The potential of Sentinel-1 and Sentinel-2 remote sensing products for monitoring smallholder maize farms in support of the Sustainable Development Goals

by

Zinhle Olga Mashaba-Munghemezulu

Submitted in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

GEOINFORMATICS

in the

Faculty of Natural and Agricultural Science

University of Pretoria

2021

© University of Pretoria

Declaration of originality

This is to declare that the research work presented here is entirely my own work, unless or otherwise explicitly acknowledged by citation of published and unpublished sources. This research has not previously been submitted for assessment in any form to the University of Pretoria or to any other institution for any other purposes.

Signature: Date: 2021/11/29

© University of Pretoria

The potential of Sentinel-1 and Sentinel-2 remote sensing products for monitoring smallholder maize farms in support of the Sustainable Development Goals

Author Zinhle Olga Mashaba-Munghemezulu[1,2]

Supervisors Dr. George Chirima[1,2]

Affiliations [1]Department of Geography, Geoinformatics and Meteorology University of Pretoria, Pretoria, South Africa
[2]Geoinformatics Division, Agricultural Research Council, Pretoria, South Africa

Degree Doctor of Philosophy Geoinformatics

Abstract

Food security is an issue of global concern; this has mandated research on the development of systems for monitoring of agriculture using cost effective techniques such as remote sensing. Smallholder maize farms are dominant in Africa; they produce 80% of the maize in the region. The majority of the African population lives in rural areas and their livelihoods are dependent on smallholder agriculture particularly maize production. Thus, smallholder maize production plays a vital role in combating food insecurity in rural areas. Targeting food insecurity in developing countries is one of the important objectives of the Sustainable Development Goals (SDGs). However, local planning agencies and governments do not have adequate spatial information on smallholder farmers, and this affects the monitoring of the SDGs. Additionally, these farmers are faced with economic and environmental constraints that limit their productivity. Furthermore, the estimates of total planted area are unknown in most developing countries. Techniques for undertaking such estimates are either absent or very unreliable. This study explores the use of Sentinel-1 and Sentinel-2 data products for mapping and monitoring smallholder farms with machine learning. Findings suggest that the multi-temporal approach with the application of support vector machine and extreme gradient boosting is the recommended method for mapping smallholder maize farms in comparison to single date imagery based on lower standard deviation errors. The random forest model was suitable for estimating soil

© University of Pretoria

nitrogen. Furthermore, the findings suggest that maize yields can be accurately predicted from two months before harvest. The frameworks developed in this study can be used to generate spatial agricultural information in areas where agricultural survey data are limited. We recommend the use of Sentinel-1 and Sentinel-2 in conjunction with machine learning algorithms to map smallholder maize farms to support the SDGs.

Keywords: Sustainable development goals; smallholder; maize; machine learning; Sentinel-1; Sentinel-2

© University of Pretoria

Acknowledgements

I would like to thank everyone and all the organizations that were instrumental in making this research possible. The funding from the Agricultural Research Council Institute for Natural Resources and Engineering, National Research Foundation, University of Pretoria, Spatial Business Intelligence—SIQ and GeoTerraImages made it possible to collect field data and share the research through scientific publications.

I thank Dr George Chirima for supervising the project single-handedly. He was inspirational and supportive throughout the research. He provided critical comments timeously which improved my research skills greatly and he never limited my research ambitions. His guidance was not only in science, but also in life, which was valuable. My colleagues at the Agricultural Research Council—Bonolo Mosuwe, Jillie Masemola, Sabelo Mazibuko, Reneilwe Maake, Kgaogelo Mogano, Eric Economon and Wonga Masiza created a welcoming environment at work. The Geoinformatics division at the Agricultural Research Council and Center for Geoinformation Science at the University of Pretoria are acknowledged for hosting this research.

I thank my husband Dr Cilence Munghemezulu for his unconditional love, which was marked by our wedding during the course of my doctoral studies. We encountered many challenges both physically and emotionally but we supported each other through all of them. He understood when I was putting in long hours to finish my thesis, which became the main priority. My parents (Jude Mashaba and Dorothy Mashaba), the Mashaba family, Mathe family and in-laws are acknowledged for their support. The prayers from our family especially my grandmother Rose Mathe and their faith helped us through the tough times. I am blessed beyond measure to have my sisters Wandile Mashaba and Lungile Mashaba that love me so much.

© University of Pretoria

Publications

Peer-reviewed publications:

1. **Mashaba-Munghemezulu, Z.**, Chirima, G.J. and Munghemezulu, C., 2021. Mapping Smallholder Maize Farms Using Multi-Temporal Sentinel-1 Data in Support of the Sustainable Development Goals. ***Remote Sensing***, *13*(9), 1666. https://doi.org/10.3390/rs13091666

2. **Mashaba-Munghemezulu, Z.**, Chirima, G.J. and Munghemezulu, C., 2021. Delineating Smallholder Maize Farms from Sentinel-1 Coupled with Sentinel-2 Data Using Machine Learning. ***Sustainability***, *13*(9), 4728. https://doi.org/10.3390/su13094728

3. **Mashaba-Munghemezulu, Z.**, Chirima, G.J., Munghemezulu, C., 2021. Modeling the Spatial Distribution of Soil Nitrogen Content at Smallholder Maize Farms Using Machine Learning Regression and Sentinel-2 Data. ***Sustainability***, *13*(11591). https://doi.org/10.3390/su132111591

© University of Pretoria

Table of Contents

© University of Pretoria

© University of Pretoria

© University of Pretoria

List of Figures

© University of Pretoria

© University of Pretoria

List of Tables

© University of Pretoria

List of Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Networks
ASP	Aspect
CA	Catchment Area
CART	Classification and Regression Trees
CDL	Cropland Data Layer
CHIRPS	Climate Hazards Infrared Precipitation with Stations
DEM	Digital Elevation Model
EL	Elevation
EOS	Earth Observation System
ESA	European Space Agency
GB	Gradient Boosting
GEE	Google Earth Engine
GEOS	Group on Earth Observations
GIS	Geographic Information Systems
GPS	Global Positioning System
GRD	Ground Range Detected
LAI	Leaf Area Index
LST	Land Surface Temperature
MODIS	Moderate-Resolution Imaging Spectroradiometer
MS	Model Stacking
NASS	National Agricultural Statistics Service
OA	Overall Accuracy
PCA	Principal Component Analysis
RF	Random Forest
ROI	Regions of Interest
SAR	Synthetic Aperture Radar
SDGs	Sustainable Development Goals
SLP	Slope
SNAP	Sentinel Application Platform
SPOT	Satellite Pour l'Observation de la Terre
SRTM	Shuttle Radar Topography Mission
ST	Model Stacking
SVM	Support Vector Machine
USDA	United States Department of Agriculture
VH	Vertical Transmit and Horizontal Received
VV	Vertical Transmit and Vertical Received
Xgboost/ XG	Extreme Gradient Boosting

© University of Pretoria

Chapter 1
General Introduction

1.1. Introduction

The continuous reliance of developing countries on smallholder farms for food security requires effective monitoring and improved management practices. Smallholder farms play a crucial role in combating hunger in developing countries (Charman and Hodge, 2007; FAO, 2016). However, smallholder farms continue to be threatened by climate variability and climate change, a rising demand for food due to population growth, and changes in land use management (Jari and Fraser, 2009; Calatayud et al., 2014).

Smallholder maize farms are important in South Africa for the production of animal feed and are an important human staple food. The major producers of maize are the Free State, Northwest, and Mpumalanga provinces. White maize is for human consumption and yellow maize is for animal feeding (DAFF, 2016). Products such as fuel and starch (used for shoe polish, glue, fireworks, paint) are also derived from maize (Du Plessis, 2003).

Rural communities are often solely reliant on smallholder farms and a majority of the farmers lack formal education, which hinders them from accessing digital information such as climate forecasts and satellite data for crop monitoring purposes. Local agricultural governments and municipalities often fail to provide necessary support to the farmers, mainly due to the lack of geospatial information such as annual crop layers that can assist with planning and resource allocation. These factors contribute towards the low productivity of theses farms.

Earth Observation System (EOS) provides a cost effective opportunity to monitor and manage smallholder farms. Essential crop parameters (e.g., biophysical, crop production area, crop yields) can be estimated with reasonable accuracies using remote sensing technologies. This information can be used to manage essential crops (e.g., maize, wheat) effectively and to improve management practices (e.g., irrigation, monitoring of production, and mobilization of resources from governmental departments to the farmers in need) (Liu et al., 2020; Chakhar et al., 2020).

© University of Pretoria

1.2. Remote Sensing for Agriculture

A remote sensing system consists of instrumentation, processing and analysis designed to measure, monitor, and predict the physical, chemical, and biological aspects of the Earth system (Liang and Wang, 2019). The instrumentation component consists of sensors that are designed to measure different properties of the Earth system. The processing and analysis components consist of information extraction from the satellite data and scientific interpretation. Recently, advanced techniques are used to extract information from the data; these include Machine Learning algorithms and Artificial Intelligence (AI) (Mitchell, 1997; Haupt et al., 2008).

The sensors can be broadly categorized into optical and microwave sensors. These sensors only observe Earth systems in a specific range of wavelengths. For example, Sentinel-2 and Landsat-8 consist of optical satellites that operate in the visible spectrum and extend to near-Infrared and thermal wavelengths. Other satellites like Sentinel-1 and RADARSAT operate in a microwave region of the spectrum. Satellites such as Orbiting Carbon Observatory-2 (OCO-2) are designed to monitor carbon emissions. The amalgamation of different satellite missions is particularly important in the agricultural sector. For example, the sector contributes approximately 52% of global anthropogenic methane emissions and 84% of global nitrous oxide emission (Smith et al., 2008). The impact of agricultural activities on climate change, ecosystem services, and environmental sustainability can be monitored by using remote sensing systems at different spatial and temporal scales.

Table 1 lists the properties of Sentinel-1A/B and Sentinel-2A/B satellites. The Sentinel-2 satellite sensor has a temporal resolution of about 5 days at the Equator and a spatial resolution from 10 m to 60 m for selected bands. This sensor offers additional red-edge bands that are not available from other sensors such as Landsat satellites. These bands allow for crop chlorophyll and nitrogen estimations (Delegido et al., 2011; Clevers and Gitelson, 2013). The Sentinel-1 satellites offer an improved temporal resolution of about 6 days compared to ERS-1/2 and ENVISAT Advanced Synthetic Aperture Radar (ASAR) and a 10 m spatial resolution. Figure 1 depicts the differences between Sentinel-1 and Sentinel-2 observations. The Vertical Transmit

© University of Pretoria

and Vertical Received (VV) and Vertical Transmit and Horizontal Received (VH) polarizations look different to the RGB composite.

Table 1. Properties of freely available Sentinel-1 and Sentinel-2 platforms.

Spectral Band/Polarization	Central Wavelength (nm)	Bandwidth (nm)	Spatial Resolution (m)
Sentinel-1			
Vertical transmit and vertical receive (VV)	55,465,763	-	10
Vertical transmit and horizontal receive (VH)	55,465,763	-	10
Sentinel-2 MSI			
1-Coastal Aerosol	442	21	60
2–Blue	490	65	10
3–Green	560	35	10
4–Red	665	30	10
5–Vegetation Red Edge	705	15	20
6–Vegetation Red Edge	740	15	20
7–Vegetation Red Edge	783	20	20
8–Near-Infrared	842	115	10
8a–Vegetation Red Edge	865	20	20
9-Water Vapour	945	31	60
10-Short-wave infrared-Cirrus	1373	91	60
11–Short-wave Infrared	1610	90	20
12–Short-wave Infrared	2190	180	20

© University of Pretoria

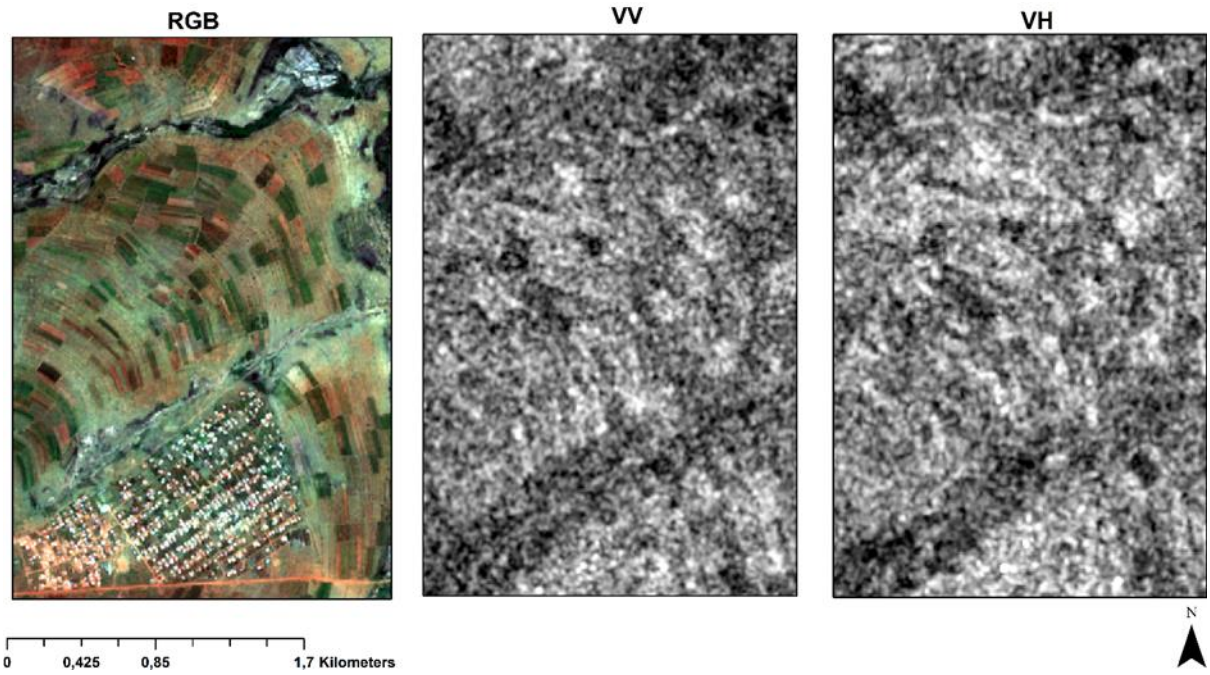

Figure 1. Visualization of Sentinel-1 and Sentinel-2 data for selected smallholder maize farms in Limpopo Province. The RGB composite is derived from the visible spectrum, while the VV and VH polarizations are derived from the microwave spectrum.

Various optical and radar data have been used in agriculture to extract useful information. Wang et al., (2020) used Landsat data archives to map maize and soybean with the aim of updating the Cropland Data Layer (CDL) in the United States. Ahmad et al., (2020) used Landsat-8 and Landsat-7 Enhanced Thematic Mapper Plus (ETM+) to model interannual variability of maize yields in Pakistan. Le Page et al., (2020) used Sentinel-1 soil moisture products to detect irrigation events in maize fields. Recently, data fusion of Sentinel-1 and Sentinel-2 has proven to yield better results in agricultural applications. For example, Van Tricht et al., (2018) used both Sentinel-1 and Senitnel-2 data to map different crop types in Belgium and obtained an overall accuracy 82% and the authors concluded that data fusion always yielded better results compared to a single sensor application. Mashaba-Munghemezulu et al., (2021) used both Sentinel-1 and Sentinel-2 to map smallholder maize farms and concluded that single date Sentinel-1 image data was not sufficient to map smallholder maize farms. However, the data fusion approach significantly improved the results by approximately 20% in accuracy. The noticeable improvements offered by both Sentinel-1 and Sentinel-2 are expected since both sensors capture different

© University of Pretoria

information about crops and thus complementing one another. For example, Sentinel-1 data provides structural, textural, and volumetric information about the crop, while Sentinel-2 data offers crop biomass information. This information can be used to discriminate crop types or study different growth stages of different crops (for example, Veloso et al., 2017).

1.3. Sustainable Development Goals and Earth Observation Systems

The 17 Sustainable Development Goals (SDGs) are a common blueprint between all the United National Member States, which were agreed on in 2015. These goals aim to promote peace and prosperity for people and the planet to ensure a sustainable future for all and the planet (Richard, 2015). Figure 2 depicts all the 17 SDGs, SDG number 2 (zero hunger) is of particular importance to this study. It has 8 targets and target number 2.3 outlines the need to double agricultural productivity and incomes of small-scale or smallholder food producers by 2030. However, the pandemic caused by the Coronavirus Disease (COVID-19) has resulted in a decline in economic activities around the world. This has caused enormous suffering to the most vulnerable including smallholder farmers in rural communities. This pandemic and other factors such as climate continue to threaten the 2030 agenda set by the United Nations Member States.

The SDGs cover problems on a regional to global scale that are very difficult to solve using conventional techniques of data capturing, monitoring, and modeling. Therefore, measuring environmental changes accurately such as quantifying planted areas at regional or global scale requires trans-disciplinary approaches. Additionally, computational resources are necessary, and the Earth Observation Systems (EOS) that can capture environmental variables with high temporal and spatial resolutions are needed. Data generated from EOS may include observation of the planet Earth on different electromagnetic frequencies (e.g., ultraviolet, visible, infrared and microwave). These frequencies allow for physical, chemical, climate, and biochemical parameter estimation of the planet Earth. The temporal and spatial resolutions of the data provided by EOS at different frequencies, makes it a suitable source of data to

© University of Pretoria

address some of the SDGs with very high accuracy and contribute towards monitoring and reporting on SDG targets.

Figure 2. The Sustainable Development Goals (Source: https://sdgs.un.org/goals).

Kavvada et al. (2020) outlined the importance of Earth Observation data in delivering on the SDGs, particularly, goals 6 (Clean Water and Sanitation), 14 (Life below Water) and 15 (Life on Land). These goals have been identified by the Group on Earth Observations (GEOS) 2020–2021 Work Plan on SDGs as areas that require attention in terms of development of methodologies and lack of data in some areas. Kavvada et al. (2020) identified additional areas where EOS can provide an indirect contribution to other SDGs such as sustainable economic growth by providing population distribution or urban structures. For example, Cochran et al. (2020) used a remote sensing-based ecosystem services platform (EnviroAtlas) to address SDG numbers 6, 11, and 15. This platform can be used to monitor water levels, land cover, and other socio-economic variables such as population density. These variables are used to report on certain SDGs indicators at different governmental levels. Hakimdavar et al., (2020) used a remote sensing approach to monitor water-related ecosystems in support of the SDG number 6.

EOS has a critical role to play in food security, especially in developing countries where ground-based infrastructure such as meteorological stations, internet connectivity is a big challenge. Kogan (2019) outlined the importance of remote

© University of Pretoria

sensing in food security. The study highlighted the use of remote sensing to monitor crop health status, monitor the impact of drought on crops, model crop yield and insurance index. These applications prove that EOS can be used to enhance food security.

1.4. Bibliometric Review of Remote Sensing and Maize Mapping

Remote sensing has been underutilized for applications concerning smallholder farms. Table 2 lists the number of retrieved articles, books, and book chapters from a bibliometric search using common key words from the two widely used databases in scientific research, i.e., Scopus and Web of Science. The results generally indicate that an average of 1807 articles were published involving the use of remote sensing for maize crops at different spatial scales. The research generally involves using remote sensing to monitor crops, classify crop types, and estimate crop biophysical parameters at different spatial scales using different remote sensing sensors (e.g., Moderate-Resolution Imaging Spectroradiometer (MODIS), Landsat, and Sentinel-1/2) (e.g., Karthikeyan et al. (2020), Mufungizi et al. (2020), Skakun et al. (2021), Ji et al. (2021)). This high number of research outputs was mainly due to the general search, using remote sensing and maize as keywords.

An average number of 39 papers were retrieved when the smallholder keyword was added to the search method. This generally shows that there is a need for more research focused on smallholder farms using remote sensing data to address SDG number 2. The bibliometric analysis also revealed that researchers from the United States of America and China produced most research involving remote sensing and maize, with a combined 819 authors contributing to this research area, whereas the African continent had only 34 authors contributing in total (Figure 3). This is concerning as smallholder maize farms contribute significant proportions to providing a sustainable staple food source for developing countries (Charman and Hodge, 2007; FAO, 2016).

Figure 4 illustrates the annual scientific contribution to literature concerning the use of remote sensing and its application to maize crop. The increase in the number

© University of Pretoria

of scientific outputs can be attributed to freely available global satellite data (such as Landsat and Sentinel series), availability of surface reflectance data, improvements in hardware and software to process satellite data, and continuous improvement in internet access and reduction in associated costs. These factors made it possible for researchers, especially in developing countries to be able to conduct research using these remote sensing technologies.

Table 2. A bibliometric search result of common result phrases and the associated number of documents retrieved. Time limit was set to 2020.

Search Criteria (limited to article, book chapter, and book)	**Scopus**	**Web of Science Core Collection**
TITLE-ABS-KEY (remote AND sensing AND maize OR corn)	1672	1942
TITLE-ABS-KEY (remote AND sensing AND sdgs)	49	66
TITLE-ABS-KEY (remote AND sensing AND sdgs AND maize OR corn)	1	1
TITLE-ABS-KEY (remote AND sensing AND maize OR corn AND smallholder)	35	43

© University of Pretoria

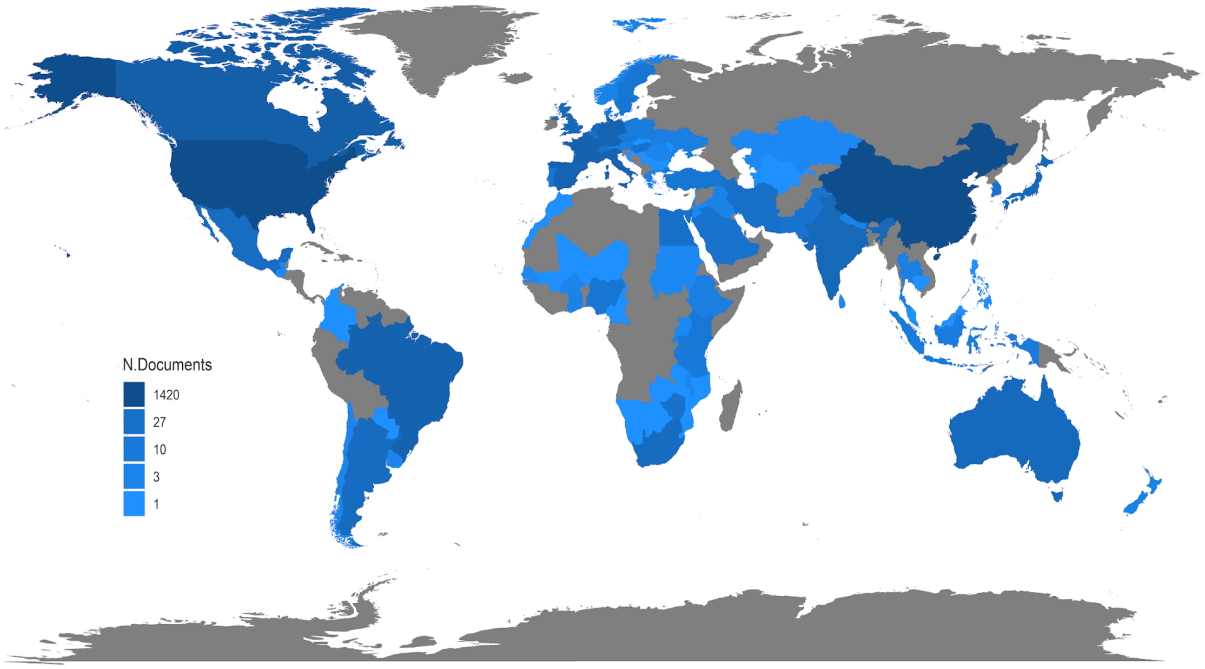

Figure 3. Map illustrating active countries that produced research related to our first search in Table 2 (grey areas have no data captured).

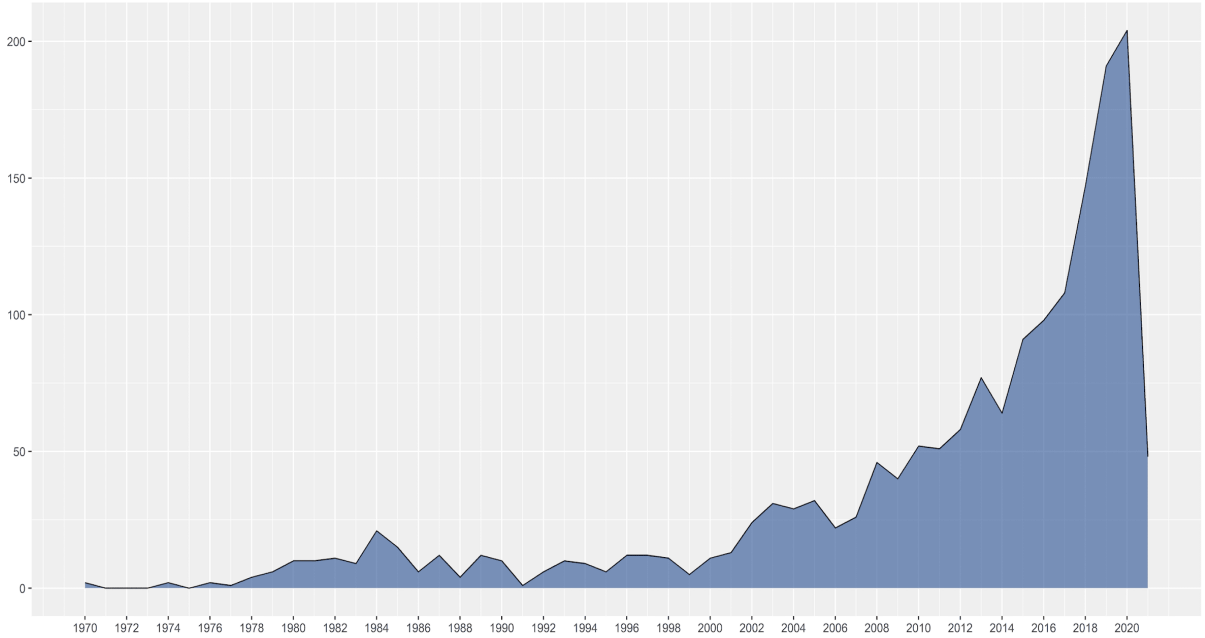

Figure 4. The exponential increase in the annual scientific production in papers/chapters published from 1970-2020 (x axis – year and y-axis – number of publications).

The Remote Sensing of Environment and Remote Sensing journals are the most important sources of information related to remote sensing research and maize (Table 3). The top 20 most relevant sources are international journals i.e., journals from

© University of Pretoria

developed countries such as Europe, United States and China. Among other factors, authors from developing countries often find it difficult to publish in international journals due lack of funding to support high publication fees (Salager-Meyer, 2008). For example, Remote Sensing of Environment and Remote Sensing journals publication fees are USD 3950 and USD 2180, respectively. These fees are equivalent to some of the research grants being given to emerging researchers in developing countries. One way to improve the contribution of researchers from developing countries in international journals is to promote cross-border collaboration.

EOS has matured enough to provide accurate information to smallholder farmers to enhance their food production under erratic climate variability and climate change, hence contributing towards SDG number 2. Figure 5 illustrates commonly used words i.e., remote sensing, maize (*Zea mays*) and crops were the most used words from 1672 publications. This is the testament to the importance of EOS in supporting SDGs.

© University of Pretoria

Table 3. The top two most relevant sources are Remote Sensing of Environment and Remote Sensing Journals.

Journal	Most Relevant Sources
Remote Sensing of Environment	150
Remote Sensing	115
Transactions of the Chinese Society of Agricultural Engineering	102
International Journal of Remote Sensing	80
IEEE Transactions on Geoscience and Remote Sensing	45
Agricultural and Forest Meteorology	32
Computers and Electronics in Agriculture	32
Transactions of the Chinese Society for Agricultural Machinery	31
Spectroscopy and Spectral Analysis	30
IEEE Journal of Selected Topics in Applied Earth Observation	29
International Journal of Applied Earth Observation	24
Precision Agriculture	24
ISPRS Journal of Photogrammetry and Remote Sensing	22
Journal of Applied Remote Sensing	22
Canadian Journal of Remote Sensing	20
Agronomy Journal	20
Field Crops Research	18
Transactions of the American Society of Agriculture	18
Sensors (Switzerland)	15

© University of Pretoria

Figure 5. Word cloud summarizing most used words within the titles of papers in the period January 1970 to December 2020.

1.5. Remote Sensing Data Analysis using Machine Learning Algorithms

Machine Learning is a subfield of computer science in which algorithms are developed to learn from and make predictions on data. The algorithms are dynamic as they can build non-linear models based on the training data to make data-driven decision (Bishop, 2006; Scheunders et al., 2018). Broadly, there are two general categories within machine learning—supervised and non-supervised approaches. The supervised category generally requires sample training data to develop the model and make appropriate predictions. The predictions can be discrete in nature, this is referred to as classification or the predictions can be continuous, and this is referred to as regression. The unsupervised category finds structures or patterns within the data using predefined procedures without any training data; examples of this include clustering or Principal Component Analysis (PCA). Both categories have huge potential in analyzing remote sensing data and extracting useful information or hidden patterns (Camps-Valls and Bruzzone, 2009).

© University of Pretoria

Remote sensing data and analysis strategies have evolved over the last few decades (Scheunders et al., 2018). For example, analysis of time-series satellite data may involve stacking multiple bands and analyzing them at once. This creates complex data structures that might be very difficult for traditional methods to extract any useful information. Machine learning algorithms are designed to deal with complex data structures regardless of the number of variables/features that may be included (Lary et al., 2015). Much research has been done using remote sensing and machine learning algorithms. Shao et al., (2021) used Random Forest (RF) to estimate maize k_c coefficients together with Unmanned Aerial Vehicle (UAV) and Leaf Area Index (LAI) data. The authors obtained a correlation coefficient of 0.65; such results could be used in precision irrigation management. Other authors have applied Machine learning algorithms for crop disease detection (Rangarajan et al., 2018), weed detection (dos Santos Ferreira et al., 2017) and yield prediction (Paudel et al., 2021). More information can be found in Benos et al., 2021 and references therein.

Python programming language (https://www.python.org/) has developed into a widely used tool that can implement machine learning algorithms. The Scikit-Learn open-source library (https://scikit-learn.org/stable/) is specifically designed to implement machine learning algorithms to solve a wide range of problems including remote sensing applications (Pedregosa et al., 2011). This library contains machine learning algorithms such as supervised (e.g., Support Vector Machine (SVM), Artificial Neural Networks (ANN), Gradient Boosting, and RF and unsupervised (e.g., clustering and PCA)). This library together with other freely available libraries were used in this study to implement machine learning algorithms to solve specific problems that are described in later chapters. Anaconda platform (https://www.anaconda.com/) was used for implementation and developmental package management. Due to Big Data problems in remote sensing (Huang et al., 2018), a high computationally intensive system was used for this project. A Ryzen 9 3900, 12 cores processor at 3.8 GHz and 128 GB Random Access Memory (RAM) computer was used.

© University of Pretoria

1.6. Problem Statement

Most local governments in developing countries still lack spatial agricultural information on smallholder farms. Spatial agricultural information includes crop type maps, soil nutrient information maps, and crop yield estimates on an annual basis. Such information is normally available for commercial farms but not available for smallholder farms. Local governments usually appoint extension officers to assist and collect smallholder farm information. However, these extension officers are poorly trained in basic Geographic Information Systems (GIS) and statistics to generate useful spatial information. There is a great need for accurate crop type, soil, and crop yield estimates information on an annual basis to support the SDGs and to monitor food production in rural communities. Such information can benefit local governments by informing their policy decision-making and implementation strategies. Remote sensing data such as Sentinel-1 and Sentinel-2 offer unprecedented opportunities for the use of local governments. This application will have an impact on the socio-economic status of rural communities in developing countries.

Other research studies have used Synthetic Aperture Radar (SAR) data to map maize fields. For example, Abubakar et al. (2020) used multi-temporal Sentinel-1 and Sentinel-2 to map smallholder farms in Nigeria using a stacking approach of different Sentinel data combinations. The authors applied Support Vector Machine (SVM) and Random Forest (RF) algorithms and achieved an overall accuracy of more than 90% for both algorithms. However, the authors did not provide the estimated production area for maize, which is the most important parameter for SDG number-2 reporting and food security monitoring. Jin et al. (2019) used multi-temporal Sentinel-1 and Sentinel-2 to also map maize production areas and estimate yield using the Google Earth Engine (GEE) platform in Tanzania and Kenya. Seasonal median composites, radar backscatter and optical surface reflectance were used to build an RF classifier and they obtained accuracies of more than 70%. Polly et al. (2019) used both Sentinel-1 and Sentinel-2 to map maize in Rwanda and noted that Sentinel-1 had a poor performance, which resulted in overestimating the maize production area compared to the Sentinel-2 data. All authors acknowledge that smallholder farms are difficult to map due to their small size and heterogeneous characteristics that can affect the

© University of Pretoria

spectral/backscatter signal. They also encourage the use of Sentinel-1 multi-temporal data since this approach can be used in all weather conditions and the resolution of 10 m is currently sufficient to contribute towards SDGs with relatively high accuracy.

Smallholder farms such as the one depicted in Figure 6, are generally poorly managed and receive very little government support. Due to poor farm management practices, weeds and pests affect the farms. Smallholder farms are often rain-fed, as the farmers do not have resources to implement irrigation systems, which make them vulnerable to erratic climate variability and climate change. Some of the farms are in remote areas and very difficult to access. The soil nutrient status of these farms are often not monitored for example the soil nitrogen content is often not managed with fertilization which can hinder maize growth. All these factors contribute towards low yields, and this has an impact on the livelihood of villagers.

Figure 6. Example of a smallholder maize farm, the red circle indicates areas where seeds did not emerge from the soil. Lack of equal row spacing and weeds can be seen in the photograph.

The use of remote sensing and machine learning to map and monitor crops can greatly enhance food security in developing countries. Applications of remote sensing in smallholder farm settings has been very limited, with only an average of 39 journal papers published between 1979 to 2020 period according to Scopus database (Table

© University of Pretoria

2) that involves remote sensing and maize and smallholder. Therefore, the use of remote sensing and machine learning in smallholder farms has not been fully explored yet.

1.7. Aim and Research Objectives

The aim of the study is to use Sentinel-1 and Sentinel-2 remote sensing data to map and monitor smallholder maize farms in support of the SDGs number-2 based on machine learning algorithms. The developed framework and associated models in this study can be used in other areas with similar settings to generate spatial agricultural information, especially in areas where such information is lacking such as developing countries and rural communities.

Research objectives are to:

1. Evaluate both Sentinel-1 and Sentinel-2 single date imagery to delineate smallholder maize farms using machine learning algorithms.
2. Develop an innovative approach using Sentinel-1 time-series data and machine learning algorithms (integrating both supervised and unsupervised methods) to map smallholder maize farms.
3. Investigate the utility of machine learning regression for spatial predictions of soil nitrogen content in smallholder maize farms.
4. Develop procedures using Sentinel-1 data to model maize yield in complex environments.

1.8. Significance of the Research

The commonly used optical techniques for crop monitoring are limited by cloud cover, sun illumination, soil properties at low plant cover, and have a low spatial resolution, which limits their applications during the summer rainfall period (Jiao et al., 2011; Inoue et al., 2014). In crop modelling activities, crop monitoring models developed using optical data often have gaps due to the periods of haze and clouds (Jiao et al., 2011). New generation techniques such as Synthetic Aperture Radar (SAR) based on the microwave wavelength can overcome these limitations and operate in all-weather

© University of Pretoria

condition or environment at an improved spatial resolution (Moran et al., 2002). However, research using this new technique is lacking in Africa as compared to optical techniques.

This research focuses on using both microwave, optical techniques and ground-based measurements for crop mapping, soil nutrient mapping, and yield estimation applications. Ground based measurements of the spectral characteristics of vegetation are used to derive narrowband indices, which are more accurate when there is high biomass and high leaf area as compared to broadband indices derived from air borne platforms, which get saturated under these conditions (Aparicio et al., 2002; Hansen and Schjoerring, 2003). In most studies, data from space borne platforms are used. These have a coarse spectral and spatial resolution, which results in weak relationships between vegetation parameters and spectral data as compared to ground-based measurements (Goel et al., 2003). However, ground based measurements are not widely used in crop monitoring and crop modelling activities. This study aims to overcome this gap. The contribution to science of this research is in assimilating different types of remotely sensed data (Sentinel-1 and Sentinel-2) in conjunction to machine learning models for agricultural applications, which is an approach not commonly used for smallholder maize farms. In addition, the study proposed novel approaches to generate two critical statistics, quantifying area planted and mapping the distribution of soil nitrogen, for the highly heterogeneous smallholder farming systems in Africa.

Maize contributes substantially to developing countries especially Africa and Latin America as a food source and for nutritional security (Shiferaw et al., 2011). Although maize production has increased due to improved crop varieties, increased fertilizer inputs, water and pesticides; climate change presents a challenge for maize in Southern Africa (Evenson and Gollin, 2003; Lobell et al., 2011). Techniques developed using remote sensing are beneficial for agricultural applications because they provide timely information on the phenological changes and the development of vegetation (Veloso et al., 2017). Additionally, new generation sensors such as SAR have cloud penetration abilities, operate in all weather conditions, can acquire images

© University of Pretoria

during the day or night and have a high spatial resolution (Moran et al., 2002). These characteristics are important for maize growth monitoring during high rainfall months.

The Agricultural Research Council (ARC) in South Africa, where this study was funded, will be generating these products (crop type, soil, yield maps) on an annual basis to assist local communities and local governments to monitor smallholder farms and improve on decision-making processes and inform relevant policies. This initiative will ensure a systematic and accurate approach towards providing necessary information to improve food security and SDGs reporting. The developed framework and associated machine learning models can be extended to other areas for local governments to benefit from spatial agricultural information that is generated from this study

1.9. General Study Area Description

The Sekhukhune district where this study is based (Figure 8) is located on the southeastern part of Limpopo province, South Africa. The main economic activities in Sekhukhune are agriculture, mining and tourism. This district shares borders with the Waterberg, Capricorn, Vhembe, and Mopani municipal districts. The economic drivers of the Waterberg district are mining, agriculture, game and cattle farming, secondary activities include manufacturing and service industries (WDM, 2015). The Capricorn district hosts Polokwane, which is the capital city of Limpopo wherein is the Central Business District (CBD), industrial areas, social services, residential areas, recreational land and smallholdings (CDM, 2015). This district is the largest contributor (24%) towards the economy of Limpopo in comparison to the other four districts (CDM, 2018). Within the Vhembe district, the main activities are tourism, livestock and crop farming (VDM, 2015). The Mopani district contributes to the economy of Limpopo through agriculture, mining, tourism and manufacturing (MDM, 2019).

The Climate Hazards Infrared Precipitation with Stations (CHIRPS) rainfall product indicate that the southern parts of the study area receives more rainfall compared to areas in the northern part. A maximum and minimum average rainfall of 11.22 mm and 3.65 mm were recorded between 2000 and 2020 period (Funk *et al*.,

© University of Pretoria

2015). Figure 7 depicts elevation that ranges between 748 m and 1759 m and some areas have slopes that can reach 51.35º (Farr *et al*., 2007). The study area receives summer rainfall (i.e., from October to April) and less rainfall is recorded during the winter periods (i.e., from May to September, Figure 8). Figure 9 depicts the annual average rainfall over time; the variations could be linked to drought events. For example, the 2015 drought that affected most agricultural sectors in South Africa was also captured by the CHIRPS data that indicates an annual average rainfall of 6.6 mm. According to Du Plessis, (2003) maize water requirement ranges between 450 mm to 600 mm per season.

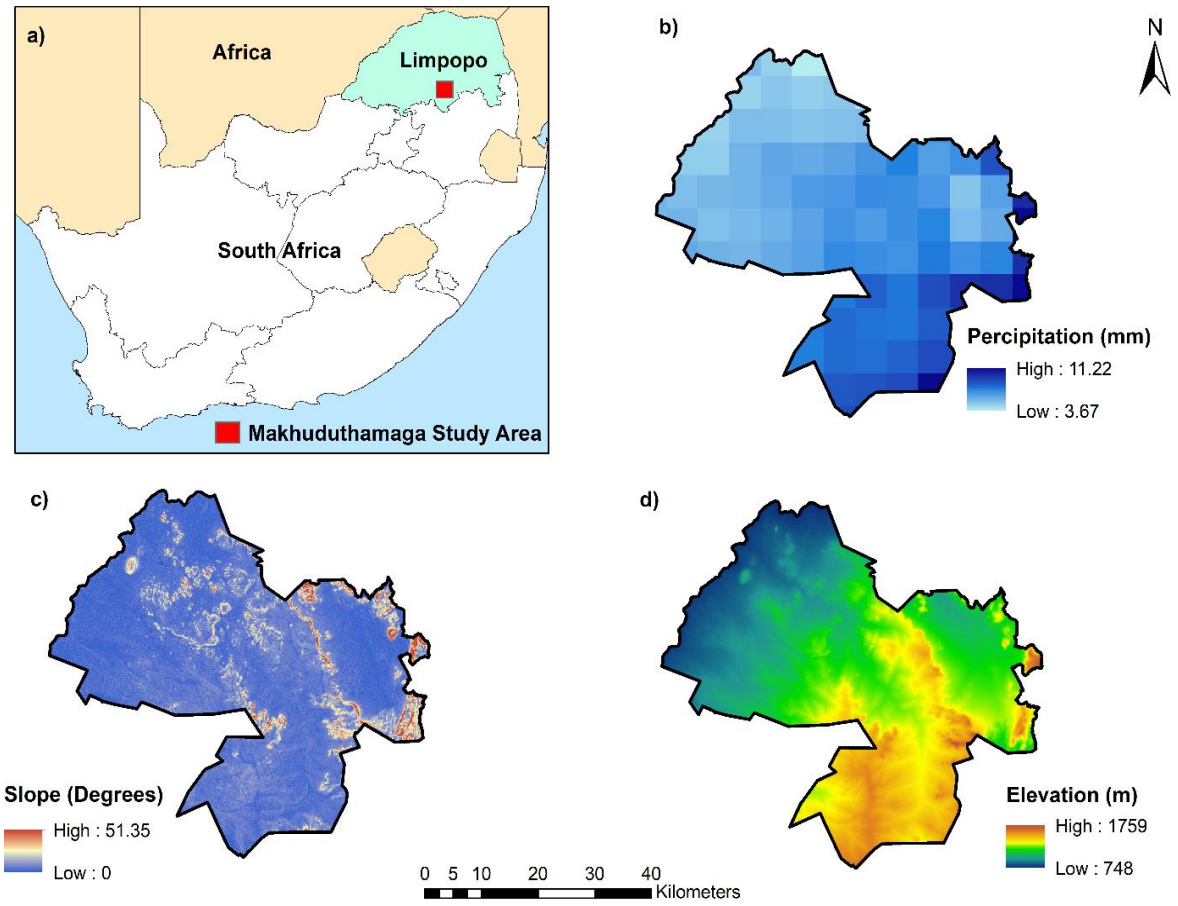

Figure 7. (a) Insert map of the study area, located in Limpopo province of South Africa. (b) Long-term average rainfall for the period 2000-2020 from CHIRPS data product. (c) and (d) Slope and elevation data, respectively (data sources: Funk *et al*., 2015 and Farr *et al*., 2007).

© University of Pretoria

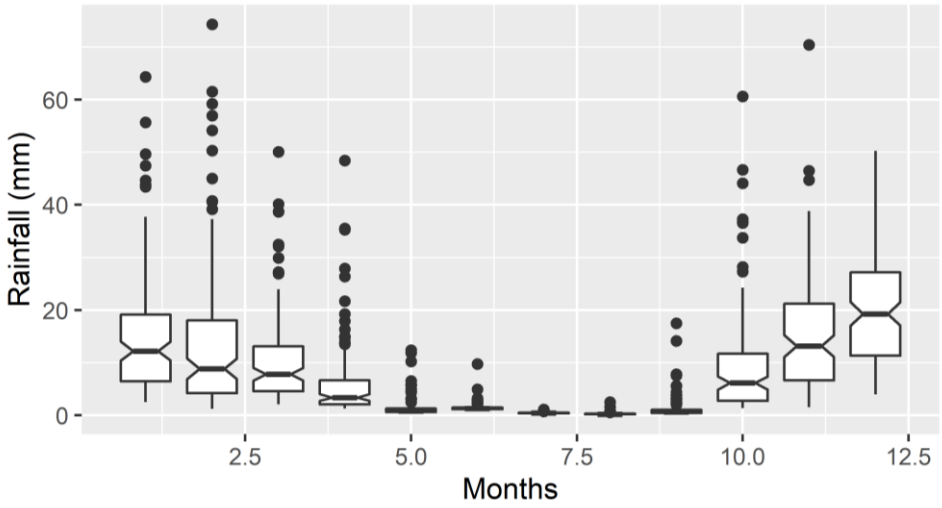

Figure 8. Monthly average precipitation records as observed by CHIRPS data product for the Sekhukhune district from year 2000 to 2020 (data source: Funk *et al*., 2015).

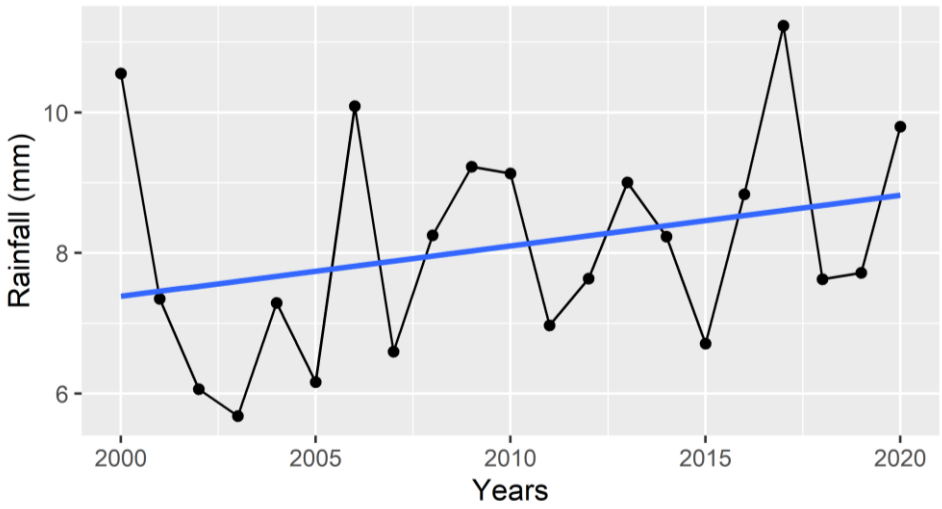

Figure 9. Annual average precipitation trend as measured by CHIRPS between 2000 and 2020 for the Sekhukhune district (data source: Funk *et al*., 2015). Linear trend suggest that generally, there is an increase in mean annual rainfall between 2000 and 2020 period.

1.10. Thesis structure

Each chapter is structured as standalone entity to deal with a specific objective as outlined in section 1.7. Different Sentinel data sets and machine learning algorithms that are suitable to the problem are used in each chapter. Due to this adopted thesis structure; there will be repetition of certain sections within the chapters. Details of each chapter are summarized as follows:

© University of Pretoria

Chapter 1: "*General Introduction*". This chapter provides a general introduction of the remote sensing in agriculture and the link of Earth Observation Systems to Sustainable Development Goals. A bibliometric review is provided to explore previous research in remote sensing for smallholder maize farms. The objective of bibliometric review was to highlight the contributions, limitations, gaps, and opportunities for satellite remote sensing techniques to the smallholder farming landscape. Furthermore, the review shows that many remote sensing datasets and data analysis techniques such as mapping of farm parcels, distribution of nitrogen, and crop types, and machine learning have not been sufficiently investigated in the smallholder farming landscape particularly in Africa and Asia.The significance of this research, problem statement, research aim, and objectives are also provided.

Chapter 2: "*Examining the utility of single date Sentinel-1 and Sentinel-2 data for Delineating Smallholder Maize Farms*". This chapter evaluates Sentinel-1 and Sentinel-2 single date data to delineate smallholder maize farms. The bibliometric analysis revealed a gap in research focusing on smallholder farmers applying remote sensing techniques. These farms have been challenging to map with conventional techniques due to their small sizes and fragmented nature. New generation Sentinel-1 and Sentinel-2 imagery are combined for mapping and estimating the planted areas for the farms. This strategy was explored to minimize the satellite data needed during the planting season which simplifies data archiving and data processing. Selected experiments are designed to explore which data combinations are suitable to delineate smallholder maize farms. Random Forest, Support Vector Machine algorithms and model stacking approach are used in each experiment. The results are evaluated using statistical metrics. Classification maps are produced using the optimal performing experiment for each model. The planted areas for maize are validated with field collected data. This chapter shows that the single-date Sentinel-1 data are insufficient to map smallholder maize farms. However, single-date Sentinel-1 integrated with Sentinel-2 data are sufficient in mapping smallholder farms and estimating their planted area.

Chapter 3: "*Mapping Smallholder Maize Farms Using Multi-Temporal Sentinel-1 Data*". A time series of Sentinel-1 imagery are used to map smallholder maize farms. The multi-temporal approach was investigated to compare with the single-date approach in Chapter 2 and determine which technique produces accurate results. The

© University of Pretoria

Principal Component Analysis data reduction method is applied to the Sentinel-1 multi-temporal data. Selected components are then used in the binary classification process to map smallholder maize farms and non-maize areas. Support Vector Machine and Extreme Gradient Boosting algorithms are used. The results are evaluated by using statistical matrices. The maize crops are mapped and the planted areas are determined. The results show that multi-temporal approach is better in mapping smallholder farms and it produced a much lower standard deviation estimate for the estimated areas in comparison to the single date approach.

Chapter 4: "*Modeling the spatial distribution of soil nitrogen content at smallholder maize farms using machine learning regression and Sentinel-2 data*". This chapter uses machine learning algorithms (Extreme Gradient Boosting, Gradient Boosting, Random Forest) in a regression format to map total soil nitrogen in the smallholder maize farms identified in Chapter 2 and Chapter 3. This application is particularly important for these farms as previous research has revealed that soil nitrogen deficiencies are limiting for maize growth (Xu et al., 2018). Four data types are integrated—Sentinel-2, environmental variables, soil indices, and vegetation indices in different experiments. The optimal data combinations for each algorithm are used model total soil nitrogen in smallholder maize farms. Rigorous model validation is done using commonly known statistical matrices. Spatial distribution maps are then created. Recommendations by crop consultants, extension services, and fertilizer dealers can benefit from using nitrogen content maps.

Chapter 5: *"Early Season Spatial Estimation of Smallholder Maize Yield based on Machine Learning".* This chapter applies machine learning algorithms—Extreme Gradient Boosting and Random Forest in a regression format to identify the optimal time for estimating maize yield in smallholder maize farms. Mainly Sentinel-1 data are used with minimal field collected data for model development. The ideal time is identified using model evaluation metrics. This procedure is necessary for forecasting maize yield to plan for maize shortages and surpluses. Important features for both machine learning models are determined. Model validation is then done for the developed model. Maize yield maps are then generated. Findings from this Chapter contribute directly to SDG 2 which aims on improving food security.

Chapter 6: *"Synthesis".* This chapter summarizes the main findings from each objective and provides recommendations for future work.

© University of Pretoria

Chapter 2

Examining the utility of single date Sentinel-1 and Sentinel-2 data for Delineating Smallholder Maize Farms

Based on: Mashaba-Munghemezulu, Z., Chirima, G.J. and Munghemezulu, C., 2021. Delineating Smallholder Maize Farms from Sentinel-1 Coupled with Sentinel-2 Data Using Machine Learning. *Sustainability*, 13(9), 4728.

Abstract

Rural communities rely on smallholder maize farms for subsistence agriculture, the main driver of local economic activity and food security. However, their planted area estimates are unknown in most developing countries. This study explores the use of Sentinel-1 and Sentinel-2 data to map smallholder maize farms. The random forest (RF), support vector (SVM) machine learning algorithms and model stacking (ST) were applied. Results show that the classification of combined Sentinel-1 and Sentinel-2 data improved the RF, SVM and ST algorithms by 24.2%, 8.7%, and 9.1%, respectively, compared to the classification of Sentinel-1 data individually. Similarities in the estimated areas (7001.35 ± 1.2 ha for RF, 7926.03 ± 0.7 ha for SVM and 7099.59 ± 0.8 ha for ST) show that machine learning can estimate smallholder maize areas with high accuracies. The study concludes that the single-date Sentinel-1 data were insufficient to map smallholder maize farms. However, single-date Sentinel-1 combined with Sentinel-2 data were sufficient in mapping smallholder farms. These results can be used to support the generation and validation of national crop statistics, thus contributing to food security.

Keywords: Sentinel-1; Sentinel-2; smallholder; maize; machine learning

© University of Pretoria

2.1. Introduction

Maize (*Zea mays L.*) is an essential cereal crop worldwide for food consumption, animal feed, and the production of industrial products such as biofuels (Ranum et al., 2014). Developed countries consume lower quantities of maize compared to developing countries (Asia, Latin America and Africa), which are reliant on maize (FAO, 2019). Smallholder farmers account for 80% of the maize produced as a staple crop in Africa (FAO, 2016). However, global climate forecasts have reported that Africa could be one of the most susceptible regions to the effects of climate change by 2050. This phenomenon will cause growing water shortages and scarcity of suitable land, which will affect the production of cereal crops including maize (Knox et al., 2012; Misra, 2014). Smallholder maize farms are important for the livelihoods of rural communities in Africa who depend on agriculture for food security and their local economic activities. These farmers are faced with problems such as inadequate rainfall due to droughts; they often have poor soils and limited irrigation infrastructure, which hinder their maximum productivity (Giller et al., 2006). Although these problems prevail in smallholder farms, there is an increasing demand for maize as a consequence of population growth (Santpoort, 2020). The disparity between declining maize supply and increasing demand for maize makes it necessary to develop a methodology to map smallholder maize farms and their sizes. Information about the areal extent of smallholder farms will guide governments when dispersing aid to them, inform land-use policies, and provide an indication of the current food security status, especially in vulnerable rural communities. The information provided by this project will enhance initiatives of local governments to provide spatial information regarding agricultural land-use by rural communities, as reliable information is lacking in most developing countries.

The use of remotely sensed data presents an opportunity for mapping the widely disparate smallholder farms and generating spatial information that can support policy implementation and enhance food security planning. Remote sensing technologies are able to collect data over a wide area in near-real time (Homolova et al., 2013). Additionally, the spatial distribution of crops on other areas within a study location that was not visited can be mapped. However, the use of remote sensing data for mapping

© University of Pretoria

smallholder farms has limitations. The coarse spatial resolution of remote sensing products such as Moderate-Resolution Imaging Spectroradiometer (MODIS) and Landsat is not sufficient to map smallholder farmland plots because of their small size of ±2 ha. Additionally, Landsat 8 has a revisit cycle of 16 days, which is insufficient to capture phenological changes for smallholder farms. Other remote sensing products such as Worldview, PlanetScope, RapidEye, and Satellite Pour l'Observation de la Terre (SPOT) have the required spatial resolution but are not freely available, and have a limited spatial coverage (Belward and Skøien, 2012). Hence, there is a need to explore the use of Sentinel-1 and Sentinel-2 data, which are freely available and have an improved spectral and spatial resolution.

The Sentinel-1 and Sentinel-2 sensors were launched for different applications amongst others, monitoring land-use/land-cover change and agricultural applications (Veloso et al., 2017). These sensors have a shorter revisit time of 10–12 days and a spatial resolution of 10-60 m (Drusch et al., 2012). Sentinel-2 is an optical sensor, which captures changes in land cover and provides a means to estimate crop area. However, the optical data from Sentinel-2 are susceptible to cloud cover or rainy weather, which limits the data availability during the cropping season (Asner, 2001). Radar imagery from Sentinel-1 overcomes the above shortfall; data are unobstructed by clouds or weather. These data have not been explored extensively for agricultural applications in comparison to optical data because of their complex data structure (Torbick et al., 2017).

The combined use of both Sentinel-1 and Sentinel-2 has the advantage of capturing both the spectral and textural information; this improves classification results, according to Cai et al. (2019). Dobson (1995) also observed that other Synthetic Aperture Radar (SAR) data such as ERS-1 and JERS-1 are also sensitive to the structural properties, soil moisture, and above-ground biomass of vegetation. Studies combining both Sentinel-1 and Sentinel-2, such as that of Van Tricht et al. (2018), have found overall accuracies (OA) between 75 and 82% when mapping maize and other land-cover classes with the application of Random Forest (RF) classification. Sonobe et al. (2017) used a kernel-based extreme learning machine to map maize and other crop types with Sentinel-1 and Sentinel-2 data. Their study

© University of Pretoria

achieved an overall classification of 96.8%. To our knowledge, limited studies have explored the potential offered by combining radar and optical data to address smallholder crop classification/mapping in a rural setting.

We examined the utility of Sentinel-1 to mapping smallholder areas under maize. We determined the outcome of integrating optical bands and vegetation indices derived from Sentinel-2 on the Sentinel-1 polarizations through a series of classification experiments for mapping maize areas. The RF algorithm, Support Vector Machine (SVM) algorithm and model stack (ST) were applied to each experiment. These machine learning algorithms are selected specifically because they have a superior discrimination capacity between different classes, suitable for noisy data and can be applied to limited samples (Belgiu and Drăguţ, 2016; Cooner et al., 2016). These distinguishing characteristics of the selected models have the potential of resolving issues with mapping fragmented inhomogeneous smallholder farms. Thereby, we achieved the overall aim of the study in developing a framework to enhance the delineation of smallholder maize farms using Sentinel-1, Sentinel-2 and vegetation indices.

2.2. Materials and Methods

2.2.1 Study Area

The field data were collected from the Makhuduthamaga district in Limpopo, South Africa (Figure 10). This area experiences rainfall during the warmer months of October to March and with a mean annual rainfall of 536 mm. The fields have an average elevation of 1333 m above mean sea level. The temperatures can drop to 7 °C in winter but can be as high as 35 °C in summer according to the records from the automatic weather stations of the Agricultural Research Council. This area was selected as a case study because most of the rural population are smallholder maize farmers; they farm primarily for subsistence and partially for selling in local markets (SDM, 2019). Specific regions of interest (ROI) were delineated for investigation based on the locations of the smallholder maize farms. The ROI was obtained from the local government department of agriculture (DAFF), where they were developed through survey campaigns. The ROI was used to generate an improved estimate of the area

© University of Pretoria

covered by smallholder farms by eliminating built-up areas, which can host households with backyard maize gardens leading to an overestimation of the planted areas. These households consume their maize before harvest-time.

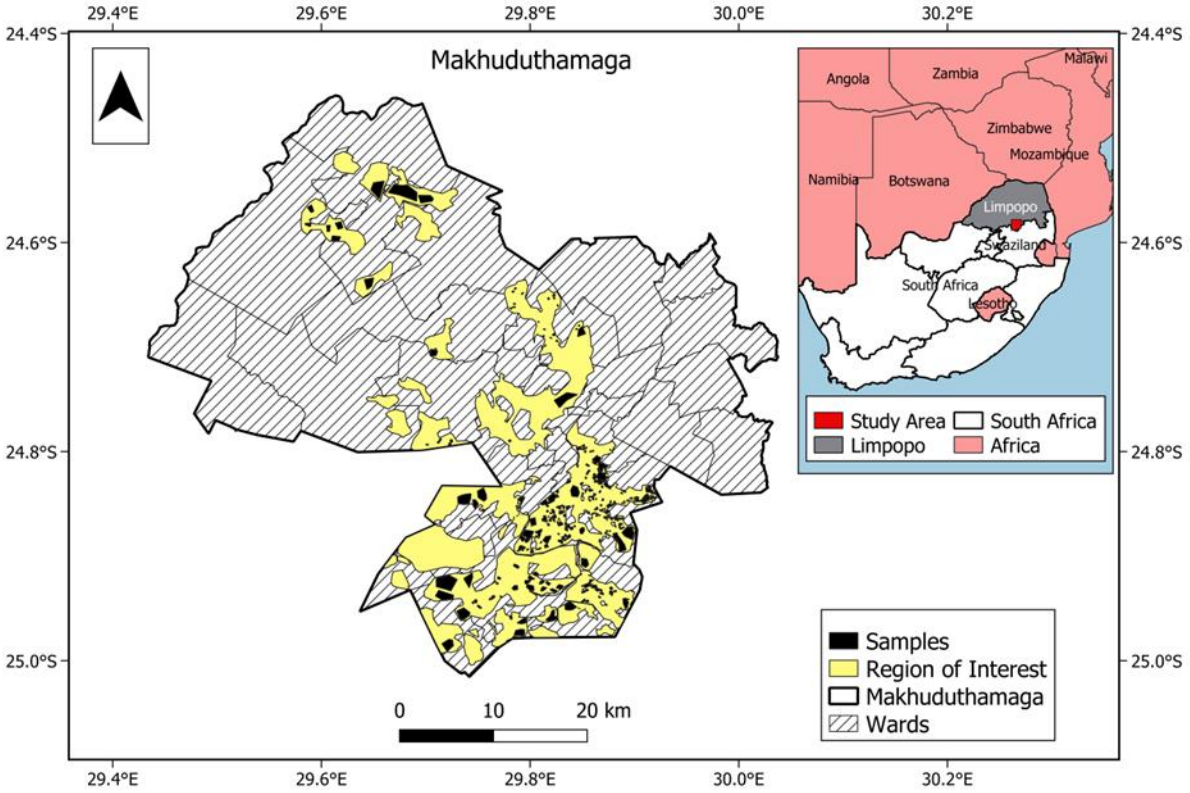

Figure 10. Location of Makhuduthamaga study area within Limpopo province, South Africa.

2.2.2 Field Data Collection

Field surveys for the collection of training and validation data for different landcover types within the ROI occurred from 18 to 21 February 2019. This period was selected because maize had the maximum green biomass at this time and could be discriminated more clearly in comparison to other land-cover types (Pervez and Brown, 2010). A handheld Garmin Global Positioning System (GPS) device was used to collect waypoints of different land-cover classes, applying a purposive sampling approach. The classes considered were maize (19.72%), bare land (50.01%), vegetation (30.23%) and water (0.0%), which are the dominant classes in the study area. The bare land, vegetation and water classes were amalgamated to form the non-maize areas and the maize areas were used as well. This approach of using only two classes of (1) maize and (2) non-maize areas reduces the classification errors from

© University of Pretoria

incorporating different land-cover classes individually. For example, there were fewer pixels for water in the study area in comparison to bare land and vegetated areas; using this as a separate class has the potential of introducing errors depending on the sensitivity of the classifier. Ground-based validation samples for 18 smallholder maize farms were collected using a GPS. The samples were not used as training data for classification.

2.2.3 Sentinel-1 Data Acquisition and Pre-Processing

Sentinel-1 Level-1 Ground Range Detected (GRD) data described in Table 4 were acquired from the Copernicus Open Access Hub. The Interferometric Wide (IW) image for 20 February 2019 was used; this consisted of the vertical transmit and vertical receive (VV) and vertical transmit and horizontal receive (VH) polarized backscatter values (in decibels) in a 10 m spatial resolution. Pre-processing of the radar images was done using the Sentinel Application Platform (SNAP). The orbit file was applied to update the orbit state vectors in the metadata file. Then, radiometric calibration was performed to convert the intensity values into sigma nought values. Speckle filtering was implemented to remove the granular noise caused by the interference of waves reflected from many scatterers. The Lee filter was applied at a 7 × 7 window size as it was found to be superior in preserving the edges, linear features, point target and texture (Lee et al., 1994). Range Doppler terrain correction was done to correct for geometric distortions caused by topography such as foreshortening and shadows; the Shuttle Radar Topography Mission (SRTM) 3-sec Digital Elevation Model (DEM) was used for this purpose (Loew and Mauser, 2007). The backscatter values were converted into decibels, and then the VH and VV polarizations were used to generate the VV/VH ratio.

© University of Pretoria

Table 4. Specifications of the Sentinel-1 and Sentinel-2 MSI data used in this study.

Spectral Band/Polarization	Central Wavelength (nm)	Bandwidth (nm)	Spatial Resolution (m)
Sentinel-1			
Vertical transmit and vertical receive (VV)	55,465,763	-	10
Vertical transmit and horizontal receive (VH)	55,465,763	-	10
Sentinel-2 MSI			
2–Blue	490	65	10
3–Green	560	35	10
4–Red	665	30	10
5–Vegetation Red Edge (RE1)	705	15	20
6–Vegetation Red Edge (RE2)	740	15	20
7–Vegetation Red Edge (RE3)	783	20	20
8–Near-Infrared (NIR)	842	115	10
8a–Vegetation Red Edge (RE4)	865	20	20
11–Short-wave Infrared (SWIR1)	1610	90	20
12–Short-wave Infrared (SWIR2)	2190	180	20

2.2.4 Sentinel-2 Data Acquisition and Pre-Processing

The Sentinel-2 Level-1C image for 26 February 2019 was acquired from the Copernicus Open Access Hub. The Sentinel-2 images were pre-processed using the Sen2Cor plugin in SNAP to convert them from the top of atmosphere reflectance units to the bottom of atmosphere reflectance (ESA, 2018). The bands, which were used are summarized in Table 4. The SWIR and vegetation red edge bands were rescaled to 10 m resolution. The indices depicted in Table 5 were derived. These indices are necessary to be investigated for mapping smallholder farms because they cover a broad part of the electromagnetic spectrum (NIR, red and green) in comparison to only using the Normalized Difference Index (NDVI). Additionally, they are sensitive to changes in soil background; they enhance the green vegetation signal, reduce the saturation effect of NDVI and are sensitive to chlorophyll content (Jordan, 1969;

© University of Pretoria

Tucker et al., 1979; Huete, 1988; Crippen, 1990; Rougean and Breon, 1995; Chen, 1996; Gitelson et al., 1996; Rondeaux et al., 1996; Broge and Leblanc, 2000; Haboudane et al., 2004).

© University of Pretoria

Table 5. Vegetation indices computed from Sentinel-2 imagery.

Vegetation Index	Equation	Justification	Reference
Difference Vegetation Index (DVI)	$DVI = NIR - Red$	Distinguishes between maize and soil.	Tucker et al., (1979)
Green Normalized Difference Vegetation Index (GNDVI)	$GNDVI = (NIR - Green) / (NIR + Green)$	More sensitive to chlorophyll concentration than NDVI.	Gitelson et al., (1996)
Infrared Percentage Vegetation Index (IPVI)	$IPVI = NIR / (NIR + Red)$	Similar to NDVI, but it is computationally faster.	Crippen (1990)
Modified Simple Ratio (MSR)	$MSR = \left(\frac{NIR}{Red} - 1\right) / \left(\left(\frac{NIR}{Red}\right)^{1/2} + 1\right)$	Minimizes the effects of variable soil reflectance.	Chen (1996)
Modified Triangular Vegetation Index (MTVI1)	$MTVI1 = 1.2 \times \left[1.2 \times \left(NIR - Green\right) - 2.5 \times \left(Red - Green\right)\right]$	Predicting maize green LAI (leaf area index).	Haboudane et al., (2004)
Modified Triangular Vegetation Index-Modified (MTVI2)	$MTVI2 = \frac{1.5 \times \left[1.2 \times \left(NIR - Green\right) - 2.5 \times \left(Red - Green\right)\right]}{\sqrt{\left(2 \times NIR + 1\right)^2 - \left(6 \times NIR - 5 \times \sqrt{Red}\right) - 0.5}}$	Better predictor of maize green LAI than MTVI1, and it accounts for soil background.	Haboudane et al., (2004)
Normalized Difference Vegetation Index (NDVI)	$NDVI = \frac{NIR - Red}{NIR + Red}$	Sensitive to maize greenness. However, it can saturate in dense vegetation when LAI becomes very high.	Tucker et al., (1979)

© University of Pretoria

Table 5 continued.

Vegetation Index	Equation	Justification	Reference
Optimized Soil Adjusted Vegetation Index (OSAVI)	$OSAVI = \frac{NIR - Red}{NIR + Red + 0.16}$	Eliminates the effect of the soil background.	Rondeaux et al., (1996)
Renormalized Difference Vegetation Index (RDVI)	$RDVI = \frac{NIR - Red}{\sqrt{NIR + Red}}$	Detects maize and is not sensitive to the effects of soil and sun viewing geometry.	Rougean and Breon, (1995)
Soil Adjusted Vegetation Index (SAVI)	$SAVI = \frac{\left((1+L)\times(NIR - Red)\right)}{NIR + Red + L}$	The SAVI index is similar to NDVI, but it reduces the influence of soil.	Huete (1988)
Simple Ratio (SR)	$SR = \frac{NIR}{Red}$	Detects healthy maize. However, it can saturate in densely vegetated maize plots when LAI becomes very high.	Jordan (1969)
Triangular Vegetation Index (TVI)	$TVI = 0.5\times\left[120\times(NIR - Green) - 200\times(Red - Green)\right]$	Detects green maize biomass and chlorophyll.	Broge and Leblanc, (2000)

© University of Pretoria

2.2.5 Classification Algorithms

Three different approaches were applied for mapping the smallholder farms, namely, RF, SVM and ST. The RF algorithm is a non-parametric decision tree ensemble classifier (Breiman, 2001). This classifier consists of a large number of classification and regression trees (CART), where each pixel is classified using a majority voting system. The RF algorithm trains each tree using an independently drawn subset of the original data using bootstrapping or bagging, and determines the number of features to be used at each node through an evaluation of a random vector (Breiman, 2001). One tuning parameter was defined for RF, the number of trees to grow (ntree), and the rest of the parameters are set to default values. In this study, the ntree was 150; this minimized the Out of Bag error, similar to Rodriguez-Galiano et al. (2012). The RF algorithm was selected because it can handle high dimensional data, is less sensitive to over-fitting and makes no distribution assumptions (Armitage and Ober, 2010; Immitizer et al., 2012; Belgiu and Drăguţ, 2016).

The SVM algorithm is also a non-parametric supervised learning classifier. The SVM uses the kernel function to transform training data into a high dimensional feature space, and to identify the optimal hyperplane that maximizes the distance between the separating hyperplane and the nearest sampling points (Cortes and Vapnik, 1995; Mountrakis et al., 2011; Mirik et al., 2013). The radial basis kernel was applied for SVM because of its good performance in previous studies (Knorn et al., 2009; Huang et al., 2002). The regularization parameter, gamma value and kernel coefficient had to be defined for the classifier. In this study, the regularization parameter was 100, the gamma value was 0.01 and kernel coefficient was 0, similar to Kumar et al. (2015). The SVM algorithm was selected because it does not make assumptions of the probability distribution and is not sensitive to training sample size (Mountrakis et al., 2011). A grid-search method was used to find these optimum turning parameters for both SVM and RF.

Model stacking was applied; it collates the predictions generated by different machine learning algorithms and uses them to generate a second-level learning classifier (Wolpert, 1992). In this study, the RF and SVM classifier were stacked, and

© University of Pretoria

the Logistic Classifier was used to combine the results. This ensemble model was applied because it has the ability to increase the predictive capacity of the two classifiers instead of using them independently (Wolpert, 1992).

Although RF has a variable importance measure, the permutation feature importance measurement was applied in this study to determine the importance of the predictors in each experiment, since previous studies have shown that RF variable importance has variations in ranking predictors as different iterations are performed (Millard and Richardson, 2015). The permutation feature importance allows different trained models (RF, SVM and ST) to assess feature importance. The algorithm computes reference scores s for the selected model on experimental datasets D. This reference score is the overall accuracy of the classifier. The features *j* in the datasets D are randomly shuffled to generate a corrupted version of the data $\tilde{D}_{k,j}$. The scores $s_{k,j}$ are computed on the corrupted datasets $\tilde{D}_{k,j}$. The feature importance i_j is then computed for feature f_j according to Equation 1

$$i_j = s - \frac{1}{K}\sum_{k=1}^{k} s_{k,j}. \quad (1)$$

2.2.6 Experimental Design

These samples were randomly separated into training (80% of the data) and testing (20% of the data) (Zhao, 2019). The training data were used for classification, whereas the testing data were used to evaluate the models. The vegetation indices in Table 5 were derived for use during classification. Then, classification experiments depicted in Table 6 were set for the classification algorithms based on different combinations (data configurations). These experimental set-ups were adopted to investigate the best approach for mapping smallholder maize with Sentinel-1 and Sentinel-2 data.

© University of Pretoria

Table 6. Combinations (data configurations) for the four experiments.

Experiment Number	Combinations	Description
1	VH, VV, VV/VH	Sentinel-1 polarization
2	VH, VV, VV/VH, DVI, GNDVI, IPVI, MSR, MTVI1, MTVI2, NDVI, OSAVI, RDVI, SAVI, SR, TVI	Sentinel-1 polarization and vegetation indices
3	VH, VV, VV/VH, DVI, GNDVI, IPVI, MSR, MTVI1, MTVI2, NDVI, OSAVI,RDVI, SAVI, SR, TVI, IPVI, 2, 3, 4, 5, 6 7, 8, 8a, 11, 12	Sentinel-1 polarization, vegetation indices and Sentinel-2 bands
4	VH, VV, VV/VH, 2, 3, 4, 5, 6 7, 8, 8a, 11, 12	Sentinel-1 polarization, and Sentinel-2 bands

2.2.7 Classification Model Evaluation and Planted Maize Area Estimation

Model evaluation was done to select the ideal model for estimating the maize areas. The matrices used were the OA, kappa coefficient of agreement ($\hat{k}$), cross-validation, precision, recall and F1-Score. The OA is the total classification accuracy and values close to 1 indicate that a classification is accurate; this is computed according to Equation 2. The OA was adjusted using the procedure of Olofsson et al. (2013) to account for classification errors. The $\hat{k}$ is calculated according to Equation 3 where *k* is the land-cover classes in the confusion matrix, x_{i+} and x_{+j} represent the marginal total for row *i* and column *j*. x_{ii} represents the number of observations in the row *i* and column *i* and *N* represents the total number of samples. $\hat{k}$ values > 0.8 represent a strong agreement between the classification map and the ground reference data. $\hat{k}$ values between 0.4 and 0.8 represent moderate agreement and $\hat{k}$ values < 0.4 represent poor agreement (Congalton and Green, 2008). The equations for both matrices are given as:

$$\text{Overall accuracy} = \frac{\sum_{i=1}^{k} x_{ii}}{N}, \tag{2}$$

© University of Pretoria

$$\hat{k} = \frac{N\sum_{i=1}^{k} x_{ii} - \sum_{i=1}^{k}\left(x_{i+} \times x_{+j}\right)}{N^2 - \sum_{i=1}^{k}\left(x_{i+} \times x_{+j}\right)}. \quad (3)$$

The K-fold cross-validation method was then applied (Efron, 1983). This method divides the training data randomly into K-folds or subsets (in this study a standard value of 10 was used), where one of the subsets is used as a test data set and the other K-1 is used as a training data set used to fit the model. This process is repeated *i* times, and the calculated average accuracy is computed for the testing data. The accuracy statistic was used during cross-validation, where values close to 1 indicate a high probability that a sample is correctly classified. The standard deviation of each accuracy value is also computed in each iteration, and the average standard deviation is indicated using a +/− attached to the cross-validation accuracy. The precision, recall and F1-Score were computed to determine the rate at which the pixels were correctly classified. The classifier performs well if the precision, recall and F1-Score are close to 1 (Kuhn et al., 2017).

Classification confidence was evaluated using McNemar's test to compare each of the models together (McNemar, 1947). We tested the hypothesis that the two models perform the same. When the Chi-squared values are less or equal to 3.84, the models have the same error at a 95% confidence level. However, one model is superior if the Chi-squared values are greater than 3.84.

The areas derived from the classification map were adjusted to account for classification error, and the 95% confidence interval was computed to compare the three models (Olofsson et al., 2013). These areas were compared to the areas derived from 18 maize farms measured during fieldwork to get an indication of how accurately the models estimate maize-planted areas using a regression equation. The p-value (p) and Pearson correlation coefficient (R) are used to evaluate the accuracy.

© University of Pretoria

2.3. Results

2.3.1 Classification Model Evaluation

The performances of the three algorithms applied in this study are presented in Table 7. The experiment with the lowest accuracies was experiment 1, containing the Sentinel-1 polarizations independently. This experiment had an accuracy of between 0.68–0.85 and a cross-validation score of between 0.65–0.69 for the three algorithms. Furthermore, the precision (0.65–0.69), recall (0.60–0.71) and F1-Score (0.64) for this experiment were considerably lower than all the other experiments. The kappa values also indicate moderate agreement between models and the reference data. However, there was a notable increase in accuracy by adding vegetation indices to the Sentinel-1 polarizations. The vegetation indices increased the accuracies by 24.2% for RF, 8.7% for SVM and 9.1% for ST. Although there was a reasonable improvement in model performance (precision of 0.925–0.929, recall of 0.926–0.930 and F1-Score of 0.925–0.930) in this experiment, adding Sentinel-2 bands improved the performance further by 5.9% for RF, 5.7% for SVM and 5.8% for ST in experiment 3. The best-performing experiment for all algorithms was experiment 4 with Sentinel-1 polarization and Sentinel-2 bands. This experiment had the highest accuracy (0.99) and was the most accurate (cross-validation: 0.91–0.92, precision: 0.99, recall: 0.99, and F1-Score: 0.99). McNemar's test (Table 8) confirmed that all three algorithms had a different performance in experiments 1–3. However, the performances of the algorithms were similar for the ST-RF combination but different for the ST-SVM combination in experiment 4.

© University of Pretoria

Table 7. The model performance statistics for the three classification (RF-Random Forest, SVM-Support Vector Machine, ST-Model Stack) algorithms in different experimental setups**.**

Experiment	Algorithm	Overall Accuracy	Cross-Validation	Precision	Recall	F1-Score	Kappa
1	RF	0.679	0.647 +/− 0.131	0.652	0.660	0.637	0.509
	SVM	0.845	0.688 +/− 0.127	0.693	0.706	0.640	0.526
	ST	0.841	0.689 +/− 0.128	0.674	0.703	0.637	0.523
2	RF	0.921	0.869 +/− 0.118	0.926	0.927	0.926	0.885
	SVM	0.932	0.873 +/− 0.112	0.925	0.926	0.925	0.884
	ST	0.932	0.870 +/− 0.109	0.929	0.930	0.930	0.890
3	RF	0.980	0.903 +/− 0.127	0.983	0.983	0.983	0.972
	SVM	0.989	0.883 +/− 0.106	0.991	0.991	0.991	0.986
	ST	0.990	0.899 +/− 0.137	0.991	0.991	0.991	0.986
4	RF	0.987	0.907 +/− 0.132	0.989	0.989	0.989	0.982
	SVM	0.991	0.914 +/− 0.082	0.992	0.992	0.992	0.988
	ST	0.991	0.921 +/− 0.112	0.991	0.991	0.991	0.986

Table 8. McNemar's test results for the ST–RF and ST–SVM combinations for experiments 1–4.

Combination	Chi-Squared	*p*-Value
ST1–RF1	4396.2	0
ST1–SVM1	430	1.7×10^{-95}
ST2–RF2	120	6.3×10^{-28}
ST2–SVM2	516.5	2.4×10^{-114}
ST3–RF3	6.3	0.0002
ST3–SVM3	34.5	4.2×10^{-9}
ST4–RF4	0.05	0.83
ST4–SVM4	9.3	0.0002

© University of Pretoria

2.3.2 Variable Importance

The variable importance was determined for the experiments in Table 6 using the permutation feature importance algorithm (Millard and Richardson, 2015). The experiments (Figure 11) varied in terms of the most important predictors depending on the input data. In experiment 1, the VH polarization had the highest importance; however, when integrating other predictors (e.g., experiments 3 and 4), the VV polarization had a higher importance over the other polarizations. The DVI outperformed all the other vegetation indices, followed by GNDVI in experiment 2. The most important bands in experiments 3 and 4 were the blue, red-edge and short-wave infrared (SWIR) bands. Additionally, the Sentinel-2 spectral bands took the highest priority in terms of importance in comparison to the Sentinel-1 polarizations.

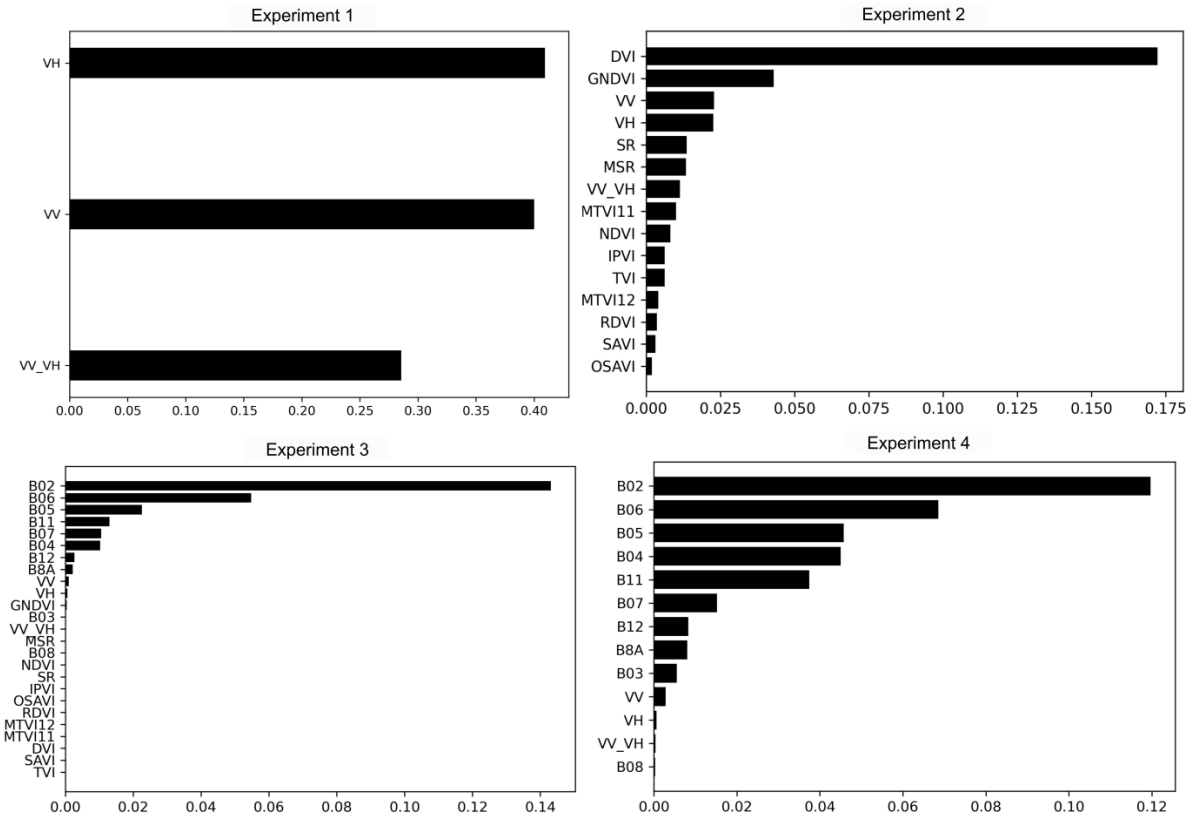

Figure 11. Variable importance plot for the four experiments.

© University of Pretoria

2.3.3 Mapping and Area Estimates for Maize

The 95% confidence interval was computed for the maize and non-maize areas within the study area. There was a relatively small variation between the total areas classified by the three algorithms for maize in Table 9. The RF algorithm had a discrepancy of 6% when compared to SVM, and 0.7% when compared to ST for the maize-planted areas. The ST algorithm had a variation of 5.5% in comparison to SVM. The areas classified as planted with maize had a lower error (0.7–1.2 ha) in comparison to the other areas which were not maize (1.2–1.88 ha) based on the 95% confidence interval. The RF algorithm had the lowest accuracy of ±1.2 ha when estimating maize areas, and SVM had the highest accuracy of ±0.7 ha.

Table 9. Estimated areas based on experiment 4 generated by the three classifiers for maize-planted areas and non-maize areas.

Algorithm	Land Cover	Total Area (ha)	95% Confidence Interval (ha)
RF	Maize	7001.35	1.236
	Non-Maize	33,496.05	1.884
SVM	Maize	7926.03	0.735
	Non-Maize	32,571.37	1.242
ST	Maize	7099.59	0.819
	Non-Maize	33,397.81	1.202

The classified areas for 18 smallholder maize farms were related to the field measured area at the same farms in Figure 12. There was a positive relationship, which was significant at a 95% confidence interval ($p < 0.05$) between the classified areas and field measured areas. The correlation coefficients obtained by the RF, ST and SVM algorithms are 0.51, 0.78 and 0.84, respectively, indicating higher agreement with the field measurements.

© University of Pretoria

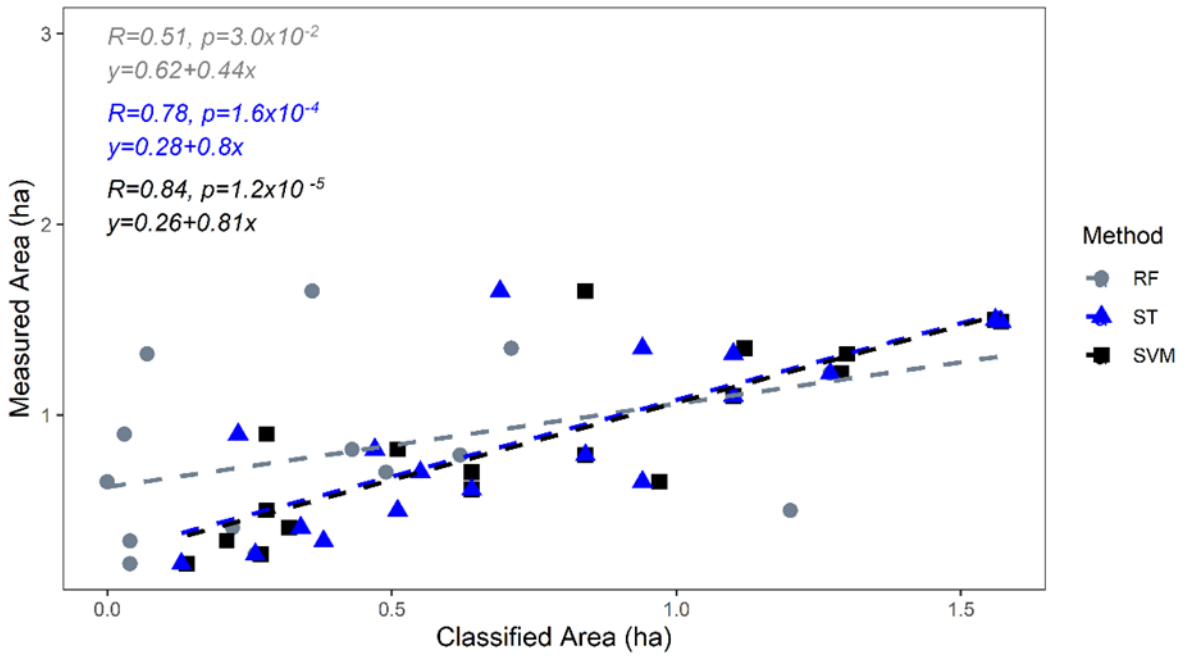

Figure 12. Linear regression models for the field measured areas (y) compared to the classified areas (x) for the best-performing experiment (experiment 4).

The three algorithms were used to generate the classification maps in Figure 13b–d depicting the spatial patterns of the two classes considered within the ROI. These maps compared well with the true color composite satellite image in Figure 4a for Sentinel-2. The classification maps generated by SVM, RF and ST were similar. The maize-planted areas were concentrated in the southern part of the Makhudutamaga district. The crop maps derived in this study are fundamental for crop forecasting and crop yield estimation at the end of the season. Changes induced by natural phenomena, such as climate variability and their effects on crop production, can be understood with the use of crop maps.

© University of Pretoria

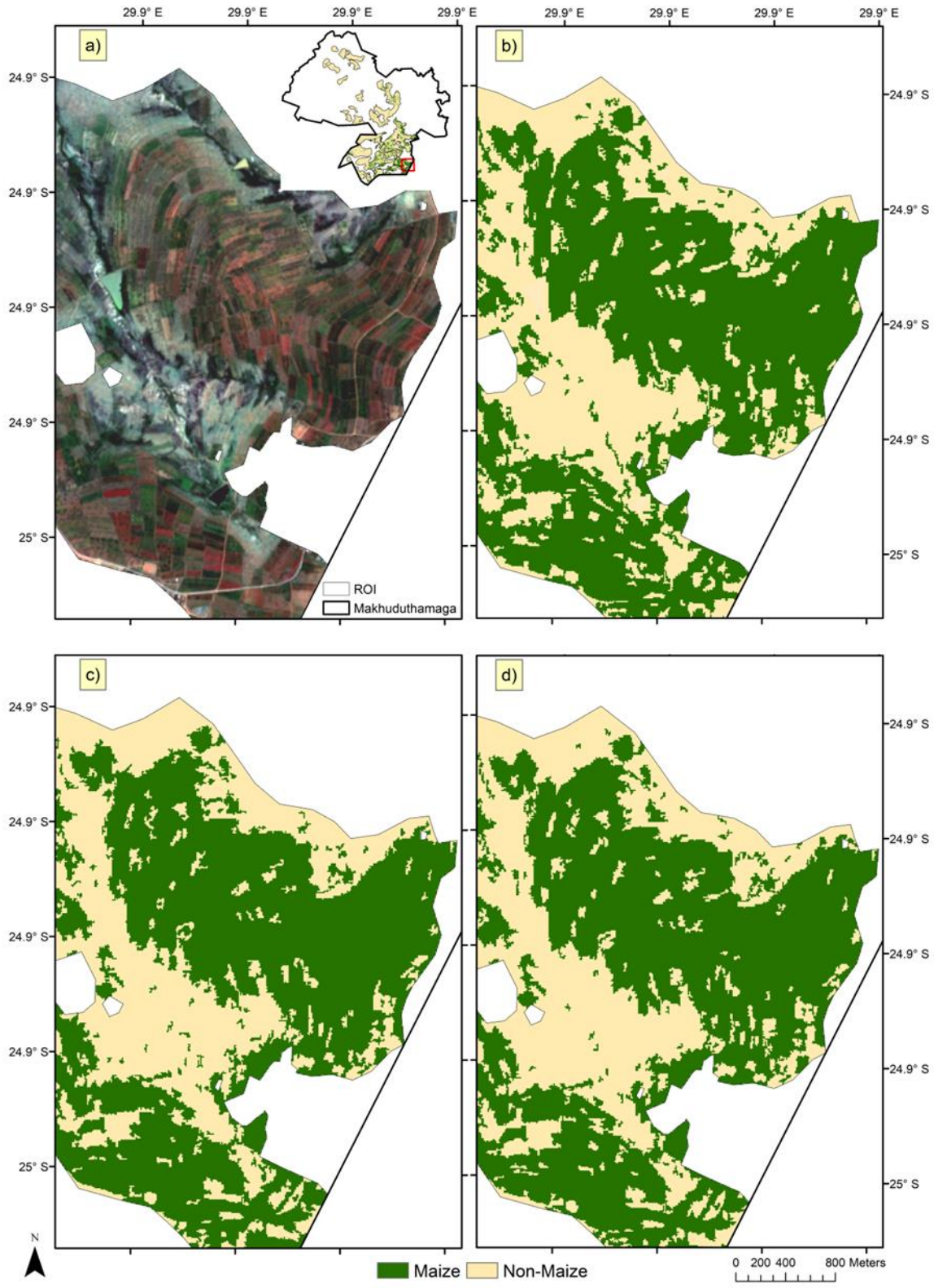

Figure 13. Classification maps for the optimal performing models in experiment 4, where (a) is the true color composite, (b) is RF, (c) is SVM and (d) is ST.

© University of Pretoria

2.4. Discussion

This study assessed the applicability of Sentinel-1, Sentinel-2 and derived vegetation indices for mapping smallholder maize in Makhudutamaga, Limpopo Province. Classification experiments were set to evaluate the performance of three machine learning algorithms. The variable importance measures were employed to investigate, which predictors had the most influence in each experiment. The best performing algorithms were then used for estimating and mapping the maize-planted areas. Findings suggest that integrating Sentinel-1 and Sentinel-2 is ideal for mapping smallholder maize farms with the application of machine learning algorithms.

Contrary to our expectations, the use of single-date Sentinel-1 radar data was not effective for mapping smallholder maize farms. The data combination consisting of Sentinel-1 polarizations exclusively had a low OA ranging from 67.9% to 84.5%, with RF being the worst performing classifier. These results are similar to those of Abubakar et al. (2020), who observed an OA of 78.9% when mapping smallholder maize using Sentinel-1 data by applying SVM. However, Useya and Chen (2019) reported an OA of 46% with RF and 40% with K-means classification when mapping smallholder maize farms and other crops with Sentinel-1 single-date data. The poor performance of the Sentinel-1 C-band data could be because of its shorter wavelength, which decreases canopy penetration in comparison to L-band SAR, which has a longer wavelength (Duguay et al., 2015; Khosravi et al., 2017). The inconsistencies in the planting pattern in the smallholder farms, such as a lack of equal row spacing, differences in the plant densities, leaf area index and crop heights in the study area, detract from the performance of the Sentinel-1 data because, according to Inoue et al. (2002), C-band data are sensitive to changes in biomass.

The integration of Sentinel-1, Sentinel-2 and vegetation indices were ideal for detecting smallholder maize farms, similar to previous studies in comparison to using Sentinel-1 data independently. Experiments 2, 3 and 4 show a clear increase in performance measures, in both OA and cross-validation scores. These values are more consistent and similar to each other, indicating the positive impact of radar-optical fusion on classification accuracy. Other studies such as that of Van Tricht et al.

© University of Pretoria

(2018) achieved OAs between 75 and 82% when mapping maize and other land-cover classes with the application of Sentinel-1 and Sentinel-2 data. Abubakar et al. (2020) achieved an OA of 97% when mapping smallholder maize with vegetation indices, Sentinel-1 and Sentinel-2 data. The high accuracies attained in this current study are attributed to the use of ideal locations of the electromagnetic spectrum such as the red-edge and SWIR. Furthermore, the vegetation indices applied in the current study reduce background effects (soils and other classes such as buildings), thereby enhancing the detection of crops and vegetation classes (Jordan, 1969; Tucker et al., 1979; Huete, 1988; Crippen, 1990; Rougean and Breon, 1995; Chen, 1996; Gitelson et al., 1996; Rondeaux et al., 1996; Broge and Leblanc, 2000; Haboudane et al., 2004).

The differences in performance of the SVM, RF and ST algorithms were expected. For example, Ouzemou et al. (2018) reported different OAs of 89.3%, 85.3% and 57.2% for RF, SVM and Spectral Angle Mapper (SAM) for crop type mapping with Landsat 8 data. Sonobe et al. (2014) found that SVM (OA of 89.1%) had a superior performance than RF (OA of 87.8%) and CART (OA of 81.2%) algorithms for classifying crops with TerraSAR-X data. These differences can be induced by various factors. In this study, the first experiment had the lowest accuracies; notably, RF had a low performance. This is because RF has been shown to be highly sensitive to small number of training input data in previous studies, in comparison to SVM and ST (Foody et al., 2006; Thanh Noi and Kappas, 2018). All three algorithms had high accuracies in the four experiments, possibly because the ROI used for training focused on maize-planted areas. This approach reduced the effects of using multiple land-cover classes individually which has a potential to lower the classification accuracy.

The variable importance results indicating the superiority of the VV polarization, DVI, GNDVI, blue band, red-edge and SWIR bands for mapping maize were expected. Forkuor et al. (2014) found that the VV band was superior to the VH band derived from TerraSAR-X for crop mapping applications. Deschamps et al. (2012) used Sentinel-1 data for crop classification and observed that the VV band was important for crop classification. However, other studies, for example Inglada, et al. (2016) and Arias et al. (2020), have reported that the VH band is more important than the VV bands for

© University of Pretoria

mapping crops because it captures the volume scattering from the crop canopy structure (McNairn et al., 2009). These results suggest that it is important to evaluate the polarizations based on the locality where they are applied. The finding that DVI and GNDVI are the most important indices, when using radar data and vegetation indices for crop classification, highlights the importance of evaluating different indices instead of relying on the commonly used NDVI index. The blue band, red-edge and SWIR bands have proven to be important in previous studies (Immitzer et al., 2016; Sonobe et al., 2018; Yi et al., 2020).). These bands capture the biochemical properties, water content and residue cover of different crop types that improves their detection (Zhang et al., 2020). In experiment 2, the OSAVI index was the least important variable. However, this seems to change in experiment 3, where this index ranked higher than RDVI, MTV12, MTV11, DVI, SAVI and TVI. This may be due to the correlation of these bands with the raw Sentinel-2 bands in experiment 3, while the indices in experiment 2 have a lower correlation between them.

The RF and ST algorithms had a relatively small difference of 0.7% when estimating the total planted maize area class, while the SVM algorithm seems to have overestimated the planted maize area by approximately 6% compared to the results from other algorithms. Even though SVM had a higher correlation coefficient than the RF and ST algorithms, we could not conclude that the SVM was the better estimator since the validation samples are relatively small. More validation data are required to provide more information on the performance of each algorithm in relation to ground-measured areas. However, since all algorithms have similar positive values of correlation coefficients, we can conclude that these algorithms can be used to estimate smallholder maize farmed areas. Unfortunately, official agricultural statistics such as production areas are not available in our study area, and could have been used to validate these observations.

The findings of this study are applicable to the Sustainable Development Goals (SDG), specifically, SDG number 2 (Zero Hunger), target 2.4 and indicator 2.4.1, which concern mitigating factors that affect agricultural production, ensuring sustainable agriculture and increasing the proportion of agricultural area under production (Richard, 2015). The agricultural production area is of great importance, as it informs

© University of Pretoria

local government and related stakeholders about agricultural activities and provides means by which production can be forecasted. The production area is one of the important indicators of food insecurity, especially in developing countries such as South Africa. Thus, this study contributes towards this SDG by using remote sensing data to accurately map production areas for smallholder maize farms. The spatial information generated can be used by local government to assist smallholder farms and policy implementation (Richard, 2015).

The limitations of this study were that a limited number of sample points were collected during fieldwork due to the undulating nature of the terrain, high cost to conduct the fieldwork and prominent mountainous areas, which were not accessible for data collection. This small sample size affects the statistical robustness of results (Foody, 2009). Secondly, the poor farm management practices of smallholder farmers such as weeds and patches of grass growing in some of the farms affect the spectral signature of maize and decrease the accuracy at which they can be detected with remotely sensed imagery. Thirdly, the use of red-edge indices, which have demonstrated some potential in improving the detection of vegetation in previous studies, should be explored (Forkuor et al., 2018; Kim and Yeom, 2014).

2.5. Conclusion

The overall aim of the study was to develop a framework to enhance the delineation of smallholder maize areas using single-date Sentinel-1, Sentinel-2 and derived vegetation indices. The results showed that single-date Sentinel-1 on its own was not sufficient in mapping planted maize fields. When Sentinel-2 data were integrated with Sentinel-1 data, an improvement of 24.2%, 8.7% and 9.1% for RF, SVM and ST algorithms, respectively, were observed. Machine learning proved to have a high capacity to estimate smallholder maize-planted areas (7001.35 ± 1.2 ha for RF, 7926.03 ± 0.7 ha for SVM and 7099.59 ± 0.8 ha for ST). The framework used in this study can be applied when evaluating different algorithms for mapping smallholder farms. The crop maps derived in this study are fundamental for crop monitoring, land-use policies and aiding food security planning activities.

© University of Pretoria

Chapter 3

Mapping Smallholder Maize Farms Using Multi-Temporal Sentinel-1 Data

Based on: Mashaba-Munghemezulu, Z., Chirima, G.J. and Munghemezulu, C., 2021. Mapping Smallholder Maize Farms Using Multi-Temporal Sentinel-1 Data in Support of the Sustainable Development Goals. *Remote Sensing,* 13(9), 1666.

Abstract

Reducing food insecurity in developing countries is one of the crucial targets of the Sustainable Development Goals (SDGs). Smallholder farmers play a crucial role in combating food insecurity. However, local planning agencies and governments do not have adequate spatial information on smallholder farmers, and this affects the monitoring of the SDGs. This study utilized Sentinel-1 multi-temporal data to develop a framework for mapping smallholder maize farms and to estimate maize production area as a parameter for supporting the SDGs. We used Principal Component Analysis (PCA) to pixel fuse the multi-temporal data to only three components for each polarization (vertical transmit and vertical receive (VV), vertical transmit and horizontal receive (VH), and VV/VH), which explained more than 70% of the information. The Support Vector Machine (SVM) and Extreme Gradient Boosting (Xgboost) algorithms were used at model-level feature fusion to classify the data. The results show that the adopted strategy of two-stage image fusion was sufficient to map the distribution and estimate production areas for smallholder farms. An overall accuracy of more than 90% for both SVM and Xgboost algorithms was achieved. There was a 3% difference in production area estimation observed between the two algorithms. This framework can be used to generate spatial agricultural information in areas where agricultural survey data are limited and for areas that are affected by cloud coverage. We recommend the use of Sentinel-1 multi-temporal data in conjunction with machine learning algorithms to map smallholder maize farms to support the SDGs.

Keywords: Sustainable Development Goals; smallholder; maize; Sentinel-1; principal component analysis; SVM; Xgboost

© University of Pretoria

3.1 Introduction

The United Nations agreed on 17 Sustainable Development Goals (SDGs) in 2015 with the aim of ensuring peace and prosperity for the people and the planet (Richard, 2015). The SDG number 2—end hunger, achieve food security and improve nutrition, and promote sustainable agriculture—aims to address this global crisis. Smallholder farming is one of the vehicles that can be used to achieve this goal (Abraham and Pingali, 2020). Smallholder farms are in most cases the only main source of reasonable income and food security for rural livelihoods in most developing countries. To achieve this goal, spatial agricultural information such as the spatial distribution of smallholder farms and production area estimates are pre-requisites. The production area estimates provide a quantitative measure in which food security can be forecasted in rural communities. Local governments can alleviate starvation and provide targeted relief efforts by using this information. Food security in developing countries remains a big challenge that the world is currently facing (Charman and Hodge, 2007; FAO, 2018). In Africa, smallholder farmers produce 80% of the maize in the regions, which forms part of the staple diet (FAO, 2016). The smallholder maize farmers of Africa are faced with environmental problems such as insufficient rainfall because of droughts, insect pest infestations, and infertile soils due to a multitude of reasons (e.g., monoculture, desertification, salinization, and degradation) (Jari and Fraser, 2009; Aliber and Hall, 2012; Calatayud et al., 2014). Additionally, economic issues such as the use of outdated technologies, limited market opportunities, and limited access to capital are prevalent in smallholder farms (FAO, 2016; Giller et al, 2006). These issues coupled with an increase in demand for maize products have contributed to food insecurity, particularly in rural communities that are reliant on maize (Santpoort, 2020).

Remote sensing data offers opportunities to monitor and map smallholder farms because they are able to capture their heterogeneous and complex characteristics (Kogan, 2018). Optical remote sensing has been used to map agricultural fields (Liu et al., 2020, Chakhar et al., 2020). However, clouds and cloud shadows remain a big challenge in extracting phenological parameters of crops during the growth stages and mapping crop fields using a multi-temporal approach due to data gaps (Baret et al., 2013). Radar data have emerged as one of the best remote sensing tools that can be

© University of Pretoria

used to map agricultural crops without being affected by clouds (Campbell and Wynne, 2011). Previously, this data type was limited to specific regions and campaigns (Woodhouse, 2017). The Sentinel-1A/B Synthetic Aperture Radar (SAR) C-band satellites were launched by the European Space Agency (ESA) with a wider coverage (Attema, 2007). Applications of SAR data in agricultural crop mapping have increased over the years; this was mainly driven by free access to the data and improved spatial (10 m) and temporal resolutions with a global coverage. The smallholder farms are generally less than 2 ha in size, which makes it difficult to map them with coarse resolution sensors (Jain et al., 2013). Therefore, the characteristics of Sentinel-1 sensors make it a suitable tool for agricultural applications (McNairn and Brisco, 2004).

Different authors have used a Sentinel-1 multi-temporal approach to map agricultural crops. Useya and Chen (2019) used Sentinel-1 data to map smallholder maize and wheat farms in Zimbabwe. The authors used model-level data fusion (i.e., data were stacked and used as input into the models) and achieved overall accuracies of 99% and 95% for different study area sites. Kenduiywo et al. (2018) applied a Dynamic Conditional Random Fields classification procedure on multi-temporal Sentinel-1 images to map different kinds of crops (maize, potato, sugar beet, wheat, and other classes). The authors were able to map maize with a producer and user accuracies of 93.74% and 90.04%, respectively. Whelen and Siqueira (2018) used comprehensive Sentinel-1 multi-temporal data to identify agricultural land cover types. They concluded that vertical transmit and vertical receive (VV) and vertical transmit and horizontal receive (VH) polarizations individually and combined were able to provide an accuracy of above 90% over North Dakota. All authors mention the problem of “Big Data” when dealing with Sentinel-1 multi-temporal images due to the increase in dimensionality. Therefore, processing multi-temporal satellite data requires more computational resources. McNairn and Brisco (2004) provide a detailed review on the applications of C-band polarimetric SAR for agricultural applications.

The use of the Principal Component Analysis (PCA) technique on Sentinel-1 to enhance the detection of smallholder maize farms has not yet been fully established. The PCA is a simple but powerful multivariate technique that transforms inter-correlated variables into a set of new linearly orthogonal (non-correlated) variables

© University of Pretoria

called principal components, and these components have maximum variance (Abdi and Williams, 2010). The condition of maximum variance is an added advantage to the classification algorithms as this can allow determination of decision boundaries with ease, therefore enhancing the detection of different classes. Meanwhile, a stacking approach such as the one used by Abubakar et al. (2020), Jin et al. (2019), and Useya and Chen (2019) may result in class overlap due to inter-correlated bands that may exist within the stacked datasets. This can lead to potential misclassification of different classes. Readers should consult Canty (2014) for more details on PCA formulation.

In this study, we used multi-temporal images of Sentinel-1 to develop a framework to map smallholder maize farms using well-known machine learning algorithms (Support Vector Machine—SVM and Extreme Gradient Boosting—Xgboost) under a complex environment. The strengths of these algorithms are that: (1) the SVM algorithm can handle high dimensional data using a few training samples (Chakhar et al., 2020). (2) The Xgboost algorithm runs at an improved computational speed, which is advantageous when processing multi-temporal images for the maize planting season (Chen and Guestrin, 2016). (3) Additionally, both algorithms have a good feature identification capacity and are non-parametric (Chakhar et al., 2020; Whelen and Siqueira, 2018; Piiroinen et al., 2015). The two-stage image fusion approach was applied. Firstly, pixel-level fusion was done; the purpose of this first stage is to reduce computational demands on the system by reducing the dimensions of the datasets using PCA. Secondly, model-level fusion was done; this second stage uses sufficient principal components for all the reduced polarizations as input into the classifying algorithms.

Generally, this approach has been used mainly in hyperspectral remote sensing image classification or change detection analysis (Licciardi et al., 2011; Chatziantoniou et al., 2017). It has not yet been applied to Sentinel-1 to map smallholder maize farms and estimation of their production areas. The approach was tested on a rural community in Makhuduthamaga, Limpopo province of South Africa. This region is dominated by smallholder maize farms and most farmers farm for subsistence.

© University of Pretoria

3.1 Materials and Methods

3.2.1 Study Area and Field Data Collection

Limpopo province is located on the northern part of South Africa. This province hosts Makhuduthamaga (Figure 14), which is the focus of this study. The area has rural villages that focus on smallholder maize farming (SDM, 2019). Hence, due to the dominance of smallholder farms in the area, it was selected as a case study. Weather stations from the Agricultural Research Council located in Nchabeleng, Ga-Rantho, and Leeuwkraal have recorded an average annual rainfall of 536 mm and an average annual temperature of 7 °C in winter and 35 °C in summer. Makhuduthamaga has an undulating topography with rock habitats in the form of rock outcrops, rocky ridges, rocky flats, and rocky refugia (Siebert et al., 2003).

Field surveys for the collection of training and validation data for different land cover types within the smallholder boundaries occurred from 18 to 21 February 2019. A handheld Garmin Global Positioning System (GPS) device which has a positional accuracy of 1.5 meters (on average mode) used to capture the coordinates of different land cover classes. The dominant land cover classes in the study area were captured; these include maize, bare land, and vegetation. The bare land and vegetation classes were combined to generate training samples (n = 9895 pixels) for the non-planted areas. The maize class consisted of n = 9802 pixels training samples. The samples were randomly selected into 80% training and 20% validation for each class. Constraining the land cover classes to two classes reduced the potential of classification errors from using the classes individually due to the variations in the natural occurrence of certain features. Limiting the area of investigation to the smallholder boundary excluded the farming activities in residents' backyards, thus only land that was demarcated as smallholder farmland was considered. A total of 18 smallholder farms were randomly selected in the field for validation purposes. Their areas were measured using a GPS. Most of these farms do not have proper access roads, which made it difficult to survey more farms.

© University of Pretoria

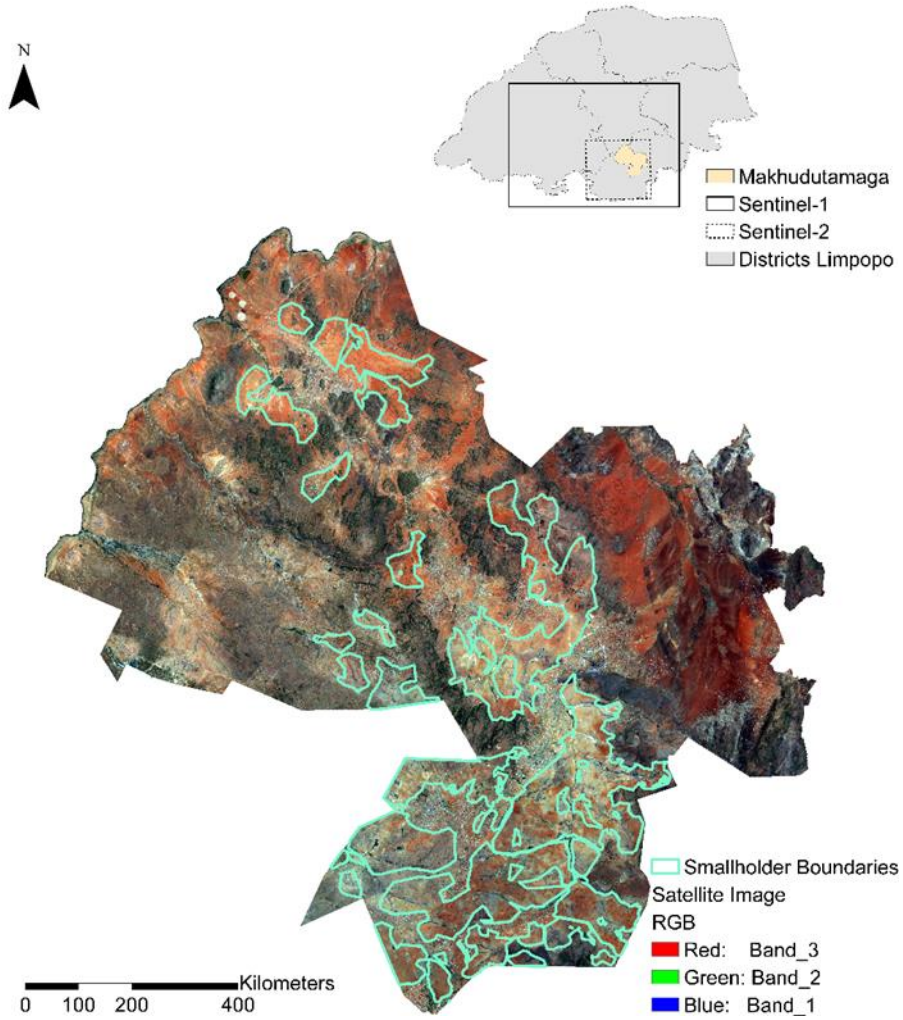

Figure 14. Study area location map for Makhuduthamaga in Limpopo, South Africa.

3.2.2 Sentinel-1 Data Acquisition and Pre-Processing

Sentinel-1 consists of a constellation of two satellites—Sentinel-1A and Sentinel-1B—which carry C-band SAR instruments to observe the Earth's surface. Sentinel-1 has a frequent repeat cycle of 12 days and the repeat cycle of the two-satellite constellation can offer a 6 day repeat cycle depending on the availability of observations from both of them and the location (Torres et al., 2012). The advantages of this configuration in the current study is that Sentinel-1 can capture the spatio-temporal variations of smallholder farms. This study used Sentinel-1 Level-1 Ground Range Detected (GRD) images, which cover the maize cropping season (November 2018–July 2019) inclusive of all the smallholder farms. These images were 22 in total, and they were acquired

© University of Pretoria

from the Copernicus Open Access Hub in the Interferometric Wide (IW) mode. Both the VV and VH polarizations with a 10 m spatial resolution were used.

Pre-processing of the radar images was done in the Sentinel Application Platform (SNAP) according to Filipponi (2019). Firstly, the orbit files were applied to update the orbit state vectors in the metadata files. Secondly, radiometric calibration was done by applying annotated image calibration constants to convert the intensity values into sigma nought values. Thirdly, speckle filtering was performed to reduce the granular noise caused by many scatters. Fourthly, the geometric distortions caused by topography were corrected for using the Range Doppler terrain correction with a 3 sec Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM). Finally, the two polarizations (VV and VH) were converted from a linear scale to a decibel scale and the ratio VV/VH was calculated.

Figure 15 illustrates the mean polarizations for selected planted maize farms and non-planted maize areas during the planting season. The mean backscatter values for the VV, VH, and VV/VH polarizations for maize are −13.66, −20.14, and 0.68 dB, respectively. The aggregated class has mean values of −14.83, −20.67, and 0.72 dB for VV, VH, and VV/VH polarizations, respectively. The VH polarization has the highest variance of 6.31 dB compared to VV polarization with 2.65 dB and the VV/VH ratio with 0.0009 dB. The VH polarization seems to respond more effectively to the growing stages of maize. A similar observation was made by Son et al. (2017) when they studied the rice crop also using Sentinel-1 data. This response is attributed to an increase in the volumetric structure of maize, which increases multiple reflections of the incoming signal.

© University of Pretoria

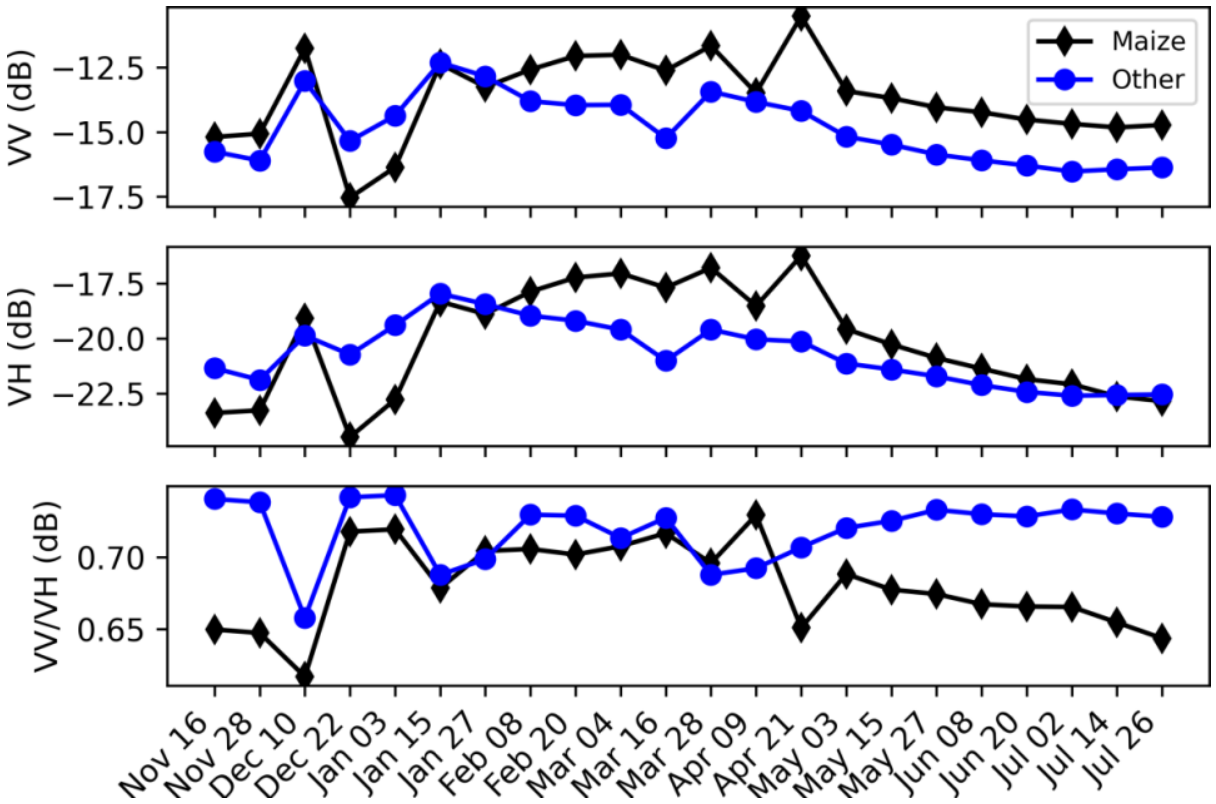

Figure 15. The mean raw VV, VH, and VV/VH backscatter profiles. The extracted polarizations are for maize crops and other classes, which refers to aggregated bare soil and grasslands.

3.2.3 Machine Learning Algorithms

The SVMs are advanced non-parametric statistical learning kernel-based algorithms commonly used in classification of remote sensing data (Foody and Mathur, 2004; Wan and Chang, 2019). Training data are projected into a higher-dimensional space using a linear/kernel-based function to optimally separate classes (Son et al., 2017). Parameters that optimally define the linear/non-linear hyperplane to separate the target classes are determined through an optimization problem. New data are evaluated based on the defined hyperplane constraints and categorized accordingly. The SVM requires regularization parameters that assist in tuning the model. These are C and gamma values, which were determined by the grid search method. In this study, the regularization parameter was 100, the gamma value was 0.01, and a Radial Basis Function kernel was used. A comprehensive review of the tuning method can be found in Mountrakis et al. (2011).

© University of Pretoria

The Xgboost is part of the classification and regression ensemble gradient boosting machines (e.g., Gradient Boosting and AdaBoost). This boosting technique is an improved version of Gradient Boosting and AdaBoost because it has a higher computational efficiency and improved capacity to deal with over-fitting. For example, Xgboost grows trees parallel to each other, whereas the original Gradient Boosting model builds the trees in a series configuration (Chen and Guestrin, 2016; Polly et al., 2019; Nobre et al., 2019). Boosting uses many weak classifiers to produce a powerful classifier in an additive manner. The classifiers are trained on the weighted versions of the training sample; misclassified data are given more weight during the iteration process so that the next step focuses on the misclassified data (Chen and Guestrin, 2016). The predictions improve over time and the final predictions are decided through a majority voting process to create vigorous predictions. This algorithm contains a rigorous number of regularization parameters that can be tuned to improve predictions and minimize overfitting (Georganos et al., 2018). These parameters are also determined using a grid search method.

3.2.4 Experimental Design

The experimental design scheme is illustrated in Figure 16. The first stage (i) involves preparation and pre-processing of Sentinel-1 images as described in Section 3.2.2. The second stage (ii) pixel-based PCA image fusion, which reduces the dimensions of the multi-temporal Sentinel-1 images into only 3 bands (i.e., principal components 1, 2, and 3). The bands describe more than 70% of the information contained in the multi-temporal images (Figure 17). The selection criteria were motivated by reducing the computational demands on the system, without compromising on the accuracy of the results. The third stage (iii) entails model-level data fusion and application of SVM and Xgboost classification algorithms. The last stage (iv) is the generation of the classification map. The second stage (ii) is necessary to reduce computational demands and Random Access Memory requirements. The third stage (iii) involves model-level data fusion using machine learning algorithms as described in Section 3.2.3. The data were separated into training (80%) and testing (20%) (Zhao et al., 2019). The performance and results of the algorithms in different experiments are compared using well-known evaluation metrics. This experiment was implemented

© University of Pretoria

using a Python programming platform on a Ryzen 9 3900, 12 cores processor at 3.8 GHz and 128 GB RAM computer.

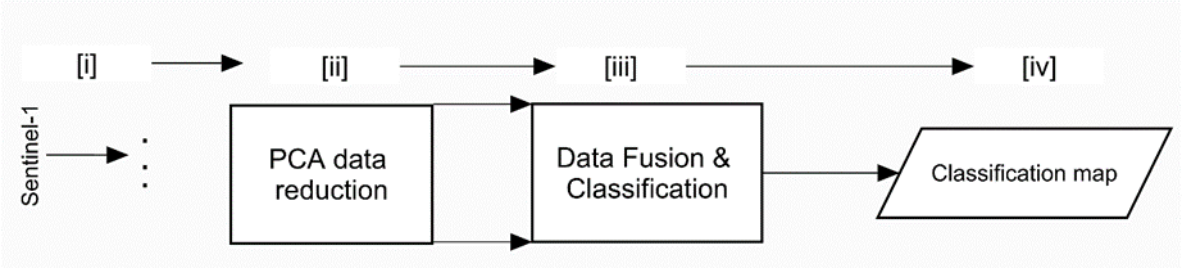

Figure 16. Schematic illustration of the experimental design.

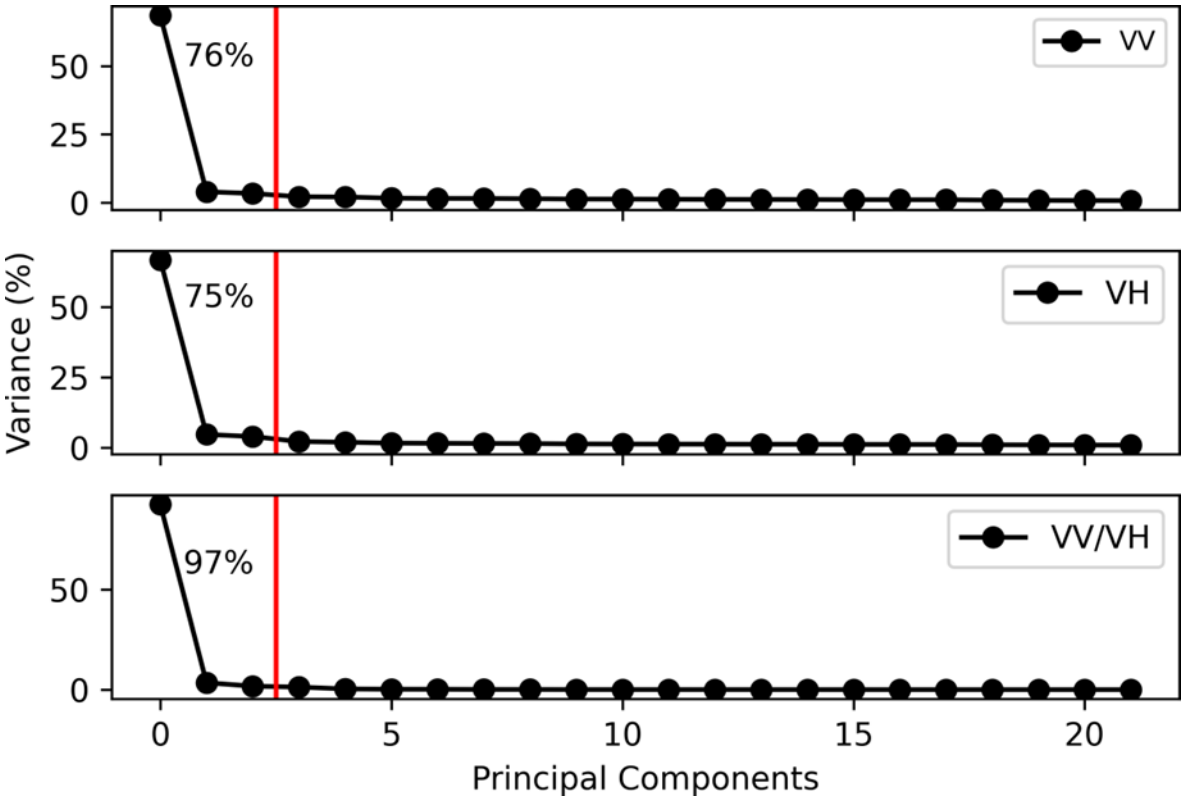

Figure 17. The variance explained by the VV, VH, and VV/VH components. The first three components, which explained greater than 70% of the variance, were selected.

© University of Pretoria

3.2.5 Accuracy Assessment and Smallholder Maize Area Estimation

The models and experiments were evaluated using standard statistical analysis, i.e., confusion matrix, cross-validation, overall accuracy, precision, recall, F1-Score, and McNemar's test (Skiena, 2017; Aggarwal, 2014). These statistical measures have been used to evaluate different machine learning algorithms, such examples include, but are not limited to, Petropoulos et al. (2012), Tong et al. (2020) and Cucho-Padin et al. (2019). The confusion matrix is constructed by comparing the results from the classification algorithm with the reference data collected in the field (Lewis and Brown, 2001). The matrix can be used to derive accuracy statistics for the map. Such statistical values include overall accuracy, kappa coefficient of agreement ($\hat{k}$), and conditional kappa coefficient of agreement ($\hat{k}_i$). These values are computed using Equations (1), (2), and (3 according to Congalton and Green (2008)):

$$\text{Overall accuracy} = \frac{\sum_{i=1}^{k} x_{ii}}{N}, \tag{1}$$

$$\hat{k} = \frac{N\sum_{i=1}^{k} x_{ii} - \sum_{i=1}^{k}\left(x_{i+} \times x_{+j}\right)}{N^2 - \sum_{i=1}^{k}\left(x_{i+} \times x_{+j}\right)}, \tag{2}$$

$$\hat{k}_i = \frac{N\left(x_{ii}\right) - \left(x_{i+} \times x_{+j}\right)}{N\left(x_{i+}\right) - \left(x_{i+} \times x_{+j}\right)} \tag{3}$$

where k is the land cover classes in the confusion matrix, x_{i+} and x_{+j} represent marginal total for row *i* and column *j*. x_{ii} represents the number of observations in row *i* and column *i*. *N* represents total number of samples.The overall accuracy describes the proportion of the area mapped correctly. It provides a user with a probability that a randomly selected location on a map is correctly classified (Olofsson et al., 2013). Kappa values that are more than 80% indicate good agreement between the reference and derived classification map. The $\hat{k}$ measures the overall level of agreement between the reference data and the model data. The $\hat{k}_i$ allows computation of the

© University of Pretoria

level of agreement between the reference data and the model data for a specific class *i*.

Precision measures the ability of the algorithm not to label a true positive sample (tp) or a sample that is false positive (fp). Recall measures the ability of the algorithm to find all the true positives, and false negative is represented by fn. F1-Score is the harmonic mean calculated from both precision and recall values. These statistical values are calculated according to Equation (4) (Lewis and Brown, 2001; Davis and Goadrich, 2006; Flach, P.; Kull, 2015):

$$\begin{aligned} precision &= \frac{tp}{tp + fp'}, \\ recall &= \frac{tp}{tp + fn'}, \\ F1-Score &= (1+\beta^2)\frac{precision * recall}{\beta^2 * precision * recall}, where \beta = 1. \end{aligned} \tag{4}$$

Cross-validation is another statistical method used to evaluate the performance of the model by dividing the data into k-folds (e.g., a standard value of 10 folds was used); the algorithm uses one set of data as training and the other sets are used to evaluate the model. During this iterative process, the accuracy score is calculated. The final cross-validation value is derived of the average accuracies from each iterative process. The superiority and significance between the SVM and Xgboost algorithms for each experiment were evaluated using a nonparametric McNemar's statistical test (McNemar, 1974; Edwards, 1948; De Leeuw et al., 2006). The test is based on chi-square (χ^2) statistics, calculated using Equation (5):

$$\chi^2 = \frac{\left(|f_{12} - f_{21}| - 1\right)^2}{\left(f_{12} + f_{21}\right)}, \tag{5}$$

where f_{12} denotes the number of cases that are wrongly classified by Model 1 but correctly classified by Model 2, and f_{21} denotes the number of cases that are correctly classified by Model 1 but wrongly classified by Model 2 (Manandhar et al., 2009). This was computed from two contingency matrices from the two algorithms that were being tested.

© University of Pretoria

The unbiased proportional mapped areas were estimated using the method described by Olofsson et al. (2013). This method takes into account errors in misclassifications as reported in a confusion matrix. The mapped areas are estimated at 95% confidence intervals, and this is useful in providing error margins for the estimated areas for the end-users. Additional validation of the classification models' ability to estimate smallholder maize was done. The areas measured at 18 smallholder farms were compared to the estimated areas from the SVM and Xgboost algorithms through a linear regression model. The p-value (p) and Pearson correlation coefficient (R) were derived to evaluate the agreement.

3.2 Results

3.3.1 Accuracy Assessment

A two-stage data fusion approach was used in this study, utilizing a time-series of Sentinel-1 polarization datasets. The SVM and Xgboost accuracy assessment results are listed in Table 10. The SVM has an overall accuracy of 97.1%, cross-validation score value of 89%, kappa value of 93.3%, and the conditional kappa coefficient of agreement of 90.54% and 95.7% for maize and non-planted classes, respectively. The Xgboost has an overall accuracy of 96.8%, cross-validation score value of 96%, kappa value of 92.6%, and conditional kappa coefficient of agreement of 90.4% and 94.4% for maize and non-planted classes, respectively. The maize classified pixels were similar for both classifiers based on the confusion matrix. The precision, recall, and F1-Score values for both algorithms have similar values that are more than 90% for both classes. It can also be noted that the recall for the planted maize class in both cases is approximately 3.7% lower compared to the precision score value. This observation is also supported by the kappa statistic and suggests that the planted maize class is less accurately classified compared to the non-planted class.

© University of Pretoria

Table 10. Accuracy assessment produced for the Sentinel-1 multi-temporal classification using the Support Vector Machine (SVM) and Extreme Gradient Boosting (Xgboost) algorithms.

Model	**Overall Accuracy**	**Cross-Validation**	**Confusion Matrix**	
SVM	0.971	0.89 +/−0,05	Planted Maize 20 139 628	Non-Planted 1457 50 790
Xgboost	0.968	0.96 +/−0.02	Planted Maize 20 115 825	Non-Planted 1481 50 593

	SVM		**Xgboost**	
Classes	Planted Maize	Non-Planted	Planted Maize	Non-Planted
Precision	0.970	0.972	0.961	0.972
Recall	0.933	0.988	0.931	0.984
F1-Score	0.951	0.980	0.946	0.978

These results show that the SVM and Xgboost produced an acceptable performance in mapping smallholder farms and illustrated the capability of two-stage image fusion employed in this study. In particular, both algorithms classified the non-planted area class better by approximately 5% compared to the planted maize class. The cross-validation score indicates that the Xgboost algorithm is more consistent and stable compared to the SVM algorithm. The Xgboost algorithm cross-validation score outperformed the SVM algorithm cross-validation score by 7%. This is in contrast with the other statistical measures (Table 10), which seem to suggest that SVM has outperformed the Xgboost algorithm.

© University of Pretoria

In situations where statistical evaluation matrices seem to contradict each other, a non-parametric statistical test must be conducted. In this case, a McNemar's significance test was applied. If the estimated test statistic is lower than the critical chi-square table value (i.e., 3.84 at 95% confidence level), the null-hypothesis is rejected and it is concluded that there is no significant difference between the two model results (Sahin and Colkesen, 2019). The McNemar's chi-square value of 64.62 and *p*-value of 9.085×10^{-16} were obtained by comparing the two algorithms. Therefore, the null-hypothesis was rejected and the conclusion was that the two results are statistically different from each other.

3.3.2 Variable Importance

Permutation variable importance was used to compute variable importance using the two estimators (SVM and Xgboost). The permutation algorithm can be defined to be the decrease in a model score when a single feature value is randomly shuffled (Breiman, 2001). Variable importance for each VV, VH, and VV/VH PCA 1, 2, and 3 polarizations are depicted in Figure 18. The VH and VV PCA polarizations formed the top six most important variables and the least important variables were the VV/VH PCA ratios. Specifically, the VH PCA 3 received the highest score, followed by the VV PCA 1 and VH PCA 2. The dominance of VH and VV polarizations was expected. Figure 19 depicts the VV, VH, and VV/VH PCA polarization composites. Smallholder maize farms are clearly enhanced by the VV and VH PCA polarization composites compared to the VV/VH polarizations.

© University of Pretoria

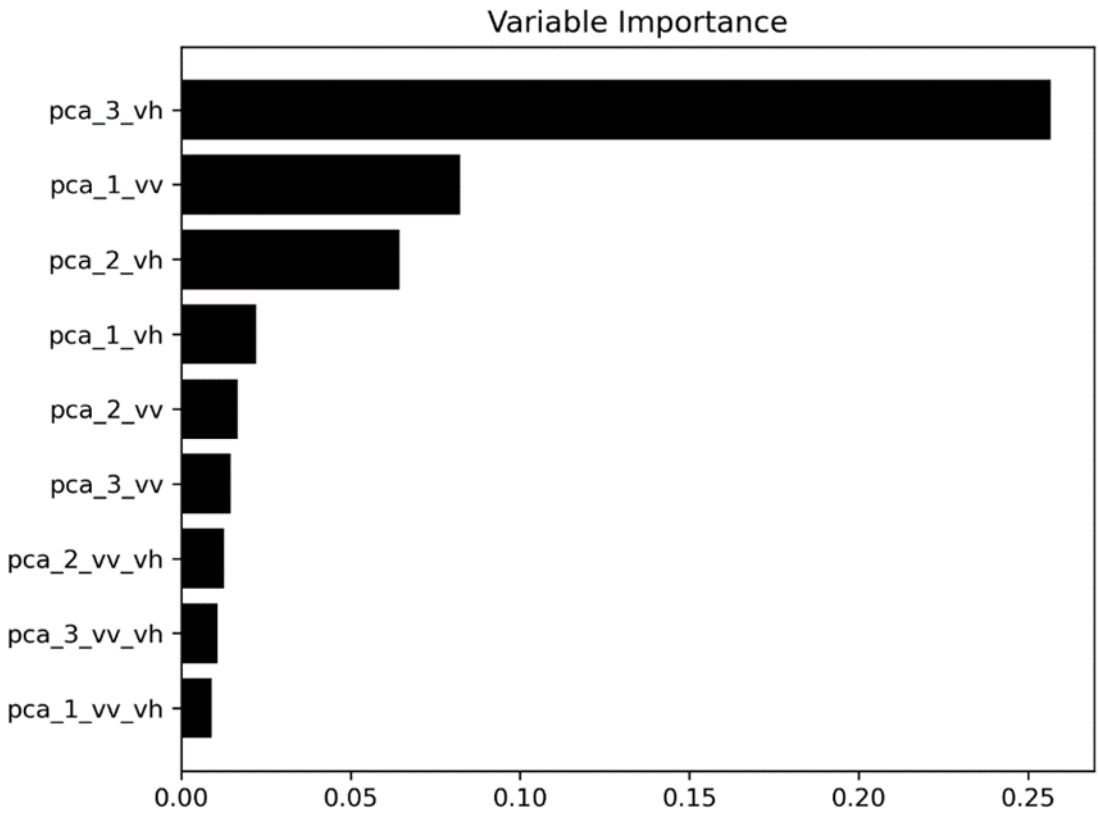

Figure 18. Permutation importance scores for the Principal Component Analysis (PCA)-derived images used in the analysis. PCA 3 for the VH polarization is the most important variable in our study. The same results were obtained for the two estimators.

© University of Pretoria

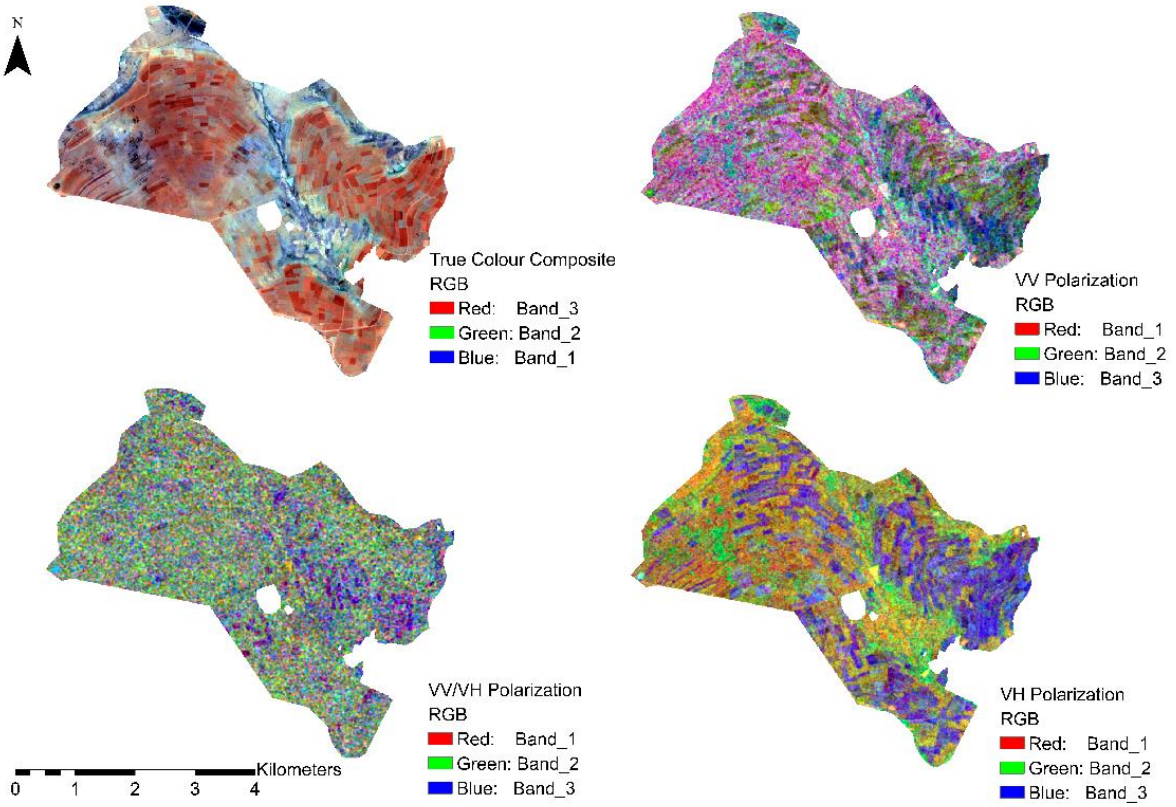

Figure 19. Examples of the PCA images for different polarizations derived from Sentinel-1 datasets. The PCA VH polarization composite seems to visually enhance smallholder farms compared to other polarizations.

3.3.3 Mapping and Area Estimate for Smallholder Maize Farms

The maps for the maize planted areas produced by the SVM and Xgboost algorithms are depicted in Figure 20. The classification maps reveal the spatial distribution of the smallholder maize farms in our study area. It can be seen that most farmers that planted maize during the 2018/2019 season are from the south eastern part of Makhuduthamaga. These observations are consistent with both maps that were produced by the two algorithms. Visual inspection reveals no obvious disagreement between the two maps as predicted by the SVM and Xgboost algorithms.

The unbiased proportional areas were generated. The SVM algorithm estimated the planted maize class to be 7073.558 ± 0.01 ha and the non-planted class was estimated to be 33420.96 ± 0.01 ha. Meanwhile, the Xgboost estimated the planted maize class area to be 7303.32 ± 0.180 ha and the non-planted class was estimated

© University of Pretoria

to be 33191.2 ± 0.820 ha. It is worth noting that the SVM algorithm has better error margins (0.01 ha) for both classes compared to the Xgboost algorithm, which has error margins of 0.18 and 0.82 ha for the planted maize class and non-planted areas, respectively. The areas for the 18 smallholder farms (Figure 21) compared well with those generated by the classification models. The SVM classifier had a better fit (R = 0.89) in comparison with the Xgboost algorithm (R = 0.84). The linear model was an ideal fit for the data. The positive relationship was significant at a 95% level ($p < 0.5$).

© University of Pretoria

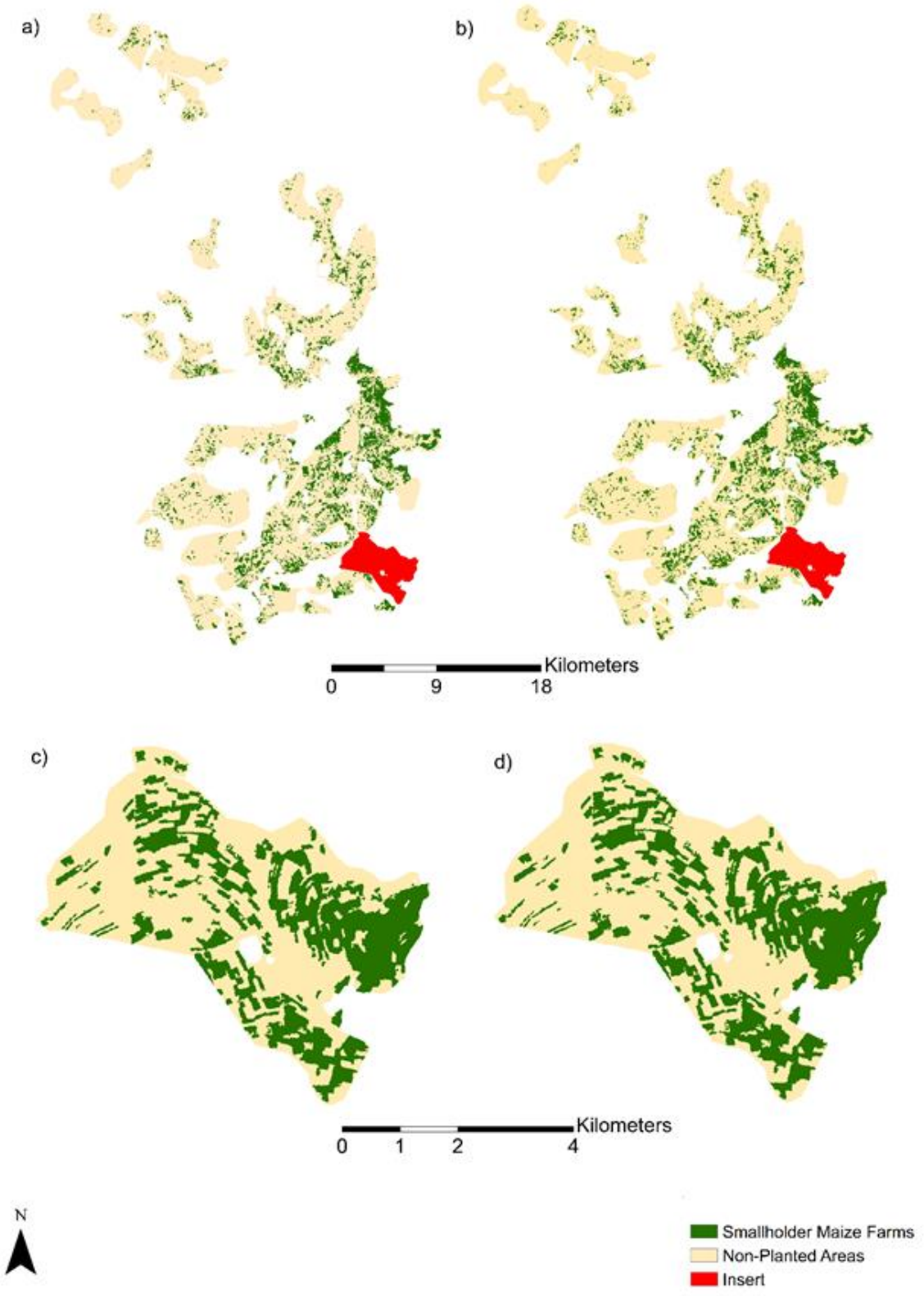

Figure 20. Planted maize crop maps produced by the SVM (a) and Xgboost (b) algorithms. Insert maps for SVM and Xgboost are represented by (c) and (d), respectively.

© University of Pretoria

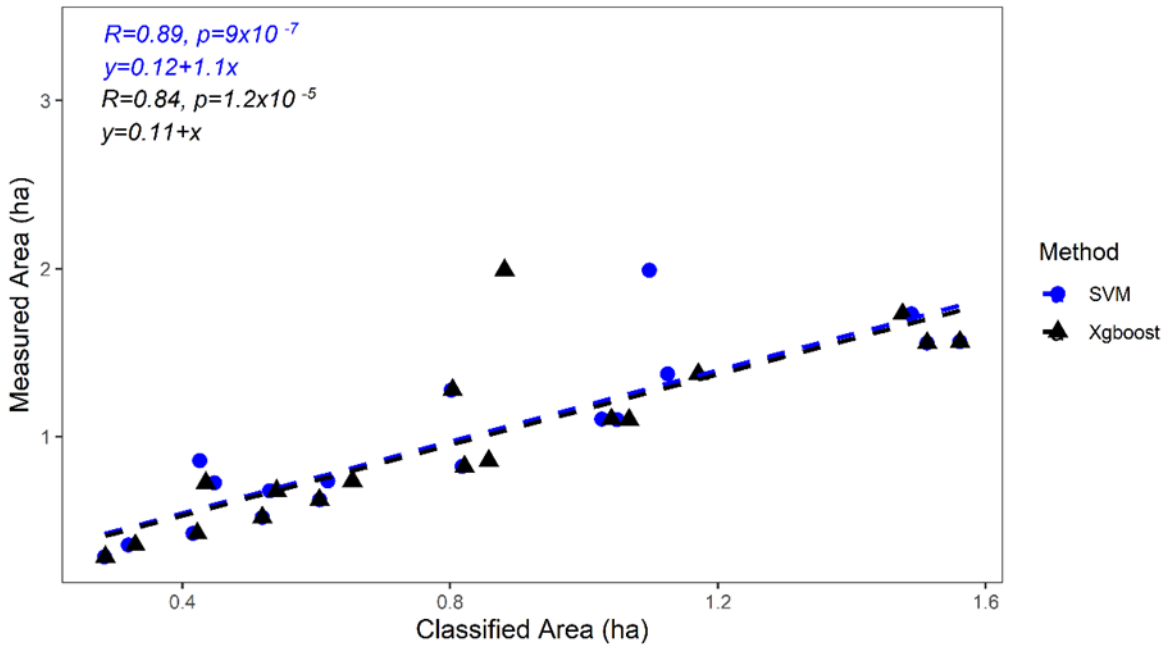

Figure 21. Comparison of the field measured areas (y) to those generated by the classification models (x) applying the SVM and Xgboost algorithms.

3.3 Discussion

This study used Sentinel-1 multi-temporal datasets to map smallholder maize farm spatial distribution and to estimate maize production area for the maize crop. A two-staged image fusion technique was employed. The first stage involved using a pixel-based PCA technique to transform the original multi-temporal backscatter values into three component images that explained more than 70% of the information. This was done for the VV, VH, and VV/VH polarizations. The second stage involved model-level fusion, where all the components were used as input features into the machine learning algorithms. The SVM and Xgboost algorithms were used as classifiers to map the distribution of the maize farms and production area in Makhuduthamaga of Limpopo province, South Africa. This study found that Sentinel-1 had a high capacity to map smallholder maize planted areas with the application of machine learning algorithms. Furthermore, the two processing strategies used in this study detected smallholder maize farms with acceptable accuracy.

The accuracy assessment results were also expected. The overall accuracies were better than 90%, the cross-validation scores were greater than 85%, and the

© University of Pretoria

kappa coefficient of agreement and conditional kappa coefficient of agreement were all better than 90%. These results confirm the suitability of our approach in mapping smallholder farms using Sentinel-1 multi-temporal datasets. Other studies such as Ndikumana et al. (2018) used Sentinel-1 multi-temporal data to map agricultural crops by applying a Deep Recurrent Neural Network and obtained favorable results that were better than 85% in accuracy. The SVM and Xgboost algorithms estimated maize production areas to be 7073.558 ± 0.01 ha and 7303.32 ± 0.180 ha, respectively. These values are relatively comparable to each other and SVM seems to have smaller error margins at a 95% confidence level and slightly higher overall accuracy than the Xgboost. However, for cross-validation scores, the Xgboost performed better. McNemar's test showed that the results from the two algorithms were statistically different from each other. Other authors have evaluated different machine learning algorithms and obtained mixed performance indicators. Aguilar et al. (2018) used different ensemble classifiers (Random Forest, SVM, and Majority Voting) to map smallholder farming systems based on the cloud-based multi-temporal approach and obtained overall accuracies ranging from 60% to 72%. Dong et al. (2018) used Xgboost algorithms together with Decision Tree, Random Forest, and SVM to map land cover using Gaofen-3 Polarimetric SAR (PolSAR) data and obtained overall accuracies ranging from 88.4% to 93%. Zhong et al. (2019) used machine learning algorithms and Deep Neural Network algorithms to map crop types and found that a Convolutional Neural Network model achieved 85.5%, while the Xgboost achieved 82.4% in overall accuracy under a multi-temporal classification scenario. Overall, the results produced by the classification algorithms compared favorably with the ground-based measured areas. Both algorithms had an agreement of more than 80%.

There are a few factors that may have contributed to the mapping errors as produced by the two algorithms and the radar data. Examples of these include, but are not limited to, loss of information during the PCA data reduction stage, backscatter mixing, and different planting patterns. Dimensional reduction may have contributed to the mapping errors. However, the three components that were kept for each polarization at more than 70% proved sufficient in our study. According to Woodhouse (2017), backscatter intensity is sensitive to variations in scattering geometry, distribution of scatterer size, surface reflectivity beneath the canopy, leaf area density,

© University of Pretoria

row structure, and orientation relative to the range domain of the radar. Smallholder farms normally practice crop mixing, un-equal row planting patterns, and lack of irrigation systems. These practices can influence the backscatter intensity from maize. Scattering from nearby vegetation, such as grass and soil-canopy multiple scattering, can also contribute towards misclassification.

We showed that the PCA data reduction method can be used to facilitate the mapping of smallholder maize farms. Machine learning algorithms require data that can be separable to successfully classify data into their respective classes [50]. The PCA provides this by decorrelating the multi-temporal backscatter values into components that describe unique information for different classes, therefore enhancing the probability of accurate classification. Maize can grow up to an average height of 2 m and the structural volume of the crop also increases as the leaves also grow. This makes it possible to map maize with radar data, since the VV and VH polarizations are sensitive to vegetation structure and volumetric changes (McNairn et al., 2009). A frequent revisit of Sentinel-1 of 10–12 days and its high spatial resolution of 10 meters can capture the phenological stages of maize (Torres, 2012). The increase in backscatter intensity for the maize class makes it possible to map smallholder farms in complex environments. PCA also suppresses other classes with low variable backscatter over time; these classes include grasslands and bare soil in our study area. The PCA image composites provide clear examples, where the advantage of the first stage of image fusion used in this study can be seen (Figure 16). The high level of importance of VH and VV polarizations were expected. Other studies, such as Arias et al. (2020), illustrated that the VH, VV, and VH/VV polarizations ranked differently depending on the type of crop that was investigated. VH polarization was more suitable for rice and rapeseed discrimination, VV polarization was more suitable for alfalfa, and the VH/VV ratio was suitable for discriminating crops from different seasons.

The results can be used to generate spatial agricultural information such as estimating crop production areas and their spatial distributions in areas where survey datasets are not available, such as in our study. The results can be used to inform local government about the levels of agricultural activities in rural communities, thus

© University of Pretoria

providing ways to forecast food shortages and improve food security. The use of Sentinel-1 multi-temporal data provides an opportunity to afford this critical information regardless of the environmental conditions such as clouds or lack of extensive reference data. These results can also be used to contribute towards the SDG number 2. We therefore recommend the use of Sentinel-1 multi-temporal data to map smallholder farms at a provincial scale. More studies need to be done to explore the phase and amplitude data extracted from the backscatter intensities and their contribution to the accuracy of classifying smallholder farms. Different image fusion techniques and multi-sensor data fusion should also be explored.

The limitation of this study was that there were no agricultural statistics to independently validate the areas obtained by the machine learning algorithms. These validation data are normally collected by local agricultural departments. For example, the United States Department of Agriculture (USDA) uses remote sensing and extensive reference data provided by the National Agricultural Statistics Service (NASS) to generate the crop layer and associated statistics (Boryan et al., 2011). In areas with limited reference data, such as smallholder farms in developing countries, remote sensing technology provides a sustainable way to generate agricultural statistics with reasonable accuracies (Jain et al., 2013; Useya and Chen, 2019). Processing multi-temporal data requires computational resources that are otherwise not easily accessible in developing countries. The Google Earth Engine (GEE) and other platforms provide an alternative solution to process data online, and these platforms allow for large-scale data processing at a relatively low cost. For example, Jin et al. (2019) used the GEE platform to process Sentinel-1 data to map smallholder maize farms.

Future work should focus on testing this approach in different areas where smallholder farms are dominant. The response and efficiency of this approach should also be tested on different crop types. The operational model should be developed to consider the time domain to allow forecasting smallholder maize production areas. The phase and amplitude data from multi-temporal Sentinel-1 data and multi-sensor data should be explored in mapping smallholder farms in the future. These research

© University of Pretoria

opportunities will ensure that remote sensing technology can be fully utilized to support SDGs.

3.4 Conclusion

This study presented Sentinel-1 multi-temporal data for mapping smallholder maize farms' spatial distribution and estimated production areas. The two-stage image fusion approach was adopted. The SVM and Xgboost machine learning algorithms were applied. The results revealed that most smallholder farms in our study area are distributed in the south eastern part of Makhuduthamaga. The algorithms provided comparable statistical evaluation results. However, McNemar's test showed that the results from the two algorithms were statistically different from each other. The SVM and Xgboost algorithms estimated maize production areas to be 7073.558 ± 0.01 ha and 7303.32 ± 0.180 ha, respectively, for the region. The classified areas for selected farms compared favorably with the measured areas in the field and the SVM classifier had a better fit (R = 0.89) in comparison with the Xgboost algorithm (R = 0.84). The SVM algorithm seems to have generally performed better than the Xgboost algorithm. The use of multi-temporal Sentinel-1 with a two-stage image fusion approach proved to be effective in mapping smallholder farms. This framework can be used to support the SDGs and to provide spatial agricultural information to inform policy design and implementation by local government. Different seasons and different crop types should be tested using this approach, including extraction of phase and amplitude data from multi-temporal Sentinel-1 data. Multi-sensor data fusion should be explored to improve the mapping of smallholder farms in the future.

© University of Pretoria

Chapter 4

Modeling the Spatial Distribution of Soil Nitrogen Content at Smallholder Maize Farms Using Machine Learning Regression and Sentinel-2 Data

Based on: Mashaba-Munghemezulu, Z., Chirima, G.J., Munghemezulu, C., 2021. Modeling the Spatial Distribution of Soil Nitrogen Content at Smallholder Maize Farms Using Machine Learning Regression and Sentinel-2 Data. *Sustainability*, 13(11591).

Abstract

Nitrogen is one of the key nutrients that indicate soil quality and an important component for plant development. Accurate knowledge and management of soil nitrogen is crucial for food security in rural communities, especially for smallholder maize farms. However, less research has been done on generating digital soil nitrogen maps for these farmers. This study examines the utility of Sentinel-2 satellite data and environmental variables to map soil nitrogen at smallholder maize farms. Three machine learning algorithms—random forest (RF), gradient boosting (GB), and extreme gradient boosting (XG) were investigated for this purpose. The findings indicate that the RF (R^2=0.90, RMSE=0.0076%) model performs slightly better than the GB (R^2=0.88, RMSE=0.0083%) and XG (R^2=0.89, RMSE=0.0077%) models. Furthermore, the variable importance measure showed that the Sentinel-2 bands, particularly the red and red-edge bands have a superior performance in comparison to the environmental variables and soil indices. The digital maps generated in this study show the high capability of Sentinel-2 satellite data to generate accurate nitrogen content maps with the application of machine learning. The developed framework can be implemented to map the spatial pattern of soil nitrogen. This will also contribute to soil fertility interventions and nitrogen fertilization management to improve food security in rural communities. This application contributes to the Sustainable Development Goal number 2.

Keywords: satellite data; random forest; gradient boosting; extreme gradient boosting; soil fertility; digital mapping

© University of Pretoria

4.1. Introduction

Improving soil nutrient management at smallholder maize (Zea mays L.) farms is imperative for ensuring food security in developing countries. Smallholder maize farms are crucial for the livelihoods of rural communities in Africa who depend on agriculture for food security and their local economic activities. Amongst the most important nutrients is nitrogen, not only is it a component of the chlorophyll molecule but is also essential for maize growth, quality, and yield (Sinclair and Muchow, 1995; Otto, 2016; Chlingaryan et al., 2018). The soil is one of the most important nitrogen reservoir in the terrestrial ecosystems (Batjes, 1996). Developing frameworks to map the spatial variability of soil nitrogen is necessary for the local government, farmers, and stakeholders to identify nitrogen excesses or deficiencies. Such information will guide soil fertility interventions at smallholder farms. In the long term, improved soil nitrogen content management will enhance maize productivity (Lemcoff and Loomis, 1986; Osterholz et al., 2017). This application is particularly important for resource limited smallholder maize farms such as those in developing countries, for example South Africa, which have reported sub-optimal yields, infertile land, and land degradation in previous studies (Shi and Tao, 2014; Fischer and Hajdu, 2015).

Several soil databases and sources are available that archive soil nutrient information for South Africa. Examples of these include the Africa Soil Information Service (AfSIS) which archives soil nutrient maps at a 250 m spatial resolution for Africa (http://africasoils.net/). The Harmonized World Soil Database (HWSD) nutrient map, which has a spatial resolution of 1 km (Jones and Thornton, 2015). Other products such as the SOTER-based soil parameter estimates (SOTWIS) product for Southern Africa have a 1:2 M (million) scale resolution (Batjes, 2004). The soil Atlas of Africa dataset for soil groups has a 1:3 M scale resolution (Jones et al., 2013). Although these products are available, they have a coarse spatial resolution to guide soil nutrient management efforts at smallholder farms, which are typically 0.5-2 ha in size. These types of farms are often fragmented and heterogeneous in most parts of the world including South Africa, which necessitates the use of improved resolution data for digital soil mapping (Chivasa, et al., 2017).

© University of Pretoria

The Sentinel-2 mission has sensor capabilities with a potential to estimate soil nutrients at smallholder farms. This satellite has an improved spatial resolution of 10-60 m, wide swath of 290 m and a frequent revisit cycle of 5-10 days (Drusch et al., 2012). Additionally, the Sen-tinel-2 data is compatible with Landsat-8 and Satellite Pour l'Observation de la Terre (SPOT) data (Wang et al., 2017). The difference between Sentinel-2 and other medium resolution sensors such as Landsat-8 is the presence of the red-edge band region in Sentinel-2. The red-edge region lies between the red and near infrared portions of the electromagnetic spectrum and is distinguished by a sharp increase in vegetation reflectance (Filella and Penuelas, 1994). This current study relies on soil and vegetation indices derived from strategic locations of the electromagnetic spectrum to estimate the soil nitrogen content for smallholder maize farms.

Different techniques have been applied for digital soil mapping. The commonly used models are multiple linear regression (Shi et al., 2013), principal component analysis regression (Yang et al., 2016), generalized additive model (de Brogniez et al., 2015) and kriging (Xu, 2018). Recently, machine learning algorithms (support vector machines, decision trees, random forests, artificial neural networks) are widely used in remote sensing studies (Friedl and Brodley, 1997; Chang and Islam, 2000; Heumann, 2011; Wang et al., 2016). These algorithms are beneficial because they can learn from limited data and reduce errors through an adaptive learning process (Belgiu and Drăguţ, 2016; Cooner et al., 2016). However, studies using these techniques for soil nitrogen mapping at smallholder maize farms are lacking (Xu, 2018). Particularly, as machine learning algorithms are not universally applicable in different environments. This necessitates the evaluation of different machine learning algorithms for applicability in our own context to understand the distribution of soil nitrogen content at the locality.

This paper uses the random forest (RF) algorithm, gradient boosting algorithm (GB) and extreme gradient boosting (XG) machine learning algorithm in a regression format. These algorithms were used because they can deal with noisy, high-dimensional and non-linear data (Izquierdo-Verdiguier, 2014; Li, 2016). The algorithms are applied to Sentinel-2 imagery to predict the spatial patterns of soil

© University of Pretoria

nitrogen content at selected smallholder maize farms in Makhuduthamaga district, South Africa. The study addresses the following specific research questions: 1) What are the relationship between soil nitrogen content and different predictor variables? 2) How effective are the selected machine learning algorithms in predicting soil nitrogen content? 3) Which predictor variables are fundamental for modelling soil nitrogen content? Lastly, 4) What is the spatial distribution pattern of soil nitrogen at smallholder maize farms?

4.2. Material and Methods

The overview of the methodological approach used in this study is summarized in Figure 22. The Sentinel-2 imageries were pre-processed to correct for atmospheric effects and band indices were calculated. Ancillary data describing the environmental variables and some of the Sentinel-2 bands were resampled to 10 m. Nine experiments with different data configurations were conducted using the Sentinel-2 bands, spectral indices and environmental variables. Three machine learning regression algorithms – RF, GB, and XG were then applied in each experiment using 70% of the nitrogen content measurements for training the model. The remaining 30% of the data was used for model evaluation with commonly used statistical metrics. Variable importance for the predictors was determined from the scores derived by the three machine learning regression models. Lastly, the spatial pattern of soil nitrogen at the smallholder maize farms was mapped.

© University of Pretoria

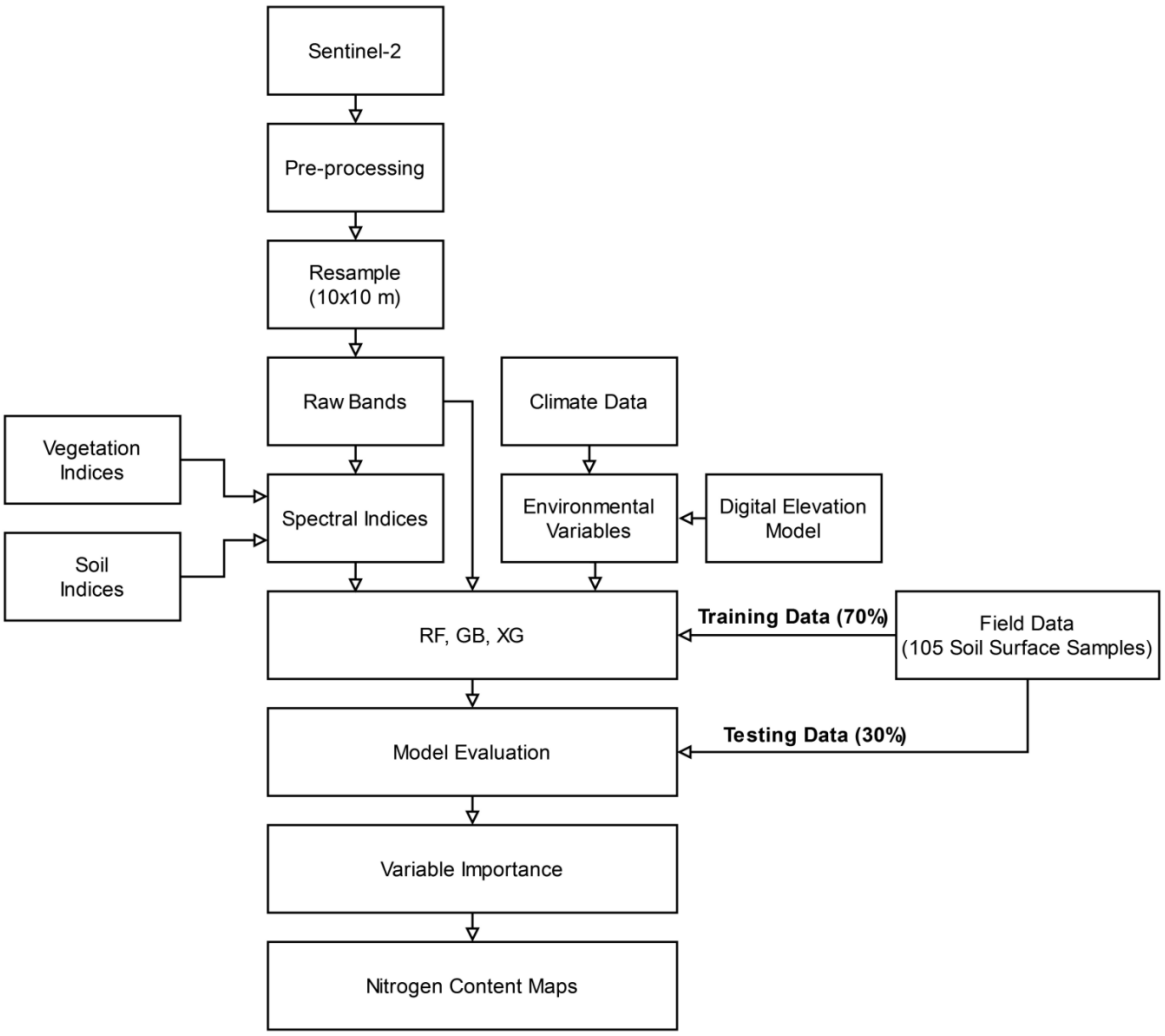

Figure 22. The proposed methodological framework for mapping soil nitrogen content at smallholder maize farms.

4.2.1. Study Area

Soil nitrogen samples were collected from the smallholder maize farms of Makhudutamaga district located in the Northern Part of South Africa (Figure 23). This district has a low elevation (799-1047 m) in the north western part and a higher elevation (1295-1791 m) in the central and southern parts. The topography is undulating with rock habitats such as rock outcrops, rocky ridges and rocky refugia (Siebert, 2003). This district was selected because most of the rural population are smallholder maize farmers; they farm mainly for subsistence and partially for selling in local markets. Smallholder maize production is predominant in the Southern part of

© University of Pretoria

the district (SDM, 2019). The farmers add manure to their fields in November. Maize is planted during December and January. The growing period is between February to May. Harvesting takes place in June and no maize is present in the smallholder farms during July-November. The smallholder farms in the district are rain-fed. The annual rainfall is 536 mm and an average annual temperature of 7°C in winter and 35°C in summer according to the Agricultural Research Council stations located in Nchabeleng, Ga-Rantho and Leeuwkraal areas.

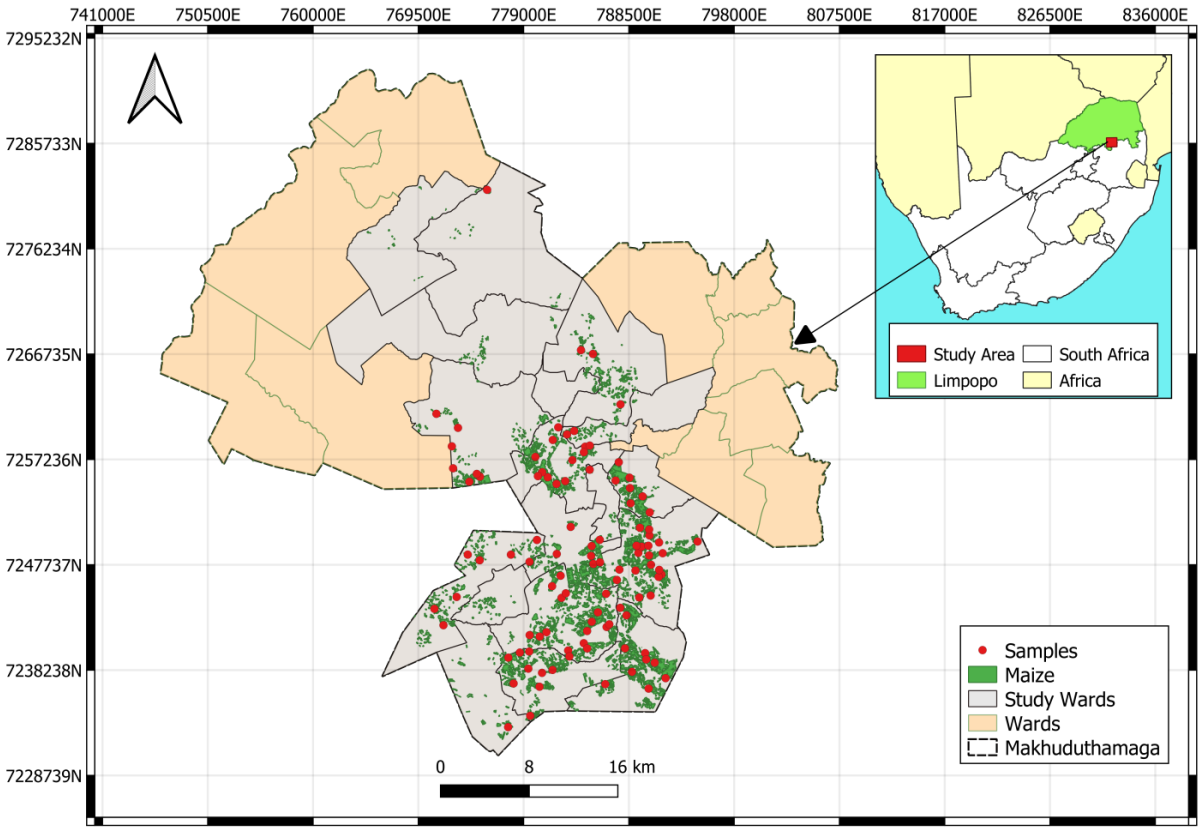

Figure 23. The location of the study wards and smallholder maize farms that are considered for soil nitrogen data collection in Makhuduthamaga district, South Africa.

4.2.2. Field Data Collection and Laboratory Analysis

A total of 105 soil surface samples were collected from the topsoil layer (0-20 cm) at the smallholder maize sample farms during 14-17 May 2019 corresponding to a period of low rainfall. The positions for each sample were captured with a handheld Global Positioning System (GPS). The samples were then processed at the Agricultural

© University of Pretoria

Research Council Analytical Laboratory where they were air-dried at room temperature (25 °C), crushed, and passed through a 2 mm sieve to remove coarse soil materials such as gravel or plant roots. The soil total nitrogen content was then determined through analytical processing with the Kjeldahl digestion method. The soil properties are summarized in Table 11 according to the dominant soil type at the top (Haplic Acrisols) and least dominant soil at the bottom (Lithic Leptosols). These were extracted from the Harmonized world soil database (Jones and Thornton, 2015).

Table 11. Soil attributes for the dominant soil types in smallholder farms.

Soil Type	Topsoil sand fraction (%)	Topsoil silt fraction (%)	Topsoil clay fraction (%)	Topsoil Texture	pH (H2O)	Bulk Density (kg/dm^3)	Organic Carbon (% weight)
Haplic Acrisols	57	19	24	Sand clay loam	5.1	1.4	0.8
Ferric Luvisols	65	18	17	Sandy loam	6.4	1.5	0.6
Lithic Leptosols	43	29	28	Clay loam	7.5	1.3	0.4

4.2.3. Sentinel-2 Data Acquisition and Pre-processing

We used Sentinel-2 MSI level-1C (L1C) data acquired from the Copernicus Open Access Hub. The image for 17 May 2019 was used in this study. This image covered the field sampling date and was appropriate considering that the image was cloud free. The L1C product images consist of top-of-atmosphere (TOA) reflectance after radiometric correction and geometric corrections (ortho-rectification and spatial registration) with a sub-pixel accuracy (https://sentinel.esa.int). Sentinel-2 MSI has 13 bands, which have different spatial resolutions. This study made used of 10 bands (visible, near-infrared, red-edge, and shortwave infrared) as summarized in Table 12 and excluded the bands which are related to water and atmosphere elements. The Sentinel-2 TOA images were pre-processed with Sen2Cor plugin in Sentinel Application Platform (SNAP) to convert them to bottom-of-atmosphere reflectance (BOA) and the 20 m bands were resampled to a 10 m spatial resolution.

© University of Pretoria

Table 12. Sentinel-2 multi-spectral bands used in this study Drusch et al., (2012).

Variable	Description		
Raw bands	**Central Wavelength (nm)**	**Bandwidth (nm)**	**Spatial Resolution (m)**
B2–Blue	490	65	10
B3–Green	560	35	10
B4–Red	665	30	10
B5–RE1	705	15	20
B6–RE2	740	15	20
B7–RE3	783	20	20
B8–NIR	842	115	10
B8a–RE4	865	20	20
B11–SWIR1	1610	90	20
B12–SWIR2	2190	180	20

Note: Red Edge (RE), Near Infrared (NIR), Short Wave Infrared (SWIR)

4.2.4. Spectral Indices

Spectral indices were generated from the Sentinel-2 bands. The vegetation indices that are included in the current study were selected by fitting the RF, XG, and GB machine learning regression models. Vegetation indices that optimized the coefficient of determination (R^2) in relation to the nitrogen content for each model were retained. This procedure was done because similar studies have reported a diverse range of vegetation indices (Wang et al., 2018; Xu, 2018; Mandal, 2016). The vegetation indices evaluated based on the RE were the: Normalized Difference Vegetation Index RE 1, 2 and 3 narrow (NDVIRE1n, NDVIRE2n, NDVIRE3n), Normalized Difference Vegetation Index RE 1 (NDRE1), Normalized Difference Vegetation Index RE 1 modified (NDRE1m), Modified Simple Ratio RE (MSRRE), Chlorophyll Index RE (CLRE) and Normalized Difference Vegetation Index RE (NDVIRE). Other indices based on the NIR, SWIR1, SWIR2 and visible parts of the electromagnetic spectrum were also evaluated. These indices included the Plant Senescence Reflectance Index (PSRI), Enhanced Vegetation Index (EVI) and the Green Normalized Difference Vegetation Index (GNDVI). Additionally, the Difference Vegetation Index (DVI), Normalized Difference Water Index (NDWI), Renormalized Difference Vegetation Index (RDVI), Normalized Difference Vegetation Index (NDVI), Optimized Soil

© University of Pretoria

Adjusted Vegetation Index (OSAVI), Soil Adjusted Vegetation Index (SAVI) and Triangular Vegetation Index (TVI) were also evaluated. The final spectral indices used in this study are summarized in Table 13.

Table 13. The collection of spectral indices considered in this study.

Vegetation Indices	Equation	Source	Property
PSRI	$\frac{(\mathrm{Re}d - Green)}{RE2}$	Merzlyak et al. (1999)	Senescence-induced reflectance changes
NDVIRE1n	$\frac{(RE4 - RE1)}{(RE4 + RE1)}$	Fernández-Manso et al. (2016)	Sparse biomass
NDVIRE2n	$\frac{(\mathrm{R}E4 - RE2)}{(\mathrm{R}E4 + RE2)}$	Fernández-Manso et al. (2016)	Sparse biomass
NDVIRE3n	$\frac{(RE4 - RE3)}{(RE4 + RE3)}$	Fernández-Manso et al. (2016)	Sparse biomass
MSRRE	$\frac{(NIR / RE1) - 1}{\sqrt{(NIR / RE1) + 1}}$	Chen (1996)	Correction for leaf specular reflection
EVI	$2.5 * \frac{(NIR - \mathrm{Re}d)}{(NIR + 6 * \mathrm{Re}d - 7.5 * Blue) + 1}$	Miura et al. (2000)	Chlorophyll sensitive
GNDVI	$\frac{(NIR - Green)}{(NIR + Green)}$	Gitelson et al., (1996)	Chlorophyll sensitive
Soil Indices	**Equation**	**Source**	**Property**
BI	$\left(\frac{(\mathrm{Re}d^2 + Green^2 + Blue^2)}{3}\right)^{0.5}$	Madeira et al., (1997); Mandal (2016)	Average reflectance magnitude
CI	$\frac{(\mathrm{Re}d - Green)}{(\mathrm{Re}d + Green)}$	Madeira et al., (1997); Mandal (2016)	Soil Colour
HI	$\frac{(2 * \mathrm{Re}d - Green - Blue)}{(Green - Blue)}$	Madeira et al., (1997); Mandal (2016)	Primary Colours
RI	$\frac{\mathrm{Re}d^2}{(Blue * Green^3)}$	Bullard and White, (2002)	Hematite content
SI	$\frac{(Red - Blue)}{(Red + Blue)}$	Madeira et al., (1997); Mandal (2016)	Spectral slope

Note: Brightness Index (BI), Coloration Index (CI), Hue Index (HI), Redness Index (RI), Saturation Index (SI)

© University of Pretoria

4.2.5. Environmental Variables

Different datasets in Table 14 were used to describe the environmental variables needed to estimate nitrogen content. These included the slope, elevation, aspect, catchment area, topographic wetness index (TWI), precipitation, and temperature. The ASTER digital elevation model (DEM) with a 30 m spatial resolution was used to extract the terrain variables. This product was used because it is freely available and was closer to the 10 m spatial resolution of Sentinel-2 data. The ASTER DEM tiles were mosaicked and resampled to a 10 m resolution using a bilinear interpolation in the R software. Then, the DEM, slope, aspect, catchment area and TWI were derived. The JAXA Earth Observation Research Center precipitation and Landsat land surface temperature (LST) covering 7 years from 2013 to 2019 were used. This period was selected based on the continuity of the Landsat LST collection. These images were also resampled to a 10 m resolution. The environmental variables have shown to be valuable in previous studies for modeling nitrogen content (Chlingaryan et al., 2018; Wang et al., 2018).

Table 14. The list of selected environmental variables used in this study.

Environmental variables	Units	Source	Property
Slope (SLP)	Degrees	Wu et al., (2008)	Rise or fall of the land surface
Elevation (EL)	Meters	Wu et al., (2008)	Distance above sea level
Aspect (ASP)	Degrees	Wu et al., (2008)	Direction of terrain
Catchment area (CA)	Square Meters	Wu et al., (2008)	Flow accumulation
TWI	-	Sörensen et al., (2006)	Soil moisture
Precipitation (RAIN)	Millimeter/hour	Kubota et al., (2007)	Rainfall
LST	Kelvin	Ermida et al., (2020)	Temperature

© University of Pretoria

4.2.6. Machine learning regression models

4.2.6.1 Random Forest Regression

Random Forest is a bagging ensemble learning method (Breimann, 2001). This algorithm can be applied to both classification and regression problems. The principle of RF regression is to predict a continuous response variable using a bootstrapping method based on the classification and regression trees. Decision tree models are fitted to the data. Whereby, every tree is trained using different bootstrap samples from the training data, referred to as in-bag samples. The final model is generated by averaging the individual tree outputs (Breimann, 2001). Samples that are not used in the bootstrap are referred to as the out-of-bag samples; these can be used for model evaluation and variable importance (Pal, 2005). The RF is applied in this study because of its superior performance capabilities. RF can handle high dimensional data, requires relatively few tuning parameters, and processes non-linear data without overestimation (Hutengs and Vohland, 2016). The tuning parameters necessary to train the RF model (number of trees and features) were determined using Gridsearch method in Python; further details can be obtained in Lerman (1980). Variable importance for the RF algorithm was determined using the built-in Python variable importance measure for RF; readers are referred to Dangeti (2017) for further details on this procedure.

4.2.6.2 Gradient Boosting Regression

Gradient boosting is an ensemble-based decision tree machine learning method developed by Friedman (2001). This method can be adapted for both regression and classification problems. The purpose of gradient boosting is to improve the performance of weak learners to achieve over random guessing (Zemel and Pitassi, 2001). At each iteration, a new regression tree is trained to improve the loss function determined by the steepest gradient. This procedure reduces the model residuals along the gradient direction. The results of the individual regression trees are combined to give the final result (Friedman, 2001). The gradient boosting algorithm is applied in the present study because it can handle unbalanced data and it is robust to

© University of Pretoria

outliers (Wei et al., 2019). The parameters needed for gradient boosting are the number of trees, number of features for the best split, maximum depth, learning rate, and the minimum number of samples required at a leaf node. These were optimized using Gridsearch method. Variable importance for the GB algorithm was determined using the built-in Python variable importance measure for GB; readers are referred to Dangeti (2017) for further details on this procedure.

© University of Pretoria

4.2.6.3 Extreme Gradient Boosting Regression

The Extreme Gradient Boosting algorithm is part of the classification and regression ensemble gradient boosting machine algorithms. This model can be applied for both classification and regression problems (Chen and Guestrin, 2016). The XG uses additive training strategies, the first learning phase is fitted to the entire input dataset and the second phase is fitted to the residuals. This procedure enhances the performance of weak supervised learning. The fitting process is done repeatedly until the stopping criteria is achieved (Chen and Guestrin, 2016). The XG algorithm was applied because it overcomes problems with overfitting and has an optimized performance (Georganos et al., 2018). This algorithm requires a rigorous number of regularization parameters; these were determined using Gridsearch. Variable importance for the XG algorithm was determined using the built-in Python variable importance measure for XG; readers are referred to Dangeti (2017) for further details on this procedure.

4.2.6.4 Experiments

We investigated the effect of different feature variables for modeling nitrogen content in smallholder maize farms. The data was split into 70% training and 30% testing. Three RF, GB, and XG models with different combinations of variables summarized in Table 15 were implemented. The experiments consisted of: (1) raw bands, (2) raw bands + vegetation indices, (3) raw bands + soil indices, (4) raw bands + environmental variables, (5) raw bands + vegetation indices + soil indices + environmental variables, (6) raw bands + vegetation indices + soil indices, (7) raw bands + vegetation indices + environmental variables, (8) raw bands + soil indices + environmental variables and (9) raw bands + environmental variables + soil indices.

© University of Pretoria

Table 15. The different data configurations for the nine machine learning regression experiments.

Experiment	Number of variables	Data configuration
1	10	Raw bands
2	17	Raw bands and vegetation indices
3	15	Raw bands and soil indices
4	17	Raw bands and environmental variables
5	29	Raw bands, vegetation indices, soil indices and environmental variables
6	22	Raw bands, vegetation indices and soil indices
7	24	Raw bands, vegetation indices and environmental variables
8	22	Raw bands, soil indices and environmental variables
9	19	Raw bands, environmental variables and soil indices

4.2.7. Model Evaluation

The predictive performances of the RF, GB, and XG models were evaluated using validation indices. These included the fraction of predictions within a factor of two (FAC2), mean absolute error (MAE), mean bias error (MBE), root mean square error (RMSE), Pearson correlation (r), R^2 and Cross Validation (CV) as shown in Equations (1) – (7):

$$FAC2 : 0.5 \leq \frac{P_i}{O_i} \leq 2.0 \tag{1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | P_i - O_i | \tag{2}$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n} (P_i - O_i) \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(P_i - O_i\right)^2} \tag{4}$$

$$r = \frac{1}{(n-1)}\sum_{i=1}^{n} \left(\frac{P_i - \overline{P}}{\sigma_p}\right)\left(\frac{O_i - \overline{O}}{\sigma_o}\right) \tag{5}$$

© University of Pretoria

$$R^2 = \frac{\sum_{i=1}^{n}\left(P_i - \overline{O}_i\right)^2}{\sum_{i=1}^{n}\left(P_i - \overline{O}_i\right)^2} \quad (6)$$

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} R_i \quad (7)$$

where n represents the number of sample points, P_i represents the predicted soil nitrogen content, O_i represents the observed soil nitrogen content in site i respectively, and σ represents the standard deviation. The reader is directed to Carslaw and Ropkins (2012) for further information on these model evaluation matrices. The Taylor diagram was derived using the Openair package in R software (Taylor, 2001).

4.3. Results

4.3.1. Statistical analysis for soil nitrogen content measurements

Different vegetation indices (Figure 24) described in Section 4.2.4 were evaluated to retain indices that perform optimally for soil nitrogen content estimation. The RF, XG, and GB models were used to relate the vegetation indices to soil nitrogen. The PSRI, NDVIRE1n, EVI, NDVIRE2n, NDVIRE3n, GNDVI, and MSRRE were retained for further analysis. These vegetation indices were strongly related to the soil nitrogen content with an R^2 of 0.62 to 0.81. The soil nitrogen content measurements collected at the smallholder maize farms are characterized in Table 16. The nitrogen content was low for the farms, ranging from 0.014%-0.088%. The mean is lower than the standard deviation, which shows that the data are clustered closely around the mean. The mean is greater than the median, indicating a positively skewed distribution similar to the skewness value of 1.42 (Cumming and Calin-Jageman, 2016). The nitrogen content measurements were related to each of the variables in the regression experiments through a correlation matrix (Table 15). The MSRRE, NDVIRE1-3n, EVI, LST, and TWI had positive relationships with the soil nitrogen content. The remaining variables had a negative relationship with soil nitrogen. The PSRI, NDVIRE1-3n, EVI, CI, BI, SI, RI, B4-B12 where strongly related to the soil nitrogen content. However, the SLP, CA, ASP, DEM, TWI, LST and RAIN had a weak relationship with soil nitrogen. Moderate relationships where observed for the HI, B3 and soil nitrogen.

© University of Pretoria

Multicollinearity was identified between the vegetation indices, soil indices and raw bands. These variables were highly linearly related.

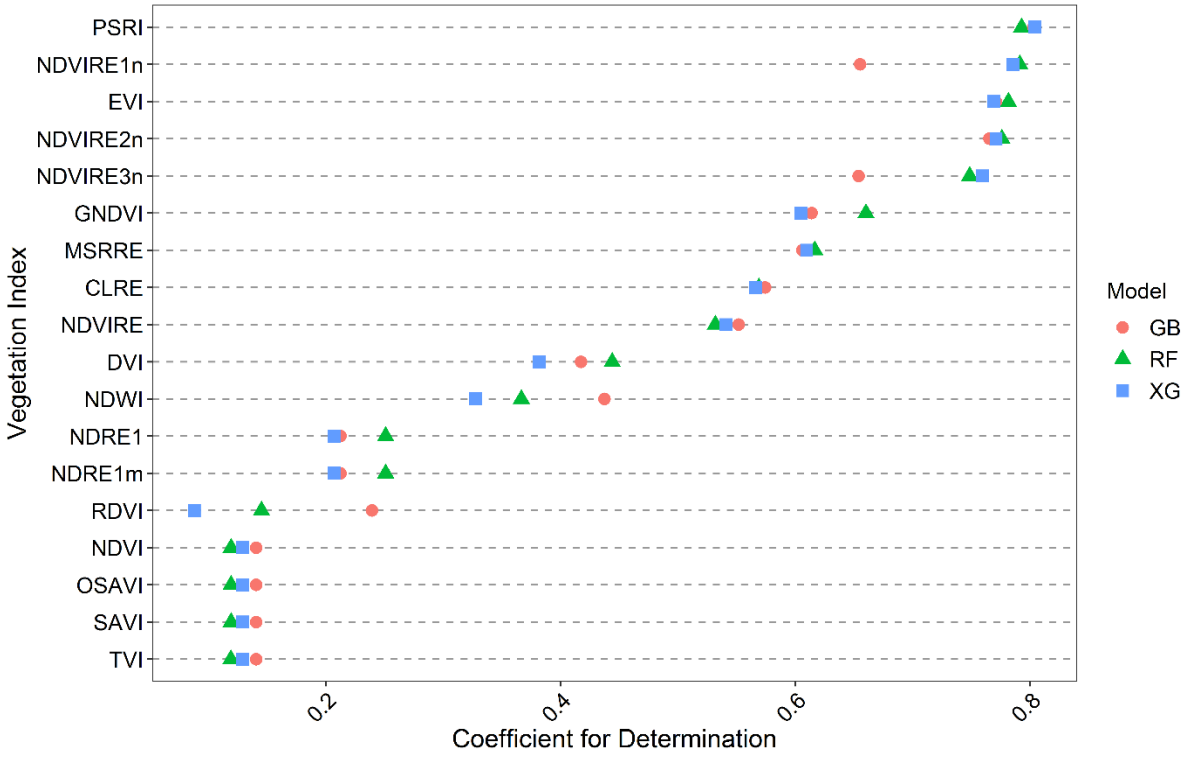

Figure 24. Vegetation indices evaluated for mapping soil nitrogen content.

© University of Pretoria

Table 16. Statistical analysis for the soil nitrogen content samples.

Soil Nitrogen							
a) Descriptive Statistics							
	Count	Minimum (%)	Maximum (%)	Mean (%)	Median (%)	Standard Deviation	Skewness
Nitrogen	105	0.014	0.088	0.033	0.025	0.019	1.424
b) Correlation							
Variable	R	Variable	r	Variable	R	Variable	r
MSRRE	0.579	CI	-0.713	B6	-0.899	TWI	0.081
PSRI	-0.793	BI	-0.798	B7	-0.894	DEM	-0.292
NDVIRE3n	0.835	SI	-0.804	B8	-0.883	ASP	-0.011
NDVIRE2n	0.840	RI	-0.748	B8A	-0.889	CA	-0.024
NDVIRE1n	0.737	B2	-0.061	B11	-0.883	SLP	-0.154
EVI	0.838	B3	-0.463	B12	-0.870		
GNDVI	-0,757	B4	-0.884	RAIN	-0.268		
HI	-0.591	B5	-0.898	LST	0.117		

© University of Pretoria

4.3.2. Model evaluation

The model performance statistics derived from the testing data (n = 32 samples) are summarized in Table 17. The best performing model from all experiments was the RF model for experiment 4. This model had the highest accuracy for soil nitrogen content estimation based on the lowest values for RMSE and MAE (RMSE = 0.0076% and MAE=0.0054%) and the highest r and R^2 (r = 0.95 and R^2 = 0.90). The predicted soil nitrogen values were smaller than the observed values based on the MBE (MBE=-0.0013%). Additionally, this model had a FAC2=1, indicating a perfect model similar to the FAC2 values for the other experiments. The least optimal performing model overall was the XG model for experiment 6 containing the raw bands, soil indices, and vegetation indices. This model had a high error rate based on the high RMSE and MAE (RMSE = 0.0090% and MAE=0.0063%) and the lowest r and R^2 (r = 0.9149 and R^2 = 0.8371). Furthermore, this model overestimated the soil nitrogen content based on the MBE (MBE = 0.0004%). The raw bands and environmental variables were sufficient to model soil nitrogen content with the RF (RF4) and GB (GB4) model. However, additional soil indices were needed in XG (XG8) for estimating soil nitrogen more accurately.

© University of Pretoria

Table 17. Model evaluation statistics for the three machine learning models in different experiments.

Model	FAC2	MAE (%)	MBE (%)	RMSE (%)	R	R^2	CV
RF1	0.9688	0.0067	0.0012	0.0086	0.9324	0.8694	0.7563
RF2	0.9688	0.0061	0.0000	0.0086	0.9302	0.8653	0.8079
RF3	0.9688	0.0071	0.0004	0.0092	0.9204	0.8472	0.7891
RF4	1.0000	0.0054	-0.0013	0.0076	0.9486	0.8998	0.6625
RF5	1.0000	0.0066	-0.0007	0.0086	0.9232	0.8523	0.7720
RF6	0.9688	0.0063	-0.0003	0.0089	0.9256	0.8568	0.6604
RF7	1.0000	0.0053	0.0000	0.0080	0.9433	0.8898	0.7104
RF8	1.0000	0.0059	0.0002	0.0083	0.9368	0.8775	0.6885
RF9	1.0000	0.0056	0.0000	0.0082	0.9395	0.8827	0.8645
GB1	0.9688	0.0070	0.0007	0.0092	0.9210	0.8482	0.5325
GB2	1.0000	0.0059	-0.0001	0.0084	0.9348	0.8739	0.6670
GB3	1.0000	0.0068	-0.0003	0.0092	0.9177	0.8423	0.6124
GB4	1.0000	0.0061	0.0001	0.0083	0.9369	0.8778	0.6354
GB5	1.0000	0.0061	0.0000	0.0084	0.9347	0.8737	0.7043
GB6	1.0000	0.0062	-0.0006	0.0087	0.9298	0.8645	0.7942
GB7	1.0000	0.0060	0.0002	0.0084	0.9336	0.8716	0.7734
GB8	0.9688	0.0064	-0.0009	0.0094	0.9172	0.8413	0.7556
GB9	1.0000	0.0058	0.0008	0.0083	0.9315	0.8676	0.7296
XG1	0.9688	0.0062	0.0003	0.0084	0.9311	0.8669	0.5671
XG2	0.9688	0.0057	0.0001	0.0085	0.9257	0.8569	0.8546
XG3	0.9688	0.0065	0.0005	0.0089	0.9227	0.8513	0.5970
XG4	1.0000	0.0062	0.0004	0.0088	0.9221	0.8502	0.5711
XG5	1.0000	0.0059	0.0004	0.0081	0.9352	0.8747	0.6121
XG6	0.9688	0.0063	0.0004	0.0090	0.9149	0.8371	0.6367
XG7	1.0000	0.0061	0.0007	0.0087	0.9234	0.8527	0.6453
XG8	1.0000	0.0054	0.0003	0.0077	0.9434	0.8900	0.5954
XG9	0.9688	0.0058	0.0002	0.0086	0.9300	0.8648	0.5839

Note: Random forest experiment number (RFx), gradient boosting experiment number (GBx), extreme gradient boosting experiment number (XGx) defined in Table 15.

© University of Pretoria

The Taylor diagram in Figure 25 was used to verify the model performance. All models had high correlation coefficients ranging from 0.91 to 0.95 and they plotted close to the observed reference value at the origin. Additionally, they had a similar performance shown by the clustering of points with the same location on the Taylor diagram (Taylor, 2001). However, the RF4 model had a slightly better performance compared to the other models based on the lowest standard deviation and root mean squared (RMS) error. The correlation coefficient was also high for this model, signifying a good fit between the observed and predicted values. The XG8 and GB4 models were the optimal performing models for the XG and GB models. They had a considerably lower standard deviation and RMS values but a high correlation. Additionally, the predicted values from these models were closer to the observed values.

Scatterplots were constructed for optimal performing RF, GB, and XG models to relate the observed and predicted soil nitrogen content in Figure 26. The data points are close to the diagonal line for all three models, indicating a good agreement between the observed and predicted values. The RF4 model had a slightly better performance R2 (R^2 = 0.90) than the other models and was statistically significant ($p=1.6x10^{-16}$) at a 95% confidence interval. The GB and XG models had similar R^2 values (R^2=0.88 and R^2=0.89). However, GB had a higher p-value of $3.1x10^{-15}$ in comparison to XG with a p-value of $6.3x10^{-16}$. Both models were statistically significant at a 95% confidence interval

© University of Pretoria

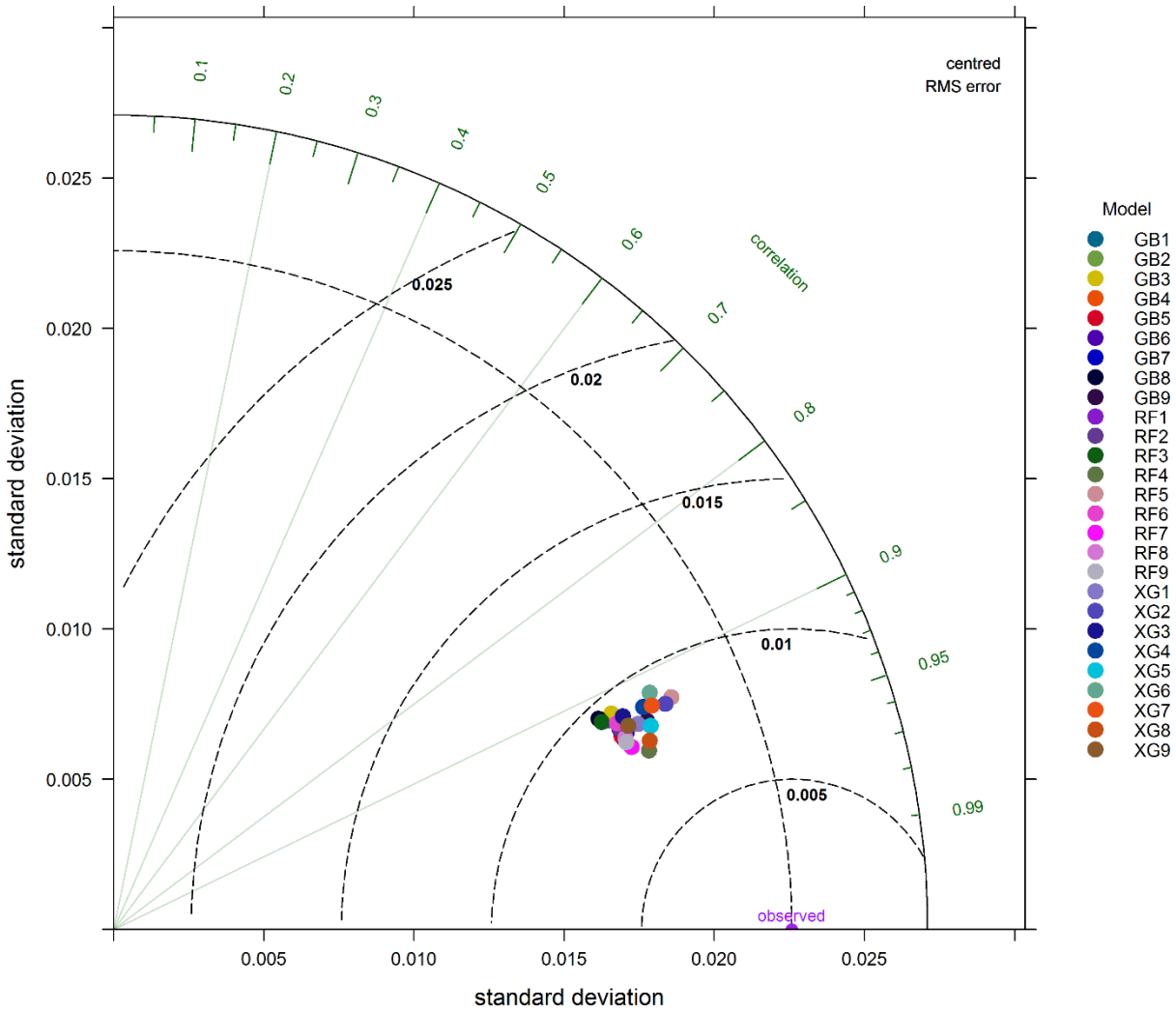

Figure 25. Taylor diagram for the nine experiments applying the three machine learning models.

© University of Pretoria

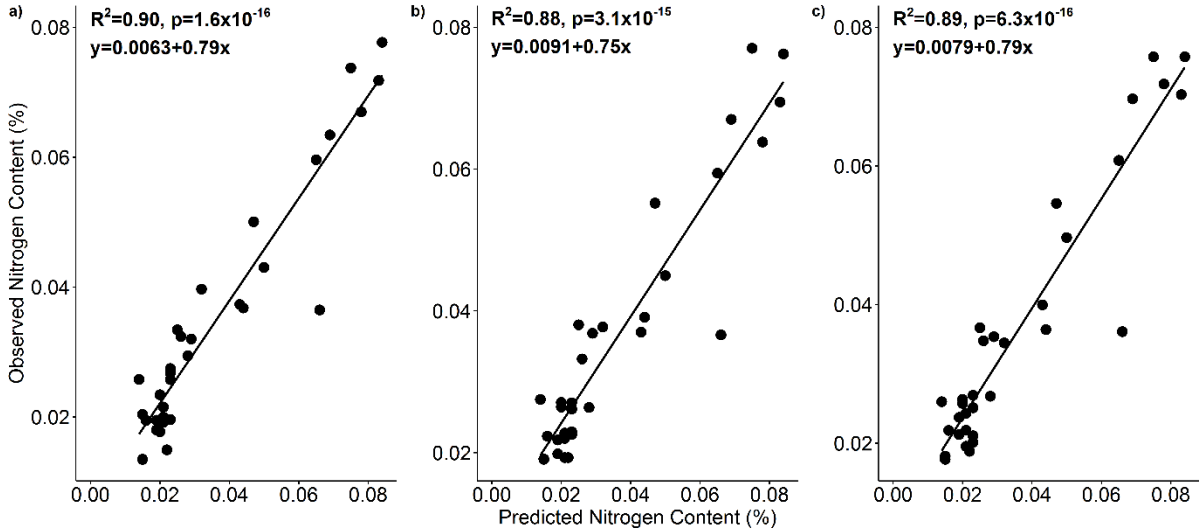

Figure 26. The relationship between observed soil nitrogen and predicted soil nitrogen where a) is RF4, b) is GB4 and c) is XG8.

4.3.3. Variable Importance

The importance of the predictor variables was determined for the most robust RF, GB, and XG models. All three models in Figure 27 varied in terms of predictor importance. The most important predictors for RF were B7, B5, B6, and B4. These were derived from experiment 4. The GB model ranked B4, B6, B5 and B12 highly from experiment 4. The B4 band was important in the XG model followed by the CI, and B5 in experiment 8. The RF model had a more even distribution of predictor importance in comparison to GB and XG where there is a greater contrast between the important (highest 4) and least important predictors (after the highest 4 predictors).

© University of Pretoria

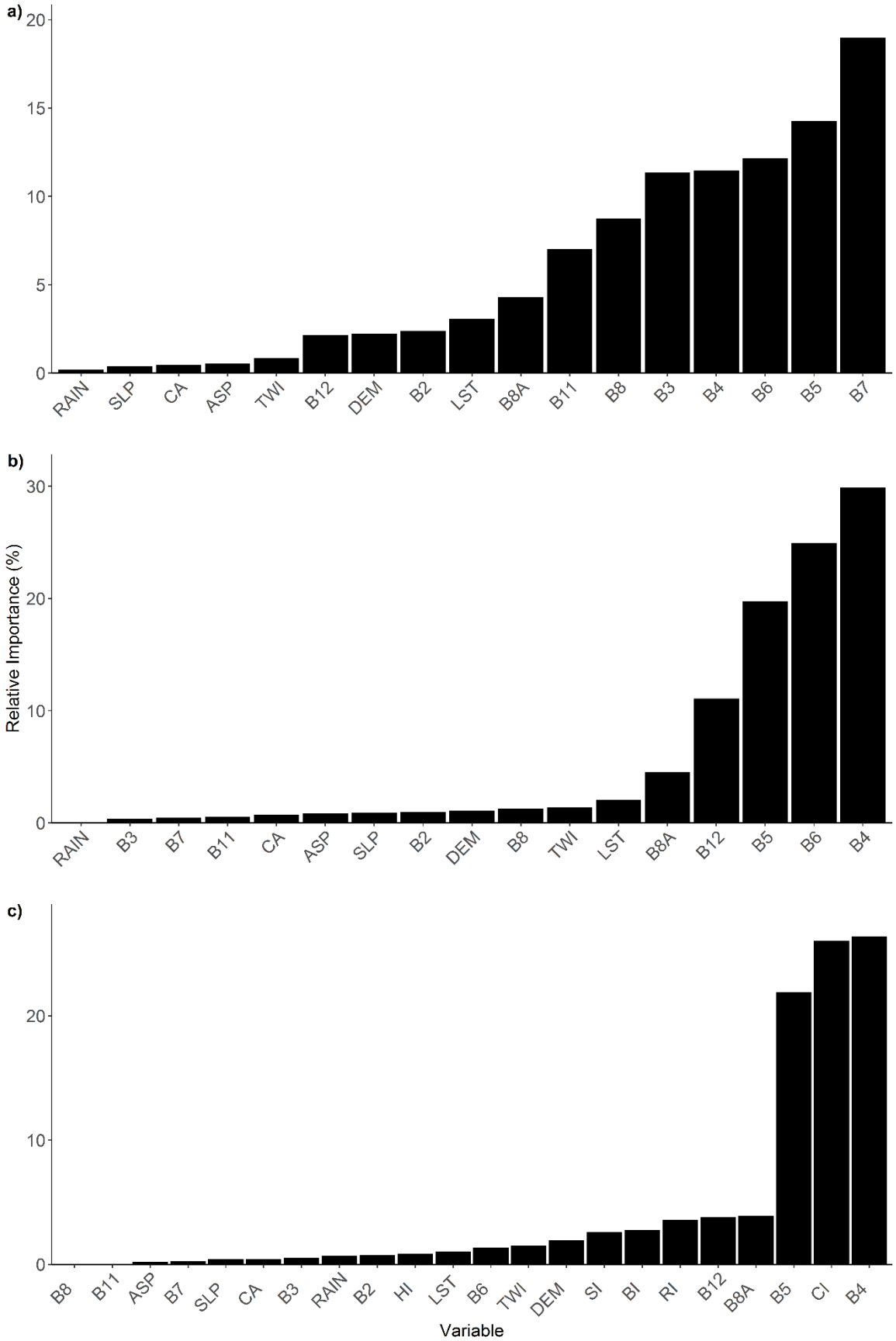

Figure 27. The ranking of variables for predicting soil nitrogen content with a) RF4, b) GB4 and c) XG8 algorithms.

© University of Pretoria

4.3.4. Mapping soil nitrogen content for smallholder maize farms

The spatial distribution of soil nitrogen was mapped in Figure 28 - Figure 30. There were differences in the spatial distribution of nitrogen for the smallholder maize farms. The smallholder farms in the central and southeastern part of the study area had a lower nitrogen content. However, the farms in the southern part of the study area had a higher nitrogen content. The maps generated by the RF and XG algorithms were similar, but, GB overestimated the nitrogen content.

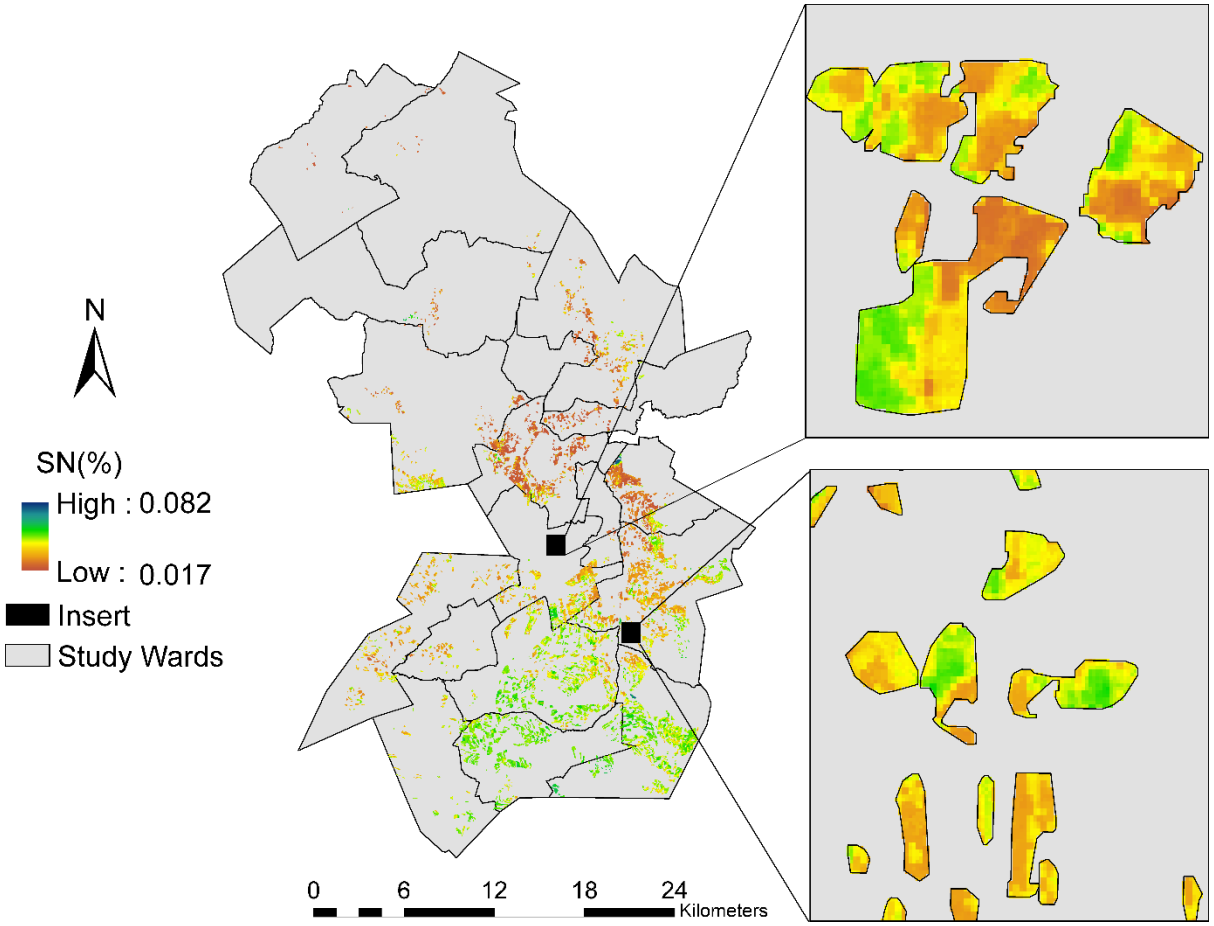

Figure 28. The spatial distribution of soil nitrogen mapped with the random forest model for experiment 4.

© University of Pretoria

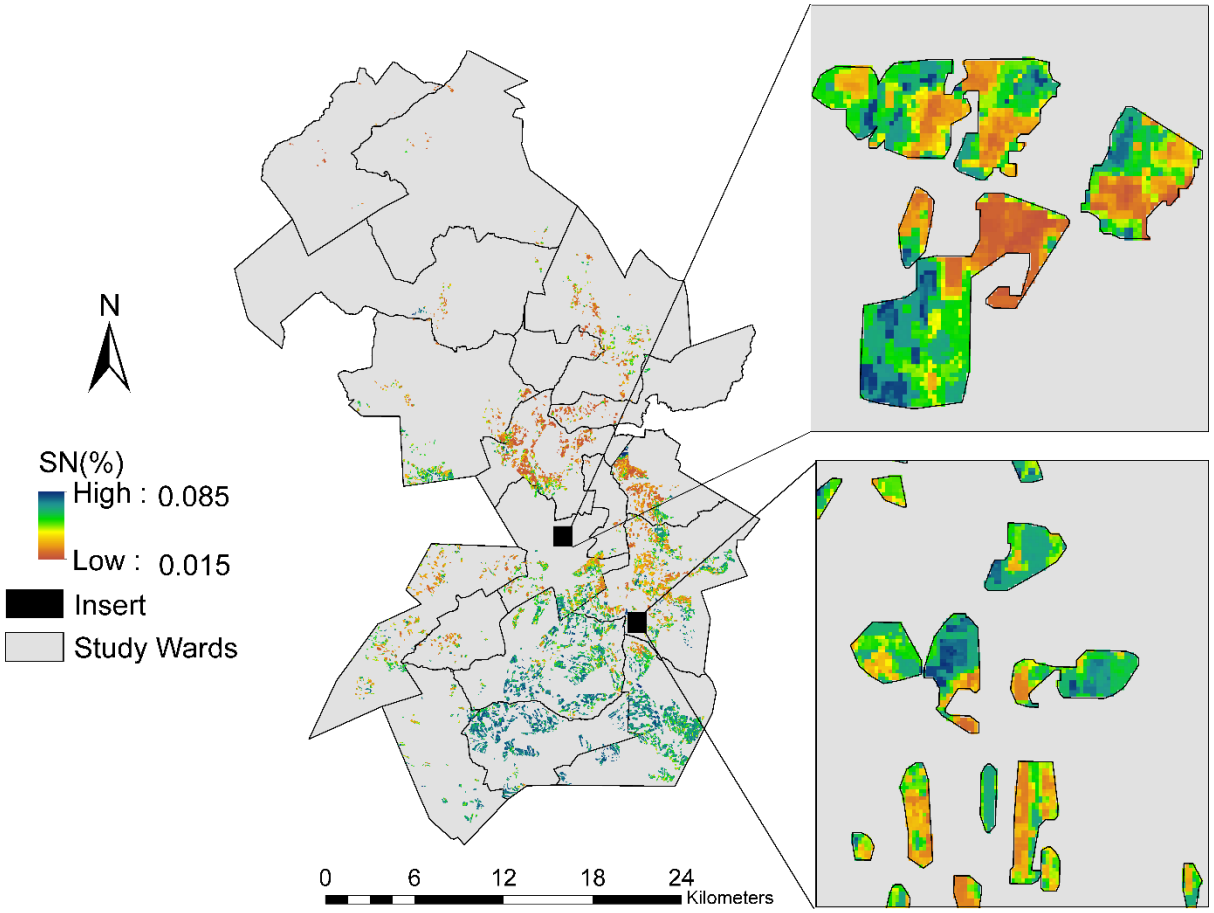

Figure 29. The spatial distribution of soil nitrogen mapped with the gradient boosting model for experiment 4.

© University of Pretoria

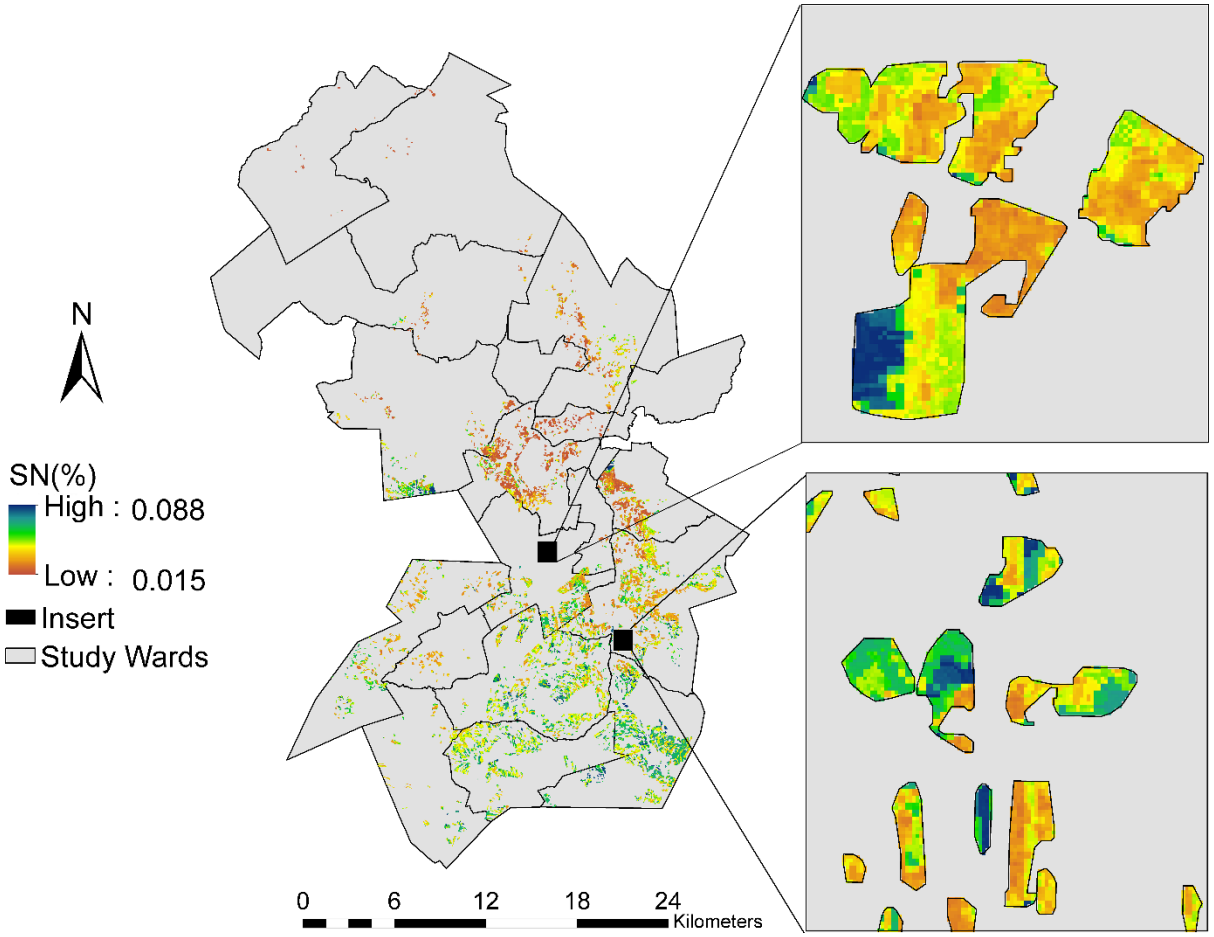

Figure 30. Distribution map of soil nitrogen obtained using the XG model is for experiment 8.

4.4. Discussion

This study assessed the applicability of Sentinel-2 bands, derived soil and vegetation indices and environmental data for predicting soil nitrogen in the smallholder maize farms of Makhuduthamaga district. Descriptive statistics were generated for the collected soil nitrogen content samples. Experiments were used to evaluate the performance of RF, GB, and XG machine learning algorithms in a regression format. The variable importance measure for each algorithm was used to determine which predictors had the most influence. The best performing algorithms in each experiment were then used for mapping the nitrogen content. The results showed that the Sentinel-2 bands and environmental variables have a superior performance when estimating the soil nitrogen content in comparison to the vegetation indices and soil indices.

© University of Pretoria

Findings from the descriptive statistics indicate that nitrogen content is low (0.014%-0.088%) for the smallholder maize farms. This is expected because the smallholder farms within the study area rarely apply nitrogen fertilization and a small proportion of the farmers use cow manure as fertilizer. For example, Nyamugara et al. (2005) conducted experiments for three years and found that the combination of cow manure and nitrogen fertilizers in smallholder maize farms in Zimbabwe improves soil nitrogen content which increases maize yield. Furthermore, data exploration in our study revealed that multicollinearity was present when relating soil nitrogen content to the different predictor variables. The presence of multicollinearity implies that the application of multiple linear regression with these variables to predict the soil nitrogen content would be unreliable (Mansfield and Helms, 1982). Multicollinearity introduces large variances in the least squares estimators (regression coefficients), lowers the quality of the resulting parameter estimates, and the variables have a low information content (Farrar and Glauber, 1967). The main advantage of the machine learning techniques, applied in the present study, is that they are less prone to multicollinearity problems. For example, Jaya et al. (2020) found that the artificial neural network model had a lower bias, mean squared error and minimized residuals in comparison to a multiple linear regression model when multicollinearity was present. Additionally, Farrell et al. (2019) study observed that multicollinearity removal and correlation removal did not reduce the performance of RF and support vector machine substantially. The robustness of machine learning could be due to the adaptive learning process used by the models which reduces errors (Belgiu and Drăguţ, 2016; Cooner et al., 2016). For example, RF uses bagging, XG uses additive training strategies, and GB reduces the model residuals along the gradient direction, which minimizes the multicollinearity problem.

Three predictive models were evaluated. Findings show that the RF model performs better than the GB and XG models when estimating soil nitrogen at smallholder maize farms in our study area. These results are similar to other studies which show the high capacity of RF in mapping soil nitrogen content (Jeong et al., 2017; Sorenson et al., 2017; Xu et al., 2018; Zhang et al., 2019; Deng et al. 2020; López-Calderón et al. 2020). Furthermore, the findings suggest that the XG model needs more input variables to model soil nitrogen content in comparison to GB and

© University of Pretoria

RF. This can be attributed to the implementation of the models, the XG algorithm is sensitive to outliers because the individual learners are in series format and RF is not sensitive to outliers because it is a parallel implementation of multiple decision trees (Li et al., 2019). In terms of variability, this study found an R^2 of 0.87-0.90, RMSE of 0.0086%-0.0092% and CV of 0.66-0.81 with RF which is the most robust model. Our results are similar to López-Calderón et al. (2020) that found an R^2 of 0.77 and a mean square error of 0.15 % when predicting soil total nitrogen content applying RF for forage maize with UAV imagery. Additionally, Sorenson et al., (2017) used field reflectance spectroscopy for estimating soil nitrogen content and reported a cross-validation RMSE of 0.62% and R^2 of 0.78 with RF for reclaimed soils. Furthermore, Deng et al. (2020) found a cross validation R^2=0.65 and RMSE=0.43 g kg^{-1} with RF applied on MODIS data when estimating soil nitrogen content for croplands. Contrary to our findings, Xu et al. (2018) reported an adjusted R^2 of 0.49 and RMSE of 125.71 mg kg^{-1} with Landsat 8 data applying RF to predict soil nitrogen at smallholder farmlands planting different crops. Jeong et al., (2017) observed an R^2=0.552 and RMSE=1.131 mg g^{-1} when applying RF soil nitrogen content estimation in a complex terrain with Landsat TM data. These differences in findings can be induced by the input variables or other factors such as whether the soil is completely bare or has plant coverage which can influence the predicted soil nitrogen content. For example, the study by Beguin et al. (2017) found that the input predictors affect the predictive capacity of models predicting soil properties. Other studies such as Zhang et al., (2019) observed different performance for the digital soil map generated in a vegetated condition (R^2=0.67) and completely bare soil condition (R^2=0.80) with RF.

Variable importance was done to determine the most important predictors for estimating soil nitrogen content at smallholder maize farms. The results showed that the Sentinel-2 bands have an advantage when estimating soil nitrogen content. However, environmental variables had a lower ranking and additional soil indices were necessary in the XG model. These findings are similar to other studies that found that spectral bands are more important than environmental variables (Zhang et al., 2019; Forkuor et al., 2017, Zhou et al., 2019). However, some studies showed contrasting results and the environmental variables had the highest ranking (Wang et al. 2018; Zhou et al., 2020). The differences in findings are attributed to variations in the model

© University of Pretoria

input variables in these studies. For example, most of these studies have used Landsat optical data for mapping soil nitrogen content which does not have the RE bands that Sentinel-2 has, which the current study incorporated. Additionally, the presence of maize crops within the smallholder farms in the current study could have contributed to the higher importance of the red-edge bands. These bands are sensitive to variations in chlorophyll content, differences in the leaf structure and plant biomass (Miura et al. 2000; Fernández-Manso et al., 2016). The radiation from the red-edge penetrates deeper into the crop canopy and leaves in comparison to visible light due to lower chlorophyll absorption in the visible region (Li et al., 2014). Xu et al. (2018) also found that red-edge spectral bands are important when estimating soil total nitrogen in smallholder farms that have different crops planted**.** These studies prove that red-edge bands have a high capability to estimate total nitrogen content accurately in smallholder farms that have crop cover. The high importance of the CI and RI amongst the soil indices was expected within the study area because most of the soils are red soils which have a high iron oxide content possibly related to haematite which the RI is sensitive to (Bullard and White, 2002). The most important predictors were LST, DEM, and TWI for the environmental variables. The LST affects the spatial distribution of soil nitrogen through its effect on soil temperature, thereby affecting the process of nitrogen mineralization (Knoepp and Swank, 2002). The DEM is important because elevation plays a role in the microclimate, runoff, evaporation and transpiration (Baxter and Olivier, 2005). The TWI is an indicator of soil moisture distribution (Sörensen et al., 2006). Soil moisture conditions, in addition of course to soil nutrients, are determinants of crop vigor and development. The distinction between highly ranked predictors and low ranking predictors in the GB and XG models shows that further exploration of the influence of the predictors on model performance can be done for both models for model optimization.

The spatial distribution of soil nitrogen was mapped. The resulting spatial maps produced from the three algorithms were similar. This finding proved the high capability of machine learning to estimate soil nitrogen content in smallholder maize farms. The soil nitrogen maps generated in this study can be used as a tool to guide decision making for smallholder farms. Recommendations by crop consultants, extension services, and fertilizer dealers can also benefit from using nitrogen content

© University of Pretoria

maps. Government initiatives providing farmers with agricultural inputs can use such maps to determine the soil nitrogen content at the farms and the proportion of fertilizer to use because different fertilizer quantities affect maize yield differently as shown by Nyamugara et al. (2005). Improved levels of soil nitrogen content at smallholder farms will increase maize yields, thereby, improving food security (Sinclair and Muchow, 1995; Otto, 2016; Chlingaryan et al., 2018). This application contributes to the Sustainable Development Goals (SDG) number 2 (Zero Hunger), target 2.4 and indicator 2.4.1, which are concerned with mitigating factors that affect agricultural production, ensuring sustainable agriculture and increasing the proportion of agricultural area under production (Richard, 2015).

The main limitation of this study is that a small number of farms were visited for field data collection due to the high cost for laboratory processing of samples and fieldwork. This study recommends further exploration of Sentinel-1 and Sentinel-2 data for estimating soil nitrogen in smallholder farms (Zhang et al., 2019; Zhou et al., 2019; Zhou et al., 2020). Studies focusing on smallholder farms are lacking especially in an African context and these farms are important for food security and rural livelihoods (Shi and Tao, 2014; Fischer and Hajdu, 2015). Training programs are recommended for the smallholder farms to improve the awareness of farmers on chemical fertilization. For example, nitrogen is essential when the crop is actively growing, but nitrogen application before that time can lead to losses through leaching or subsurface flow (Poffenbarger et al., 2018). Other more cost-effective alternatives to nitrogen fertilizers such as leguminous trees and shrubs grown with maize are recommended for smallholder farms in resource poor areas. These will provide nitrogen-rich residues that increase soil fertility (FAO, 2016).

4.5. Conclusion

This study was aimed at assessing Sentinel-2 bands, derived soil and vegetation indices and environmental variables for predicting soil nitrogen in smallholder maize farms applying machine learning regression. Different predictor variables were related to soil nitrogen content. The red, red-edge and short-wave infrared bands were strongly related to soil nitrogen with correlations of 0.89-0.90. The machine learning

© University of Pretoria

models applied in this study (RF, GB, and XG) were suitable for the data because multicollinearity was present between the predictors, which these models dealt with effectively. Model evaluation results show that machine learning models have a high predictive capacity in estimating soil nitrogen (R^2=0.84-0.90 and RMSE=0.0076-0.0094%) in smallholder farms. Variable importance revealed that the Sentinel-2 bands, particularly the red and red-edge bands are fundamental for modeling soil nitrogen in all three models. The soil nitrogen maps generated in this study can be used as a tool to guide decision making for smallholder farms. Recommendations by governments, extension services and fertilizer dealers can also benefit from using such maps. These maps are useful to establish nitrogen management plans in the smallholder farms, which will increase maize yields, thereby, improving food security.

© University of Pretoria

Chapter 5

Early Season Spatial Estimation of Smallholder Maize Yield based on Machine Learning

Abstract

Food security is an issue of global concern; this has mandated the monitoring of agricultural systems using cost effective techniques such as remote sensing. Smallholder maize farms are dominant in Africa, they produce 80% of the maize in the region. These farmers are faced with economic and environmental issues that limit their productivity. The utility of Sentinel-1 and minimal field collected soil samples are investigated for predicting smallholder maize yield early in the season. Two machine learning models were tested—random forest (RF) and extreme gradient boosting (XG). The findings suggest that maize yield can be accurately predicted from two months before harvest. However, the accuracies of the models were low (R^2 of 0.2-0.41), this was expected for smallholder farms. These farms are fragmented and usually have non-homogeneous planting patterns. The VV_December, soil nitrogen content, VH_April and VH_March were identified as important variables for estimating maize yield. The yield maps generated in this study can be used to contribute to the Sustainable Development Goals (SDG) number 2 (Zero Hunger). Thereby, ensuring food security and improved policy implementation in rural communities.

Keywords: food security, maize, yield estimation, Sentinel-1

© University of Pretoria

5.1. Introduction

The Food and Agriculture Organization (FAO) state of the World Series has identified food security as an issue of global concern of the 21st century (FAO, 2018). Furthermore, the Sustainable Development Goal (SDG) number 2: End hunger, achieve food security, improve nutrition, and promote sustainable agriculture, aims to address this global crisis (SDG, 2019). Maize is an important cereal crop worldwide for different uses such as human consumption, animal feeding and industrial production (Orhun, 2013; Ranum et al., 2014). However, maize production is expected to decrease worldwide under conventional management due to climate variability (Parry et al., 2004; Zougmoré et al., 2018). Smallholder maize farms that contribute 80% of the maize produced in the rural communities of Africa are also threatened by this phenomenon (FAO, 2016). However, the increasing demand for maize products in rural areas where it is a primary food source has contributed to food insecurity in these communities (Santpoort, 2020). Therefore, frameworks to predict smallholder maize yields before harvest are imperative. These frameworks will provide an indication of the expected maize yield and aid in planning for maize shortages and surpluses to ensure food security, especially in rural communities.

Satellite data is an indispensible tool for maize yield estimation (Prasad et al., 2006; Sibley et al., 2014; Yao et al., 2015). However, the high density of clouds during the rainy season when smallholder maize is cultivated limits satellite data from optical sensors. Thus, this study uses synthetic aperture radar data (SAR), which can penetrate cloud cover to overcome this constraint. Few studies have been done on the use of SAR for crop monitoring, mapping and biophysical parameter estimation (Wu et al., 2010; Satalino et al., 2013; McNairn and Shang, 2016; Zhou et al., 2017). Additionally, the lack of freely available SAR images and complexity of the data have limited its development in crop yield estimation (Torbick et al., 2017). Sentinel-1 is equipped with SAR sensors, which have a shorter revisit time of 6 days and an improved spatial resolution (5 m x 20 m) in the interferometric wide-swath mode that has a broad potential in estimating crop yield (Torres et al., 2012).

© University of Pretoria

Mainly two types of approaches have been used to predict maize yields, these comprise of crop simulation models and statistical models. Crop simulation models simulate crop yield using meteorological data and farm management data (Whisler et al., 1986; Lal et al., 1993). Different crop models have been developed for maize yield predictions these include but are not limited to: DSSAT, APSIM, WOFOST and Hybrid maize (Yang et al., 2004; Liu et al., 2011; Archontoulis et al., 2014; Cheng et al., 2016). The drawback of using these models on a regional scale is the sparse distribution of weather stations especially in Africa (Van Wart et al., 2013). Additionally, crop simulation models provide results that are point based (Brisson et al., 1992). Furthermore, most of the smallholder farmers in Africa rarely record their farm management information (Gommes et al., 1998). An alternative approach is the statistical approach, this method relies on developing an empirical relationship between weather data and historical yield records to develop future forecasts (Matsumura et al, 2015). Linear regression approaches have been used widely in this approach for maize yield prediction (Rojas et al., 2007; Golam et al., 2011; Matsumura et al., 2015). The disadvantage of these models is that they often do not explain the soil–plant–atmospheric interactions (Chivasa et al., 2017). Additionally, their spatial generalization to other areas is low and they can be easily affected by multicollinearity (Farrar and Glauber, 1967; Mkhabela et al., 2005).

Advanced algorithms such as machine learning can be applied for estimating crop yields. Machine learning algorithms are widely used because they can learn from limited data and reduce errors through an adaptive learning process (Belgiu and Drăguţ, 2016; Cooner et al., 2016). Previous studies have reported varying model performances when different machine learning models were compared. For example, Kang et al. (2020) found that the extreme gradient boosting (XG) algorithm was better than the lasso, support vector machine (SVM), random forest (RF), long-short term memory and convolutional neural network algorithms when estimating maize yield. Chen et al., (2021) reported that Cubist was slightly better than RF, SVM, and XG for predicting maize yield. Furthermore, Meng et al., (2021) observed that RF and adaptive boosting performed better than linear regression (LR), K-nearest neighbor (KNN), SVM and Gaussian process (GP) regression for estimating maize yield. This necessitates the evaluation of different machine learning algorithms to find the suitable

© University of Pretoria

techniques for estimating the smallholder maize yields within the study area. This current study uses the RF and GP machine learning algorithms maize yield forecasting because they can deal with noisy, high-dimensional and non-linear data (Izquierdo-Verdiguier, 2014; Li, 2016). The research questions addressed in this study are: 1) Which time window is optimal for maize yield estimation? 2) Which features are important for maize yield estimation? and 3) What is the spatial pattern of smallholder yield?

5.2. Materials and methods

5.2.1 Study Area

Maize and soil samples were collected in Makhuduthamaga district (Figure 31) located in Limpopo province of South Africa. This district is dominated by smallholder maize farms that depend on their produce for sustenance, thus, it was selected as a case study (SDM, 2019). The area experiences summer rainfall and the mean annual rainfall is 536 mm. The minimum mean annual temperature can reach 7°C and the maximum mean annual temperature can reach 35°C according to the automatic weather stations of the Agricultural Research Council. The topography is undulating with rock habitats such as rock outcrops, rocky ridges and rocky refugia (Siebert et al., 2003). The dominant soil types are—haplic acrisols, ferric luvisols and lithic leptosols (Jones and Thornton, 2015).

© University of Pretoria

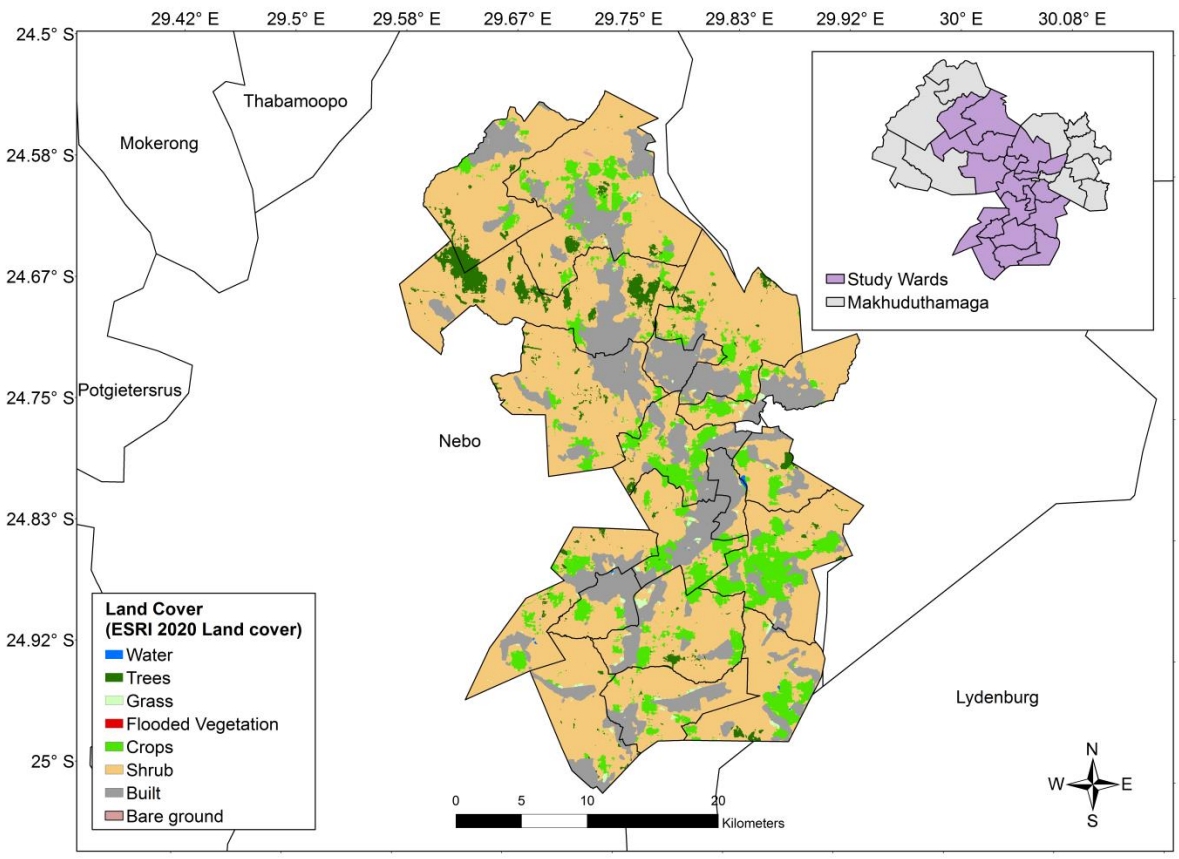

Figure 31. Dominant land cover classes within the study wards in Makhuduthamaga.

5.2.2 Satellite data: Sentinel-1

Sentinel-1 imagery were acquired from the Google Earth Engine 'COPERNICUS/S1_GRD' image collection. This collection consists of the Level-1 Ground Range Detection (GRD) scenes (GEE, 2021). The data is preprocessed using the Sentinel-1 toolbox to generate the backscatter coefficient using key preprocessing steps before uploading to GEE. These steps consist of applying the obit file to update the orbit metadata. The GRD border noise removal is done to remove low intensity noise and invalid data. Thermal noise removal takes place to remove additive noise. Then, radiometric calibration is done for the computation of backscatter intensity. Lastly, the terrain correction is done to remove the geometric distortions caused by topography (Filipponi, 2019). Both the Vertical transmit and vertical receive (VV) and Vertical transmit and horizontal receive (VH) polarizations were used in this study as described in Table 18. The Sentinel-1 data covered the maize cropping season (December 2018 - 30 June 2019).

© University of Pretoria

Table 18. The characteristics of the Sentinel-1 data.

Polarizations	Bandwidth (nm)	Spatial Resolution (m)
Vertical transmit and vertical receive (VV)	-	10
Vertical transmit and horizontal receive (VH)	-	10

5.2.3 Maize yield data

The maize yield samples were collected in June for 2018/19. The yield was then determined in tons per hectare using the method by Sapkota et al., (2016) based on the grain weight. The parameters measured at each point per field were the number of kernel rows per ear, number of ears per square meter (determined from a 1 meter quadrant), kernel weight in grams and the 1000 grain weight. The Global Positioning System (GPS) locations of 104 maize samples were captured. The phenological calendar for maize in the study area is depicted in Figure 32. Maize is planted during December and January. The maize then grows for four months between February to May. Harvesting takes place in June and no maize is present in the smallholder farms during July-November. The maize from the smallholder farms in the study area is rain-fed.

© University of Pretoria

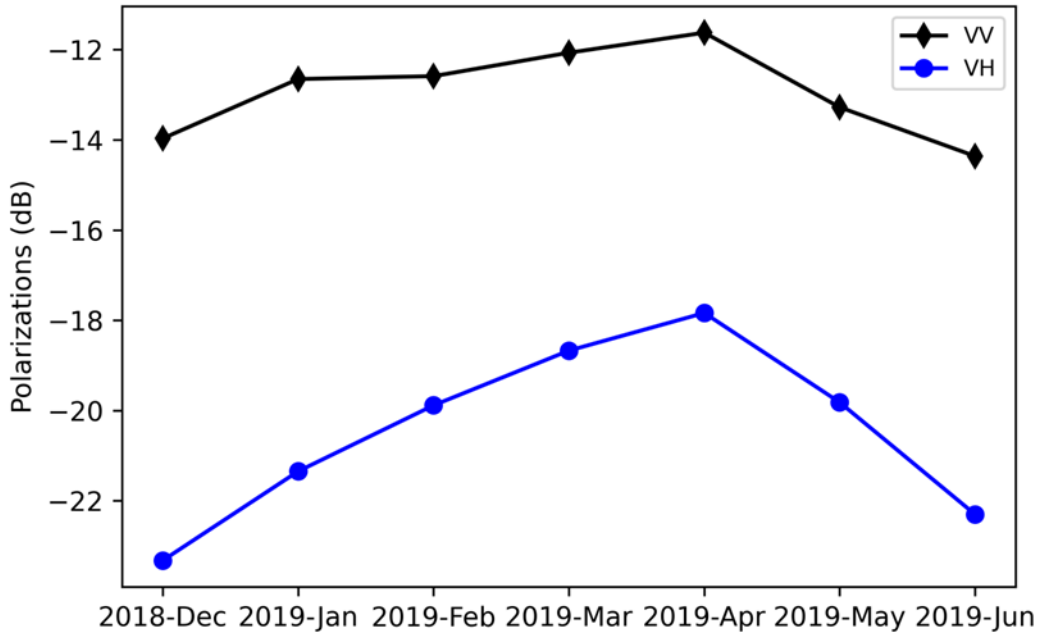

Figure 32. The time series evolution of the VV and VV polarizations during the planting season.

5.2.4 Soil Data

Soil samples were collected in the topsoil of smallholder maize farms in May 2019. There were 105 samples in total that were sent to the analytical laboratory of the Agricultural Research Council of South Africa to determine the nitrogen content and pH. Incorporating soil nutrient information is necessary because a lack of the appropriate amount and form of crop nutrients is a major crop productivity constraint in the third world (Hussain et al., 2006).

5.2.5 Machine learning regression models

Both RF and XG machine learning regression models were applied for maize yield estimation. The RF classifier is an ensemble learning algorithm consisting of a combination of classifiers. Each pixel is assigned to a specific class using a majority voting system. The RF algorithm trains each tree using an independently drawn sample of the original data using bootstrapping or bagging and determines the number of features to be used at each node by evaluating a random vector (Breiman, 2001).

© University of Pretoria

This algorithm was used because it is insensitive to noise or overtraining. The RF model has other advantages such as determining the variable importance and is less computationally extensive (Rodriguez-Galiano et al., 2012). This model was implemented in Python and feature importance was determined using the built-in function for RF.

The XG model is a scalable tree boosting algorithm proposed by Chen et al., (2015). The framework for this model was developed from the gradient tree boosting system by Friedman et al., (2000) and Friedman (2001). This algorithm is effective and uses additive training process to develop strong learners. The model is fitted to the entire datasets and adjusted using the residuals. The additional regularization term helps to smooth the final learnt weights to avoid over-fitting (Chen and Guestrin, 2016). The algorithm can learn the desired model from complex datasets, supports parallel computing, which enables it to reduce computational time. This model was implemented in Python and feature importance was determined using the built-in function for XG.

5.2.6 Metrics for model evaluation

Model evaluation for the predictive performance of the machine learning algorithms was done using common model evaluation metrics. This included the mean absolute error (MAE), mean bias error (MBE), root mean squared error (RMSE) and coefficient for determination (R^2) as shown in Equation (1) – (3):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | P_i - O_i | \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(P_i - O_i\right)^2} \quad (2)$$

$$R^2 = \frac{\sum_{i=1}^{n}\left(P_i - \overline{O}_i\right)^2}{\sum_{i=1}^{n}\left(P_i - \overline{O}_i\right)^2} \quad (3)$$

© University of Pretoria

where n represents the number of sample points, P_i represents the predicted soil nitrogen content and O_i represents the observed soil nitrogen content in site i respectively.

5.2.7 Experiments

The effectiveness of Sentinel-1 and other ancillary soil data were evaluated for estimating smallholder maize yield. These datasets were tested to identify the earliest window when maize yields can be predicted accurately. Five windows summarized in Table 19 were considered for experimentation: December-January, December-February, December-March, December-April, December-May and Dec-June. The data was split 90% (93 samples) for training and 10% (11 samples) for validation.

Table 19. The configurations of the six datasets.

Dataset	Number of predictors	Months	Data type
1	6	Dec-Jan	VV, VH, pH, Nitrogen
2	8	Dec-Feb	VV, VH, pH, Nitrogen
3	10	Dec-Mar	VV, VH, pH, Nitrogen
4	12	Dec-Apr	VV, VH, pH, Nitrogen
5	14	Dec-May	VV, VH, pH, Nitrogen
6	16	Dec-June	VV, VH, pH, Nitrogen

5.3. Results

5.3.1 Identifying the Earliest Time Window to Predict Smallholder Maize Yield

Statistical analysis were done to determine the earliest time window when smallholder maize yield can be predicted with RF and XG machine learning regression models. A multidimensional time series of Sentinel-1 images and soil information were related to maize yield for this purpose. The optimal time window for yield modelling was the same for both models. Based on Figure 33 to Figure 35, maize yields can be predicted accurately two months before harvest. The ideal time window is in April with a data cube including data from December (sowing period). The R-squared value is high for the Dec-April dataset, indicating a stronger relationship between maize yield and the

© University of Pretoria

predictor variables. Additionally, XG (R^2: 0.41) performed slightly better than RF (R^2: 0.30) based on the higher R-squared value. Both the RMSE and MAE were lower for the XG (RMSE: 0.41 t/ha and MAE: 0.46 t/ha) model in comparison to RF (RMSE: 0.48 t/ha and MAE: 0.51 t/ha) during this time. The model accuracies decreased after the peak in maize growth for both models.

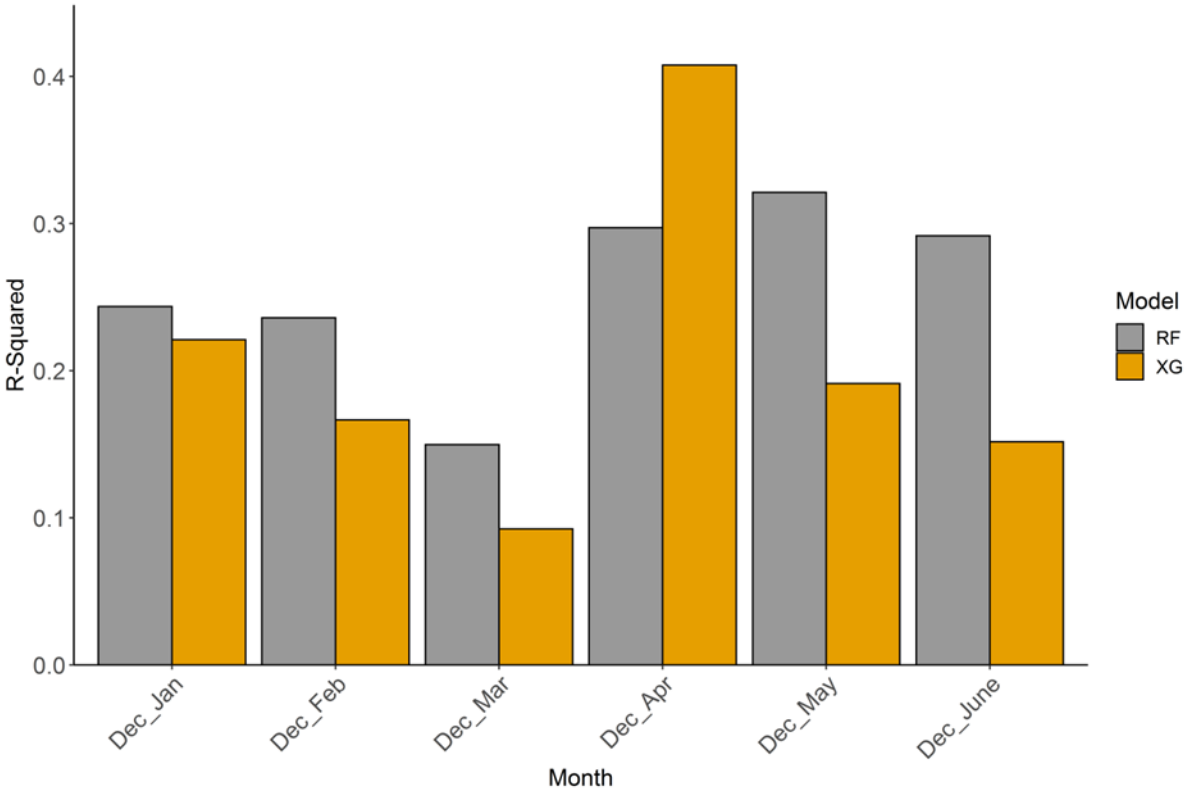

Figure 33. The R-Squared values for the two machine learning models.

© University of Pretoria

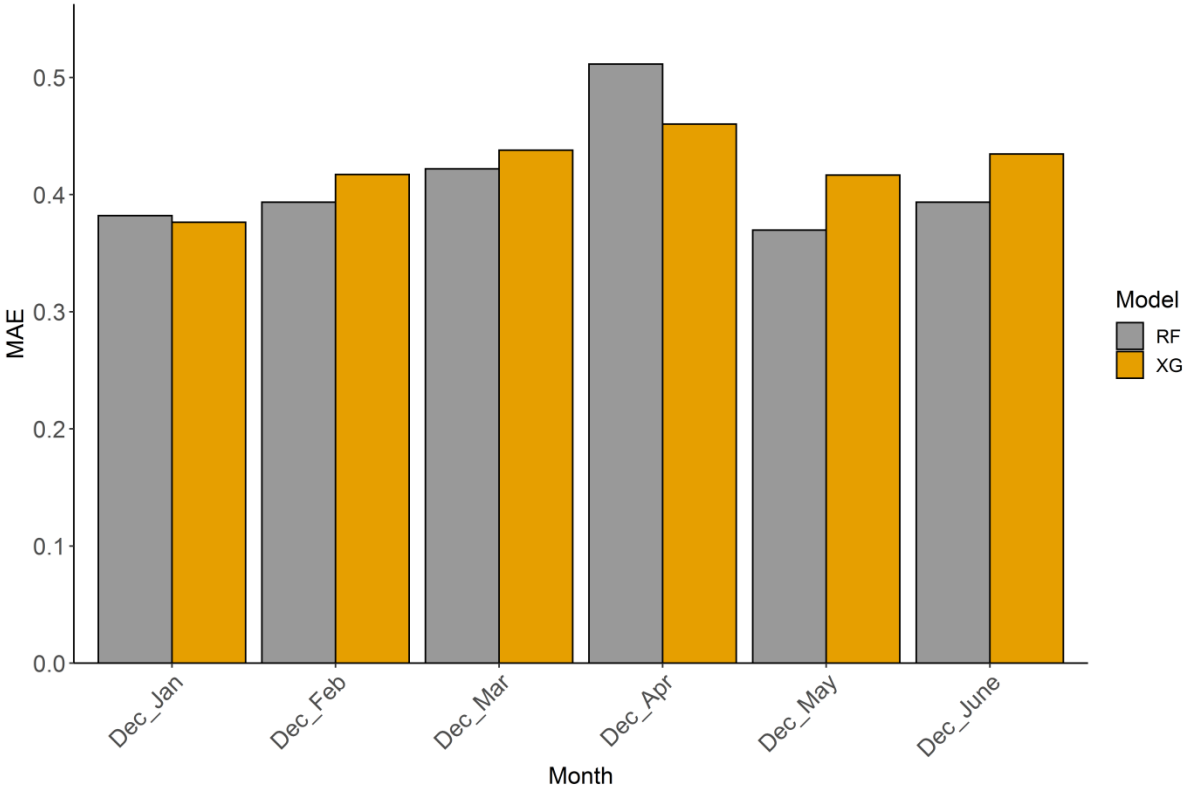

Figure 34. The MAE for the RF and XG machine learning models.

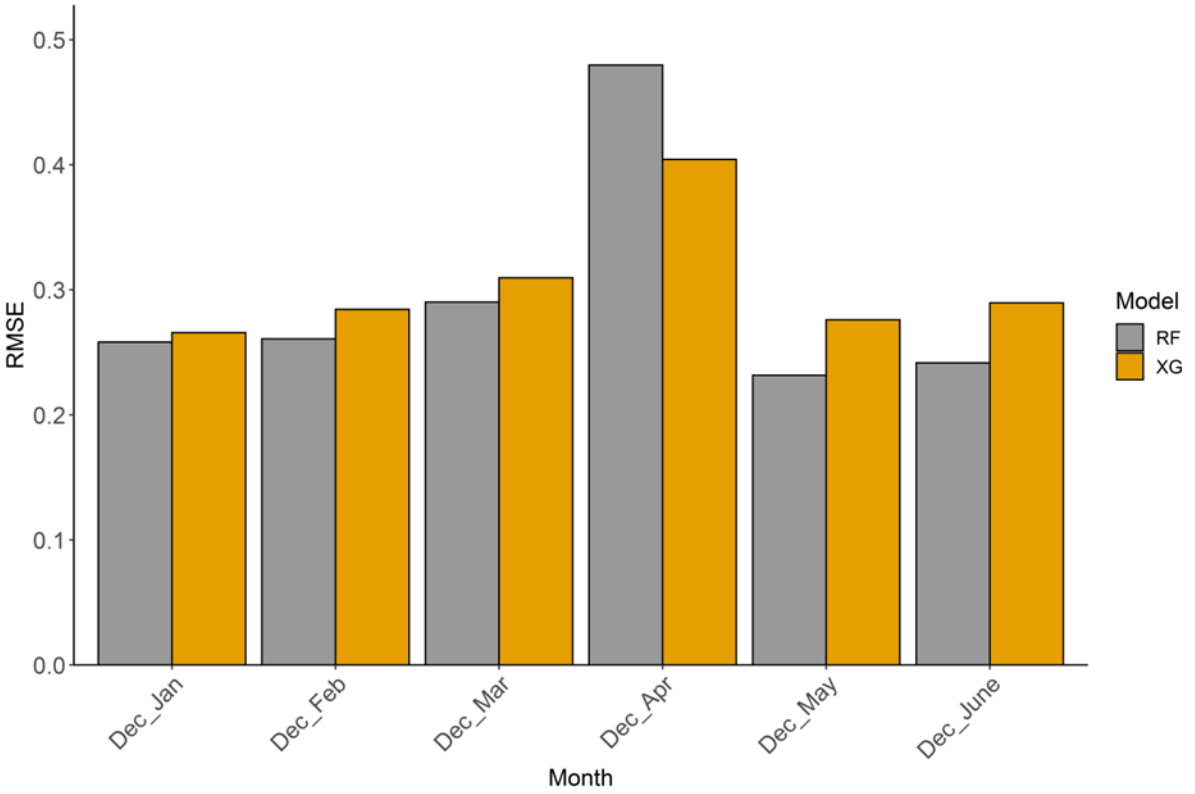

Figure 35. The RMSE for the machine learning regression models.

© University of Pretoria

5.3.2 Feature importance

Feature importance was done to determine the most important variables for modeling smallholder maize yield in April using RF and XG (Figure 36 and Figure 37). The models were similar in terms of feature ranking, both models ranked the VV_December, soil nitrogen content, VH_April and VH_March as important features. The RF model had a clear distinction of the three most important features, which had a ranking greater than 10%. However, RF needed more input data in comparison to XG where the VH_Dec, VV_February and VV_January were not important for the model.

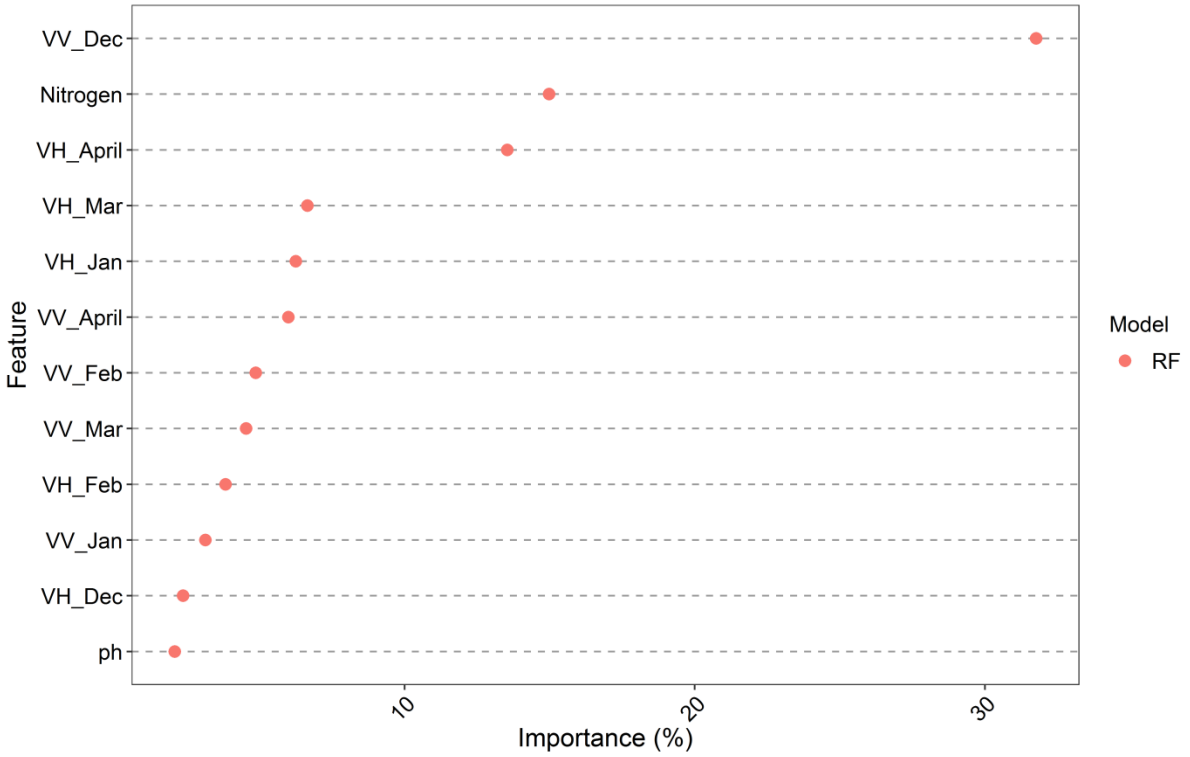

Figure 36. Feature importance plot for RF based on the December to April data cube.

© University of Pretoria

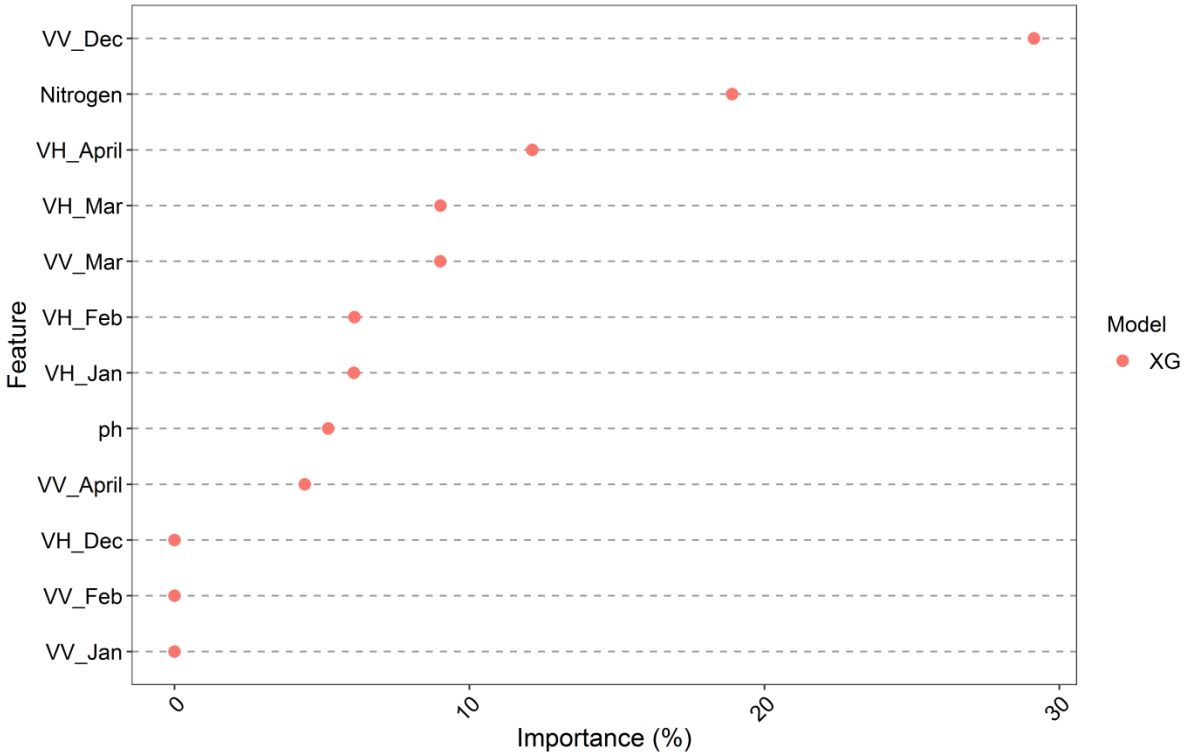

Figure 37. Feature importance for XG generated from the December-April data.

5.3.3 Model validation

Scatterplots were generated for the December-April window applying RF and XG (Figure 38). These are based on 10% of the samples, which were not used for model training. The data points are scattered away from the diagonal indicating a weak relationship between the observed yield (field collected samples) and predicted yield from the regression models. The XG model (R^2 of 0.49) had an improved performace in comparison to RF (R^2 of 0.33). Both models were significant at a 95% confidence interval based on p-values of 0.016 for XG and 0.062 for RF.

© University of Pretoria

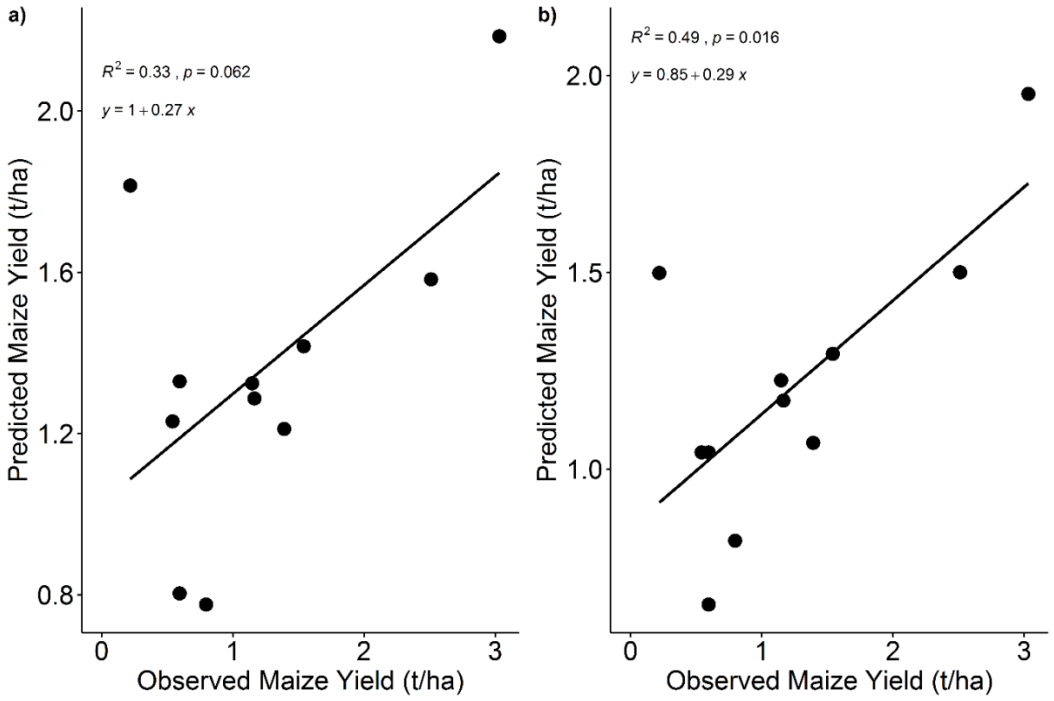

Figure 38. The observed and predicted maize yield where a) is RF and b) is XG.

5.3.4 Smallholder Maize Yield Maps

The maize yield maps are represented in Figure 39 and Figure 40 for RF and XG models. The XG model had a better distinction between the high yielding and low yielding areas in comparison to RF. High yielding areas are located on the south eastern part of the study wards and the low yielding areas are on the northern part. The south eastern part has a favourable climate, fertile soils and more smallholder maize planting areas in comparison to the northern part.

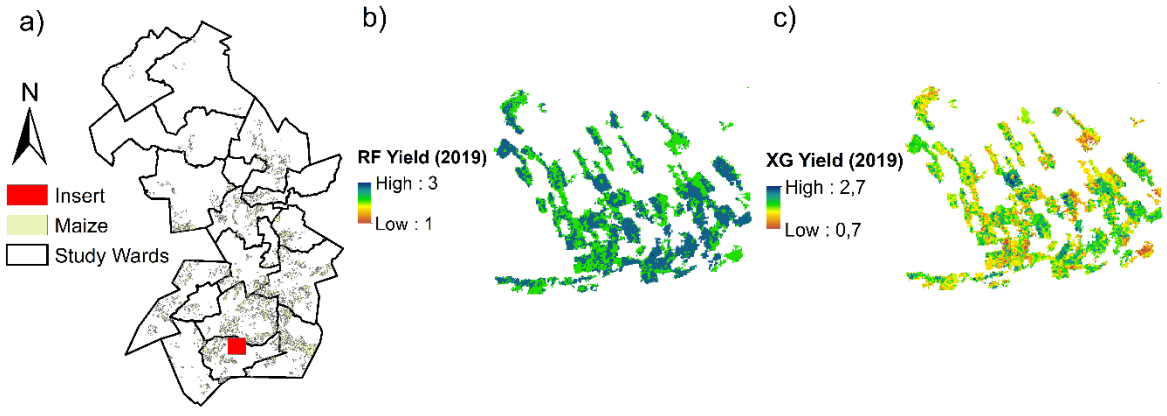

Figure 39. Insert maps for the maize yield classification generated by RF and XG.

© University of Pretoria

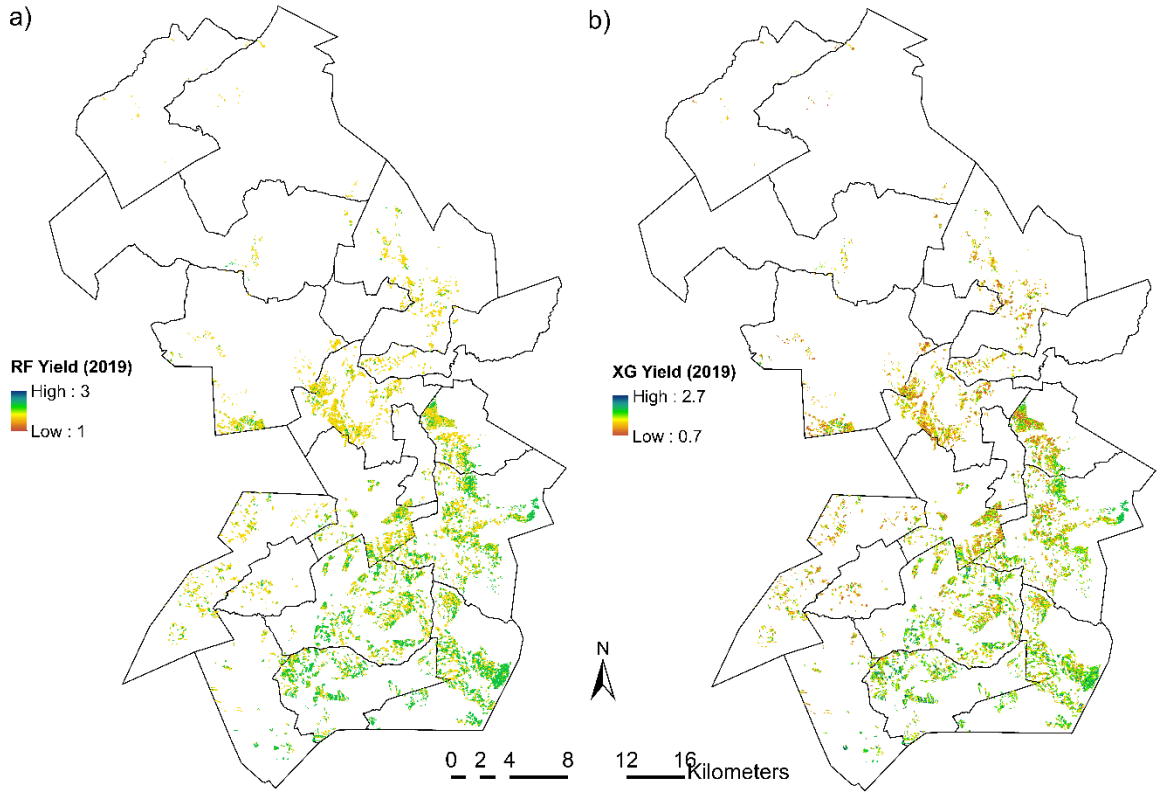

Figure 40. The spatial distribution of smallholder maize yield within the study wards.

5.4. Discussion

The aim of this study was to develop a framework based on Sentinel-1 data with the application of machine learning models, which can be easily adapted due to minimal data inputs for estimating smallholder maize yield. The optimal window for developing the yield model was identified. Feature importance was done for both RF and XG machine learning regression models. The spatial pattern of maize yield was mapped. Results from the present study indicate that maize can be estimated accurately in advance before harvest. Furthermore, the model performances show that there is scope for improving the Sentinel-1 based model to be better suited for heterogeneous smallholder farms.

Findings from this study reveal that the ideal time window for maize yield estimation is in April (two months before harvest), corresponding to the peak growing period. These findings are similar to what other authors have observed. For example,

© University of Pretoria

Leroux et al., (2019) determined that the period two months before harvest is optimal for maize yield estimation in West Africa. Ma et al., (2021) found the same period to be important for maize yield estimation in the US Corn Belt. The peak growing period reflects maximum greenness for maize, which is important for crop yield estimation (Bolton and Friedl, 2013; Rembold et al., 2013; Liu et al., 2020). In our case, the use of both the VV and VH backscatter for Sentinel-1 captured different properties for maize. The VV polarization is sensitive to soil moisture and the VH polarization increases with an increase in leaf area index and biomass. Both polarizations decrease after harvest similar to Vreugdenhil et al., (2018). The temporal evolution of profile explains the importance of the VV_December, VH_April and VH_March polarizations for both models during the feature importance analysis.

Contrary to the findings of other studies, which observed good performances for maize yield estimation (R^2: 0.69-0.89) (Fieuzal et al., 2017; Ouattara et al., 2020), the current study found an R^2 of 0.41 for XG and an R^2 of 0.31 for RF. Different factors resulted in a poor model performance. For example, the smallholder fields in the study area are heterogeneous due to poor farm management practices. There are bare patches within the farms, a lack of equal row spacing and the farmers rely on rainfall for crop production. Other factors which limit the yield of the smallholder farms are: acidic soils, lack of a use of suitable maize seeds and suboptimal fertilization. The differences in model performances were expected for XG and RF. Zhang et al., (2020) observed that the XG (R^2=0.77) model had a slightly better performance than RF (R^2=0.76) through the combined use of optical, fluorescence and thermal Satellite data. Chen et al., (2021) found that the RF model performs better (correlation coefficient of R=0.94) in comparison to the XG model (R=0.85) with the use of multisource data. Spatial statistics for crop yield are not routinely measured which meant that the observed yields could not be compared to official statistics. However, the field collected yield samples were related to those generated by the machine learning model.

The limitations of this study were that a small number of maize fields were sampled. Multisource data from Google Earth Engine (GEE) could not be integrated in the machine learning models due to the small sizes of the farms which need high

© University of Pretoria

resolution data to resolve. Potential GEE datasets which could have improved the model performance would be climate and soil moisture data. The climate data is important because climate influences the yields of widely grown cereal crops such as wheat, barley and maize. For example, Lobell and Field (2007) used 41 years of climate data and crop yields, the study showed that climate variability reduces crop yields (2-3% losses in maize, wheat and barley yields globally). Furthermore, Matiu et al., (2017) used 53 years of temperature data, standardized precipitation evapotranspiration index and related it to crop yields (maize, rice, soy beans and wheat yield). That study found that drought decreases the yield of maize by 11.6% globally. The soil moisture is important for smallholder maize farms because they depend on rain for crop production. Insufficient soil moisture affects seedling rooting and emergence at the beginning of the season (Yang et al., 2021). This study recommends the exploration of integrating Sentinel-1 and Sentinel-2 data which was not possible in the current study due to the high density of cloud cover during the cropping season.

5.5. Conclusion

This paper assessed the potential application of machine learning regression with Sentinel-1A and a limited dataset for smallholder maize yield estimation. Findings suggest that the period 2 months before harvest is optimal for early season maize yield estimation. Furthermore, the different input features were ranked. This analysis revealed that the VV_December, soil nitrogen content, VH_April and VH_March are important in the model. Validation results showed a poor relationship between the observed and predicted yields (R^2 =0.49 for XG and R^2=0.33 for RF), which shows the need for more studies focused on model optimization for smallholder maize farms. The resulting maps are important for managing maize supply and demand to improve food security in rural areas.

© University of Pretoria

Chapter 6
Synthesis

The Sustainable Development Goals (SDGs) were developed as part of the 2030 agenda for sustainable development. This initiative is aimed at ensuring zero hunger, promoting justice, reducing inequality, amongst other critical themes. Smallholder farmers contribute significantly towards the zero-hunger theme, and in most developing countries, the smallholder farming sector maybe the only practical platform for meeting SDG2. In such cases, this sector is the only means of food production and economic livelihood. However, smallholder maize farmers have a smaller size (0.5-2 ha) and are located in remote areas that are difficult to access. This poses challenges to local governments to generate continuous spatial agricultural information to support decision making process. Such information is crucial in policy development, monitoring, and implementation. Remote sensing satellites with improved spatial and temporal resolutions such as the Sentinel-1 and Sentinel-2 offer unprecedented opportunities to contribute towards documenting production, and generating the lacking agricultural statics for the smallholder sector, which is required for planning, encouraging development, and hence contribute to the many different SDGs including zero hunger. This thesis explored and developed models using Sentinel data and machine learning algorithms to support SDG 2 and smallholder maize farmers.

The aim of the study was to use Sentinel-1 and Sentinel-2 remote sensing data to map and monitor smallholder maize farms in support of the SDGs number-2 based on machine learning algorithms. In this chapter, we summarize the research findings, links between the chapters and conclusions based on the set objectives in section 1.7. A summary of each objective is given below:

1. Evaluate both Sentinel-1 and Sentinel-2 single date imagery to delineate smallholder maize farms using machine learning algorithms

For most developing countries, the literature survey had shown the dire lack of credible information on levels agricultural production, and areal extents for the smallholder farming sector. This key information is required in order for governments to allocate resources and improve this sector. Thus, this chapter explored the use of both

© University of Pretoria

Sentinel-1 and Sentinel-2 single date imagery to delineate smallholder maize farms. The results showed that single-date Sentinel-1 on its own was not sufficient in mapping planted maize fields. When Sentinel-2 data were integrated with Sentinel-1 data, an improvement of 24.2%, 8.7% and 9.1% for random forest (RF), support vector machine (SVM) and model stack (ST) algorithms, respectively, were observed. Machine learning proved to have a high capacity to estimate smallholder maize-planted areas (7001.35 ± 1.2 ha for RF, 7926.03 ± 0.7 ha for SVM and 7099.59 ± 0.8 ha for ST). The framework used in this study can be applied when evaluating different algorithms for mapping smallholder farms. The crop maps derived in this study are fundamental for crop monitoring, land-use policies and aiding food security planning activities. The integration of Sentinel-1 and Sentinel-2 is optimal for delineating smallholder farms and area estimates with single date imagery. This approach can be used in resource scarce areas.

2. **Develop an innovative approach using Sentinel-1 time-series data and machine learning algorithms (integrating both supervised and unsupervised methods) to map smallholder maize farms.**

This objective utilized Sentinel-1 multi-temporal data for mapping smallholder maize farms' spatial distribution and estimate production areas. The two-stage image fusion approach was adopted. The multi-temporal approach was investigated to compare with the single-date approach in Chapter 2 and determine which technique produces accurate results. The SVM and extreme gradient boosting (XG) machine learning algorithms were applied. The results revealed that most smallholder farms in our study area are distributed in the south eastern part of Makhuduthamaga. The algorithms provided comparable statistical evaluation results. However, McNemar's test showed that the results from the two algorithms were statistically different from each other. The SVM and XG algorithms estimated maize production areas to be 7073.558 ± 0.01 ha and 7303.32 ± 0.180 ha, respectively, for the region. The classified areas for selected farms compared favorably with the measured areas in the field and the SVM classifier had a better fit (R = 0.89) in comparison with the XG algorithm (R = 0.84). The SVM algorithm seems to have generally performed better than the XG algorithm. The use

© University of Pretoria

of multi-temporal Sentinel-1 with a two-stage image fusion approach proved to be effective in mapping smallholder farms.

3. Investigate the utility of machine learning regression for spatial predictions of soil nitrogen content in smallholder maize farms.

This study was aimed at assessing Sentinel-2 bands, derived soil and vegetation indices and environmental variables for predicting soil nitrogen in smallholder maize farms applying machine learning regression. This procedure was done for the farms identified in Chapter 2 and Chapter 3. Previous research has shown that soil nitrogen deficiencies are limiting for maize growth, thus it was important to establish a framework to monitor it (Xu et al., 2018). Different predictor variables were related to soil nitrogen content. The red, red-edge and short-wave infrared bands were strongly related to soil nitrogen with correlations of 0.89-0.90. The machine learning models applied in this study (RF, GB, and XG) were suitable for the data because multicollinearity was present between the predictors, which these models dealt with effectively. Model evaluation results show that machine learning models have a high predictive capacity in estimating soil nitrogen (R^2=0.84-0.90 and RMSE=0.0076-0.0094%) in smallholder farms. Variable importance revealed that the Sentinel-2 bands, particularly the red and red-edge bands are fundamental for modeling soil nitrogen in all three models.

4. Develop procedures using Sentinel-1 data to model maize yield in complex environments.

This objective was aimed at developing Sentinel-1 based procedures with minimal field collected data for mapping smallholder maize yield early in the season using machine learning. This procedure is important for forecasting maize yield in for ensured food security and thus contributing to SDG2. The farms identified in Chapter 2 and Chapter 3 were considered. Two machine learning models were tested—RF and XG. The findings showed that maize yield can be predicted accurately from two months before harvest. The model accuracies were low (R^2 of 0.2-0.41), this was expected for smallholder farms. These farms are fragmented and usually have non-homogeneous

© University of Pretoria

planting patterns. Feature importance showed that VV_December, soil nitrogen content, VH_April and VH_March are important variables for estimating maize yield.

Summary of the scientific contribution

Chapter 1 provided a general introduction and a direction for the thesis. Describing the importance of remote sensing technology in supporting SDGs within the context of smallholder maize farms. Chapter 2 and Chapter 3 explored techniques to accurately identify smallholder maize farms within the complex environments. A single date and multi-temporal approaches were studied to identify smallholder maize farms and estimate planted areas. The two approaches yielded comparable results, however there were noticeable differences between the models i.e., (1) a single date approach produced high variance between the model estimates and each estimate of the planted area had higher standard deviation. (2) While a multi-temporal approach produced smaller variance between the estimates of planted areas and each model had much lower standard deviation estimates compared to the single date approach. Therefore, the multi-temporal approach is the recommended method for mapping smallholder maize farms. However, this method is computationally intensive and not easy to implement, while the single date method is relatively easy to implement and requires less computation resources. Chapter 4 used machine learning regression methods to estimate spatial distribution of the total soil nitrogen content at smallholder maize farms. Nitrogen is one of the most important nutrients in the soil for plants. The ability of Sentinel-2 data to support production of nitrogen maps with a medium spatial resolution (10 m) is an important contribution towards achieving SDGs number 2. Especially, in developing countries such as South Africa where such information is lacking. This framework can be used to provide spatial agricultural information and the associated statistics to inform policy design and implementation by local government. Recommendations by governments, extension services and fertilizer dealers can also benefit from using such maps. These maps are useful to establish nitrogen management plans in the smallholder farms, which will increase maize yields, thereby, improving food security. The use of Sentinel-1 and minimal field collected data in Chapter 5 has not been extensively explored for smallholder maize yield estimates which is necessary for African farmers which are resource scarce and face issues of

© University of Pretoria

food security. Early season yield predictions will aid in planning for maize supply and demand.

Future work

- The developed framework should be tested at different seasons, different climatic zones, and different crop types to assess its robustness under these different conditions.
- A Multi-temporal approach in mapping crop types should include phase, mean, and amplitude data as extracted from the multi-temporal images per pixel. This additional information could potentially improve classification results.
- Application of deep leaning algorithms should be explored (e.g., Convolutional Neural Network) as these algorithms have a potential to outperform traditional machine learning algorithms.
- Model optimization should be done for the yield model derived for smallholder farms and the integration of high spatial resolution data should be explored.
- To upscale the framework of monitoring smallholder crop farms, it is recommended that the Python code be implemented/converted to Java scripting to allow easy implementation on the cloud platforms such as Google Earth Engine platform (https://earthengine.google.com/). Cloud computing will execute the code much faster than the Ryzen 9 system that was adopted in this study.
- Other data products should be developed to contribute to monitoring of SDG 2 for example land suitability maps using Sentinel data which will guide farmers of where optimal areas are located for smallholder maize production.

© University of Pretoria

References

1. Abdi, H. and Williams, L.J., 2010. Principal component analysis: Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2 (4), 433–459.
2. Abraham, M. and Pingali, P., 2020. Transforming Smallholder Agriculture to Achieve the SDGs. In *The Role of Smallholder Farms in Food and Nutrition Security*; Gomez y Paloma, S., Riesgo, L., Louhichi, K., Eds.; Berlin/Heidelberg, Germany: Springer International Publishing, 173–209.
3. Abubakar, G.A., Wang, K., Shahtahamssebi, A., Xue, X., Belete, M., Gudo, A.J.A., Mohamed Shuka, K.A. and Gan, M., 2020. Mapping Maize Fields by Using Multi-Temporal Sentinel-1A and Sentinel-2A Images in Makarfi, Northern Nigeria, Africa. *Sustainability*, 12(6), 2539.
4. Aggarwal, C.C., ed., 2014. *Data classification: algorithms and applications*. Florida, USA: CRC Press, Taylor & Francis Group.
5. Aguilar, R., Zurita-Milla, R., Izquierdo-Verdiguier, E., and A. de By, R., 2018. A Cloud-Based Multi-Temporal Ensemble Classifier to Map Smallholder Farming Systems. *Remote Sensing*, 10(5), 729.
6. Ahmad, I., Singh, A., Fahad, M. and Waqas, M.M., 2020. Remote sensing-based framework to predict and assess the interannual variability of maize yields in Pakistan using Landsat imagery. *Computers and Electronics in Agriculture*, 178, 105732.
7. Aliber, M. and Hall, R., 2012. Support for smallholder farmers in South Africa: Challenges of scale and strategy. *Development Southern Africa*, 29(4), 548-562.
8. Archontoulis, S.V., Miguez, F.E. and Moore, K.J., 2014. Evaluating APSIM maize, soil water, soil nitrogen, manure, and soil temperature modules in the Midwestern United States. *Agronomy Journal*, 106(3), 1025-1040.
9. Arias, M., Campo-Bescós, M.Á., and Álvarez-Mozos, J., 2020. Crop Classification Based on Temporal Signatures of Sentinel-1 Observations over Navarre Province, Spain. *Remote Sensing*, 12(2), 278.
10. Armitage, D.W. and Ober, H.K., 2010. A comparison of supervised learning techniques in the classification of bat echolocation calls. *Ecological Informatics*, 5 (6), 465–473.
11. Asner, G.P., 2001. Cloud cover in Landsat observations of the Brazilian Amazon. *International Journal of Remote Sensing*, 22 (18), 3855–3862.
12. Attema, E., Davidson, M., Floury, N., Levrini, G., Rosich-Tell, B., Rommen, B., and Snoeij, P., 2007. Sentinel-1 ESA's new European SAR mission. In: *Remote Sensing*; Meynart, R., Neeck, S. P., Shimoda, H., and Habib, S., Eds.; Florence, Italy: SPIE, 674403.
13. Baret, F., Weiss, M., Lacaze, R., Camacho, F., Makhmara, H., Pacholcyzk, P. and Smets, B., 2013. GEOV1: LAI and FAPAR essential climate variables and FCOVER global time series capitalizing over existing products. Part1: Principles of development and production. *Remote Sensing of Environment*, 137, 299-309.
14. Beguin, J., Fuglstad, G.A., Mansuy, N. and Paré, D., 2017. Predicting soil properties in the Canadian boreal forest with limited data: Comparison of spatial and non-spatial statistical approaches. *Geoderma*, *306*, 195-205.
15. Belgiu, M. and Drăguţ, L., 2016. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 24–31.
16. Belward, A. S. and Skøien, J. O., 2015. Who launched what, when and why; trends in global land-cover observation capacity from civilian earth observation satellites. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103, 115–128.
17. Benos, L., Tagarakis, A.C., Dolias, G., Berruto, R., Kateris, D. and Bochtis, D., 2021. Machine Learning in Agriculture: A Comprehensive Updated Review. *Sensors*, 21(11), 3758.

© University of Pretoria

18. Bishop Christopher, M., 2006. Pattern recognition and machine learning. *Information science and statistics New York: Springer*.
19. Bolton, D.K. and Friedl, M.A., 2013. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agricultural and Forest Meteorology*, *173*, 74-84.
20. Boryan, C., Yang, Z., Mueller, R., and Craig, M., 2011. Monitoring US agriculture: the US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program. *Geocarto International*, 26 (5), 341–358.
21. Breiman, L., 2001. Random Forests. *Machine Learning*, 45 (1), 5–32.
22. Brisson, N., Seguin, B. and Bertuzzi, P., 1992. Agrometeorological soil water balance for crop simulation models. *Agricultural and forest meteorology*, 59(3-4), 267-287.
23. Broge, N.H. and Leblanc, E., 2001. Comparing prediction power and stability of broadband and hyperspectral vegetation indices for estimation of green leaf area index and canopy chlorophyll density. *Remote Sensing of Environment*, *76*(2), 156-172.
24. Batjes, N.H., 1996. Total carbon and nitrogen in the soils of the world. *European Journal of Soil Science*, 47(2).
25. Batjes, N.H., 2004. SOTER-based soil parameter estimates for Southern Africa. Report 2004/04. ISRIC – World Soil Information, Wageningen.
26. Baxter, S.J. and Oliver, M.A., 2005. The spatial prediction of soil mineral N and potentially available N using elevation. *Geoderma*, *128*(3-4), 325-339.
27. Bullard, J.E. and White, K., 2002. Quantifying iron oxide coatings on dune sands using spectrometric measurements: An example from the Simpson-Strzelecki Desert, Australia. *Journal of Geophysical Research: Solid Earth*, 107(B6), ECV-5.
28. Cai, Y., Lin, H., and Zhang, M., 2019. Mapping paddy rice by the object-based random forest method using time series Sentinel-1/Sentinel-2 data. *Advances in Space Research*, 64 (11), 2233–2244.
29. Calatayud, P.A., Le Ru, B.P., Van den Berg, J. and Schulthess, F., 2014. Ecology of the African maize stalk borer, Busseola fusca (Lepidoptera: Noctuidae) with special reference to insect-plant interactions. *Insects*, 5(3), 539-563.
30. Campbell, J.B.; Wynne, R.H., 2011. *Introduction to Remote Sensing*, 5th ed.; New York, USA: Guilford Press.
31. Camps-Valls, G. and Bruzzone, L. eds., 2009. *Kernel methods for remote sensing data analysis*. John Wiley & Sons.
32. Canty, M.J. 2014. *Image Analysis, Classification and Change Detection in Remote Sensing: With Algorithms for ENVI/IDL and Python*, 3rd ed.; Florida, USA: CRC Press.
33. Capricorn district municipality (CDM), 2015. Capricorn District Municipality: 2016/17-2021 Final Draft IDP/Budget. Capricorn district municipality, Limpopo.
34. Capricorn district municipality (CDM), 2018. Annual performance report 2017/18. Capricorn district municipality, Limpopo.
35. Carslaw, D.C. and Ropkins, K., 2012. Openair—an R package for air quality data analysis. *Environmental Modelling & Software*, *27*, 52-61.
36. Chakhar, A., Ortega-Terol, D., Hernández-López, D., Ballesteros, R., Ortega, J.F. and Moreno, M.A., 2020. Assessing the accuracy of multiple classification algorithms for crop classification using Landsat-8 and Sentinel-2 data. *Remote Sensing*, 12(11), 1735.
37. Chang, D.H. and Islam, S., 2000. Estimation of soil physical properties using remote sensing and artificial neural network. *Remote Sensing of Environment*, 74(3), 534-544.
38. Charman, A. and Hodge, J., 2007. Food Security in the SADC Region: An Assessment of National Trade Strategy in the Context of the 2001–03 Food Crisis. In: *Food Insecurity, Vulnerability and Human Rights Failure*; Guha-Khasnobis, B., Acharya, S.S., Davis, B., Eds.; Studies in Development Economics and Policy; London, United Kingdom: Palgrave Macmillan, 58–81.
39. Chatziantoniou, A., Psomiadis, E. and Petropoulos, G.P., 2017. Co-Orbital Sentinel 1 and 2 for LULC mapping with emphasis on wetlands in a mediterranean setting based on machine learning. *Remote Sensing*, 9(12), 1259.

© University of Pretoria

40. Chlingaryan, A., Sukkarieh, S. and Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, 151, 61-69.
41. Chen, J.M., 1996. Evaluation of Vegetation Indices and a Modified Simple Ratio for Boreal Applications. *Canadian Journal of Remote Sensing*, 22 (3), 229–242.
42. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y. and Cho, H., 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2*, *1*(4), 1-4.
43. Chen, T. and Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA: ACM, 785–794.
44. Chen, X., Feng, L., Yao, R., Wu, X., Sun, J. and Gong, W., 2021. Prediction of Maize Yield at the City Level in China Using Multi-Source Data. *Remote Sensing*, 13(1), 146.
45. Cheng, Z., Meng, J. and Wang, Y., 2016. Improving spring maize yield estimation at field scale by assimilating time-series HJ-1 CCD data into the WOFOST model using a new method with fast algorithms. *Remote Sensing*, *8*(4), 303.
46. Chivasa, W., 2017. Application of remote sensing in estimating maize grain yield in heterogeneous African agricultural landscapes. *International Journal of Remote Sensing*, *38*(23), 6816-6845.
47. Clevers, J.G. and Gitelson, A.A., 2013. Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on Sentinel-2 and-3. *International Journal of Applied Earth Observation and Geoinformation*, 23, 344-351.
48. Cochran, F., Daniel, J., Jackson, L. and Neale, A., 2020. Earth observation-based ecosystem services indicators for national and subnational reporting of the sustainable development goals. *Remote Sensing of Environment*, 244, 111796.
49. Congalton, R.G. and Green, K. 2008. A*ssessing the Accuracy of Remotely Sensed Data Principles and Practices*, 2nd ed; Florida, USA: CRS Press.
50. Cooner, A.J., Shao, Y. and Campbell, J.B., 2016. Detection of urban damage using remote sensing and machine learning algorithms: Revisiting the 2010 Haiti earthquake. *Remote Sensing*, 8(10), 868.
51. Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine Learning*, 20(3), 273–297.
52. Crippen, R., 1990. Calculating the vegetation index faster. *Remote Sensing of Environment*, 34(1), 71–73.
53. Cucho-Padin, G., Loayza, H., Palacios, S., Balcazar, M., Carbajal, M., and Quiroz, R., 2020. Development of low-cost remote sensing tools and methods for supporting smallholder agriculture. *Applied Geomatics*, 12(3), 247–263.
54. Cumming, G. and Calin-Jageman, R., 2016. *Introduction to the new statistics: Estimation, open science, and beyond*. Routledge, New York.
55. Dangeti, P., 2017. Statistics for Machine Learning: Techniques for Exploring Supervised, Unsupervised, and Reinforcement Learning Models with Python and R, Packt Publishing: Birmingham, UK, ISBN 9781788295758.
56. Davis, J. and Goadrich, M., 2006. The relationship between Precision-Recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, USA, 233–240.
57. De Brogniez, D., Ballabio, C., Stevens, A., Jones, R.J.A., Montanarella, L. and van Wesemael, B., 2015. A map of the topsoil organic carbon content of Europe generated by a generalized additive model. *European Journal of Soil Science*, 66(1), 121-134.
58. De Leeuw, J., Jia, H., Yang, L., Liu, X., Schmidt, K., and Skidmore, A. K., 2006. Comparing accuracy assessments to infer superiority of image classification methods. *International Journal of Remote Sensing*, 27(1), 223–232.
59. Delegido, J., Verrelst, J., Alonso, L. and Moreno, J., 2011. Evaluation of sentinel-2 red-edge bands for empirical estimation of green LAI and chlorophyll content. *Sensors*, 11(7), 7063-7081.

© University of Pretoria

60. Deines, J.M., Patel, R., Liang, S.Z., Dado, W. and Lobell, D.B., 2021. A million kernels of truth: Insights into scalable satellite maize yield mapping and yield gap analysis from an extensive ground dataset in the US Corn Belt. *Remote Sensing of Environment*, 253, 112174.
61. Deng, X., Ma, W., Ren, Z., Zhang, M., Grieneisen, M.L., Chen, X., Fei, X., Qin, F., Zhan, Y. and Lv, X., 2020. Spatial and temporal trends of soil total nitrogen and C/N ratio for croplands of East China. *Geoderma*, 361, 114035.
62. Deschamps, B., McNairn, H., Shang, J. and Jiao, X., 2012. Towards operational radar-only crop type classification: comparison of a traditional decision tree with a random forest classifier. *Canadian Journal of Remote Sensing*, 38(1), 60-68.
63. Dobson, M.C., Ulaby, F.T. and Pierce, L.E., 1995. Land-cover classification and estimation of terrain attributes using synthetic aperture radar. *Remote sensing of Environment*, 51(1), 199-214.
64. Dong, H., Xu, X., Wang, L. and Pu, F., 2018. Gaofen-3 PolSAR Image Classification via XGBoost and Polarimetric Spatial Information. *Sensors*, 18(2), 611.
65. Dos Santos Ferreira, A., Freitas, D.M., da Silva, G.G., Pistori, H. and Folhes, M.T., 2017. Weed detection in soybean crops using ConvNets. *Computers and Electronics in Agriculture*, 143, 314-324.
66. Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., and Bargellini, P., 2012. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment*, 120, 25–36.
67. Duguay, Y., Bernier, M., Lévesque, E., and Tremblay, B., 2015. Potential of C and X Band SAR for Shrub Growth Monitoring in Sub-Arctic Environments. *Remote Sensing*, 7(7), 9410–9430.
68. Du Plessis, J., 2003. *Maize production*. Pretoria, South Africa: Department of Agriculture.
69. Edwards, A.L., 1948. Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. *Psychometrika*, 13(3), 185–187.
70. Efron, B., 1983. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *Journal of the American Statistical Association*, 78(382), 316–331.
71. Elith, J., Leathwick, J. R., and Hastie, T., 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4), 802–813.
72. Ermida, S.L., Soares, P., Mantas, V., Göttsche, F.M. and Trigo, I.F., 2020. Google earth engine open-source code for land surface temperature estimation from the landsat series. *Remote Sensing*, 12(9), 1471.
73. ESA, 2018. *Sen2Cor*, Paris, France: ESA.
74. Fang, D., Zhang, X., Yu, Q., Jin, T.C. and Tian, L., 2018. A novel method for carbon dioxide emission forecasting based on improved Gaussian processes regression. *Journal of Cleaner Production*, 173, 143-150.
75. FAO, 2016. *Save and grow in practice: maize, rice and wheat, a guide to sustainable cereal production*, Rome, Italy: FAO.
76. FAO, 2016. *Food and Agriculture Organization of the United Nations OECD-FAO Agricultural Outlook 2016–2025*, Rome, Italy: FAO.
77. FAO, 2018. *Building Climate Resilience for Food Security and Nutrition*, Rome, Italy: FAO.
78. FAO, 2019. *Global Information and Early Warning System on Food and Agriculture Crop Prospects and Food Situation*, Rome, Italy: FAO.
79. Farrar, D.E. and Glauber, R.R., 1967. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, 92-107.
80. Farr, T.G., Rosen, P.A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., and Alsdorf, D.E., 2007. The shuttle radar topography mission: Reviews of Geophysics, (45)2, RG2004.

© University of Pretoria

81. Farrell, A., Wang, G., Rush, S.A., Martin, J.A., Belant, J.L., Butler, A.B. and Godwin, D., 2019. Machine learning of large-scale spatial distributions of wild turkeys with high-dimensional environmental data. *Ecology and Evolution*, 9(10), 5938-5949.
82. Fernández-Manso, A., Fernández-Manso, O. and Quintano, C., 2016. SENTINEL-2A red-edge spectral indices suitability for discriminating burn severity. *International Journal of Applied Earth Observation and Geoinformation*, 50, 170-175.
83. Fieuzal, R., Sicre, C.M. and Baup, F., 2017. Estimation of corn yield using multi-temporal optical and radar satellite data and artificial neural networks. *International Journal of Applied Earth Observation and Geoinformation*, 57, 14-23.
84. Filella, I. and Penuelas, J., 1994. The red edge position and shape as indicators of plant chlorophyll content, biomass and hydric status. *International Journal of Remote Sensing*, 15(7), 1459-1470.
85. Filipponi, F., 2021. Sentinel-1 GRD Preprocessing Workflow. In: *Proceedings of the 3rd International Electronic Conference on Remote Sensing*, Roma, Italy: ECRS, 18, 11.
86. Fischer, K. and Hajdu, F., 2015. Does raising maize yields lead to poverty reduction? A case study of the Massive Food Production Programme in South Africa. *Land Use Policy*, 46, 304-313.
87. Flach, P.; Kull, M., 2015. *Precision-Recall-Gain Curves: PR Analysis Done Right*. Advances in Neural Information Processing Systems; Bristol, United Kingdom: NIPS.
88. Foody, G.M. and Mathur, A., 2004. Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification. *Remote Sensing of Environment*, 93 (1–2), 107–117.
89. Foody, G.M. and Mathur, A., 2006. The use of small training sets containing mixed pixels for accurate hard image classification: Training on mixed spectral responses for classification by a SVM. *Remote Sensing of Environment*, 103(2), 179–189.
90. Foody, G.M., 2009. Sample size determination for image classification accuracy assessment and comparison. *International Journal of Remote Sensing*, 30 (20), 5273–5291.
91. Forkuor, G., Conrad, C., Thiel, M., Ullmann, T., and Zoungrana, E., 2014. Integration of Optical and Synthetic Aperture Radar Imagery for Improving Crop Mapping in Northwestern Benin, West Africa. *Remote Sensing*, 6 (7), 6472–6499.
92. Forkuor, G., Hounkpatin, O.K., Welp, G. and Thiel, M., 2017. High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: a comparison of machine learning and multiple linear regression models. *PloS one*, 12(1), e0170478.
93. Forkuor, G., Dimobe, K., Serme, I. and 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 & Remote Sensing*, 55(3), 331-354.
94. Friedl, M.A. and Brodley, C.E., 1997. Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, 61(3), 399-409.
95. Friedman, J., Hastie, T. and Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, *28*(2), 337-407.
96. Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189-1232.
97. GEE (2021). Sentinel-1 Algorithms. Available online: https://developers.google.com/earth-engine/sentinel1 (accessed on 24 August 2021).
98. Georganos, S., Grippa, T., Vanhuysse, S., Lennert, M., Shimoni, M., and Wolff, E., 2018. Very High Resolution Object-Based Land Use–Land Cover Urban Classification Using Extreme Gradient Boosting. *IEEE Geoscience and Remote Sensing Letters*, 15 (4), 607–611.
99. Giller, K.E., Rowe, E.C., de Ridder, N. and van Keulen, H., 2006. Resource use dynamics and interactions in the tropics: Scaling up in space and time. *Agricultural Systems*, 88(1), 8-27.

© University of Pretoria

100. Gitelson, A.A., Kaufman, Y.J., and Merzlyak, M.N., 1996. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sensing of Environment*, 58 (3), 289–298.
101. Golam, F., Farhana, N., Zain, M.F., Majid, N.A., Rahman, M.M., Rahman, M.M. and Kadir, M.A., 2011. Grain yield and associated traits of maize (Zea mays L.) genotypes in Malaysian tropical environment. *African Journal of Agricultural Research*, 6(28), 6147-6154.
102. Gommes, R. 1998. “Overview of Variables Used Agrometeorological Crop Forecasting Tools. 323-326.” In Agrometeorological Application for Regional Crop Monitoring and Production Assessment, edited by D. Rijks, J. M. Terres, and P. Vossen, 516. Italy: Space Research Institute, EC Joint Research Centre, European Commission.
103. Haboudane, D., Miller, J.R., Pattey, E., Zarco-Tejada, P.J. and Strachan, I.B., 2004. Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture. *Remote Sensing of Environment*, *90*(3), 337-352.
104. Hakimdavar, R., Hubbard, A., Policelli, F., Pickens, A., Hansen, M., Fatoyinbo, T., Lagomasino, D., Pahlevan, N., Unninayar, S., Kavvada, A. and Carroll, M., 2020. Monitoring water-related ecosystems with earth observation data in support of Sustainable Development Goal (SDG) 6 reporting. *Remote Sensing*, 12(10), 1634.
105. Haupt, S.E., Pasini, A. and Marzban, C. eds., 2008. Artificial intelligence methods in the environmental sciences. Springer Science & Business Media.
106. Heumann, B.W., 2011. An object-based classification of mangroves using a hybrid decision tree—Support vector machine approach. *Remote Sensing*, 3(11), 2440-2460.
107. Homolova, L., Malenovský, Z., Clevers, J.G., Garcia-Santos, G. and Schaepman, M.E., 2013. Review of optical-based remote sensing for plant trait mapping. *Ecological Complexity*, *15*, 1-16.
108. Huang, C., Davis, L.S. and Townshend, J.R.G., 2002. An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, *23*(4), 725-749.
109. Huang, Y., Chen, Z.X., Tao, Y.U., Huang, X.Z. and Gu, X.F., 2018. Agricultural remote sensing big data: Management and applications. *Journal of Integrative Agriculture*, 17(9), 1915-1931.
110. Huete, A.R., 1988. A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment*, 25(3), 295-309.
111. Hussain, M.Z., Khan, M.A. and Ahmad, S.R., 2006. Micro-nutrients status of Bannu Basin soils. *Sarhad Journal of Agriculture*, 22(2), 283.
112. Hutengs, C. and Vohland, M., 2016. Downscaling land surface temperatures at regional scales with random forest regression. *Remote Sensing of Environment*, 178, 127-141.
113. Immitzer, M., Vuolo, F., and Atzberger, C., 2016. First Experience with Sentinel-2 Data for Crop and Tree Species Classifications in Central Europe. *Remote Sensing*, 8(3), 166.
114. Inglada, J., Vincent, A., Arias, M., and Marais-Sicre, C., 2016. Improved Early Crop Type Identification By Joint Use of High Temporal Resolution SAR And Optical Image Time Series. *Remote Sensing*, 8(5).
115. Inoue, Y., Kurosu, T., Maeno, H., Uratsuka, S., Kozu, T., Dabrowska-Zielinska, K. and Qi, J., 2002. Season-long daily measurements of multifrequency (Ka, Ku, X, C, and L) and full-polarization backscatter signatures over paddy rice field and their relationship with biological variables. *Remote Sensing of Environment*, *81*(2-3), 194-204.
116. Izquierdo-Verdiguier, E., Gómez-Chova, L., Bruzzone, L. and Camps-Valls, G., 2014. Semisupervised kernel feature extraction for remote sensing image analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 52(9), 5567-5578.
117. Jeong, G., Oeverdieck, H., Park, S.J., Huwe, B. and Ließ, M., 2017. Spatial soil nutrients prediction using three supervised learning methods for assessment of land potentials in complex terrain. *Catena*, 154, 73-84.

© University of Pretoria

118. Jain, M., Mondal, P., DeFries, R.S., Small, C. and Galford, G.L., 2013. Mapping cropping intensity of smallholder farms: A comparison of methods using multiple sensors. *Remote Sensing of Environment*, 134, 210-223.
119. Jari, B. and Fraser, G.C.G., 2009. An analysis of institutional and technical factors influencing agricultural marketing amongst smallholder farmers in the Kat River Valley, Eastern Cape Province, South Africa. *African Journal of Agricultural Research*, 4(11), 1129-1137.
120. Jaya, I.G.N.M., Ruchjana, B. and Abdullah, A., 2020. Comparison Of Different Bayesian And Machine Learning Methods In Handling Multicollinearity Problem: A Monte Carlo Simulation Study. *ARPN J. Eng. Appl. Sci*, 15(18), 1998-2011.
121. Ji, Z., Pan, Y., Zhu, X., Wang, J. and Li, Q., 2021. Prediction of Crop Yield Using Phenological Information Extracted from Remote Sensing Vegetation Index. *Sensors*, 21(4), 1406.
122. Jin, Z., Azzari, G., You, C., Di Tommaso, S., Aston, S., Burke, M. and Lobell, D.B., 2019. Smallholder maize area and yield mapping at national scales with Google Earth Engine. *Remote Sensing of Environment*, 228, 115-128.
123. Jones, A., Breuning-Madsen, H., Brossard, M., Dampha, A., Deckers, J., Dewitte, O., Gallali, T., Hallett, S., Jones, R., Kilasara, M., Le Roux, P., Michéli, E., Montanarella, L., Spaargaren, O., Thiombiano, L., Van Ranst, E., Yemefack, M., Zougmore, R., (eds.), 2013. Soil Atlas of Africa. European Commission, Publications Office of the European Union, Luxembourg. 176 pp. ISBN 978-92-79-26715-4.
124. Jones, P.G. and Thornton, P.K., 2015. Representative soil profiles for the Harmonized World Soil Database at different spatial resolutions for agricultural modelling applications. *Agricultural Systems*, *139*, 93-99.
125. Jordan, C.F., 1969. Derivation of Leaf-Area Index from Quality of Light on the Forest Floor. *Ecology*, 50(4), 663–666.
126. Kang, Y., Ozdogan, M., Zhu, X., Ye, Z., Hain, C. and Anderson, M., 2020. Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest. *Environmental Research Letters*, 15(6), 064005.
127. Karthikeyan, L., Chawla, I. and Mishra, A.K., 2020. A review of remote sensing applications in agriculture for food security: Crop growth and yield, irrigation, and crop losses. *Journal of Hydrology*, 586, 124905.
128. Kavvada, A., Metternicht, G., Kerblat, F., Mudau, N., Haldorson, M., Laldaparsad, S., Friedl, L., Held, A. and Chuvieco, E., 2020. Towards delivering on the Sustainable Development Goals using Earth observations. *Remote Sensing of Environment*, 247, 111930.
129. Kenduiywo, B.K., Bargiel, D. and Soergel, U., 2018. Crop-type mapping from a sequence of Sentinel 1 images. I*nternational Journal of Remote Sensing*, 39(19), 6383-6404.
130. Khosravi, I., Safari, A., and Homayouni, S., 2017. Separability analysis of multifrequency SAR polarimetric features for land cover classification. *Remote Sensing Letters*, 8 (12), 1152–1161.
131. Kim, H.-O. and Yeom, J.-M., 2014. Effect of red-edge and texture features for object-based paddy rice crop classification using RapidEye multi-spectral satellite image data. *International Journal of Remote Sensing*, 1–23.
132. Knoepp, J.D. and Swank, W.T., 2002. Using soil temperature and moisture to predict forest soil nitrogen mineralization. *Biology and Fertility of Soils*, 36(3), 177-182.
133. Knorn, J., Rabe, A., Radeloff, V. C., Kuemmerle, T., Kozak, J., and Hostert, P., 2009. Land cover mapping of large areas using chain classification of neighboring Landsat satellite images. *Remote Sensing of Environment*, 113 (5), 957–964.
134. Knox, J., Hess, T., Daccache, A., and Wheeler, T., 2012. Climate change impacts on crop productivity in Africa and South Asia. *Environmental Research Letters*, 7 (3), 034032.

© University of Pretoria

135. Kogan, F. 2018. Food Security: The Twenty-First Century Issue. In: *Promoting the Sustainable Development Goals in North American Cities*; Metzler, J.B., Ed.; Berlin/Heidelberg, Germany: Springer, 9–22.
136. Kogan, F., 2019. Remote sensing for food security. Springer International Publishing.
137. Kubota, T., Shige, S., Hashizume, H., Aonashi, K., Takahashi, N., Seto, S., Hirose, M., Takayabu, Y.N., Ushio, T., Nakagawa, K. and Iwanami, K., 2007. Global precipitation map using satellite-borne microwave radiometers by the GSMaP project: Production and validation. *IEEE Transactions on Geoscience and Remote Sensing*, *45*(7), 2259-2275.
138. Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., and Cooper, T., 2017. *Caret: Classification and Regression Training, R Package Version 6.0-76*, Available online: http://cran.r-project.org/package=caret (accessed on 2 June 2020).
139. Kumar, P., Gupta, D. K., Mishra, V. N., and Prasad, R., 2015. Comparison of support vector machine, artificial neural network, and spectral angle mapper algorithms for crop classification using LISS IV data. *International Journal of Remote Sensing*, 36 (6), 1604–1617.
140. Lal, H., Hoogenboom, G., Calixte, J.P., Jones, J.W. and Beinroth, F.H., 1993. Using crop simulation models and GIS for regional productivity analysis. *Transactions of the ASAE*, 36(1), 175-184.
141. Lary, D.J., Alavi, A.H., Gandomi, A.H. and Walker, A.L., 2015. Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, 30, 1e9.
142. Lee, J.S., Jurkevich, L., Dewaele, P., Wambacq, P., and Oosterlinck, A., 1994. Speckle filtering of synthetic aperture radar images: A review. *Remote Sensing Reviews*, 8 (4), 313–340.
143. Lemcoff, J.H. and Loomis, R.S., 1986. Nitrogen influences on yield determination in maize. *Crop Science*, 26(5), 1017-1022.
144. Le Page, M., Jarlan, L., El Hajj, M.M., Zribi, M., Baghdadi, N. and Boone, A., 2020. Potential for the detection of irrigation events on maize plots using sentinel-1 soil moisture products. *Remote Sensing*, 12(10), 1621.
145. Lerman, P.M., 1980. Fitting segmented regression models by grid search. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(1), 77-84.
146. Leroux, L., Castets, M., Baron, C., Escorihuela, M.J., Bégué, A. and Seen, D.L., 2019. Maize yield estimation in West Africa from crop process-induced combinations of multi-domain remote sensing indices. *European Journal of Agronomy*, 108, 11-26.
147. Lewis, H.G. and Brown, M., 2001. A generalized confusion matrix for assessing area estimates from remotely sensed data. *International Journal of Remote Sensing*, 22 (16), 3223–3235.
148. Liang, S. and Wang, J. eds., 2019. Advanced remote sensing: terrestrial information extraction and applications. Academic Press.
149. Licciardi, G., Marpu, P.R., Chanussot, J. and Benediktsson, J.A., 2011. Linear versus nonlinear PCA for the classification of hyperspectral data based on the extended morphological profiles. *IEEE Geoscience and Remote Sensing Letters*, 9(3), 447-451.
150. Li, F., Miao, Y., Feng, G., Yuan, F., Yue, S., Gao, X., Liu, Y., Liu, B., Ustin, S.L., Chen, X., 2014. Improving Estimation of Summer Maize Nitrogen Status with Red Edge-Based Spectral Vegetation Indices. *Field Crops Research, 157*, 111–123.
151. Li, X., Chen, W., Cheng, X. and Wang, L., 2016. A comparison of machine learning algorithms for mapping of complex surface-mined and agricultural landscapes using ZiYuan-3 stereo satellite imagery. *Remote Sensing*, 8(6), 514.
152. Li, Y., Li, C., Li, M. and Liu, Z., 2019. Influence of variable selection and forest type on forest aboveground biomass estimation using machine learning algorithms. *Forests*, *10*(12), 1073.
153. Liu, H.L., Yang, J.Y., Drury, C.A., Reynolds, W.D., Tan, C.S., Bai, Y.L., He, P., Jin, J. and Hoogenboom, G., 2011. Using the DSSAT-CERES-Maize model to simulate crop yield and nitrogen cycling in fields under long-term continuous maize production. *Nutrient Cycling in Agroecosystems*, 89(3), 313-328.

© University of Pretoria

154. Liu, J., Huffman, T., Qian, B., Shang, J., Li, Q., Dong, T., Davidson, A. and Jing, Q., 2020. Crop yield estimation in the Canadian prairies using Terra/MODIS-derived crop metrics. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 2685-2697.
155. Liu, L., Xiao, X., Qin, Y., Wang, J., Xu, X., Hu, Y. and Qiao, Z., 2020. Mapping cropping intensity in China using time series Landsat and Sentinel-2 images and Google Earth Engine. *Remote Sensing of Environment*, 239, 111624.
156. Lobell, D.B. and Field, C.B., 2007. Global scale climate–crop yield relationships and the impacts of recent warming. *Environmental Research Letters*, 2(1), 014002.
157. Loew, A. and Mauser, W., 2007. Generation of geometrically and radiometrically terrain corrected SAR image products. *Remote Sensing of Environment*, 106 (3), 337–349.
158. López-Calderón, M.J., Estrada-Ávalos, J., Rodríguez-Moreno, V.M., Mauricio-Ruvalcaba, J.E., Martínez-Sifuentes, A.R., Delgado-Ramírez, G. and Miguel-Valle, E., 2020. Estimation of Total Nitrogen Content in Forage Maize (Zea mays L.) Using Spectral Indices: Analysis by Random Forest. *Agriculture*, 10(10), 451.
159. Ma, Y., Zhang, Z., Kang, Y. and Özdoğan, M., 2021. Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach. *Remote Sensing of Environment*, 259, 112408.
160. Madeira, J., Bedidi, A., Cervelle, B., Pouget, M. and Flay, N., 1997. Visible spectrometric indices of hematite (Hm) and goethite (Gt) content in lateritic soils: the application of a Thematic Mapper (TM) image for soil-mapping in Brasilia, Brazil. *International Journal of Remote Sensing*, 18(13), 2835-2852.
161. Manandhar, R., Odeh, I., and Ancev, T., 2009. Improving the Accuracy of Land Use and Land Cover Classification of Landsat Data Using Post-Classification Enhancement. *Remote Sensing*, 1(3), 330–344.
162. Mandal, U.K., 2016. Spectral color indices based geospatial modeling of soil organic matter in Chitwan district, Nepal. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 41.
163. Mansfield, E.R. and Helms, B.P., 1982. Detecting multicollinearity. *The American Statistician*, 36(3a), 158-160.
164. Mateo-Sanchis, A., Piles, M., Muñoz-Marí, J., Adsuara, J.E., Pérez-Suay, A. and Camps-Valls, G., 2019. Synergistic integration of optical and microwave satellite data for crop yield estimation. *Remote Sensing of Environment*, 234, 111460.
165. Matiu, M., Ankerst, D.P. and Menzel, A., 2017. Interactions between temperature and drought in global and regional crop yield variability during 1961-2014. *PloS one*, 12(5), e0178339.
166. Matsumura, K., Gaitan, C.F., Sugimoto, K., Cannon, A.J. and Hsieh, W.W., 2015. Maize yield forecasting by linear regression and artificial neural networks in Jilin, China. *The Journal of Agricultural Science*, 153(3), 399-410.
167. McNairn, H. and Brisco, B., 2004. The application of C-band polarimetric SAR for agriculture: A review. *Canadian Journal of Remote Sensing*, 30(3), 525-542.
168. McNairn, H., Champagne, C., Shang, J., Holmstrom, D., and Reichert, G., 2009. Integration of optical and Synthetic Aperture Radar (SAR) imagery for delivering operational annual crop inventories. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(5), 434–449.
169. McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157.
170. Mkhabela, M.S., Mkhabela, M.S. and Mashinini, N.N., 2005. Early maize yield forecasting in the four agro-ecological regions of Swaziland using NDVI data derived from NOAA's-AVHRR. *Agricultural and Forest Meteorology*, 129(1-2), 1-9.
171. Meng, L., Liu, H., L Ustin, S. and Zhang, X., 2021. Predicting Maize Yield at the Plot Scale of Different Fertilizer Systems by Multi-Source Data and Machine Learning Methods. *Remote Sensing*, 13(18), 3760.

© University of Pretoria

172. Merzlyak, M.N., Gitelson, A.A., Chivkunova, O.B. and Rakitin, V.Y., 1999. Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening. *Physiologia Plantarum*, 106(1), 135-141.
173. Millard, K. and Richardson, M., 2015. On the importance of training data sample selection in random forest image classification: A case study in peatland ecosystem mapping. *Remote Sensing*, *7*(7), 8489-8515.
174. Mirik, M., Ansley, R.J., Steddom, K., Jones, D., Rush, C., Michels, G., and Elliott, N., 2013. Remote Distinction of A Noxious Weed (Musk Thistle: CarduusNutans) Using Airborne Hyperspectral Imagery and the Support Vector Machine Classifier. *Remote Sensing*, 5(2), 612–630.
175. Misra, A.K., 2014. Climate change and challenges of water and food security. *International Journal of Sustainable Built Environment*, 3(1), 153–165.
176. Mitchell, T.M., 1997. Machine learning.
177. Miura, T., Huete, A.R. and Yoshioka, H., 2000. Evaluation of sensor calibration uncertainties on vegetation indices for MODIS. *IEEE Transactions on Geoscience and Remote Sensing*, 38(3), 1399-1409.
178. Mopani District Municipality (MDM), 2019. Integrated Development Plan 2016-2021. Mopani District Municipality, Limpopo.
179. Mountrakis, G., Im, J., and Ogole, C., 2011. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3), 247–259.
180. Mufungizi, A.A., Musakwa, W. and Gumbo, T., 2020. A Land Suitability Analysis of the Vhembe District, South Africa, the Case of Maize and Sorghum. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 1023-1030.
181. Ndikumana, E., Ho Tong Minh, D., Baghdadi, N., Courault, D., and Hossard, L., 2018. Deep Recurrent Neural Network for Agricultural Classification using multitemporal SAR Sentinel-1 for Camargue, France. *Remote Sensing*, 10(8), 1217.
182. Nobre, J. and Neves, R.F., 2019. Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets. *Expert Systems with Applications*, 125, 181-194.
183. Nyamangara, J., Mudhara, M. and Giller, K.E., 2005. Effectiveness of cattle manure and nitrogen fertilizer application on the agronomic and economic performance of maize. *South African Journal of Plant and Soil*, 22(1), 59-63.
184. Ogutu, G.E., Franssen, W.H., Supit, I., Omondi, P. and Hutjes, R.W., 2018. Probabilistic maize yield prediction over East Africa using dynamic ensemble seasonal climate forecasts. *Agricultural and Forest Meteorology*, 250, 243-261.
185. Olofsson, P., Foody, G. M., Stehman, S. V., and Woodcock, C. E., 2013. Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. *Remote Sensing of Environment*, 129, 122–131.
186. Orhun, G.E., 2013. Maize for life. *International Journal of Food Science and Nutrition Engineering*, 3(2), 13-16.
187. Osterholz, W.R., Rinot, O., Liebman, M. and Castellano, M.J., 2017. Can mineralization of soil organic nitrogen meet maize nitrogen demand?. *Plant and Soil*, 415(1), 73-84.
188. Otto, R., Castro, S.A.Q., Mariano, E., Castro, S.G.Q., Franco, H.C.J. and Trivelin, P.C.O., 2016. Nitrogen use efficiency for sugarcane-biofuel production: what is next?. *Bioenergy Research*, 9(4), 1272-1289.
189. Ouattara, B., Forkuor, G., Zoungrana, B.J., Dimobe, K., Danumah, J., Saley, B. and Tondoh, J.E., 2020. Crops monitoring and yield estimation using sentinel products in semi-arid smallholder irrigation schemes. *International Journal of Remote Sensing*, 41(17), 6527-6549.
190. Ouzemou, J.E., El Harti, A., Lhissou, R., El Moujahid, A., Bouch, N., El Ouazzani, R., Bachaoui, E.M. and El Ghmari, A., 2018. Crop type mapping from pansharpened

© University of Pretoria

Landsat 8 NDVI data: A case of a highly fragmented and intensive agricultural system. *Remote Sensing Applications: Society and Environment*, 11, 94-103.

191. Pal, M., 2005. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1), 217-222.
192. Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C. and Athanasiadis, I.N., 2021. Machine learning for large-scale crop yield forecasting. *Agricultural Systems*, 187, 103016.
193. Parry ML, Rosenzweig C, Iglesias A, Livermore M, Fischer G. 2004. Effects of climate change on global food production under SRES emissions and socio-economic scenarios. *Glob Environ Change*. 14(1):53–67.
194. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *The Journal of machine Learning Research*, 12, 2825-2830.
195. Pérez-Cruz, F., Van Vaerenbergh, S., Murillo-Fuentes, J.J., Lázaro-Gredilla, M. and Santamaria, I., 2013. Gaussian processes for nonlinear signal processing: An overview of recent advances. IEEE Signal Processing Magazine, 30(4), 40-50.
196. Pervez, M.S. and Brown, J.F., 2010. Mapping Irrigated Lands at 250-m Scale by Merging MODIS Data and National Agricultural Statistics. *Remote Sensing*, 2(10), 2388–2412.
197. Petropoulos, G.P., Kalaitzidis, C., and Prasad Vadrevu, K., 2012. Support vector machines and object-based classification for obtaining land-use/cover cartography from Hyperion hyperspectral imagery. *Computers & Geosciences*, 41, 99–107.
198. Piiroinen, R., Heiskanen, J., Mõttus, M., and Pellikka, P., 2015. Classification of crops across heterogeneous agricultural landscape in Kenya using AisaEAGLE imaging spectroscopy data. International *Journal of Applied Earth Observation and Geoinformation*, 39, 1–8.
199. Poffenbarger, H.J., Sawyer, J.E., Barker, D.W., Olk, D.C., Six, J. and Castellano, M.J., 2018. Legacy effects of long-term nitrogen fertilizer application on the fate of nitrogen fertilizer inputs in continuous maize. *Agriculture, Ecosystems & Environment*, *265*, 544-555.
200. Polly, J., Hegarty-Craver, M., Rineer, J., O'Neil, M., Lapidus, D., Beach, R., and Temple, D., 2019. The use of Sentinel-1 and -2 data for monitoring maize production in Rwanda. In: *Remote Sensing for Agriculture, Ecosystems, and Hydrology XXI*; Neale, C. M. and Maltese, A., Eds.; France: SPIE, 72.
201. Prasad, A.K., Chai, L., Singh, R.P. and Kafatos, M., 2006. Crop yield estimation model for Iowa using remote sensing and surface parameters. *International Journal of Applied Earth Observation and Geoinformation*, 8(1), 26-33.
202. Price, J.C., 1994. How unique are spectral signatures?. *Remote Sensing of Environment*, 49(3), 181-186.
203. Rangarajan, A.K., Purushothaman, R. and Ramesh, A., 2018. Tomato crop disease classification using pre-trained deep learning algorithm. *Procedia computer science*, 133, 1040-1047.
204. Ranum, P., Peña-Rosas, J.P., and Garcia-Casal, M.N., 2014. Global maize production, utilization, and consumption. *Annals of the New York Academy of Sciences*, 1312(1), 105–112.
205. Rembold, F., Atzberger, C., Savin, I. and Rojas, O., 2013. Using low resolution satellite imagery for yield prediction and yield anomaly detection. Remote Sensing, 5(4), 1704-1733.
206. Richard, C., 2015. *Water for a sustainable world*. The United Nations world water development report. Paris: UNESCO.
207. Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., and Rigol-Sanchez, J.P., 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67, 93–104.

© University of Pretoria

208. Rojas, O., 2007. Operational maize yield model development and validation based on remote sensing and agro-meteorological data in Kenya. *International Journal of Remote Sensing*, *28*(17), 3775-3793.
209. Rondeaux, G., Steven, M., and Baret, F., 1996. Optimization of soil-adjusted vegetation indices. *Remote Sensing of Environment*, 55(2), 95–107.
210. Roujean, J.-L. and Breon, F.-M., 1995. Estimating PAR absorbed by vegetation from bidirectional reflectance measurements. *Remote Sensing of Environment*, 51(3), 375–384.
211. Sahin, E. and Colkesen, I., 2019. Performance analysis of advanced decision tree-based ensemble learning algorithms for landslide susceptibility mapping. *Geocarto International*, 1–23.
212. Salager-Meyer, F., 2008. Scientific publishing in developing countries: Challenges for the future. *Journal of English for Academic Purposes*, 7(2), 121-132.
213. Satalino, G., Balenzano, A., Mattia, F. and Davidson, M.W., 2013. C-band SAR data for mapping crops dominated by surface or volume scattering. *IEEE Geoscience and Remote Sensing Letters*, *11*(2), 384-388.
214. Santpoort, R., 2020. The drivers of maize area expansion in Sub-Saharan Africa. How policies to boost maize production overlook the interests of smallholder farmers. *Land*, 9(3), 68.
215. Sapkota, T.B., Jat, M.L., Jat, R.K., Kapoor, P. and Stirling, C., 2016. Yield estimation of food and non-food crops in smallholder production systems. In *Methods for measuring greenhouse gas balances and evaluating mitigation options in smallholder agriculture* (163-174). Springer, Cham.
216. Scheunders, P., Tuia, D. and Moser, G., 2018. Contributions of machine learning to remote sensing data analysis. In *Data processing and analysis methodology* (199-243). Elsevier.
217. SDG. Sustainable Development Goals; United Nations: New York, NY, USA, 2019.
218. SDM, 2019. *Greater Sekhukhune Cross Border District Municipality Integrated Development Plan: 2019/20*; Limpopo, South Africa: SDM.
219. Shao, G., Han, W., Zhang, H., Liu, S., Wang, Y., Zhang, L. and Cui, X., 2021. Mapping maize crop coefficient Kc using random forest algorithm based on leaf area index and UAV-based multispectral vegetation indices. *Agricultural Water Management*, 252, 106906.
220. Shi, T., Cui, L., Wang, J., Fei, T., Chen, Y. and Wu, G., 2013. Comparison of multivariate methods for estimating soil total nitrogen with visible/near-infrared spectroscopy. *Plant and Soil*, 366(1), 363-375.
221. Sibley, A.M., Grassini, P., Thomas, N.E., Cassman, K.G. and Lobell, D.B., 2014. Testing remote sensing approaches for assessing yield variability among maize fields. *Agronomy Journal*, 106(1), 24-32.
222. Siebert, S.J., Van Wyk, A.E., Bredenkamp, G.J. and Siebert, F., 2003. Vegetation of the rock habitats of the Sekhukhuneland Centre of Plant Endemism, South Africa. *Bothalia Pretoria*, 33(2), 207-228.
223. Sinclair, T.R. and Muchow, R.C., 1995. Effect of nitrogen supply on maize yield: I. Modeling physiological responses. *Agronomy Journal*, 87(4), 632-641.
224. Skakun, S., Kalecinski, N.I., Brown, M.G., Johnson, D.M., Vermote, E.F., Roger, J.C. and Franch, B., 2021. Assessing within-Field Corn and Soybean Yield Variability from WorldView-3, Planet, Sentinel-2, and Landsat 8 Satellite Imagery. *Remote Sensing*, 13(5), 872.
225. Skiena, S.S., 2017. Machine Learning. In: *The Data Science Design Manual*; Skiena, S.S., Ed.; Texts in Computer Science; Cham, Germany: Springer International Publishing, 351–390.
226. Smith, P., Martino, D., Cai, Z., Gwary, D., Janzen, H., Kumar, P., McCarl, B., Ogle, S., O'Mara, F., Rice, C. and Scholes, B., 2008. Greenhouse gas mitigation in agriculture.

© University of Pretoria

Philosophical transactions of the royal Society B: Biological Sciences, 363(1492), 789-813.

227. Son, N.T., Chen, C.F., Chen, C.R. and Minh, V.Q., 2017. Assessment of Sentinel-1A data for rice crop classification using random forests and support vector machines. *Geocarto International*, 33(6), 587-601.
228. Sonobe, R., Tani, H., Wang, X., Kobayashi, N. and Shimamura, H., 2014. Parameter tuning in the support vector machine and random forest and their performances in cross- and same-year crop classification using TerraSAR-X. *International Journal of Remote Sensing*, 35(23), 7898-7909.
229. Sonobe, R., Yamaya, Y., Tani, H., Wang, X., Kobayashi, N. and Mochizuki, K.I., 2017. Assessing the suitability of data from Sentinel-1A and 2A for crop classification. *GIScience & Remote Sensing*, 54(6), 918-938.
230. Sonobe, R., Yamaya, Y., Tani, H., Wang, X., Kobayashi, N. and Mochizuki, K.I., 2018. Crop classification from Sentinel-2-derived vegetation indices using ensemble learning. *Journal of Applied Remote Sensing*, 12(2), 026019.
231. Sörensen, R., Zinko, U. and Seibert, J., 2006. On the calculation of the topographic wetness index: evaluation of different methods based on field observations. *Hydrology and Earth System Sciences*, *10*(1), 101-112.
232. Sorenson, P.T., Small, C., Tappert, M.C., Quideau, S.A., Drozdowski, B., Underwood, A. and Janz, A., 2017. Monitoring organic carbon, total nitrogen, and pH for reclaimed soils using field reflectance spectroscopy. *Canadian Journal of Soil Science*, 97(2), 241-248.
233. Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, *106*(D7), 7183-7192.
234. Thanh Noi, P. and Kappas, M., 2018. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. *Sensors*, 18(1), 18.
235. Tong, X.-Y., Xia, G.-S., Lu, Q., Shen, H., Li, S., You, S., and Zhang, L., 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, 237, 111322.
236. Torbick, N., Chowdhury, D., Salas, W. and Qi, J., 2017. Monitoring rice agriculture across myanmar using time series Sentinel-1 assisted by Landsat-8 and PALSAR-2. *Remote Sensing*, 9(2), 119.
237. Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., Traver, I. N., Deghaye, P., Duesmann, B., Rosich, B., Miranda, N., Bruno, C., L'Abbate, M., Croci, R., Pietropaolo, A., Huchler, M., and Rostan, F., 2012. GMES Sentinel-1 mission. *Remote Sensing of Environment*, 120, 9–24.
238. Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of Environment*, 8(2), 127-150.
239. Useya, J. and Chen, S., 2019. Exploring the Potential of Mapping Cropping Patterns on Smallholder Scale Croplands Using Sentinel-1 SAR Data. *Chinese Geographical Science*, 29(4), 626-639.
240. Van Tricht, K., Gobin, A., Gilliams, S. and Piccard, I., 2018. Synergistic use of radar Sentinel-1 and optical Sentinel-2 imagery for crop mapping: a case study for Belgium. *Remote Sensing*, 10(10), 164.
241. Van Wart, J., Kersebaum, K.C., Peng, S., Milner, M. and Cassman, K.G., 2013. Estimating crop yield potential at regional to national scales. *Field Crops Research*, 143, 34-43.
242. Veloso, A., Mermoz, S., Bouvet, A., Le Toan, T., Planells, M., Dejoux, J.F. and 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.
243. Vhembe District Municipality (VDM), 2015. 2016/17 IDP Review final draft. Vhembe District Municipality, Limpopo.

© University of Pretoria

244. Vreugdenhil, M., Wagner, W., Bauer-Marschallinger, B., Pfeil, I., Teubner, I., Rüdiger, C. and Strauss, P., 2018. Sensitivity of Sentinel-1 backscatter to vegetation dynamics: An Austrian case study. *Remote Sensing*, 10(9), 1396.
245. Wan, S. and Chang, S.-H., 2019. Crop classification with WorldView-2 imagery using Support Vector Machine comparing texture analysis approaches and grey relational analysis in Jianan Plain, Taiwan. *International Journal of Remote Sensing*, 40(21), 8076–8092.
246. Wang, Q., Blackburn, G.A., Onojeghuo, A.O., Dash, J., Zhou, L., Zhang, Y. and Atkinson, P.M., 2017. Fusion of Landsat 8 OLI and Sentinel-2 MSI data. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7), 3885-3899.
247. Wang, S., Adhikari, K., Wang, Q., Jin, X. and Li, H., 2018. Role of environmental variables in the spatial distribution of soil carbon (C), nitrogen (N), and C: N ratio from the northeastern coastal agroecosystems in China. *Ecological Indicators*, 84, 263-272.
248. Wang, S., Di Tommaso, S., Deines, J.M. and Lobell, D.B., 2020. Mapping twenty years of corn and soybean across the US Midwest using the Landsat archive. *Scientific Data*, 7(1), 1-14.
249. Waterberg district municipality (WDM), 2015. Waterberg district municipality draft 2016/17 IDP. Waterberg district municipality, Limpopo.
250. Wei, Z., Meng, Y., Zhang, W., Peng, J. and Meng, L., 2019. Downscaling SMAP soil moisture estimation with gradient boosting decision tree regression over the Tibetan Plateau. *Remote Sensing of Environment*, 225, 30-44.
251. Whelen, T. and Siqueira, P., 2018. Time-series classification of Sentinel-1 agricultural data over North Dakota. *Remote Sensing Letters*, 9(5), 411-420.
252. Whisler, F.D., Acock, B., Baker, D.N., Fye, R.E., Hodges, H.F., Lambert, J.R., Lemmon, H.E., McKinion, J.M. and Reddy, V.R., 1986. Crop simulation models in agronomic systems. *Advances in agronomy*, 40, 141-208.
253. Williams, C.K. and Barber, D., 1998. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12), 1342-1351.
254. Wolpert, D.H., 1992. Stacked generalization. *Neural Networks*, 5(2), 241–259.
255. Woodhouse, I.H., 2017. *Introduction to Microwave Remote Sensing*, Florida, USA: CRC Press.
256. Wu, F., Wang, C., Zhang, H., Zhang, B. and Tang, Y., 2010. Rice crop monitoring in South China with RADARSAT-2 quad-polarization SAR data. *IEEE Geoscience and Remote Sensing Letters*, 8(2), 196-200.
257. Wu, S., Li, J. and Huang, G.H., 2008. A study on DEM-derived primary topographic attributes for hydrologic applications: Sensitivity to elevation data resolution. *Applied Geography*, 28(3), 210-223.
258. Yang, H.S., Dobermann, A., Lindquist, J.L., Walters, D.T., Arkebauer, T.J. and Cassman, K.G., 2004. Hybrid-maize—a maize simulation model that combines two crop modeling approaches. *Field Crops Research*, 87(2-3), 131-154.
259. Yang, J., Gong, W., Shi, S., Du, L., Sun, J. and Song, S.L., 2016. Estimation of nitrogen content based on fluorescence spectrum and principal component analysis in paddy rice. *Plant, Soil and Environment*, 62(4), 178-183.
260. Yang, M., Wang, G., Lazin, R., Shen, X. and Anagnostou, E., 2021. Impact of planting time soil moisture on cereal crop yield in the Upper Blue Nile Basin: A novel insight towards agricultural water management. *Agricultural Water Management*, *243*, 106430.
261. Yao, F., Tang, Y., Wang, P. and Zhang, J., 2015. Estimation of maize yield by using a process-based model and remote sensing data in the Northeast China Plain. *Physics and Chemistry of the Earth, Parts A/B/C*, 87, 142-152.
262. Yi, Z., Jia, L., and Chen, Q., 2020. Crop Classification Using Multi-Temporal Sentinel-2 Data in the Shiyang River Basin of China. *Remote Sensing*, 12(24), 4052.
263. Xu, Y., Smith, S.E., Grunwald, S., Abd-Elrahman, A., Wani, S.P. and Nair, V.D., 2018. Estimating soil total nitrogen in smallholder farm settings using remote sensing spectral indices and regression kriging. *Catena*, 163, 111-122.

© University of Pretoria

264. Zemel, R.S. and Pitassi, T., 2001. A gradient-based boosting algorithm for regression problems. *Advances in Neural Information Processing Systems*, 696-702. Ouyang, L., 2019. Mapping stocks of soil total nitrogen using remote sensing data: A comparison of random forest models with different predictors. *Computers and Electronics in Agriculture*, 160, 23-30.
265. Zhao, H., Chen, Z., Jiang, H., Jing, W., Sun, L. and Feng, M., 2019. Evaluation of three deep learning models for early crop classification using sentinel-1A imagery time series—A case study in Zhanjiang, China. *Remote Sensing*, 11(22), 2673.
266. Zhong, L., Hu, L., and Zhou, H., 2019. Deep learning based multi-temporal crop classification. *Remote Sensing of Environment*, 221, 430–443.
267. Zhou, T., Pan, J., Zhang, P., Wei, S. and Han, T., 2017. Mapping winter wheat with multi-temporal SAR and optical images in an urban agricultural region. *Sensors*, 17(6), 1210.
268. Zhang, H., Kang, J., Xu, X. and Zhang, L., 2020. Accessing the temporal and spectral features in crop type mapping using multi-temporal Sentinel-2 imagery: A case study of Yi'an County, Heilongjiang province, China. *Computers and Electronics in Agriculture*, 176, 105618.
269. Zhang, L., Zhang, Z., Luo, Y., Cao, J. and Tao, F., 2020. Combining optical, fluorescence, thermal satellite, and environmental data to predict county-level maize yield in China using machine learning approaches. *Remote Sensing*, 12(1), 21.
270. Zhou, T., Geng, Y., Chen, J., Sun, C., Haase, D. and Lausch, A., 2019. Mapping of Soil Total Nitrogen Content in the Middle Reaches of the Heihe River Basin in China Using Multi-Source Remote Sensing-Derived Variables. *Remote Sensing*, 11(24), 2934.
271. Zhou, T., Geng, Y., Chen, J., Pan, J., Haase, D. and Lausch, A., 2020. High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, Sentinel-1 and Sentinel-2 data based on machine learning algorithms. *Science of The Total Environment*, 729, 138244.
272. Zougmoré, R.B., Partey, S.T., Ouédraogo, M., Torquebiau, E. and Campbell, B.M., 2018. Facing climate variability in sub-Saharan Africa: analysis of climate-smart agriculture opportunities to manage climate-related risks. *Cahiers Agricultures (TSI)*, 27(3), 1-9.

© University of Pretoria