detection ability of Striga occurrence. Since Striga swiftly invades agroecological landscapes, we also suggest landscape-scale monitoring of Striga using multi-date Sentinel-2 data. Long-term remote sensing applications are effective in defining spatial trends and the evolution of the distribution of invasive weeds over time. Furthermore, future studies must consider the presence of several components crops within intercropped systems that could potentially impact the satellite signatures obtained within croplands.

Chapter 4

Landscape-scale hierarchical cropland and subpixel Striga detection

Photo: courtesy of Bester Tawona Mudereri 2018

This chapter is based on:

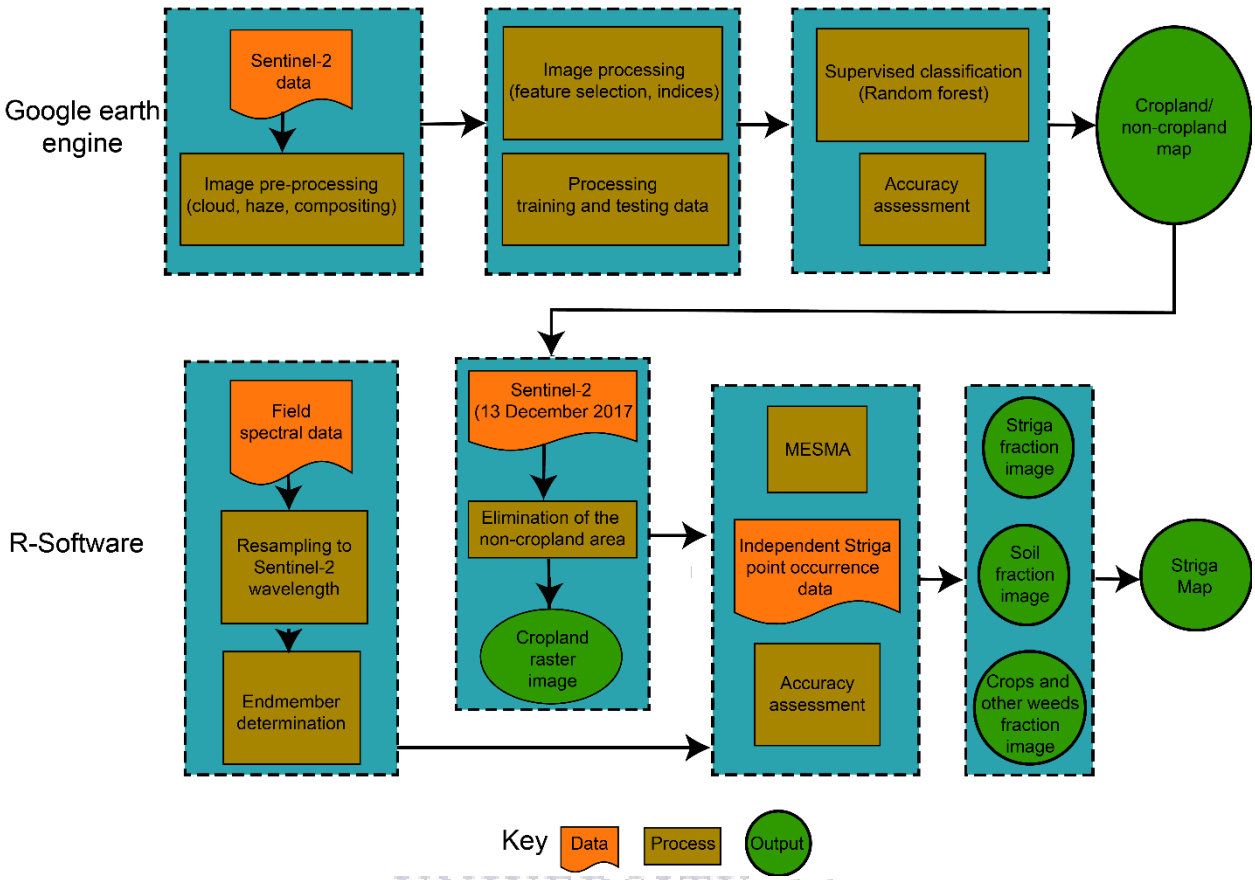
Mudereri, B.T., Adel-Rahman, E.M., Dube, T., Niassy, S., Khan, Z., Tonnang H.E.Z., Landmann, T., (2020). *A two-step approach for detecting Striga in a complex agroecological system using Sentinel-2 data.* Science of the Total Environment. https://doi.org/10.1016/j.scitotenv.2020.143151

Abstract

Information on weed occurrence within croplands is vital but is often unavailable to support weeding practices and improve cropland productivity assessments. Until now, few studies have been conducted to estimate weed abundances within agroecological systems from spaceborne images over wide-area landscapes, particularly for the genus Striga. Therefore, this chapter attempts to increase the detection capacity of Striga at subpixel size using spaceborne high-resolution imagery. A two-step classification approach was used to detect Striga (*Striga hermonthica*) weed occurrence within croplands in Rongo, Kenya. Firstly, multidate and multiyear Sentinel-2 data (2017 to 2018) were utilized to map cropland and non-cropland areas using a random forest classification algorithm within the Google Earth Engine. The non-cropland class was thereafter masked out from a single date Sentinel-2 image of the 13th of December 2017. The remaining cropland area was then used in a subpixel multiple endmember spectral mixture analysis (MESMA) to detect Striga occurrence and infestation using endmembers (EMs) obtained from the in-situ hyperspectral data. The gathered in-situ hyperspectral data were resampled to the spectral waveband configurations of Sentinel-2 and three representative EMs were inferred, namely: (1) Striga, (2) crop and other weeds, and (3) soil. Overall classification accuracies of 88% and 78% for the pixel-based cropland mapping and subpixel Striga detection were achieved, respectively. Furthermore, an F-score (0.84) and a root mean square error (0.0075) showed that the MESMA subpixel algorithm provides plausible results for predicting the relative abundance of Striga within each Sentinel-2 pixel at a landscape scale. The capability of MESMA together with a cropland classification hierarchical approach was thus proven to be well suited for Striga detection in a heterogeneous agroecological system. These results can be used to guide in the adaptation, mitigation, and remediation of already infested areas, thereby avoiding further Striga infestation of new croplands.

Keywords: endmember selection; Google Earth Engine; spectral mixture modeling; invasive weeds; Africa

Graphic abstract



4.1 Introduction

There is a great need for precise information on the occurrence of harmful weeds within croplands to enhance scaling out of interventions and mitigation of crop production constraints, that may impede agricultural products' supply chains (Mutanga et al., 2017). Striga is a parasitic weed of global economic importance that penetrates the roots of the cereal crops after germination and withdraws nutrients and water thereby suffocating and out-competing their cereal crop hosts (Hassanali et al., 2008). Consequently, the productivity of economically fundamental host crops like maize, sorghum, and rice can deteriorate considerably, causing extensive grain yield reductions and unprecedented socioeconomic impacts (Ejeta and Gressel, 2007). Striga weed detection using remote sensing techniques such as the in-situ hyperspectral methods (Mudereri et al., 2020a) or high-to-medium spatial resolution satellite data (i.e. PlanetScope and Sentinel-2) are essential tools to help identify weed distributional patterns and priority areas at plot and/or field scales (Mudereri et al., 2019a). Spectral responses in small-scale agroecological systems, typical for Africa, exhibit a large intra- and inter-field variability which results in detection errors when discriminating between co-occurring crops and weeds (i.e. maize and Striga) (Mudereri et al., 2019a). Specifically, the heterogeneity, fragmentation, and complex crop cycles (Xiong et al., 2017a) of African cropping systems hinder the detection of hidden spectral properties of the Striga weed, when pixel-based detection approaches are employed.

On the other hand, the complex structure of remotely sensed spectral information often shadows analytical spectral characteristics and obstructs the comprehensive characterization of targeted constraints within the croplands (Somers et al., 2011). However, studies have shown that the mixed pixels (e.g. crop and Striga) can be decomposed using subpixel classification methods i.e. spectral mixture analysis (SMA: Powell et al., 2007). Numerous variants of SMA to model sub-pixel fractions have been used by other studies such as the simple spectral mixture analysis (sSMA: Somers et al., 2011), Monte Carlo spectral mixture model (AutoMCU: Asner and Lobell, 2000), Bayesian spectral mixture analysis (BSMA: Song, 2005), linear mixture analysis (Heinz and Chang, 2001) and multiple endmember spectral mixture analysis (MESMA: Roberts et al., 1998). These methods are all set and operationalized by: (1) determining how many pure spectra known as endmembers (EMs) are present in an image data, (2) identifying the biophysical nature of each of those EMs within a pixel, and (3) estimating the fractional abundances of each EM in a pixel

(Somers et al., 2011).

MESMA is, by far, the most extensively used SMA method that yields a relatively low classification error and provides high detection accuracies when compared to the other SMA methods (Degerickx et al., 2019). The MESMA approach has been used in a wide range of applications including characterization of urban environments (Franke et al., 2009), mapping fire (Quintano et al., 2013), plant species mapping (Roberts et al., 1998), mapping of marshes (Li et al., 2005; Rosso et al., 2005) and classification of agricultural agroecosystems (Njenga, 2016), among others. However, no, or few studies have yet to use MESMA-based fraction images to estimate Striga (*Striga hermonthica*) occurrence or severity. Thus, this study is the first attempt to test the efficacy of MESMA in estimating and mapping Striga occurrence and fractions within croplands in Africa using a test site in Kenya.

To effectively perform a subpixel classification method, Degerickx et al. (2019) suggested using *a-priori* classification or segmentation along with SMA approaches to constrain EMs to certain segments of the image to ensure a reduction in the spectral confusion. To detect invasive weeds like Striga that infest croplands using image data and subpixel SMA, one would need to first classify the croplands, before employing any SMA like MESMA. Mapping of croplands using a singledate or multidate multispectral images and machine learning classification algorithms is well documented in the literature (Belgiu and Csillik, 2018; Gumma et al., 2019). Among the machine learning algorithms used in most of these studies, the random forest (RF) performs relatively better than the other methods for delineating croplands (Belgiu and Csillik, 2018; Immitzer et al., 2016).

In this chapter, the Google Earth Engine (GEE) was used to execute accurate and localized cropland mapping. The strength and versatility of the GEE and the subpixel MESMA to manipulate Sentinel-2 multitemporal data was exploited to execute a two-step cropland and Striga weed classification approach. This approach was necessitated by the need to constrain the classification of Striga occurrence within the cropland area to considerably reduce the number of EMs. Various studies in agriculture monitoring and management have used the GEE platform to add and curate their data while exploiting Google's cloud resources to undertake all the processing procedures (Kelley et al., 2018; Landmann et al., 2019). The objective of this chapter was thus to advance the feasibility of Striga weed detection using the subpixel MESMA within croplands

derived using the high resolution (10 m) freely available multispectral Sentinel-2 images and their respectively derived vegetation indices (VIs) in the small-holder farming system.

4.2 Study area

The study was conducted in the Rongo sub-county (coordinates: 0^0 39`12``S; $34^0$ 35`40``E and 0^0 59`16``S; $34^0$ 37`21``E), covering an area of ~ 213 km^2. Rongo is in the Migori county of western Kenya, occurring at an altitude of 1 470 m (Figure 4.1). A tropical and subtropical climate characterizes the study area with a bimodal rainfall distribution. Average annual rainfall of 1 600 mm across the two rainy seasons i.e. during 'short rains' season spanning November to January and 'long rains' season occurring between March and June characterizes the study area. The study area experiences relative humidity ranges of between 50% and 70% throughout the year, while the annual average temperature is 20.6^0C.

In Rongo, the savanna grassland biome dominates the natural ecosystems, often flanked with deciduous and exotic forest vegetation. On the other hand, the agricultural cropping system in the study area is mainly mixed small-scale farms, with an average field size of 0.1 ha. Maize, bean, cassava, green gram, groundnut, and fruits such as avocado, banana, pawpaw, mango, and indigenous vegetables as food crops are common in Rongo for predominantly subsistence purposes. Sugarcane grown mainly in the southern region of the study area is the main cash crop, among others. Rainfall variability, insect pests, and the invasive Striga weed are the major crop production constraints in the study area.

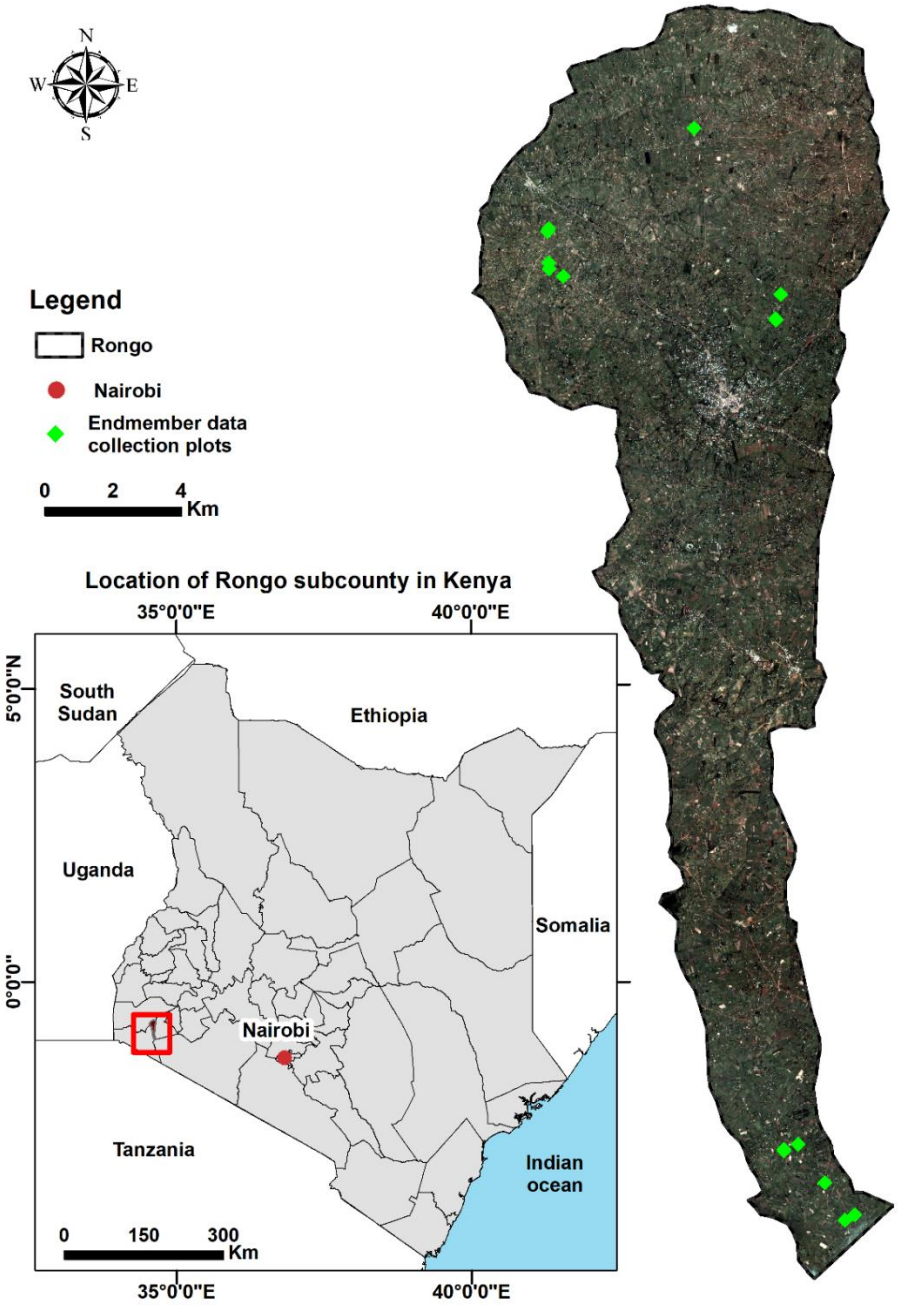


Figure 4.1: The location of Rongo in Kenya. The light green diamonds show the location where the spectral endmembers were extracted. The background layer is the Sentinel-2 image of the 13th of December 2017 displayed in the true color using the red, green, and blue (RGB) band combination as Sentinel-2 bands 4,3, and 2, respectively.

4.3 Methods

4.3.1 Two-step hierarchical classification approach

The proposed methodology uses a two-step hierarchical approach to detect the occurrence of Striga i.e. (1) Sentinel-2 time-series composites to detect and distinguish cropland and non-cropland over

a series of agricultural growing seasons using the random forest classifier in a semi-automatic approach in GEE and (2) spectral unmixing of the derived cropland using a single-date (13th December 2017) Sentinel-2 image and MESMA in R-software to assess the occurrence of Striga. The generalized workflow of the classification approach used in this study is shown in Figure 4.2. The details of each step are thereafter described in the subsequent sub-sections.

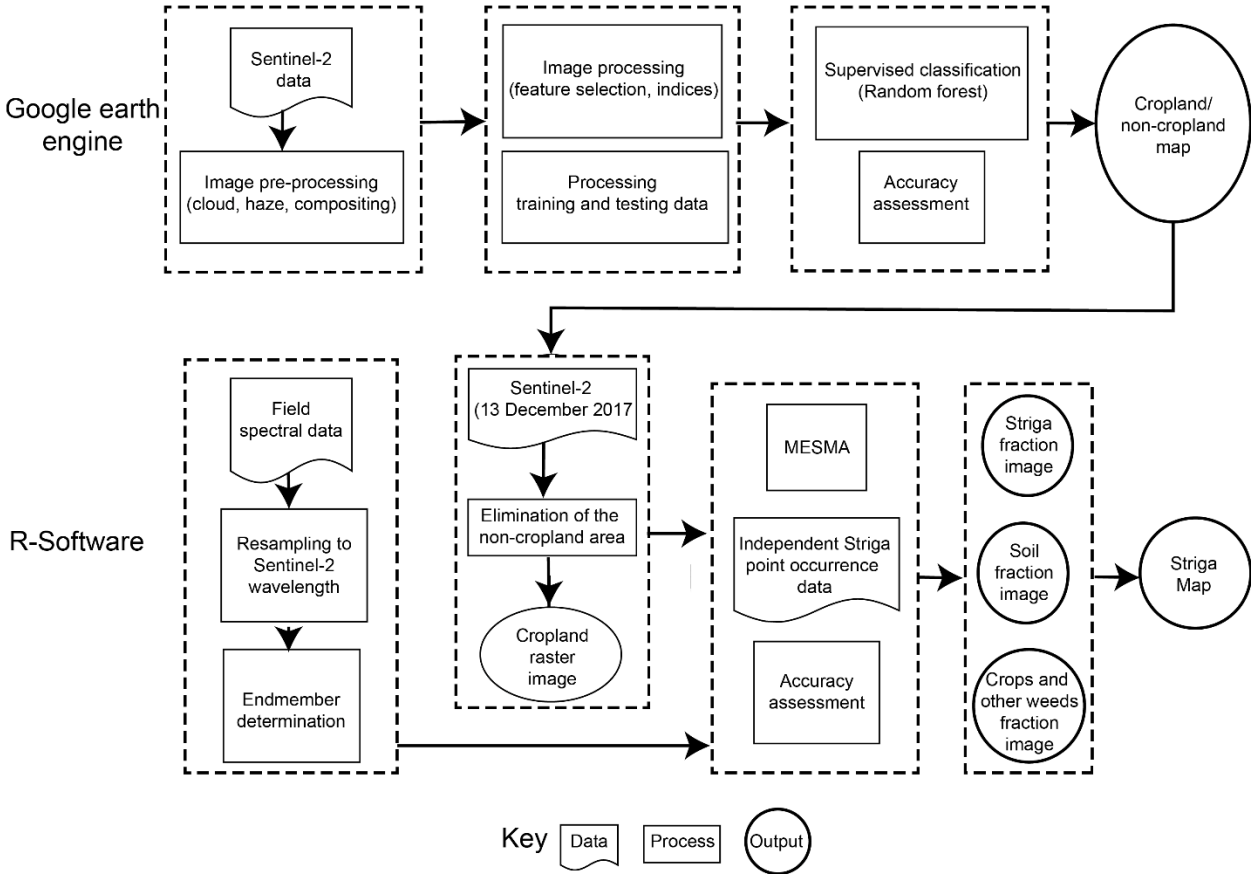


Figure 4.2: The general workflow of the two-step hierarchical approach for detecting Striga weed

4.3.2 Striga and land use/ cover (LULC) field data collection

Striga 'presence' (n = 52) or 'absence' (n = 20), together with cropland data were obtained from field surveys conducted between the 12th and 16th of December 2017 which coincided with the peak Striga period and the maximum phenological stage of the maize in the study area. A global positioning system (GPS) instrument (±3 m accuracy) was used to locate the field Striga reference data. Also, reference data on different land use and land cover (LULC) classes i.e. bare land (including fallow and abandoned cropland), built-up area, natural vegetation (forests, shrubland, and grasslands), and water were collected through digitizing regions of interest (ROI) on Google

Earth® (Chemura et al., 2017b) and are summarized in Table 4.1. The administrative ward boundaries were used as strata in a stratified random sampling approach, to determine areas of Striga 'presence' and 'absence' and other LULC classes and reduce sampling bias.

The Striga and LULC classes were arranged into two main classes for classification i.e. cropland and non-cropland to create a crop mask. A total of 260 sample polygons consisting of 5 890 pixels for the two classes were derived and divided into 77% for image classification training and 23% for testing the accuracy of the classification (Table 4.1). To perform a pixel-based classification, the sample points at a pixel scale were organized in a Google fusion table and retrieved in the GEE, then the corresponding input sample pixel values were extracted at the Sentinel-2 image pixel resolution (i.e. 10 m).

Table 4.1: Striga and Land use/ cover (LULC) classes and samples sizes used in the classification of the cropland and Striga in the Rongo study area, Kenya

Striga/ LULC class	Description	Cropland class	Sample polygon	Train pixels	Test pixels
Striga infested field	Crop fields with Striga	Cropland	52		
Non-Striga infested fields	Crop fields without Striga	Cropland	20		
Other crops	Other non-maize cropland	Cropland	46		
		Total	118	2362	720
Built-up	Human-made constructions	Non-cropland	52		
Bare land	Surfaces without vegetation	Non-cropland	21		
Natural vegetation	Woodland and grassland	Non-cropland	35		
Water	Water bodies	Non-cropland	34		
		Total	142	2158	650
Total			260	4520	1370

4.3.3 Cropland mask generation in GEE

The use of the GEE was employed to map cropland and generate a cropland mask to reduce the propagation of classification error. The freely accessible cropland maps are derived over large areas at coarse spatial resolution hence have many limitations such as inability to account for fragmentation, failure to delineate small farms, and to adequately capture the influence of mixed cropping systems which leads to confusion with other land cover types (Oliphant et al., 2019). As a result, such cropland maps are inadequate and not useful in areas where crop fields are small (< 0.1 ha) and within mixed cropping systems such as those in the small-holder farming sector which characterize the present study area in Kenya.

4.3.3.1 Sentinel-2 data compositing

Image processing and analysis for generating the non-cropland mask was implemented in the GEE. The procedures executed in GEE included image cloud filtering, vegetation indices calculation, normalization for illumination effects (i.e. shade), median compositing, machine learning classifier parameterization, creating the final cropland classification map, and assessing the accuracies of the non-cropland product. A relatively cloud-free layer stack image input was formed by stacking the Sentinel-2 image collection within two consecutive years (i.e. between 01st of January 2017 and 31st of December 2018) and then applying a cloud mask. Pixel-based image compositing is a common procedure to condense the number of pixels with redundant and invalid data due to the atmospheric interference, shadow, or other noise remaining after pre-processing (Bey et al., 2020).

The median compositing method has been reported to be computationally and technically less demanding and provides the best results when compared to other pixel-based image compositing methods such as (i) maximum ratio value, (ii) annual greenness pixel, (iii) best pixel based on the distance to the nearest cloud, and (iv) seasonal greenest pixel (Bey et al., 2020). The median values were used for each band to capture the variability between seasons since the median value is often used instead of the mean to handle outliers and have the most representative pixel value over the entire period.

Five vegetation indices (VIs) namely normalized difference vegetation index (NDVI: Rouse et al., 1974), modified soil adjusted vegetation index (MSAVI: Qi et al., 1994), enhanced vegetation index (EVI: Huete et al., 2002), and two red-edge (RE) vegetation indices i.e. RE-NDVI (Sibanda et al., 2019) and RE-EVI were derived from the bands of Sentinel-2 and added to the spectral band composite. These indices were selected as they are designed to capture the sensitivity of vegetation features while minimizing the influence of the confounding factors such as atmospheric effects and soil reflectance (Bannari et al., 1995; Xue and Su, 2017). These indices have also been reported by other studies as the best in capturing vegetation variabilities over time (Chemura et al., 2017c; Sibanda et al., 2019). These indices are calculated as shown in Equations 4.1 – 4.5:

$$\text{NDVI} = \frac{\rho NIR - \rho Red}{\rho NIR + \rho Red} \quad (4.1)$$

$$\text{MSAVI} = \frac{2\rho NIR+1-\sqrt{(2\rho NIR+1)^2-8\,(\rho NIR-\rho Red)}}{2} \quad (4.2)$$

$$\text{EVI} = 2.5 \text{ x } \frac{\rho NIR-\rho Red}{\rho NIR+6\,X\,\rho Red-7.5\,X\,\rho Blue+1} \quad (4.3)$$

where *ρNIR*, *ρRed ρBlue* in Equations (4.1), (4.2) and (4.3) represent the near-infrared red and blue reflectance values respectively for a given pixel.

$$\text{RE-NDVI} = \frac{\rho NIR-\rho RE2}{\rho NIR+\rho RE2} \quad (4.4)$$

$$\text{RE-EVI} = 2.5 \text{ x } \frac{\rho NIR-\rho RE2}{\rho NIR+6\,X\,\rho RE2-7.5\,X\,\rho Blue+1} \quad (4.5)$$

Where *ρRE2* in Equations (4.4) and (4.5) represent Sentinel-2 band 6 or red-edge 2 (RE2) reflectance values for each given pixel.

4.3.3.2 Random forest classification

The RF algorithm (Breiman, 2001) was used to classify the cropland and non-cropland area in GEE using Sentinel-2 wavebands and the additional five VIs. RF is an ensemble machine learning classifier that has accomplished effective classification and prediction results in many remote sensing studies, including cropland mapping (Belgiu and Csillik, 2018; Oliphant et al., 2019). A detailed explanation of RF and its efficiency in remote sensing is provided in Abdel-Rahman et al. (2014). RF builds compound decision trees (*ntree*) in a machine learning ensemble algorithm approach for classification and regression (Breiman, 2001). For each of the decision trees, a bootstrap sample (2/3 of the original data referred to as "in bag" data) is grown. These randomly sampled subsets are used to split several nodes of these decision trees using random subsets of variables for classification (*mtry*) with the default *mtry* value calculated as the square root of the total number of variables (Abdel-Rahman et al., 2014a). The class with the majority votes from all the generated trees is then provided as the final class prediction (e.g. cropland or non-cropland) (Mudereri et al., 2019b). In this chapter, the default *ntree* and *mtry* settings, *viz*. 500 trees and 3 *mtry* were used to perform the pixel-based RF classification for cropland mapping. On the other hand, the algorithm was trained using the independent 70%-pixel samples (4 520) as shown in Table 4.1. The non-cropland class was masked out from the Sentinel-2 image of the 13th of

December 2017 and the remaining cropland class was then used to perform MESMA as described in section 4.3.3.

4.3.3 Endmember selection and collection

The most vital step to a successful and valid SMA encompasses determining the number, type, and matching spectral signatures of EMs (Song, 2005; Somers et al., 2011). These spectral signatures of the EMs are obtainable either directly from the satellite image data reference spectra or spectral libraries, and portable spectroradiometers (Landmann, 2003; Somers et al., 2011).

Three EM spectra within the cropland class were identified and used in the MESMA namely: (1) flowering Striga (2) crop and other weeds, and (3) soil. These three EMs were collected using canopy-level in-situ hyperspectral data collected with the ASD FieldSpec® Handheld 2™ spectroradiometer (HH2: ASD, 2010) using a 1 m x 1 m quadrat. A full description of how these in-situ hyperspectral data were collected is provided in Mudereri et al. (2020a).

The ASD used for the EM data collection acquires reflected radiation in 325–1075 nm of the electromagnetic spectrum with a built-in 2 nm sampling resolution (ASD, 2010). The collected spectra are automatically resampled to a 1 nm spectral range. A bare optical input at a nadir full conical angle field of view (25^{o}) was employed at 1 m above the maize crop to collect the data. This setting allows the instrument to capture spectra in an area of ~0.5 m in diameter on the target (FieldSpec, 2017). This target area unit was enough for capturing the spectral signal of the three EMs i.e. Striga, crops with other weeds, and soil.

The captured spectra were then filtered using the 'noiseFiltering' function and smoothened using the 'Savitzky–Golay' filter in R software (R Core Team, 2020) using the 'hsdar' package (Lukas et al., 2018). The spectral resampling to the configuration of Sentinel-2 was then conducted using the spectral response function, i.e. 'SpectralResampling' in the 'hsdar' package. The values corresponding to each band were used as the input to the EM values for running the MESMA in R software (R Core Team, 2020).

4.3.4 Striga detection using multiple endmember spectral mixture analysis (MESMA)

The cropland area of the Sentinel-2 image of the 13th of December 2017 that was used in the

MESMA was preprocessed in the GEE as described in section 4.3.3.1 of this chapter. The choice of the image sensing date was influenced by the coincidence with the period of field reference data collection. The MESMA models the spectral data as linear combinations of pure spectra, called EMs while allowing the types and number of EMs to vary on a per-pixel basis (Roberts et al., 1998). Thus, each class is characterized by unique sets of EMs and their fractions. For each pixel, MESMA iteratively runs several candidate models (Franke et al., 2009) and the best model is assessed based on pixel fractions identified, their residuals, and having the smallest root mean square error (RMSE) when compared to the spectral curve of the pixel (Degerickx et al., 2019). The models are thus applied to the entire image on a per-pixel basis. In this study, the MESMA algorithm within the 'RStoolbox' package (Leutner et al., 2019) in R-software (R Core Team, 2020) was used. The 'RStoolbox' package for MESMA uses the non-negative least squares (NNLS) regression through a sequential coordinate-wise algorithm. The MESMA was executed using the default settings as follows: method NNLS, 400 iterations, and tolerance of 1 x 10^{-9}. The outputs from the MESMA algorithm are individual bands representing the estimated pixel fraction and probability of occurrence of each of the tested EMs per pixel (0–1) and RMSE. The base equations of MESMA and RMSE are given by equations 4.6 – 4.8 below:

$$\text{Li}\lambda = \sum_{k=1}^{N} C * L_{k\lambda} + \in_{i\lambda}) \quad (4.6)$$

and

$$\sum_{K=1}^{N} f_{ki} = 1 \quad (4.7)$$

$$\text{RMSE} = \left[\sum_{K=1}^{Z} \frac{(\in_{i\lambda})^2}{Z}\right]^2 \quad (4.8)$$

where a mixed pixel Liλ from location i is modeled as the sum of N EMs, $L_{k\lambda}$, each covering a fraction f_{ki} of the pixel. The residual term $\in_{i\lambda}$ describes the unmodeled portion of the radiance, and the chosen model for each pixel is the one that minimizes the RMSE over the included number of bands used in unmixing the Z.

4.3.5 Accuracy assessment

Accuracy assessment and validation are a key component of any thematic map production, particularly when using remotely sensed data (Dube et al., 2019). The process demands randomly sampled high-quality reference data sets enumerated at suitable spatial and temporal scales. The

classification results of each of the hierarchical steps i.e. cropland and Striga maps were compared to random samples of validation pixels to assess the classification accuracy. Due to the unavailability of percentage coverage reference observations for Striga, soil, and other crops and weeds classes to compare with the predicted class fractional coverage, the accuracy of MESMA results was validated using the standard and traditional classification confusion matrix method. A pixel in the MESMA map was classified as the material (i.e. Striga, soil, or crops with other weeds) with the highest fractional coverage. The soil and crops with other weeds classes were then combined as non-Striga.

Specifically, the performance of the RF classifier used in cropland mapping and MESMA were validated using the accuracy assessment metrics, derived from the respective confusion matrices (i.e. overall accuracy (OA), user's accuracy (UA), producer's accuracy (PA), and the F-score) and RMSE. The F-score evaluates the accuracy of a class using the precision (positively classified values) and recall (the number of relevant instances that were actually classified, also called sensitivity) (Graesser and Ramankutty, 2017; Kyalo et al., 2017), while RMSE measures the spread of residuals around the regression line (line of best fit). For each of the mapping approaches (i.e. cropland mapping in GEE and Striga occurrence mapping using MESMA), the error matrices that provided all the four metrics i.e. OA, UA, PA, and F-score were established.

4.4 Results

4.4.1 Cropland mapping

Approximately 54% (114.3 km^2) of the Rongo area was classified as cropland while the remaining area of 98.7 km^2 (46%) was non-cropland. Remarkably, the forest areas, roads, bare land, built-up areas, and their boundaries were accurately detected. The results indicate that the croplands in Rongo are diverse and irregular in shape, direction, and sizes (Figure 4.3). Due to the inter-annual variability, some of the croplands were fallow or probably abandoned as shown from the Google high-resolution imagery. However, the classified image managed to capture the inter-seasonal variability.

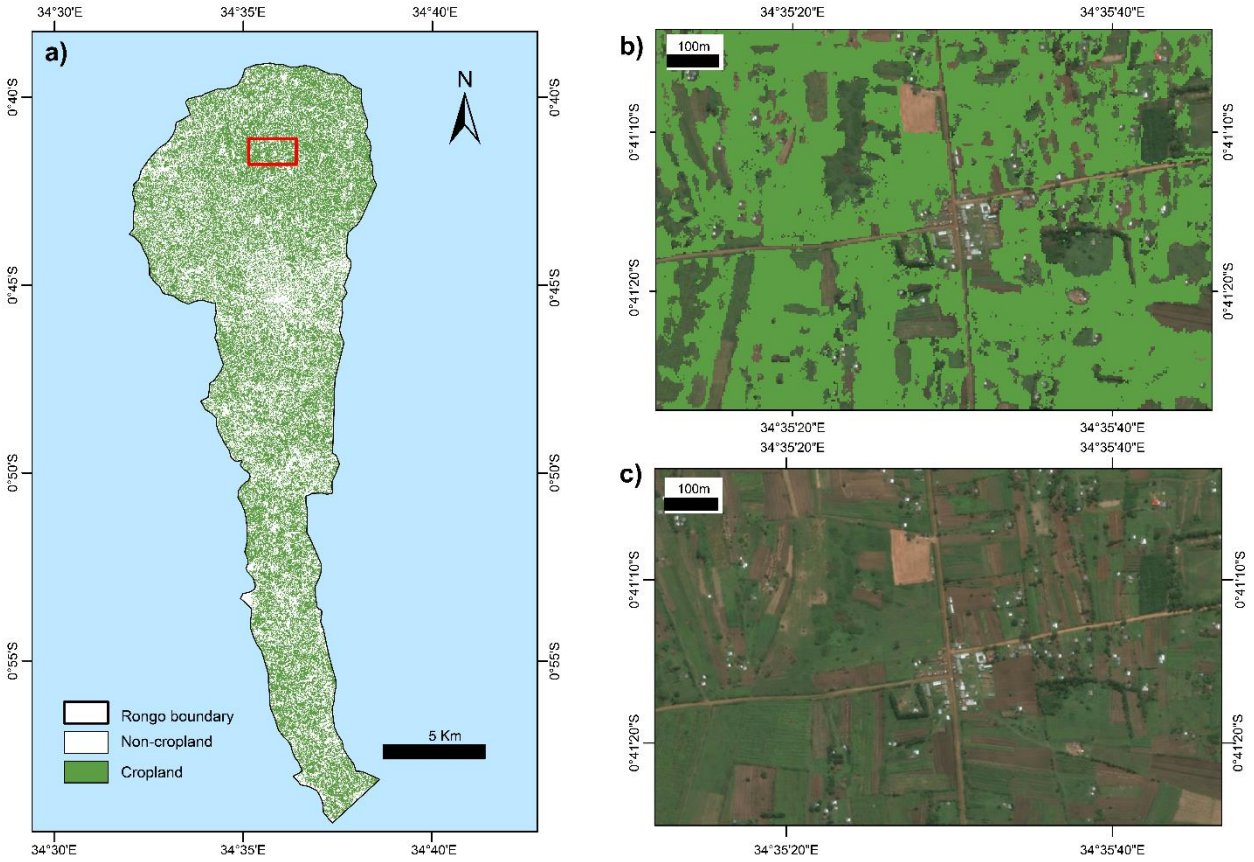


Figure 4.3: Visual comparison of the 10 m cropland map extents derived using Sentinel-2 data and random forest classifier in GEE a) Rongo sub-county, b) a subset cropland extent shows crop fields in light green color overlaid on a Google Earth image, and c) a reference Google Earth image zoomed-out from the red rectangle in a).

Table **4.2** shows that the overall cropland classification accuracy was comparatively high (88%) with F-scores closer to 1 (0.87 and 0.89 for the cropland and non-cropland classes, respectively). Table 4.2 also shows that the error of omission for the cropland class was 4% (PA = 96%) and the error of commission was 21% (UA = 79%), while for non-cropland the error of omission was 17% (PA = 83%) and error of commission omission was 3% (UA = 97%).

Table 4.2: Classification confusion matrix for the cropland and non-cropland classes mapped in the Rongo sub-county using Sentinel-2 data and random forest classifier in GEE.

		Reference data			
		Cropland	Non-cropland	Total	UA (%)
Map data	Cropland	527	143	670	79
	Non-cropland	24	676	700	97
	Total	551	819	1370	
	PA (%)	96	83		
	F-score	0.87	0.89		
	OA				
	88%				

4.4.2 Endmembers used in Striga detection using MESMA

Figure 4.4 demonstrates the variation in spectral responses of the three selected EMs within the 8 resampled Sentinel-2 bands that correspond with the wavelength range of the ASD instrument used. The RE2, RE3, NIR, and NIRn revealed spectrally distinguishable differences (± 0.3) emanating from the different compositions of the classes. However, there is no substantial difference in the EM values within the same class (e.g. Striga EM) across the four stated bands (i.e. RE2, RE3, NIR, and NIRn). On the other hand, the visible bands (blue, green, and red) and the RE1 did not show substantial differences for both within classes and across the three EMs (Figure 4.4).

The Striga EM demonstrated the influence of the Striga flower compaction and color by having the highest EM value in the red and higher than the crops and other weeds EM in RE1. In contrast, the soil EM shows higher reflectance values in the blue and red spectral bands. The crops and other weeds EM dominated the other 5 bands (i.e. green, RE2, NIR, and NIRn).

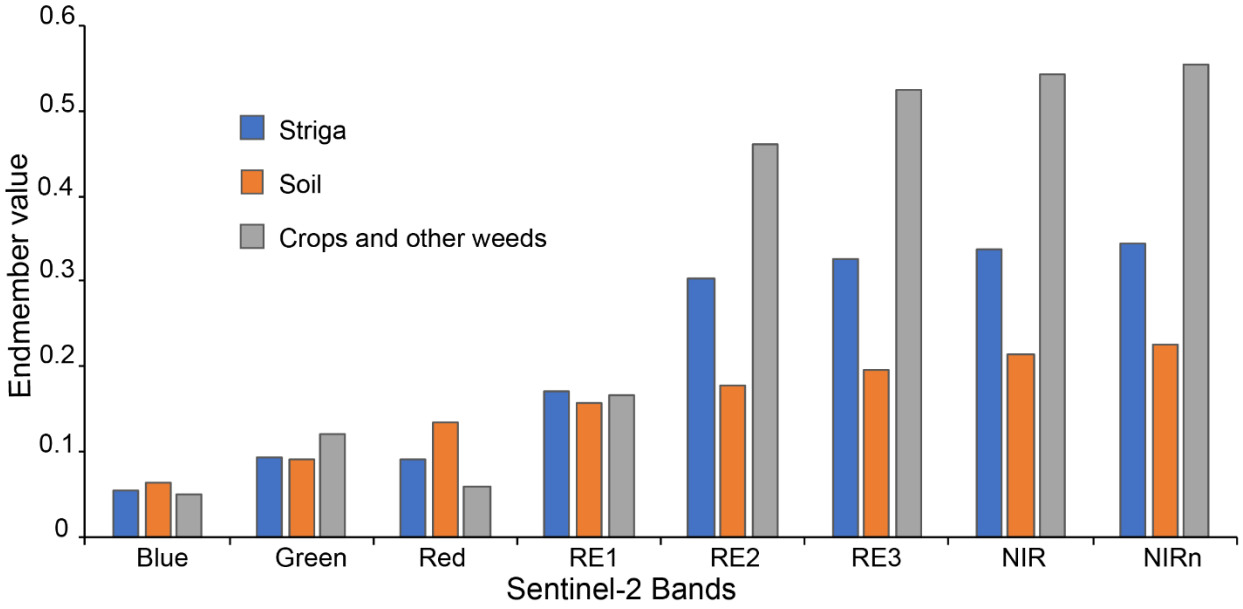


Figure 4.4: Three endmembers and their respective values derived from the resampled eight Sentinel-2 spectral bands used in the multiple endmember spectral mixture analysis (MESMA) for Striga detection. The eight bands correspond to the blue, green, red, RE1, RE2, RE3, NIR, and NIRn waveband areas.

4.4.3 Frequency of pixel fractions of the three endmembers

Figure 4.5 shows the distribution and frequency of pixel fractions across the three EMs tested in this study. The crops and other weeds EM dominated most of the pixels compared to those of Striga and soil EMs in the study area. Although many of the pixels and the larger pixel fractions were occupied by crops and other weeds, results show fewer pixels that had crop and other weeds fractions that are > 0.6. The huge density of Striga pixels was between 0.1– 0.4 pixel fractions with > 250 000 pixels having zero fractions of Striga. On the contrary, there were very few pixels that exhibited close to zero value of the crops and other weeds fractions in the entire study area. The proportion of soil EM, when compared to the other two EMs, was very low as shown by the few pixels representing the soil fraction and very low soil fractions within those few available soil representative pixels. Most of the pixels had a low RMSE (< 0.01) and a maximum RMSE of 0.024 across all the pixels was obtained. Remarkably, this RMSE value represents almost 100% of the whole study area. The average RMSE (0.0075) revealed that there were generally low EM prediction errors using MESMA for the entire study area.

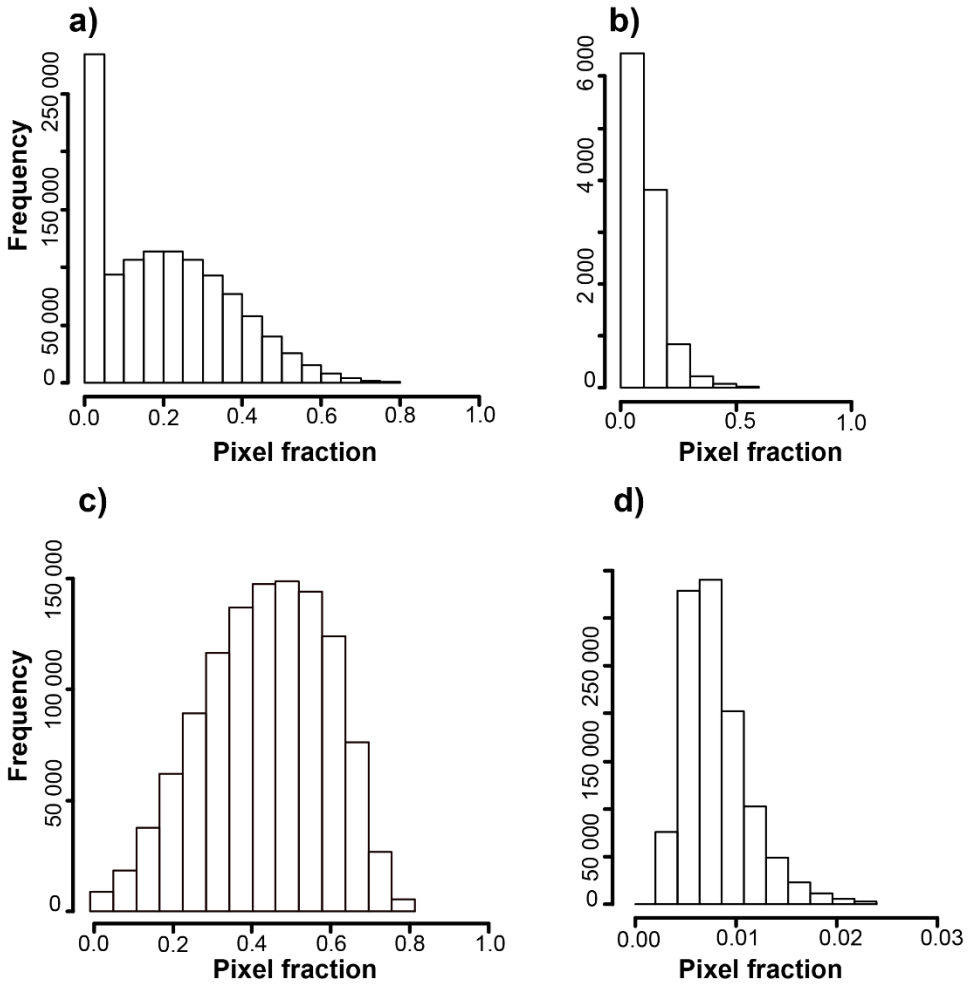


Figure 4.5: Frequency of the pixel fractions of the probability of occurrence values within the endmember fraction images of a) Striga, b) soil, c) crops and other weeds, and d) root mean square error (RMSE)

MESMA generated four fraction images for the three EMs and their corresponding RMSE. Figure 4.6 shows the classification results from the MESMA of the fraction images of the three EMs. The results showed that the RMSE was generally low across the entire study area, except for the central areas of Rongo. Striga occurrence was high in the north and northwestern sides of Rongo as visually shown by the red sub-pixels in Figure 4.6.

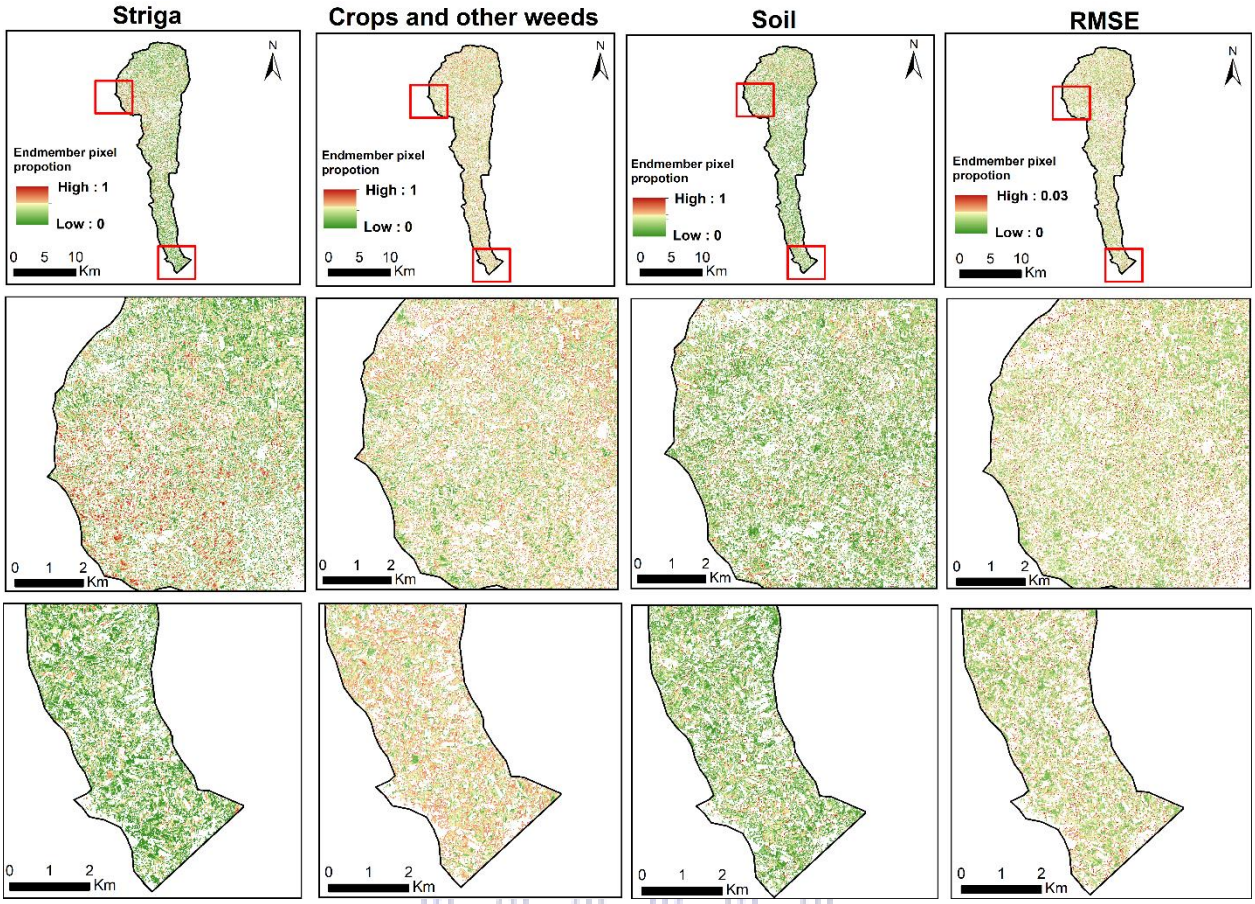


Figure 4.6: Classification results from the MESMA showing the fraction images of the three endmembers from the lowest 0 (green) to the highest proportion of 1 (red) and the root mean square error (RMSE) from lowest 0 (green) to the highest 0.03 (red). The images in the four columns represent from left to right: Striga (extreme left column), crops and other weeds, soil, and the respective RMSE (extreme right column). The zoomed-out images show example regions of high Striga proportions (second row from the top) and very low Striga proportions (third row from the top)

4.4.4 Striga infestation in Rongo sub-country

The false-color image of the three EMs predicting the pixel fraction using MESMA predicted that high proportions of Striga occur in the northwestern region of Rongo (Figure 4.7b), while in the southern region it is generally low (Figure 4.7c). Interestingly, Figure 4.7 shows that some fields were completely free of Striga infestation while in some other fields the entire field (~ 30 m x 30 m) was completely occupied by Striga.

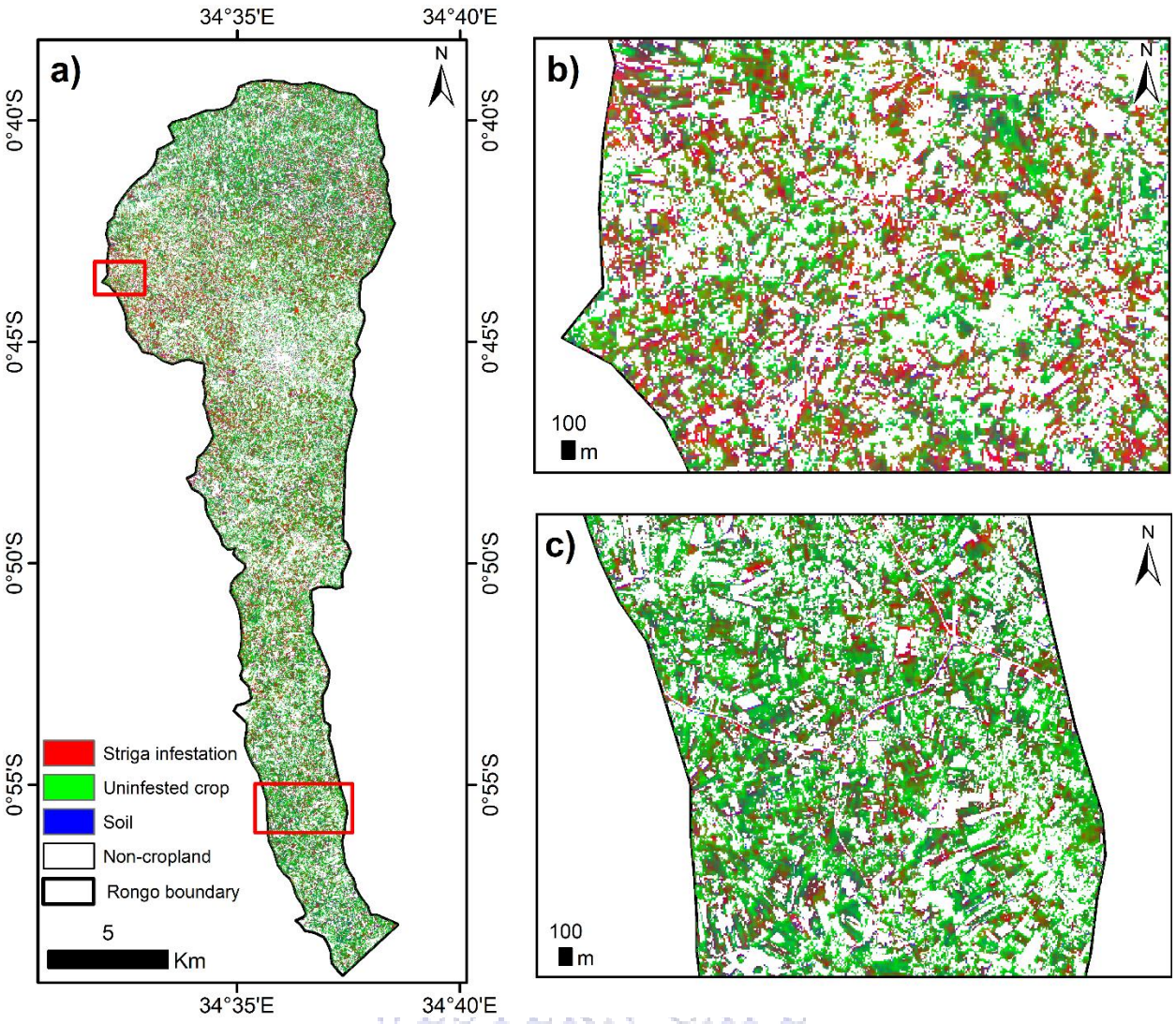


Figure 4.7: A graphical representation of the fractions of the three endmembers i.e. Striga, soil and crops, and other weeds that were derived from MESMA with a) showing the entire study area, b) zoomed section of the study area with the top red box, and c) zoomed-out section of the study area with the bottom red box

4.4.5 Accuracy assessment of Striga detection using MESMA

The Striga classification accuracy assessment focused primarily on two classes of interest i.e. Striga infested and non-infested crop fields. The predicted occurrence of Striga was in good agreement with the occurrence reference points with an OA of 78%, PA of 79% (error of omission of 21%), and UA of 89% (error of commission of 11%) as shown in Table 4.3. An F-score of 0.84 for mapping the Striga infestation showed that the MESMA algorithm successfully predicted the occurrence of Striga and its relative abundance within each pixel (Table 4.3). The PA and UA for the non-Striga fields were lower than those obtained for the Striga-infested fields.

Table 4.3: Classification confusion matrix for the Striga and non-Striga classes detected in the Rongo sub-county using Sentinel-2 data and MESMA

		Reference data			
		Cropland	Non-Striga	Total	UA (%)
Map data	Striga	41	5	46	89
	Non-Striga	11	15	26	58
	Total	52	20	71	
	PA (%)	79	75		
	F-score	0.84	0.65		
OA (%) = 78					

4.5 Discussion

This chapter leveraged the strength and versatility of RF and the robust spatiotemporal resolution of Sentinel-2 imagery in the cloud-based GEE for cropland mapping. Five spectral indices i.e. NDVI, EVI, MSAVI, the red-edge NDVI, the red-edge EVI, and vegetation red-edge bands (bands 5, 6, 7, and 8a) were used for the cropland mapping. Thus, adding the indices, red-edge bands and the spatial resolution of the Sentinel-2 sensor permitted the successful delineation of croplands with high precision, including where the agricultural landscapes are very heterogeneous, fragmented and the fields are small (< 0.1 ha). The use of these indices ensured stable and meaningful comparisons of seasonal and inter-annual changes in vegetation growth and activity thus achieving satisfactory classification results (Belgiu and Csillik, 2018). Other studies have also reported the added-value of the improved Sentinel-2 spectral and spatial resolution for estimating leaf area index (Sibanda et al., 2019), mapping LULC (Forkuor et al., 2018), Striga weed (Mudereri et al., 2019a), and in cropland mapping (Xiong et al., 2017b).

Many recent studies have reported very high accuracies in cropland mapping using Sentinel-2 imagery (Forkuor et al., 2018; Noi and Kappas, 2018). Belgiu and Csillik, (2018) reported a 96.19% OA, using Sentinel-2 in cropland mapping, while Sibanda et al., (2015) confirmed high agreements in quantifying above-ground biomass, using Sentinel-2 data. The differences between the cropland mapping accuracies observed in the present study and those obtained from other studies could have emanated from the heterogeneous nature of the cropping system and differences in the weeding regimes by the diverse small-holder farmers. This heterogeneity results in multiple spectra and texture captured by the sensors. Additionally, most of the farmers in Rongo practice mixed cropping within small field plots (< 0.1 ha) resulting in the diversity of field sizes,

orientation, and shapes. These factors could have caused the 12% error margin realized in the MESMA classification approach used in this study.

Thus, the UA (commission errors) were significantly lower than the PA (omission errors) for cropland mapping. This was mainly because when training the RF algorithm, we aimed to capture many croplands, thus ensuring low omission errors for the cropland class. Thus, some non-croplands were included as croplands, resulting in higher commission errors for the cropland class. In principle, an algorithm must balance PA and UA (Belgiu and Csillik, 2018; Oliphant et al., 2019). This method is similar to the approach used by Oliphant et al., (2019) who also championed reduction of the error of commission by optimizing the RF algorithm, intending to capture as much cropland area as possible. Similarly, this study aimed to collectively capture all croplands including fallow croplands to holistically test the occurrence of the devastating Striga weed. The target to reduce the propagation of the error to the MESMA process was achieved, thus the 4% cropland omission error is plausible given the damage caused by Striga.

The results produced from the accuracy assessment of the MESMA were assuring, considering that the validation reference data used was entirely independent of the MESMA model, unlike in the standard accuracy assessment procedure where the model internally splits the reference data into training and testing samples. The PA produced using an independent point dataset was relatively high showing relatively low errors of omission in the classification. Other studies have also reported high accuracies when using EMs derived using field spectroradiometer dataset (Landmann, 2003).

Although the three EMs used in this chapter produced comparable accuracies, it is still a challenge to concretely locate enough EMs and their variability in different croplands to represent the heterogeneity of croplands and their various growth stages over the entire image (Franke et al., 2009). We observed that spectral data from the same class had varying spectral characteristics perhaps attributable to crop age variations, insect pests, and disease damage, varying management practices, cropping system, or atmospheric effects (Degerickx et al., 2019). This, however, often leads to spectral variations and errors. Thus, it can be concluded that a spectrum that has been tested to adequately model other spectra within the library might not perform with the same precision when tested on a satellite image (Song, 2005). These disparities are often introduced by

fluxes in the brightness caused by bidirectional reflectance that is exacerbated by large viewing angles and broadband wavelengths when the in-situ EMs are upscaled to satellite-based imagery resolutions (Franke et al., 2009; Rosso et al., 2005). In comparison to other studies, the brightness-effects did not appear to have caused a major impact on the Striga detection results, since MESMA automatically integrates shade as an endmember.

Additionally, the novel two-step hierarchical approach for cropland mapping using RF and Striga detection using MESMA improved classification accuracies as the suppression of the spatial constraints (noise) from other classes minimized the spectral confusion (Degerickx et al., 2019). Therefore, applying a few EMs to a spatially constrained agroecosystem reduced the spectral densities and produced relatively high mapping accuracies at low complexity levels. Hence, this hierarchical MESMA approach accounts for the spectral pixel-to-pixel variability of agroecological systems through the simultaneous control of the spatial classification dimension (Franke et al., 2009). Additionally, in contrast with the other classification algorithms, MESMA provides the physical measurement of material contributions in a pixel or vegetation abundance (Li et al., 2005; Rosso et al., 2005). Thus, besides providing information about the occurrence of Striga in Rongo, this study quantified the magnitude and severity of the Striga infestation at suitable spatial resolutions (10 m x 10 m) that have never been provided before. Further studies are nevertheless desirable, to evaluate the performance of other classifiers and compare them to the Striga detection results derived from the approach of the present study.

The results of the MESMA for detecting Striga revealed that most of the crop fields in Rongo were infested with the Striga weed. The results obtained in this study concurred with the reports from other studies that reported the occurrence of Striga in western Kenya (Atera et al., 2013; Oswald et al., 2001) and in Rongo, mainly in the north and northwestern sites of Rongo. This information is critical for agricultural planning in Rongo as most of the farms in the northwestern sites of Rongo practice little crop rotation and concentrate on maize throughout the year on degraded soils. Therefore, there is a need for awareness among the farmers to initiate crop rotation and to improve soil fertility by incorporating cover crops that add organic matter, use green manure, or growing legumes that enhance nitrogen fixation in the soil. This is particularly different from the southern region where most of the farms are sugarcane plantations hence the low levels of Striga infestation observed.

4.6 Conclusions

In this chapter, a two-step hierarchical approach was employed for mapping cropland using the RF classifier in GEE and the Striga weed using MESMA on Sentinel-2 data within a heterogeneous agroecological system in the Rongo sub-county in Kenya. This chapter confirmed the effectiveness of the GEE as a data curation and cropland characterization platform using RF. The high potential of the MESMA algorithm to decompose mixed pixels and detect Striga occurrence and infestation levels were demonstrated. Therefore, the masking out of other non-target classes for different land cover classes before the implementation of MESMA allowed for a more focused and spatially adjusted spectral unmixing procedure that proved to enhance the classification accuracy results and reduced spectral confusion. Besides, deriving EMs from the in-situ spectrometric dataset provided a more realistic array spectrum of Striga rather than deriving the EMs from the Sentinel-2 image itself. Future studies should compare the use of Striga EMs from other sources and the respective EM selection mechanisms such as the endmember average root mean squared error (EAR), the minimum average spectral angle (MASA), and the count based endmember selection (CoB) that have been tested in other studies and are known to provide a filter to get pure spectra for use as EMs. Furthermore, Striga infestation fraction estimates could be empirically tested in areas of low, moderate, and high infestation, using the MESMA and cropland mapping hierarchical approach performed in this chapter.

Chapter 5

Landscape-scale integrative modeling approach for predicting Striga invasion risk

Photo: courtesy of Bester Tawona Mudereri 2018

This chapter is based on:

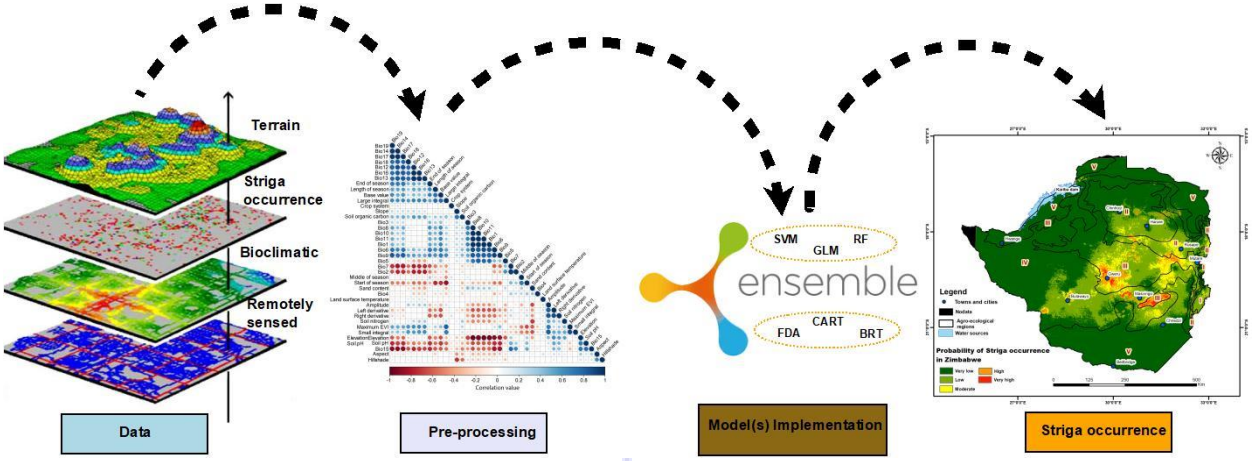
Mudereri, B. T., Abdel-Rahman, E. M., Dube, T., Landmann, T., Khan, Z. R., Kimathi, E., Owino, R., Niassy, S. (2020). *Multi-source spatial data-based invasion risk modeling of Striga (Striga asiatica) in Zimbabwe*. GIScience & Remote Sensing 57: 553–571. https://doi.org/10.1080/15481603.2020.1744250.

Abstract

Monitoring of destructive invasive weeds such as those from the genus Striga requires accurate, near real-time predictions and integrated assessment techniques to enable better surveillance and consistent assessment initiatives. Thus, in this chapter, the potential ecological niche of Striga (*Striga asiatica*) weed was predicted in Zimbabwe, to identify and understand its propagation and map potentially vulnerable cropping areas. Vegetation phenology from remote sensing, bioclimatic, and other environmental variables (i.e. cropping system, edaphic, land surface temperature, and terrain) were used as predictors. Six machine learning modeling techniques and the ensemble model were evaluated on their suitability to predict current and future Striga weed distributional patterns. The mentioned predictors (n = 40) were integrated into six models with 'presence-only' training and evaluation data, collected in Zimbabwe over the period between the 12th and 28th of March 2018. The area under the curve (AUC) and true skill statistic (TSS) were used to validate the performance of the Striga modeling outputs. The results showed that the ensemble model had the strongest Striga occurrence predictive power (AUC = 0.98; TSS = 0.93) when compared to the other modeling algorithms. Temperature seasonality (Bio4), the maximum temperature of the warmest month (Bio5), and precipitation seasonality (Bio15) were determined to be the most dominant bioclimatic variables influencing Striga occurrence. 'Start of the season' and 'season minimum value' of the 'enhanced vegetation index base value' were the most relevant remote sensing-based variables. Based on projected climate change scenarios, the chapter showed that up to 2050, the suitable area for Striga propagation will increase by ~ 0.73% in Zimbabwe. The present work demonstrated the importance of integrating multi-source data in predicting possible crop production restraints due to weed propagation. The results can enhance national preparedness and management strategies, specifically, if current future risks areas can be identified for early intervention and containment

Keywords: climate variability; food security; machine learning; niche modeling; remote sensing; sub-Saharan Africa, Striga weeds

Graphic abstract



5.1 Introduction

The genus *Striga* is composed of several species of parasitic weeds of global economic importance that cause socioeconomic losses of over US$ 1 billion in sub-Saharan Africa (Spallek et al. 2013). *Striga hermonthica* and *S. asiatica* are the most prevalent among the Striga species with *S. asiatica* reported affecting approximately 40% of arable land in the region (Cochrane and Press, 1997). These two species thrive in climatic conditions, which are also favorable for most cereal crops like maize, sorghum, millet, and rice (Khan et al., 2014). These cereal crops are the most important staple crops for the majority of the African population, with maize being grown for food purposes by more than 300 million people out of an estimated one billion population in sub-Saharan Africa (Sasson, 2012).

Striga weeds attach themselves to the roots of the cereal crops, after germination and outcompete their hosts for space, nutrients, and water. As a result, the growth and development of the host crops deteriorate, causing considerable yield reduction (Ejeta and Gressel, 2007). In some areas in Africa, the scourge of Striga has reached epidemic magnitudes, causing a desperate scenario, mostly to poor small-scale farmers (Mandumbu et al. 2017a). The most common response practice in such scenarios is for farmers to abandon the land and look for new croplands, a very labor-intensive task that inevitably contributes to cropland expansion and severe environmental degradation.

In this chapter, the probability of occurrence of Striga (i.e. *S. asiatica*) was predicted in Zimbabwe, using vegetation phenology from remote sensing, bioclimatic, other remotely sensed variables (i.e. cropping system, edaphic, land surface temperature, and terrain), empirical machine learning and ecological niche modeling approaches. The advent of these freely available earth observation 'big data' from multiple sources and machine learning algorithms permit access to a new paradigm of immense opportunities to understand the earth and agroecological systems over time and space (Cian et al. 2018). This allows for comprehensive statistical analysis on large temporal resolution data using the phenological characteristics hidden in these time-series data (Landmann et al., 2019). These various time-series inherent characteristics in the 'big data', are often concealed in single snapshot remotely sensed imagery (Cian et al. 2018; Ochungo et al. 2019).

Thus, this multi-source remotely sensed data, coupled with advanced and efficient machine

learning and ecological niche modeling approaches provide a cost-effective, timely, robust, and very accurate platform to map and predict the occurrence of invasive weeds like Striga (Thamaga and Dube 2019; Jafarian et al.2019; Sibanda et al. 2019). In particular, mapping flowering Striga, i.e. *S. hermonthica* using in-situ methods, high spatial resolution satellite data, and machine learning has been proven to be largely possible and achievable at the plot (Mudereri et al. 2020a) and field scale (Mudereri et al. 2019a). However, the potential of using these remotely sensed data to detect and map the risk posed by the understory Striga such as the *S. asiatica*, which exists completely covered underneath crop canopies has not been attempted anywhere, more so at landscape scales or by using multi-source data. This information deficit is mainly attributable to the heterogeneous nature of the agro-natural landscapes in Africa and the multiple spectral responses obtained from crop fields that are infested with understory Striga weed, which cause enormous errors in their detection and mapping (Mudereri et al. 2019a).

When used in species distribution modeling, machine learning and ecological niche models correlate the present location ('presence-only' or 'presence-absence' data) of a species with the appropriate predictor variables (e.g. environmental variables), thereby providing a statistical link between the spatial differences of the predictor variables and the dispersion of the species in the environment, in this case, Striga (Ayebare et al., 2018). Accuracy of the machine learning and ecological niche modeling relies on the precision and distribution of the 'presence-only' data tied with a careful selection of ecological and climatic predictor variables (Elith et al. 2010). However, it can be inferred that there is no universally best machine learning algorithm, which warrants the scoping into the best predictive model and the best predictor combinations for species distribution (Guo et al., 2019).

Therefore, identifying robust machine learning and ecological niche modeling algorithms that can select the most relevant predictor variables from multiple ecological covariates to predict the occurrence, propagation, and distribution of the understory Striga species such as boosted regression trees (BRT: Friedman, 1999), classification and regression trees (CART: Breiman et al., 1984), flexible discriminant analysis (FDA: Fisher, 1936), generalized linear model (GLM: Nelder and Wedderburn, 1972), random forest (RF: Breiman, 2001) and support vector machines (SVM: Vapnik, 1979), is crucial. Moreover, integrating the remotely sensed and bioclimatic data in such machine learning and ecological niche modeling has been reported in other studies as the

best way to produce reliable and accurate results by harnessing the vast information provided by the intrinsic phenological vegetation metrics and the external influence of climatic variables (Kyalo et al., 2018; Makori et al., 2017).

However, there is a huge deficit in information regarding the use of such technology on analysis and mapping of Striga distribution and risk particularly the influence of climate change on the distribution of Striga in Africa. Many studies have investigated the effects of climate change on a range of species, showing that change in climatic conditions has a profound impact on species distribution ranges (Mbatudde et al. 2012; Wan and Wang 2019; Guan et al. 2020). To the best of the candidate's knowledge, there is no precise spatial information or scenario modeling detailing the current or probable risk that climate change will impose on the distribution, occurrence, and severity of Striga in Zimbabwe or on the entire African continent. The risk inflicted by these parasitic weeds is likely to be worsened by climate change and the inadequate adaptive or mitigation capacity, in addition to the limited impact documentation leading to inadequate preparedness (Niang et al., 2014).

Thus, identifying and controlling these invasive weeds before they can spread to new environments requires better surveillance and constant monitoring across the African countries with adequate, cost-effective tools and methods (Sibanda et al. 2019). It is, therefore, hypothesized that climate change might cause restrictions or expansions on the distribution of Striga species through altering host availability or imposing Striga intolerable or suitable climatic conditions (Cochrane and Press, 1997). Thus, the uncertainties brought by these future climate scenarios necessitate robust and accurate mapping methods and relevant environmental multi-source variables and datasets to estimate and predict the potential and actual impact of climate change on the current and future distribution of the biological niche of Striga in Zimbabwe.

5.2 Study area

The occurrence of Striga in the 10 provinces of Zimbabwe (Figure 5.1) was predicted. Zimbabwe is a landlocked country in southern Africa covering a land area of ~ 390 753 km^2, which shares borders with Botswana, Mozambique, South Africa, Zambia, and partly Namibia. It is bound within latitudes 15.6^0 and 22.4^0 South and longitudinally between 25.2^0 and 33.1^0 East (Kuri et al., 2018). Geographically, the central part of the country is located on a high plateau forming a

watershed between the two major river systems, i.e. Zambezi river in the north and Limpopo river in the south. The country has a remarkably varied climate, marked by the differences in latitude which characterizes the wide-ranging rainfall patterns and extensive agronomic activities. It is situated within the tropics and experiences the short cold, dry season between May and September, while the period November to April is marked by heavy rainfall (Mudereri et al. 2019a).

Zimbabwe is subdivided into five agroecological regions that vary in temperature, rainfall, soil, and agricultural potential (Table 5.1). These five agroecological regions include regions I and II referred to as the Highveld; region III which is Middleveld, while region IV and V are referred to as the Lowveld (Sungirai et al., 2018). In Zimbabwe, the lowest annual rainfall is 400 mm which is received in region V and the highest amount (1 200 mm) is received in the region I. The mean annual temperature ranges from 16^0 C in regions I and II to ~ 26–35^0 C in the southern Lowveld (Kuri et al., 2019). Approximately, 11% of the country is arable land with ~ 0.31% of that arable land being continuously under different crops such as maize, wheat, sorghum, and millet. Of these crops, maize is the most economically important and commonly grown cereal crop at both small- and large-scales (Kuri et al., 2018).

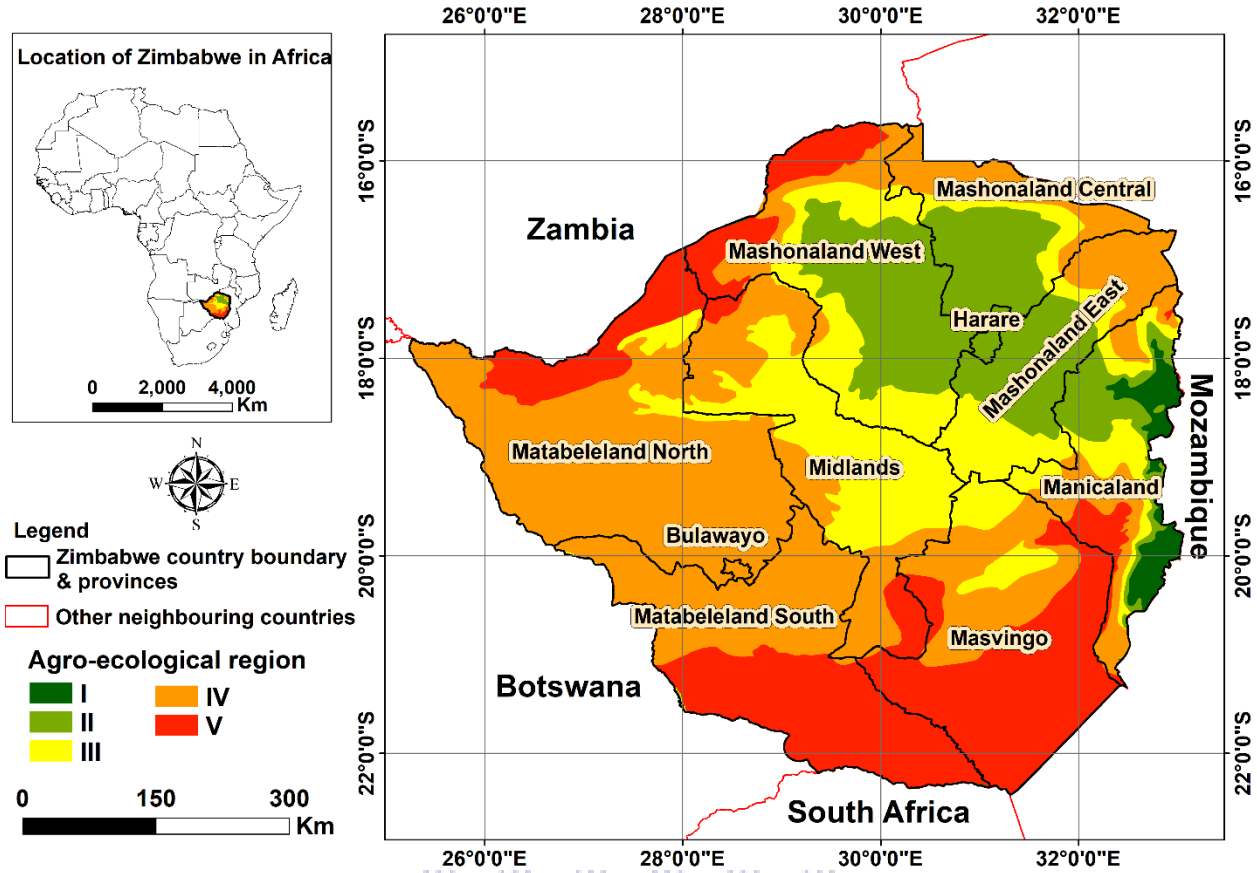


Figure 5.1: Location of Zimbabwe in Africa and the relative location and boundaries of the five agroecological regions of the country which characterize the study area. *See* Table 5.1 for a detailed description of the agroecological regions.

Table 5.1: Characteristics of the five agroecological regions of Zimbabwe (FAO and ACFD, 1999; Mugandani et al., 2012).

Agroecological region	Average annual rainfall (mm year^{-1})	Mean maximum temperature (^{0}C)	Dominant soil type	Agriculture potential and farming system
I	>1000	16–19	Acrisols, Ferralsols	Suitable for dairy farming forestry, tea, coffee, fruit, beef, and maize production
II	700–1050	19–23	Cambisols, Luvisols, Arenosols	Suitable for intensive farming, based on maize, tobacco, cotton, and livestock
III	500–800	23–26	Arenosols	Suitable for intensive farming, based on maize, tobacco, cotton, and livestock
IV	450–650	19–26	Leptosols, Vertisols, Solonetz	Semi-extensive region. Suitable for farm systems based on livestock and resistant fodder crops. Forestry, wildlife/tourism
V	<450	26–32	Leptosols, Vertisols, Solonetz	Extensive farming region. Suitable for extensive cattle ranching. Zambezi Valley is infested with tsetse fly. Forestry, wildlife/tourism

5.3 Methodology

Figure 5.2 shows a flowchart that explains the entire methodology adopted in the present chapter for modeling Striga invasion risk.

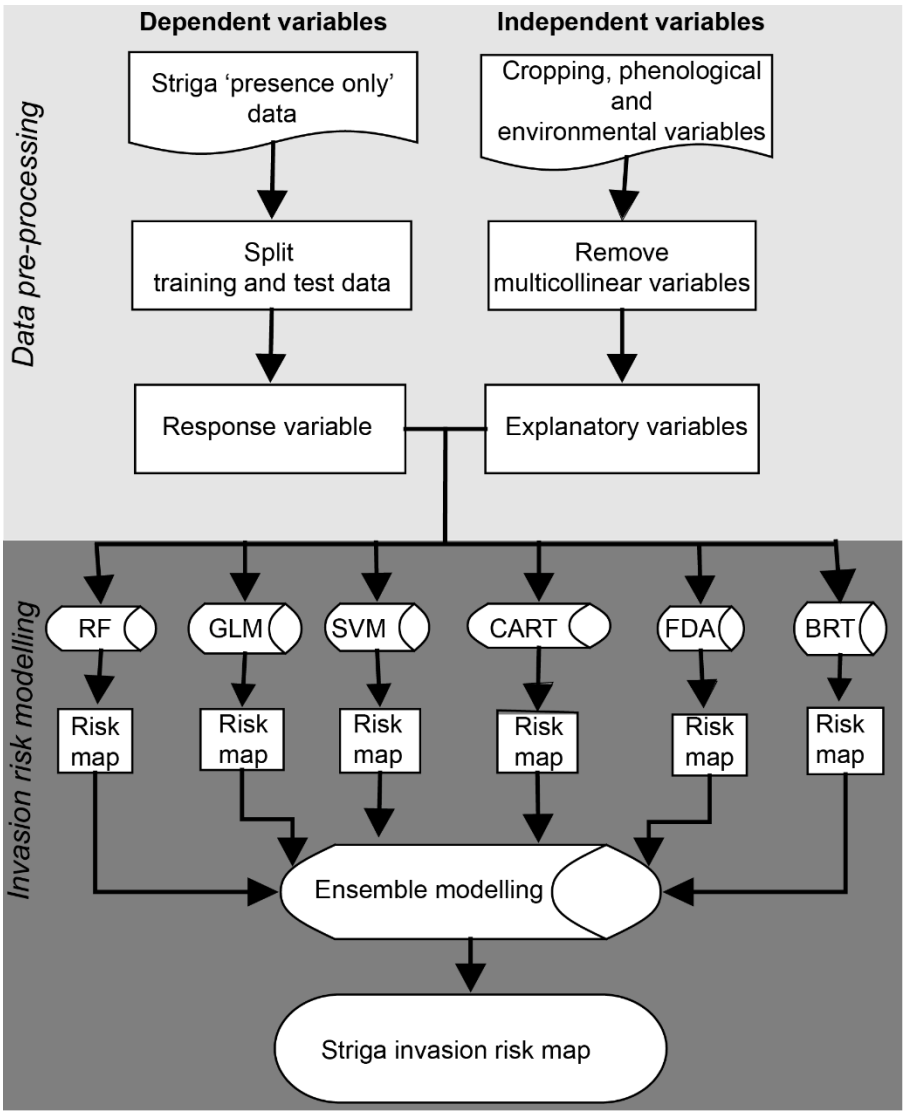


Figure 5.2: Flow diagram of the methodology of Striga invasion risk modeling. The six models used are random forest (RF), generalized linear model (GLM), support vector machines (SVM), classification and regression trees (CART), flexible discriminant analysis (FDA), and boosted regression trees (BRT).

5.3.1 Striga occurrence data collection

The Striga 'presence-only' data were collected between the period 12th and 28th of March 2018. The data collection period matched with the flowering phase of Striga in Zimbabwe. The flowering phase was targeted since this is the best time to differentiate Striga from other co-occurring weeds (Mudereri et al. 2020a). Reference 'presence-only' data were gathered within maize croplands, which in the study area were mainly mono and mixed maize cropping systems. The mixed cropping system is mostly a combination of maize with groundnuts, round-nuts, or beans. A purposive sampling approach was employed using local expert knowledge (i.e. extension officers and

farmers) to assist in the identification of the Striga infested fields. A handheld global positioning system (GPS) device with an error margin of ±3 m was used to locate the reference control points. The edge-effect was avoided by collecting the Striga 'presence-only' data at the center of the field (sampling unit: 30 m x 50 m). A total of 50 'presence-only' Striga reference data were collected covering the six districts in Midlands and Masvingo provinces (Figure 5.1); namely Bikita, Chivi, Gweru, Masvingo, Shurugwi, and Zaka (agroecological regions III and IV). These points were spread across the different elevation gradients (400–1 600 m a.s.l), except for the highest elevation in the Eastern highlands (> 1 600 a.s.l). Agroecological region V, which was not sampled, is the region mainly reserved for livestock production in Zimbabwe and is characterized by non-arable land and pastures (*see* Table 5.1).

5.3.2 Predictor variables

The predictor variables that were used in this chapter were grouped into two main categories, i.e. bioclimatic and remotely sensed variables (Table 5.2 Table 5.3). Variable spatial and temporal resolutions are a key notion in determining a dataset's fitness for a given use as they influence the pattern that can be observed during the analysis (Degbelo and Kuhn, 2018). However, Csillag et al. (1992) pointed out that there is no single best resolution when combining environmental variables of varying resolutions. In this chapter, the variables ranged in pixel size from 1 km x 1 km to approximately 250 m x 250 m spatial resolution. This variation influences the integration of multiresolution variables within models.

The variation was, therefore, counteracted by resampling all the datasets to the lowest spatial resolution of 250 m x 250 m pixel size. In addition, the vegetation phenological variables were derived from multidate input data, while other remotely sensed and bioclimatic variables were generic. Thus, only their respective derivatives (i.e. output variables) from the TIMESAT model were used as inputs in the invasion risk modeling analysis to offset the effect of the temporal variation. Therefore, there is no expectation of any influence of the spatial or temporal resolution on the accuracy of the models. It is further worth noting that all the predictor variables are freely available.

5.3.2.1 Bioclimatic variables

In this chapter, 19 bioclimatic variables (Table 5.2) that are freely downloadable from the

WorldClim platform (www.worldclim.org) at ~1 km x 1 km spatial resolution were used to determine the key climatic conditions influencing the distribution of Striga in Zimbabwe for both the current and future climate scenarios. Four representative concentration pathways (RCPs) were set by the IPCC using the total radioactive forcing of values 2.6, 4.5, 6.0, and 8.5 watt/m^2 (IPCC, 2014). The current bioclimatic data (1950–2000) and a one-time step of the future climate data of the maximum emission (RCP8.5) for the CO_2 concentrations predicted for 2050 (average of predictions for 2041–2060: Guan et al. 2020) were used. The future climatic data were obtained from the fourth version of the community climate system model (CCSM4), which is one of the models that provide the most efficient global future climate projections (Mohammadi et al., 2019; Mudereri et al., 2020b). All these bioclimatic variables were clipped to the Zimbabwean country boundary and resampled to 250 m x 250 m pixel size, to match the size and extents of the remotely sensed variables.

5.3.2.2 Remotely sensed variables

A total of five remotely sensed variable categories was used i.e. cropping system, edaphic, land surface temperature, terrain, and vegetation phenology (vegetation seasonality characteristics) (Table 5.3). These variables were selected because they are reported in several studies as key determinants of Striga distribution (Mandumbu et al. 2017a; Oswald et al. 2001; Parker 2009; Mudereri et al. 2019a). All the remotely sensed variables that were not in the 250 m x 250 m spatial resolution were standardized and resampled to this pixel size.

***Cropping system variable*.** The cropping system variable provided by the study of Landmann et al. (2019) was used. The variable was obtained at 30 m x 30 m pixel size with three categorical classes: rainfed wildland, rainfed cropland, and irrigated cropland. The cropping system variable for Zimbabwe was derived at 97% accuracy (Landmann et al., 2019) using vegetation harmonics of the time-series NDVI derived from Landsat 8 operational land imager (OLI) images.

Edaphic variables. Four soil properties that were downloaded from https://www.isric.org/explore and referred to as the 'AfSoilGrids50m' (Hengl et al., 2015) were used. Specifically, the Africa soil grids produced at 0–30 cm depth with a spatial resolution of 250 m x 250 m was used. The soil characteristics data includes total soil nitrogen (N) (mg/kg: ppm), soil pH, soil organic carbon (SOC) in g/kg, and sand content (50-2000 µm) in g/100 g (Hengl et al., 2015). These variables

were chosen because they broadly influence soil fertility, and thus the potential occurrence of Striga. Several studies have established that depleted soil fertility leads to rapid propagation and thriving of Striga within croplands (Ekeleme et al., 2014; Yoneyama et al., 2007).

Land surface temperature. 'Day time' land surface temperature climate modeling grid (LST_Day_CMG) available in K, simulated from the moderate resolution imaging spectro-radiometer (MODIS) data and available at https://lpdaac.usgs.gov/products/mod11c2v006/, (Wan et al. 2015) was used. Specifically, the 'multi-day' MOD11C2 LST product of 5.6 x 5.6 km spatial resolution available from the year 2000 to the present was used. LST was chosen because the cereal chemical that triggers Striga germination requires optimal temperature (i.e. $22^0 – 30^0$C) and Striga seeds also need an optimum soil temperature range to germinate (Rich and Ejeta, 2007). It is therefore postulated that the surface fluxes measured by LST would be one of the proxy key variables that immensely predict the potential germination of Striga seeds.

Terrain variables. The terrain variables were calculated from the shuttle radar topographic mission (SRTM) data which are available as 3 arc seconds (~ 90 m resolution) digital elevation model (DEM). The vertical error of the DEM was less than 16 m, which was sufficient for the intended purpose (CGIAR-CSI, 2019). In addition to the elevation, other terrain variables (aspect, hill-shade, and slope) were derived using the 'terrain analysis' plugin in QGIS (QGIS Development Team, 2019). The influence of elevation, slope, hill-shade, and aspect on soil type, soil moisture content, soil fertility, soil temperature, and runoff among other factors were anticipated to influence the occurrence and propagation of Striga weed. Striga has been reported by other studies to be tolerant of a wide range of altitudes from sea level to ~2 480 m a.s.l (Cochrane and Press, 1997).

Vegetation phenological variables. Vegetation phenological variables were estimated from 250 m, MODIS 16-day EVI time-series composites. A 6-year observation period between 2012 and 2018 was used. The vegetation phenological variables were computed using the TIMESAT software (Eklundh and Jönsson 2017; Jönsson and Eklundh 2004, 2002). TIMESAT enumerates phenological harmonics that occur within a time-series satellite dataset by superimposing localized equations to the time-series data points. Curve smoothing functions are thereafter applied to the model to extract the seasonal vegetation phenological parameters from these vast multitemporal

data dimensions. This consecutively reduces the influence of residual signal noise produced by the variation in the raw EVI time-series data (Hentze et al. 2016; Makori et al. 2017). A Savitzky-Golay filter was employed to smoothen the fitting curves and removed outliers using a 3- and 5-point window over 2 fitting steps, 3.0 adaptation strength without a spike or amplitude cut-offs, a 0.0 season cut-off, and a 20% threshold for the beginning and end of the season following a procedure described in Makori et al. (2017). Using this protocol, 11 vegetation phenological variables (Table 5.3) were extracted. For a detailed explanation of the calculated variables and how the TIMESAT algorithm operates, readers are referred to Eklundh and Jönsson (2017).

5.3.3 Collinearity test of variables used in the ecological niche modeling

A 2-stage variable elimination criterion was employed using the variance inflation factor (VIF) and the Person correlation coefficient. VIF detects multicollinearity by taking each predictor and regressing it against the other variables in a multiple linear regression analysis (Plant, 2012). The resulting R^2 value obtained from this multiple regression analysis is then replaced in the VIF calculation formula as shown in Equation 5.1.

$$VIF_i = \frac{1}{1 - R_i^2} \quad (5.1)$$

Where i is the predictor

In this chapter, the 'vifcor' function in the 'usdm' package available in R (Naimi et al., 2014; R Core Team, 2019) was chosen to perform the variable elimination. The 'vifcor' function iteratively selects pairs of variables with high linear correlation, then eliminates the one with the highest VIF. A threshold of th = 0.7 was set, which represents a Pearson correlation coefficient ($r \geq 0.7$) following Kyalo et al. (2018) recommendation. In principle, a VIF value greater than 10 is evidence of the collinearity problem within a model (Dormann et al., 2013). Although some of the variables from the VIF calculation process showed low VIF values, the correlation matrix (Figure 5.3) revealed further correlations among some of the variables. Therefore, from the variables with low VIF, the variables that have been reported in the literature to be of ecological significance (Mandumbu 2017) were selected. The variable elimination procedure resulted in a selection of 21 eligible variables from 40 bioclimatic and remotely sensed variables. The 21 variables that were used in the final modeling procedure are highlighted in bold in Table 5.2 Table 5.3.

Table 5.2: Bioclimatic variables used in the species distribution models for Striga occurrence prediction and their variance inflation factor (VIF) values. The variables in bold were used in the final Striga occurrence prediction after eliminating highly correlated ones.

BioClim Code	Environmental variable description	Unit	VIF value
Bio1	**Annual mean temperature**	**^{0}C**	**7.30**
Bio2	Mean diurnal range [mean of monthly (max temp–min temp)]	^{0}C	3.05
Bio3	**Iso-thermality (Bio2/Bio7) (×100)**		**3.25**
Bio4	**Temperature seasonality (standard deviation ×100)**		**2.00**
Bio5	**Maximum temperature of the warmest month**	**^{0}C**	**8.21**
Bio6	Min temperature of the coldest month	^{0}C	10.23
Bio7	Temperature annual range (Bio5–Bio6)	^{0}C	12.63
Bio8	Mean temperature of wettest quarter	^{0}C	5.04
Bio9	Mean temperature of driest quarter	^{0}C	1.92
Bio10	Mean temperature of warmest quarter	^{0}C	8.73
Bio11	Mean temperature of coldest quarter	^{0}C	1.62
Bio12	Annual precipitation	mm	2.76
Bio13	Precipitation of wettest month	mm	7.27
Bio14	Precipitation of driest month	mm	3.91
Bio15	**Precipitation seasonality (coefficient of variation)**	**mm**	**4.32**
Bio16	Precipitation of wettest quarter	mm	4.70
Bio17	Precipitation of driest quarter	mm	6.33
Bio18	**Precipitation of warmest quarter**	**mm**	5.41
Bio19	Precipitation of coldest quarter	mm	2.58

Table 5.3: Remotely sensed variables used in the species distribution models for Striga occurrence prediction and their variance inflation factor (VIF) values. The variables in bold were used in the final Striga occurrence prediction model after eliminating highly correlated ones. EVI is the enhanced vegetation index.

Variable	Description	Units	VIF Value
	Cropping system		
Cropping system	**Irrigated or rain-fed cropland/wildland**	**Categorical**	**1.20**
	Edaphic variables		
Sand content	**Quantity of sand in the soil**	**g/100g**	**2.44**
Soil nitrogen (N)	**The total amount of nitrogen in the soil**	**mg/kg**	**2.50**
Soil organic carbon	**Organic matter present in the soil**	**g/kg**	**1.89**
Soil pH	**Acidity or alkalinity of the soil**	**pH value**	**2.07**
	Land surface temperature (LST)		
LST	**Land surface temperature**	**K**	**1.07**
	Seasonality variables		
Amplitude	The difference between maximum EVI and base value	EVI value	1.09
Base value	**Minimum EVI value**	**EVI value**	**1.60**
End of season	**EVI value at the time of the end of season**	**EVI value**	**1.02**
Large integral	**Integral of the season from start to end**		**1.58**
Left derivative	**Rate of EVI value increase at the start of the season**	**%**	**1.23**
Length of season	Time-lapse from start to end of season	Days	6.88
Maximum EVI	Maximum EVI value in season	EVI value	1.05
Middle of season	Absolute value at the middle of the season	EVI value	2.19
Right derivative	Rate of EVI value increase at the end of season	%	1.99
Small integral	Integral of the season and base value from start to end of season		2.47
Start of season	**EVI value at the beginning of the season**	**EVI value**	**1.04**
	Terrain variables		
Aspect	**Slope direction**	**Degrees**	**1.15**
Elevation	**Ground height above sea level**	**m**	**4.46**
Hillshade	**Shading the sun effect**		**1.22**
Slope	**Ground steepness**	**%**	**1.76**

Collinearity amongst the predictor variables in most ecological niche modeling causes instability and volatility of the model parameterization and performance (Dormann et al., 2013). The variables correlation matrix using the Pearson correlation coefficient is shown in Figure 5.3.

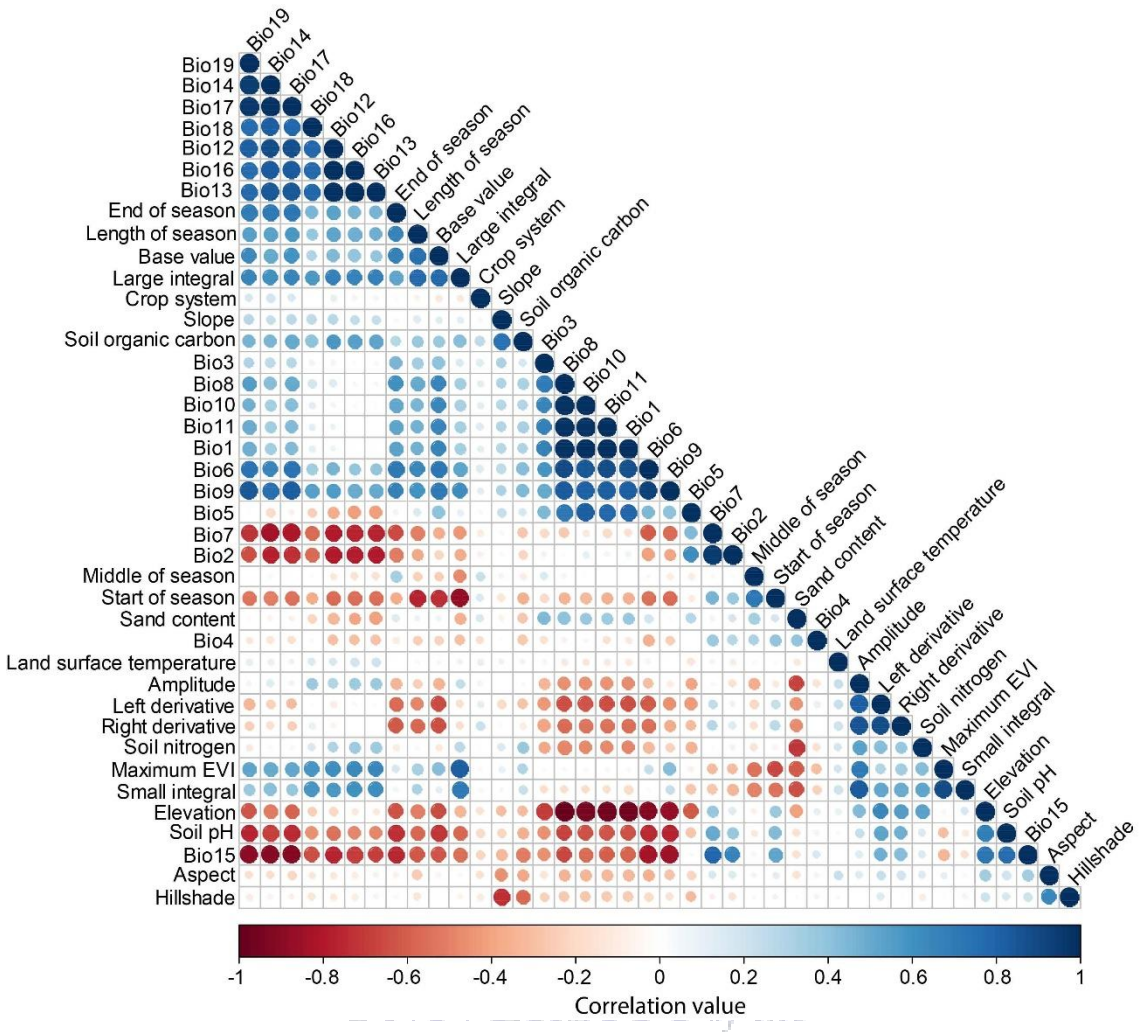


Figure 5.3: Collinearity matrix for ecological niche models' predictor variables. Darker shades of blue and red color indicate high variable collinearity, while lighter shades indicate low collinearity. Similarly, the smaller the circle, the lower the correlation value.

5.3.4 Species distribution models implementation

Striga occurrence predictive models were built using the 'sdm' package (Naimi and Araújo, 2016) performed in R (R Core Team, 2019). The 50 'presence-only' points data that were collected in the field against 1 000 pseudo-absence points generated using the 'sdmdata' function inherent in the 'sdm' package were used. Stockwell and Peterson (2002) concluded that when using machine learning methods for species niche predictions, the accuracy for predicting the occurrence of a species at a location, was 90% of maximum within 10 sample points, and was near maximal at 50 data points. Therefore, the Striga 'presence-only' sample size (i.e. 50) was within the sufficient sample size required for accurate predictions when ecological niche modeling is used at a national

scale (Stockwell and Peterson, 2002). Often, obtaining real 'presence-absence' data is logistically impractical; however like other ecological niche models, 'sdm' allows for the use of background pseudo-absence data (Guan et al., 2020; Mudereri et al., 2020b). The package 'sdm' combines diverse executions of ecological niche models (n = 15) within a single platform and uses the same 'presence-only' and pseudo-absence data by applying an object-oriented reproducible and extensible framework for ecological niche modeling in R (Naimi and Araújo, 2016). In the present chapter, only 6 of the 15 modeling techniques in 'sdm' as follows: CART, BRT, RF, FDA, GLM, and SVM were used and inter-compared.

The CART model grows a single decision tree based on the binary partitioning algorithm, which splits the data until it is homogenous, using a hierarchical structure and regression tree (Breiman et al., 1984). Similarly, the BRT model uses the same decision tree approach but improves from the use of a single regression tree by combining multiple decision trees in a process called boosting (Elith et al. 2008). On the other hand, the RF uses these multiple decision trees and randomly grows a forest of decision trees, then predictions are conducted through majority voting for the class with the highest number of votes among these multiple grown trees (Bangira et al., 2019). FDA is a non-parametric multiple regression and additive technique and the GLM uses a linear regression approach (Nelder and Wedderburn, 1972), while SVM uses a hyperplane to estimate the divergence of class groupings for the prediction (Hastie et al. 1994; Vapnik 1979). These six algorithms were selected in this chapter because they are widely used in conducting complex output predictions with relatively high modeling accuracies for regression and classification (Abdel-Rahman et al. 2013a; Makaya et al. 2019; Mosomtai et al. 2016; Tesfamichael et al. 2018). A summary of these models' execution syntax and their corresponding packages used by 'sdm' in the parallel model simulations is provided in Table 5.4.

Table 5.4: R software packages used by 'sdm' in the parallel execution of the six models; namely (a) boosted regression trees (BRT), (b) classification and regression trees (CART), (c) flexible discriminant analysis (FDA), (d) generalized linear model (GLM), (e) random forest (RF) and (f) support vector machines (SVM)

Algorithm	Syntax code in 'sdm'	Package used	Reference
Boosted regression trees	'brt'	gbm	(Elith et al. 2008)
Classification and regression trees	'cart'	rpart	(Breiman et al., 1984)
Flexible discriminant analysis	'fda'	earth	(Hastie et al., 1994)
Generalized linear regression	'glm'	glmnet	(Friedman et al., 2010)
Random forest	'rf'	randomForest	(Liaw et al., 2002)
Support vector machines	'svm'	Kernlab	(Karatzoglou et al., 2004)

An ensemble projection approach was used to harmonize the variations produced by the different model predictions. Ensembles have been reported to have superior predictive performance as compared to individual models (Hao et al., 2019). The ensemble modeling fits and maximizes the prediction accuracy with higher reliability as it binds together the highest performance of all the models that have the most acceptable precision and accuracy. The function 'ensemble' within the 'sdm' package was used to harmonize the results of the Striga occurrence prediction amongst the six modeling algorithms using the TSS weighted average approach (Naimi and Araújo, 2016). Compared to using the most intuitive approach, which applies the mean or median, the weighted average improves the model's predictive ability (Naimi and Araújo 2016; Jafarian et al. 2019). However, the weighted average requires validation of the selected modeling algorithms before inclusion in the 'sdm' (Hao et al., 2019). For the present chapter, the threshold of TSS = 0.7 was set for the models to qualify for inclusion in the ensemble as generally a TSS score of > 0.7 shows high agreements between the predictor variable and the independent data (Allouche et al. 2006).

For consistency, the same approach was used to perform the Striga occurrence predictions for both the current and future climate scenarios. Since this study was more focused on the influence of climate change on the distribution of Striga, only the selected climatic variables were varied, but all the other variables were assumed to remain constant in the future and if they would be differences, it is assumed that they would not to be significant enough to cause major variances to the obtained results.

The predicted suitable area for the probability of Striga occurrence for both the current and future scenarios was further calculated using an average thresholding value of all the models. Using this

value, a binary raster image of the 'presence' (occurrence) and 'absence' classes for the whole study area was created. The total number of pixels to estimate the total coverage of the predicted area against the unsuitable area was used.

5.3.5 Models' accuracy validation

The accuracy and variable importance of the models were tested using a 10-fold cross-validation approach. The relative variable contribution was used to the model using the inbuilt randomly split 'independent test dataset' option available in the 'sdm' package. This was automated to universally apply to each of the six models. The performance of the six models was further validated using the receiver operating curve (ROC) by analyzing the area under the curve (AUC) and true skills statistic (TSS: Allouche et al. 2006; Guan et al. 2020). The ROC is a graphical representation of how well the model fits the data points. The values for the AUC range between 0 and 1. Models whose predictions are 100% inaccurate have an AUC of 0, while those with perfect prediction have an AUC of 1. In general, AUC values ≥ 0.7 demonstrate high model prediction performances (Mohammadi et al., 2019).

On the other hand, TSS combines sensitivity and specificity to explain the model commission and omission errors (Kyalo et al., 2018). The values of TSS range between −1 to +1, where +1 demonstrates a perfect agreement between the observed and the predicted Striga occurrence, while values ≤ 0 indicate no agreements or that most of the predictions for the Striga occurrence were produced by chance (Allouche et al., 2006). The weighted average of the TSS was therefore used to perform the ensemble predictions. TSS was chosen since it is a relatively reliable measure instead of the AUC and chi-squared (X^2) statistics which are somewhat biased and highly sensitive to the proportional extent of the predicted presence observations (Kyalo et al., 2018). The ensemble model merges the strengths of these ecological niche modeling approaches while minimizing their weaknesses (Araújo et al., 2019; Guan et al., 2020)

$$\text{TSS} = \textit{Sensitivity} + \textit{Specificity} - 1 \tag{5.1}$$

$$Sensitivity = \frac{a}{a+b} \tag{5.2}$$

$$Specificity = \frac{d}{c+d} \tag{5.3}$$

where *a* is true positive, *b* is a false negative, *c* is false positive, and *d* is true negative.

The output maps from the six models and their respective ensembles were imported into a geographical information system (GIS) environment for further analysis. Based on a suggestion by Abdelaal et al. (2019), the probability maps were reclassified into five classes of Striga probability of occurrence. These classes were: (i) very low probability (≤ 0.05), (ii) low probability (0.051–0.10), (iii) moderate probability (0.11–0.30), (iv) high probability (0.31–0.50), and (v) very high probability (≥ 0.50).

5.4 Results

5.4.1 Models' accuracy, comparison, and validation

The VIF statistic of the predictor variables that were included in the modeling approach using the six models is summarized in Table 5.2 and Table 5.3. The lowest values of VIF were related to remotely sensed variables i.e. end of the season (1.02), the start of the season (1.04), LST (1.07), and amplitude (1.09), while bioclimatic variables had higher values of VIF such as Bio5 (8.21), Bio1 (7.30), Bio18 (5.41) and Bio15 (4.32). However, these values were not large enough to warrant these variables to be eliminated from the modeling. On the other hand, the VIF values for Bio6 and Bio7 were 10.23 and 12.63, respectively. These variables had VIF values greater than 10, so they were excluded from the modeling analysis.

Using the ROC, the patterns of the smoothened graphs of the ten replicated ROCs showed that RF and GLM were relatively consistent in their prediction amongst the model replicates compared to the other models (Figure 5.4). Some of the replicated graphs using different sets of data folds for the CART, FDA, and BRT models were below or closer to the one-to-one line (the black dotted line in Figure 5.4).

All the models generally showed relatively high accuracy in predicting Striga occurrence in Zimbabwe, with all the models producing acceptable accuracies *viz.* AUC > 0.85 and TSS > 0.70. Further, the models' predictive performance, as indicated by AUC and TSS values revealed that RF had the highest values of AUC (0.98) and TSS (0.92) (Figure 5.4). The FDA model produced the lowest AUC (0.87) and TSS (0.72) scores. It is, however, observed variations where models such as the GLM had a higher AUC, but lower TSS in comparison to other models such as CART.

Nonetheless, the FDA performed the least using the AUC and the TSS accuracy measures.

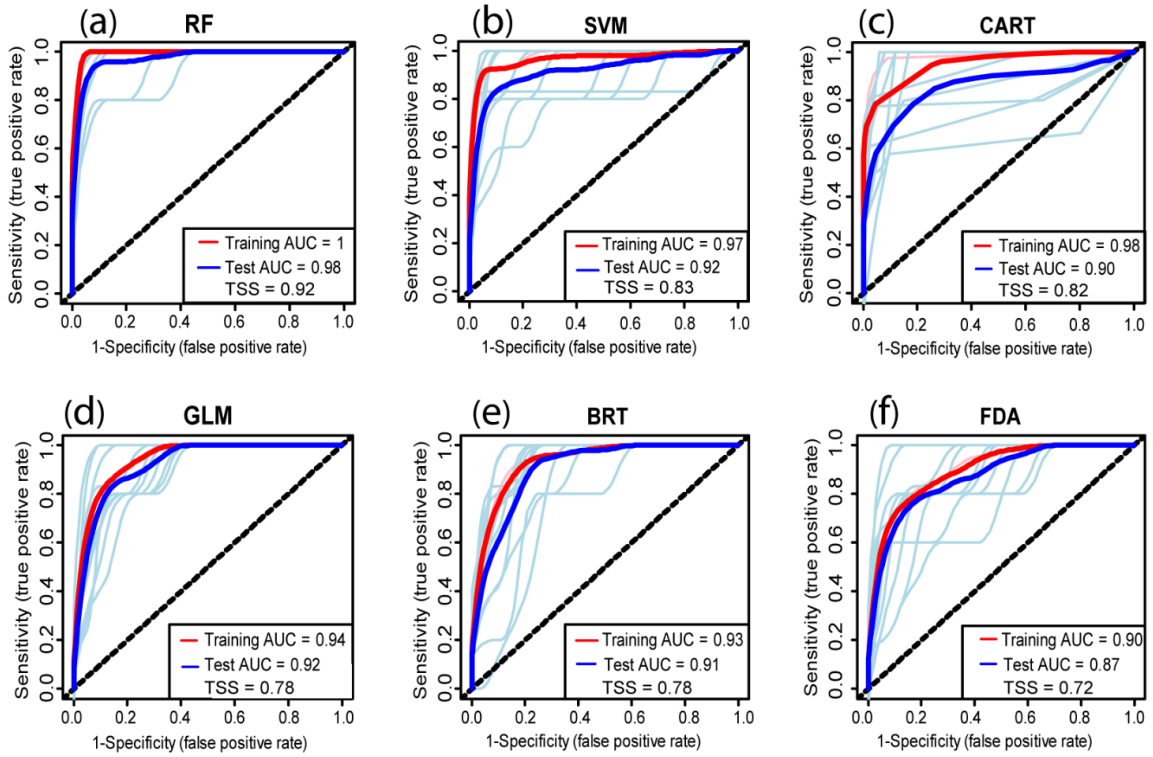


Figure 5.4: Results of the receiver operating curve (ROC) for the six machine learning and ecological niche model algorithms used to predict Striga occurrence in Zimbabwe; namely: (a) random forest (RF), (b) support vector machines (SVM), (c) classification and regression trees (CART), (d) generalized linear model (GLM), (e) boosted regression trees (BRT), and (f) flexible discriminant analysis (FDA). The red curve represents the smoothened mean area under the curve (AUC) using the training data, while the blue curve depicts the smoothened mean AUC using the test data from the 10-fold cross-validation sampling. The cyan curves show the 10-fold replicated model runs using the training data.

5.4.2 Variable importance using the current climate scenario

A total of 5 out of the 21 predictor variables appeared in the ten most relevant variables for all six models. The five predictor variables regarded as very relevant by all the models are 'base value', 'start of season', 'temperature seasonality' (Bio4), 'maximum temperature of the warmest month' (Bio5), and 'precipitation seasonality' (Bio15). The respective variable contributions in the different models are summarized in Figure 5.5. The Bio5 variable appeared twice as the most relevant variable i.e. for the RF and BRT models, while the Bio1 variable also appeared twice as the most relevant variable for the GLM and FDA models. Bio4 and Bio15 were also selected as important predictors for the CART and SVM models, respectively.

Further analysis of the variable importance revealed that the bioclimatic and seasonality

parameters dominated the most relevant list while the edaphic, terrain, LST, and cropping system were not particularly relevant across the six algorithms tested. Terrain variables appeared in all the models at different contribution levels; however, elevation appeared more frequently than the other terrain variables. Similarly, regarding the edaphic variables, soil organic carbon and sand content dominated their category with varying contributions across the six models. The cropping system variable appeared once under the BRT model, however, with a very low contribution to the entire model. LST did not appear among the ten most important variables for the six models. RF, which had the highest accuracy (AUC = 0.98) amongst the other models, selected Bio5, Bio4, and Bio15 as the most relevant predictor variables for the estimating occurrence probability of Striga in Zimbabwe (Figure 5.5).

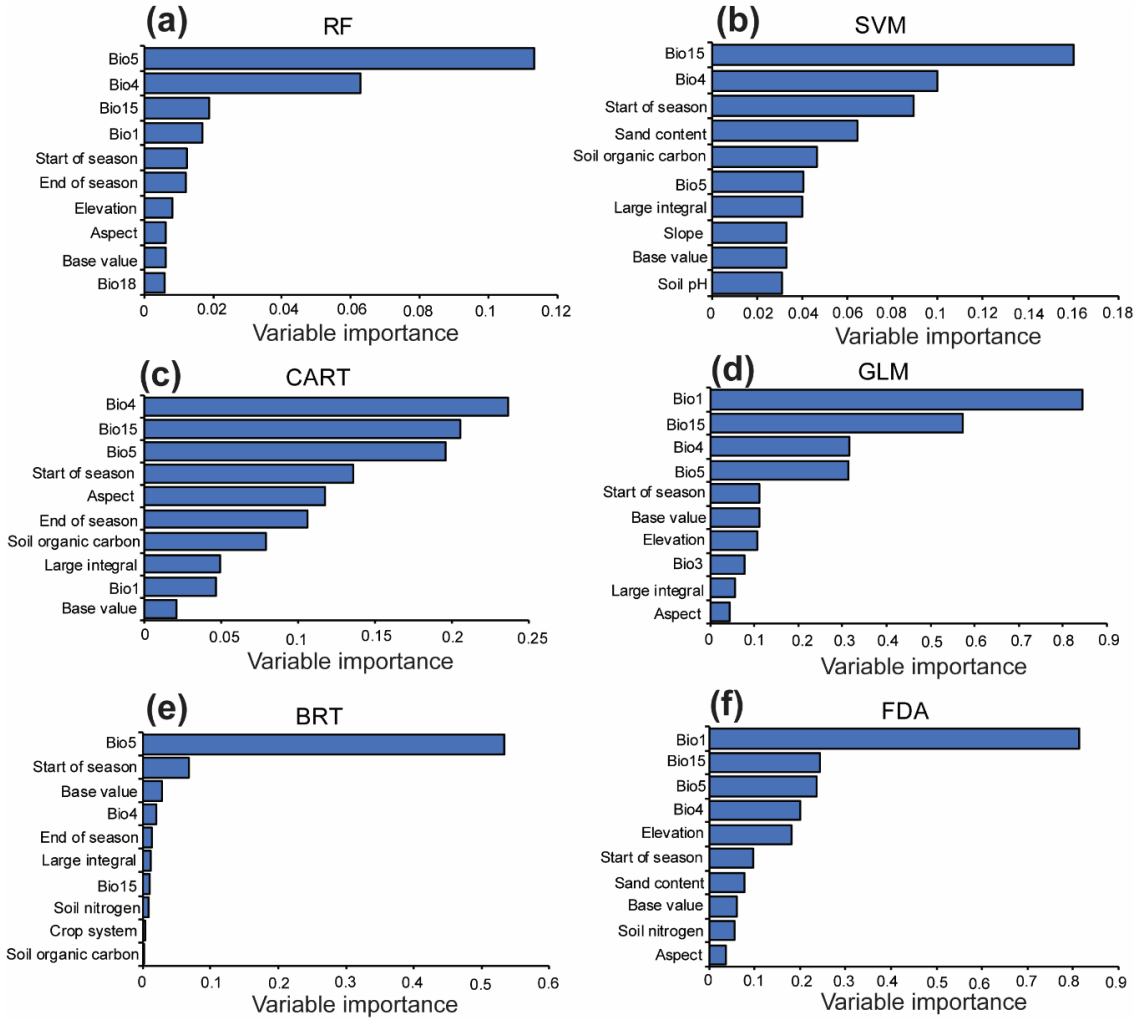


Figure 5.5: The ten most important variables for the six ecological niche model algorithms used to predict Striga occurrence in Zimbabwe; namely (a) random forest (RF), (b) support vector machines (SVM), (c) classification and

regression trees (CART), (d) generalized linear model (GLM), (e) boosted regression trees (BRT), and (f) flexible discriminant analysis (FDA).

5.4.3 Ecological niche models for predicting the occurrence of Striga using the current climate scenario

The six ecological niche models using the 21 predictor variables exhibited varied results for predicting Striga probability of occurrence (Figure 5.6). However, all six models predicted the ecological niche and Striga occurrence to be within the central plateau (mainly ecological region II, III, and IV) of the country's main watershed as shown by the warmer colors (yellow, orange, and red) in Figure 5.6. Areas close to the boundaries of Zimbabwe (ecological region V) represented by the cooler colors (green) were predicted to be relatively safe and free from potential Striga infestation using SVM and CART, while in the eastern highlands of Zimbabwe (i.e. ecological region I and II), the occurrence of Striga was predicted using RF, FDA, and BRT.

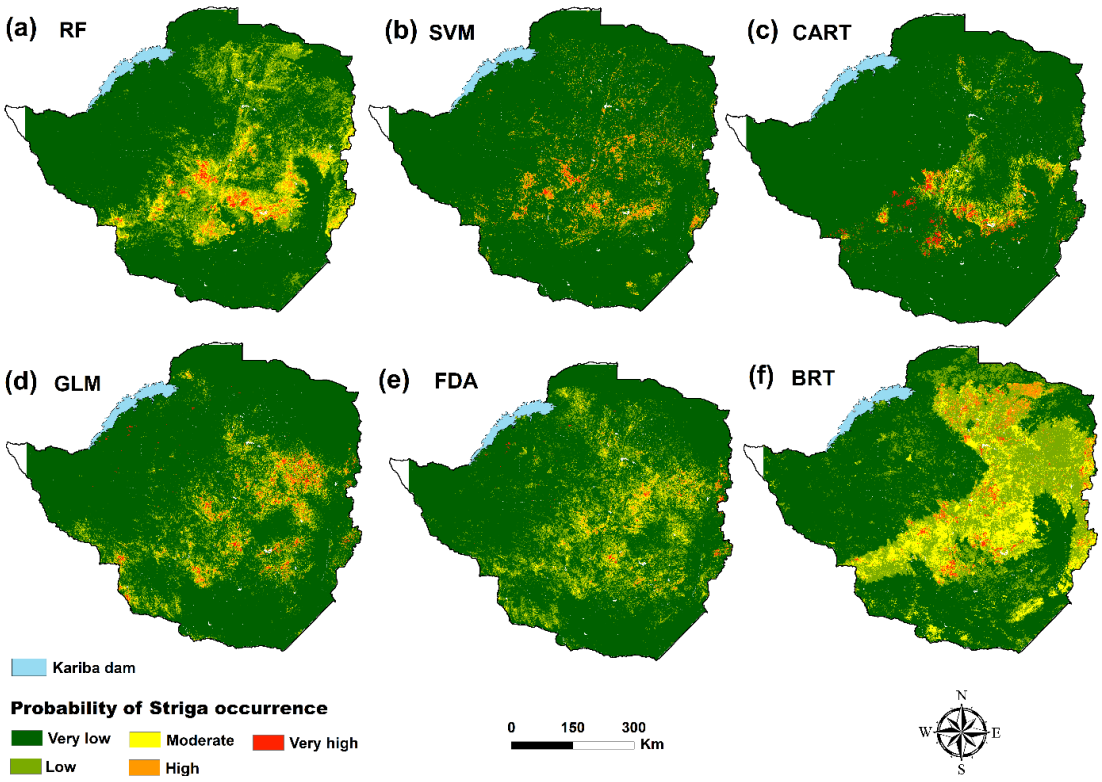


Figure 5.6: Striga probability of occurrence using the current remotely sensed and bioclimatic variables and the six ecological niche model algorithms: (a) random forest (RF), (b) support vector machines (SVM), (c) classification and regression trees (CART), (d) generalized linear model (GLM), (e) flexible discriminant analysis (FDA), and (f) boosted regression trees (BRT).

5.4.4 Ensemble projection for predicting the occurrence of Striga using the current climate scenario

The results of the ensemble projection of the six models using the current climate scenario (1950–2000) combined the best predictions of all the models and estimated the overall Striga probability of occurrence (Figure 5.7). The highest prevalence and probability of occurrence was predicted to be in the Midlands and Masvingo provinces which are in regions III and IV that are regions with very low intensity of irrigation agriculture. However, some relatively similar predictions were also observed in Matabeleland North province towards the Kariba dam which has climate characteristics of ecological region III. Similarly, parts of the provinces of Manicaland (ecological region I and II), Bulawayo (ecological region IV), and Mashonaland East (ecological region II and III) exhibited moderate, high to very high probabilities of potential Striga incidences. The highest probability of occurrence was observed in agroecological regions I, II, III, and IV, whereas very little to none Striga probabilities of occurrence were predicted in region V. Interestingly, the ensemble model was precise (AUC = 0.98) in predicting Striga occurrence following the boundaries of region V where the Striga occurrence is predicted to very low. The area towards the west of region IV was predicted to have a very low Striga probability of occurrence, whereas the central and eastern areas within region IV were predicted to have high to very high incidences of Striga occurrence. In general, the warmer colors also showed that the Striga probability of occurrence is skewed towards the central and eastern regions of the country with relatively high-altitudes (800 m–1 600 m a.s.l) compared to the low-altitude areas (< 800 m a.s.l) on the west, south, and north (Figure 5.7).

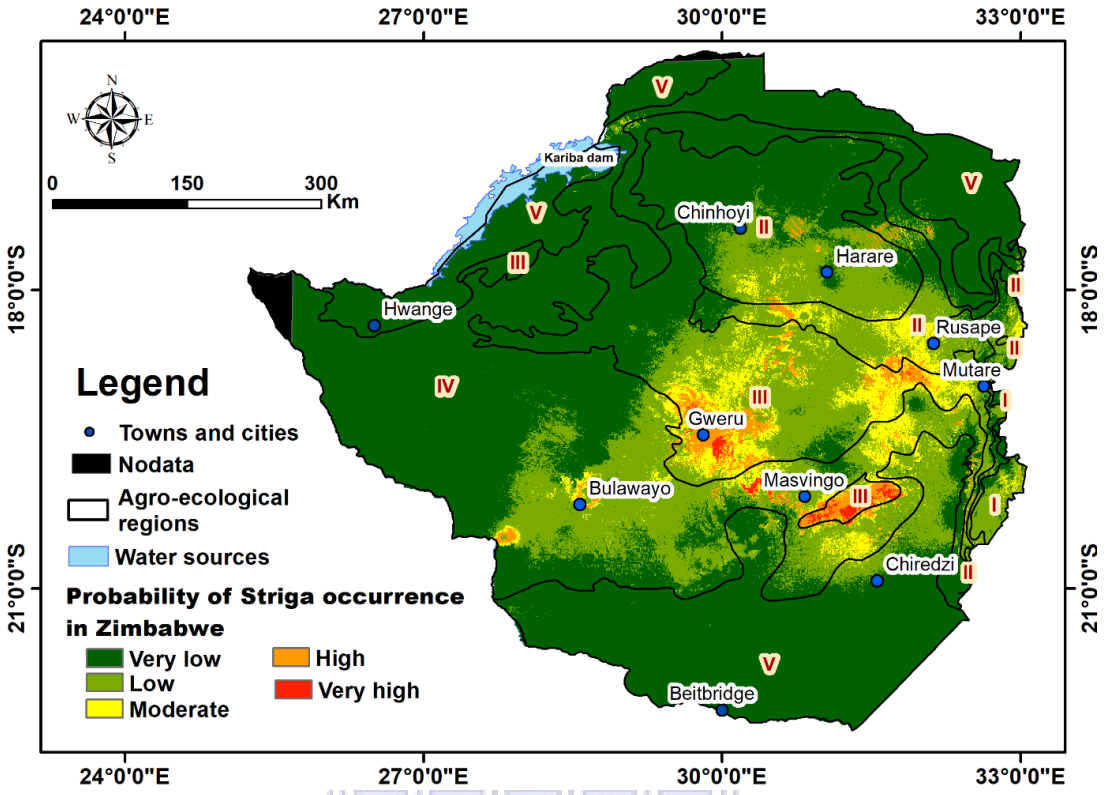


Figure 5.7: Current Striga probability of occurrence predicted using ensemble projection and the weighted average of the true skill statistics (TSS) of the six prediction models, *viz.* random forest, support vector machines, classification and regression trees, generalized linear model, flexible discriminant analysis, and boosted regression trees ecological niche model algorithms.

5.4.5 Comparison of the ensemble predictions using the current and future climate scenarios

Slight differences in the suitable area between the current and future climate scenarios for Striga occurrence in Zimbabwe were detected. It is observed that Striga occurrence would shift towards the North, i.e. Mashonaland West and East, which are in ecological region II (Figure 5.8) and will be reduced in the southern regions of the country, i.e. Matabeleland North and South occurring in ecological region IV (Figure 5.8). The future model predicted a very high increase in the area that shall be suitable for Striga, particularly for Masvingo and Midlands provinces which are in the ecological region III. It is noted that the intensity of the severity as evidenced by the increase of most areas from the moderate class to very high probability was particularly in ecological region III.

The current area suitable for Striga occurrence in Zimbabwe is 7.4% of the total area, while an increase of ~ 0.73% is likely by the end of 2050 due to climate change. Therefore, the approximate

area currently suitable for the occurrence of Striga is 28 916 km^2, while it is expected to increase to 31 768 km^2 (8.13%) by the year 2050 using the maximum carbon emission scenario (RCP8.5). The estimated increase of the area occupied by Striga due to climate change by the year 2050 is 2 852 km^2 which is an estimated gradual increase rate of ~95 km^2/yr^{-1} over 30 years.

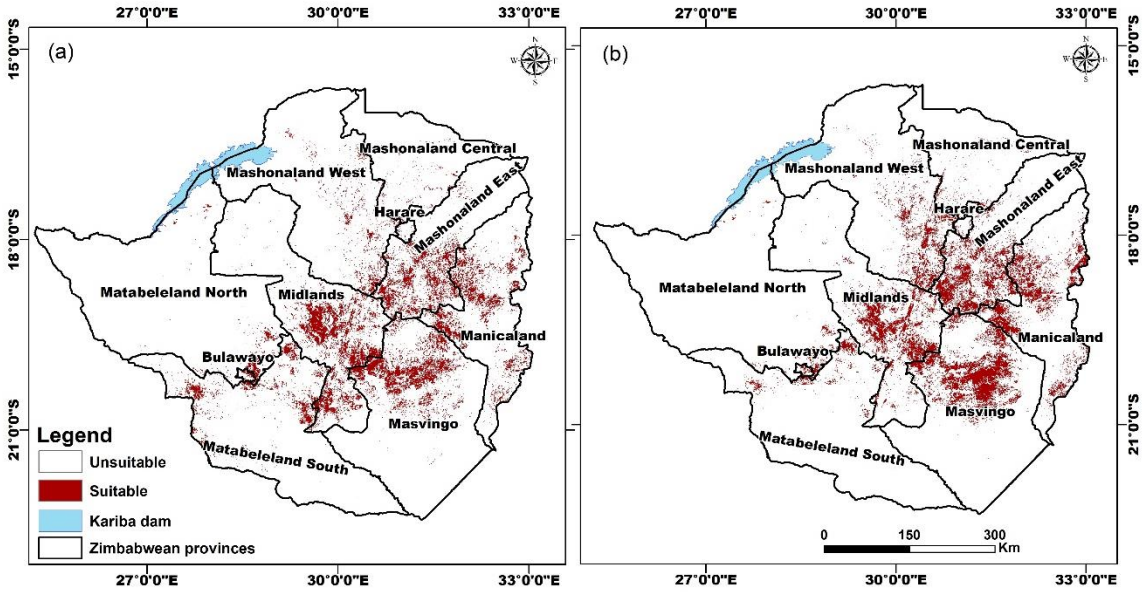


Figure 5.8: Striga probability of occurrence predicted using the representative concentration pathway (RCP:8.5), ensemble projection, and the weighted average of the true skill statistics (TSS) of six ecological niche model algorithms, *viz*. random forest, support vector machines, classification and regression trees, generalized linear model, flexible discriminant analysis, and boosted regression trees. (a) Current (1950–2000) and (b) future (2041–2060) climate scenario.

5.5 Discussion

In this chapter, six machine learning and ecological niche modeling models were used to predict the current Striga probability of occurrence in Zimbabwe. The best practice standards for ecological niche modeling were followed by assessing the quality of the response variables, predictor variables, model evaluation ideals, and building multiple models using the same data following the protocols suggested by Araújo et al. (2019). Although a sufficient sample size required for accurate predictions when using ecological niche models at a national scale (Stockwell and Peterson, 2002) was used, the performance of some models like FDA and GLM which require a relatively large sample size could have been reduced.

5.5.1 Model performances

Generally, predictive models with AUC and TSS values larger than 0.7 suggest good predictive and simulation performance (Elith et al. 2010). In this chapter, AUC and TSS values for all the six models were above the 0.7 threshold, demonstrating that the models were exceptional for simulating the distribution of Striga in Zimbabwe. As expected, the model accuracies and the predicted areas differed across the six models, since models depend on different mathematical functions and tuning parameters (Araújo et al., 2019). Using AUC and TSS, the results pointed to RF as the best predictive model, which was consistent with the hypothesis of the present chapter. Based on the obtained AUC and TSS results from this chapter, we recommend the use of the ensemble, RF, SVM, and CART for Striga predictive modeling using multi-source data. These recommended methods have also been used and suggested by many researchers as the best for simulating predictions for invasive weed species occurrence and mapping their geographical niches (Mudereri et al. 2019a; Tesfamichael et al. 2018; Landmann et al. 2020; Guan et al. 2020).

Importantly, it is, however, noted that there were huge overlaps and similarities in the areas anticipated to be suitable for Striga occurrence. These varied outputs and accuracy results are in agreement with other studies that have used the multiple models approach in ecological niche modeling (Hao et al. 2019; Jafarian et al. 2019; Mohammadi et al. 2019; Guan et al. 2020). Jafarian et al. (2019) used four predictive models to simulate the occurrence of five dominant plant species in Iran and concluded that the ensemble method yielded high predictive power compared to the individual models. On the other hand, Mohammadi et al. (2019) also compared MaxEnt and 'sdm' to predict two rodent species and they established that all models were comparable and demonstrated high predictive power. Similarly, in this chapter, there is no convincing indication to prove that one model is significantly better than the other. Therefore, regarding the future investigations that will focus on the accuracy of Striga occurrence and prediction, it is recommended to include several models in an ensemble approach to reduce the modeling uncertainties.

5.5.2 Striga probability of occurrence in the current climate scenario and under climate change

Fundamentally, input data preparation is key to determine and improve the accuracy and

dependability of the outputs derived from predictive models (Araújo et al., 2019). ecological niche models reflect the deep interrelationships and interactions among species and their environmental parameters. Using the package 'usdm' and the 'vifcor' function provided an easy and practical way of eliminating the correlated variables systematically (Jafarian et al. 2019). Specifically, the use of VIF as a measure of collinearity and elimination of conflating variables improved the accuracy of the modeling results. This is per other studies that have successfully employed VIF to select a few noncorrelated predicted variables (Abdelaal et al., 2019; Muposhi et al., 2016). The non-conflating variables (n = 21) that were finally used in the modeling experiments were crucial in explaining the occurrence of the Striga weed.

Notwithstanding, the important variables that were selected by the models for mapping Striga occurrence were local and not global. That means the variables were only relevant for modeling Striga in Zimbabwe and not somewhere else on the globe. As anticipated, the results showed that the interrelationship between temperature (Bio1, Bio4, and Bio5) and precipitation (Bio15) was central in defining the ecological niche of the Striga weed. This concurred with the results reported by Cotter and Sauerborn (2012), who alluded to the variation in the current and future distribution of Striga in the entire African continent to be influenced by mean annual temperature (Bio1). However, the future Striga prediction models should be interpreted with some caution as they were not yet validated.

Striga requires both optimum rainfall and temperature for germination, growth, propagation, and simultaneously the growth of its hosts i.e. cereal crops (Mandumbu et al. 2017b). However, extreme temperature and heavy rainfall conditions limit the propagation of Striga (Rich and Ejeta, 2007); hence, the very low probability of Striga occurrence in ecological region V of Zimbabwe. Region V, which is mostly on the borders of Zimbabwe, experiences very high temperatures and low rainfall making it unsuitable for crop production, hence the unavailability of cereals that are Striga hosts. However, with the increase in temperatures anticipated through climate change, most farmers in all agroecological regions of Zimbabwe are likely to shift to planting C_4 crops like sorghum and millet which are drought-resistant but are attractive to the occurrence of Striga (Mandumbu, 2017).

Agroecological regions I–IV experience very high to moderate rainfall and temperature compared

to ecological region V (Mugandani et al., 2012). This could have been the reason for the high prediction of Striga occurrence in these regions. These regions have varied climatic conditions, but the modeling and mapping results showed that ecological regions II, III, and IV have the most optimum climatic conditions for the germination, growth, reproduction, and spread of Striga species. Because of the immense dependence on the distribution of Striga on climatic variables, future climate conditions may greatly determine the suitable niche for Striga (Mandumbu et al. 2017a; Cotter and Sauerborn 2012).

Because of climate change, the increase in carbon dioxide and temperature changes are likely to lead to an increase in the germination and spread rates of Striga in areas that were once non-Striga occurrence areas. The results agreed with the perception reported by Mandumbu et al. (2017a), who argued that the future increase in temperature would increase the breaking of dormancy of Striga seeds, thereby increasing its germination rate. As could be seen from the results, Striga shall occupy new adjacent areas to the already infested areas, mostly in regions III and IV of Zimbabwe. These areas are predicted to have temperature ranges between 20^0C and 35^0C. This phenomenon is likely to result in increased areas occupied by Striga and enhance the intensity and severity of the crop losses caused by this weed. Additionally, as the temperature continues to increase in the future, crops that are currently not affected by Striga such as the winter wheat may eventually become susceptible to the weed infestation (Mandumbu et al. 2017a). Therefore, any efforts targeted at curbing the spread of Striga in the future should focus on areas with the likelihood of temperature increase and a reduced amount of rainfall.

Soil N is reported to constrain the germination of Striga seeds by reducing the production of Strigolactones, the chemical that is produced by the host plants to simulate the germination of Striga seeds, while soil N also increases the vegetative growth of the host plant (Ekeleme et al., 2014). Notwithstanding, the results indicated that the edaphic factors that were tested (i.e. soil N, soil pH, soil organic C, and soil sand content) had little effects on the Striga modeling accuracy, probably because of the interplay between them and precipitation (i.e. Bio15). To the best of the candidate knowledge, soil moisture affects nutrient motility, particularly nitrogen, which is mostly unstable and susceptible to leaching (Yoneyama et al., 2007).

It is worth noting that precipitation could be a proxy for soil moisture that is an important predictor

variable for triggering the germination of Striga seeds. It is expected that the edaphic factor variables and precipitation are intimately interlinked. Hence, they tend to result in low performance of each other in predictive modeling, due to the relatively high correlation. However, various studies argue that degraded soils of high acidity promote the growth and proliferation of Striga weed (Midega et al. 2017; Larsson 2012; Yoneyama et al. 2007). On the other hand, it is reported that the degradation of soils and increasing its soil pH is expected to worsen with climate change (Mandumbu et al. 2017a). As earlier mentioned, in response to the low soil fertility and drought, farmers are likely to shift to C_4 plants which are more tolerant of droughts and high temperatures but are more susceptible to the Striga infestation.

This chapter also shows that 'start of season' and EVI 'base value' were among the most important predictor variables in all the six ecological niche modeling algorithms. Striga depends on the availability of the host cereal crops for its germination, survival, and propagation. The minimum level of greenness in the whole season and the level of greenness at the start of the season can be described by the EVI values which foretell crop planting date and crop health (e.g. Striga infestation rate). Similarly, the minimum value of EVI during the season signifies the crop health status. Thus, the start of season and EVI base value can be very relevant variables to predict the occurrence of Striga in semi-arid environments.

5.5.3 Implications of the study

Modeling the potential distribution of weeds such as Striga is useful in agricultural management systems in areas most likely to be susceptible to invasion and colonization. This chapter supports national scale preparedness and management strategies for the protection of key crops from diseases and pests, especially in the face of climate change and variability. Furthermore, results from the present chapter show that using ecological niche modeling is one of the most reliable and central tools for determining the fundamental and realized niche of Striga within a geographical space. This chapter showed that Striga spread and propagation is likely to be within the adjacent areas in the ecological region III of Zimbabwe.

Although this cannot precisely be empirically derived from the current models, it is inferred and anticipate that wind, water, animal, and human movement shall be the main modes of Striga seed spread. Striga plants are highly productive with the potential to produce between 10 000 and 200

000 seeds per plant (Ejeta and Butler 1993). These seeds are of lightweight (~ 4-7 µg per seed), which makes them easy to disperse by wind water or animals (Mandumbu et al. 2017a; Wan and Wang 2019). Similarly, farmers within the same or adjacent areas are likely to exchange farming equipment, thereby promoting the spread of Striga seeds. Additionally, since the soils in the ecological region III of Zimbabwe are deep Kalahari sands, further degradation of these soils which leads to losses in soil fertility and ultimately could promote the spread of Striga (Yoneyama et al., 2007) was anticipated. To combat Striga occurrence and spread particularly in regions II, III, and IV of Zimbabwe, the results of this chapter can be utilized for guiding the implementation and upscaling of 'push-pull' technology (PPT).

PPT is a climate-smart integrated farming system that uses the legume *Desmodium* as an intercrop to combat the reproduction cycle of Striga and repel insect pests i.e. stemborers and fall armyworm (Khan et al., 2014). *Desmodium* secretes a set of compounds that promote the suicidal germination of Striga and effectively inhibits the possibilities of the Striga to attach their roots to the roots of the host plant. Interested readers are referred to Khan et al. (2006) and Pickett et al. (2014), for elaborate information on the PPT farming system.

5.6 Conclusions

The machine learning ecological niche models i.e. RF, CART, SVM, BRT, GLM, FDA, and their respective ensemble were compared for predicting the probability of Striga occurrence in Zimbabwe using multi-source bioclimatic and remotely sensed data. It is established that RF, CART, SVM, and the ensemble approach, yield the most accurate Striga occurrence prediction results in Zimbabwe. The results showed that temperature and precipitation are the key drivers of the occurrence of Striga. Also, the Striga epidemic in Zimbabwe is highly likely to worsen and spread into new areas where it was not initially found, particularly in the ecological regions I to IV of the country. Therefore, immediate and palliative action is critical to contain and manage the spread and intensity of Striga in Zimbabwe.

The results could help researchers, policymakers, extension officers, and various other stakeholders to employ and implement effective and early Striga management options to contain and eradicate the weed. Because this chapter employed a synoptic approach at a national scale using datasets at a coarse spatial resolution (250 m x 250 m pixel size), future studies should focus

on developing localized early warning advisory platforms and high resolution (i.e. submeter) remotely sensed observations to detect and monitor Striga infestation and density. Specifically, the use of UAVs should be investigated for appropriate use to early detect Striga occurrence and suitable habitats before its flowering stages. Although Zimbabwe was used as a case study, the modeling results can be upscaled in other African countries that possess similar agroecological characteristics.

Chapter 6

Striga weed detection and modeling across different agroecological farming systems: A synthesis

6.1 Summary of findings

The results of the study showed that in-situ hyperspectral remote sensing is an effective means to discriminate Striga weeds and provided new avenues to track weed invasions and improve their characterization in complex heterogeneous mixed cropping systems. Specifically, this research demonstrated the unique capabilities and application of in-situ hyperspectral remotely sensed data, as one of the critical baseline data sources for the detection and discrimination of co-occurring vegetation that exhibits similar spectra such as Striga occurrence and other weeds and crop classes. A spectral library demonstrating the spectral behavior of Striga at a canopy level was generated from the in-situ hyperspectral data collected, using a FieldSpec® Handheld 2™ analytical spectral device. Coupling the in-situ hyperspectral wavebands and their visible and near-infrared (VNIR) derived indices with machine learning discriminant algorithms (i.e. in this case random forest (RF), linear discriminant analysis (LDA), gradient boosting (GB), and support vector machines (SVM)) demonstrated plausible capability and high accuracy for discriminating among different levels of Striga (i.e. *S. hermonthica*) infestations in maize fields in western Kenya.

Further, Sentinel-2 waveband configurations simulated from the in-situ hyperspectral data demonstrated credible capabilities for use in mapping and discriminating Striga infestation in heterogeneous cereal crop fields. Despite the challenges due to spectral behavior similarities among Striga and the co-occurring plants, seven Striga infestation classes based on three flowering Striga classes (low, moderate, and high) against two background endmembers were successfully detected and distinguished. The use of canopy-level in-situ hyperspectral data to predict the presence and level of Striga infestation based on flowering characteristics provided the critical windows for Striga weed detection and mapping. This was particularly possible during the peak flowering period. However, the reliance on this phenological stage hinders the timely Striga weed mapping for meaningful interventions before crop damage.

Also, the study showed that the RF model yielded slightly better Striga weed detection and mapping results compared with other different machine learning algorithms. This was the case from the use of both the in-situ hyperspectral and resampled Sentinel-2 multispectral wavebands. This was necessitated by the improvement brought by using the reduced dimensionality of the hyperspectral data based on the guided regularized random forest (GRRF). These positive findings from the use of in-situ hyperspectral data prompted the interest to test the utility of Sentinel-2

imagery to detect Striga occurrence at a landscape scale.

Moreover, the use of Sentinel-2 and the constellation of dove nanosatellites i.e. PlanetScope data, in Striga weed (*S. hermonthica*) mapping within intercropped maize fields proved to be relatively important in mapping the Striga occurrence, although PlanetScope data demonstrated slightly higher Striga detection capacity. These results were expected since PlanetScope has a higher spatial resolution (3 m x 3 m) when compared to the Sentinel-2 (10 m x 10 m) spatial resolution. Nonetheless, these findings proved that Sentinel-2 can operationally be used in Striga weed mapping as it is readily and freely available. However, errors of commission between Striga and non-Striga fields caused by similarities in the spectral behaviors of the crops and other confounding factors in this study prompted further investigations using a hierarchical approach. The hierarchical approach involved masking out croplands and using subpixel-unmixing algorithms to improve the detectability of Striga. Therefore, a two-step hierarchical approach was employed, using multidate and multiyear (2017 to 2018) Sentinel-2 data to firstly classify and obtain a cropland mask using RF in cloud-based Google Earth Engine (GEE), and then depict Striga occurrence and infestation using the subpixel multiple endmember spectral mixture analysis (MESMA). The hierarchical classification approach proved to be well suitable for Striga weed detection in a heterogeneous agroecological system.

In addition, the high potential of the MESMA algorithm to decompose mixed pixels and detect Striga occurrence and infestation levels were demonstrated. Despite the few misclassifications realized when using the two-step hierarchical approach, the masking out of the other non-target classes belonging to other different land cover/ land use classes before the implementation of MESMA demonstrated to be a more focused and spatially adjusted spectral unmixing procedure that proved to enhance classification accuracy and reduce spectral confusion. Also, the effectiveness of the GEE as a data curation and cropland mapping platform was inferred. Additionally, deriving endmembers (EMs) using in-situ hyperspectral data provided more realistic Striga spectra rather than deriving the EMs from the image itself in light of the noise from non-target features that often hinders the acquisition of high-quality images. Despite the good performance of the two-step hierarchical method, the approach was found inapplicable for *S. asiatica* detection and mapping as this weed occurs completely covered underneath the cereal crops. The use of ecological niche machine learning and the ensemble modeling techniques is

fundamental for integrating multi-source data and predicting possible crop production restraints due to the Striga weed propagation. It is thus commendable to use the ensemble modeling approach to ensure harmonization and reduction of predictive errors from individual models when a multi-model approach is used. The study integrated bioclimatic and remotely sensed variables to predict Striga occurrence using six ecological niche machine learning algorithms and their ensemble model. It was established that RF, SVM, classification, and regression trees (CART), and the ensemble models performed the best for mapping Striga (*S. asiatica*) occurrence. Furthermore, temperature and precipitation were depicted as the key drivers of the Striga occurrence in agroecological systems, with projections indicating that that by the year 2050, the suitable area for Striga propagation will extend into new agroecological regions.

Overall, this study provides baseline information required for Striga weed management to reduce further cereal crop losses in Africa.

6.2 Conclusions

The use of the relatively new multispectral systems and image processing techniques, integrated with non-parametric machine learning algorithms, accurately improves Striga detection and mapping. This is particularly necessary for resource-scarce areas, a previously challenging task with old traditional multispectral systems. Based on the findings in the different chapters of this study, the following was concluded:

1. Levels of Striga infestation (low, moderate, and high) can accurately be detected and predicted using machine learning algorithms and in-situ hyperspectral data at a plot level,

2. The narrow-waveband spectral variables; WBI, ARI, ARI2, Datt4, DDN, and PRI.CI2, 415 nm, 548 nm, 551 nm, 556 nm, 568 nm, 578 nm, 657 nm, 677 nm, and 1 060 nm plus the broad wavebands of Sentinel-2; band 3, band 4, band 9, band 1, band 5 and band 2 (in order of importance) are very relevant for predicting Striga infestation in maize fields in semi-arid agroecosystems,

3. The machine learning GRRF algorithm is deemed to be a powerful feature selection method that can be used to reduce the redundancy in the complex multi-dimensional 1-

nm hyperspectral datasets,

4. The relatively new multispectral Sentinel-2 sensor remains a potential primary data source for estimating Striga occurrence and infestation level, particularly when integrated with ancillary datasets such as bioclimatic data at multiple mapping scales,

5. Besides, the study has demonstrated the strength associated with machine learning algorithms for discriminating Striga at multiple scales within heterogeneous cropping systems such as those used by the small-holder farmers in Africa,

6. The GEE is an effective data curation and cropland characterization platform using RF,

7. The MESMA algorithm provides the commendable potential to decompose mixed pixels and detect Striga occurrence and infestation levels at both field and landscape scales,

8. The masking out of non-target classes of different land cover classes before implementing the MESMA algorithm, allows for a more focused and spatially adjusted spectral unmixing procedure, and enhances classification of the target class (i.e. Striga) and reduces spectral confusion,

9. This study revealed that it is not feasible to directly detect *Striga asiatica* using satellite imagery because this Striga species grows underneath the crops. Thus, for *S. asiatica*, remotely sensed ancillary data and machine learning algorithms have plausible predictive abilities of the ecological niche and potentially susceptible areas of Striga invasion. Specifically, the RF, CART, SVM, and the ensemble modeling approaches yield the most accurate Striga occurrence prediction results, and

10. Overall, the findings of the present study demonstrated the importance of the earth observation and geospatial modeling algorithms as important and powerful tools that help to detect and map Striga weeds within heterogenous croplands at different mapping scales. This study underscores the utility of multi-source remotely sensed data for providing invaluable datasets for regional-scale Striga occurrence predictions.

6.4 The future

Multi-source ancillary data and the relatively new multispectral systems present an effective and robust primary data source required for characterizing Striga weeds in small-holder farms. The findings of this study provide the necessary insight and motivation to the agricultural, remote sensing, ecology, and environment communities to shift towards embracing the use of the cheap and readily-available remotely sensed data necessary for reliable and accurate monitoring of the Striga weed in data-scarce environments, where the use of hyperspectral and high-resolution airborne systems remains a challenge due to the associated acquisition and processing costs. This study, therefore, suggests the following recommendations for future research:

1. For more precise and reliable research-based outputs for monitoring Striga at a global scale, earth observation and geospatial modeling tools could be assessed to detect and map the condition of the infested crops rather than targeting the Striga weed itself,

2. Although the use of the field hyperspectral data in vegetation studies is not new, the spectral library generated in this study could be applied in other areas of similar climatic conditions. Further, future studies should compare the use of EMs from other sources like satellite imagery to test the effectiveness of other EM selection mechanisms such as the endmember average root mean squared error (EAR), the minimum average spectral angle (MASA), and the count based endmember selection (CoB) that have been tested in other studies and shown to provide pure spectra for use as EMs,

3. The EMs (i.e. spectral libraries) developed in this study can be used to craft and design a Striga-specific vegetation index that could be up-scaled to detect and map Striga weed using Satellite-based multispectral systems such as Sentinel-2,

4. Models for Striga infestation fraction estimates developed in this study could be tested for better precision in areas of low, moderate, and high Striga infestations using high resolution airborne- and UAV multispectral data,

5. Striga detection and mapping approaches developed in the present study should be further upscaled and projected at different points in space and time. In other words,

models developed in the present study should be tested in other sites or the same sites but under different cropping seasons, and

6. The Striga prediction, detection, and mapping modeling frameworks developed in the present study should be integrated with crowdsourced ‘big data’ and the advanced artificial intelligence (AI) algorithms to develop cloud-based dashboards and mobile applications (apps) for effective and near-real-time Striga monitoring at various scales.

References

Abdel-Rahman, E.M., Ahmed, F.B., Ismail, R., 2013a. Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. International Journal of Remote Sensing 34, 712–728. https://doi.org/10.1080/01431161.2012.713142

Abdel-Rahman, E.M., Landmann, T., Kyalo, R., Ong'amo, G., Mwalusepo, S., Sulieman, S., Ru, B. Le, 2016. Predicting stem borer density in maize using RapidEye data and generalized linear models. International Journal of Applied Earth Observation and Geoinformation 57, 61–74. https://doi.org/10.1016/j.jag.2016.12.008

Abdel-Rahman, E.M., Mutanga, O., Adam, E., Ismail, R., 2014a. Detecting *Sirex noctilio* grey-attacked and lightning-struck pine trees using airborne hyperspectral data, random forest and support vector machines classifiers. ISPRS Journal of Photogrammetry and Remote Sensing 88, 48–59. https://doi.org/10.1016/j.isprsjprs.2013.11.013

Abdel-Rahman, E.M., Mutanga, O., Odindi, J., Adam, E., Odindo, A., Ismail, R., 2014b. A comparison of partial least squares (PLS) and sparse PLS regressions for predicting yield of Swiss chard grown under different irrigation water sources using hyperspectral data. Computers and Electronics in Agriculture 106, 11–19. https://doi.org/10.1016/j.compag.2014.05.001

Abdel-Rahman, E.M., Way, M., Ahmed, F., Ismail, R., Adam, E., 2013b. Estimation of thrips (*Fulmekiola serrata* Kobus) density in sugarcane using leaf-level hyperspectral data. South African Journal of Plant and Soil 30, 91–96. https://doi.org/10.1080/02571862.2013.803616

Abdelaal, M., Fois, M., Fenu, G., Bacchetta, G., 2019. Using MaxEnt modeling to predict the potential distribution of the endemic plant *Rosa arabica* Crép. in Egypt. Ecological Informatics 50, 68–75. https://doi.org/10.1016/j.ecoinf.2019.01.003

Abegunde, V.O., Sibanda, M., Obi, A., 2019. The dynamics of climate change adaptation in sub-Saharan Africa: A review of climate-smart agriculture among small-scale farmers. Climate 7. https://doi.org/10.3390/cli7110132

Adam, E., Deng, H., Odindi, J., Abdel-Rahman, E.M., Mutanga, O., 2017. Detecting the early

stage of *Phaeosphaeria* leaf spot infestations in maize crop using in situ hyperspectral data and guided regularized random forest algorithm. Journal of Spectroscopy 2017. https://doi.org/10.1155/2017/6961387

AGRA, 2017. Africa agriculture status report: The business of smallholder agriculture in sub-Saharan Africa (Issue 5). Nairobi, Kenya: Alliance for a Green Revolution in Africa (AGRA), Issue No. 5. Alliance for a Green Revolution in Africa.

Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology 43, 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x

Araújo, M.B., Anderson, R.P., Barbosa, M.A., Beale, C.M., Dormann, C.F., Regan, E., Garcia, R.A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R.B., Zimmermann, N.E., Rahbek, C., 2019. Standards for distribution models in biodiversity assessments. Science advances 5, 1–12. https://doi.org/10.1126/sciadv.aat4858

ASD, 2010. FieldSpec® HandHeld 2 user manual. http://www.geo-informatie.nl/courses/grs60312/material2017/manuals/600860-dHH2Manual.pdf 1–140.

Asner, G.P., Lobell, D.B., 2000. A biogeophysical approach for automated SWIR unmixing of soils and vegetation. Remote Sensing of Environment 74, 99–112. https://doi.org/10.1016/S0034-4257(00)00126-7

Atera, E.A., Ishii, T., Onyango, J.C., Itoh, K., Azuma, T., 2013. Striga infestation in Kenya : Status, distribution and management options 2, 99–108. https://doi.org/10.5539/sar.v2n2p99

Ayebare, S., Mdegela, R.H., Nyakarahuka, L., Ndimuligo, S.A., Mwakapeje, E.R., Nonga, H.E., Skjerve, E., Mosomtai, G., 2018. Ecological niche modeling as a tool for prediction of the potential geographic distribution of *Bacillus anthracis* spores in Tanzania. International Journal of Infectious Diseases 79, 142–151. https://doi.org/10.1016/j.ijid.2018.11.367

Badu-Apraku, B., Yallou, C., Oyekunle, M., Akinwale, R., Aweke, G., Kamara, A., 2015. Consistency of performance of early-maturing maize cultivars in Striga infested and Striga free environments. Canadian Journal of Plant Science 95, 1073–1084. https://doi.org/10.4141/cjps-2015-056

Baloloy, A.., Blanco, A.., Candido, C.., Argamosa, R.J.., Dumalag, J.B.L.., Dimapilis, L.L.., Paringit, E.., 2018. Estimation of mangrove forest aboveground biomass using multispectral bands, vegetation indices and biophysical variables derived from optical satellite imageries: RapidEye, Planetscope and Sentinel-2. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 4, 29–36. https://doi.org/https://doi.org/10.5194/isprs-annals-IV-3-29-2018

Bangira, T., Van Niekerk, A., Menenti, M., Alfieri, S.M., 2019. Comparing thresholding with machine learning classifiers for mapping complex water. Remote Sensing 1–21. https://doi.org/10.3390/rs11111351

Bannari, A., Morin, D., Bonn, F., Huete, A.R., 1995. A review of vegetation indices. Remote Sensing Reviews 13, 95–120. https://doi.org/10.1080/02757259509532298

Baugh, W.M., Groeneveld, D.P., 2006. Broadband vegetation index performance evaluated for a low-cover environment. International Journal of Remote Sensing 27, 4715–4730. https://doi.org/10.1080/01431160600758543

Belgiu, M., Csillik, O., 2018. Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis. Remote Sensing of Environment 204, 509–523. https://doi.org/10.1016/j.rse.2017.10.005

Best, R.. G., Wehde, M.. E., Linder, R.., 1981. Spectral reflectance of hydrophytes. Remote Sensing of Environment 11, 27–35. https://doi.org/10.1016/0034-4257(81)90004-3

Bey, A., Jetimane, J., Lisboa, S.N., Ribeiro, N., Sitoe, A., Meyfroidt, P., 2020. Mapping smallholder and large-scale cropland dynamics with a flexible classification system and pixel-based composites in an emerging frontier of Mozambique. Remote Sensing of Environment 239, 111611. https://doi.org/10.1016/j.rse.2019.111611

Blackburn, G.A., 2007. Hyperspectral remote sensing of plant pigments. Journal of Experimental Botany 58, 855–867. https://doi.org/10.1093/jxb/erl123

Breiman, L., 2002. Manual on setting up, using, and understanding random forests v3. 1. Technical Report, http://oz.berkeley.edu/users/breiman, Statistics Department University of California Berkeley, … 29. https://doi.org/10.2776/85168

Breiman, L, 2001. Random forests. Machine learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., J.Stone, C., Olshen, R.A., 1984. Classification algorithms and regression Trees. Wadsworth Statistics/Probability 246–280.

Brereton, R.G., Lloyd, G.R., 2010. Support vector machines for classification and regression. The Analyst 135, 230–267. https://doi.org/10.1039/B918972F

Campbell, J., Wynne, R., 2007. Introduction to Remote sensing, 5th ed, Journal of Experimental Psychology: General. Guiford Press, New York.

Cao, J., Liu, K., Zhu, Y., Li, J., He, Z., 2018. Identifying mangrove species using field close-range snapshot hyperspectral imaging and machine-learning techniques. Remote Sensing 10, 2047. https://doi.org/10.3390/rs10122047

Carvalho, S., Schlerf, M., van der Puttena, W.H., Skidmore, A.K., 2013. Hyperspectral reflectance of leaves and flowers of an outbreak species discriminates season and successional stage of vegetation. International Journal of Applied Earth Observation and Geoinformation 24, 32–41. https://doi.org/10.1016/j.jag.2013.01.005

CGIAR-CSI, 2019. SRTM. http://srtm.csi.cgiar.org/ (accessed 3.31.19).

Chemura, A., Mutanga, O., Dube, T., 2017a. Integrating age in the detection and mapping of incongruous patches in coffee (*Coffea arabica*) plantations using multi-temporal Landsat 8 NDVI anomalies. International Journal of Applied Earth Observation and Geoinformation 57, 1–13. https://doi.org/http://dx.doi.org/10.1016/j.jag.2016.12.007

Chemura, A., Mutanga, O., Dube, T., 2017b. Separability of coffee leaf rust infection levels with machine learning methods at Sentinel-2 MSI spectral resolutions. Precision Agriculture 18, 859–881. https://doi.org/https://link.springer.com/article/10.1007/s11119-016-9495-0

Chemura, A., Mutanga, O., Odindi, J., 2017c. Empirical modeling of leaf chlorophyll content in coffee (*Coffea Arabica*) plantations with Sentinel-2 MSI Data: Effects of spectral settings, spatial resolution, and crop canopy cover. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 10, 5541–5550. https://doi.org/10.1109/JSTARS.2017.2750325

Chemura, A., Mutanga, O., Odindi, J., Kutywayo, D., 2018. Mapping spatial variability of foliar nitrogen in coffee (*Coffea arabica* L.) plantations with multispectral Sentinel-2 MSI data. ISPRS Journal of Photogrammetry and Remote Sensing 138, 1–11. https://doi.org/10.1016/j.isprsjprs.2018.02.004

Chen, J., Shen, M., Zhu, X., Tang, Y., 2009. Indicator of flower status derived from in situ hyperspectral measurement in an alpine meadow on the Tibetan Plateau. Ecological Indicators 9, 818–823. https://doi.org/10.1016/j.ecolind.2008.09.009

Cian, F., Marconcini, M., Ceccato, P., 2018. Normalized difference flood index for rapid flood mapping: Taking advantage of EO big data. Remote Sensing of Environment 209, 712–730. https://doi.org/10.1016/j.rse.2018.03.006

Cochrane, V., Press, M.C., 1997. Geographical distribution and aspects of the ecology of the hemiparasitic angiosperm *Striga asiatica* (L .) Kuntze : A herbarium study. Journal of Tropical Ecology 13, 371–380. https://doi.org/10.1017/S0266467400010579

Cooley, S.W., Smith, L.C., Stepan, L., Mascaro, J., 2017. Tracking dynamic northern surface water changes with high-frequency planet CubeSat imagery. Remote Sensing 9, 1–21. https://doi.org/10.3390/rs9121306

Cotter, M., Sauerborn, J., 2012. Understanding the present distribution of the parasitic weed *Striga hermonthica* and predicting its potential future geographic distribution in the light of climate change. Julius -Kühn-Archiv 13, 630–636. https://doi.org/10.5073/jka.2012.434.082

Csillag, F., Kummert, Á., Kertész, M., 1992. Resolution, accuracy and attributes: Approaches for environmental geographical information systems. Computers, Environment and Urban Systems 16, 289–297. https://doi.org/10.1016/0198-9715(92)90010-O

de Castro, A.I., Torres-Sánchez, J., Peña, J.M., Jiménez-Brenes, F.M., Csillik, O., López-Granados, F., 2018. An automatic random forest-OBIA algorithm for early weed mapping between and within crop rows using UAV imagery. Remote Sensing 10, 1–21. https://doi.org/10.3390/rs10020285

De Groote, H., 2007. Striga economics. In: Ejeta, G. and J. Gressel (eds.), Intergrating new technologies for Striga Control: Towards ending the witch-hunt. World Scientific Publishing Company, Singapore, pp. 265–280. https://doi.org/10.1142/9789812771506

Degbelo, A., Kuhn, W., 2018. Spatial and temporal resolution of geographic information: an observation-based theory. Open Geospatial Data, Software and Standards 3. https://doi.org/10.1186/s40965-018-0053-8

Degerickx, J., Roberts, D.A., Somers, B., 2019. Enhancing the performance of Multiple Endmember Spectral Mixture Analysis (MESMA) for urban land cover mapping using airborne lidar data and band selection. Remote Sensing of Environment 221, 260–273. https://doi.org/10.1016/j.rse.2018.11.026

Deng, H., 2013. Guided Random Forest in the RRF Package 1–3.

Deng, H., Runger, G., 2013. Gene selection with guided regularized random forest. Pattern Recognition 46, 3483–3489. https://doi.org/10.1016/j.patcog.2013.05.018

Dhau, I., Adam, E., Mutanga, O., Ayisi, K., Abdel-Rahman, E.M., Odindi, J., Masocha, M., 2018. Testing the capability of spectral resolution of the new multispectral sensors on detecting the severity of grey leaf spot disease in maize crop. Geocarto International 33, 1223–1236. https://doi.org/http://doi.org/10.1080/10106049.2017.1343391

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., Mcclean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., Lautenbach, S., 2013. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. Ecography 36, 027–046. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Dube, T., Mutanga, O., 2015. Evaluating the utility of the medium-spatial resolution Landsat 8 multispectral sensor in quantifying aboveground biomass in uMgeni catchment, South Africa. ISPRS Journal of Photogrammetry and Remote Sensing 101, 36–46. https://doi.org/http://dx.doi.org/10.1016/j.isprsjprs.2014.11.001

Dube, T., Mutanga, O., Abdel-Rahman, E.M., Ismail, R., Slotow, R., 2015. Predicting *Eucalyptus* spp. stand volume in Zululand, South Africa: an analysis using a stochastic gradient boosting regression ensemble with multi-source data sets. International Journal of Remote Sensing 36, 3751–3772. https://doi.org/https://doi.org/10.1080/01431161.2015.1070316

Dube, T., Mutanga, O., Elhadi, A., Ismail, R., 2014. Intra-and-inter species biomass prediction in a plantation forest: Testing the utility of high spatial resolution spaceborne multispectral

rapideye sensor and advanced machine learning algorithms. Sensors (Switzerland) 14, 15348–15370. https://doi.org/10.3390/s140815348

Dube, T., Pandit, S., Shoko, C., Ramoelo, A., Mazvimavi, D., Dalu, T., Dube, T., Pandit, S., Shoko, C., Ramoelo, A., Mazvimavi, D., Dalu, T., 2019. Numerical assessments of Leaf Area Index in tropical savanna rangelands, South Africa using Landsat 8 OLI derived metrics and in-situ measurements. Remote Sensing 2019, Vol. 11, Page 829 11, 829. https://doi.org/https://doi.org/10.3390/rs11070829

Ejeta, G., Butler, L.G., 1993. Host-parasite interactions throughout the *Striga* life cycle, and their contributions to Striga resistance. African Crop Science Journal.

Ejeta, G., Gressel, J., 2007. Integrating new technologies for striga control: Towards ending the Witch-Hunt. World Scientific, Singapore. https://doi.org/10.1142/6470

Ekeleme, F., Jibrin, J.M., Kamara, A.Y., Oluoch, M., Samndi, A.M., Fagge, A.A., 2014. Assessment of the relationship between soil properties, *Striga hermonthica* infestation and the on-farm yields of maize in the dry Savannas of Nigeria. Crop Protection 66, 90–97. https://doi.org/10.1016/j.cropro.2014.09.001

Eklundh, L., Jönsson, P., 2017. TIMESAT 3.3 with seasonal trend decomposition and parallel processing Software Manual. Lund and Malmo University, Sweden 1–92.

Elith, J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J., Phillips, S.J., 2010. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17, 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. Journal of Animal Ecology 77, 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Evangelista, P.H., Stohlgren, T.J., Morisette, J.T., Kumar, S., 2009. Mapping invasive tamarisk (Tamarix): A comparison of single-scene and time-series analyses of remotely sensed data. Remote Sensing 1, 519–533. https://doi.org/10.3390/rs1030519

FAO, IFAD, UNICEF, WFP, and WHO., 2019. The state of food security and nutrition in the world 2019. Safeguarding against economic slowdowns and downturns. Rome, FAO Licence: CC BY-NC-SA 3.0 IGO.

FAO, IFAD, UNICEF, WFP, WHO., 2018. The state of food security and nutrition in the world 2018. Building climate resilience for food security and nutrition. Rome, FAO. Licence: CC BY-NC-SA 3.0 IGO, Building climate resilience for food security and nutrition. https://doi.org/10.1093/cjres/rst006

FAO, IFAD, UNICEF, WFP. and WHO., 2017. T The state of food security and nutrition in the world 2017. Building resilience for peace and food security. Rome, FAO.

FAO, 2019. Crop prospects and food situation, quaterly global report No.3, september 2019. Rome, FAO.

FAO, ACFD, 1999. A fertilizer strategy for Zimbabwe, Food and Agriculture Organisation of the United Nations (FAO). Rome, FAO.

FieldSpec, 2017. FieldSpec HandHeld2: Light theory and measurements using the Fieldspec HandHeld 2 portable Spectroradiometer. https://www.malvernpanalytical.com/en/learn/knowledge-center/user-manuals/Copy_of_fieldspec-handheld-2-light-theory-and-measurement.html.

Fisher, R.., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Forkuor, G., Dimobe, K., Serme, I., Tondoh, J.E., 2018. Landsat-8 vs. Sentinel-2: examining the added value of Sentinel-2's red-edge bands to land-use and land-cover mapping in Burkina Faso. GIScience and Remote Sensing 55, 331–354. https://doi.org/10.1080/15481603.2017.1370169

Franke, J., Roberts, D.A., Halligan, K., Menz, G., 2009. Hierarchical Multiple Endmember Spectral Mixture Analysis (MESMA) of hyperspectral imagery for urban environments. Remote Sensing of Environment 113, 1712–1723. https://doi.org/10.1016/j.rse.2009.03.018

Friedman, J., 1999. Greedy function approximation: A Gradient Boosting Machine 39, 561–563.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for Generalized Linear Models via coordinate descent. Journal of Statistical Software 33, 1–22. https://doi.org/URL http://www.jstatsoft.org/v33/i01/10

Ge, S., Everitt, J., Carruthers, R., Gong, P., Anderson, G., 2006. Hyperspectral characteristics of

canopy components and structure for phenological assessment of an invasive weed. Environmental Monitoring and Assessment 120, 109–126. https://doi.org/10.1007/s10661-005-9052-1

Graesser, J., Ramankutty, N., 2017. Detection of cropland field parcels from Landsat imagery. Remote Sensing of Environment 201, 165–180. https://doi.org/10.1016/j.rse.2017.08.027

Greenwell, B., Boehmke, B., Cunningham, J., Developers, G., 2019. gbm: Generalized Boosted Regression Models. R package version 2.1.5. CRAN Repository: https://cran.r-project.org/.

Große-Stoltenberg, A., Hellmann, C., Werner, C., Oldeland, J., Thiele, J., 2016. Evaluation of continuous VNIR-SWIR spectra versus narrowband hyperspectral indices to discriminate the invasive *Acacia longifolia* within a mediterranean dune ecosystem. Remote Sensing 8. https://doi.org/10.3390/rs8040334

Guan, B. cai, Guo, H. jing, Chen, S. si, Li, D. ming, Liu, X., Gong, X., Ge, G., 2020. Shifting ranges of eleven invasive alien plants in China in the face of climate change. Ecological Informatics 55, 101024. https://doi.org/10.1016/j.ecoinf.2019.101024

Gumma, M.K., Thenkabail, P.S., Teluguntla, P.G., Oliphant, A., Xiong, J., Giri, C., Pyla, V., Dixit, S., Whitbread, A.M., 2019. Agricultural cropland extent and areas of South Asia derived using Landsat satellite 30-m time-series big-data using random forest machine learning algorithms on the Google Earth Engine cloud. GIScience and Remote Sensing 00, 1–21. https://doi.org/10.1080/15481603.2019.1690780

Guo, Y., Li, X., Zhao, Z., Nawaz, Z., 2019. Predicting the impacts of climate change, soils and vegetation types on the geographic distribution of *Polyporus umbellatus* in China. Science of the Total Environment 648, 1–11. https://doi.org/10.1016/j.scitotenv.2018.07.465

Hadjimitsis, D.G., Papadavid, G., Agapiou, A., Themistocleous, K., Hadjimitsis, M.G., Retalis, A., Michaelides, S., Chrysoulakis, N., Toulios, L., I. Clayton, C.R., 2010. Atmospheric correction for satellite remotely sensed data intended for agricultural applications: Impact on vegetation indices. Natural Hazards and Earth System Science 10, 89–95. https://doi.org/10.5194/nhess-10-89-2010

Han, H., Guo, X., Yu, H., 2017. Variable selection using mean decrease accuracy and mean decrease Gini based on random forest. Proceedings of the IEEE International Conference on

Software Engineering and Service Sciences, ICSESS 219–224. https://doi.org/10.1109/ICSESS.2016.7883053

Hao, T., Elith, J., Guillera-Arroita, G., Lahoz-Monfort, J.J., 2019. A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. Diversity and Distributions 25, 839–852. https://doi.org/10.1111/ddi.12892

Hassanali, A., Herren, H., Khan, Z.R., Pickett, J.A., Woodcock, C.M., 2008. Integrated pest management: The push-pull approach for controlling insect pests and weeds of cereals, and its potential for other agricultural systems including animal husbandry. Philosophical Transactions of the Royal Society B: Biological Sciences 363, 611–621. https://doi.org/10.1098/rstb.2007.2173

Hastie, T., Tibshirani, R., Buja, A., 1994. Flexible discriminant analysis by optimal scoring. Journal of the American Statistical Association 89, 1255–1270. https://doi.org/10.1080/01621459.1994.10476866

He, K.S., Rocchini, D., Neteler, M., Nagendra, H., 2011. Benefits of hyperspectral remote sensing for tracking plant invasions. Diversity and Distributions 17, 381–392. https://doi.org/10.1111/j.1472-4642.2011.00761.x

Heinz, D.C., Chang, C.I., 2001. Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. IEEE Transactions on Geoscience and Remote Sensing 39, 529–545. https://doi.org/10.1109/36.911111

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., De Jesus, J.M., Tamene, L., Tondoh, J.E., 2015. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. PLoS ONE 10, 1–26. https://doi.org/10.1371/journal.pone.0125814

Hentze, K., Thonfeld, F., Menz, G., 2016. Evaluating crop area mapping from modis time-series as an assessment tool for Zimbabwe's "fast track land reform programme." PLoS ONE 11, 1–22. https://doi.org/10.1371/journal.pone.0156630

Ho, P.-G.P., 2009. Geoscience and remote sensing. https://www.harrisgeospatial.com/Support/Maintenance-Detail/ArtMID/13350/ArticleID/16162/Vegetation-Analysis-Using-Vegetation-Indices-in-

ENVI (accessed 2.20.19).

Holloway, J., Mengersen, K., 2018. Statistical machine learning methods and remote sensing for sustainable development goals: A Review. Remote Sensing 10, 1365. https://doi.org/https://doi.org/10.3390/rs10091365

Houborg, R., McCabe, M.F., 2018a. A cubesat enabled spatio-temporal enhancement method (CESTEM) utilizing Planet, Landsat and MODIS data. Remote Sensing of Environment 209, 211–226. https://doi.org/10.1016/j.rse.2018.02.067

Houborg, R., McCabe, M.F., 2018b. Daily retrieval of NDVI and LAI at 3 m resolution via the fusion of CubeSat, Landsat, and MODIS data. Remote Sensing 10. https://doi.org/10.3390/rs10060890

Huang, J., Wei, Chen, Zhang, Y., Blackburn, G.A., Wang, X., Wei, Chuanwen, Wang, J., 2015. Meta-analysis of the detection of plant pigment concentrations using hyperspectral remotely sensed data. PLoS ONE 10, 1–26. https://doi.org/10.1371/journal.pone.0137029

Huete, A., Didan, K., Miura, T.., Rodriguez, E.., Gao, X., Ferreira, L.G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sensing of Environment 83, 195–213. https://doi.org/10.1016/S0034-4257(02)00096-2

Immitzer, M., Vuolo, F., Atzberger, C., 2016. First experience with Sentinel-2 data for crop and tree species classifications in central Europe. Remote Sensing 8. https://doi.org/https://doi.org/10.3390/rs8030166

IPCC, 2014. Climate change 2014. Synthesis report. Versión inglés, Climate Change 2014: synthesis report. Contribution of working groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change. https://doi.org/10.1017/CBO9781107415324

Jafarian, Z., Kargar, M., Bahreini, Z., 2019. Which spatial distribution model best predicts the occurrence of dominant species in semi-arid rangeland of northern Iran? Ecological Informatics 50, 33–42. https://doi.org/10.1016/j.ecoinf.2018.12.011

Jia, K., Wu, B., Tian, Y., Li, Q., Du, X., 2011. Spectral discrimination of opium poppy using field spectrometry. IEEE Transactions on Geoscience and Remote Sensing 49, 3414–3422.

https://doi.org/10.1109/TGRS.2011.2126582

Jönsson, P., Eklundh, L., 2004. TIMESAT - A program for analyzing time-series of satellite sensor data. Computers and Geosciences 30, 833–845. https://doi.org/10.1016/j.cageo.2004.05.006

Jönsson, P., Eklundh, L., 2002. Seasonality extraction by function fitting to time-series of satellite sensor data. IEEE Transactions on Geoscience and Remote Sensing 40, 1824–1832. https://doi.org/10.1109/TGRS.2002.802519

Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., 2004. Kenlab - An S4 package for kernel methods in R. Journal of Statistical Software 11, 1–20.

Kelley, L.C., Pitcher, L., Bacon, C., 2018. Using Google Earth Engine to map complex shade-grown coffee landscapes in northern Nicaragua. Remote Sensing 10, 952. https://doi.org/10.3390/rs10060952

Khadka, K.K., James, D.A., 2017. Modeling and mapping the current and future climatic-niche of endangered Himalayan musk deer. Ecological Informatics 40, 1–7. https://doi.org/10.1016/j.ecoinf.2017.04.009

Khan, Z.R., Hassanali, A., Overholt, W., Khamis, T.M., Hooper, A.M., Pickett, J.A., Wadhams, L.J., Woodcock, C.M., 2002. Control of witchweed *Striga hermonthica* by intercropping with *Desmodium* spp., and the mechanism defined as allelopathic. Journal of Chemical Ecology 28, 1871–1885. https://doi.org/10.1023/A:1020525521180

Khan, Z.R., Midega, C.A.O., Amudavi, D.M., Hassanali, A., Pickett, J.A., 2008. On-farm evaluation of the "push-pull" technology for the control of stemborers and striga weed on maize in western Kenya. Field Crops Research 106, 224–233. https://doi.org/10.1016/j.fcr.2007.12.002

Khan, Z.R., Midega, C.A.O., Hassanali, A., Pickett, J.A., Wadhams, L.J., 2007. Assessment of different legumes for the control of *Striga hermonthica* in maize and sorghum. Crop Science 47, 730–736. https://doi.org/10.2135/cropsci2006.07.0487

Khan, Zeyaur R, Midega, C.A.O., Pittchar, J.O., Murage, A.W., Birkett, M.A., Bruce, T.J.A., Pickett, J.A., 2014. Achieving food security for one million sub-Saharan African poor through push–pull innovation by 2020. Philosophical Transactions of the Royal Society B: Biological

Sciences 369, 20120284. https://doi.org/10.1098/rstb.2012.0284

Khan, Zeyaur R., Midega, C.A.O., Pittchar, J.O., Pickett, J.A., 2014. Exploiting phytochemicals for developing sustainable crop protection strategies to withstand climate change: Example from Africa, in: Advances in Plant Biopesticides. Springer India, pp. 35–46. https://doi.org/DOI 10.1007/978-81-322-2006-0_3

Khan, Z.R., Pickett, J.A., Wadhams, L.J., Hassanali, A., Midega, C.A.O., 2006. Combined control of *Striga hermonthica* and stemborers by maize-Desmodium spp. intercrops. Crop Protection 25, 989–995. https://doi.org/10.1016/j.cropro.2006.01.008

Kuhn, M., Wing, J., Weston, S., Williams, A., Chris, A., Engelhardt, Tony, C., Mayer, Z., Kenke, B., The R CoreTeam, I., Michael, B., Reynald, L., Andrew, Z., Luca, S., Yuan, T., 2018. Caret: Classification and regression training. R package version 6.0-81.

Kumar, A., Manjunath, K.R., Meenakshi, Bala, R., Suda, R.K., Singh, R.D., Panigrahy, S., 2013. Field hyperspectral data analysis for discriminating spectral behavior of tea plantations under various management practices. International Journal of Applied Earth Observation and Geoinformation 23, 352–359. https://doi.org/10.1016/j.jag.2012.10.006

Kumar, L., Schmidt, K.S., Dury, S., Skidmore, A.K., 2001. Imaging spectroscopy and vegetation science, in: van der Meer, F.D., de Jong, S.M. (Eds.), Imaging spectrometry: Basic principles and prospective applications, remote sensing and digital image processing : Vol. 4. Kluwer Academic, Dordrecht, The Netherlands, pp. 111–155.

Kuri, F., Masocha, M., Murwira, A., Murwira, K.S., 2019. Differential impact of remotely sensed dry dekads on maize yield in Zimbabwe. Geocarto International 0, 1–11. https://doi.org/10.1080/10106049.2019.1583774

Kuri, F., Murwira, A., Murwira, K.S., Masocha, M., 2018. Accounting for phenology in maize yield prediction using remotely sensed dry dekads. Geocarto International 33, 723–736. https://doi.org/10.1080/10106049.2017.1299798

Kyalo, R., Abdel-Rahman, E., Mohamed, S., Ekesi, S., Christian, B., Landmann, T., 2018. Importance of remotely-sensed vegetation variables for predicting the spatial distribution of African citrus triozid (*Trioza erytreae*) in Kenya. ISPRS International Journal of Geo-Information 7, 429. https://doi.org/10.3390/ijgi7110429

Kyalo, R., Abdel-Rahman, E.M., Subramanian, S., Nyasani, J.O., Thiel, M., Jozani, H., Borgemeister, C., Landmann, T., 2017. Maize cropping systems mapping using RapidEye observations in agro-ecological landscapes in Kenya. Sensors 17, 2537. https://doi.org/https://doi.org/10.3390/s17112537

Landmann, T., 2003. Characterizing subpixel Landsat ETM+ fire severity on experimental fires in the Kruger National Park, South Africa. South African Journal of Science 99, 357–360.

Landmann, T., Eidmann, D., Cornish, N., Franke, J., Siebert, S., 2019. Optimizing harmonics from Landsat time series data: the case of mapping rainfed and irrigated agriculture in Zimbabwe. Remote Sensing Letters 10, 1038–1046. https://doi.org/10.1080/2150704X.2019.1648901

Landmann, T., Piiroinen, R., Makori, D.M., Abdel-Rahman, E.M., Makau, S., Pellikka, P., Raina, S.K., 2015. Application of hyperspectral remote sensing for flower mapping in African savannas. Remote Sensing of Environment 166, 50–60. https://doi.org/10.1016/j.rse.2015.06.006

Landmann, T., Dubovyk, O., Ghazaryan, G., Kimani, J., Abdel-Rahman, E.M., 2020. Wide-area invasive species propagation mapping is possible using phenometric trends. ISPRS Journal of Photogrammetry and Remote Sensing 159, 1–12. https://doi.org/10.1016/j.isprsjprs.2019.10.016

Larsson, M., 2012. Soil fertility status and *Striga hermonthica* infestation relationship due to management practices in Western Kenya. Swedish University of Agricultural Sciences.

Leutner, B., Horning, N., Schwalb-willmann, J., 2019. Tools for remote sensing data analysis. Rpackage version 0.2.4.

Li, L., Ustin, S.L., Lay, M., 2005. Application of multiple endmember spectral mixture analysis (MESMA) to AVIRIS imagery for coastal salt marsh mapping: A case study in China Camp, CA, USA. International Journal of Remote Sensing 26, 5193–5207. https://doi.org/10.1080/01431160500218911

Liaw, A., Wiener, M., Weiner, M., 2002. Classification and regression by randomForest. R news 2, 18–22. https://doi.org/10.1177/154405910408300516

Litchenthaler, H.., Buschmann, C., 2001. Chlorophylls and carotenoids measurement and UV-vis

characterization 1–8. https://doi.org/10.1002/0471709085.ch21

Lukas, W., Lehnert, Hanna, M., Joerg, B., 2018. hsdar: Manage, analyse and simulate hyperspectral data in R. R package version 0.7.2.

Makaya, N.P., Mutanga, O., Kiala, Z., Dube, T., Seutloali, K.E., 2019. Assessing the potential of Sentinel-2 MSI sensor in detecting and mapping the spatial distribution of gullies in a communal grazing landscape. Physics and Chemistry of the Earth 112, 66–74. https://doi.org/10.1016/j.pce.2019.02.001

Makori, D., Mutanga, O., Irungu, J., Mosomtai, G., Odindi, J., Makau, S., Nkoba, K., Abdel-Rahman, E., Fombong, A., Ongus, J., Raina, S., Landmann, T., 2017. Predicting spatial distribution of key honeybee pests in Kenya using remotely sensed and bioclimatic variables: Key honeybee pests distribution models. ISPRS International Journal of Geo-Information 6, 66. https://doi.org/10.3390/ijgi6030066

Makurira, H., 2010. Rainfed agriculture in Sub-Saharan Africa. Water productivity in rainfed Agriculture 181, 9–21. https://doi.org/10.1201/b10823-3

Mandanici, E., Bitelli, G., 2016. Preliminary comparison of Sentinel-2 and Landsat 8 imagery for a combined use. Remote Sensing 8. https://doi.org/10.3390/rs8121014

Mandumbu, R., 2017. The stability of tolerance of Sorghum spp to *Striga asiatica* L . Kuntze under diverse conditions and existence of pre- attachment resistance. University of Fort Hare South Africa.

Mandumbu, R., Mutengwa, C., Mabasa, S., Mwenje, E., 2017a. Predictions of the striga scourge under new climate in Southern Africa: A perspective. Journal of Biological Sciences 17, 194–201. https://doi.org/10.3923/jbs.2017.194.201

Mandumbu, R., Mutengwa, C., Mabasa, S., Mwenje, E., Gotosa, J., Munyati, V.T., 2017b. The parasitic weeds scourge in northern Zimbabwe: effects of soil degradation, hosts and food security implications to rural farmers. Scientia Agriculturae 20. https://doi.org/10.15192/pscp.sa.2017.20.3.8691

Masocha, M., Dube, T., Mpofu, N.T., Chimunhu, S., 2018. Accuracy assessment of MODIS active fire products in southern African savannah woodlands. African Journal of Ecology 56, 563–

571. https://doi.org/10.1111/aje.12494

Matongera, T.N., Mutanga, O., Dube, T., Sibanda, M., 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. https://doi.org/10.1016/j.jag.2016.12.006

Maxwell, A.E., Warner, T.A., Fang, F., 2018. Implementation of machine-learning classification in remote sensing: An applied review. International Journal of Remote Sensing 39, 2784–2817. https://doi.org/https://doi.org/10.1080/01431161.2018.1433343

Mbatudde, M., Mwanjololo, M., Kakudidi, E.K., Dalitz, H., 2012. Modelling the potential distribution of endangered *Prunus africana* (Hook . f .) Kalkm . in East Africa. African Journal of Ecology 50, 393–403. https://doi.org/10.1111/j.1365-2028.2012.01327.x

Midega, C.A.O., Wasonga, C.J., Hooper, A.M., Pickett, J.A., Khan, Z.R., 2017. Drought-tolerant *Desmodium* species effectively suppress parasitic striga weed and improve cereal grain yields in western Kenya. Crop Protection 98, 94–101. https://doi.org/10.1016/j.cropro.2017.03.018

Mirik, M., Ansley, R.J., Steddom, K., Jones, D.C., Rush, C.M., Michels, G.J., Elliott, N.C., 2013. Remote distinction of a noxious weed (Musk Thistle: *Carduus Nutans*) using airborne hyperspectral imagery and the support vector machine classifier. Remote Sensing 5, 612–630. https://doi.org/10.3390/rs5020612

Mohammadi, S., Ebrahimi, E., Shahriari Moghadam, M., Bosso, L., 2019. Modelling current and future potential distributions of two desert *jerboas* under climate change in Iran. Ecological Informatics 52, 7–13. https://doi.org/10.1016/j.ecoinf.2019.04.003

Moore, F.C., Baldos, U., Hertel, T., Diaz, D., 2017. New science of climate change impacts on agriculture implies higher social cost of carbon. Nature Communications 8, 1–8. https://doi.org/10.1038/s41467-017-01792-x

Mosomtai, G., Evander, M., Sandström, P., Ahlm, C., Sang, R., Hassan, O.A., Affognon, H., Landmann, T., 2016. Association of ecological factors with Rift Valley fever occurrence and mapping of risk zones in Kenya. International Journal of Infectious Diseases 46, 49–55. https://doi.org/10.1016/j.ijid.2016.03.013

Mudereri, B.T., Chitata, T., Mukanga, C., Mupfiga, E.T., Gwatirisa, C., Dube, T., 2019a. Can biophysical parameters derived from Sentinel-2 spaceborne sensor improve land cover characterization in semi-arid regions? Geocarto International. https://doi.org/10.1080/10106049.2019.1695956

Mudereri, B.T., Dube, T., Adel-Rahman, E.M., Niassy, S., Kimathi, E., Khan, Z., Landmann, T., 2019b. A comparative analysis of PlanetScope and Sentinel-2 space-borne sensors in mapping Striga weed using Guided Regularised Random Forest classification ensemble . ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-2/W13, 701–708. https://doi.org/10.5194/isprs-archives-XLII-2-W13-701-2019

Mudereri, B.T., Dube, T., Niassy, S., Kimathi, E., Landmann, T., Khan, Z., Abdel-Rahman, E.M., 2020a. Is it possible to discern Striga weed (*Striga hermonthica*) infestation levels in maize agro-ecological systems using in-situ spectroscopy? International Journal of Applied Earth Observation and Geoinformation 85, 102008. https://doi.org/10.1016/j.jag.2019.102008

Mudereri, B.T., Mukanga, C., Mupfiga, E.T., Gwatirisa, C., Kimathi, E., Chitata, T., 2020b. Analysis of potentially suitable habitat within migration connections of an intra-African migrant-the Blue Swallow (*Hirundo atrocaerulea*). Ecological Informatics 57, 101082. https://doi.org/10.1016/j.ecoinf.2020.101082

Mugandani, R., Wuta, M., Makarau, A., Chipindu, B., 2012. RE-Classification of Agro-ecological regions of Zimbabwe in conformity with climate variability and change. African Crop Science Journal 20, 361–369. https://doi.org/https://pdfs.semanticscholar.org/d24a/4e02f534cf8e9f8705690c7a5ba19145becc.pdf

Muposhi, V.K., Gandiwa, E., Chemura, A., Bartels, P., Makuza, S.M., Madiri, T.H., 2016. Habitat heterogeneity variably influences habitat selection by wild herbivores in a semi-arid tropical savanna ecosystem. PLoS ONE 11. https://doi.org/10.1371/journal.pone.0163084

Mureriwa, N., Adam, E., Sahu, A., Tesfamichael, S., 2016. Examining the spectral separability of prosopis glandulosa from co-existent species using field spectral measurement and guided regularized random forest. Remote Sensing 8. https://doi.org/10.3390/rs8020144

Mutanga, O., Dube, T., Galal, O., 2017. Remote sensing of crop health for food security in Africa: Potentials and constraints. Remote Sensing Applications: Society and Environment 8, 231–239. https://doi.org/10.1016/j.rsase.2017.10.004

Naimi, B., Araújo, M.B., 2016. Sdm: A reproducible and extensible R platform for species distribution modelling. Ecography 39, 368–375. https://doi.org/10.1111/ecog.01881

Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., Toxopeus, A.G., 2014. Where is positional uncertainty a problem for species distribution modelling? Ecography 37, 191–203. https://doi.org/10.1111/j.1600-0587.2013.00205.x

Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized Linear Models. Journal of the Royal Statistical Society. Series A: General 135, 370–384. https://doi.org/10.2307/2344614

Nhamo, L., Matchaya, G., Mabhaudhi, T., Nhlengethwa, S., Nhemachena, C., Mpandeli, S., 2019. Cereal production trends under climate change: Impacts and adaptation strategies in Southern Africa. Agriculture (Switzerland) 9, 1–16. https://doi.org/10.3390/agriculture9020030

Niang, I., Ruppel, O.C., Abdrabo, M.A., Essel, A., Lennard, C., Padgham, J., Urquhart, P., 2014. Africa, in: Barros, V.R., C.B. Field, D.J. Dokken, M.D. Mastrandrea, K.J. Mach, T.E. Bilir, M. Chatterjee, K.L. Ebi, Y.O. Estrada, R.C. Genova, B. Girma, E.S. Kissel, A.N. Levy, S. MacCracken, P.R. Mastrandrea, and L.L.W. (eds. . (Ed.), Climate Change 2014: Impacts, adaptation, and vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Barros,. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 1199–1265. https://doi.org/10.1017/CBO9781107415386.002

Njenga, G.W., 2016. Multiple Endmember Spectral Mixture Analysis (MESMA) on multi-temporal VHR images for weed detection in smallholder farms . 116. University of Twente

Noi, P.T., Kappas, M., 2018. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. Sensors (Switzerland) 18. https://doi.org/10.3390/s18010018

Ochungo, P., Veldtman, R., Abdel-Rahman, E.M., Raina, S., Muli, E., Landmann, T., 2019. Multi-sensor mapping of honey bee habitats and fragmentation in agro-ecological landscapes in Eastern Kenya. Geocarto International 0, 1–22.

https://doi.org/10.1080/10106049.2019.1629645

OECD/FAO, 2016. Agriculture in Sub-Saharan Africa. Prospects and challenges for the next decade, OECD-FAO Agricultural Outlook 2016-2025. Rome. https://doi.org/10.1787/agr_outlook-2016-en

Oliphant, A.J., Thenkabail, P.S., Teluguntla, P., Xiong, J., Krishna, M., Congalton, R.G., Yadav, K., 2019. Mapping cropland extent of Southeast and Northeast Asia using multi-year time-series Landsat 30-m data using a random forest classifier on the Google Earth Engine Cloud. International Journal of Applied Earth Observation and Geoinformation 81, 110–124. https://doi.org/10.1016/j.jag.2018.11.014

Oswald, A., 2005. Striga control - Technologies and their dissemination. Crop Protection 24, 333–342. https://doi.org/10.1016/j.cropro.2004.09.003

Oswald, A., Ransom, J.K., 2001. Striga control and improved farm productivity using crop rotation. Crop Protection 20, 113–120. https://doi.org/10.1016/S0261-2194(00)00063-6

Oswald, A., Ransom, J.K., Kroschel, J., Sauerborn, J., 2001. Transplanting maize and sorghum reduces *Striga hermonthica* damage. Weed Science 49, 346–353. https://doi.org/10.1614/0043-1745(2001)049[0346:TMASRS]2.0.CO;2

Pal, M., Mather, P.M., 2005. Support vector machines for classification in remote sensing. International Journal of Remote Sensing 26, 1007–1011. https://doi.org/10.1080/01431160512331314083

Parker, C., 2009. Observations on the current status of *Orobanche* and *Striga* problems worldwide. Pest Management Science 65, 453–459. https://doi.org/10.1002/ps.1713

Peña, J.M., Torres-Sánchez, J., de Castro, A.I., Kelly, M., López-Granados, F., 2013. Weed mapping in early-season maize fields using object-based analysis of unmanned aerial vehicle (UAV) Images. PLoS ONE 8, 1–11. https://doi.org/10.1371/journal.pone.0077151

Phalke, A.R., Özdoğan, M., 2018. Large area cropland extent mapping with Landsat data and a generalized classifier. Remote Sensing of Environment 219, 180–195. https://doi.org/10.1016/j.rse.2018.09.025

Pickett, J.A., Woodcock, C.M., Midega, C.A.O., Khan, Z.R., 2014. Push-pull farming systems.

Current Opinion in Biotechnology 26, 125–132. https://doi.org/10.1016/j.copbio.2013.12.006

Planet, 2018. Planet — Planet Imagery Products. https://www.planet.com/products/planet-imagery/

Plant, R.E., 2012. Spatial data analysis in ecology and agriculture using R. CRC Press: Taylor and Francis Group, Califonia.

Powell, R.L., Roberts, D.A., Dennison, P.E., Hess, L.L., 2007. Subpixel mapping of urban land cover using multiple endmember spectral mixture analysis : Manaus , Brazil 106, 253–267. https://doi.org/10.1016/j.rse.2006.09.005

Prăvălie, R., Piticar, A., Roșca, B., Sfîcă, L., Bandoc, G., Tiscovschi, A., Patriche, C., 2019. Spatio-temporal changes of the climatic water balance in Romania as a response to precipitation and reference evapotranspiration trends during 1961–2013. Catena 172, 295–312. https://doi.org/10.1016/j.catena.2018.08.028

QGIS Development Team, 2019. QGIS Geographic Information System. Open Source Geospatial Foundation Project. https://doi.org/Http://qgis.osgeo.org,

Qi, J., Chehbouni, A., Huete, A.R., Kerr, Y.H., Sorooshian, S., 1994. A modified soil adjusted vegetation index. Remote Sensing of Environment 48, 119–126. https://doi.org/10.1016/0034-4257(94)90134-1

Qiao, C., Daneshfar, B., Davidson, A.M., 2017. The application of discriminant analysis for mapping cereals and pasture using object-based features. International Journal of Remote Sensing 38, 5546–5568. https://doi.org/10.1080/01431161.2017.1325530

Quintano, C., Fernández-Manso, A., Roberts, D.A., 2013. Multiple Endmember Spectral Mixture Analysis (MESMA) to map burn severity levels from Landsat images in Mediterranean countries. Remote Sensing of Environment 136, 76–88. https://doi.org/10.1016/j.rse.2013.04.017

R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria. URL https://www.R-project.org/.

R Core Team, 2020. R: A language and environment for statistical computing. R Foundation for

Statistical Computing,Vienna, Austria. URL https://www.R-project.org/.

Rakotoarisoa, M.A., Iafrate, M., Paschali, M., 2012. Why has Africa become a net food importer? Explaining Africa agricultural and food trade deficits, trade and market division, Food And Agriculture Organisation of the United Nations. Rome, FAO.

Ramoelo, A., Cho, M., Mathieu, R., Skidmore, A.K., 2015. Potential of Sentinel-2 spectral configuration to assess rangeland quality. Journal of Applied Remote Sensing 9, 094096. https://doi.org/10.1117/1.JRS.9.094096

Ransom, J., Kanampiu, F., Gressel, J., De Groote, H., Burnet, M., Odhiambo, G., 2012. Herbicide applied to imidazolinone resistant-maize seed as a *Striga* control option for small-scale African farmers. Weed Science 60, 283–289. https://doi.org/10.1614/WS-D-11-00060.1

Rich, P.J., Ejeta, G., 2007. Biology of host-parasite interactions in Striga species, in: Intergrating new technologies for *Striga* control: Towards ending the witch-hunt. World Scientific Publishing Company, Singapore, pp. 19–32.

Rispail, N., Dita, M.A., González-Verdejo, C., Pérez-De-Luque, A., Castillejo, M.A., Prats, E., Román, B., Jorrín, J., Rubiales, D., 2007. Plant resistance to parasitic plants: Molecular approaches to an old foe: Research review. New Phytologist 173, 703–712. https://doi.org/10.1111/j.1469-8137.2007.01980.x

Roberts, D.A., Gardner, M., Church, R., Ustin, S., Scheer, G., Green, R.O., 1998. Mapping chaparral in the Santa Monica Mountains using multiple endmember spectral mixture models. Remote Sensing of Environment 65, 267–279. https://doi.org/10.1016/S0034-4257(98)00037-6

Rodenburg, J., Demont, M., Zwart, S.J., Bastiaans, L., 2016. Parasitic weed incidence and related economic losses in rice in Africa. Agriculture, Ecosystems and Environment 235, 306–317. https://doi.org/10.1016/j.agee.2016.10.020

Rodenburg, J., Riches, C.R., Kayeke, J.M., 2010. Addressing current and future problems of parasitic weeds in rice. Crop Protection 29, 210–221. https://doi.org/10.1016/j.cropro.2009.10.015

Rosso, P.H., Ustin, S.L., Hastings, A., 2005. Mapping marshland vegetation of San Francisco Bay,

California, using hyperspectral data. International Journal of Remote Sensing 26, 5169–5191. https://doi.org/10.1080/01431160500218770

Rouse, J.W., Hass, R.H., Schell, J.A., Deering, D.W., 1974. Monitoring vegetation systems in the great plains with ERTS. Third Earth Resources Technology Satellite (ERTS) symposium 1, 309–317. https://doi.org/citeulike-article-id:12009708

Royimani, L., Mutanga, O., Odindi, J., Zolo, K.S., Sibanda, M., Dube, T., 2019. 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. https://doi.org/10.1016/j.rsase.2018.11.007

Samejima, H., Babiker, A.G., Takikawa, H., Sasaki, M., Sugimoto, Y., 2016. Practicality of the suicidal germination approach for controlling *Striga hermonthica*. Pest management science 72, 2035–2042. https://doi.org/10.1002/ps.4215

Sasson, A., 2012. Food security for Africa: An urgent global challenge. Agriculture and Food Security 1, 1–16. https://doi.org/10.1186/2048-7010-1-2

Schmidt, K.S., Skidmore, A.K., 2003. Spectral discrimination of vegetation types in a coastal wetland. Remote Sensing of Environment 85, 92–108. https://doi.org/10.1016/S0034-4257(02)00196-7

Scholes, J.D., Press, M.C., 2008. Striga infestation of cereal crops - an unsolved problem in resource limited agriculture. Current Opinion in Plant Biology 11, 180–186. https://doi.org/10.1016/j.pbi.2008.02.004

Shoko, C., Mutanga, O., 2017. Examining the strength of the newly-launched Sentinel 2 MSI sensor in detecting and discriminating subtle differences between C3 and C4 grass species. ISPRS Journal of Photogrammetry and Remote Sensing 129, 32–40. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2017.04.016016

Sibanda, M., Mutanga, O., Dube, T., S Vundla, T., L Mafongoya, P., 2019. Estimating LAI and mapping canopy storage capacity for hydrological applications in wattle infested ecosystems using Sentinel-2 MSI derived red edge bands. GIScience and Remote Sensing 56, 68–86. https://doi.org/10.1080/15481603.2018.1492213

Sibanda, M., Mutanga, O., Rouget, M., 2015a. 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. https://doi.org/http://dx.doi.org/10.1016/j.isprsjprs.2015.10.005

Sibanda, M., Mutanga, O., Rouget, M., Odindi, J., 2015b. Exploring the potential of *in situ* hyperspectral data and multivariate techniques in discriminating different fertilizer treatments in grasslands. Journal of Applied Remote Sensing 9, 096033. https://doi.org/10.1117/1.JRS.9.096033

Sims, D.A., Gamon, J.A., 2002. Relationships between leaf pigment content and spectral reflectance across a wide range of species 81, 337–354. https://doi.org/10.1016/S0034-4257(02)00010-X

Smith, A.M., Blackshaw, R.E., 2003. Weed: Crop discrimination using remote sensing: A detached leaf experiment. Weed Technology 17, 811–820. https://doi.org/10.2307/3989767

Somers, B., Asner, G.P., Tits, L., Coppin, P., 2011. Remote sensing of environment endmember variability in spectral mixture analysis : A review. Remote Sensing of Environment 115, 1603–1616. https://doi.org/10.1016/j.rse.2011.03.003

Song, C., 2005. Spectral mixture analysis for subpixel vegetation fractions in the urban environment: How to incorporate endmember variability? Remote Sensing of Environment 95, 248–263. https://doi.org/10.1016/j.rse.2005.01.002

Spallek, T., Mutuku, M., Shirasu, K., 2013. The genus *Striga*: A witch profile. Molecular Plant Pathology 14, 861–869. https://doi.org/10.1111/mpp.12058

Stockwell, D.R.B., Peterson, A.T., 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling 148, 1–13. https://doi.org/10.1016/S0304-3800(01)00388-X

Sultan, B., Defrance, D., Iizumi, T., 2019. Evidence of crop production losses in West Africa due to historical global warming in two crop models. Scientific Reports 9, 1–15. https://doi.org/10.1038/s41598-019-49167-0

Sungirai, M., Moyo, D.Z., De Clercq, P., Madder, M., Vanwambeke, S.O., De Clercq, E.M., 2018. Modelling the distribution of *Rhipicephalus microplus* and *R. decoloratus* in Zimbabwe.

Veterinary Parasitology: Regional Studies and Reports 14, 41–49. https://doi.org/10.1016/j.vprsr.2018.08.006

Tesfamichael, S.G., Newete, S.W., Adam, E., Dubula, B., 2018. Field spectroradiometer and simulated multispectral bands for discriminating invasive species from morphologically similar cohabitant plants. GIScience and Remote Sensing 55, 417–436. https://doi.org/10.1080/15481603.2017.1396658

Thamaga, K.H., Dube, T., 2019. Understanding seasonal dynamics of invasive water hyacinth (*Eichhornia crassipes*) in the Greater Letaba river system using Sentinel-2 satellite data. GIScience and Remote Sensing 56, 1355–1377. https://doi.org/10.1080/15481603.2019.1646988

Thamaga, K.H., Dube, T., 2018. Remote sensing of invasive water hyacinth (*Eichhornia crassipes*): A review on applications and challenges. Remote Sensing Applications: Society and Environment 10, 36–46. https://doi.org/10.1016/j.rsase.2018.02.005

Thenkabail, P., Mariotto, I., Gumma, M., 2013. Selection of hyperspectral narrowbands (HNBs) and composition of hyperspectral two-band vegetation indices (HVIs) for biophysical characterization and. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 6, 427–439. https://doi.org/10.1109/JSTARS.2013.2252601

Unachukwu, N.N., Menkir, A., Rabbi, I.Y., Oluoch, M., Muranaka, S., Elzein, A., Odhiambo, G., Farombi, E.O., Gedil, M., 2017. Genetic diversity and population structure of S*triga hermonthica* populations from Kenya and Nigeria. Weed Research 57, 293–302. https://doi.org/10.1111/wre.12260

Vapnik, V., 1979. Estimation of dependences based on empirical data. Nauka, Moscow, translation: Springer Verlag, New York, 1982) 27, 5165–5184.

Veloso, A., Mermoz, S., Bouvet, A., Le Toan, T., Planells, M., Dejoux, J.F., Ceschia, E., 2017. Understanding the temporal behavior of crops using Sentinel-1 and Sentinel-2-like data for agricultural applications. Remote Sensing of Environment 199, 415–426. https://doi.org/10.1016/j.rse.2017.07.015

Venables, W.N., Ripley, B.D., 2002. Modern applied statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

Vrieling, A., Meroni, M., Darvishzadeh, R., Skidmore, A.K., Wang, T., Zurita-Milla, R., Oosterbeek, K., O'Connor, B., Paganini, M., 2018. Vegetation phenology from Sentinel-2 and field cameras for a Dutch barrier island. Remote Sensing of Environment 215. https://doi.org/10.1016/j.rse.2018.03.014

Wan, J.Z., Wang, C.J., 2019. Determining key monitoring areas for the 10 most important weed species under a changing climate. Science of the Total Environment 683, 568–577. https://doi.org/10.1016/j.scitotenv.2019.05.175

Wan, Z., Hook, S., Hulley, G., 2015. MOD11C2 MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 0.05Deg CMG V006 [Data set]. NASA EOSDIS Land Processes DAAC. https://doi.org/doi: 10.5067/MODIS/MOD11C2.006

Worldbank, 2018. Rural population: Sub-Saharan Africa https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS?locations=ZG.

Xiong, J., Thenkabail, P.S., Gumma, M.K., Teluguntla, P., Poehnelt, J., Congalton, R.G., Yadav, K., Thau, D., 2017a. Automated cropland mapping of continental Africa using Google Earth Engine cloud computing. ISPRS Journal of Photogrammetry and Remote Sensing 126, 225–244. https://doi.org/10.1016/j.isprsjprs.2017.01.019

Xiong, J., Thenkabail, P.S., Tilton, J.C., Gumma, M.K., Teluguntla, P., Oliphant, A., Congalton, R.G., Yadav, K., Gorelick, N., 2017b. Nominal 30-m cropland extent map of continental Africa by integrating pixel-based and object-based algorithms using Sentinel-2 and Landsat-8 data on google earth engine. Remote Sensing 9, 1–27. https://doi.org/https://doi.org/10.3390/rs9101065

Xue, J., Su, B., 2017. Significant remote sensing vegetation indices: A review of developments and applications. Journal of Sensors 2017. https://doi.org/10.1155/2017/1353691

Yoneyama, Kaori, Yoneyama, Koichi, Takeuchi, Y., Sekimoto, H., 2007. Phosphorus deficiency in red clover promotes exudation of orobanchol, the signal for mycorrhizal symbionts and germination stimulant for root parasites. Planta 225, 1031–1038. https://doi.org/10.1007/s00425-006-0410-1

Yue, S., Wenjiang, H., Huichun, Y., Chao, R., Naichen, X., Yun, G., Yingying, D., Dailiang, P., 2018. Partial least square discriminant analysis based on normalized two-Stage vegetation

indices for mapping damage from rice diseases using PlanetScope datasets 1–16. https://doi.org/10.3390/s18061901