Mapping Arable Field Fractions with Muultisensor Remote Sensing Data-Driven Gradient Boosted and Classical GAMs

YEBELAY GONFA GRAGN

July, 2021

SUPERVISORS:

Dr. M.T. Marshall

Dr. C.A.J.M. De Bie

Mapping Arable Field Fractions with Multisensor Remote Sensing Data-Driven Gradient Boosted and Classical GAMs

YEBELAY GONFA GRAGN

Enschede, The Netherlands, July, 2021

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformatics

SUPERVISORS:

Dr. M.T. Marshall

Dr. C.A.J.M. De Bie

THESIS ASSESSMENT BOARD:
Dr. A. Vrieling (Chair)
Dr. M. Belgiu (External Examiner, University of Twente)

ABSTRACT

Information on crop production estimates is the basis for supporting the current and future food security initiatives, especially for developing countries. However, for most developing countries obtaining crop production estimates is a challenge due to several reasons. One of the reasons is a challenge in identifying and extracting information about the extent and location of agricultural areas. These agriculture areas have different characteristics that make them challenging to quantify their extent. For instance, in Ethiopia's Oromia region, the arable fields are characterized by small size, irregular shape, often trees inside fields, irregular cropping patterns, and heterogeneity in weather conditions. These challenges increase the uncertainty in the delineation of the field extent. In this research, a combined method is developed to map arable field fractions with opensource earth observation data that minimize the uncertainty in estimating the field extent. Gradient boosted and Classical GAM models used with Sentinel-1 backscatter matrices, Sentinel-2 optical, topographic features, and hyper temporal images (i.e., Proba-V). The hyper-temporal imagery is used primarily to extrapolate a 1km NDVI (i.e., a 1km arable field fraction map is extrapolated and used as an input variable for the model). The hyper-temporal images are also used for identifying the wet and dry seasons for downloading Sentinel-1&2 image features. Eight Sentinel-1 image features (i.e., dry and wet season VV (Vertical transmit, Vertical receive), VH (Vertical transmit, Horizontal receive), VV/VH ratio, and NRMP) and eleven Sentinel-2 optical image features (i.e., three red-edge bands, 2 SWIR, two dry and wet season NDVI, two dry and wet season Normalized Difference Tillage Index (NDTI), and two dry and wet season Land Surface Wetness Index (LSWI)) are used in the model. In addition to Sentinel-1&2 image features, topographic variables (i.e., elevation, slope, relative DEM, and topographic wetness index) are included in the model. Gradient boosted regression is used to select the most important predictor variables, and the Classical GAM is used to predict arable field fractions from these important predictor variables. Based on the boosted GAM model and stability selection, six informative variables (i.e., dry season VH, elevation, red edge (Band-5), dry season VV/VH ratio, Slope, and a 1km arable field estimate) out of twenty-four explanatory variables are selected. The overall deviance of the model was 87%. The partial deviance explained by Sentinel-1 dry season VH was 33.3% which is the most explanatory variable in discriminating arable field fractions. The partial deviance of elevation, Band-5, 1km arable field fractions, slope and dry season VV/VH ratio was 17.2%, 14.4%,13.3%,6.2% and 3.34% respectively. Classical GAM is fitted with the most informative variables selected using the Gradient boost and stability selection method. Finally, a 20m arable field fraction map was extrapolated for the Oromia region. The developed method can be applied to extrapolate 20m arable field fractions for the rest regional states of Ethiopia and country-level wall-to-wall mapping by considering agroecological variations.

Keywords: Hyper temporal Image, Sentinel-1 backscatter matrices, Sentinel-2 optical, Informative variables, Extrapolation

ACKNOWLEDGEMENTS

First of all, I would like to thank the almighty God for His mercies and blessings.

I want to thank my first advisor, Dr. Michael Marshal, for his continuous support, patience, and guidance in every aspect of the research, especially when I have questions about the modeling process. He also gave me an insightful comment, and his invaluable guidance helped me to write this thesis. I want to thank also my second supervisor, Dr.Kees De Bie, for his patience, unreserved support, and motivation have deeply encouraged me to finish my thesis. He is supportive in guiding me in every aspect of this project especially related to the data issues. During this Covid 19, I faced some problems, and he is the one who gave me the first call. I would never forget his saying, "Take your bicycle, just go outside and relax." I also want to thank Orange Knowledge Program (OKP) for covering the full cost of my study and enabling me to join the University of Twente, one of the world's best geospatial science institutes. Last but not least, I would like to thank my love, Mare, for her support and encouragement.

TABLE OF CONTENTS

Contents

LIST OF ACRONYM AND ABBREVIATION

ACMA	Automated Cropland Mapping Algorithm
AIC	Akaike Information Criteria
ANN	Artificial Neural Network
AUC	Area Under the Curve
CPZ	Crop Production Zone
CSA	Central Statistics Agency
DEM	Digital Elevation Model
ERDAS	Earth Resource Data Analysis System
ESA	European Space Agency
ETM+	Enhanced Thematic Mapper Plus
EVI	Enhanced Vegetation Index
FAO	Food and Agriculture Organization
FEWS NET	Famine Early Warning Systems Network
GAM	Generalized Additive Model
GBR	Gradient Boosted Regression
GDP	Gross Domestic Product
GEE	Google Earth Engine
GLC	Global Land Cover
IDL ENVI	Interactive Data Visualization Environmental for Visualizing Images
IDW	Inverse Distance Weight
IHSN	International Household Survey Network
ISO	International Organization for Standardization
LSWI	Land Surface Water Index
MODIS	Moderate Resolution Imaging Spectroradiometer
NDTI	Normalized Difference Tillage Index
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
NIR	Near-Infrared
NRPB	Normalized Procedure Ratio between Bands
OCHA	Office for the Coordination of Humanitarian Affairs
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
RSS	Residual Sum of Squares
SAR	Synthetic Aperture Radar
SDG	Sustainable Development Goals
SPSS	Statistical Package for the Social Sciences
SRTM	Shuttle Radar Topography Mission
SVM	Support Vector Machines
SWIR	Short-Wave Infrared
TWI	Topographic Wetness Index
UNDP	United Nations Development Programme
VH	Vertical transmit Horizontal receive
VV	Vertical transmit Vertical receive

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Background Information and Justification

Information on Crop Production estimates at district to national levels is a base for supporting the current and future planning of food security initiatives (See et al., 2015). Agricultural information like arable land area, yield, and crop production estimates are crucial and the backbone for the development of the agricultural sector (FAO, 2017). This information is vital to Ethiopia, with 12 million smallholder farmers that contribute to 95 percent of the country's agricultural production (FAO, 2018). Most of the agricultural products are produced by smallholder farmers (Getahun, 2020). Ethiopia's estimated total population is 100.5 million in 2015 (UNDP, 2018), and the economy depends on the agricultural sector, specifically on rainfed agriculture and smallholder farming (Demeke & Ferede, 2004). Therefore, to increase resilience in the economy and increase agricultural productivity, reliable agricultural information routinely plays a significant role.

Agriculture production determines food availability, one of the five pillars of food security (Muzari, 2016). According to Muzari (2016), reducing poverty and food insecurity and increasing agricultural productivity have positive relationships. For instance, Thirtle & Piesse. (2003) indicate that the increment of 1% in crop yields reduces the number of economically poor people by 0.72% in Sub-Saharan Africa. Additionally, Jenkins (2005) finds a positive relationship between cereal crops and GDP per capita in developing countries. Therefore, agricultural productivity has a significant link with food security and depends on agricultural areas' productivity.

Identifying and extracting information about the extent and location of agricultural areas are the fundamental steps for productive and sustainable agriculture (See et al., 2013). Under SDG (Sustainability Development Goal) 2.4.1, "the cropland within productive and sustainable agriculture are those farms that support the sustainability of the three dimensions (i.e., economic, social, and environmental dimensions)" (FAO, 2020b). The SDG 2.4.1 critically emphasized the importance of measuring the proportion of extent of land under sustainable agriculture and the overall extent of agricultural land area. Therefore, it is crucial to measure the countries agricultural area (i.e., arable field) to achieve sustainability in agriculture.

The term arable field follows the definition of arable land (FAO, 2016), which states that" *land under temporary crops (double-cropped areas counted only), and land under temporarily fallow (under less than five years).*" After identifying the arable field, we can estimate the agricultural production, which is the actual yield per crop area (i.e., arable field) with the unit of Kgm^{-2} (Xiong et al., 2016). The authors indicate the essentiality of extracting arable fields to provide accurate arable field extent as baseline information, and this information is crucial for most African countries due to the absence of high-resolution crop land products.

In Sub-Saharan Africa, arable fields size is very small (i.e., <2ha), there is also heterogeneity and sometimes indistinct field patterns (Debats et al., 2016). In Ethiopia, arable fields have different characteristics such as

an irregularity in crop calendar practice, higher topographic variation, non-contrasting irregular fields, and heterogeneity in weather conditions over shorter distances (Mohammed et al., 2020).

In the Oromia region, Ethiopia, agricultural information obtained from the Central Statistics Agency (CSA), and the information was collected in 2002 by the traditional way of surveying techniques (IHSN, 2006); it is still challenging to conduct agricultural censuses throughout the country. Furthermore, traditional techniques are expensive, time-consuming, and labor-intensive (Marshall et al., 2019). For instance, the central statistics agency provides district-level information about the main agricultural crop products such as Teff (i.e., local crop name), barley, wheat, maize, and sorghum (Taffese et al., 2013). The expensiveness, labor intensiveness, and data type characteristics become a challenge to access timely and reliable agricultural information.

The challenges in obtaining agricultural information can be acquired using the current data acquisition technology like remote sensing and mapping through using potential machine learning techniques. Different remote sensing data can be used for arable field mapping, and we have to choose which remote sensing data can fit our purposes. For instance, moderate resolution (20-30m) lacks some capability to map smallholder farms (McCarty et al., 2017). A high-resolution image can alleviate the coarse and moderate resolution image's feature identification problem, which is mixed pixel-spatial heterogeneity, and still, only a few scholars evaluate their effectiveness in mapping small fields (Crommelinck et al., 2016). According to Persello et al. (2019), accurate information of agricultural boundaries is possible using very high resolution (<5m) images of worldview 2/3. Generally, it's advantageous to consider the temporal, spatial, and spectral characteristics of remote sensing data to map the arable field. It is possible to use a different satellite image to get spectral, spatial, and temporal advantages.

In general, Mapping arable fields requires a combination of remote sensing data with spatial and temporal characteristics additional to mapping methods(Haack & Bechdol, 1999). The current study focuses on the use of multisensor data like Sentinel-1 (i.e., backscatter matrices), sentinel-2 (i.e., red edge bands), topographic features (i.e., elevation, slope, Topographic Wetness Index), and 1km Proba-V NDVI. A 1km Proba-V is used to differentiate the wet and dry season periods and to prepare a 1km arable field estimate which is the important variable for predicting a 20m arable field fractions. The main advantage of using Sentinel-1 image is obtaining cloud-free imageries (ESA, 2014). The sentinel-2 optical data is also important because the dry season red edge band can discriminate the vegetation cover from other land-use features (i.e., arable fields, water bodies, and urban land use) (Sun et al., 2020). Topographic data is used in this study because arable fields have strong relationships with topographic features (i.e., elevation and slope) (Husak et al., 2008).

In this research, two GAMs models are used. The first one is gradient boosted GAM for selecting the most influential input variables. The second one is Classical GAM used for predicting a 20m arable field probability estimate. The classical GAM has flexibility in the statistical distribution of the data (Murase et al., 2009) and uses quasibinomial distribution. The quasibinomial distribution considers overdispersion (i.e., the variability of the input datasets, especially if there are zero and one value in the input data set) (Elder et al., 1999). The research explores the potential of Sentinel-1, Sentinel-2, and topographic features in estimating 20m arable field fractions in the Oromia region, Ethiopia. Still know the regional level a 20m arable field fractions are not extrapolated by using sentine1 microwave, sentinel-2 optical, and topographic features for the Oromia

region. The developed method with multisensor remote sensing data can be applied to extrapolate 20m arable field fractions for the rest regional states' of Ethiopia as well as country-level wall-to-wall mapping of arable fields by considering agroecological zones and farming practices.

1.2. Research Objectives and Questions

1.2.1. General Objective

The general objective is to map a 20m arable field probabilities in Ethiopia's fragmented landscape using Sentinel-1 microwave, Sentinel-2 optical, and a combination of Boosted and Classical GAM models.

1.2.2. Specific Objectives and Questions

Subobjective 1: To stratify the landscape along with general crop phenological cycles or crop productions (CPZs) with 1km Proba-V NDVI

Subobjective 2: To estimate coarse (1km) resolution arable field fractions by integrating the a1km NDVI classes with agricultural statistics data.

Subobjective 3: To determine the relative importance of Sentinel-1 (i.e., Backscatter matrices), Sentinel-2 (i.e., Red edge & SWIR Bands), a 1km field fraction estimate and topographic features in estimating a 20m arable field fractions by using Gradient Boosted Generalized Additive Model.

Q 1. What percentage of field fraction variability can be explained by using a 1km field fraction estimates?

Q 2. What is the relative importance of Sentinel-1 backscatter matrices in estimating 20m arable field probabilities?

Q 3. What is the relative importance of Sentinel-2 Red edge bands, dry and wet season NDVI in estimating 20m arable field fractions?

Q 4. Do topographic features outperform other predictor variables in estimating 20m arable field fractions?

Subobjective 4: To evaluate the Gradient Boosted and classical GAM model in predicting arable field Fractions using out-of-sample data.

Q 5. Does the Gradient Boosted outperform Classical GAM in predicting arable field fractions using out-of-sample data?

1.3. Research Hypothesis

1.3.1. H1 = Sentinel-1 backscatter matrices (i.e., dry season VH and VV/VH ratio) have a relatively higher importance in estimating 20m arable field fractions

H0 = sentinel-1 backscatter matrices have relatively low importance in estimating 20m arable field fractions.

1.3.2. H1 = Sentinel-2 dry season NDVI variable have a relatively higher importance in estimating 20m arable field fractions.

H0 = Sentinel-2 dry season NDVI variable have relatively lower importance in estimating 20m arable field fractions.

1.3.3. H1 = Sentinel-2 wet season NDVI variable have a relatively higher importance in estimating 20m arable field fractions.
H0 = Sentinel-2 wet season NDVI variables have relatively lower importance in estimating a 20m arable field fractions.

1.3.4. H1= The feature importance rank of red edge bands of sentinel-2 is higher than the overall model variables.

H0= The feature importance rank of red edge bands of sentinel-2 is lower than the overall model variables.

1.3.5. H1= elevation and slope have relatively a higher feature importance rank in the estimation of 20m arable field probabilities.

H0= elevation and slope have relatively a lower feature importance rank for the estimation of 20m arable field fractions.

1.3.6. H1 = The Gradient Boosted outperform Classical GAM in predicting 20m arable field fractions.
H0 = The Boosted and Classical GAM effectively estimate 20m arable field fractions within tolerable R^2 and AUC.

2. LITERATURE REVIEW

2.1 Multisensor Remote Sensing data

Remote sensing is one of the current technologies that offer techniques in acquiring information related to agriculture, especially for providing agricultural information (Huang et al., 2018). Remote sensing has an advantage over the traditional way of acquiring information related to agriculture(Xie et al., 2008). The authors indicated one of the advantages of remote sensing is providing large spatial and repetitive coverage datasets. The other advantage of remote sensing is the accessibility of freely available datasets. Obtaining vital information from different spatial, spectral, and temporal resolutions requires integrating multisensor remote sensing data(Kulo, 2020). The aim of multisensor remote sensing data integration is to combine various remote sensing data from different sources with different spatial, temporal, and radiometric resolutions to deliver timely and reliable information is crucial for various mapping (Pastorino et al., 2021). Therefore, by integrating multisensory remote sensing data, we can identify and collect information about the arable field and estimate the total crop production. In this paper, multisensor remote sensing data like Hyper-temporal image, Sentinel-1 Microwave, Sentinel-2 optical, and topographic data are used to estimate 20m arable field fractions.

2.1.1 Hyper-temporal Image Analysis

A hyper-temporal image is one of the hyper-temporal data that can be collected from fine temporal resolution (Scarrott et al., 2019). Hyper-temporal image is used to map potential land cover, for instance, for large areas cropland mapping and help to differentiate forest, pasture, and shrublands (Craig, 2001). Hypertemporal image analysis depends on the time-based spectral originated from large observational dates like hyper temporal image data that uses profile patterns to classify vegetation (strata) over a landscape (De Bie et al., 2008). For instance, identifying agricultural areas with different vegetation classes and cropland that follow different crop calendars is possible using hyper-temporal images (De Bie et al., 2008). A good example of a hyper-temporal image is Proba-V, a global vegetation monitoring satellite with a 100m to 1km spatial resolution. Proba-V provides information on a 10-day temporal basis (Taffese et al., 2013). Identifying agroecological zones is possible based on a careful study of the temporal profile. Hyper temporal images like Proba-V, MODIS, and Spot vegetation enable us to understand the temporal dynamics by a trade of the spatial resolution (i.e., a loss regarding spatial resolution). Hyper temporal images exist in a courser resolution (>=250) and inefficient in identifying small to medium agricultural fields. Identifying fields is challenging due to the mixed pixel effect (Taffese et al., 2013). Not only courser resolution satellite imagery (e.g., Proba-V and MODIS) are used to produce land cover maps using the temporal advantage, and moderate resolution satellites like Landsat 8 use the time series of NDVI for mapping crop phenology (Salik & Karacabey, 2019).

2.1.2 Sentinel-1 Microwave

Sentinel-1 is a constellation of two satellites, namely Sentinel-1A and Sentinel-1B (Filgueiras et al., 2019). The temporal resolution of sentinel-1 is six days, and it is one of the types of SAR (Synthetic Aperture Radar) sensors that provide time series backscatter matrices. "Backscatter is the portion of the outgoing radar signal that the target redirects directly back towards the radar antenna" (ESA, 2017). According to Nasirzadehdizaji et al. (2021), the backscatter signal is returned to the radar mostly influenced by vegetation and soil properties. In general, the radar data mostly used for agricultural purposes because of the penetration capacity of the radar signal into vegetation canopy and other natural features (Fawwaz et al., 2015). Vreugdenhil et al. (2020) also show the properties of backscatter matrices with the characteristics of vegetation and bare soil (Figure 1). The authors indicated the horizontal backscatter (H) in vegetation greater than the backscatter in bare soils (Figure 1b). In VV (Vertical transmit, Vertical receive) polarization, the energy is equally scattered in all directions, and we can say that volume scattering is much higher in vegetation (Figure 1c). When energy is scattered in all directions, the relationship between the backscatter and the incidence angle becomes flat (Figure 1d). When the incidence angle increases, the backscatter volume will decrease (Figure 1a), and the relationship between the backscatter matrices and the incidence angle becomes steeper.

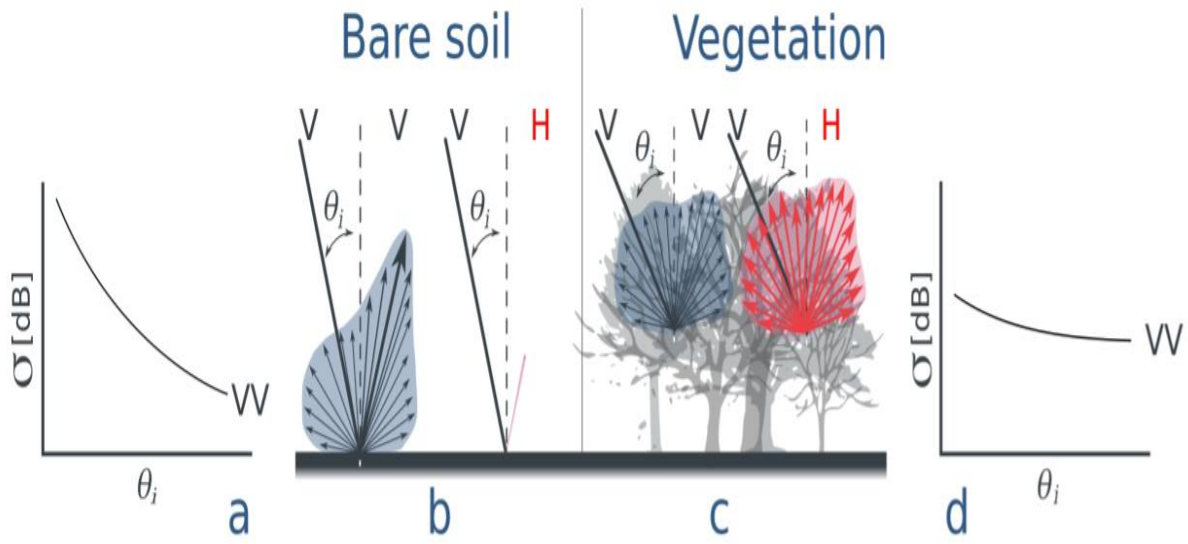


Figure 1 The relationship of backscatter with bare soil and vegetation.

Source: Adapted from Vreugdenhil et al. (2020)

Sentinel-1 also provides information and potentially discriminates crop types by considering the backscatter intensity. For instance, SAR intensity in VH polarization can differentiate crop types (Sun et al., 2020). The single and dual-polarization mode of Sentinel-1 backscatter matrices has different application areas and different backscatter responses from the land features (Table 1).

Table 1 Backscatter matrices and their application areas

Backscatter matrices Name	Spatial resolution/Exporting scale in GEE (m)	Description	Application and backscatter matrices characteristics with the land features
HH Single Co-polarization	10/20	Horizontal transmit/Horizontal receive	Mapping flooded vegetation and water classification (López-Caloca et al., 2018)
HV Dual-band Cross-polarization	10/20	Horizontal transmit/vertical receive	Applied on the areas which have weak backscatter signals, like sea surfaces (López-Caloca et al., 2018)
VV Single Co-polarization	10/20	Vertical transmit/ Vertical receive	Very sensitive to crop structure (Lemoine, 2018) and very sensitive to crop phenological stages (Nasirzadehdizaji et al., 2021a).
VH Dual-band Cross-polarization	10/20	Vertical transmit/Horizontal receive	VH Increase with increasing of the vegetation cover and vice versa (Lemoine, 2018) and VH increase with leaf development of crop due to increase in volume scattering (Khabbazan et al., 2019)
VV/VH	20	Backscatter ratio	Important for pasture classification (Nicolau et al., 2021)

The looking angle of a radar image also has a relationship with soil moisture, residue moisture, and residue cover (Mc Nairn et al., 2001). The authors indicated that the look angle radar also has different responses for different crop residues; for instance, the backscatter has different responses for barely and corn residues. The study used the backscatter signal as the response variable and the soil moisture, crop residue moisture, and crop cover as independent variables (Table 2). According to Mc Nairn et al. (2001), the backscatter (i.e., C-VH) distinguishes the crop residue moisture and crop cover within the range of looking angle from 30^0 up to 50^0. This looking angle of the backscatter is similar to the Sentinel-1 incidence angle range (i.e., 29.1^0-46^0 with the interferometric wide swath mode) (ESA, 2021).

Table 2 Backscatter matrices and their relationship with Soil moisture, Residue moisture, and Cover.

Multiple regression results for corn residue plots					
Scatterometer Configuration		Multiple Regression Coefficient (R)	Independent Variables		
			Soil Moisture (0-3 cm)	Residue Moisture	Residue Cover
Look Direction Parallel to Residue Row Direction					
C-HH	20	0.448			✓ *
	30	0.678	✓	✓ *	
	40	0.833	✓	✓ *	
	50	0.813	✓	✓ *	
C-VV	20	0.764		✓	✓ *
	30	0.714	✓	✓ *	
	40	0.746		✓ *	
	50	0.763		✓ *	
C-VH	20	0.838	✓	✓ *	
	30	0.843	✓	✓ *	
	40	0.887	✓	✓ *	✓
	50	0.865	✓	✓ *	✓
Multiple regression results for barley residue plots					
Look Direction Parallel to Residue Row Direction					
C-HH	20	0.701	✓	✓	
	30	0.758			
	40	NS	✓		
	50	0.499	✓		✓ *
C-VV	20	0.717	✓ *	✓	✓
	30	0.827	✓ *	✓	✓
	40	0.752	✓ *	✓	
	50	0.825	✓ *	✓	
C-VH	20	0.741	✓ *	✓	
	30	0.777	✓	✓ *	✓
	40	0.751	✓	✓ *	
	50	0.745	✓ *	✓	

✓ Indicates significance at p value<.05

* Indicates the largest contribution

Source: Adapted from Mc Nairn et al., (2001)

2.1.3 Sentinel-2 Optical

Sentinel-2 optical data has thirteen bands: one coastal aerosol, three visible, Red edge, NIR, and SWIR bands (He & Yokoya, 2018) and accessed with a different spatial resolution (Table 3). Sentinel 2 satellite data are a potential source of arable field mapping. According to Gumma et al. (2020), arable field areas of different crop types mapping are possible using a sentinel 2 NDVI time series. In addition to the NDVI index, another index can be calculated from the sentinel-2 image; for instance, NDTI is vital for crop type classification and improve classification accuracy (Zhang et al., 2020). Arable field mapping even possible by using the spectral bands of Sentinel-2 (i.e., Band 11 &12) and its spectral indices (i.e., NDWI) (Sun et al., 2020). Band 11 and 12 are also essential in fine crop classification (Zhang et al., 2020a). The use of all available sentinel-2 bands gives more information and improves classification accuracy (Qiu et al., 2017). Based on the above evidence from the literature, Sentinel-2 Red and NIR are selected for creating vegetation indices. In addition to this, Red edge bands and SWIR bands are used as additional input variables.

Table 3 Sentinel 2 image characteristics (the bands with the red colors are used in this study)

Spectral Band/Names	Center Wavelength (mm)	Band width (mm)	Spatial resolution
B1 - Coastal aerosol	443	20	60
B2 - Blue	490	65	10
B3 - Green	560	35	10
B4 - Red	665	30	10
B5 - Vegetation Red Edge	705	15	20
B6 – Vegetation Red Edge	740	15	20
B7 – Vegetation Red Edge	783	20	20
B8 - NIR	842	115	10
B8a – Narrow NIR	865	20	20
B9 – Water vapor	945	20	60
B10 – SWIR-Cirrus	1375	30	60
B11 - SWIR	1610	90	20
B12 - SWIR	2190	180	20

Source: Adapted from He & Yokoya. (2018)

2.1.4 Topographic Data

Other supplementary remote sensing data sets, like Topographic data (i.e., Elevation and Topographic wetness index), can also provide information in mapping an arable field. Topographic features are essential variables for the estimation of the crop area. For instance, high field fractions exist on the higher elevations (i.e., within the range of 1500 up to 2300m) in some parts of the Oromia region, and after this range, there are lower arable field fractions (Mohammed et al., 2020). In addition to this, Husak et al. (2008) also indicate in the fragmented landscape of Ethiopia; the maximum field fractions exist in higher elevation (i.e., 2000m). According to Wilson et al. (2016), the topographic wetness index of the topographic features also discriminates crop types. The topographic wetness index strongly correlates with species' vegetation composition (Moeslund et al., 2013). Not only the variables also the choice of the method play a crucial role in arable field mapping.

2.2 Available Methods and Models for Mapping Arable field

Most research is conducted to map cropland areas using remote sensing technologies and a specific method that covers continental levels. According to See et al. (2013), the cropland maps are prepared using crowdsourcing technology and combined with Google Earth imagery and Geo-wiki. A campaign organized for volunteers to collect cultivation areas, and the crop map produced using simple inverse distance weight (IDW) techniques. This map lacks detailed information on the extent of cropland because of its courser resolution (i.e., 1 km), and the method only considers the geographical location of cropland features. The cropland also can be mapped at the continental level. According to Xiong et al. (2017), cropland mapping across Africa is possible using the Moderate Resolution Imaging Spectroradiometer (MODIS) NDVI data. And they show the possibility of mapping the cropland area using an automated cropland mapping algorithm (ACMA), which is implemented in GEE. The authors used a mobile application to collect ground truth through using a mobile application and additional reference data from literature reviews. In addition to the ground data, they used MODIS NDVI to create cropland layers (i.e., cropland extent, cropping intensities) for 2014. They use this reference year for producing an automated cropland layer for the years 2003 up to 2014 by using MODIS NDVI data. Finally, they compare the 2014 automated result with census-based crop land data, and they found that there is under estimation of crop land area.

Different scholars also use different machine learning algorithms for data analysis. Machine learning algorithms like Random Forest, Artificial Neural Network (ANN), and Support Vector Machines (SVM), and a combination of spectral indices used to map crop types (Taffesse et al., 2013). A random forest was also used with time series enhanced vegetation index (EVI) extracted from Landsat 7 ETM+ to produce a crop classification map, and it is noted that with a limited training sample, the random forest classification accuracy is strongly affected (Tatsumi et al., 2015). According to Zheng et al. (2015), a support vector machine (SVM) used for mapping crop type by using time series of NDVI extracted from the Landsat image and applied two types of training data collection approach (i.e., stratified sampling and intelligent selection

approach). The authors concluded that the SVM shows a greater performance when using the intelligent selection approach than the stratified approach.

Additional to the above scholars, Mohammed et al. (2020) show the possibility of the crop area estimation using the Generalized additive model (GAM), which is not a machine learning technique. The authors use the GAM model by using five predictor variables: Landsat dry and wet season NDVI, topographic information (i.e., elevation and slope), and a 1km Field fraction map to estimate crop area. The model uses a classic GAM function (i.e., the Binomial link function).

Therefore, we need to consider some of the characteristics of machine learning algorithms and GAM models. Machine learning algorithms have different characteristics for the given datasets, and machine learning algorithms face various challenges (Zoubin, 2018). For instance, a random forest (bagging approach) is sometimes hard to interpret results but in gradient boosted algorithms; the results are more interpretable (Niklas Donges, 2019). Mostly there are two types of ensemble methods bagging and boosting. The random forest, which follows the bagging approach, builds an ensemble of independent trees based on the majority vote. In contrast, the gradient boosted model, a boosting method, builds a tree in a sequential manner (Bradley & Brandon, 2020). On the other hand, the GAM model has flexibility in the statistical distribution of the data (Murase et al., 2009) and uses quasibinomial distribution. This research uses Gradient Boosted (i.e., Beta link function) for variable selection and classical GAM to predict the 20m arable field probabilities.

2.3 Arable Field Mapping: The Current State and Knowledge Gap

The choice and method of data collection can have a significant impact on arable field estimation. The quantification of arable field estimation is done by using field-based statistical surveys, remote sensing, and a combination of agricultural statistics surveys with remote sensing data.

There are several land cover and cropland maps produced by using remote sensing data. However, some maps like MODIS and GLC2000 land cover maps have uncertainty incorrectly estimate the crop area (Eggen et al., 2016). According to the author, MODIS land cover underestimates cropland, whereas GLC-200 overestimates the cropland area. Crop maps were produced using GLC2000, MODIS Land Cover, MODIS crop Likelihood, and Africover and combining with national agricultural data for sub-Saharan Africa. The produced crop maps have a courser resolution (i.e., 1km), and there is a higher error rate of omission and commission in identifying crop areas (Eggen et al., 2016). Additionally, the authors indicate such maps are inconsistent in determining the extent and size of the arable field in the fragmented landscape.

On the other hand, mapping arable field using only optical images (i.e., sentinel-2 and Landsat images) is a challenge due to persistent cloud cover (Ashiagbor et al., 2020). Furthermore, the authors indicated that using only the optical images lowers the classification accuracy compared to using hybrid datasets (i.e., sentinel-1, sentinel-2, and image features). In addition to this, using a higher resolution optical remote sensing data requires a high cost (Ashiagbor et al., 2020), consumes much time, and computationally hard for extensive area mapping. Besides, measuring the arable field through conventional surveying techniques also requires higher cost, labor-intensive, collected at the district level, and crop estimation lacks spatial inaccuracies

(Marshall et al., 2019). Therefore, we need to establish a hybrid mapping method that can map the arable field by considering the integration of data, techniques, and methods.

Different mapping methods and models lack some characteristics; For instance, in models like a random forest, sometimes it is difficult to interpret the result (Hofner et al., 2014). Therefore, the research used gradient boosted regression models for variable selection because the model can select the most informative variables, and boosted model result is easy to interpret, and classical GAMs is used to estimate arable field fractions.

A few studies used SAR backscatter matrices like VV, VH, VV/VH ratio (Abdikan et al., 2018; Kumari et al., 2019), and topographic features (i.e., elevation and slope) (Husak et al., 2008; M. T. Marshall et al., 2011; Mohammed et al., 2020) to map arable fields. In addition to this, other research studies tried to map the arable field for a specific location (regional level) or the larger areas (i.e., country or continental level) with courser resolution arable field estimates. Therefore, it is important to consider the fusion of spectral, spatial, and temporal characteristics of the data to map arable fields.

The purpose of this study is to map 20m arable fields by using Sentinel 1 SAR backscatter matrices, Sentinel 2, topographic features, and a 1km arable field fraction. In the research, the Gradient boost model is used to select the most informative variable, and classical GAM is used to estimate a 20m arable field fraction for the Oromia region, Ethiopia.

3. METHODOLOGY

3.1 Study Area

The study area is located in the Oromia Region, one of Ethiopia's tenth regional states. The Oromia region surrounds Addis Ababa, Ethiopia's capital city, and Harari regional states. The total area of the region is 32,442.86 million hectares. The great African rift valley divides the region into two, and the region has different agroecological zones. In the Oromia region, there are 189 Woredas (i.e., districts).

In Ethiopia, food security is deteriorating in parts of the country, specifically in the Oromia region within different zones like Bale, Guji, east, and west Harerge zones affected by drought due to below-average rainfall patterns (FEWS NET, 2020). In addition to drought, the desert locust invasion affects the country's agricultural production by damaging significant crops and becoming a challenge for Ethiopian regional states (i.e., Oromia, Tigray, Amhara, and Somalia regions) (FAO, 2020). Mostly severe locust invasion will occur within 25 years (OCHA, 2020).

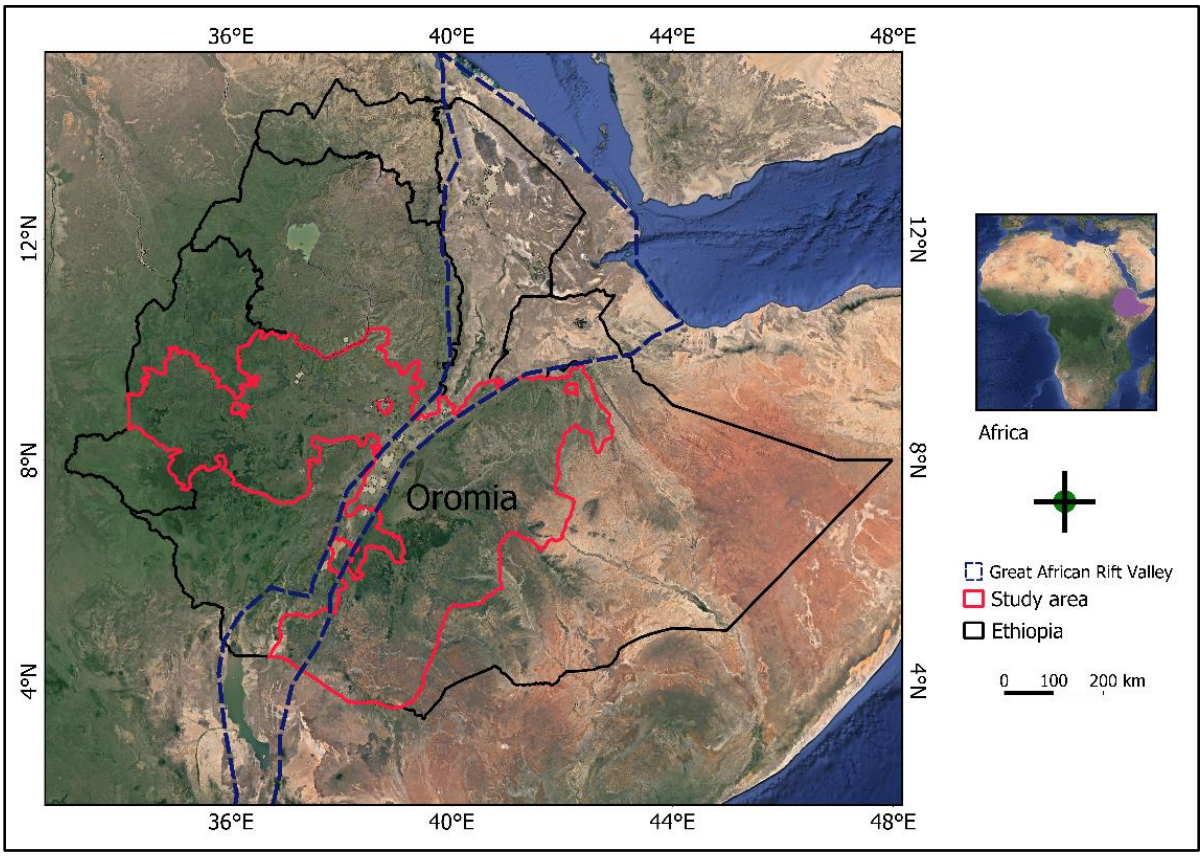


Figure 2 The Study Area

In the Oromia region, crop agriculture is characterized by small and irregularly shaped arable fields and mainly produces cereal crops for private conception and sales (Taffese et al., 2013). The Oromia region is the primary crop-producing region of Ethiopia, and the main crop produced in these regions is cereal crops, pulses, and oil crops.

Cereal Crops like maize are produced with an elevation range of 1500 to 2200m, wheat and barley grown within elevation ranges 1800 to 2200m and mid-highlands areas of the region such as Bale and Arsi zones, and pulses(i.e., bean) are grown within the range of 1400 to 2000 (Argaw, 2015). The Oromia region is a mountainous region with the highest elevation of 4387m and the lowest 308m. Mostly the eastern part of the region is considered as low land. The slope of the region ranges from 0^0 up to 78^0. The study area's central, eastern, and southern part is below 10^0 in average (Figure 3). Figure 4 shows the extracted average rainfall distribution (i.e., five years Rainfall Estimate from Rain Gauge and Satellite (CHIRPS) data) by using the sample point locations (i.e., the actual field fractions). From the graph, we can see that the rainfall pattern becomes the lowest in January and February, mostly the dry seasons of the study area.

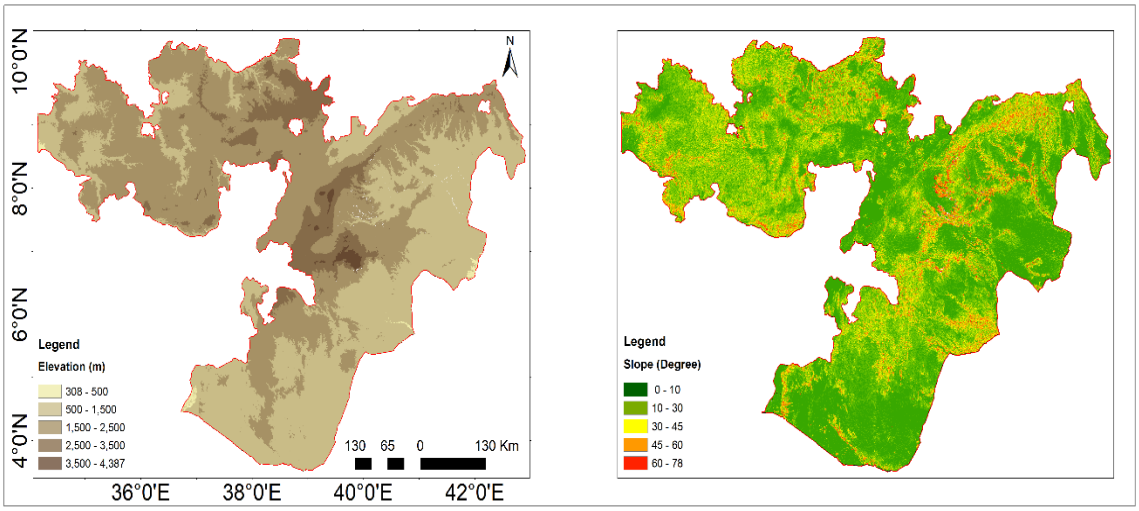


Figure 3 The topographic characteristics of the area, Elevation (Left) and Slope (Right)

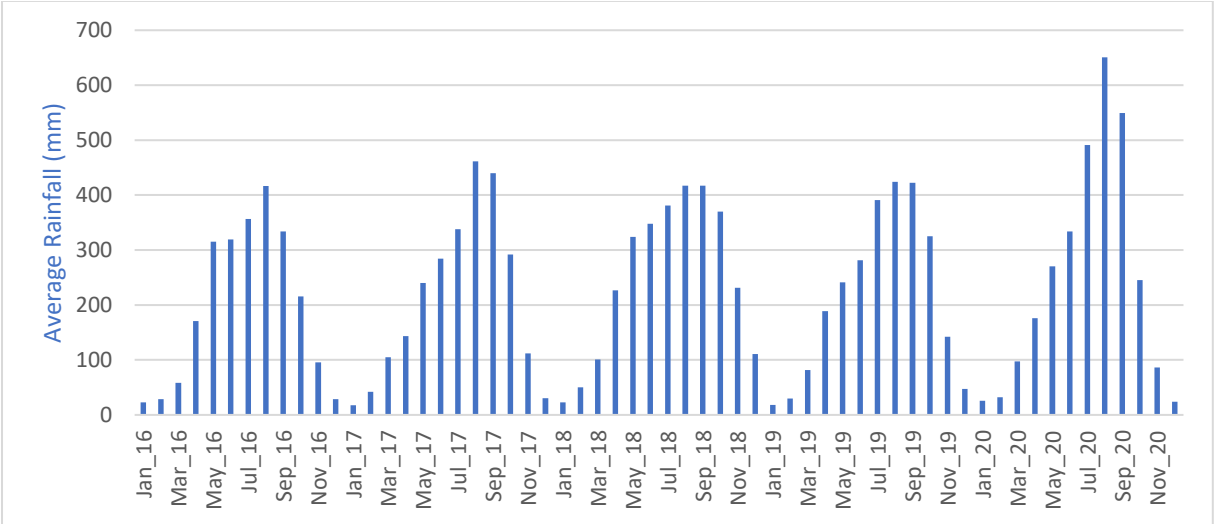


Figure 4 shows the Average Rainfall distribution on the Sample points locations.

3.2 Methodological Flowchart

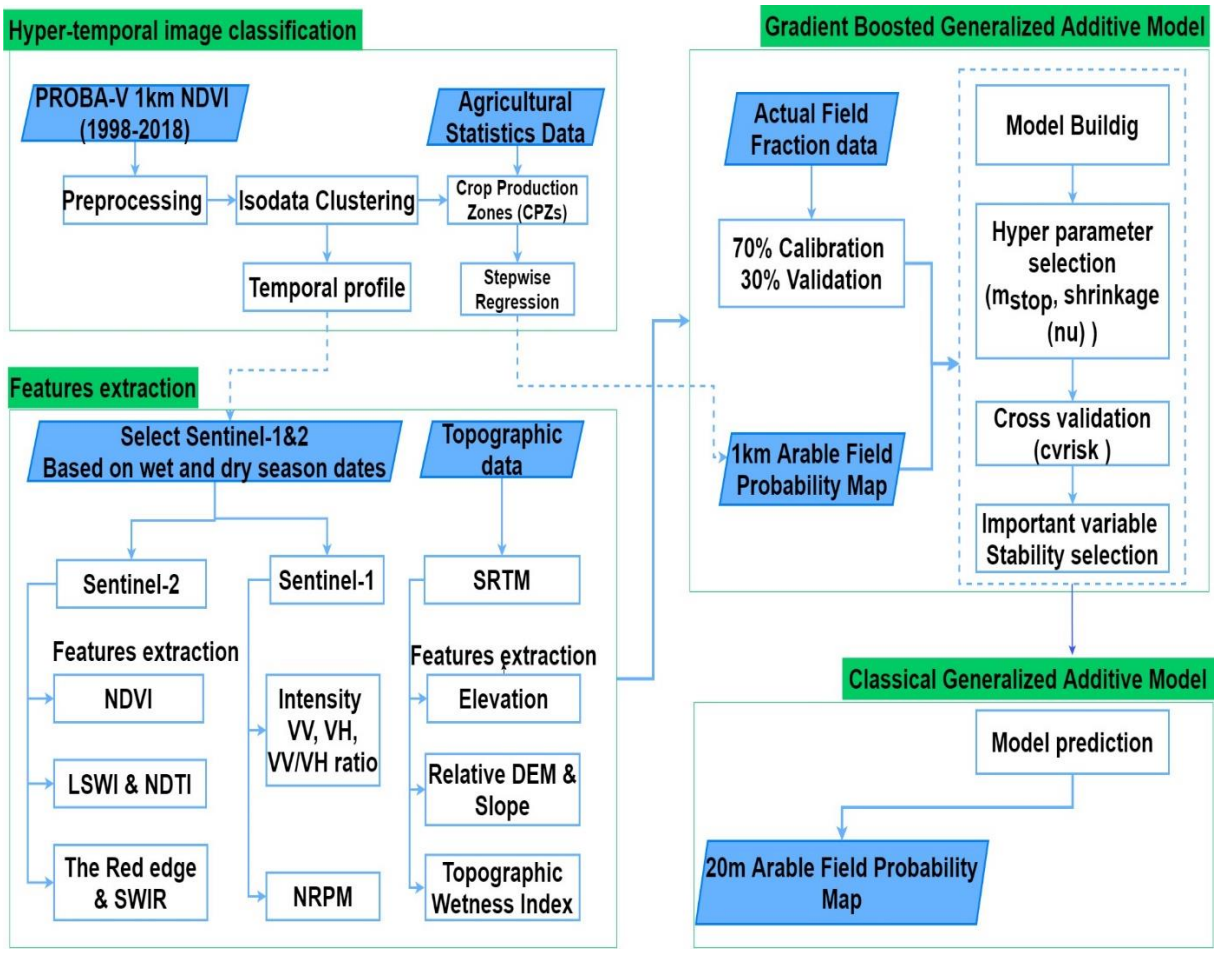


Figure 5 Methodological Flowchart

A summary of the method: the overall research workflow is presented in figure 5 and divided into three parts: hyper temporal image (i.e., a 1km Proba-V) classification, image feature extraction, and modeling process (i.e., using Gradient Boosted and Classical GAMs). In the first part, the Proba-V is classified into 200 classes using iso clustering techniques to produce an agroecological map. The produced agroecological map (CPZs) integrated with agricultural statistics data. Then stepwise regression is carried out in SPSS to estimate a 1km arable field fractions map. In addition to this, during the Isodata clustering process, the 200 classes were grouped into 66 clusters based on their relative minimum and maximum values of median monthly NDVI values to identify the wet and dry season periods. In the second part, based on the identified seasonal periods, the sentinel 1 and 2 images downloaded using GEE and image features are extracted from sentinel-1 and sentinel-2. In addition, topographic features were also extracted from the SRTM data. Gradient boost is used to select the important predictors in the third part, and the stability feature selection method is applied to the Gradient boosted method to select the most informative variable. The classical GAM is used to estimate 20m arable field fractions; Finally, the model performance was evaluated by using 321 validation points (30% of the total dataset). The detail of the method presented as follows:

3.3 Hyper temporal Image Classification

3.3.1. Data Preprocessing

The source of 1km PROBA-V NDVI imagery is from the ITC archive, a copy of Copernicus global 1km NDVI imagery. After downloading the 1km PROBA-V NDVI data, all the NDVI data stacked using ERDAS imagine software by including six ten-day images before 1998 and after 2018 to allow proper temporal filtering (De Bie, 2020). The temporal filtering method will replace DN values that are above 250 to zero (i.e., "The values are 251 for missing (Bad radiometry), 252 for cloud or shadow,253 for sea, and 255 for background (missing input data)") (De Bie, 2020). In addition to temporal filtering, the Savitsky Golay filtering (i.e., Upper-Envelope filtering through iterative smoothing process used to reduce the noisy NDVI time serious) was applied to the temporal cleaned image using IDL ENVI software (Beltran-Abaunza, 2009). Finally, the cleaned image is classified in ERDAS by using Isodata clustering.

3.3.2. Isodata Clustering

Isodata Clustering is a method of unsupervised classification that group classes based on their similarity in spectral characteristics by applying the clusters mean of the class values and stratifying the study area into different strata (Beltran-Abaunza, 2009). The PROBA-V NDVI data classified into 200 classes. The main reason for classifying the PROBA-V image into 200 classes is to capture the smallest variability of the arable field in each Ethiopian woredas (i.e., the total number of woredas is 520). In addition to this, "the choice of the 200 classes depends on the number of mixed categories within a pixel that is going to be differentiated" (De Bie, n.d.). The classification result of PROBA-V NDVI is used for two purposes.

The first one is to produce a crop production zone (CPZs) map which is a 1km arable field fraction estimate. To prepare CPZs, the classified Proba-V NDVI (i.e., 200 classes) converted into a shapefile and intersected with the Ethiopian woreda shapefile (Figure 7). The result is the NDVI classes per district shapefile and the area of NDVI clusters calculated for each class within the district. The R software is applied to automate the process, the Group_by and Summarize function is used; a parsed excel file is produced. The excel file merged with Ethiopian woreda agricultural statistics data (i.e., the arable field area per district level) https://catalog.ihsn.org/catalog/1438/related-materials. Then Stepwise regression is applied. The stepwise regression follows the following equation:

$$\text{Arable field area}_{\text{district level}} = f(\text{NDVI-Clusters area}_{\text{district level}}) \text{----------(1) (De Bie et al., 2008)} \quad (1)$$

The resulting data is used by SPSS software, and the coefficients are used as a field fraction. Finally, a 1km arable field fraction map is produced for the whole of Ethiopia, and masking is applied to reduce it to the Oromia Region. Figure 6 shows the overall detail of the method used to produce a 1km arable field fraction.

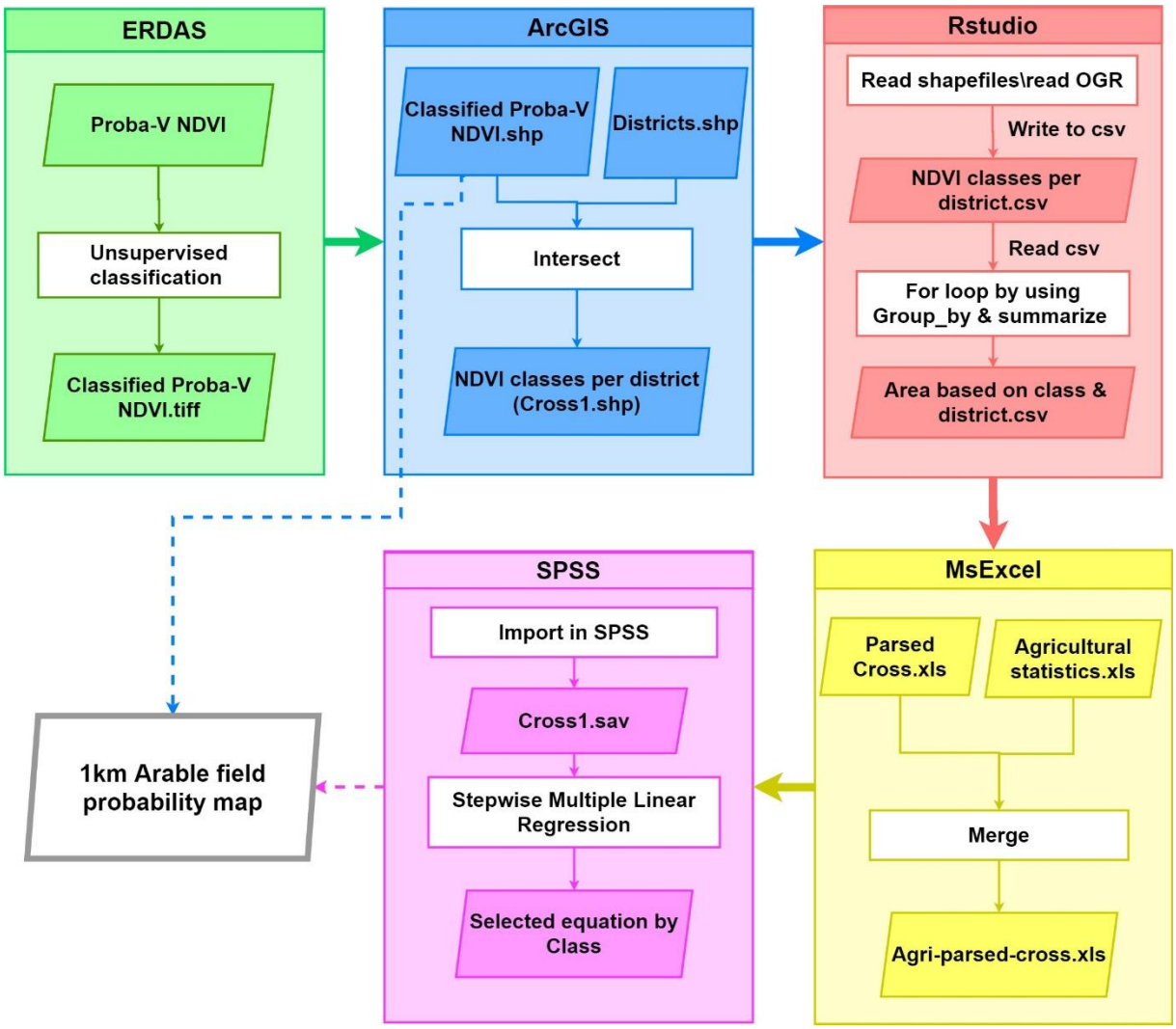


Figure 6 shows the step used to produce a 1km arable field fraction map

The second use of Isodata clustering is to create monthly NDVI profiles to identify the wet and dry seasons by using the relative minimum and maximum of the actual decadal NDVI values. After identifying the relative minimum and maximum NDVI values, the values are multiplied by 1.1 (i.e., 1.1*Lowest) and 0.9 (i.e., 0.9*Highest), respectively. The multiplication by this constant number (i.e., 1.1&0.9) helps us to obtain other relatively low or high NDVI values in addition to the minimum and maximum NDVI values. Additionally, the difference of the relative minimum and maximum is computed. After identifying the relative minimum, maximum, and difference, the conditional statement is created to assign each decadal value to high (H) and low (L). For instance, if the decadal value is less than the multiplication result (i.e., 1.1*Lowest), the lowest (L) is assigned. When the decadal value is greater than the multiplication result (i.e., 0.9*Highest), we give the highest (H). Flat (i.e., classes with "Flat" curves that indicate no growing seasons have similar characteristics throughout the year) is assigned when the difference of the maximum and minimum NDVI values are less than the minimum NDVI difference (i.e., the minimum NDVI difference equals zero). If these three conditions are not fulfilled, the decadal values are assigned a dashed line (i.e.,"-"). Then the table is sorted and grouped based on the similarity of high (H), low (L), and flat. For instance, the NDVI value 69 (i.e., the value with red color found at 6 dekad) assigned low “L” because it is below the multiplication result

(1.1*MIN), which is the threshold set to identify low "L" categories. The same is true for the high categories; if the value is greater than the multiplication result (0.9*Max), it will be assigned to High "H" (i.e., the value in the green color 174>160). The flat curves (Flat) are assigned when the difference of the multiplication results (0.9Max minus 1.1Min) less than the Min NDVI difference (i.e., zero).

Table 4 shows the identification of high, low, and flat categories

Min NDVI 0 Difference		Jan			Feb			Mar			Apr			May			Jun		
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Season CNt	Class	Band_1	Band_2	Band_3	Band_4	Band_5	Band_6	Band_7	Band_8	Band_9	Band_10	Band_11	Band_12	Band_13	Band_14	Band_15	Band_16	Band_17	Band_18
0	80	79	76	74	72	70	69	68	67	67	67	67	67	69	70	71	75	80	88
0	2	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	27
0	5	43	43	43	43	42	42	42	41	41	41	41	41	41	41	41	40	40	39
0	80	--	--	--	--	--	L	L	L	L	L	L	L	L	L	--	--	--	--
0	2	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat
0	5	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat

Min NDVI 0 Difference		Jul			Aug			Sep			Oct			Nov			Dec							
		19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36					
Season CNt	Class	Band_19	Band_20	Band_21	Band_22	Band_23	Band_24	Band_25	Band_26	Band_27	Band_28	Band_29	Band_30	Band_31	Band_32	Band_33	Band_34	Band_35	Band_36	MAX	MIN	Diff	0.9Max	1.1Min
0	80	101	116	133	150	165	174	178	174	165	151	137	123	112	103	95	90	85	81	178	67	87	160	73
0	2	27	27	27	27	27	28	28	29	29	29	29	29	29	28	28	28	28	28	29	27	-3	26	29
0	5	39	38	38	38	38	39	39	40	40	40	40	41	41	41	42	42	43	43	43	38	-3	39	42
0	80	--	--	--	--	--	H	H	H	--	--	--	--	--	--	--	--	--	--					
0	2	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat					
0	5	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat	flat					

Source: Adapted from De Bie. (2020)

Then the 200 classes were grouped into 66 clusters see Annex 2. The following figure shows an example of the time period used for downloading sentinel-1&2 image features. For instance, the offseason time period follows the minimum NDVI values (i.e., 4-10 dekad), and the on-season follows the maximum NDVI (i.e., 22-30 dekad) (Figure 7).

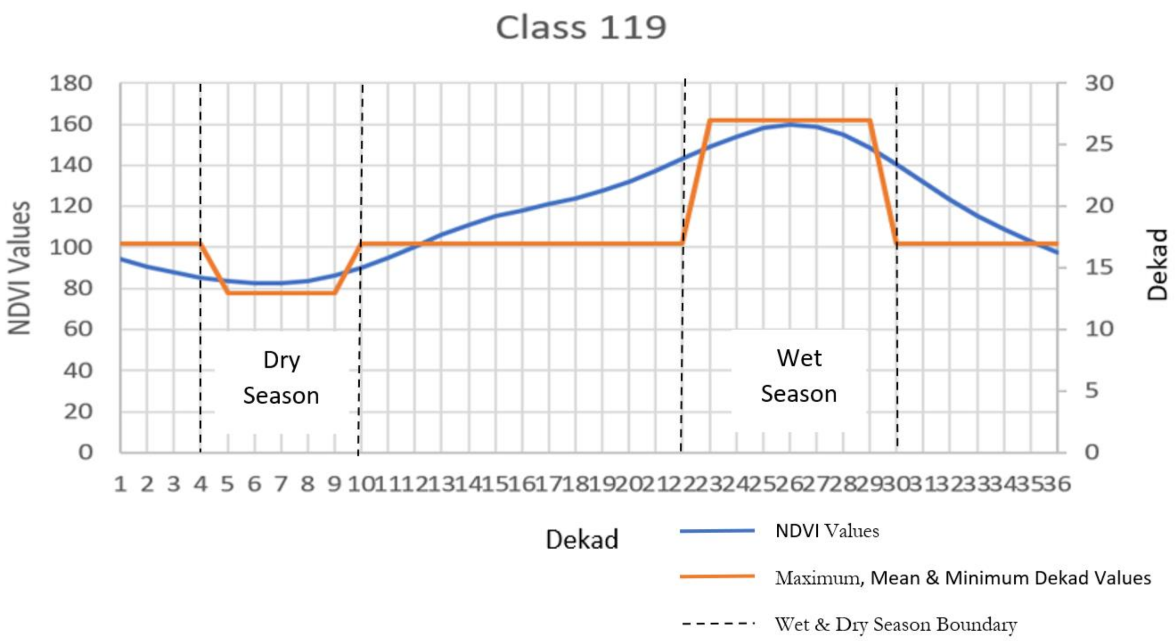


Figure 7 The periods used to download Sentinel-1&2 image features, the blue color shows the NDVI values, and the orange color indicates seasonal dates (dekad)

The periods are adapted to differentiate the wet (on the season) and dry (off-season) periods, which is essential for selecting sentinel-1 and sentinel-2 image features.

3.4 Feature Extraction

3.4.1 Sentinel-1

In this study, Sentinel-1 image features are extracted from Google Earth Engine (GEE) see Annex 3. Each Sentinel scene was already preprocessed (i.e., thermal noise removal, radiometric calibration, and terrain correction) with the sentinel-1 toolbox, and no need for further preprocessing; it is used directly. The time period used to download sentinel-1 is from 2016 up to 2020 (i.e., the date interval is chosen to align with the FAO's arable field definition), and the median values are considered. Sentinel-1 Backscatter polarization VV (Vertical-Vertical), VH (Vertical-Horizontal), and VV/VH ratio used for monitoring crops (Sun et al., 2020). The author indicated that the backscatter matrices like the VV and VH show the biomass distribution and are used for crop monitoring. In contrast, the VV/VH ratio indicates the shooting and harvest of vegetation. In addition to this, the inclusion of Normalized Procedure Ratio between Bands (NRPB) in the model increases the classification result to reach good statistical matrices in the validation and increases the model's generalization ability (Filgueiras et al., 2019). Table 5 shows the main Sentinel 1 image features used for estimating 20m arable field fractions.

Table 5 The Sentinel 1 image features

Backscatter matrices name (on and off-season)	Formula
VV and VH	intensity
VV/VH ratio	VV/VH
The Normalized procedure ratio between bands (NRPB)	intVH-intVV/intVH+intVV

3.4.2 Sentinel-2

In this study, Sentinel-2 image features are extracted from Google Earth Engine (GEE) (Table 6). Before downloading Sentinel-2 image features, the built-in cloud masking function is applied in GEE. The time period used for downloading Sentinel-2 images starting from 2016 up to 2020 and the median values are considered. From Sentinel-2, three image features from the Red edge bands (i.e., Band 5-7), two Land Surface Water Index (LSWI), two Normalized Difference Tillage Index (NDTI), two SWIR (i.e., Band 11&12) image features, and the NDVI vegetation indices (dry and wet season). The NDVI has an advantage in the discrimination of vegetation (Gumma et al., 2020). Besides, Red edge bands have a higher importance in classifying vegetation-related land cover maps (Qiu et al., 2017), and also SWIR significantly discriminates fine crop classifications (Zhang et al., 2020). The following image features are used for mapping the arable field fractions.

Table 6 The Sentinel 2 image features

Matrices Name (on and off-season)	Formula
Normalized Difference Vegetation Index (NDVI)	B8-B4/B8+B4
Land Surface Water Index (LSWI)	B8-SWIR1/B8+SWIR1
Normalized Difference Tillage Index (NDTI)	SWIR1-SWIR2/SWIR1+SWIR2
Red edge & Shortwave Infrared (SWIR) bands	The raw bands of Red edge and SWIR

3.4.3 Topographic features

The source of topographic features is the Shuttle Radar for Topographic Mission (SRTM). It is downloaded from USGS Earth Explorer. After downloading each tile of SRTM data for the whole Oromia region, all tiles are mosaicked in ArcGIS software. The SRTM data used to extract the topographic features (i.e., Topographic Wetness Index and Slope). The topographic wetness index is prepared by creating a tool in the Arc toolbox and using online topographic wetness index python scripts (i.e., the main input is SRTM) see Annex 1. The slope is also derived from the SRTM; additionally, relative DEM is a new concept developed by De Bie, and it was also created from SRTM data. The relative DEM data is accessed from the ITC archive., a new data set created from the SRTM and river shapefiles. The relative dem express the relative height of the surface by considering the river surface (i.e., a river surface has a zero meter elevation) as a starting point for the height measurement. The river shapefiles are also used for creating distance to the river input variable by using the Euclidean distance method.

3.4.4 Training and Testing Data

The training and testing data sets are obtained from a previous study. For this research, 1070 sample points are used, and these samples were prepared in the previous study by using 30m by 30m area frames' distributed randomly through the study area, and each area frame contains a grid of equally spaced 16 points (Mohammed et al., 2020). The points are assigned a label of an arable or non-arable field by using visual interpretation. The number of the arable field is divided by the total number of points to get the frames' arable field percentage. For instance, if there are 8 points labeled with arable field within the frame, to get the field fraction of the frame, we need to divide the labeled 8 points by the total number of the area frame points (i.e., 16) and the field fraction of the frame become 8/16 (0.5).

3.5 Modeling Process and Validation of the model

To achieve a better model result, we need to give more attention to the modeling process, especially in selecting the informative variables and the accuracy of each feature. Besides, the choice of the model also has a crucial impact on the final model prediction result. In this research, the classical GAM is used to

estimate 20m arable field fractions. For selecting the important variables, the Gradient boosted model is used, and to choose the most informative variables stability selection method is applied. Finally, the model validation is carried out by using out-of-sample (30% of the data) datasets.

3.5.1 Classical Generalized Additive Model (GAM)

Generalized Additive models are an extension of linear regression (i.e., Generalize Linear Model) with smooth terms (i.e., for fitting nonlinear terms) (Hastie & Tibshirani, 1986). The smooth terms imply a non-parametric regression method that adjusts the degree of smoothness. In GAMs, we can model linear, categorical, and nonlinear effects to the input data. Data issues like the normality of errors, nonlinear relationships, and autocorrelation of variables are handled by GAM models (Schmid et al., 2013). The GAM is the extension of the linear logistic regression. The linear logistic regression follows the following formula:

$$E(Y \mid X_1, X_2, \ldots\ldots, Xp) = \beta_0 + \beta_1 X_{1+\ldots\ldots+} \beta_p X_p \quad (2)$$

Where Y is a random variable, X_1, X_2,……., Xp covariant of the regression model; β_0, β_1,…. B_p is the coefficient of explanatory variables, and $E(Y \mid X_1, X_2, \ldots\ldots, X_P)$ is an estimate of the model.

The model equation of GAM extends GLM linearity, and to estimate using the GAM model, we first need to estimate the coefficient of (β_0) and $f(X_1)$ …. $f(X_p)$; an estimation of the GAM formula follows:

$$fGAM(X1 \ldots\ldots, Xp) = \beta_0 \sum_{j=1}^{p} fj(Xj) \quad (3)$$

Many researchers use classical GAM for ecological modeling (Citores et al., 2020; Maloney et al., 2012; Murase et al., 2009) and crop field mapping (Husak et al., 2008; M. T. Marshall et al., 2011; Mohammed et al., 2020). In this research, the classical GAM is used for estimating 20m arable field fractions. Many scholars use different models and multisensor remote sensing data to estimate arable field fractions.

3.5.2 The Gradient Boosted Regression

A. How Gradient Boosted Model Works?

The gradient boosted model is a method of converting an ensemble of several weak learners into strong learners; the residual of the first weak learners is fitted with the next weak learner until the residuals approach zero through the inclusion of many weak learners in the model (Greenwell and Brandon, 2020). In each iteration step, the algorithm tries to minimize the error of the first step by the second one, and finally, the model sequentially adds the result of individual steps, then a strong model is created (Yıldırım, 2020). The following Figure 8 shows how the sequential approach of the gradient boosted algorithm works.

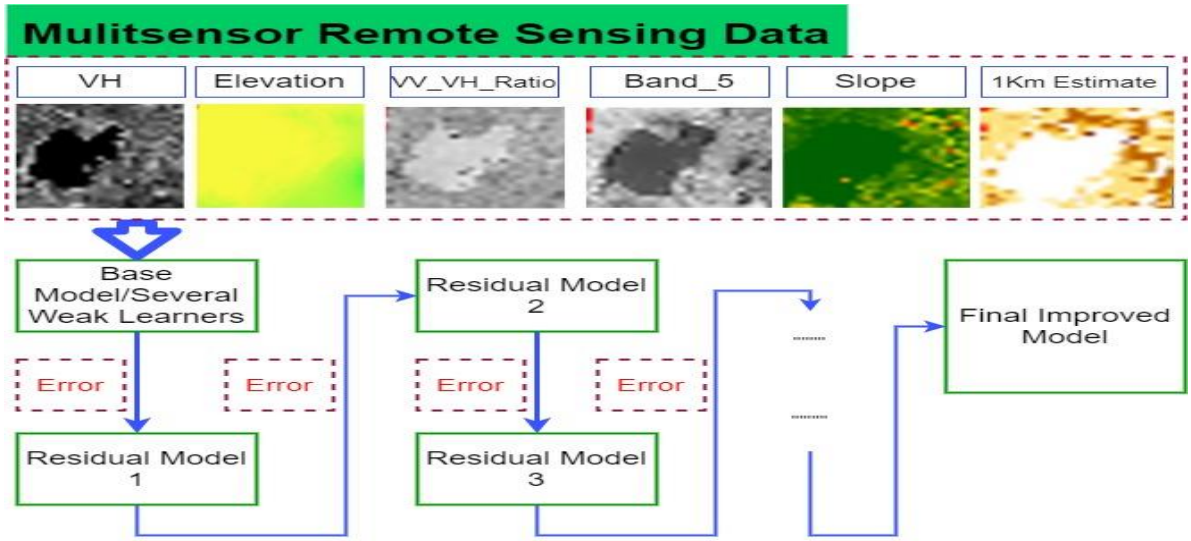


Figure 8 Sequential ensemble approach.

In general boosting can be said that stage-wise procedures (Thomas et al., 2017). When we say stage-wise, it includes the first fitted variables without excluding them from the second fitting model processes, and the base learners are sequentially linked. Therefore, understanding the gradient boosted helps us to implement the model in practical application, but we also need to give emphasis on the hyperparameter tunning or hyperparameter selection during the modeling process of gradient boosted.

The gradient Boosted regression (GBR) is used to select important input variables for predicting the arable field fractions. The Gamboost function is used for a boosted (generalized) additive model, and a model is validated by using cross-validation implemented in the cv risk function. The model uses base-learners (bbs) for fitting base models in component-wise gradient boosting in the function of Gamboost.

B. Hyperparameter selection with GBR

Hyperparameters are one of the model parameters that are set before starting to train the models. The promising model prediction results mostly depend on the choice of hyperparameter selection and predictor variables. Hyperparameter tuning is the basis for obtaining an optimal prediction result. Different algorithms use different techniques to determine the most influential hyperparameters. For instance, selecting a lower or higher learning rate will significantly impact the final model result (Yıldırım, 2020). Figure 9 shows a boosted gradient model with eight hundred iteration and different learning rate (i.e., Shrinkage) values. In the model, when we include a large learning rate, some model gives a result with some predictor variables with lower squared error. For instance, in Figure 9a, when the learning rate is one, we can get an optimal number of iteration at 28, but with a small number of predictor variables, and in Figure 9b with a 0.5 shrinkage value, we can get 93 optimal number of iteration but with a higher number of predictor variable as compared to learning rate with one. Whereas, when we incorporate a lower learning rate in the model, we are unable to reach the optimal boosting iteration (Figure 9d). The same is true for the number of iterations incorporated in our model when there are optimal iterations present in our model, the better and flexible option for achieving a good model result but with a low loss function (i.e., slower gradient decent) in finding

the local minimum. Therefore, based on this Gradient boosting model, the main hyperparameter is the shrinkage value that has an influence on finding the optimal iteration.

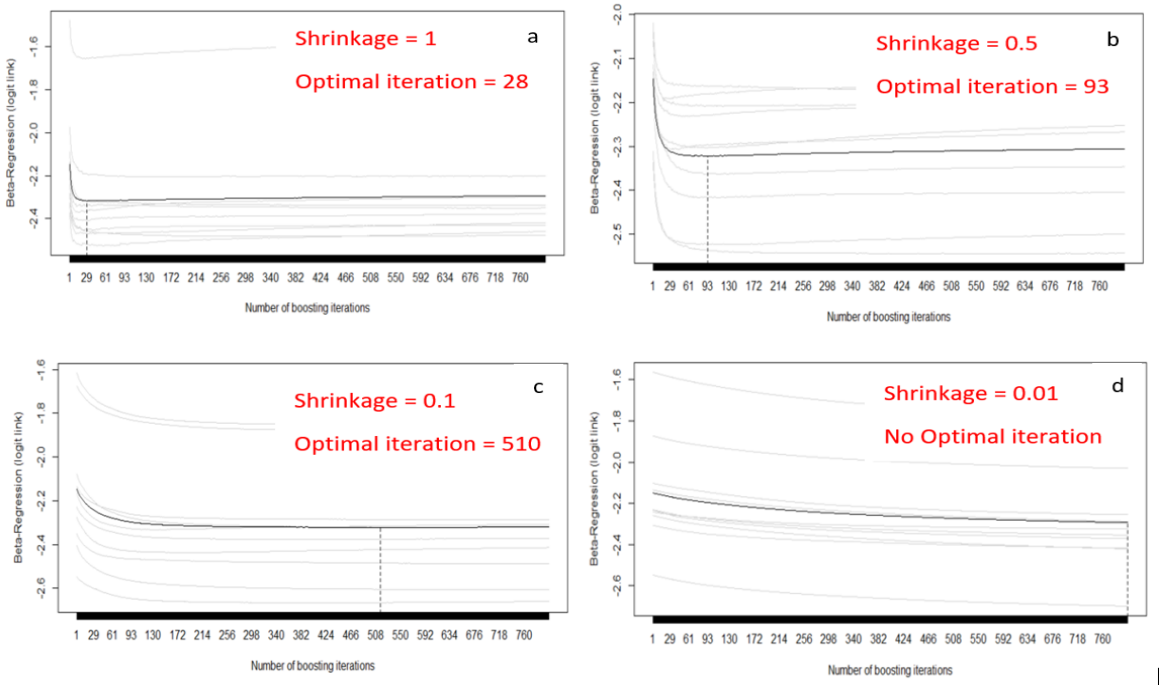


Figure 9 shows the relation between the learning rate and optimal iteration

C. Fitting Generalized Additive Model with Gamboost function

Fitting, a generalized additive model, requires a model formula with the base-learners a penalized regression splines. The general structure of the model is presented in Figure 6. In the model, the formula is created by using field fraction as the dependent variable and the predictor variables as an explanatory variable by using base learner "bbs" for nonlinear features and "bols" for linear feature (i.e., a 1km field estimate). In the model, the hyperparameters used with m_{stop} value of 800 and a shrinkage value of 0.3 applied to boost control functions. These values are selected by trying different combinations of m_{stop} and shrinkage values.

```
gamboost(formula, data = list(), na.action = na.omit, weights = NULL,
         offset = NULL, family = betaReg(), control = boost_control(),
         oobweights = NULL, baselearner = c("bbs", "bols"))
```

Figure 10 The General structure of Gradient boosted Model

D. Base learners

In the Gamboost function, the base learners are classified into three categories to capture linear or categorical effect by using bols(), smooth effect by using bbs(), and smooth surface estimation by using bspatial() functions (Hofner et al., 2014). The bols() function helps us define the ordinary least square base learner with a linear or categorical effect. In this research, the crop production zone estimate is considered a linear predictor variable because it is assumed that the field fractions and crop production zones (i.e., 1km estimate) have leaner relationships. The other 23 predictor variables are incorporated in the model as a smooth effect,

and the bbs() function is used. The spatial effect (i.e., bspatial()) is not incorporated in the model because of the spatial variability of the study area captured by crop production zones.

The research used the extracted input features as base learners from sentinel-1 (i.e., VV, VH and VV/VH ratio, and NRMP), sentinel-2 (i.e., NDVI, LSWI, NDTI, the red edge, and SWIR bands) for the gradient boosted model as an input. In addition to this, topographic features (I.e., Elevation, slope, relative DEM, and topographic wetness index), distance to river, and a 1km arable field probability map (i.e., CPZs) are also used as input for the model.

E. Cross-Validation: Early Stopping

In the Gamboost domain, cross-validation of the model was constructed by using the CVrisk() function. The number of boosting iterations is one of the hyperparameters for boosting model (i.e., denoted by m_{stop}). There are many ways to determine the early Stopping of the boosting methods, for instance, using Akaike Information Criteria (AIC) and cross-validation estimates (Hofner et al., 2014). In this research, ten-fold cross-validation is used to stop the model before overfitting.

F. Variable Reduction: Stability Selection

Selecting the most important variables is a base for achieving the best prediction result from a given algorithm. Identifying the most influential variable has an advantage for the machine learning algorithm to train quickly, increase the model's flexibility, and increase the model interpretability ability to achieve better accuracy and minimize overfitting of the model (Saurav, 2016). There are different types of variable importance selection methods like filter, wrapper, and embedded methods.

Filter methods mostly depend on selecting the variable based on the performance criteria without considering the modeling algorithm (Jović et al., 2012). These variable selection methods are very flexible and have a high speed (Bouhamed et al., 2012). One of the drawbacks of filter methods is on the response and predictor variables relation; the method does not consider the relationship between the target and the predictor variable (Raschka, 2021). Information gain, chi-square test, fisher score, correlation coefficient, and variance methods are examples of filter methods.

Wrapper methods focus on the efficiency of the chosen model algorithm for sub-setting the features; for instance, for classification, the subset depends on the classifier's performance (e.g., SVM) (Jović et al., 2012). In addition to this, wrapper methods are slower than filters method because it is more dependent on resource-demanding of a given algorithm, and forward (i.e., by adding variables) and backward (i.e., removing variables) selection methods are an example of wrapper method (Jović et al., 2012).

Embedded methods differ from wrapper and filter methods based on the way of selecting an important variable. These models incorporate variable selection during the training process (Guyon & De, 2003). For instance, Random Forest uses a variable score for selecting an important variable and ranks them based on the score. The percentage score is calculated from the mean squared error, and the feature is ranked based on the importance score (Holzinger et al., 2014).

Component-wise boosting algorithms are one solution to select predictors in high dimensional data when there are many predictor variables than the observation. Through cross-validation, one can stop the boosting algorithm before overfitting, and variable selection is carried out during the modeling process, and the optimal base-learners are selected. Thomas et al. (2017) state that during cross-validation, the selection of the stopping iteration optimizes the empirical risk on the test data (predictor risk). Sometimes when we are implementing early Stopping, the boosting algorithm tends to offers a way to perform variable selection while fitting the model. Sometimes the boosting algorithm may incorporate some noisy variables during the model fitting(Thomas et al., 2017). Some researchers use different techniques to reduce such problems by plotting the Root Mean Squared Error (RMSE) and the number of iterations. In addition, some researchers use stability selection methods to resolve the issue of noisy variables.

Stability selection is a dynamic and versatile approach that can be integrated with a different algorithm such as gradient boosted model and improve the already fitted model using the important unique variables from the gradient boosted model (Nicolai & Peter, 2010). According to the authors, the stability selection method follows five different steps to enhance model performance by reducing the noisy variable included during the modeling process. Implementation of stability selection in boosting algorithm presented in the following five steps :

1. The first step is fitting the model with a sample dataset with a specified m_{stop} and learning rate.
2. Fitting the boosted model and increase the iteration until early stoping with m_{stop} up to where the most important variable is selected.
3. Repeat the above two steps for the selected variable
4. Calculate the selection frequency per variable (i.e., base learner)
5. Select all base-learners that were selected with a frequency of at least π_{thr} (i.e., prespecified threshold values) and after the stable variables are identified.

The Gamboost function gives the rank of feature importance in gradient boost but doesn't tell us which explanatory variable is crucial for the modeling process. The boosting algorithm often includes some noisy variables stopped based on early stopping cross-validation techniques (Thomas et al., 2016). Selecting an optimal predictor variable is a base for model building and stability selection method used to select the most influential variables. In this research, the Stability selection method is applied to choose the most influential variable by using the stabsel() function and using the fitted Gradient boost model. The stability selection method in addition to the fitted Gradient boost model, also uses two main input variables for selecting the most influential input variables the cutoff value and the number of the unique variables from the result of the Gradient boost model (Hofner et al., 2015). Figure 11 shows how the stability selection works;

```
stabsel(x, cutoff, q, PFER,
        folds = subsample(model.weights(x), B = B),
        B = ifelse(sampling.type == "MB", 100, 50),
        assumption = c("unimodal", "r-concave", "none"),
        sampling.type = c("SS", "MB"),
        papply = mclapply, verbose = TRUE, FWER, eval = TRUE, ...)
```

Figure 11 Shows the Components of the Stability Selection method

Source: R package documentation.

Where "x" is the fitted model, "cutoff" is the value to be set in between 0.6 up to 0.9, "q" is the number of unique variables in the boosted Gam model and PFER (Per family error Rates) this indicates falsely selected variables.

Variable importance identification and future reduction is a base for obtaining a good model prediction result. Generally, mapping arable field fraction requires different datasets, variable reduction, and prediction models. The choice of a prediction model can have a significant impact on the estimation of arable field fractions. One can choose the prediction models based on the capacity of the model in capturing the characteristics of input datasets. For instance, classical GAM is a nice model for solving data issues.

3.6 Model Validation and Prediction

From 1070 samples, 70% (749 sample points) of the data used to train the model, and the rest 30 % (321 sample points) were used to test the model accuracy. Multiclass ROC validation criteria are used for validating the model. Multiclass AUC considers the different class values (i.e., in the field fractions, different classes are starting from zero up to one) to be calculated and provide a total of multiclass measures. Thus, the multiclass AUC considers many class values and differentiates multiclass values, whereas the classical AUC considers only two-class problems (Landgrebe & Duin, 2007).

The predicted model result depends on the accuracy and performance of the model in predicting the out-of-sample data. The final prediction is made with the most informative image features, and the selection of these image features is based on the Gamboost model and stability selection method. The accuracy of the final model result was evaluated by using multiclass AUC, R^2, and the overall deviance and deviance, explained by each predictor variable.

4. RESULT

4.1 1km Arable Field Probability Estimate

This research aims to produce a 1km arable field probability map, later used in the modeling process as an input variable. Using stepwise regression in SPSS software, the arable field probability is produced by integrating the regression coefficients as field fractions. The analysis result shows an 88.4% R^2 value (adjusted R^2). Table 7 indicates the statistical summary of the regression result. Thirty-three classes are significant from 200 Proba-V NDVI classes, and seven classes have more than 50% field fractions.

Table 7 The statistical result of Stepwise regression

NDVI Class	Coefficient	Significance
Class 156	0.740	.000
Class 110	0.607	.000
Class 177	0.585	.000
Class 103	0.509	.000
Class 194	0.372	.000
Class 184	0.325	.000
Class 66	0.304	.000
Class 172	0.234	.000
Class 160	0.319	.000
Class 114	0.791	.000
Class 132	0.482	.000
Class 197	0.232	.000
Class 124	0.726	.000
Class 187	0.874	.000
Class 193	0.205	.000
Class 170	0.357	.000
Class 129	0.407	.000
Class 100	0.449	.000
Class 183	0.372	.000
Class 80	0.249	.000
Class 143	0.200	.022
Class 20	0.132	.001
Class 149	0.382	.001
Class 145	0.424	.000
Class 58	0.247	.012
Class 97	0.366	.012
Class 56	0.261	.006
Class 26	0.328	.029
Class 120	0.051	.029
Class 174	0.172	.033
Class 135	0.259	.004

From the seven classes, two classes (i.e., Class 177 and 110) are located in the eastern and western parts of the region; this shows the similarity of agroecological characteristics. Figure 12 shows the spatial orientation of classes with more than 50% of field fractions. In Figure 12, some characteristics of arable fields of the study area, for instance, in class 103 (Figure 12(1)) and class 177 (Figure 12(3)), we can observe trees inside arable fields.

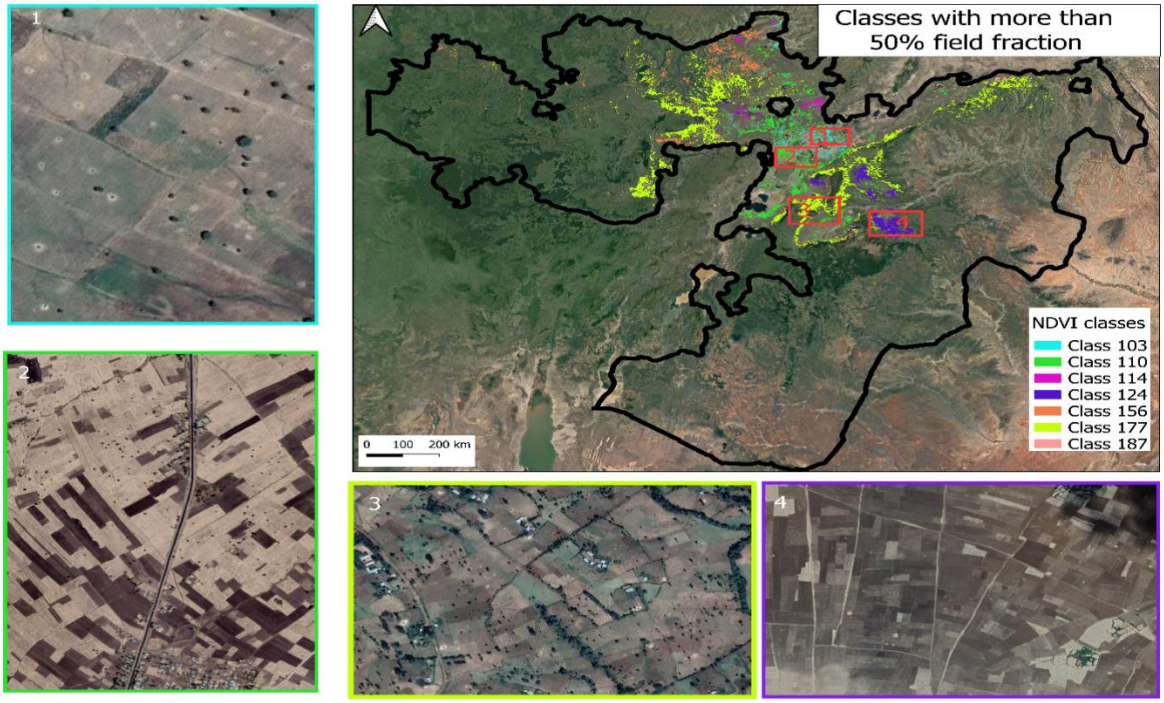


Figure 12 The classes with more than 50% field fractions

Finally, 1km arable field fractions estimates were produced for the Oromia region. The Oromia region's total statistical agricultural crop production estimate is 4,077,968 hectares, and the estimated arable field fraction is 3917033.1 hectares. Higher field fractions are concentrated in the middle of the region. In Figure 13, the inset map shows the field fractions estimates are more or less discriminate the arable field and the forest cover. The arable fields are mostly concentrated at the central part of the region.

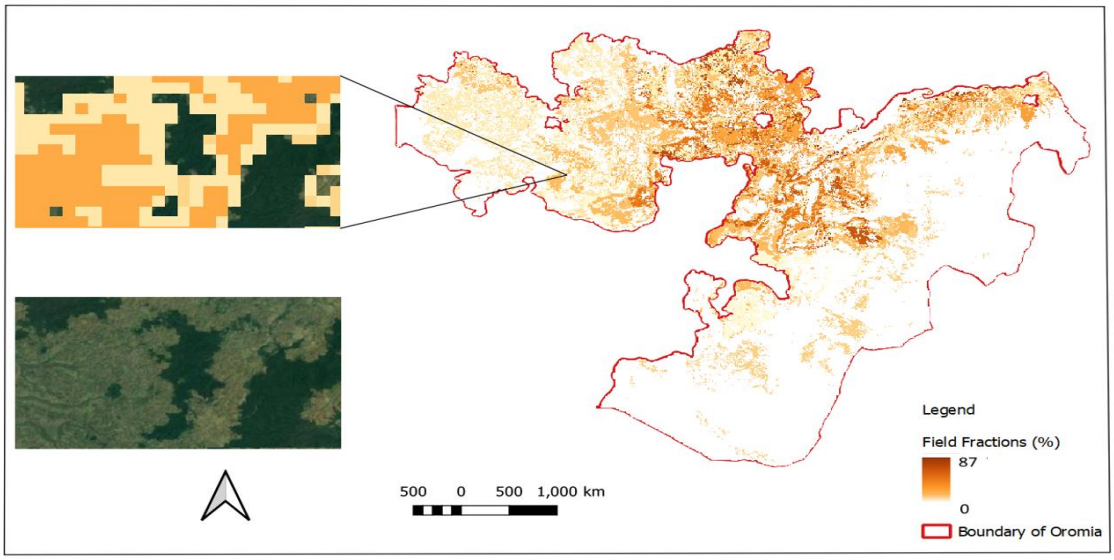
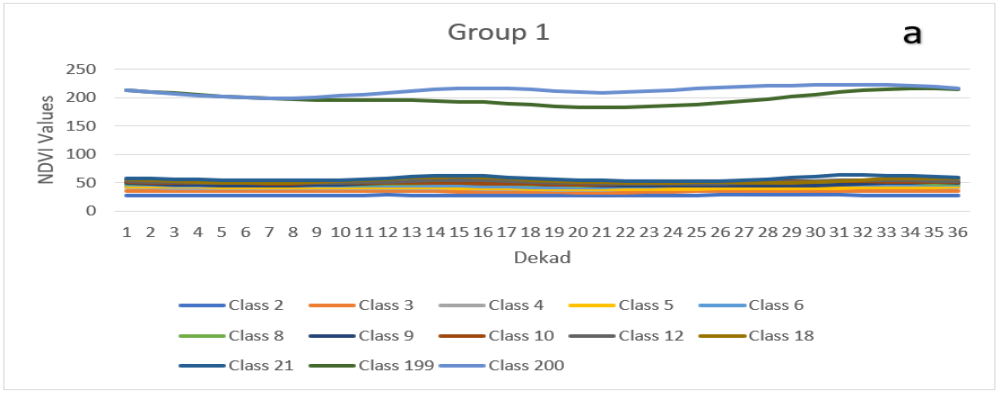
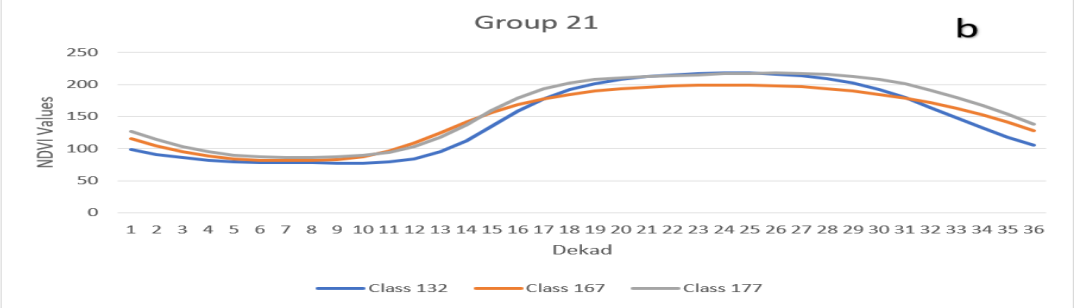


Figure 13 A 1km Arable field estimate of Ethiopia and Oromia region

4.1. Feature extraction by Grouping NDVI Clusters

Grouping NDVI classes into smaller groups can help us reduce computational time and generalize similar information from the group of classes. The main use of grouping NDVI clusters is to use the grouped clusters for extraction of Sentinel 1 and Sentinel 2 image features. In addition to this, before using the whole group of clusters, we can have the chance to exclude some groups that have different spectral responses (i.e., spectral profiles that have similar NDVI values throughout the years). The 66 groups (Annex 2) are identified based on their NDVI values by using the conditional statement (i.e., Low, high, and "flat" curves. The existence of arable fields analyzed by a careful study of the NDVI temporal profile within the group. Depending on the NDVI profile pattern, we can distinguish the type of land use land cover of the study area, and the interpretation of some of the classes are presented as follows. For instance, Figure 14a (Group 1) shows a flat curve in the highest NDVI values and in the lowest part. The lowest NDVI values could represent permanent cropland, and the highest NDVI flat curve represents a forest cover because it shows a similarly high value throughout the year. Therefore, we cannot expect arable fields from this group, and this group is excluded from the modeling process. On the other hand, Figure 14b represents the Oromia region arable field because the graph shows two periods, one for the lowest, representing the harvesting time, and the second period indicates high NDVI values, which is most probably the growing season. Figure 16b is also overlaid on Google Earth, and it indicates the existence of an arable field. In the Oromia region, we can get double-cropping seasons. Figure 14c shows the four seasons of the study area, with two dry seasons where the NDVI values are too low and two wet seasons where the NDVI values are too high. For such a season, the dates for downloading Sentinel images are chosen based on the lowest NDVI values.





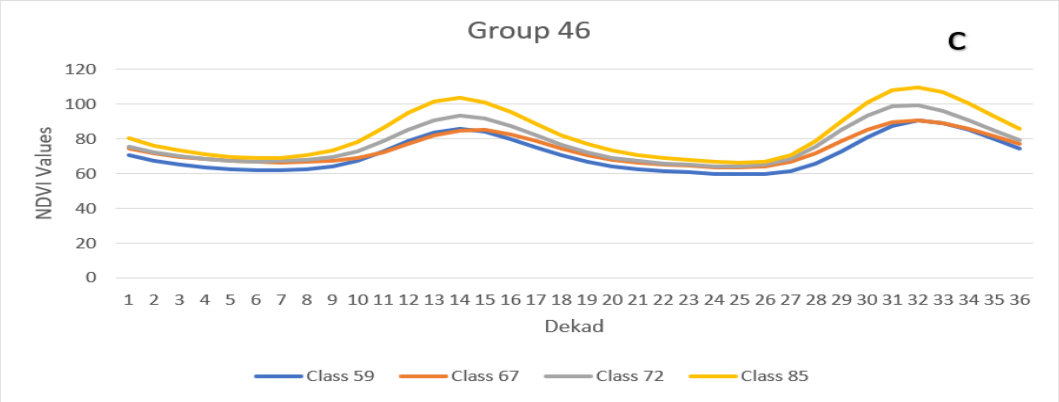


Figure 14 The spectral profile curve of Group 1(a), Group 21(b), and Group 46 (c)

4.2 Modeling process and Model Accuracy

The main goal is to find out the optimal model parameters that can lead to a good model prediction result in the modeling process. To achieve an optimal prediction result, we need to select the hyperparameters and the most informative variables. In gradient boost, early stopping is applied to stop the model before overfitting occurs, and the most explanatory variables are selected using stability selection. Then the model used the most informative variable to estimate 20m arable field fractions. The detail of results are presented as follow:

4.2.1 Early Stopping

Most modeling algorithms require hyperparameter selection to prevent the model from overfitting. In gradient boosted regression, early Stopping is used to prevent overfitting by using cross-validated predictive risk with 10-fold bootstrapping. Eight hundred boosting iterations are used, and the model stopped at 141 iterations to prevent overfitting (Figure 15).

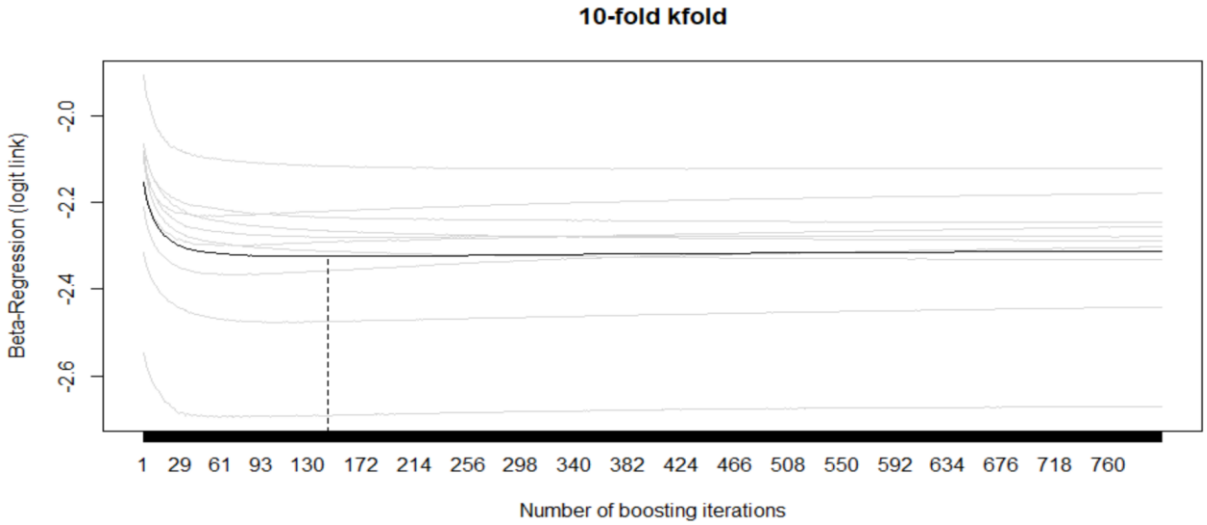


Figure 15 Cross-validation predictive risk with 10-fold (The optimal number of boosting iteration)

4.2.2 Variable Importance

The major activities in the research community are the estimation or prediction of the dependent variable from the collection of predictor variables by selecting the most informative variables (Looppe et al., 2020). The Varimp () function applied on the optimal boosted GAM model for ranking the variables according to their importance. Varimp () function in boosted regression works by selecting the most important variable based on the selection frequency of the variables. Figure 16 shows the rank of the variable based on their importance. The most influential variable is Sentinel-1 (dry_season_vh), from the total of twenty-four image features. Thus, there are ten unique (i.e., their selection frequency greater than 0.01) informative variables for mapping arable field probability, starting from dry_season_vh up to LSWI_dry variables.

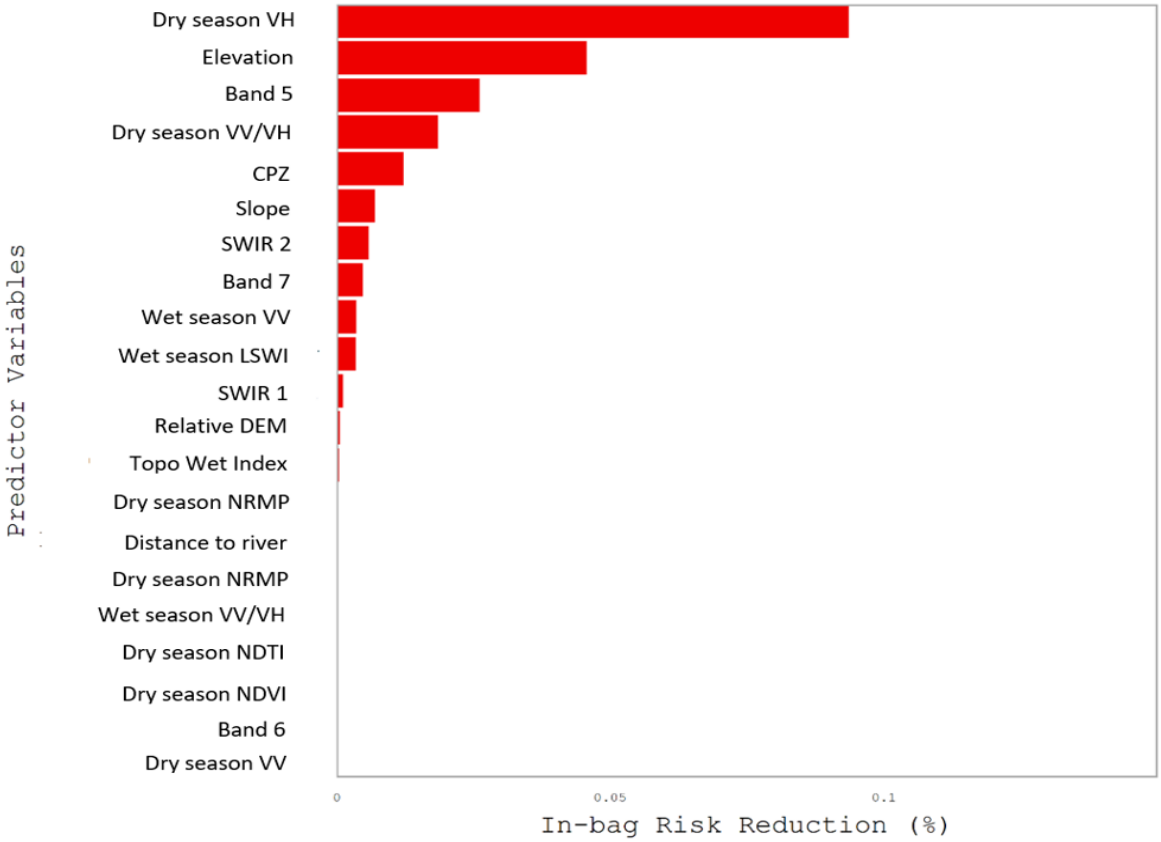


Figure 16 Feature importance of the model input variables

4.2.3 Variable Reduction Using Stability Selection Method

It is important to reduce the number of predictor variables before running the gradient boosting model prediction. Variable reduction is very important in making a prediction based on informative variables. The important variables were also identified by using the stability selection method. Six variables are selected, and the most influential variables are two sentinel one image features (i.e., dry season VH and dry season VV over VH ratio), topographic features (i.e., elevation and slope), Sentinel-2 (i.e., Band 5), and a 1km arable field probability estimates (i.e., CPZs). The red dot in the graph shows the most informative variables that

are found beyond the cut-off value (i.e., π equals 0.6). The rest variables below this cutoff line are not important for estimating a 20m arable field fraction.

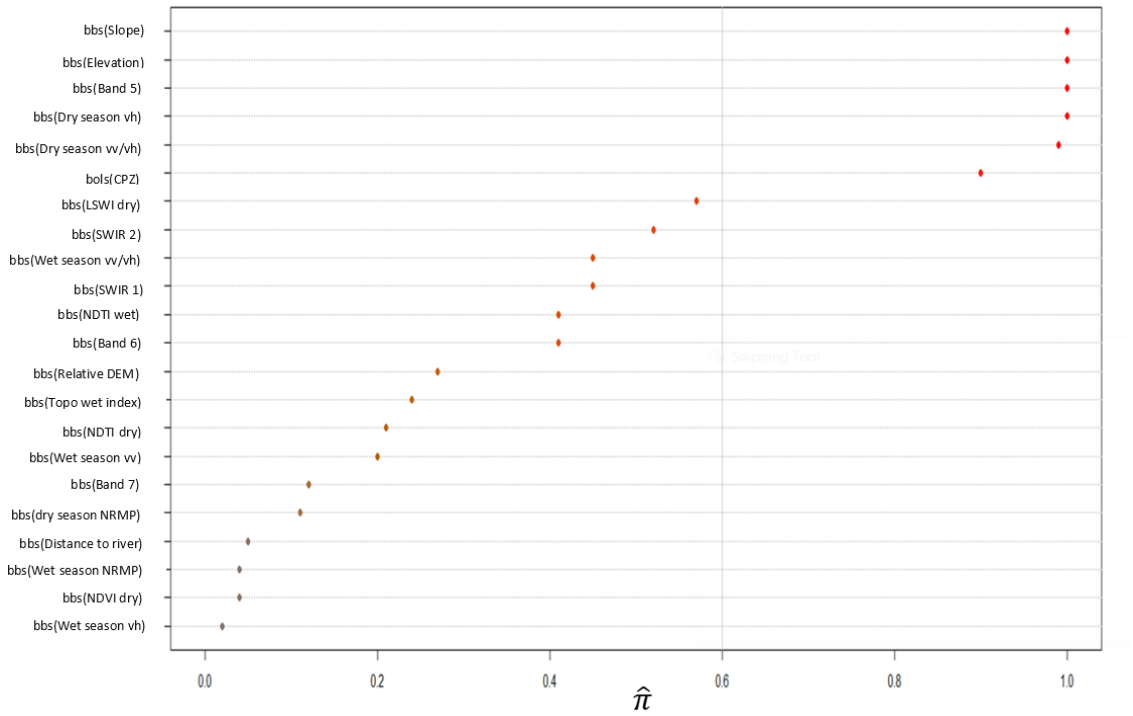


Figure 17 Feature reduction using Stability Method: where π is selection frequency; the six red above the cut-off line (0.6) are the most informative variables that are used to estimate 20m arable field fractions. The grey line represents the threshold.

4.2.4 Partial effect of the model input variables on the Actual field fractions

We analyzed the partial deviance explained by each informative variable before describing the partial effect of the model predictor variables on the actual field fractions. Figure 18 shows the partial deviance explained by each predictor variable in incremental order. The partial deviance explained by dry season VH was 33.3% which is the most informative variable. The second most important informative variable is elevation, and the partial deviance explained 17.2%. The deviance explained by Band 5, a 1km arable field fractions, and the slope was 14.4%, 13.3%, and 6.2%, respectively. The least informative variable is the dry season VV/VH ratio, and the partial deviance explained was 3.34%.

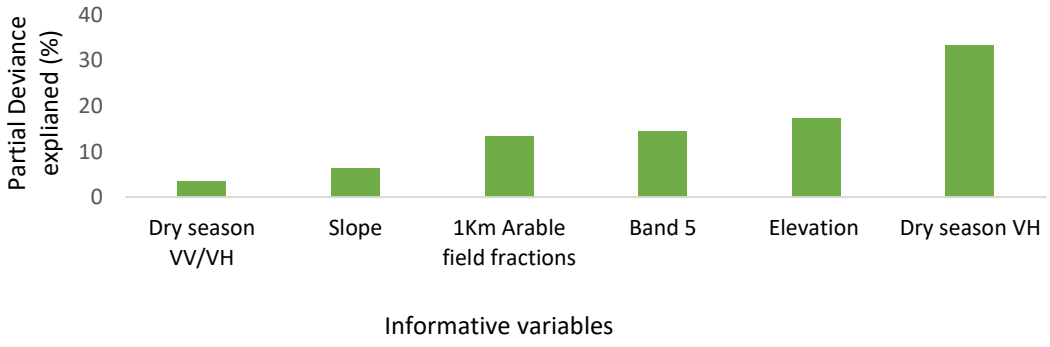
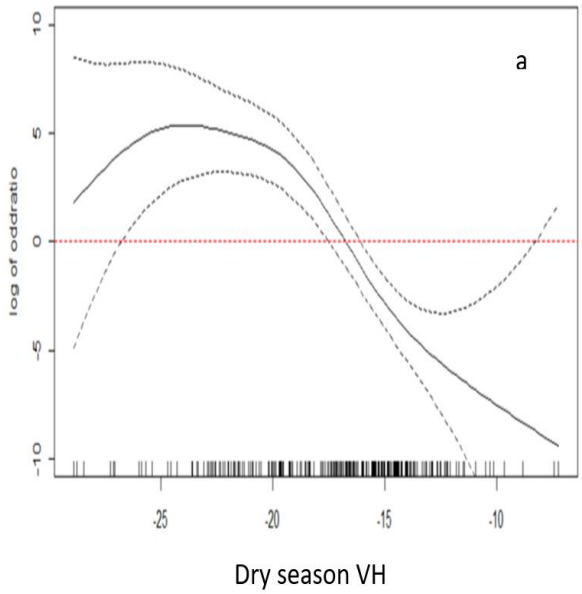
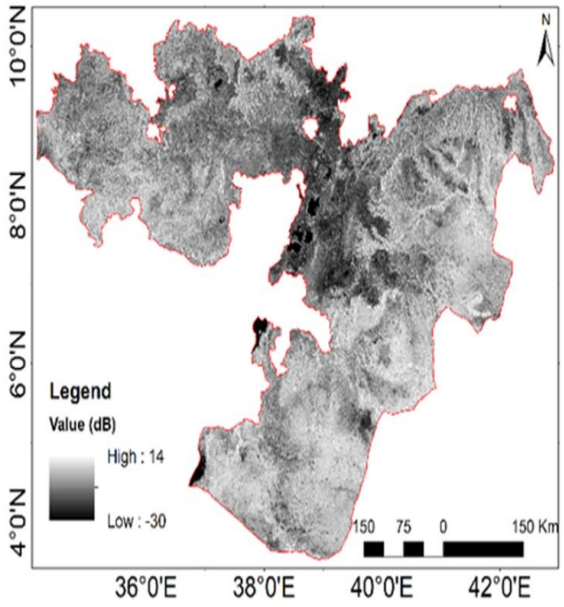


Figure 18 The partial deviance of each informative variable

One of the classical GAM advantages is to provide information about the effect of the input variable on the dependent variable (i.e., Actual field fractions). For instance, figure 19a shows the partial effect of Sentinel-1 offseason VH on the actual field fractions. As the value of dry season VH increases, the partial effect on the existence of arable field will increase up to the -25 value (i.e., the presence of high field fractions) afterward; the effect will decrease as the value of dry season VH increases, and we can get above 50% of field fraction from offseason VH backscatter matrices. As we can see from the counter map of dry season VH, the central part of the region has a weak backscatter signal. The dark tones potentially can be an arable field, whereas a very dark one is a water body, especially the reft valley lakes. Figure 19b shows that as elevation increases, the partial effect on the field fraction will increase to an elevation value around 2500m, and the partial effect decreases after 2500m. Therefore, as we can see from the graph, there is a likely chance of obtaining a 50% field fraction within the range of 1500m to 2500m. This range in the map potentially can be the yellow color range, and most of this range is expected to have arable field fractions. Figure 19c, as the dry red edge band reflectance value increases, the effect on the field fraction will also increase, and most arable fields exist in the spectral reflectance value of 0.05 up to 0.2. There is a likely chance of obtaining 50% of arable field fractions within the range of 0.05 up to 0.2 reflectance value. The red edge (Band 5) map also shows a lower spectral reflectance around the center of the region, and the dark black color indicates the forest area. Figure 19d shows the sentinel-1 ratio of VV/VH similar characteristics up to 0.6, and beyond 0.6, its effect goes down. Most fields are concentrated in the range of 0.4 up to 0.6. As the slope increases in figure 19e, the partial effect on arable field fraction will decrease. The slope and the field fractions are negatively correlated, and mostly we can obtain 50% arable field fractions within the range of 0 up to 10 degrees. Above 10-degree slope decrease with decreasing field fractions.





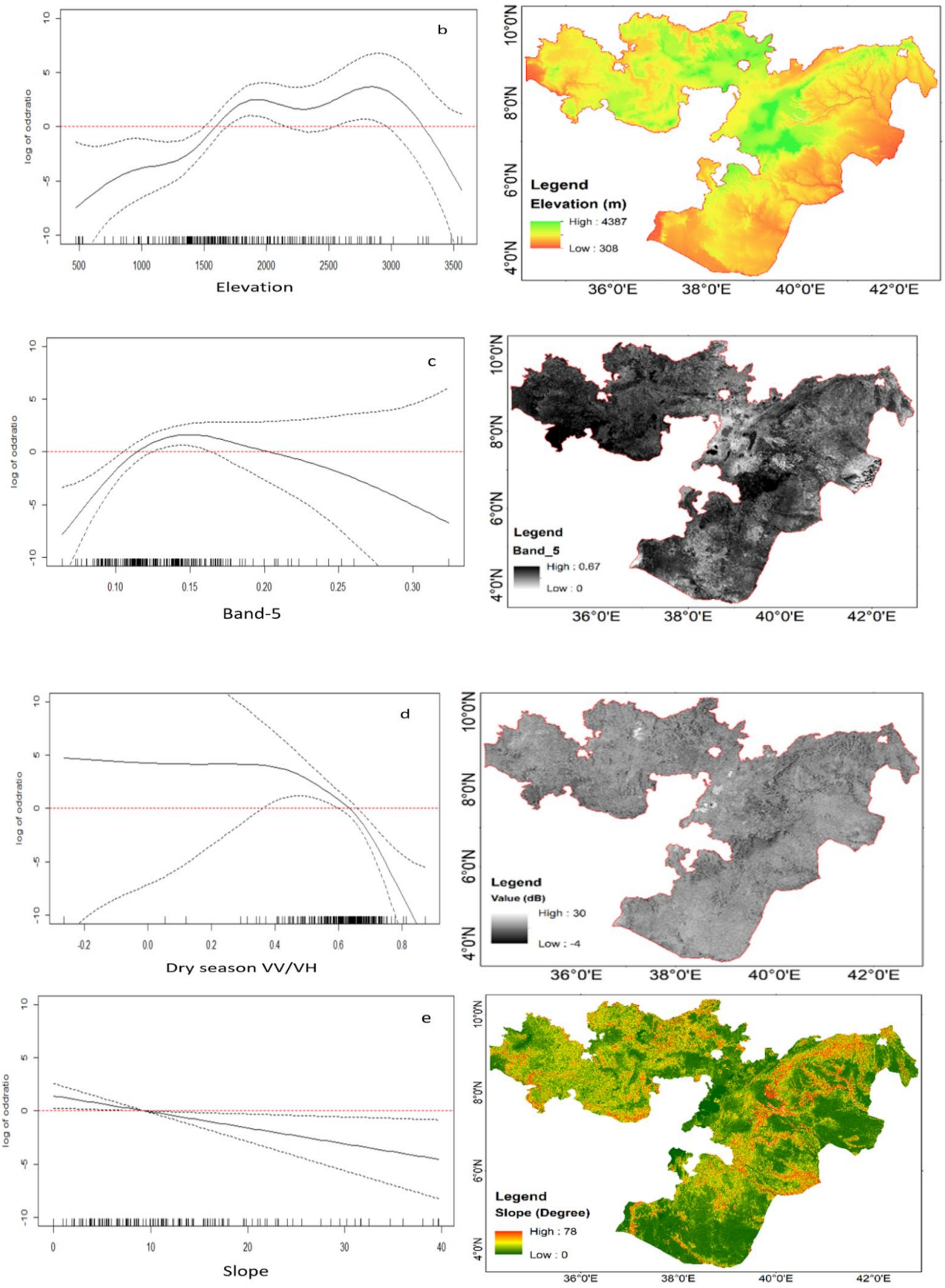


Figure 19 The partial effect of the explanatory variables on the actual field fractions (a-e) and the input variable map. The red line represents a 50% probability of field fraction, and the dots indicate the standard error.

4.2.5 Model Accuracy and Prediction

The performance of many machine learning models is evaluated based on their potential in predicting by using unseen or out-of-the-bag data. There are different model evaluation measures. In this study, the model's accuracy was evaluated by using Area Under Curve (AUC) and R^2 values evaluation techniques. A multiclass AUC evaluation technique is applied to identify AUC, and the 76% AUC value was obtained for the gradient boosted model. The model's accuracy by using the calibration data indicates a 69% R^2 value, and in the validation dataset, the R^2 value of 71%. The model accuracy of the classical Gam is the 74% AUC value and the R^2 value of 69%.

After checking the model accuracy and informative variable selection, the prediction was made on the six image features. The gradient boosted model performed well in ranking the most informative variable using early Stopping. Satbility selection method applied on the fitted Gradient boosted model and six most informative variables are identified. The classical GAM uses the six informative variables to estimate a 20m arable field fraction. The model result shows that the model captures the arable field characteristics of the study area. Figure 20 shows the 20m arable field estimate of classical GAM, and most of the arable fields are concentrated in the middle of the study area. The occurrence of arable field fraction in the southeastern part of the area is very low. The inset map shows a clear separation of forest cover and arable field fractions.

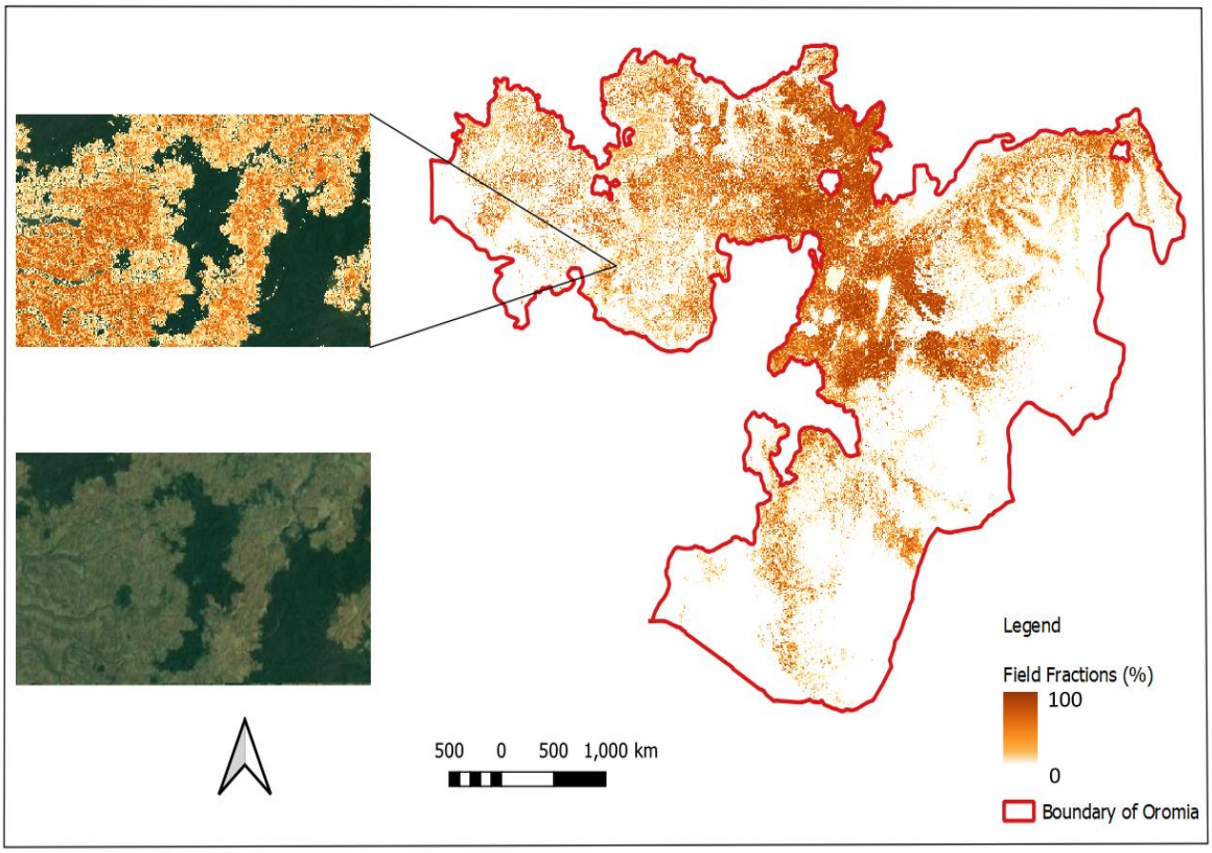


Figure 20 The Extrapolated 20m arable field fraction estimate

5. DISCUSSION

This research examined multisensor remote sensing data in estimating a 20m arable field probability estimate using the gradient boosted model for variable selection and classical Gam for prediction. The stratification of PROBA-V NDVI plays a significant role in estimating a 1km arable field fraction. In addition to this, the stratification of PROBA-V NDVI can discriminate the dry wet season of the study area. Most importantly, Sentinel 1 backscatter matrices have a higher relative importance in identifying the occurrence of arable field fraction. In this research, gradient boosted is performed well in selecting the most informative variable and classical GAM has the higher prediction on the estimation of arable field fractions. This section introduces in detail the main findings of the research based on the research questions and hypothesis.

5.1. 1km Field Fraction Estimation

Arable field estimation can be prepared by integrating hyper-temporal image (i.e., Prova-V NDVI) and agricultural district level area estimates. In this study, the hyper-temporal image plays a significant role in estimating arable field probability and in determining specific dates for downloading sentinel-1 and 2. One of the uses of the hyper-temporal image is for estimating a 1km arable field probability. Many researchers utilize this hyper-temporal image for estimating arable field probability by integrating it with agricultural statistics data. For example, De Bie et al. (2008) prepared a 1km crop field, and Mohammed et al. (2020) also uses a 1km hyper-temporal image to map the crop field. A 1km arable field probability estimate is produced in this study by integrating district-level agricultural statistics and Proba-V NDVI data. The result shows that an agreement of 89.6 adjusted R^2 and an area estimation of 3,917,033.1 hectares which accounts for a bit of difference with the reported regional level statistics of 4,077,968 hectares. This does not give any guaranty that the estimation report of the Central Statistics Agency is exact. However, such a result can support the report in determining specific areas within the district level (i.e., Woreda) rather than one estimate for the whole of Woreda, depending on the size of the district.

A 1km arable field estimate is also used as an input for the classical GAM model for estimating 20m arable field probability. The 1km arable field estimate entered into the model as a linear effect. The assumption is the actual field fraction and a 1km arable field probability estimate have a positive relationship (i.e., as the actual field fraction increases, the 1km arable field probability also increases). The result indicates the variable becomes the fifth predictor variable. Compared to Sentinel 1 and 2, topographic features a 1km estimate is the least predictor variable in estimating 20m arable field fraction. The main reason for a reduction in the importance of 1km estimate could be data issues meaning a 1km estimate contains much of zero probabilities) and modeling issues (i.e., gradient boosted model does not account for zero values meaning the model uses Betareg as a family and doesn't allow the response variable to be one or zero). When we compare the result with the classical GAM, a 1km estimate becomes the fourth informative variable, and the

deviance explained by this variable is 13.3%. Including a 1km arable field estimate in classical GAM improve the estimation of arable field fractions, whereas including a 1km arable field in the Gradient boost model does not improve the estimation of arable field fractions.

5.2. The relative importance of model input variables

Hyper temporal images also help us to download Sentinel-1 backscatter matrices. First, one should have to identify the time periods for downloading dry and wet season Sentinel-1 images. To identify these time periods, the Proba-V plays a significant role. Based on the specified periods, eight Sentinel-1 image features are downloaded and used as an input variable for the model. From those six variables, only off_season_VH and off_season_VV_VH ratio are the most informative image features in estimating a 20m arable field probability.

The backscatter matrices of Sentinel-1 play a major role in determining arable field, especially after harvesting and removing the crop's ruminant, due to interaction among the SAR signal and the soil (Nasirzadehdizaji et al., 2021b). One of the backscatter matrices of the Sentinel-1 image is the single co-polarization of the VH backscatter. In this research, the VH backscatter is identified as the influential variable in distinguishing the arable fields. Sentinel-1 (i.e., off_season_Vh) is the most informative class from the sentinel-1 image features and from all the model input variables used in the model. The model result Figure 19a shows within the range of -30dB up to -17dB, and there is the existence of an arable field. Especially, in between -25dB and -20dB, we can obtain high actual fractions (i.e., 100% field fractions). The main reason could be Sentinel-1 Vh polarization can detect the backscatter information from arable field characteristics.

Sentinel-1 Vh polarization can detect crop residue (i.e., stubble) and soil moisture. For instance, crop residue like stubble and straw are the main determinant of the radar backscatter (McNairn et al.,2002). In Ethiopia, mostly the months like October, November, and December are the harvesting periods, and after these periods, the dry seasons follow. During this dry period (February up to march) in the Oromia region, mostly arable fields are covered with crop residue and some land management practices. Even though this period is identified as dry seasons, there is also frost during the morning periods with little rain. This weather condition makes the soil moisture increase. In general, crop residue, land management practice, and some soil moisture significantly impact the sensitivity of the VH backscatter matrices. Therefore, we can conclude that the VH backscatter matrices have a significant impact in identifying Arable fields. This conclusion is supported by Sun et al., (2020) findings they indicate the VH intensity has the potential in discriminating arable field features. The Sentinel-1 VV_VH ratio also has a significant contribution in identifying arable fields. Mostly arable fields are concentrated in the range of 0.4 up to 0.6 dB. The sentinel 1 VV_VH ratio also has the potential to discriminate pasture land from other land features (Nicolau et al., 2021). Therefore, Sentinel-1 VH backscatter and VV over VH ratio significantly impact the estimation of arable field probability.

Many studies investigated the importance of topographic features for mapping arable fields. (Husak et al., 2008; M. T. Marshall et al., 2011; Mohammed et al., 2020) shows the importance of elevation and slope in estimating arable field probabilities. In this study, four topographic features were used to estimate 20m arable

field probabilities, and elevation becomes the most informative variable compared to the rest topographic features and followed by slope predictor variable. Elevation also the second most influential variable from all model input features. Primarily, the arable fields exist in the elevation range of 1500 up to 2500 and this result is similar to the previous study conducted in Oromia region (Mohammed et al., 2020). The other most informative variable is from the topographic features is the slope, and most arable field fractions exist in the range of 0 up to 10 degrees. The slope and elevation variable express the fragmented landscape characteristics of the study area and the existence of arable fields. Mostly cereal crops like Teff, wheat, and barley grow at an elevation range of 1800 to 2200, and corn grows at an elevation range of 1500 to 2200 meters. This confirms the potential of classical GAM model in determining the relationships among the explanatory variable (i.e., elevation) with the actual arable field fractions.

This study used the optical Sentinel-2 data for extracting three red-edge bands, two SWIR bands, dry and wet season NDTI, and LSWI. From these nine Sentinel-2 image features, Band 5 (Red edge 1) is the most informative variable, and the rest eight variables are not significant variables for estimating a 20m arable field probability. Mainly red edge bands can differentiate the most vegetated areas because of their spectral response of green vegetation (Weichelt et al., 2012). The dry season Band 5 (red edge) variable can exclude the more vegetated areas. Most arable fields in the study area are concentrated at the lower portion of their spectral reflectance values (i.e., with a Band 5 reflectance value of 0.1 up to 0.2). It is expected to have even lower than these spectral values for arable fields, but the study area arable field characterized in some parts with trees inside the field boundaries. This makes the rise in spectral reflectance value of the red edge band. The rest of Sentinel-2 drive products like the vegetation indices the dry, wet season NDVI, dry and wet season NDTI, and LSWI are not important explanatory variables. We can conclude the red edge band is one of the informative variables in estimating arable field fractions. Sun et al. (2020) also show the red edge (i.e., Band 5) as the most important variable for mapping arable field areas.

5.3. Model Evaluation

The gradient boosted model fitted with twenty-four explanatory variables and the response variable (i.e., Arable field fractions). From the twenty-four predictors, the boosted GAM chooses ten informative variables. To select these variables, boosted GAM used cross-validation as an early stopping mechanism for solving overfitting issues. An add-on feature (i.e., stabsel ()) function is applied on the boosted GAM to reduce feature reduction from the ten most informative variables to six predictor variables. Finally, the boosted GAM with stebsel () function identifies the most informative six variables, and the variables are used to estimate 20m arable field probabilities.

The research utilizes the advantage of the two models for estimating arable field probabilities and exploring the models' potential. Before using the classical GAM for prediction, the boosted Gam were tested in estimating arable field fractions. Table 8 shows the prediction result of classical GAM, Boosted GAM, the previous study result, and the agricultural statistics report. The final arable field fractions estimation result of

the classical GAM is 8,345,975 ha and around 600,000 ha difference from the previous study. When we compare with the reported agricultural data, the estimated area doubled the reported agricultural statistics.

Table 8 shows a 20m Arable field fractions estimation of Classical Gam, Boosted Gam, Previous study, and agricultural report

Type	R2	ROC	Area Estimation (ha)
Classical GAM	69	74	8,345,975
Boosted GAM	71	76	10327221.5
Previous study result	65	71	8,903,744
1Km Estimate	88.4	---	3917033.1
Agricultural Report	---	--	4,077,968

The result shows the overestimation of the arable field compared to statistical reports and the prediction estimate of classical GAM. As we can see from the table, the R^2 value of out-of-sample data for the classical GAM is lower than the gradient boost model. Even the Gradient boosted R^2 value is better than the previous study R^2 value. The same is true for the evaluation depending on the multiclass AUC value. The gradient boosted gam arable field fraction estimation is higher or overestimated than the classical. GAM and exceeded the classical GAM result by 1.4 million hectares. The result also doubles the area estimation of the reported agricultural statistics. The overestimation of the arable field is due to the response variable used in this model with zero and one values. In addition to this, the overestimation could be the classification of non crop areas as crop land, and the capability of the model (i.e., spectral classifier) is limited in discriminating and identifying unique signatures for mixed pixels (Husak et al., 2008). The model lacks capturing these extreme values of the response variable, and the boosted model with the family betaReg() lacks in considering the extreme values. The data issues are the main reason for choosing classical GAM. The classical GAM utilizes the quasi binomial distribution. It captures zero or one inflated response variable (i.e., The actual field fraction has over-dispersed zero and one values). The classical GAM gives a better estimation of arable field fractions than the Gradient boost estimates.

The 1km estimate of arable field fractions approximately equals the reported agricultural statistics data. Even though the estimation of a 1km estimate almost equals the reported agriculture statistics data, it may include the uncertainty with its resolution (1km). In 1Km resolution, we can get other land features, and also, it doesn't confirm the correctness of the agricultural statistics estimate of the Government.

The current and the previous study also compared based on the method, the number of input variables, and the relative importance of each variable included in both studies. The current study used two models (i.e., Gradient Boosted and Classical GAM) with twenty-four predictor variables and used six of the most informative variables to estimate a 20m arable field fractions. In contrast, in the previous study, only the

classical GAM is used with five predictor variables (i.e., a 1km arable field estimate, dry and wet season NDVI, elevation, and slope) to estimate a 30m arable field fractions. In the previous study, the most informative variable is a 1km estimate which is converted into dummy variables before running the classical GAM model. In the current study, the significancy of 1km estimate is low compared to the previous study. This variable is included in our model as a linear input variable in the Gradient boosted model and as a continuous input variable in the classical GAM model.

When we compare the partial deviance explained by each predictor (i.e., using common predictor variables in both studies) in the current and previous studies, there is a small difference in variations explained by each predictor variable. Table 9 shows the comparison of the current and previous studies based on the partial deviance of each predictor variable. In the current study, the elevation shows a higher variation explained than the previous study, but less variation is explained by a 1km arable field estimate in the current study compared to the previous study. The slope explains almost the same variation.

Table 9 shows a comparison of the current and previous studies based on the partial deviance explained by each predictor

Predictor Variable Name	Partial Deviance Explained (%)	
	Current Study	Previous Study
Elevation	17.2	12.2
A 1km arable field fractions	13.3	19.7
Slope	6.2	5.8

In General, the result suggests the gradient boosted model has higher prediction accuracy in out-of-sample data. This shows the potential of the boosted model, and it is wise to consider other possibilities to boost its capacity by increasing the number of ground sample data, especially by considering a 1km field fraction estimates (i.e., creating random sample points within the most significant classes that have more than 50% field fractions to surpass zero or one inflated actual field fractions or the response variable). In addition to this, more focus to include other explanatory variables, especially the raw bands rather than the vegetation matrices.

6. CONCLUSION AND RECOMMENDATION

6.1 CONCLUSION

This study explored the potential of multisensor remote sensing data like Sentinel-1\2 and topographic data using boosted and classical GAM models. From twenty-four predictor variables, six variables were identified by using boosted GAM model and stability selection method, which is a good model for selecting the most informative variables. The study identified Sentinel-1 backscatter matrices are the most informative variables in estimating arable field probabilities. Not only Sentinel-1 image features but also elevation and Sentinel-2 red edge bands are also the most significant predictor variables in estimating arable fields. The classical GAM model fitted with those six explanatory variables, and a 20m arable field fraction map is extrapolated. In general, the fusion of multisensor remote sensing data like Sentinel-1 microwave, Sentinel-2 optical, and topographic data is essential in estimating 20m arable field fractions.

6.2 RECOMMENDATIONS

A good Model prediction result depends on the number of input variables, training and validation dataset, selection of informative variables, and model used for the prediction. Most machine learning algorithms satisfy this requirement, like Random Forest and Gradient Boosted models. In this research, the potential of the gradient boosted model is explored, and the model overestimates 20m arable field fractions as compared to the classical GAM. The reason can be the Gradient boost model uses the family betaReg(), and this function doesn't allow the response variable to have zero and one. To maximize the prediction performance of the gradient boost model, one can explore the use of other families like zero or one inflated beta and other families to incorporate the response variable with zero and one. Further research needs to be done by incorporating suitable families that consider zero and one response values.

Generally, to increase the performance of the classical GAM and the Gradient boosted in the estimation of arable field fractions, it is good to include a 1km arable field estimate as a dummy variable rather than a linear or continuous feature. In addition to this, it needs a careful study to exclude some areas that are not belonging to arable fields by using the sixty-six Proba-V NDVI groups (i.e., in the current study, only group 1 is excluded) before the modeling process. More training and validation datasets also needed to be incorporated for obtaining an optimal estimation of arable field fractions.

LIST OF REFERENCES

Abdikan, S., Sekertekin, A., Ustunern, M., Sanli, F. B., & Nasirzadehdizaji, R. (2018). *Backscatter analysis using multi-temporal sentinel-1 sar data for crop growth of maize in Konya basin, Turkey*. https://doi.org/10.5194/isprs-archives-XLII-3-9-2018

Argaw, M. (2015). *Forestry and Land Use and Kulima Integrated Development Solutions Good Agricultural Adaptation Practices in Ethiopia Good Agricultural Adaptation Practices in Ethiopia 2*. Retrieved from www.kulima.com/agriculturaladaptation

Ashiagbor, G., Forkuo, E. K., Asante, W. A., Acheampong, E., Quaye-Ballard, J. A., Boamah, P., … Foli, E. (2020). Pixel-based and object-oriented approaches in segregating cocoa from forest in the Juabeso-Bia landscape of Ghana. *Remote Sensing Applications: Society and Environment*, *19*, 100349. https://doi.org/10.1016/j.rsase.2020.100349

Beltran-Abaunza, J. (2009). Method development to process hyper-temporal remote sensing (RS) images for change mapping. Retrieved from http://www.itc.nl/library/papers_2009/msc/gem/beltran-abaunza.pdf

Bouhamed, H., Lecroq, T., & Rebaï, A. (2012). *New Filter method for categorical variables' selection*. Retrieved from http://www-igm.univ-mlv.fr/~lecroq/articles/ijcsi2012.pdf

Bradley, B., & Brandon, G. (2020). *Chapter 12 Gradient Boosting. in Taylor and Francis (eds). Hands-On Machine Learning with R*. Retrieved from https://bradleyboehmke.github.io/HOML/gbm.html

Citores, L., Ibaibarriaga, L., Lee, D. J., Brewer, M. J., Santos, M., & Chust, G. (2020). Modelling species presence–absence in the ecological niche theory framework using shape-constrained generalized additive models. *Ecological Modelling*, *418*. https://doi.org/10.1016/j.ecolmodel.2019.108926

Craig, M. E. (National A. S. S. (2001). *The NASS Cropland Data Layer Program*. (November), 5–7. Retrieved from http://www.esri.com

Crommelinck, S., Bennett, R., Gerke, M., Nex, F., Yang, M. Y., & Vosselman, G. (2016). Review of automatic feature extraction from high-resolution optical sensor data for UAV-based cadastral mapping. *Remote Sensing*, *8*(8). https://doi.org/10.3390/rs8080689

De Bie, C. A. (2020a). W2 - Procedures and Tools to Process NDVI-Images: Spatio-temp. Analysis RS for food&water (iLecture). Retrieved from https://canvas.utwente.nl/courses/5206/pages/w2-procedures-and-tools-to-process-ndvi-images?module_item_id=147844

De Bie, C. A. (2020b). W2 - Procedures and Tools to Process NDVI-Images: Spatio-temp. Analysis RS for food&water (iLecture). Retrieved June 19, 2021, from https://canvas.utwente.nl/courses/5206/pages/w2-procedures-and-tools-to-process-ndvi-images?module_item_id=147844

De Bie, C. A. J. M. (n.d.). *Novel approaches to use rs-products for mapping and studying agricultural land use sytems*. Retrieved from https://webapps.itc.utwente.nl/librarywww/papers/0019.pdf

De Bie, C. A., Khan, M. R., Toxopeus, A. G., Venus, V., & Skidmore, A. K. (2008). Hypertemporal image analysis for crop mapping and change detection. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *37*(January). Retrieved from https://www.researchgate.net/publication/313102904_Hypertemporal_image_analysis_for_crop_mapping_and_change_detection

Debats, S. R., Luo, D., Estes, L. D., Fuchs, T. J., & Caylor, K. K. (2016). A generalized computer vision approach to mapping crop fields in heterogeneous agricultural landscapes. *Remote Sensing of Environment*, *179*, 210–221. https://doi.org/10.1016/j.rse.2016.03.010

Demeke, M., & Ferede, T. (2004). Agricultural Development in Ethiopia : Are There alternatives to Food Aid ? (Masters' Thesis). Retrieved from https://www.researchgate.net/publication/242555005

Eggen, M., Ozdogan, M., Zaitchik, B. F., & Simane, B. (2016). Land cover classification in complex and fragmented agricultural landscapes of the Ethiopian highlands. *Remote Sensing*, *8*(12).

https://doi.org/10.3390/rs8121020

Elder, J. A., Carter, W. H., Gennings, C., & Elswick, R. K. (1999). A Quasi-Likelihood Approach for Overdispersed Binomial Data When N Is Unobserved. In *Source: Journal of Agricultural, Biological, and Environmental Statistics* (Vol. 4). Retrieved from https://about.jstor.org/terms

ESA. (2014). ESA - Sentinel-1: seeing through clouds. Retrieved from https://www.esa.int/ESA_Multimedia/Videos/2014/03/Sentinel-1_seeing_through_clouds

ESA. (2017). User Guides - Sentinel-1 SAR - Sentinel Online. Retrieved from European Space Agency website: https://sentinel.esa.int/web/sentinel/user-guides/sentinel-1-sar/definitions

ESA. (2021). User Guides - Sentinel-1 SAR - Interferometric Wide Swath - Sentinel Online - Sentinel. Retrieved from https://sentinel.esa.int/web/sentinel/user-guides/sentinel-1-sar/acquisition-modes/interferometric-wide-swath

FAO. (2016). *FAOSTAT, Methods & standards*. Retrieved from http://www.fao.org/ag/agn/nutrition/Indicatorsfiles/Agriculture.pdf

FAO. (2017). *Methodology for Estimation of Crop Area and Crop Yield under Mixed and Continuous Cropping*. Retrieved from http://gsars.org/wp-content/uploads/2017/03/TR-15.03.2017-Methodology-for-Estimation-of-Crop-Area-and-Crop-Yield-under-Mixed-and-Continuous-Cropping.pdf

FAO. (2018). *Ethiopia at a glance*. Retrieved from http://www.fao.org/ethiopia/fao-in-ethiopia/ethiopia-at-a-glance/en/

FAO. (2020a). *Desert locust invasion in Ethiopia : FAO in Emergencies*. Retrieved from http://www.fao.org/emergencies/resources/photos/photo-detail/en/c/1257696/

FAO. (2020b). Sustainable Development Goals : 2.4.1 Agricultural sustainability. Retrieved from FAO website: http://www.fao.org/sustainable-development-goals/indicators/241/en/

Fawwaz T. Ulabay, Richards K., M., & Adrian K., F. (2015). Microwave Radar and Radiometric Remote Sensing [Book Reviews]. In *IEEE Geoscience and Remote Sensing Magazine* (Vol. 3). https://doi.org/10.1109/mgrs.2015.2398391

FEWS NET. (2020). *ETHIOPIA Food Security Outlook : Crisis (IPC Phase 3) outcomes likely to persist due to below-average seasonal rainfall*. Retrieved from https://reliefweb.int/sites/reliefweb.int/files/resources/ET_OL_June 2019_ January 2020 ___.pdf

Filgueiras, R., Mantovani, E. C., Althoff, D., Fernandes Filho, E. I., & da Cunha, F. F. (2019). Crop NDVI monitoring based on sentinel 1. *Remote Sensing*, *11*(12). https://doi.org/10.3390/rs11121441

Getahun, A. (2020). Smallholder Farmers Agricultural Commercialization in Ethiopia: A Review. *Agriculture, Forestry and Fisheries*, *9*(3), 67. https://doi.org/10.11648/j.aff.20200903.14

Gilles, L., Louis, W., Antonio, S., & Pierre, G. (2020). *Understanding variable importances in forests of randomized trees*. Retrieved from https://www.researchgate.net/publication/264046801_Understanding_variable_importances_in_Forests_of_randomized_trees

Gumma, M. K., Tummala, K., Dixit, S., Collivignarelli, F., Holecz, F., Kolli, R. N., & Whitbread, A. M. (2020). Crop type identification and spatial mapping using Sentinel-2 satellite data with focus on field-level information. *Geocarto International*, *0*(0), 1–17. https://doi.org/10.1080/10106049.2020.1805029

Guyon, I., & De, A. M. (2003). An Introduction to Variable and Feature Selection André Elisseeff. In *Journal of Machine Learning Research* (Vol. 3). Retrieved from https://dl.acm.org/doi/10.5555/944919.944968

Haack, B., & Bechdol, M. (1999). Multisensor remote sensing data for land use/cover mapping. *Computers, Environment and Urban Systems*, *23*(1), 53–69. https://doi.org/10.1016/S0198-9715(99)00003-4

Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. In *Statistical Science* (Vol. 1). Retrieved from https://www.jstor.org/stable/2245459?seq=1#metadata_info_tab_contents

He, W., & Yokoya, N. (2018). Multi-temporal sentinel-1 and -2 data fusion for optical Image Simulation. *ISPRS International Journal of Geo-Information*, *7*(10). https://doi.org/10.3390/ijgi7100389

Hofner, B., Boccuto, L., & Göker, M. (2015). Controlling false discoveries in high-dimensional situations: Boosting with stability selection. *BMC Bioinformatics*, *16*(1). https://doi.org/10.1186/s12859-015-0575-3

Hofner, B., Mayr, A., Robinzonov, N., & Schmid, M. (2014). Model-based boosting in R: A hands-on tutorial using the R package mboost. *Computational Statistics*, *29*(1–2), 3–35. https://doi.org/10.1007/s00180-012-0382-5

Holzinger, E. R., Szymczak, S., Malley, J., Pugh, E. W., Ling, H., Griffith, S., … Bailey-Wilson, J. E. (2014). Comparison of parametric and machine methods for variable selection in simulated Genetic Analysis Workshop 19 data. *From Genetic Analysis Workshop*, *19*, 24–26. https://doi.org/10.1186/s12919-016-0021-1

Huang, Y., Chen, Z. xin, YU, T., Huang, X. zhi, & Gu, X. fa. (2018). Agricultural remote sensing big data: Management and applications. *Journal of Integrative Agriculture*, *17*(9), 1915–1931. https://doi.org/10.1016/S2095-3119(17)61859-8

Husak, G. J., Marshall, M. T., Michaelsen, J., Pedreros, D., Funk, C., Galu, G., & Husak, C. : (2008). Crop area estimation using high and medium resolution satellite imagery in areas with complex topography. *J. Geophys. Res*, *113*, 14112. https://doi.org/10.1029/2007JD009175

IHSN. (2006). Agricultural Sample Enumeration, Area and Production 2001-2002 (1994 E.C). Retrieved from http://catalog.ihsn.org/index.php/catalog/1438

Jenkins, I. (2005). Book review: Book review. *Journal of Computer Assisted Learning*, *16*(3), 280–280. https://doi.org/10.1046/j.1365-2729.2000.00139.x

Jović, A., Brkić, K., & Bogunović, N. (2012). *A review of feature selection methods with applications*. Retrieved from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7160458

Khabbazan, S., Vermunt, P., Steele-Dunne, S., Arntz, L. R., Marinetti, C., van der Valk, D., … van der Sande, C. (2019). Crop monitoring using Sentinel-1 data: A case study from The Netherlands. *Remote Sensing*, *11*(16). https://doi.org/10.3390/rs11161887

Kulo, N. (2020). *Multisensor Remote Sensing Data Integration . Geodetski glasnik DALJINSKIH ISTRAŽIVANJA*. (January). Retrieved from https://www.researchgate.net/publication/338344898_Multisensor_Remote_Sensing_Data_Integration_Geodetski_glasnik

Kumari, M., Murthy, C. S., Pandey, V., & Bairagi, G. D. (2019). *SOYBEAN CROPLAND MAPPING USING MULTI-TEMPORAL SENTINEL-1 DATA*. https://doi.org/10.5194/isprs-archives-XLII-3-W6-109-2019

Landgrebe, T. C. W., & Duin, R. P. W. (2007). *Approximating the multiclass ROC by pairwise analysis*. https://doi.org/10.1016/j.patrec.2007.05.001

Lemoine, G. (2018). *Introduction to Sentinel-1 Sentinel data use for CAP monitoring and control Training module at the Vilnius IACS Workshop*. Retrieved from https://marswiki.jrc.ec.europa.eu/wikicap/images/6/6c/Sentinel1Training.pdf

López-Caloca, A. A., Tapia-Silva, F. O., & Rivera, G. (2018). Sentinel-1 Satellite Data as a Tool for Monitoring Inundation Areas near Urban Areas in the Mexican Tropical Wet. In *Water Challenges of an Urbanizing World*. https://doi.org/10.5772/intechopen.71395

Ma, C., Li, X., & McCabe, M. F. (2020). Retrieval of high-resolution soil moisture through combination of Sentinel-1 and Sentinel-2 data. *Remote Sensing*, *12*(14), 1–28. https://doi.org/10.3390/rs12142303

Maloney, K. O., Schmid, M., & Weller, D. E. (2012). Applying additive modelling and gradient boosting to assess the effects of watershed and reach characteristics on riverine assemblages. *Methods in Ecology and Evolution*, *3*(1), 116–128. https://doi.org/10.1111/j.2041-210X.2011.00124.x

Marshall, M., Crommelinck, S., Kohli, D., Perger, C., Yang, M. Y., Ghosh, A., … Nelson, A. (2019).

Crowd-Driven and Automated Mapping of Field Boundaries in Highly Fragmented Agricultural Landscapes of Ethiopia with Very High Spatial Resolution Imagery. https://doi.org/10.3390/rs11182082

Marshall, M. T., Crommelinck, S., Kohli, D., Perger, C., Yang, M. Y., Ghosh, A., … Nelson, A. (2019). Crowd-driven and automated mapping of field boundaries in highly fragmented agricultural landscapes of Ethiopia with very high spatial resolution imagery. *Remote Sensing*, *11*(18), 1–17. https://doi.org/10.3390/rs11182082

Marshall, M. T., Husak, G. J., Michaelsen, J., Funk, C., Pedreros, D., & Adoum, A. (2011). Testing a high-resolution satellite interpretation technique for crop area monitoring in developing countries. *International Journal of Remote Sensing*, *32*(23), 7997–8012. https://doi.org/10.1080/01431161.2010.532168

Mc Nairn, H., Duguay, C., Boisvert, J., Huffman, E., & Brisco, B. (2001). Defining the sensitivity of multi-frequency and multi-polarized radar backscatter to post-harvest crop residue. *Canadian Journal of Remote Sensing*, *27*(3), 247–263. https://doi.org/10.1080/07038992.2001.10854941

McCarty, J. L., Neigh, C. S. R., Carroll, M. L., & Wooten, M. R. (2017). Extracting smallholder cropped area in Tigray, Ethiopia with wall-to-wall sub-meter WorldView and moderate resolution Landsat 8 imagery. *Remote Sensing of Environment*, *202*, 142–151. https://doi.org/10.1016/j.rse.2017.06.040

McNairn, H., Duguay, C., Brisco, B., & Pultz, T. J. (2002). The effect of soil and crop residue characteristics on polarimetric radar response. *Remote Sensing of Environment*, *80*(2), 308–320. https://doi.org/10.1016/S0034-4257(01)00312-1

Moeslund, J. E., Bøcher, P. K., Odgaard, M. V., Svenning, J. C., Arge, Á. L., Dalgaard, Á. T., & Ejrnaes, R. (2013). Topographically controlled soil moisture drives plant diversity patterns within grasslands. *Biodivers Conserv*, *22*, 2151–2166. https://doi.org/10.1007/s10531-013-0442-3

Mohammed, I., Marshall, M. T., De Bie, C. A., Estes, L., & Nelson, A. (2020). A blended census and multiscale remote sensing approach to probabilistic cropland mapping in complex landscapes. *ISPRS Journal of Photogrammetry and Remote Sensing*, *161*, 233–245. https://doi.org/10.1016/j.isprsjprs.2020.01.024

Murase, H., Nagashima, H., Yonezaki, S., Matsukura, R., & Kitakado, T. (2009). Application of a generalized additive model (GAM) to reveal relationships between environmental factors and distributions of pelagic fish and krill: A case study in Sendai Bay, Japan. *ICES Journal of Marine Science*, *66*(6), 1417–1424. https://doi.org/10.1093/icesjms/fsp105

Muzari, W. (2016). Agricultural Productivity and Food Security in Sub-Saharan Africa. *International Journal of Science and Research (IJSR)*, *5*(1), 1769–1776. https://doi.org/10.21275/v5i1.23011601

Nasirzadehdizaji, R., Cakir, Z., Balik Sanli, F., Abdikan, S., Pepe, A., & Calò, F. (2021a). Sentinel-1 interferometric coherence and backscattering analysis for crop monitoring. *Computers and Electronics in Agriculture*, *185*, 106118. https://doi.org/10.1016/j.compag.2021.106118

Nasirzadehdizaji, R., Cakir, Z., Balik Sanli, F., Abdikan, S., Pepe, A., & Calò, F. (2021b). Sentinel-1 interferometric coherence and backscattering analysis for crop monitoring. *Computers and Electronics in Agriculture*, *185*, 106118. https://doi.org/10.1016/j.compag.2021.106118

Nicolai, M., & Peter, B. (2010). *stebility selection.pdf*. Retrieved from https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/j.1467-9868.2010.00740.x

Nicolau, A. P., Flores-Anderson, A., Griffin, R., Herndon, K., & Meyer, F. J. (2021). Assessing SAR C-band data to effectively distinguish modified land uses in a heavily disturbed Amazon forest. *International Journal of Applied Earth Observation and Geoinformation*, *94*, 102214. https://doi.org/10.1016/j.jag.2020.102214

Niklas Donges. (2019). The Random Forest Algorithm: A Complete Guide. Retrieved from https://builtin.com/data-science/random-forest-algorithm

OCHA. (2020). *Desert locusts in East Africa_ A plague of another order _ Mercy Corps*. Retrieved from https://reliefweb.int/report/ethiopia/desert-locusts-east-africa-plague-another-order

Pastorino, M., Montaldo, A., Fronda, L., Hedhli, I., Moser, G., Serpico, S. B., & Zerubia, J. (2021). Multisensor and multiresolution remote sensing image classification through a causal hierarchical markov framework and decision tree ensembles. *Remote Sensing*, *13*(5), 1–25. https://doi.org/10.3390/rs13050849

Persello, C., Tolpekin, V. A., Bergado, J. R., & de By, R. A. (2019). Delineation of agricultural fields in smallholder farms from satellite images using fully convolutional networks and combinatorial grouping. *Remote Sensing of Environment*, *231*. https://doi.org/10.1016/j.rse.2019.111253

Qiu, S., He, B., Yin, C., & Liao, Z. (2017). Assessments of Sentinel-2 vegetation red-edge spectral bands for improving land cover classification. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *42*(2W7), 871–874. https://doi.org/10.5194/isprs-archives-XLII-2-W7-871-2017

Raschka, S. (2021). What is the difference between filter, wrapper, and embedded methods for feature selection? Retrieved from https://sebastianraschka.com/faq/docs/feature_sele_categories.html

Salik, A. W., & Karacabey, E. (2019). Application of Landsat 8 Satellite Image – NDVI Time Series for Crop Phenology Mapping: Case Study Balkh and Jawzjan Regions of Afghanistan. *Journal of Graduate School of Natural and Applied Sciences*, *5*(1), 49–62. https://doi.org/10.28979/comufbed.557792

Saurav, K. (2016). Feature Selection Methods | Machine Learning. Retrieved from https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/

Scarrott, R. G., Cawkwell, F., Jessopp, M., O'Rourke, E., Cusack, C., & de Bie, K. (2019). From land to sea, a review of hypertemporal remote sensing advances to support ocean surface science. *Water (Switzerland)*, *11*(11). https://doi.org/10.3390/w11112286

Schmid, M., Wickler, F., Maloney, K. O., Mitchell, R., Fenske, N., & Mayr, A. (2013). Boosted Beta Regression. *PLoS ONE*, *8*(4). https://doi.org/10.1371/journal.pone.0061623

See, L., Fritz, S., You, L., Ramankutty, N., Herrero, M., Justice, C., … Obersteiner, M. (2015). Improved global cropland data as an essential ingredient for food security. *Global Food Security*, *4*, 37–45. https://doi.org/10.1016/j.gfs.2014.10.004

See, L., Mccallum, I., Fritz, S., Perger, C., Kraxner, F., Obersteiner, M., … Kalita, N. R. (2013). Mapping Cropland in Ethiopia Using Crowdsourcing. *International Journal of Geosciences*, *4(6)*, 6–13. https://doi.org/10.4236/ijg.2013.46A1002.

Sun, L., Chen, J., Guo, S., Deng, X., & Han, Y. (2020). Integration of time series sentinel-1 and sentinel-2 imagery for crop type mapping over oasis agricultural areas. *Remote Sensing*, *12*(1), 1–29. https://doi.org/10.3390/RS12010158

Taffese, A. S., Dorosh, P., & Gemessa, S. A. (2013). Crop production in Ethiopia: Regional patterns and trends. *Food and Agriculture in Ethiopia: Progress and Policy Challenges*, *9780812208*, 53–83. https://doi.org/10.9783/9780812208610.53

Tatsumi, K., Yamashiki, Y., Canales Torres, M. A., & Taipe, C. L. R. (2015). Crop classification of upland fields using Random forest of time-series Landsat 7 ETM+ data. *Computers and Electronics in Agriculture*, *115*, 171–179. https://doi.org/10.1016/j.compag.2015.05.001

Thirtle, C., Lin, L., & Piesse, J. (2003). The impact of research-led agricultural productivity growth on poverty reduction in Africa, Asia and Latin America. *World Development*, *31*(12), 1959–1975. https://doi.org/10.1016/j.worlddev.2003.07.001

Thomas, J., Mayr, A., Bischl, B., Schmid, M., Smith, A., & Hofner, B. (2017). *Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates*. https://doi.org/10.1007/s11222-017-9754-6

Thomas, J., Mayr, A., Bischl, B., Schmid, M., Smith, A., Hofner, B., … Hofner, B. (2016). *Stability selection for component-wise gradient boosting in multiple dimensions*. https://doi.org/10.1007/s11222-017-9754-6

UNDP. (2018). *Ethiopia National Human Development Report 2018 : Industrialization with a Human Face*.

Retrieved from http://hdr.undp.org/sites/default/files/ethiopia_national_human_development_report_2018.pdf

Vreugdenhil, A. ;, Navacchi, M. ;, Bauer-Marschallinger, C. ;, Hahn, B. ;, Steele-Dunne, S. ;, Pfeil, S. ;, … Wagner, W. (2020). Sentinel-1 cross ratio and vegetation optical depth: A comparison over. *Europe. Remote Sensing*, *12*(20), 1. https://doi.org/10.3390/rs12203404

Weichelt, H., Rosso, P., Marx, A., Reigber, S., Douglass, K., & Heynen, M. (2012). The RapidEye Red Edge Band. *White Paper*, 1–6. Retrieved from http://www.rapideye.com/upload/Red_Edge_White_Paper.pdf

Wilson, N. R., Norman, L. M., Villarreal, M., Gass, L., Tiller, R., & Salywon, A. (2016). Comparison of remote sensing indices for monitoring of desert cienegas. *Arid Land Research and Management*, *30*(4), 460–478. https://doi.org/10.1080/15324982.2016.1170076

Xie, Y., Sha, Z., & Yu, M. (2008). Remote sensing imagery in vegetation mapping: a review. *Journal of Plant Ecology*, *1*(1), 9–23. https://doi.org/10.1093/jpe/rtm005

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

Xiong, J., Thenkabail, P. S., Tilton, J. C., Gumma, M. K., Teluguntla, P., Oliphant, A., … Gorelick, N. (2016). Nominal 30-m Cropland Extent Map of Continental Africa by Integrating Pixel-Based and Object-Based Algorithms Using Sentinel-2 and Landsat-8 Data on Google Earth Engine. *Remote Sens*, *9*, 1065. https://doi.org/10.3390/rs9101065

Yıldırım, S. (2020a). Random Forests vs Gradient Boosted Decision Trees. Retrieved from Artificial Intelligence in Plain English website: https://ai.plainenglish.io/random-forests-vs-gradient-boosted-decision-trees-dd4f0ef86554

Yıldırım, S. (2020b). Random Forests vs Gradient Boosted Decision Trees | by Soner Yıldırım | Artificial Intelligence in Plain English. Retrieved from https://ai.plainenglish.io/random-forests-vs-gradient-boosted-decision-trees-dd4f0ef86554

Zhang, H., Kang, J., Xu, X., & Zhang, L. (2020a). *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*. https://doi.org/10.1016/j.compag.2020.105618

Zhang, H., Kang, J., Xu, X., & Zhang, L. (2020b). 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. https://doi.org/10.1016/j.compag.2020.105618

Zheng, B., Myint, S. W., Thenkabail, P. S., & Aggarwal, R. M. (2015). A support vector machine to identify irrigated crop types using time-series Landsat NDVI data. *International Journal of Applied Earth Observation and Geoinformation*, *34*(1), 103–112. https://doi.org/10.1016/j.jag.2014.07.002

Zoubin, G. (2018). *Automatic Machine Learning: Methods, Systems, Challenges*. Retrieved from https://www.automl.org/wp-content/uploads/2018/12/automl_book.pdf

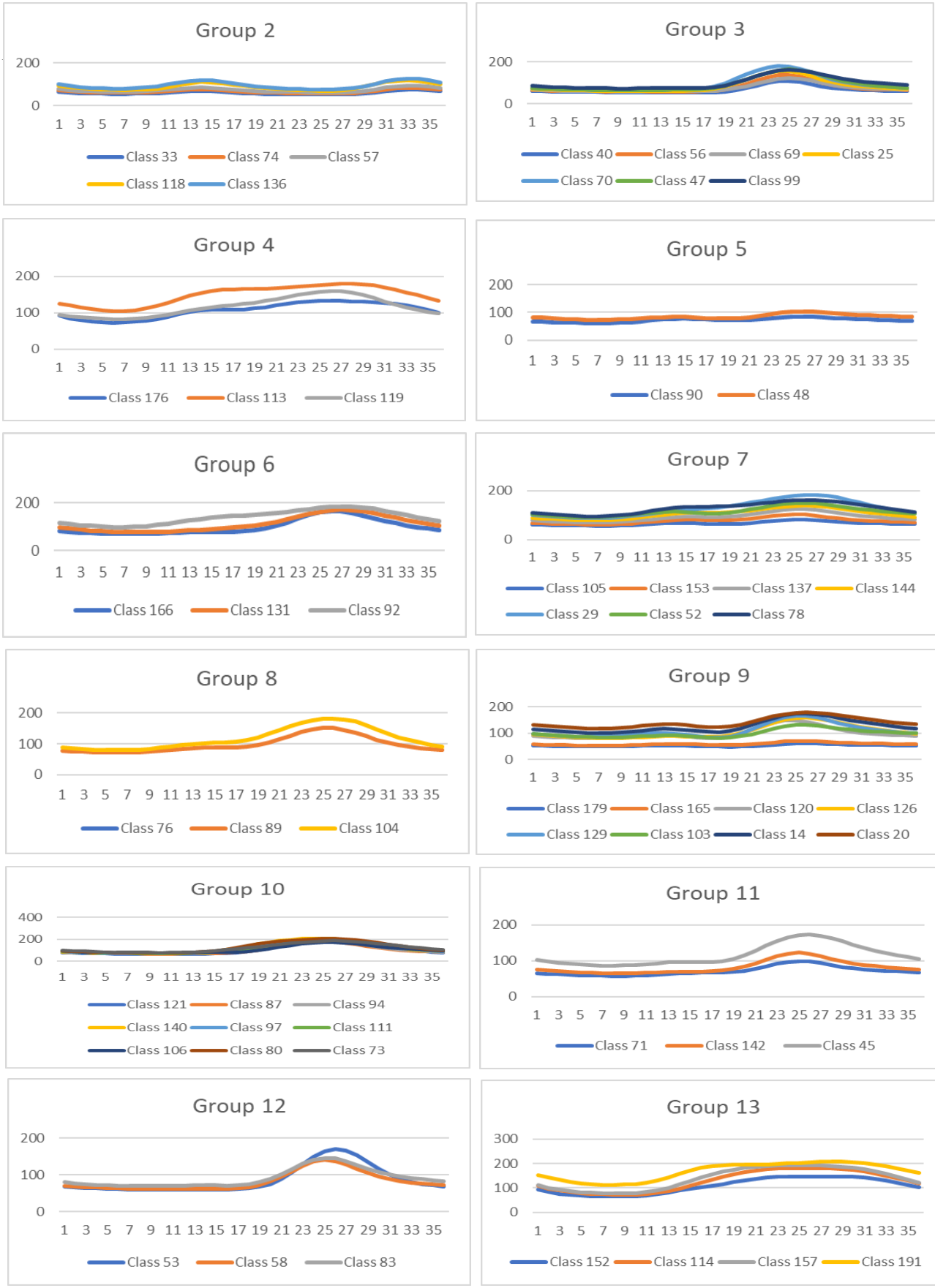
ANNEXES

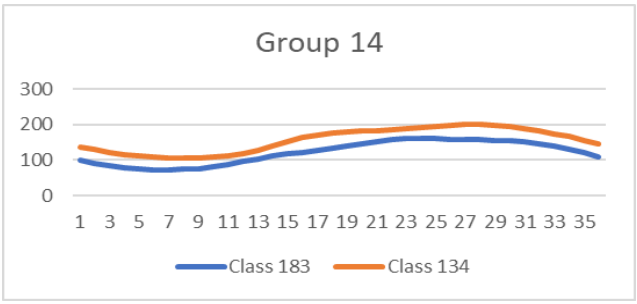
Annex 1. Code used to prepare Topographic Wetness Index (TWI)

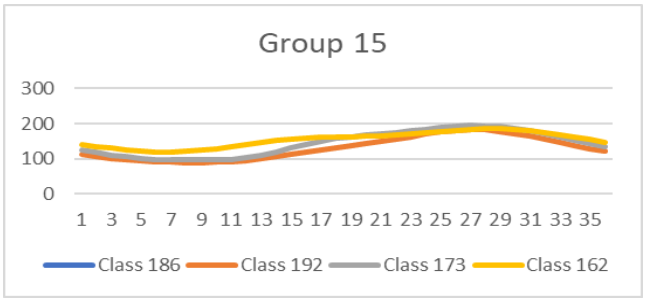
The following code is used for preparing TWI by integrating this code with Arc toolbox.

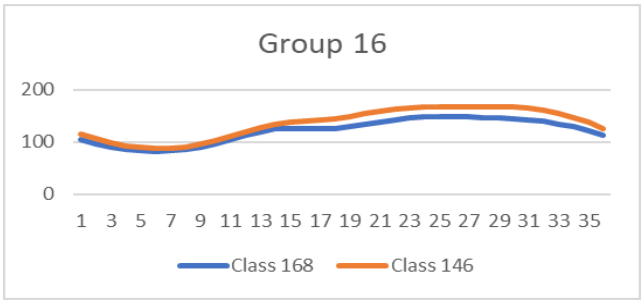
```
import arcpy, math
if __name__ == '__main__':
        arcpy.CheckOutExtension("Spatial")

        # Define workspace and set input and output files
        arcpy.env.workspace = arcpy.GetParameterAsText(0)
        inDEM = arcpy.GetParameterAsText(1)
        outTWI = arcpy.GetParameterAsText(2)

        # Intermediates
        arcpy.AddMessage("Filling DEM.\n")
        DEM_filled = arcpy.sa.Fill(inDEM)
        arcpy.AddMessage("Creating flow direction.\n")
        outFlowDirection = arcpy.sa.FlowDirection(DEM_filled, "FORCE")
        arcpy.AddMessage("Creating flow accumulation.\n")
        #outFlowAccumulation = arcpy.sa.FlowAccumulation(outFlowDirection, "", "FLOAT") + 1
        outFlowAccumulation = arcpy.sa.FlowAccumulation(outFlowDirection, "", "INTEGER") + 1

        arcpy.AddMessage("Creating slope.\n")
        slope = arcpy.sa.Slope(DEM_filled)
        arcpy.AddMessage("Converting slope in degrees to slope in radians")
        # 2Pi radians = 360 degrees
        # Pi radians = 180 degrees
        # conversion: Pi radians/180 degress
        slope_radians = slope * math.pi/180.0

        # Output
        arcpy.AddMessage("Creating TWI\n")
        TWI = arcpy.sa.Ln(outFlowAccumulation / (arcpy.sa.Tan(slope_radians)+.01))
        TWI.save(outTWI)
        arcpy.AddMessage("Saved TWI. Done.")
```

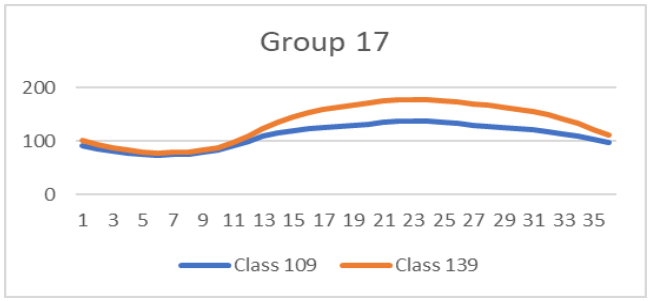
Annex 2. The NDVI groups (i.e., 66 groups)

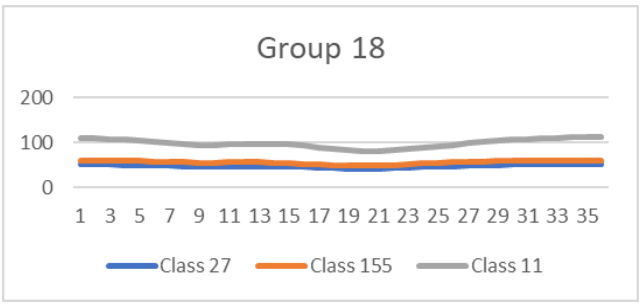


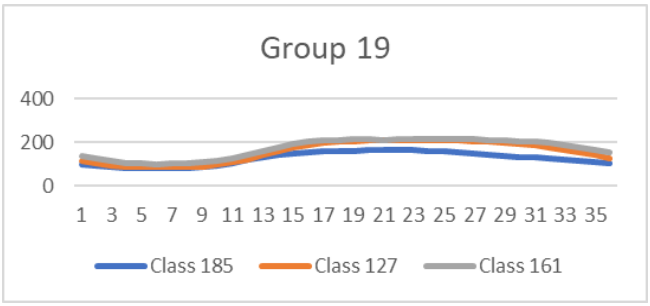

Group 14
300
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 183
Class 134

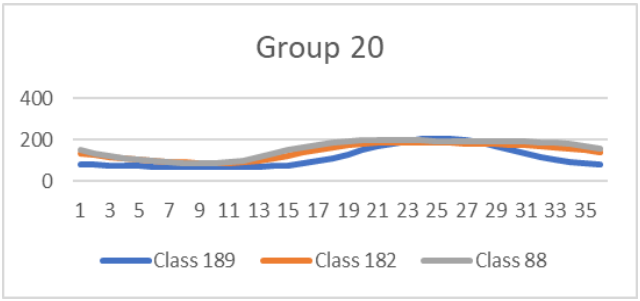

Group 15
300
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 186
Class 192
Class 173
Class 162

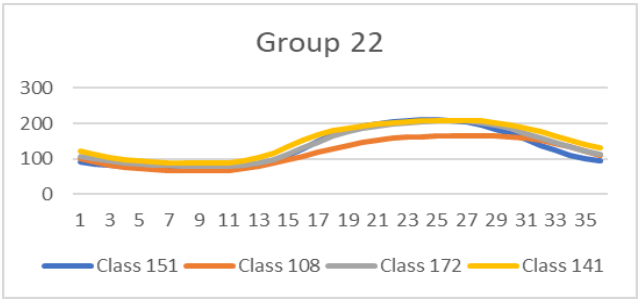

Group 16
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 168
Class 146

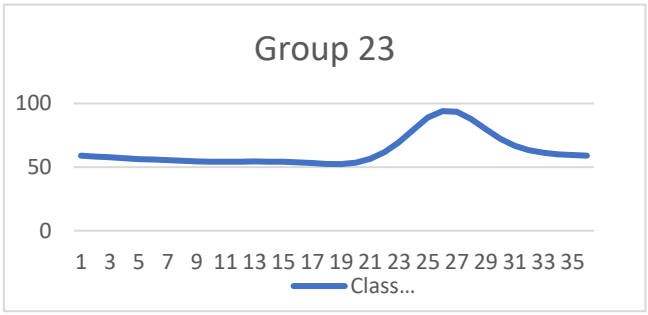

Group 17
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 109
Class 139

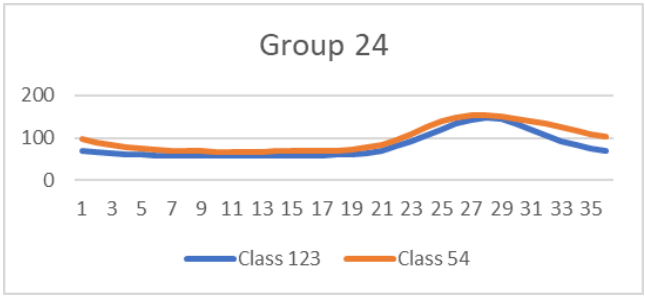

Group 18
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 27
Class 155
Class 11

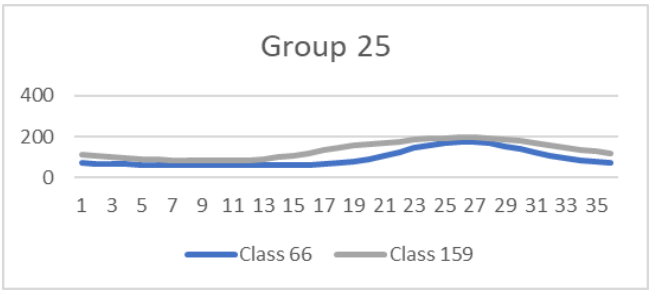

Group 19
400
200
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 185
Class 127
Class 161

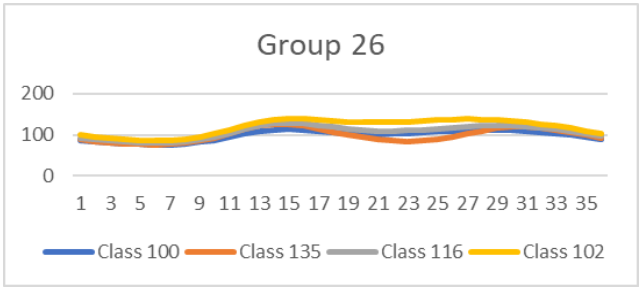

Group 20
400
200
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 189
Class 182
Class 88

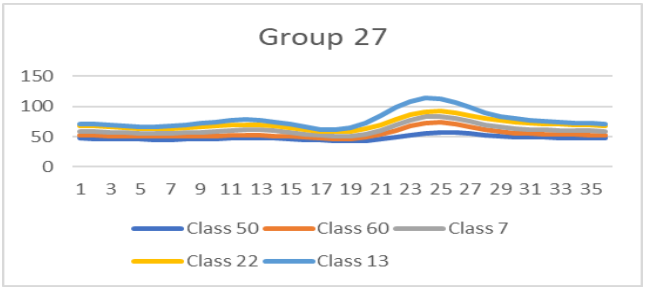

Group 22
300
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 151
Class 108
Class 172
Class 141

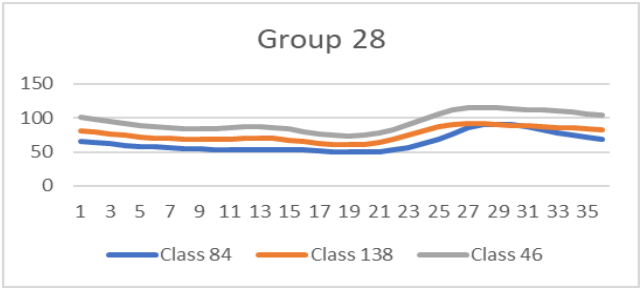

Group 23
100
50
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class...

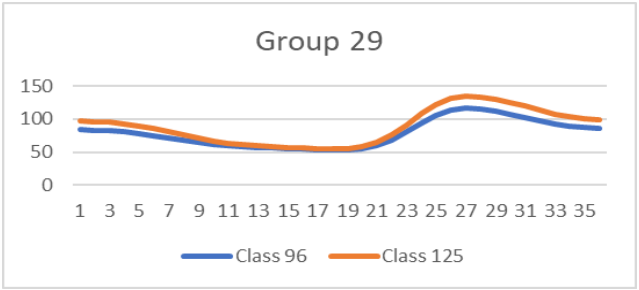

Group 24
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 123
Class 54

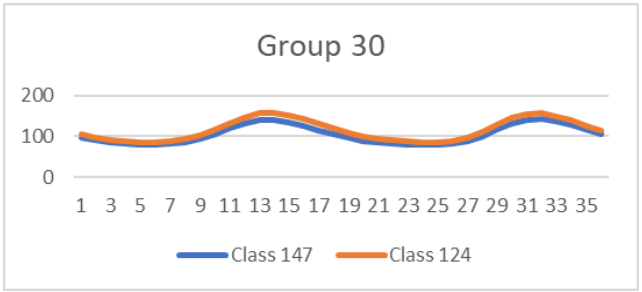

Group 25
400
200
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 66
Class 159

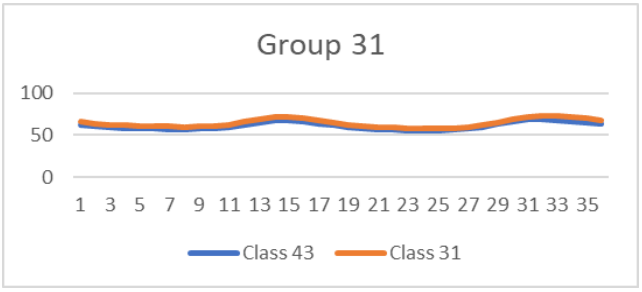

Group 26
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 100
Class 135
Class 116
Class 102

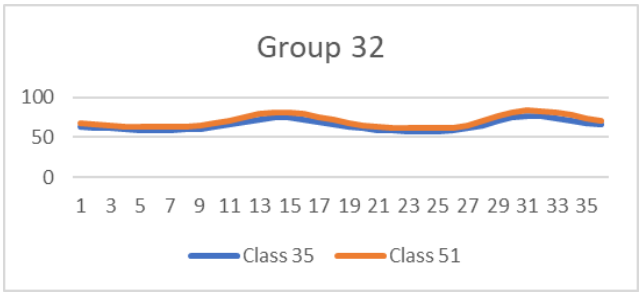

Group 27
150
100
50
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 50
Class 60
Class 7
Class 22
Class 13

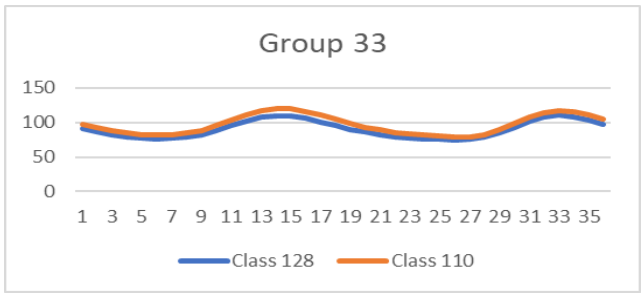

Group 28
150
100
50
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 84
Class 138
Class 46

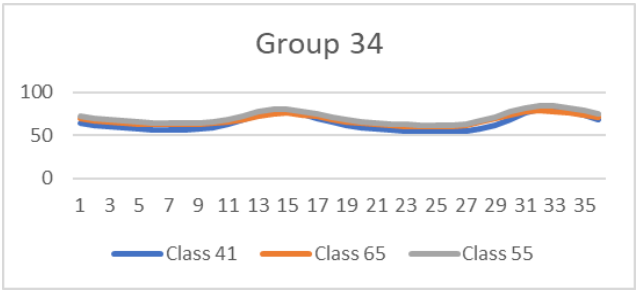

Group 29
150
100
50
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 96
Class 125

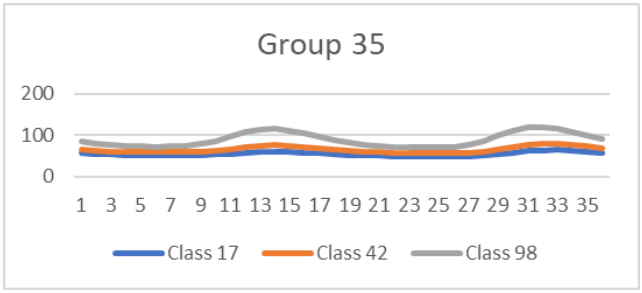

Group 30
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 147
Class 124

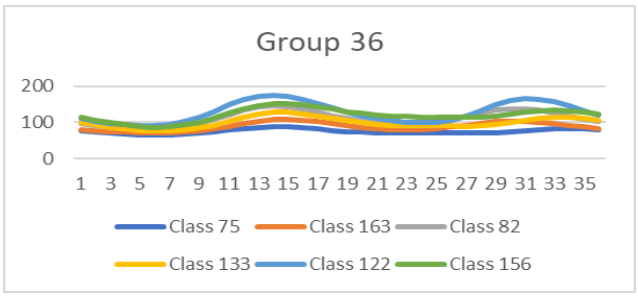

Group 31
100
50
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 43
Class 31

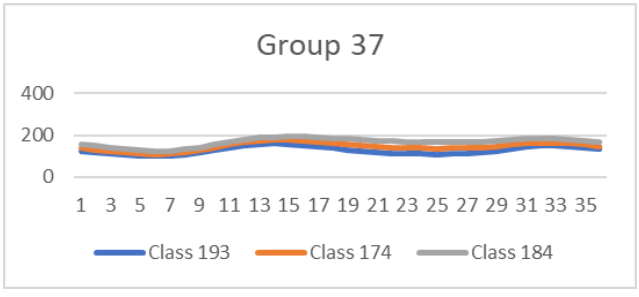

Group 32
100
50
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 35
Class 51

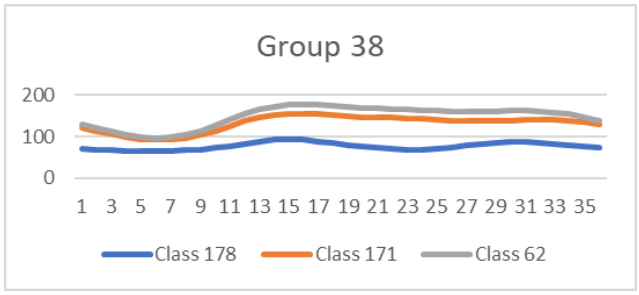

Group 33
150
100
50
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 128
Class 110

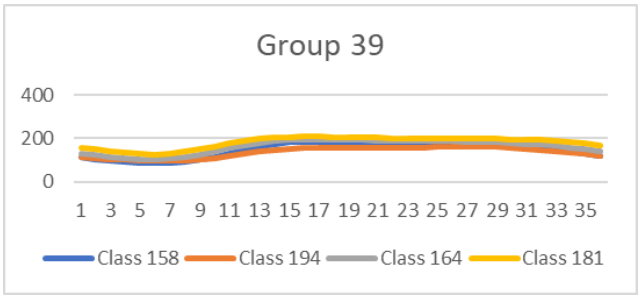

Group 34
100
50
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 41
Class 65
Class 55

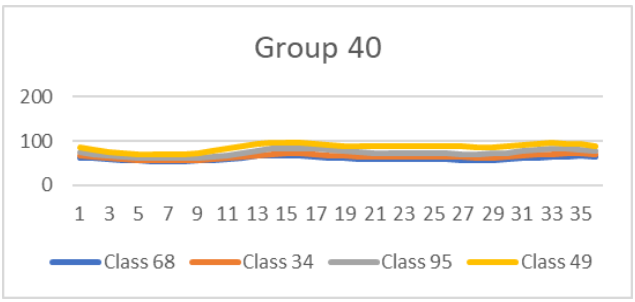

Group 35
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 17
Class 42
Class 98

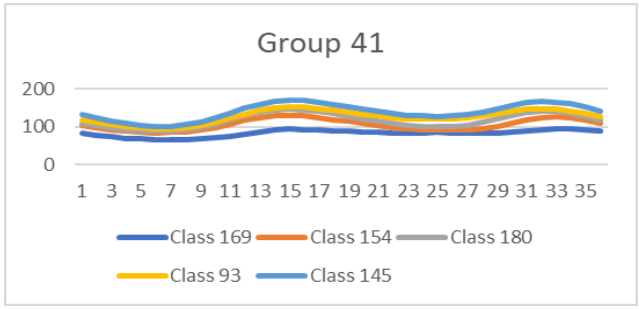

Group 36
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 75
Class 163
Class 82
Class 133
Class 122
Class 156

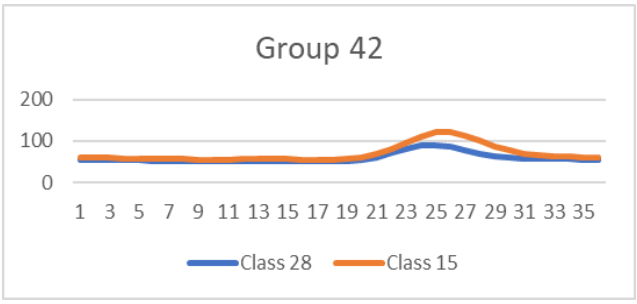

Group 37
400
200
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 193
Class 174
Class 184

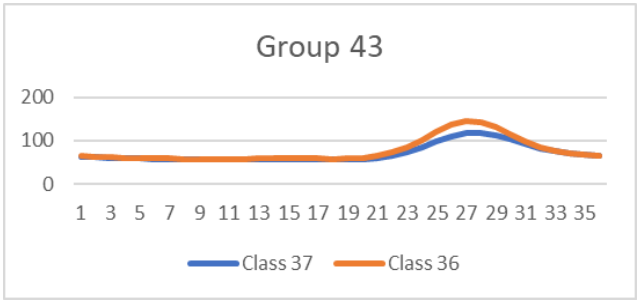

Group 38
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 178
Class 171
Class 62

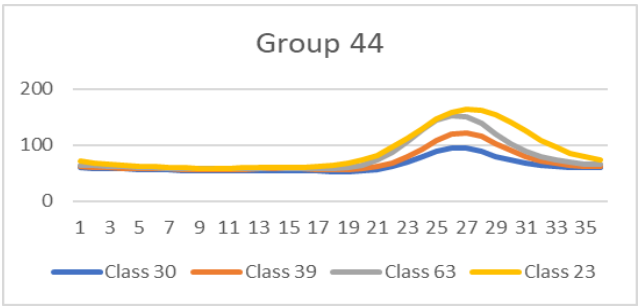

Group 39
400
200
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 158
Class 194
Class 164
Class 181

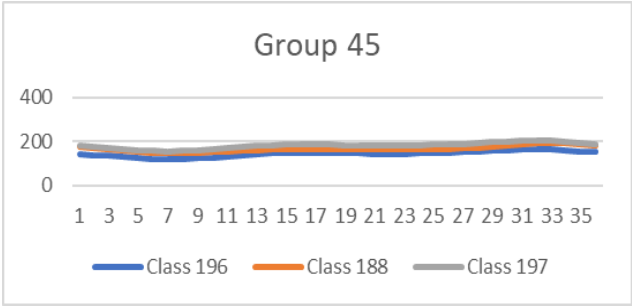

Group 40
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 68
Class 34
Class 95
Class 49

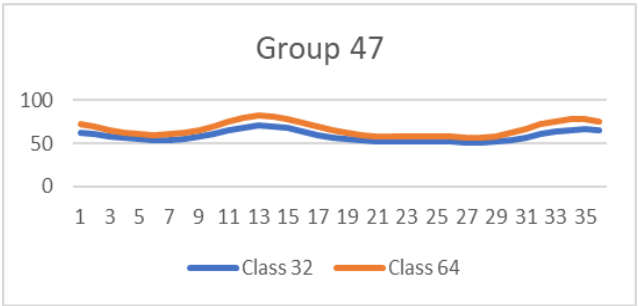

Group 41
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 169
Class 154
Class 180
Class 93
Class 145

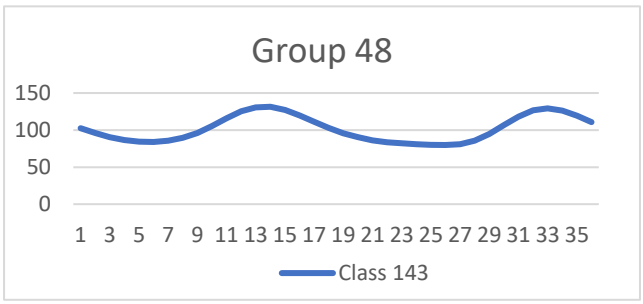

Group 42
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 28
Class 15

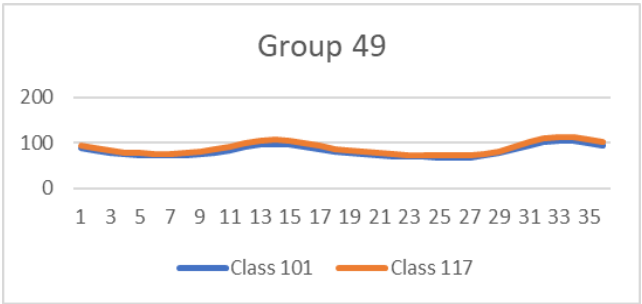

Group 43
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 37
Class 36

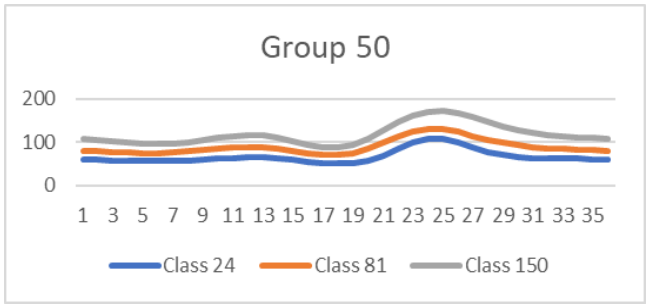

Group 44
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 30
Class 39
Class 63
Class 23

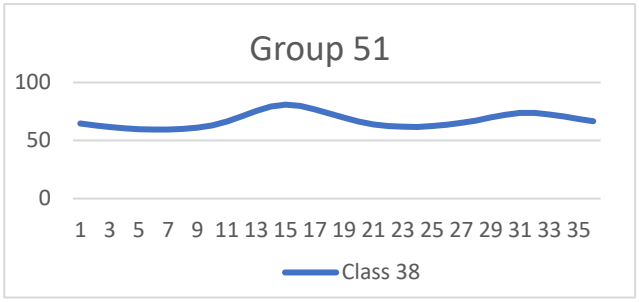

Group 45
400
200
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 196
Class 188
Class 197

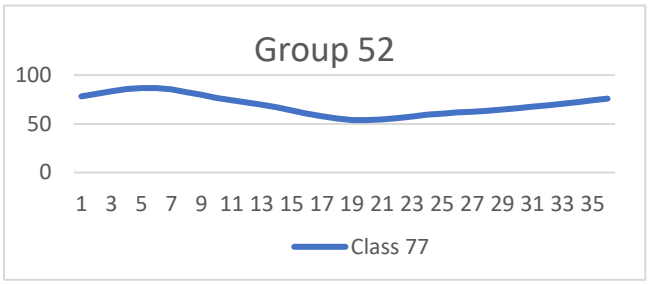

Group 47
100
50
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 32
Class 64

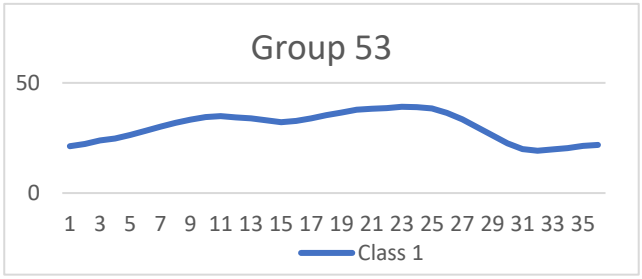

Group 48
150
100
50
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 143

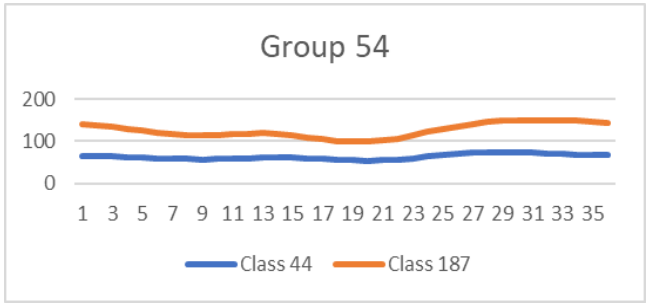

Group 49
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 101
Class 117

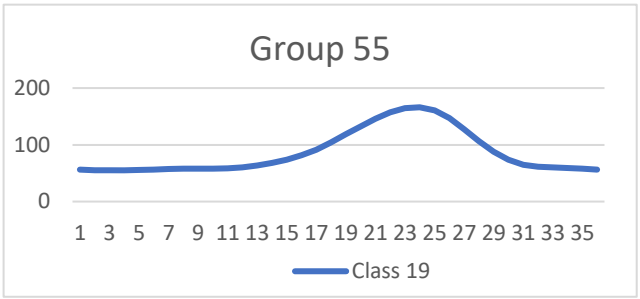

Group 50
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 24
Class 81
Class 150

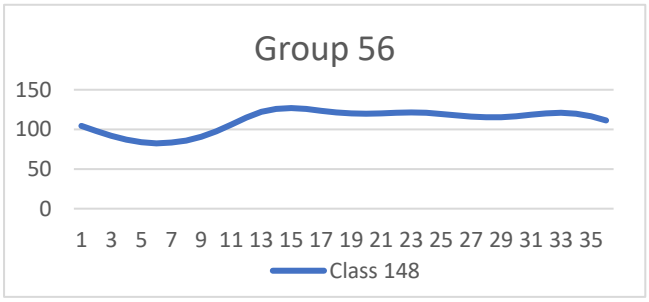

Group 51
100
50
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 38

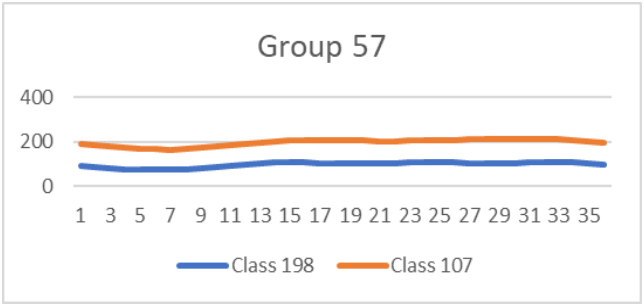

Group 52
100
50
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 77

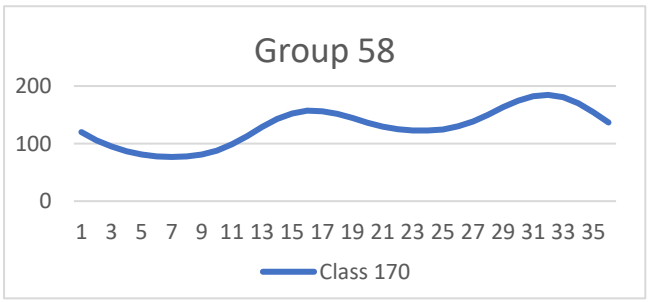

Group 53
50
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 1

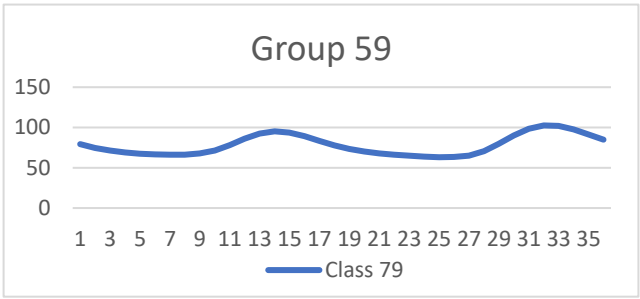

Group 54
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 44
Class 187

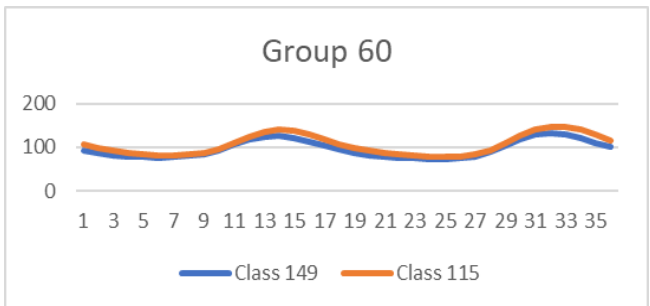

Group 55
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 19

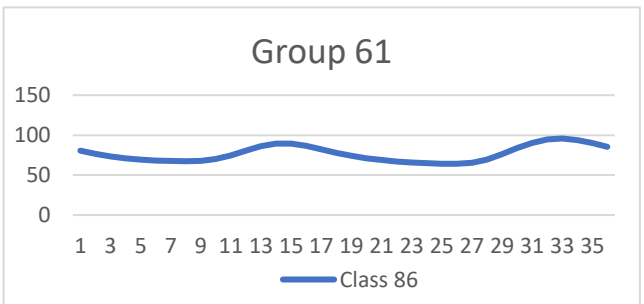

Group 56
150
100
50
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 148

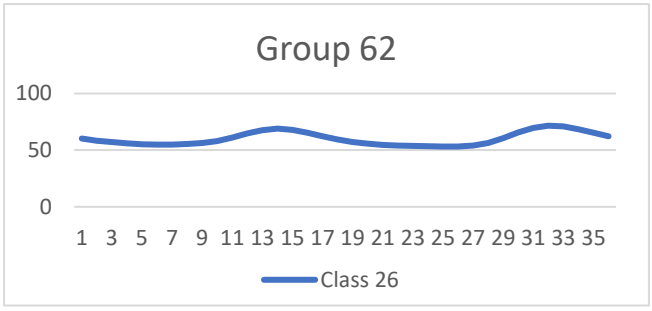

Group 57
400
200
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 198
Class 107

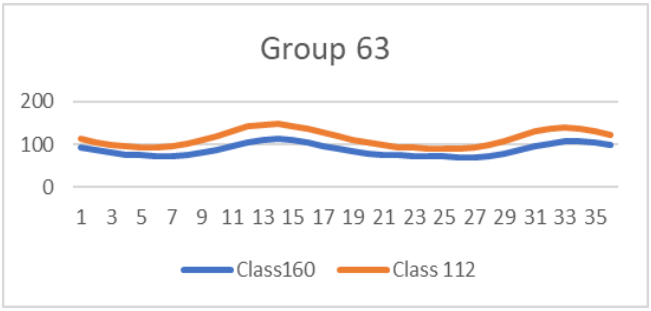

Group 58
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 170

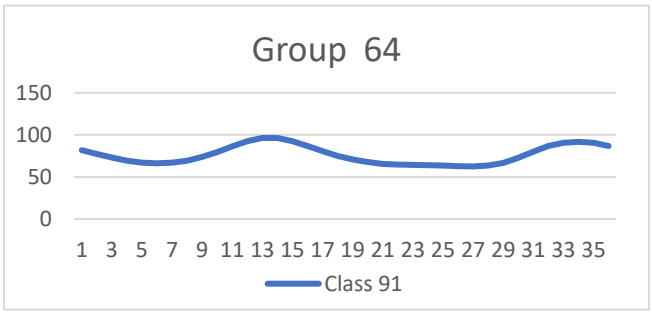

Group 59
150
100
50
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 79

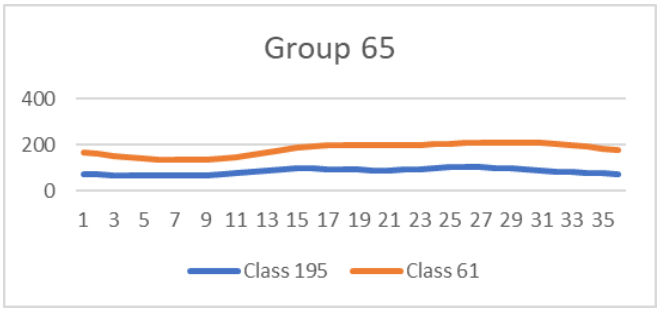

Group 60
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 149
Class 115

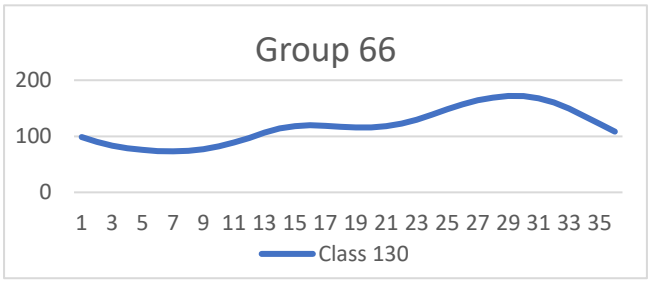

Group 61
150
100
50
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 86


Group 62
100
50
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 26


Group 63
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class160
Class 112


Group 64
150
100
50
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 91


Group 65
400
200
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 195
Class 61


Group 66
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Class 130

Annex 3. Code used to download Sentinel-1 and 2

```
// This code is used to download Sentinel image fieatures
// dry (off) season and wet (on) season image features

// oro_group1 refers to a shape file of 66 Proba-V NDVI groups with an attribute
// of gridcode (the name of the Groups)
var oro_group1 = ee.FeatureCollection('users/ygagragn/oro_dissolve');

// filter out a single group based on the gridcode
var filter = ee.Filter.inList('gridcode',[2]);
// the filltered group assigned a name group_filtered
var group_filtered = oro_group1.filter(filter);

/// VV polarization images
var imgVV = ee.ImageCollection('COPERNICUS/S1_GRD')
          .filter(ee.Filter.listContains('transmitterReceiverPolarisation', 'VV'))
          .filter(ee.Filter.eq('instrumentMode', 'IW'))
          .select('VV')
          .map(function(image) {
            var edge = image.lt(-30.0);
            var maskedImage = image.mask().and(edge.not());
            return image.updateMask(maskedImage);
          });
/// VH polarization images
var imgVH = ee.ImageCollection('COPERNICUS/S1_GRD')
          .filter(ee.Filter.listContains('transmitterReceiverPolarisation', 'VH'))
          .filter(ee.Filter.eq('instrumentMode', 'IW'))
          .select('VH')
          .map(function(image) {
            var edge = image.lt(-30.0);
            var maskedImage = image.mask().and(edge.not());
            return image.updateMask(maskedImage);
          });
// filter out descending sentinel 1 VV images
var desc = imgVV.filter(ee.Filter.eq('orbitProperties_pass', 'DESCENDING'));
// filter out descending sentinel 1 VH images
var desc_vh = imgVH.filter(ee.Filter.eq('orbitProperties_pass', 'DESCENDING'));


////////////// filtering out descending sentinel 1 VV images
/// on_season sentinel images
var onseason2016 = ee.Filter.date('2016-09-20', '2016-11-10');
var onseason2017 = ee.Filter.date('2017-09-20', '2017-11-10');
var onseason2018 = ee.Filter.date('2018-09-20', '2018-11-10');
var onseason2019 = ee.Filter.date('2019-09-20', '2019-11-10');
var onseason2020 = ee.Filter.date('2020-09-20', '2020-11-10');
///five year on season sentinel VV polarization image image collection
var on_season_vv = ee.Image.cat(
          desc.filter(onseason2016).median(),
          desc.filter(onseason2017).median(),
          desc.filter(onseason2018).median(),
          desc.filter(onseason2019).median(),
          desc.filter(onseason2020).median());
/// five year on season sentinel VH polarization image collection
var on_season_vh = ee.Image.cat(
          desc_vh.filter(onseason2016).median(),
          desc_vh.filter(onseason2017).median(),
          desc_vh.filter(onseason2018).median(),
          desc_vh.filter(onseason2019).median(),
          desc_vh.filter(onseason2020).median());
/// of season sentinel image
var ofseason2016 = ee.Filter.date('2016-01-10', '2016-03-01');
```

```
var ofseason2017 = ee.Filter.date('2017-01-10', '2017-03-01');
var ofseason2018 = ee.Filter.date('2018-01-10', '2018-03-01');
var ofseason2019 = ee.Filter.date('2019-01-10', '2019-03-01');
var ofseason2020 = ee.Filter.date('2020-01-10', '2020-03-01');
/// of season sentinel VV polarization image collection
var of_season_vv = ee.Image.cat(
        desc.filter(ofseason2016).median(),
        desc.filter(ofseason2017).median(),
        desc.filter(ofseason2018).median(),
        desc.filter(ofseason2019).median(),
        desc.filter(ofseason2020).median());
/// of season sentinel VH polarization image image collection
var of_season_vh = ee.Image.cat(
        desc_vh.filter(ofseason2016).median(),
        desc_vh.filter(ofseason2017).median(),
        desc_vh.filter(ofseason2018).median(),
        desc_vh.filter(ofseason2019).median(),
        desc_vh.filter(ofseason2020).median());

 // clip the on&of seasonsen tinel 1 VV image with Ethio boundary
var on_season_vv_clip = on_season_vv.clip(group_filtered);
var of_season_vv_clip = of_season_vv.clip(group_filtered);

 // clip the on&of seasonsentinel 1 VH image with Ethio boundary
var on_season_vh_clip = on_season_vh.clip(group_filtered);
var of_season_vh_clip = of_season_vh.clip(group_filtered);

// Claculate median of the median of VV images
var on_season_vv_final = on_season_vv_clip.reduce(ee.Reducer.median());
var of_season_vv_final = of_season_vv_clip.reduce(ee.Reducer.median());
// Claculate median of the median of VH images
var on_season_vh_final = on_season_vh_clip.reduce(ee.Reducer.median());
var of_season_vh_final = of_season_vh_clip.reduce(ee.Reducer.median());
////////////////////////////////////////////////////////////////////////////
//Calculation of the backscater matrices of semtinel 1
///// VV/VH ratio of both seasons
// dry and wet season VV/VH ratio of the
var on_season_vvvh = on_season_vv_final.divide(on_season_vh_final);
var of_season_vvvh = of_season_vv_final.divide(of_season_vh_final);
///// NRPM(Normalized procedure ratio between bands) of both season
var on_season_NRMP = on_season_vv_final.subtract(on_season_vh_final)
                    .divide(on_season_vv_final.add(on_season_vh_final));
var of_season_NRMP = of_season_vv_final.subtract(of_season_vh_final)
                    .divide(of_season_vv_final.add(of_season_vh_final));
///// stacked sentinel 1 variables
var stack_sentinel_data = on_season_vv_final.addBands(of_season_vv_final)
                          .addBands(on_season_vh_final).addBands(of_season_vh_final)
                          .addBands(on_season_vvvh).addBands(of_season_vvvh)
                          .addBands(on_season_NRMP).addBands(of_season_NRMP);
// Displaying the stacked image
Map.addLayer(stack_sentinel_data)
// Export only the selected group (i.e.,grouped_filtered )
Export.image.toDrive({
  image: stack_sentinel_data,
  description: 'group2',
  scale: 20,
  folder: 'sent_1',
  maxPixels: 1e13,
  region: group_filtered
});
```