U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

2.º
CICLO

FCUP
2020

U.PORTO

Remote Sensing and Machine Learning Tools for Vegetation Monitoring

Sofia Perestrelo de Vasconcelos C. Pereira

FC

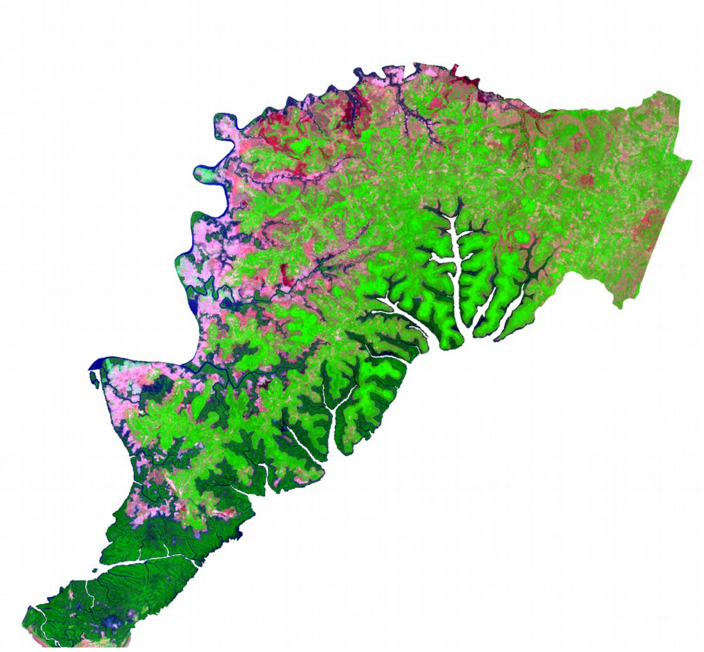
U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Remote Sensing and Machine Learning Tools for Vegetation Monitoring

Sofia Perestrelo de Vasconcelos C. Pereira

Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto em
Data Science

2020



Remote Sensing and Machine Learning Tools for Vegetation Monitoring

Sofia Perestrelo de Vasconcelos C. Pereira
Ciência de Dados (Data Science)
Departamento de Ciência de Computadores
2020

Orientador
João Pedro Pedroso, Professor Auxiliar, Faculdade de Ciências

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, ______/______/_________

Abstract

The forests of Guinea Bissau are under constant threat; they are silently being cut down and replaced by tree crops, mainly cashew. While the exports of cashew nuts greatly contribute to the gross domestic product and to support local livelihoods, the country's natural capital is depleted due to unsustainable land use. Even though measures to stop deforestation were taken, the problem is not being fully addressed and there are no systematic nor automatic means for monitoring the situation. This work presents a contribution for the development of an affordable, reliable and easy to use alternative to field monitoring. It uses remote sensing and machine learning techniques to develop models capable of automatically detecting cashew orchards in satellite images. The results obtained through a case study developed for a protected area in southern Guinea-Bissau indicate that this type of monitoring is possible when classifying satellite images for which the models are trained. However, large amounts of ground truth data and frequent updates might be necessary to build a system fully able to generalize for other years in which the model was not trained.

Keywords— Sustainability, Vegetation, Guinea-Bissau, Remote Sensing, Earth Observation, Land Cover Monitoring, Machine Learning, Supervised Learning, Python.

Resumo

A floresta da Guiné-Bissau encontra-se sob constante ameaça; é discretamente destruida para dar lugar a plantações agrícolas que contribuem fortemente para o produto interno bruto do país. As plantações de cajú são as mais frequentes. Apesar de terem sido implementadas medidas para impedir a desflorestação, estas não se encontram a ser cumpridas na sua totalidade e não existe uma forma sistemática e automatizada de monitorizar esta situação. Esta dissertação propõe uma alternativa barata, confiável e simples à monitorização baseada em trabalho de campo, recorrendo a técnicas de deteção remota e *machine learning* para desenvolver modelos capazes de detetar plantações de cajú em imagens de satélite. Os resultados obtidos neste caso de estudo desenvolvido numa área protegida no sul da Guiné-Bissau indicam que este tipo de monitorização é de facto possível para classificar imagens nas quais o modelo foi treinado. No entanto, grandes quantidades de dados de campo e actualizações frequentes podem ser necessárias para o desenvolvimento de um sistema capaz de generalizar para outros anos que não aqueles em que o modelo tenha sido treinado.

Acknowledgements

I was once told that when it comes to thesis supervisors, having someone kind was as important as having someone brilliant. With that said, I start by thanking my thesis advisor Prof. João Pedro Pedroso for sharing not only his knowledge but most importantly his time, patience and kindness. I would also like to thank my family (my parents, my brother and Francisco) and closest friends (they know who they are) for bearing with me on my worst days. A special thanks to my parents for being not only great parents but also remarkable scientists that taught me what I consider to be my most valuable skill when it comes to science, which is to think critically. I am also grateful for all the support provided by RSeT, mainly by Catarina.

Thank you all.

Dedicated to Charles Darwin, the father of modern biology.

Contents

Abstract **i**

Resumo **iii**

Acknowledgements **v**

Contents **ix**

List of Tables **xii**

List of Figures **xiv**

Listings **xiv**

Acronyms **xv**

1 Introduction **1**

1.1 Cashew Monitoring in Guinea Bissau 1

1.2 Objectives 2

1.3 Outline 2

2 Background **5**

2.1 Sustainable Development Goals 5

2.2 Sustainable Development Goals in Guinea-Bissau 6

2.3 Basic Concepts of Remote Sensing 8

2.4 Earth Observation Missions . . . 9

2.5 Sentinel 2 Multi-Spectral Images . . . 9

2.6 Land Cover and Land Use Maps . . . 12

2.7 Tools and Software . . . 13

3 Materials and Methods 17

4 Development of a cashew orchard detection and monitoring tool 19

4.1 Overview . . . 19

4.2 Data Description, Analysis and Cleaning . . . 21

4.3 Classification Years with Ground Truth Data (2019) . . . 27

4.3.1 Data pre-processing . . . 29

4.3.2 Preliminary Classification . . . 33

4.3.3 Hyperparameter optimization . . . 33

4.3.4 Test Set Classification and Land Cover Map . . . 35

4.3.5 Map post processing . . . 37

4.4 Classification of Years Without Ground Truth Data (2020) . . . 42

4.4.1 Classification extension problem . . . 43

4.4.2 Single Composite Approach . . . 45

4.4.2.1 Data Pre-Processing . . . 46

4.4.2.2 Preliminary Classification . . . 47

4.4.2.3 Hyperparameter optimization . . . 47

4.4.2.4 Test Set Classification . . . 47

4.4.3 Incorporating several years of data into training . . . 48

4.4.3.1 Data Pre-Processing . . . 49

4.4.3.2 Preliminary Classification . . . 49

4.4.3.3 Hyperparameter optimization . . . 49

4.4.3.4 Test Set Classification, Post Processing and Land Cover Maps 50

5 Conclusions and Future Work **53**

5.1 Conclusions 53

5.2 Future Work 56

A Supplementary Figures and Tables **59**

Bibliography **63**

List of Tables

1 Spectral bands of the Sentinel 2 Satellites. 10

2 Schematic representation of the data setup used for the data cleaning step. 24

3 Schematic representation of the vector data. 27

4 Number of points of each class in the data. 27

5 Schematic representation of the data setup (section 4.3). 29

6 Example of one grouped stratified cross-validation fold. 30

7 Selected features (section 4.3.). 32

8 Cross-validation scores of the RF and SVM (section 4.3) 33

9 Confusion Matrix corresponding to Figure 19. 36

10 F1-score corresponding to Figure 19. 36

11 Confusion Matrix corresponding to Figure 22 a. 40

12 F1-score corresponding to Figure 22 a. 40

13 Confusion Matrix corresponding to Figure 22 b. 41

14 F1-score corresponding to Figure 22 b. 41

15 Confusion Matrix corresponding to Figure 24. 44

16 F1-score corresponding to Figure 24. 45

17 Selected features (section 4.4.2). 47

18 Cross-validation scores of the RF and SVM (section 4.4.2) 47

19 Schematic representation of the data setup (section 4.4.3). 48

20 Selected features (section 4.4.3). 49

21 Cross-validation scores of the RF and SVM (section 4.4.3) 49

22 Confusion Matrix corresponding to Figure 27 a. 50

23 F1-score corresponding to Figure 27 a. 50

24 Confusion Matrix corresponding to Figure 27 b. 51

25 F1-score corresponding to Figure 27 b. 51

S1 Confusion Matrix corresponding to section 4.4.2 (2019). 60

S2 F1-score score corresponding to section 4.4.2 (2019). 60

S3 Confusion Matrix corresponding to section 4.4.2 (2020). 61

S4 F1-score score corresponding to section 4.4.2 (2020). 61

List of Figures

1 Sustainable development goals. . . . 6

2 The optical remote sensing process. . . . 8

3 Distribution of pixels of different types of cover in a red and near-infrared space. . . . 11

4 Percentage of atmospheric transmission in the different regions of the spectrum. . . . 12

5 Schematic representation of a multispectral image. . . . 13

6 Schematic representation of a classification pipeline. . . . 16

7 Schematic overview of chapter 4. . . . 20

8 Cantanhez National Park. . . . 21

9 Overview of the study area and the training polygons. . . . 22

10 Examples of cashew polygons in different stages of development. . . . 22

11 RGB image of the study area (full season). . . . 24

12 95% density level ellipses for all land cover classes. . . . 25

13 95% density level ellipses for class "Cashew". . . . 26

14 RGB images of the study area (monthly). . . . 28

15 Reflectance of each S2 band (monthly). . . . 30

16 Number of selected features and corresponding accuracy (section 4.3.). . . . 31

17 Optimization of the Random Forest's hyperparameters (section 4.3). . . . 34

18 Optimization of the SVM's hyperparameters (section 4.3) . . . 34

19 Land cover map 2019 (section 4.3). . . . 35

20 Two areas inside the ROI and their corresponding classification (section 4.3). 37

21 Performance metrics of the range of tested sieve thresholds (section 4.3). 37

22 175p and 300p sieves of the map in Figure 19 (section 4.3). 39

23 Zoom-in of a small area of the ROI and corresponding classification w/wo sieves. 42

24 Land cover map 2020 (section 4.4.1) 44

25 Reflectance of each S2 band (full season). 46

26 Number of Selected Features and Corresponding Accuracy (section 4.4.2). 46

27 125p sieved land cover map (section 4.4.3) 52

S1 Optimization of the SVM's hyperparameters (section 4.4.2). 59

S2 Optimization of the SVM's hyperparameters (section 4.4.3). 62

S3 Performance metrics of the range of tested sieve thresholds (section 4.4.3). 62

Acronyms

ANN	Artificial Neural Network
API	Application Programming Interface
BOA	Bottom of the Atmosphere
CBD	Convention about Biological Diversity
CSV	Comma Separated Value
EO	Earth Observation
ESA	European Space Agency
FAO	Food and Agriculture Organization
GDAL	Geospatial Data Abstraction Library
GDP	Gross Domestic Product
GEE	Google Earth Engine
GIS	Geographical Information System
GLCM	Gray-Level Co-Occurrence Matrix
MLE	Maximum Likelihood Estimator
MSI	Multi-Spectral Instrument
NASA	National Aeronautics and Space Administration
NDVI	Normalized Difference Vegetation Index
NIR	Near-infrared
REDD+	Reducing Emissions from Deforestation and Forest Degradation
RF	Random Forest
RGB	Red-Green-Blue
ROI	Region of Interest
RS	Remote Sensing
S2	Sentinel-2
SDG	Sustainable Development Goal
SVM	Support Vector Machine
SWIR	Short-wave infrared
TOA	Top of the Atmosphere
UN	United Nations
UNFCCC	United Nations Framework Convention on Climate Change
USGS	United States Geological Survey
WRI	World Resources Institute

Chapter 1

Introduction

1.1 Cashew Monitoring in Guinea Bissau

Guinea-Bissau is home to globally significant forest and savanna woodlands in a territory where two marked seasons determine vegetation appearance, a dry season between November and May, and a wet season between June and October. However, these rich and diverse ecosystems are under severe threat: deforestation has been reported as a major ecological sustainability problem in the country, largely due to uncontrolled conversion of woodlands into permanent cashew tree plantations. In such a poor country, the selling of high quality cashew nuts to installed commercial networks that export to processing factories, mainly in India, is a means of expeditiously improving the economic situation of both the rural families and the State. Cashew nuts are the main source of fast cash for the local population and the country's most exported product, representing a very large proportion of the country's Gross Domestic Product (GDP).

It should be noted that the quality of cashew nuts produced in a given stand starts decreasing after 25 years, while the hydrological equilibrium and productivity of the land become seriously compromised. Thus, the rampant uncontrolled plantation of cashew, which has been converting the country into a large tree orchard with patches of unknown extent, age, or state, threatens food security in the short-term; decreases land availability and suitability for agriculture in the medium term; and drains natural resources and biodiversity in the not-so-long term.

Despite its extreme poverty, Guinea-Bissau has invested quite a lot of effort in attempting to conserve its biodiversity and its forests, and the country is part of the United Nations (UN) climate conventions and of the convention to conserve biodiversity [1]. Nevertheless, the country's low levels of education combined with its political instability makes these policies very hard to implement and thus Guinea-Bissau has not been able to adequately control or halt deforestation and move towards sustainability.

Official information regarding cashew production in Guinea-Bissau is made available by the Food and

Agriculture Organization (FAO) [2], but since it is based on registered transactions, like exported tons of cashew, it is unreliable. In fact, due to the prevailing unregistered selling of the product, and to the lack of objective assessment of the areas occupied by the crop (which remains uncertain) the numbers provided by official agencies are likely to be seriously underestimated. Given these circumstances, land cover monitoring based on satellite remote sensing technology can become an essential aid supporting a better assessment of Guinea-Bissau's cashew plantations and production.

Sentinel, the European Space Agency (ESA) most recent Earth Observation (EO) mission provides multi-spectral images that can be used to produce land cover maps if appropriate training data are made available. This way, an approach consisting of a machine-learning algorithm fed with information derived from both EO data and sufficient ground truth geographical information can be developed and used to produce land cover maps. Such an approach is more practical than extensive field data collection; considerably less costly; faster; and complete, also providing wall-to-wall frequent coverage, while holding great potential for automation [3]. However, it poses some difficulties: forests and cashew plantations are spectrally similar and can exhibit a very similar behavior when using the currently available free optical satellite imagery, such as that from the *Sentinel* mission. Thus, advances towards automated tools, capable of producing sufficiently accurate multi-temporal land cover maps depicting cashew plantations, forests and woodlands, can assist better land use decision making and contribute to improve sustainability in Guinea-Bissau.

1.2 Objectives

The overall objective of this work is to develop an EO and machine-learning-based tool capable of producing land cover maps that accurately spot cashew orchards in Guinea-Bissau, in order to obtain a more realistic representation of the cashew areas in the country. The specific objectives are:

1. To develop a system that produces an accurate land cover map of the only year for which ground truth data is currently available (2019);
2. To explore the possibility of expanding this system so it can produce valid land cover maps for years without ground truth data.

1.3 Outline

This thesis is organized in the following manner:

- **Chapter 2** contains the background information relevant for understanding the following chapters;
- **Chapter 3** describes the methods used throughout this dissertation;

- **Chapter 4** details the developed work. It is comprised of three main sections: initially, the data is described and analyzed. Then, the process for obtaining land cover maps for the year with ground truth data (the year in which the model was trained, which is 2019) is described. Finally, the last section describes the process for obtaining land cover maps for years other than that with available ground truth data (for years in which the model was not trained). This is challenging as there is variability between images of different years, probably due to distinct weather, surface and radiometric conditions;
- **Chapter 5** contains the conclusions and some thoughts about the future work that follows this thesis.

Chapter 2

Background

Here, the main concepts underlying this dissertation are presented. Since the dissertation is very multidisciplinary, so is this chapter, which is intended to provide basic knowledge of the several domains approached in the study. First, some background on the many scopes of sustainability is presented, focusing on land use and ecosystem sustainability associated with cashew production in Guinea-Bissau. Then, information regarding EO such as space missions, remote sensing, satellite imagery, and land cover maps is addressed. Finally, an overview of the tools, software and algorithms involved in this work is provided.

2.1 Sustainable Development Goals

In 2015, the UN set 17 Sustainable Development Goals (SDGs) to be adopted by all member states and achieved by 2030. Many years of hard work of both the member states and the UN, agreements, frameworks and agendas culminated in the SDGs as we know them today. These goals cover the many aspects of sustainable development, ranging from topics more centered around economic and social sciences (gender equality, inequality, or economic growth) to others more related to public health and environmental sciences (clean water and sanitation, and zero hunger) [4] (Figure 1).

This dissertation directly approaches the **SDG No. 15– Life on Land** (sustainable management of forests and land in general, the halting of deforestation, desertification and the protection of the World's biodiversity), **SDG No. 2– Zero Hunger** (better management of the world's food resources) and **SDG No. 13– Climate Action** (arresting, or at least slowing down, climate change). In a more indirect manner, this dissertation also addresses **SDG No. 12– Responsible Consumption and Production** (promoting resource and energy efficiency) [5–8].

SUSTAINABLE DEVELOPMENT GOALS
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS

Figure 1: The seventeen Sustainable Development Goals set by the United Nations.

2.2 Sustainable Development Goals in Guinea-Bissau

Guinea Bissau is the country in which this dissertation focuses. The following section details the situation of the country regarding each of the main SDGs addressed by this work.

SDG 15- Life on Land

Guinea-Bissau is a country that despite its small area of about 36000 Km^2 is home to a fair amount of natural resources. Even though savanna woodlands are the most common type of land cover in the country, a significant percentage of its territory is covered by tropical forest [9]. Studies form the World Resources Institute (WRI) reveal that Guinea Bissau is one of the countries in the world with a higher deforestation rate [10]. At the same time, between 1975 and 2013 the country's agricultural area doubled [9], pressured by a very high population growth [11]. This is a strong indicator that the country's forests and woodlands are being taken down to generate fertile fields for agriculture.

The observed trend leads to numerous problems, starting by the destruction of valuable habitats. Every country taking part of the UN climate agreements is bound to implement policies and strategies to ensure the goals defined through the convention. Guinea-Bissau's forest is extremely rich in both plant and animal biodiversity. It is home not only to some endemic species but also to others that are threatened and/or rare. Examples of such species are *Ammannia santoi* and *Pterocarpus erinaceus* (pau-de-sangue), respectively [12]. Given the worrisome scenario of deforestation in the country, in 2015 the government implemented awareness campaigns and took legal measures to stop deforestation; however, these were not very effective [10]. In this least developed country, where poverty and political instability are the norm, food insecurity is a constant, and the type of agriculture performed may damage the soil, there is also a high risk of disturbing ecosystems and of loosing production capacity [13].

SDG 2 Zero Hunger

Initially, the relationship between the people of Guinea-Bissau and agriculture was healthy: they adopted shifting cultivation techniques, where a given area is used for about 2 to 3 years and then, when the soil starts to get depleted from nutrients, that field is abandoned and the crops are moved to another area. During the fallow periods following cultivation (periods without any cultivation) the soil regenerates, secondary forest eventually starts to develop and after 10 to 20 years the field is productive again. This type of agriculture is performed in a small scale and is generally used for subsistence.

Eventually, this type of agriculture was replaced by what is called a cash crop mindset: instead of subsistence cropping, the goal became to profit out of the crops, mostly though exportation of the resulting products. This could be a solution to combat poverty if managed for sustainability. However, fueled by despair, this type of agriculture became uncontrolled and cashew is now the most abundant cash crop in the country, corresponding to 90% of the country's exports. It is estimated that 85% of the country's population depends on cashew to survive and, like with any other mono-culture, the native vegetation is widely replaced by cashew crops. Even though the consequences these plantations have on biodiversity are not fully understood, it is a known fact that they are more susceptible to pests and diseases, and also more vulnerable to the extreme climatic episodes that come along with climate change and soil depletion [13, 14].

In the past, crop diversity in the country was much higher, with other crops such as rice, millet or maize being planted at a much larger scale. Crop diversity is crucial for food security: single crop economies are not only very vulnerable in case of extreme weather events or pests but they are also very susceptible to market shifts. Both soil depletion and single crop economies are bad for long-term sustainability and food management [13].

SDG 3 Climate Action

Guinea-Bissau is involved in the world's carbon markets, being part of the United Nations Framework Convention on Climate Change (UNFCCC), namely of the Reducing Emissions from Deforestation and Forest Degradation (REDD+) program [15]. Through this program, developing countries can discuss the future of their forests at a higher level by adopting actions for mitigating the effects and promoting adaptation to climate change, while obtaining relevant financial fluxes for sustainable land use management with improvement of local livelihoods. Nevertheless, REDD+ requires that the cause, magnitude, and location of emissions and removals of greenhouse gases from forests be characterized and periodically quantified through credible estimates at national and sub-national level, with best practices requiring the use of EO based methods [16]. Therefore, as expected, due to major data gaps and to technical and technological complexities, this has not been easy to achieve in tropical Africa [17].

Due to all the reasons described above, land cover monitoring is essential to help control Guinea-Bissau's cashew state of affairs. Some researchers think that, when reconverted early enough, cashew orchards might still be convertible to forest, which means that it is still possible to revert or at least alleviate this situation [18]

2.3 Basic Concepts of Remote Sensing

The Sun is the energy source that makes optical Remote Sensing (RS) possible. Electromagnetic radiation originating from the Sun goes through the atmosphere until it reaches the Earth's surface. However, not all the light that is emitted by the Sun towards the Earth actually reaches the ground: Part of it is absorbed and part is scattered by the atmosphere. Once the remaining radiation hits the surface of the Earth, one of three things can happen: it can either be absorbed, reflected or transmitted. The part of the radiation that is reflected by the Earth's surface will then go back again through the atmosphere, where it can be one more time scattered or absorbed by the atmosphere. The portion of the radiation left will finally reach a sensor on board a satellite.

Although many types of sensors exist, multi spectral scanners are the relevant type of scanner for this dissertation. These scanners record the reflectance values on different wavelengths in the light spectrum, and can vary a lot regarding their spatial resolution (the pixel size of the satellite image, in meters) and spectral resolution (the number and width of the channels in the light spectrum through which the sensor registers reflectance values). The radiation in each region of the electromagnetic spectrum has different percentages of atmospheric transmission, meaning that the amount of radiation that reaches the satellite might be very different depending on the wavelength. Wavelengths where the percentage of atmospheric transmission is very low are usually not useful for remote sensing of the Earth's surface [19] (Figure 2).

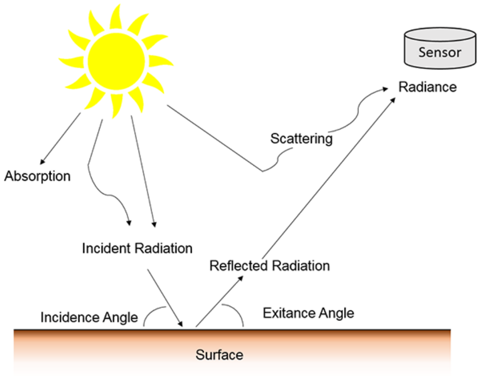


Figure 2: The optical remote sensing process. Taken from http://gsp.humboldt.edu/OLM/Courses/GSP$_2$16$_Online/lesson4-1/radiometric.html$.

Different types of land cover have different reflective properties, and thus different regions of the spectrum can be directly related to inherent properties of different types of land cover, as well as different states of a given land cover. The spectral signature observed in an image may, therefore, be used to extract biophysical information about land cover.

In the case of vegetation, analyzing the spectral signature can help distinguish between different types of vegetation or even (but more difficultly) between different states of a given type of vegetation. Like everything else, plants interact with electromagnetic energy. As photosynthetic activity takes place on the green parts of vegetation, these areas absorb light that will interact with the photosynthetic pigments, water and air spaces inside the cells of the leaves.

2.4 Earth Observation Missions

When compared to field data collection, RS is a more efficient way to monitor land cover of a given area: it allows a cost-effective, fast and frequent monitoring of extensive areas, including those that may be hard, or unsafe, to reach on the ground. Additionally, it has potential for automation. However, it may also pose some limitations, such as not only the spatial and spectral resolutions but also the temporal resolution (the amount of time needed to revisit a certain location and acquire new data).

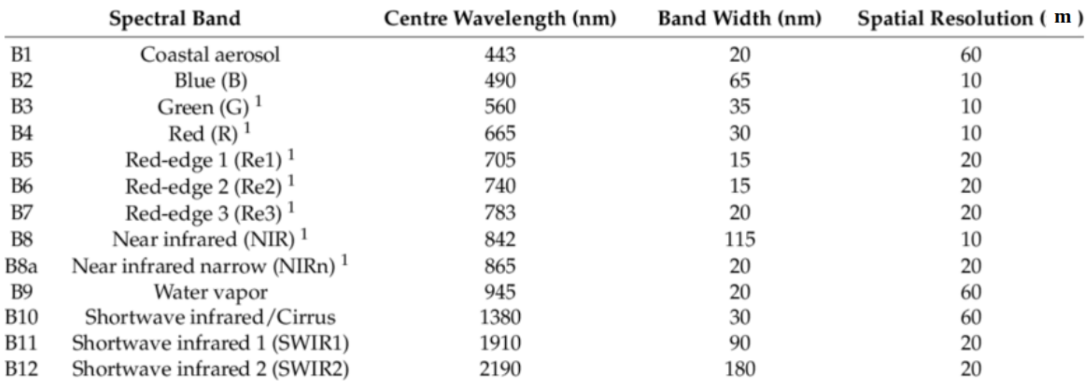
For land cover monitoring resorting to RS techniques, satellite images are used. These satellite images originate from EO space missions. The *LandSat* program, a joint effort of the National Aeronautics and Space Administration (NASA) and the United States Geological Survey (USGS) was one of the first ever EO programs carried out in the world. The first satellite of this mission was launched in 1972, and since then others have been launched. The eighth and latest satellite of the mission was launched in 2013 [20]. More recently, in 2015, the ESA initiated the *Sentinel Mission*, with the goal of replacing the previous EO missions, which were by then outdated or reaching their end, without breaking the data stream.

The new *Sentinel* program is comprised of several missions, each one focusing on a different aspect of EO (ocean, vegetation, air quality,...). Sentinel-2 (S2) is the mission relevant for this dissertation: it is a multi-spectral imaging mission, consisting of two similar polar-orbiting satellites (*Sentinel 2A* and *Sentinel 2B*), placed in orbits with an 180° lag. This setup aims at minimizing the revisiting time, which is approximately 5 days. The images acquired are the starting point for the development of products like land cover and land change detection maps, and the monitoring of geophysical variables, which are key elements for land monitoring [21].

2.5 Sentinel 2 Multi-Spectral Images

Each of the S2 satellites carries a Multi-Spectral Instrument (MSI) with 13 channels ranging from the blue region of the spectrum to the short wave infrared (Table 1). The bands are located at different regions of the spectrum, having different spatial resolutions and band widths [21]. Given that several S2 bands have a resolution of 10 m, the 20m and 60m bands can be resampled to 10 m by replicating the original reflectance values to maximize the spatial resolution.

Table 1: Spectral bands of the Sentinel 2 Satellites.



	Spectral Band	Centre Wavelength (nm)	Band Width (nm)	Spatial Resolution (m)
B1	Coastal aerosol	443	20	60
B2	Blue (B)	490	65	10
B3	Green (G) [1]	560	35	10
B4	Red (R) [1]	665	30	10
B5	Red-edge 1 (Re1) [1]	705	15	20
B6	Red-edge 2 (Re2) [1]	740	15	20
B7	Red-edge 3 (Re3) [1]	783	20	20
B8	Near infrared (NIR) [1]	842	115	10
B8a	Near infrared narrow (NIRn) [1]	865	20	20
B9	Water vapor	945	20	60
B10	Shortwave infrared/Cirrus	1380	30	60
B11	Shortwave infrared 1 (SWIR1)	1910	90	20
B12	Shortwave infrared 2 (SWIR2)	2190	180	20

S2 images contain Top of the Atmosphere (TOA) data. These data are not corrected for atmospheric absorption and scattering of solar radiation, neither in the downward trajectory from the Sun to the surface of the Earth, nor on the upward trajectory from the surface to the MSI installed on board the S2 satellites. This implies that variations in the concentration of certain atmospheric gases, aerosols, and particulates might induce slight variations in the spectral signal detected at the sensor. Since December 2018, Bottom of the Atmosphere (BOA) images derived from the corresponding TOA products started being distributed, based on atmospheric correction procedures that minimizes the effects described above. However, since this study implies images from early 2018, non-corrected imagery was used through the years of study for consistency.

Due to the inherent characteristics of vegetation, a healthy plant (more specifically, its photosynthetic pigments) will absorb light in the visible region of the spectrum (0.4-0.7 μm). In the Near-infrared (NIR) region of the spectrum (0.8-1.2 μm), plants reflect a high amount of light due to multiple scattering processes occurring inside the leaves when the radiation goes from water-filled spaces to air-filled spaces and vice versa. In the Short-wave infrared (SWIR) region (1.2-2.5μm), most of the radiation is absorbed by the water inside the plant tissues. The prominent difference in reflectance between the red and the NIR is designated by red edge and is a very unique feature of green, healthy vegetation, as a healthy plant absorbs red light needed for photosynthesis but it does not absorb light in the NIR region. If the vegetation is under some type of stress, the red edge becomes significantly less steep or even absent.

For image classification purposes, it is useful to explore the relationships between the spectral features of the land cover classes using 2D plots. The most widely used bi-spectral space is the Red-NIR, precisely because it explores the unique positioning of green vegetation (the red edge) and emphasizes its contrast with other land cover types. In the case of Sentinel 2, the Red-NIR space corresponds to bands 04 and

08, respectively) [19]. Healthy vegetation has a very steep red edge, and therefore low reflectance on the red region and high reflectance in the NIR. Drier vegetation will have higher reflectance in the red and lower in the NIR when compared to healthy vegetation (Figure 3). Figure 3 also shows where bare soil (either dry or wet) stands when it comes to the relationship between these two bands. Similarly to this analysis on the Red and NIR space and for the same reasons, it is common to perform similar analysis in the NIR and SWIR space.

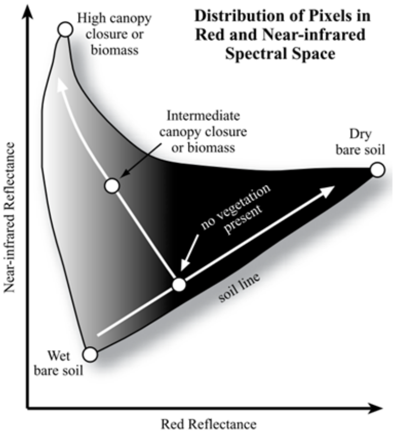

Distribution of Pixels in Red and Near-infrared Spectral Space
High canopy closure or biomass
Intermediate canopy closure or biomass
Dry bare soil
no vegetation present
soil line
Wet bare soil
Near-infrared Reflectance
Red Reflectance

Figure 3: Distribution of pixels of different types of cover in a red and near-infrared space. Adapted from 1. Jensen JR. Remote sensing of the environment: an earth resource perspective second edition. Vol. 1, Pearson Education Limited,Harlow, England. 2014. pages 333-378.

It is also important to understand that atmospheric transmissivity varies across the electromagnetic spectrum (Figure 4). Bands 01, 09 and 10 are designed to quantify atmospheric effects (such as the diffusion and absorption of light by atmospheric gases, aerosols and particles). The reflectance values observed in these bands can be used to adapt the reflectance values of the bands that are actually designed to provide information about the Earth's surface, so that they account for atmospheric effects. Effects of disturbances such as smoke (originating from fires or pollution) or marine aerosols can be detected this way and therefore accounted for. This is why bands No. 01, 09 and 10 are usually discarded in EO pipelines, since the amount of solar radiation reaching the surface (i.e. the signal) in these regions of the spectrum is small [19].

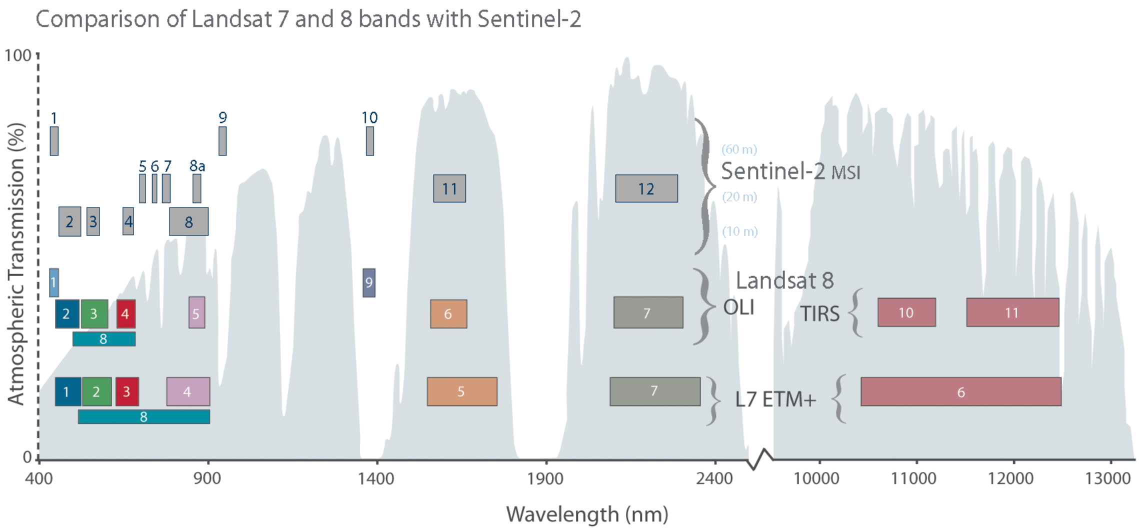


Figure 4: Percentage of atmospheric transmission in the different regions of the spectrum. The bands of several satellites, including S2 are depicted. Taken from 1. User Guides - Sentinel-2 MSI - Overview - Sentinel Online [Internet]. Available from: https://sentinel.esa.int/web/sentinel/user-guides/sentinel-2-msi/overview.

2.6 Land Cover and Land Use Maps

One of the most common uses of S2 images is the generation of land cover maps. These maps display the bio-physical cover of the earth's surface, representing the physical coverage of different types of land type classes in a given region. Several classes are usually depicted, such as water, urban areas, forest, and others. This type of information is very important in for a wide variety of uses, whether in a more scientific scope (for example, environmental, geological or social studies) or in more applied circumstances, such as disaster management (floods or droughts) or emergence response (fires) [**?**].

Both the USGS and ESA produce global land cover products; however, due to their global coverage, these are produced at a low resolution. If there is a strong research interest in a particular region of the globe, S2 can be used to produce land cover or land change maps at a much higher resolution (10m). For that, training areas are delimited (geo-referenced) either using field work or higher resolution imagery, and then classification is performed using satellite images. The pixels of these delimited areas are labeled and used as ground truth training data for the classification task. A dynamic approach can also be considered, where instead of static land cover, land cover change is assessed.

Cashew land cover monitoring is not available in the worldwide benchmark products, as it is both a very specific class and requires a resolution higher than that available in the global products. Therefore, mapping the area covered by cashew orchards in Guinea-Bissau, would result in a more accurate estimate of the actual area, not only for monitoring purposes but also to ensure that the environmental policies the country agreed upon are being fulfilled (for example, that supposedly protected forest areas are not being deforested and replaced with cashew plantations).

It is very important to understand the difference between land cover and land use: while land cover classes relate to the physical land type of a given area (such as forest, water, bare soil, etc.) and are therefore directly related to the spectral properties, land use documents what people do with the land and, therefore, may not be directly related to spectral properties. For example, some small regions inside a forest can be deforested. When it comes to land cover (i.e., in spectral terms) these regions cannot be classified as forest. However, they are still part of the forest when it comes to the use of those areas. Therefore, while land cover is respective to spectral classes, land use relates to informational classes [19].

2.7 Tools and Software

Google Earth Engine and Google Earth Pro

S2 multi-spectral images can be obtained through Google Earth Engine (GEE), which is a cloud-based platform provided by Google [22]. This platform combines open-source data catalogs of satellite imagery with computing power optimized for parallel processing of geospatial data, which is usually very heavy. Besides having Application Programming Interfaces (APIs) for both Python and JavaScript, an online Integrated Development Environment using JavaScript, called GEE Code, allows the user to perform spatial analysis, visualization tasks and download the necessary imagery. The images are usually stored as Geo-Referenced Tagged Image File Format (GeoTIFF, or GeoTIF for short) files, which is a very popular data format for raster data. One can think of a GeoTIFF file as a stack of images, one per each satellite band (Figure 5).

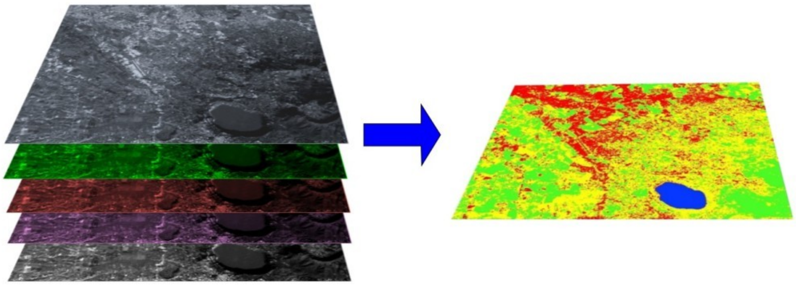


Figure 5: Schematic Representation of a multispectral image. Taken from: https://semiautomaticclassificationmanual.readthedocs.io/pt/latest/remote_sensing.html.

Besides being useful for handling single S2 images, GEE can also perform very useful tasks regarding the combination of multiple S2 images. First, as S2 images come in 100 × 100 Km^2 tiles, it is very common for a Region of Interest (ROI) to be comprised of more than one S2 image tile. When that is the case, it is necessary to assemble the tiles in order to produce a spatially continuous image, which is called mosaicking.

On the other hand, over a given period of time many images regarding one ROI can be available. Instead of simply choosing one, these overlapping images can be combined into a single image based on an aggregation function. This is called compositing. This can offer many advantages, and the aggregation function should be chosen according to the goal of the compositing process. For example, if the goal is to maximize the vegetation signal, one can use the maximum value of the Normalized Difference Vegetation Index (NDVI), an indicator of live green vegetation. This would mean that for a given pixel inside the ROI, the pixel with the highest NDVI among the overlapping images will be used, resulting in one artificial image that contains pixels originating from a stack of images. Other compositing criteria, such as picking the median value of each band per pixel (instead of a maximum or minimum) can be used and may provided other advantages; in the case of the median, the variability between composites of different years is minimized when using this criteria.

Although GEE is optimized for handling raster data, it can also be useful to process vector data. This type of data consists of georeferenced points, lines, or polygons that are used to delimit regions of the globe (for example, a country's boundaries or a highway). In this particular context, the vector data will correspond to delimited regions (polygons) belonging to each one of the classes of the land cover problem. GEE can be used for example to generate training data by intersecting a raster with a vector layer; this way, the reflectance values for the pixels belonging to each polygon are obtained, creating a data set in which each pixel is labeled and variables that correspond the reflectance values of each band are its attributes. If a detailed analysis or edition of the vector data is necessary, the resolution provided by S2 images might not be sufficient. For this, Google Earth Pro provides high resolution satellite imagery (resolution varies with date and location, as Google purchases the images from commercial providers) that allows a more detailed overview of the vector data. Unlike S2 products, this type of image can be very troublesome and expensive to get and is not available on a continuous and consistent basis, rather being available for a few specific dates only (that might vary according on the region of the globe).

GDAL, Rasterio and Scikit-Learn

GEE's computing power is very handy for handling the very large S2 data catalogs and for retrieving the necessary multi-spectral images. However, its computational power has several limitations as it is meant to serve many users around the world. Since it was clearly not designed in a data science perspective, basic elements necessary for an appropriate analysis are either not implemented (although community versions are sometimes available), nor practical or sometimes even possible. Some examples are certain types of exploratory data analysis or cross-validation. Although it does support the use of classification algorithms, it does not provide much freedom when it comes to their parameters, with hyperparameter tuning not being implemented. Therefore, it is common to use GEE just to get the necessary images and then migrate to a local machine or a cloud service (such as Google Cloud Platform) where other more appropriate tools can be used.

Both Python and R have libraries designed for the manipulation of raster data. The Geospatial Data Abstraction Library (GDAL) is the most famous code library for reading, processing and writing many types of raster data formats. It is a free software licensed by the Open Source Geospatial Foundation, written in C/C++ [23]. Although an API for Python is available, it provides little abstraction from the C API. More recently, in 2016, `Rasterio` was developed: it is a `GDAL` and `Numpy`-based Python library with the goal of providing all the necessary tools for manipulation of geospatial data using modern Python language features and performing as fast as `GDAL`'s Python bindings [24]. For the data manipulation and classification tasks, `scikit-learn` is the most widely spread Python library. Not only it provides support for numerous learning algorithms (both supervised and unsupervised) but also for several pre-processing steps, among other data handling capabilities [25] .

QGIS

GEE is useful for obtaining the necessary multi-spectral images and rasterio for importing and processing raster data in Python. However, one of the most common tools for handling geospatial data is QGIS, a free open-source Geographical Information System (GISs) software [26]. It supports visualization, analysis and editing of both raster and vector layers in many possible data formats. Although QGIS allows the realization of many raster and vector operations (most of them through GDAL), in this dissertation it was mainly used for visualization purposes, because using Python directly provides more freedom and integrates better with other performed operations.

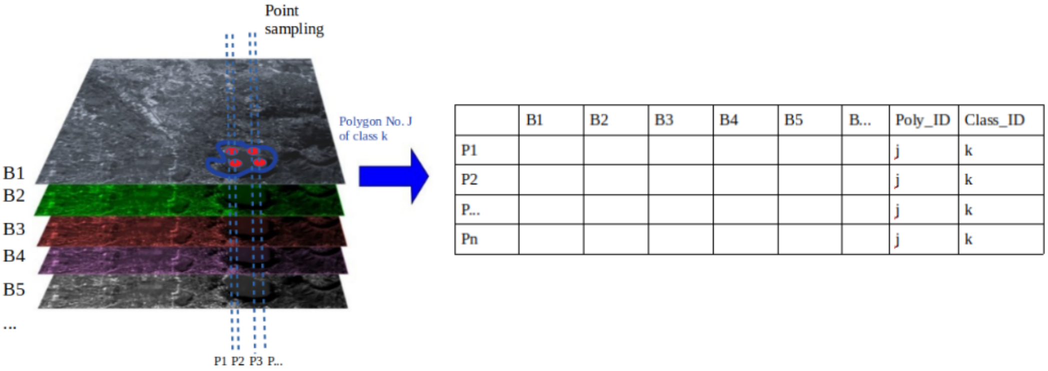
Classification Pipeline and Algorithms

As stated before, the production of land cover maps involves a classification task. These maps are made by classifying individual pixels based on their spectral properties observed through satellite imagery. Although unsupervised approaches can be used, they usually result in large errors. Therefore, supervised approaches are the most common and for these approaches, training data are needed. Training data consist of labeled pixels by class of interest in order for a supervised classification algorithm to learn. A common pipeline for obtaining these labeled pixels goes as following (Figure 6):

1. Regions of the study area corresponding to each class are delimited and labeled. This corresponds to a vector layer where each object is a delimited area (a polygon) with a label. The polygons usually undergo a rasterization step, meaning that their shape is adapted from a somewhat irregular shape to a regular shape that fits the pixel resolution of the satellite in use;
2. An intersection operation between the multi-spectral image and the vector layer is performed, in order to extract the reflectance values for all the bands of the image in each of the pixels inside the training polygons;
3. A data set containing the reflectance values for each band plus a label per pixel is now available

and fed to the classification algorithm to train;

4. A classifier is trained on the training data and is then ready to classify the rest of the pixels/vectors.



	B1	B2	B3	B4	B5	B...	Poly_ID	Class_ID
P1							j	k
P2							j	k
P...							j	k
Pn							j	k

Figure 6: Schematic Representation of a classification pipeline. Adapted from: https://semiautomaticclassificationmanual.readthedocs.io/pt/latest/remote_sensing.html.

Supervised parametric approaches (like the Maximum Likelihood Estimator (MLE)) or non-supervised approaches (like clustering) were the options available for land cover classification task in the past and are still widely used today. However, the rise of Machine Learning opened many doors regarding the production and accuracy of classification maps. Since then, many papers bench-marking algorithm performance using S2 images emerged, stating that the most commonly used algorithms nowadays are Random Forests (RFs), Support Vector Machines (SVMs) and Artificial Neural Network (ANN), usually rendering a very similar performance [27, 28]. Other algorithms (sometimes specifically designed for this purpose) are constantly being explored and developed. In order to make the classification more accurate, multi-temporal approaches can be used to take into account the dynamics of the land, which might be very useful. For example, perennial crops that dry out during the dry season of the year will exhibit a cyclic pattern when it comes to their reflective pattern.

However, training the algorithms using the reflectance values of single pixels is no longer the only option available. More recently, innovative approaches for land cover mapping emerged: instead of performing the traditional pixel-based classification, patch-based (or object-based) approaches started to become very popular. These approaches tend to be very accurate, as they use more than just single-pixel spectral information to train the classifier; they also use information like shape, homogeneity size, color and relationships with neighboring pixels [29]. With these techniques, the training data are not individually labeled pixels, but labeled patches of the image. This type of approach leans more towards computer vision and tipically requires the use of deep learning algorithms (usually Convolutional ANNs) to be accurate. Very recently, GEE started to allow exporting images in a Tensor Flow-ready format (TFRecord).

Chapter 3

Materials and Methods

GEE was used to obtain S2 TOA images from a period ranging from 2018 to 2020. GEE was also used to generate training tables by intercepting the images with the vector data in a process similar to what is shown in figure 6. The data exploratory analysis and cleaning was done using Rstudio, namely `ggplot` [30] and `plotly` [31], for plotting the graphs. In addition, Google Earth Pro was used to edit the vector data during the cleaning process. After cleaning the data, Python (with `scikit-learn` and `rasterio` being the main libraries) was used for the preprocessing steps, training the algorithm, produce land cover maps and post processing tasks. Finally, QGIS was used for laying out the maps and the Red-Green-Blue (RGB) images displayed throughout the dissertation.

The classifiers used in the classification tasks were the `scikit-learn`'s `RandomForestClassifier` and `SVC` (a RF and a classification SVM, respectively). Random Forests are an ensemble method: a large number of decision trees is generated and the resulting prediction is the majority vote of the trees in the forest (or the average of the individual prediction, in the case of a regression problem). The vote of each tree is usually weighted by their probability estimates. Decision trees, the building blocks of a RF, are non-parametric supervised learning methods that work by partitioning the feature space linearly and in a recursive way into smaller regions that will each belong to a given class. Using RFs corrects the well-known tendency of decision trees to overfit to the training data. RFs have many hyperparameters that can be tuned, and the number of trees in the forest is one of the parameters that can greatly influence the performance of the model [32]. Support Vector Machines are also non-parametric supervised learning methods: they rely on the idea of dividing the data into different classes using hyper-planes. In addition to performing linear classification, SVMs can be used for non-linear classification problems by using a kernel, which maps the non-linear data into a higher dimension feature space in which the classes become separable. Although different kernels exist, the radial basis function is the most commonly used kernel when dealing with land cover classification tasks [28]. In addition to the choice of kernel, SVMs also have hyperparameters that can be tuned. The Cost parameter controls the how much the decision boundaries are allowed to fit the data; a large Cost can lead to overfitting and vice-versa. Gamma is a kernel parameter that defines how far the influence of a single training example reaches; a large Gamma can lead to underfitting and vice-versa [33].

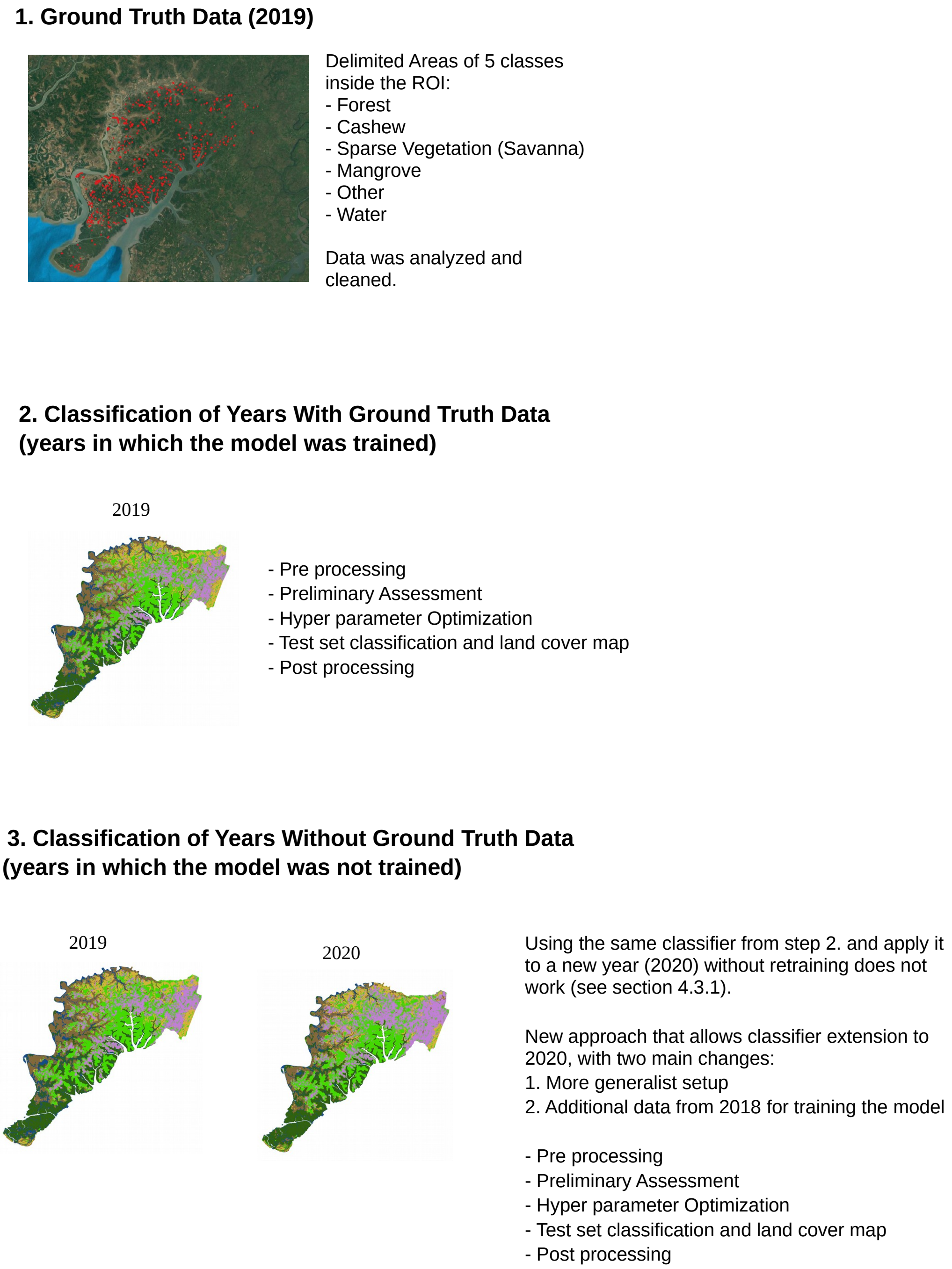
Hyperparameter tuning was done using `hyperopt`, a library providing a Bayesian optimization approach for hyperparameter optimization [34]. Such an approach is much quicker than performing grid search (trying out all possible combinations of parameters within the search space), as it focuses more on the combinations of values that are most likely to result in a good performance.

Chapter 4

Development of a cashew orchard detection and monitoring tool

4.1 Overview

Figure 7, in the next page, contains a schematic overview of the work developed throughout this chapter: First, the data is described, analyzed and cleaned. Then, the process for obtaining land cover maps for the year with ground truth data (the year in which the model was trained, which is 2019) is described. Then, it is observed that the 2019 classifier is not directly transposable for the classification of other years and therefore alternatives are explored.



1. Ground Truth Data (2019)

Delimited Areas of 5 classes inside the ROI:
- Forest
- Cashew
- Sparse Vegetation (Savanna)
- Mangrove
- Other
- Water

Data was analyzed and cleaned.

2. Classification of Years With Ground Truth Data (years in which the model was trained)

2019

- Pre processing
- Preliminary Assessment
- Hyper parameter Optimization
- Test set classification and land cover map
- Post processing

3. Classification of Years Without Ground Truth Data (years in which the model was not trained)

2019

2020

Using the same classifier from step 2. and apply it to a new year (2020) without retraining does not work (see section 4.3.1).

New approach that allows classifier extension to 2020, with two main changes:
1. More generalist setup
2. Additional data from 2018 for training the model

- Pre processing
- Preliminary Assessment
- Hyper parameter Optimization
- Test set classification and land cover map
- Post processing

Figure 7: Schematic overview of chapter 4.

4.2 Data Description, Analysis and Cleaning

Vector Data

Data from the Cantanhez National Park (Figure 8), a protected area covering an extent of more than 1000 Km^2 and located in the southeast region of the country, were provided by the non-governmental organization RSeT [35]. These data consisted of georeferenced polygons labeled with six classes: Forest, Cashew, Sparse Vegetation (including Savanna), Mangroves, Other (including mostly small villages, bare ground and paddy fields) and Water. The provided data was designed for single pixel classification tasks.

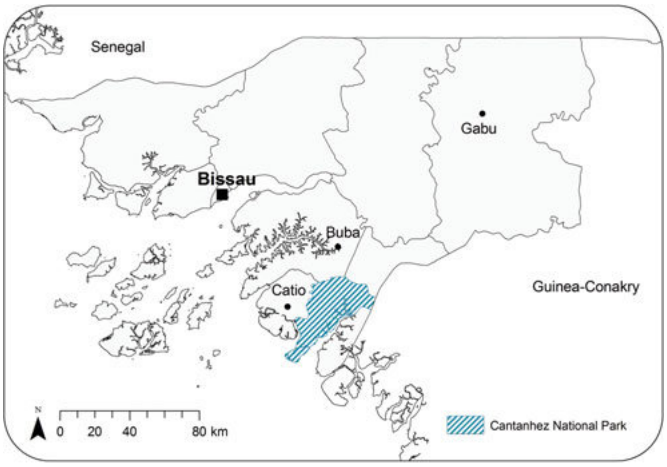


Figure 8: Cantanhez National Park in Guinea-Bissau. Taken from Sousa J, Ainslie A, Hill CM. Sorcery and nature conservation. Environ Conserv. 2018 Mar 1;45(1):905.

A grid was applied to the polygons in order to extract the data points from inside them. Since the polygons had originally been subject to a 25m rasterization (as they were being used in another study), the data points were extracted applying a 25m spaced grid inside each polygon, instead of a 10m spaced grid (which is the resolution of the sentinel imagery in use). This has no serious consequence, resulting only in a smaller number of pixels being included in the data. If the original polygons (prior to the rasterization step) were made available, they could have been rasterized with any resolution. However, it is not a good idea to re-rasterize the polygons because it can severely deform their shape and lead to classification errors, with the labels no longer valid. Each point bears a Class ID (label) and polygon ID as represented in Table 3.

The polygons were drawn using high resolution images from the year 2019 (Figure 9), available at Google Earth Pro. These polygons were drawn and labeled by domain experts. Figure 10 shows the appearance of cashew orchards in high resolution imagery, and also the 25m grid used to extract the data points. Even though cashew orchards are very homogeneous (since it's a mono-culture), there is a significant variability in the aspect of the orchards depending on the size and age of the plants: when the plants are young (on Figure 10 a), the orchards are very sparse having only a few trees. As a consequence, the spectral information in these types of orchards will likely be more close to that of "Sparse Vegetation" than to "Cashew", meaning that although the informational class is "Cashew", the spectral class might be identified as sparse vegetation by the classifier. When the plants grow slightly older and bigger they become much more predominant and exhibit a line pattern (10 b). From this stage onward, there is a convergence between spectral and informational classes, meaning that a good algorithm (provided with appropriate training data) should be able to classify the pixels as "Cashew". At an even later stage, the plants get very big and the line pattern is lost because the tree crowns overlap, which can make the classification task harder again (10 c).

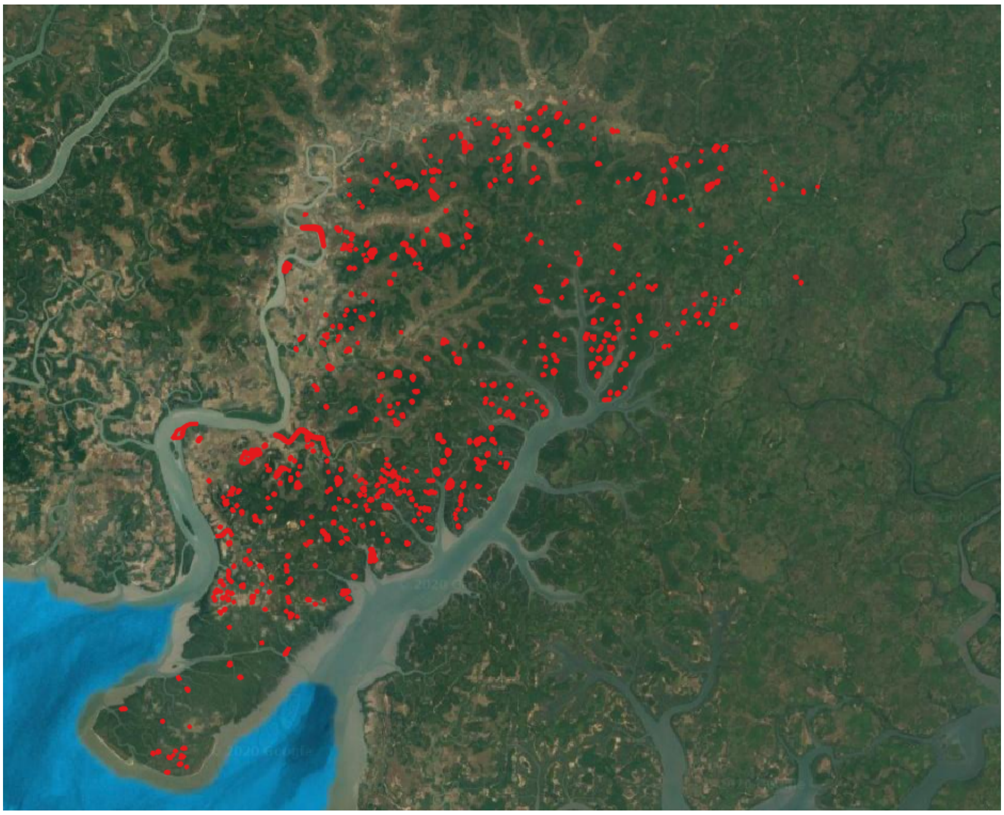
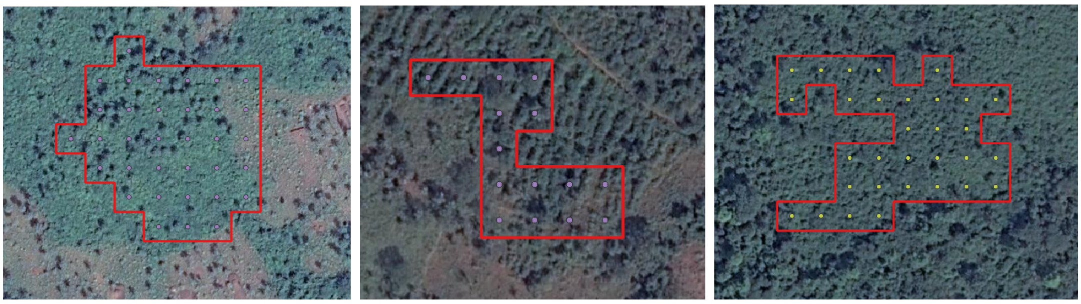


Figure 9: Overview of the study area and the training polygons.



(a)

(b)

(c)

Figure 10: Examples of cashew polygons in different stages of development: early (a), mid (b) and advanced (c). The points inside each polygon correspond to the 25m spaced grid used to extract the data points.

Raster Data

Using the provided vector data as a starting point plus a vector layer of the Cantanhez National Park to delimit the ROI, the necessary raster data can be obtained at any time using GEE. Although S2 has 13 bands in total, band 10 is not made available in GEE due to the reasons described in the previous section (low atmospheric transmission, as it was not designed for remote sensing of the surface of the earth). Therefore, each S2 image generated through GEE will be comprised of 12 stacked layers of pixels, one layer per band, meaning that each pixel will be characterized by a vector of 12 variables. In addition to the original bands of the S2 image, it is very common to derive new bands corresponding to relationships between the pre-existing bands (known as spectral indices) or to other types of information such as textural data [22].

Although forests and cashew plantations can have very similar spectral signatures, cashew is a perennial tropical tree, while the forest is made up of several deciduous trees. This means that during the dry season of the year, part of the forest's trees sheds their leaves resulting in an altered spectral signature (*ie*, pixels will look different from what they do during the wet season). Unlike the forest, cashew trees do not shed. In addition, during the dry season the herbaceous plants underneath the forest and the cashew orchards dry out, which decreases the amount of noise of the signal reaching the satellite. Due to this, it is easier to distinguish between cashew and forest areas in the drier months of the year, which in Guinea-Bissau's case range from late November to mid May [11]. Therefore, the images used in this dissertation were all taken from the period ranging from January to April. Images composites were generated using the median as a compositing criteria.

Exploratory Data Analysis and Cleaning

For preliminary data analysis, a single median composite representing the full 2019 dry season (from January to April) was generated using GEE (see section 2.7). Bands 12 (SWIR), 08 (NIR) and 04 (Red) of this composite were used to produce an artificial RGB image (Figure 11). Using these bands is a very common approach, because due to the reasons described in the background, these bands maximize the information regarding the presence/absence of vegetation in the image. In locations of the ROI with a lot of vegetation, reflectance in the NIR region is high and the appearance in a artificial color RGB is green. On the other hand, in regions where vegetation is not abundant or not present, the reflectance in the NIR region is lower but higher in the Red region, hence the red/pink appearance. This composite was later intersected with the labeled data in GEE to obtain a Comma Separated Value (CSV) file with the pixel values for each labeled point (Table 2).

The resulting CSV file was imported to RStudio and ggplot was used to generate some very useful visualizations. Ellipses corresponding a 95% density level both for all the pixels of each class (Figure 12 a-b) and for the pixels of each polygon (Figure 12 c-d) were plotted in the two most commonly evaluated band spaces (B4/B8) and (B12/B8), resulting in figures comparable with Figure 3 from the background

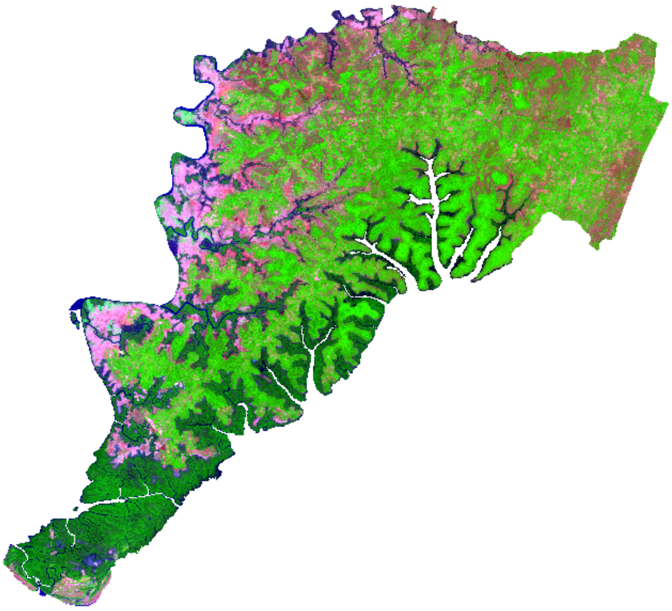
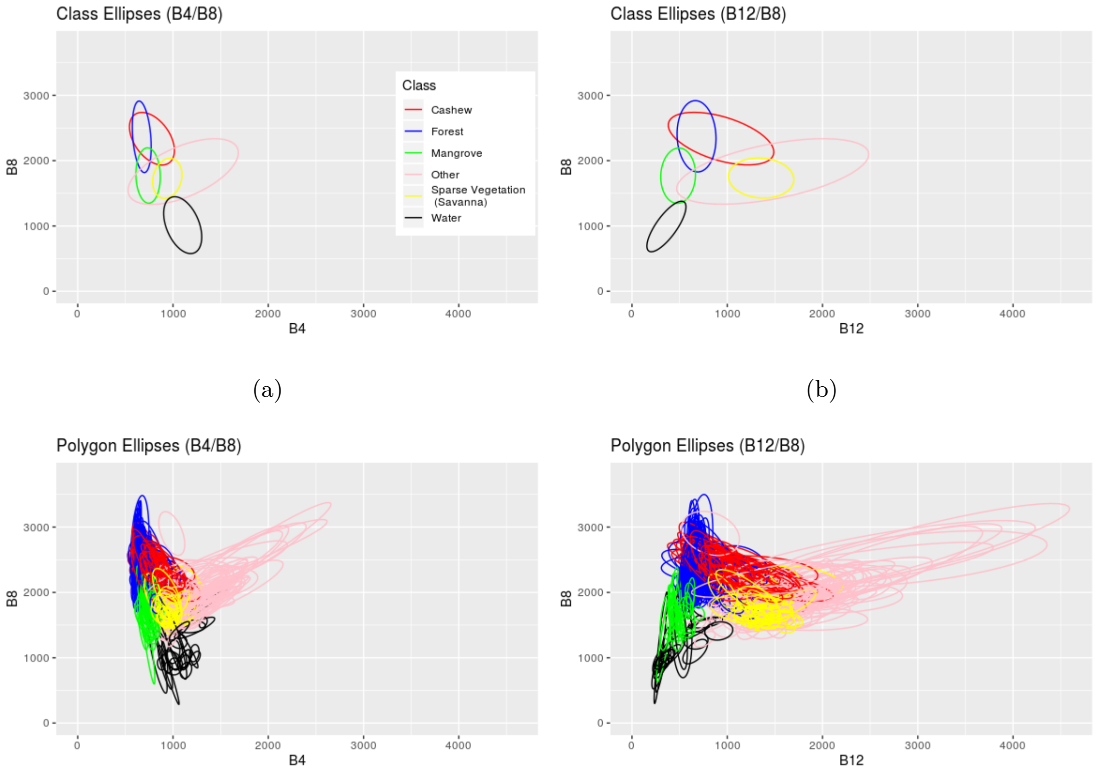


Figure 11: RGB image (artificial color) of the study area (Bands 12 (R), 08(G) and 04(B)) corresponding to a median composite made with images from the full dry season (Jan-Apr).

Table 2: Schematic representation (variables and data points) of the data setup used for the data cleaning step.

	B1	B2	B3	B4	B5	B6	B7	B8	B9	B11	B12	Poly_ID	Class_ID
P1													
P2													
Pn													

chapter. These types of graphs allow the user to have a preliminary overview of the separability of the different classes. However, looking at all the pixels of each class as a whole can be misleading, as it does not provide a sense of the internal variability of each class. Considering only these two band spaces, the sparse vegetation class is completely included inside the class "Other" and because of that one might think of removing it. However, savanna (which represents the majority of the "Sparse Vegetation" class) is the most common type of land cover in the country and therefore it would not be appropriate to remove it. Additionally, later it will become clear that the separability of these two classes is not as bad as it seems here, which indicates that for separating these two classes other sets of bands are also useful. Another thing that is very evident is the fact that the class "Other" has very high variability: this was expectable as this class is used as an "umbrella-class" for a lot of small sub-classes not relevant for this problem (paddy fields, villages and bare ground). Some overlap between the pixels of the forest and cashew classes is also visible. Even though some of this overlap might possibly not happen in other band spaces, there is a real overlap of the spectral signatures which makes this classification task challenging.

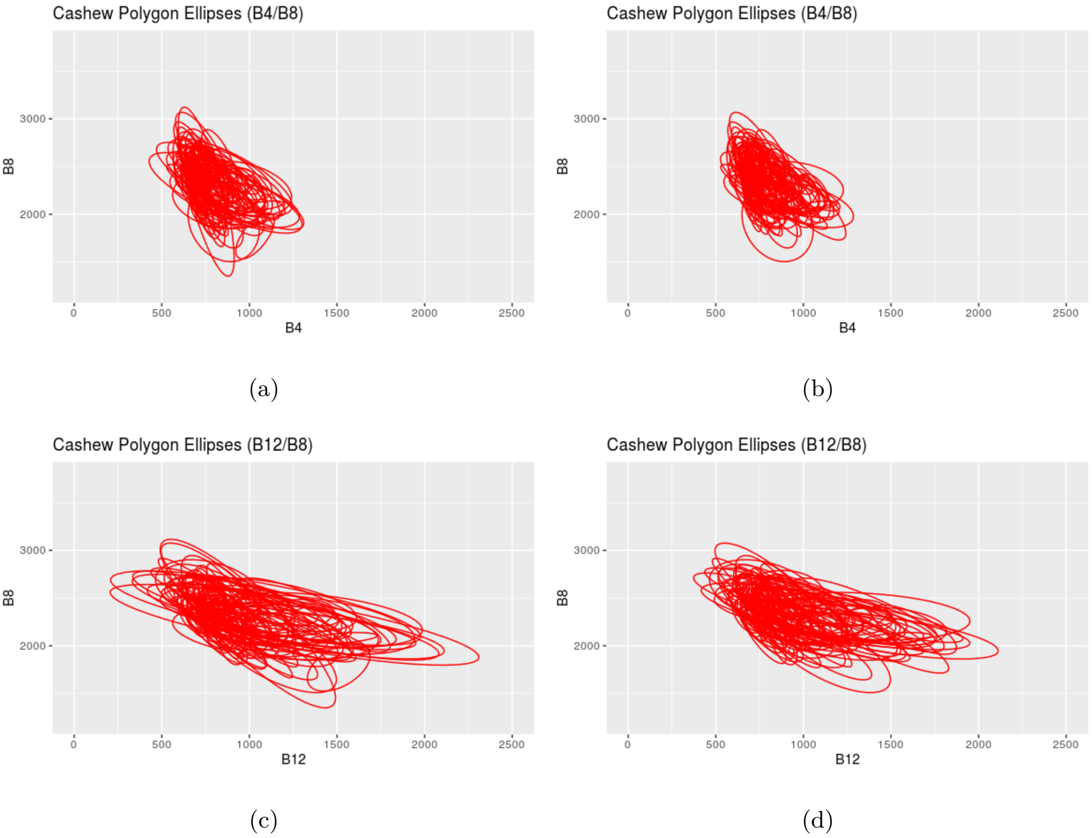


(a) (b)

(c) (d)

Figure 12: 95% density level ellipses for every class (a,b) and for every ellipse (c,d) in the B4/B8 and B12/B8 spaces.

Plotting the ellipses regarding polygons of individual classes might be very useful for cleaning the training data. One can check the polygon ID of the polygons corresponding to ellipses that seem suspicious: ellipses that are too broad or too far away from most dense region of the ellipses of a given class might indicate that the corresponding polygon is too heterogeneous or contains labeling errors. For this, `ggplot` and `plotly` were combined in order to create interactive plots of each individual class that exhibit the polygon ID on hover. This way, suspicious polygons were individually analyzed using Google Earth Pro's high resolution imagery, with a special emphasis on the "Cashew" class, which is the main focus of this dissertation. In case the inspected polygons were not adequate they were removed; if the number of polygons for a given class became severely decreased, new polygons were drawn (with the guidance of domain experts) to compensate the ones deleted. Figure 12 shows that the "Cashew" class overlaps with "Forest" (upper left related to the cashew) or "Sparse Vegetation" (lower right of the cashew) classes. In the case of the overlap with the "Forest" class, this is thought to be a "real overlap", meaning that these classes can actually be very hard to separate. However, the overlap with the "Sparse Vegetation" can indicate the presence of very heterogeneous polygons. Cashew orchards (in an informational sense) can, in fact, have large portions where the plants are still very young (Figure 10 a), and therefore the spectral signature will be more similar to that of sparse vegetation. Polygons containing cashew orchards in this situation will, therefore, present high variability in their pixels and have some spectral similarity with "Sparse Vegetation": this means that these polygons will correspond to very wide ellipses that overlap with ellipses from the "Sparse Vegetation" class. Due to this, most of the data screening focused on examining the appearance of these polygons, deleting them if they were, in fact, too heterogeneous and adding new polygons to replace the ones deleted. The result of this task can be explained through Figure 13: After the data screening (Figure 13 b and d), the cashew class ellipses are less spread out and more centered on the core compared to the ellipses prior to this step (Figure 13 a and c). Most of the improvement corresponds to the lower left region of the ellipses, which is precisely the region that overlaps with "Sparse Vegetation". Besides this, some small adjustments to the "Forest" class and also some labeling errors in other classes were taken care of.



(a)

(b)

(c)

(d)

Figure 13: 95% density level ellipses for every class (a,b) and for every ellipse (c,d) in the B4/B8 and B12/B8 spaces.

It is important to be aware of the fact that such task is always a trade-off: if a given class is too heterogeneous, the classification algorithm will have a hard time distinguishing the class from the others. However, if the class is too homogeneous, the classifier will miss pixels that are only slightly different from the norm. Even though the cashew polygons were a bit too heterogeneous, some variability was kept in them to try to overcome this limitation.

The vector data used in the following sections are this new improved "cleaner" version of the data. In total, the final data set was comprised of 25637 points of data (after removal of a small number of missing values, a process which is later detailed), extracted from 534 polygons (Tables 3 and 4). These data can now be intersected with the desired image composites to generate data similar to that on Table 2.

Table 3: Schematic representation of the vector data.

	Class_ID	Polygon_ID
P1		
P2		
P		
P 25638		

Table 4: Number of points of each class in the data.

Class	No. Pixels
Forest	8260
Cashew	3196
Sparse Veg (Savanna)	2037
Mangrove	2595
Other	3847
Water	5702

4.3 Classification Years with Ground Truth Data (2019)

In this section, the process for obtaining a classification map for the year corresponding to the ground truth data provided (2019) is detailed. For training the algorithm, only the labeled data (which consists in a very small percentage of all the pixels in the image, which can be seen through Figure 9) is needed. Therefore, in this step there is no need to work with the images themselves, which can be very large (more than 0.5 GB per image). Instead, a CSV file representing the training data can be directly exported from GEE. Only once the algorithm is trained the full image is retrieved and the algorithm applied to each pixel.

Depicting the vegetation dynamics often helps the classifier to distinguish between the different classes. For this reason, instead of considering a single composite representing the full season, four composites (one per month of the dry season, January - April) were combined (Figure 14, below). Bands 08 and 08A depict the same region of the spectrum with different spatial and spectral resolutions, and thus it only makes sense to keep one of them. Since this study is being performed at a 10m resolution, band 08 (which has that same resolution) was kept and band 08A discarded. Due to the reasons detailed in the Background section, bands 01 and 09 could also have been removed at this step. However, since there is some (although not much) atmospheric transmission in the corresponding regions of the spectrum, they were kept at this stage. Later, a variable selection algorithm will make clear if these bands are being useful for classification purposes. Each pixel is, then, characterized by four sets of eleven variables, as illustrated in Table 5 .

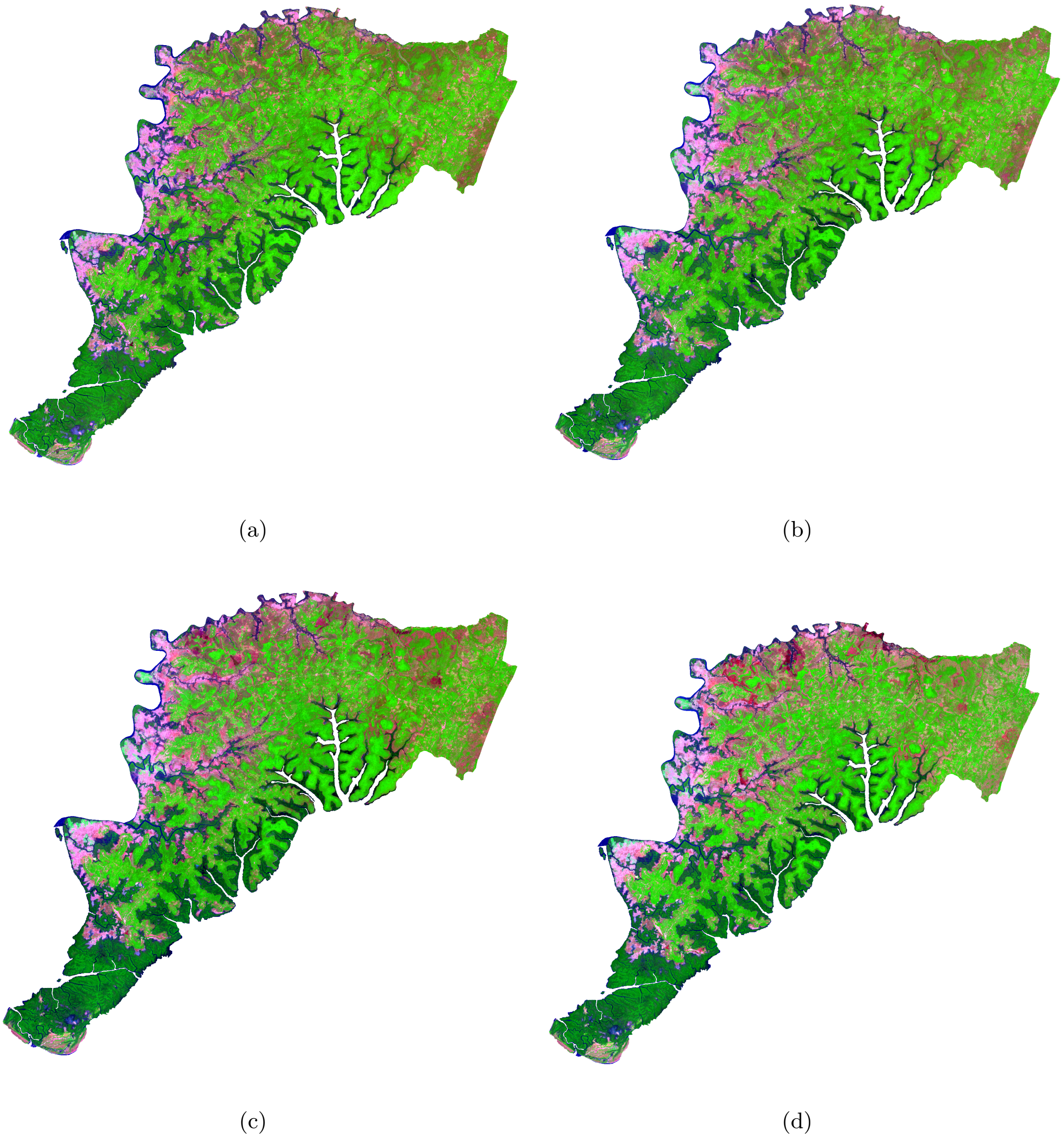


Figure 14: RGB images of the study area (Bands 12, 08 and 04) corresponding to four monthly median composite, one per each month of the dry season (Jan-Apr).

Given that the cashew orchards very often present a regular pattern (Figure 10 c), including textural information about the data can also help the classifier. For this, the GLCM was calculated. This is a square matrix centered in a given pixel in which the (i,j)th entry represents the number of times a pixel with intensity i is adjacent to a pixel with intensity j, inside the region being considered. After calculating the GLCM matrix for each band and around each pixel, multiple statistics that highlight the

Table 5: Schematic representation of the data setup (variables and data points). Gray-Level Co-Occurrence Matrix (GLCM)-related values are not represented.

	B1_J	B2_J	...	B12_J	B1_F	B2_F	...	B12_A	Poly_ID	Class_ID
P1										
P2										
P...										
Pn										

textural information contained in that region can be derived [36, 37]. GEE outputs 18 statistics, meaning that per each spectral band, 18 additional variables will be generated. Since the four composites have 11 bands each, the final feature space will be comprised of $19 \times 11 \times 4 = 836$ predictors. Variable selection techniques will be used to reduce this feature space later on.

To get a preliminary overview of the separability of the land cover classes, their spectral signatures were plotted (Fig 15). The figure shows the spectral behavior of different classes. Although for some bands the overlap can be very large (B09, for example), in others it is easier to distinguish between the different classes (B11, for example). The overlap between the "Forest" and "Cashew" classes is decreasing as the dry season goes on, meaning that the drier the vegetation, the easier it is to distinguish between these two classes.

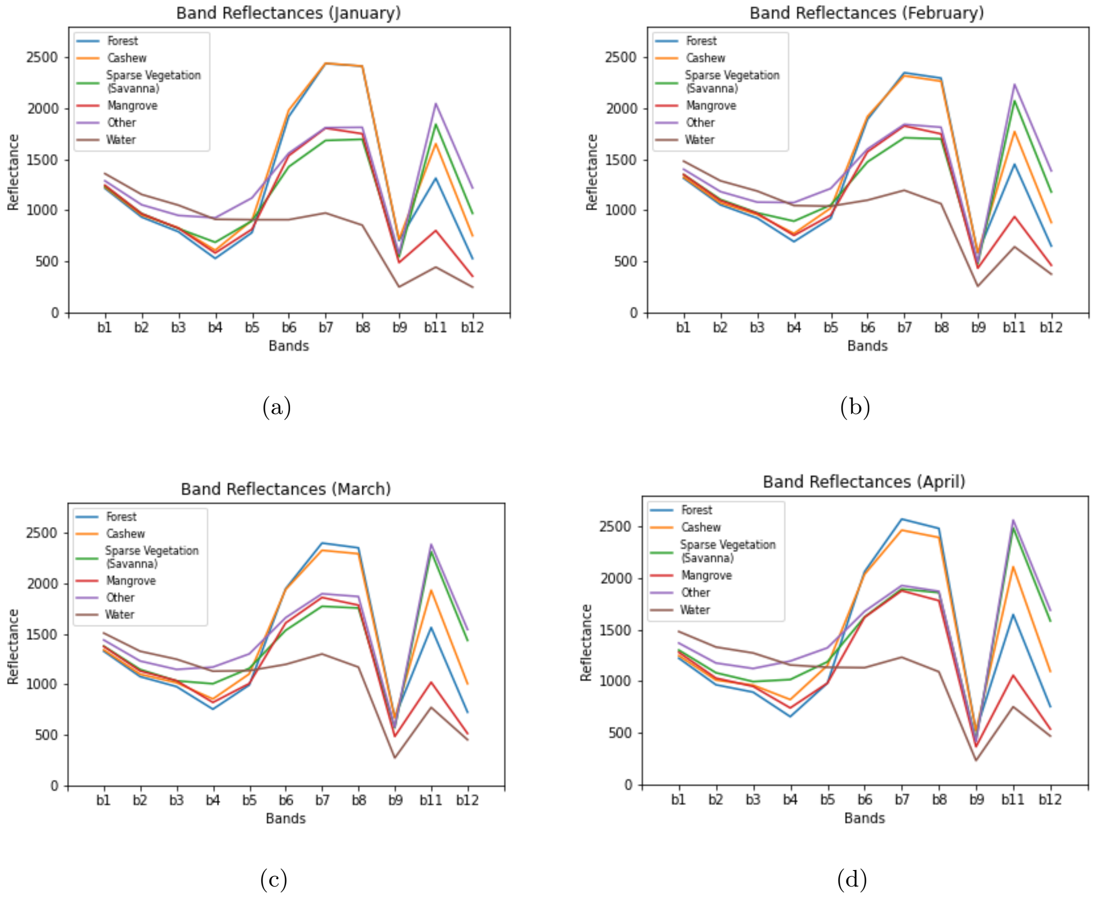
4.3.1 Data pre-processing

Missing Values

The first pre-processing step concerns the removal of a small number (=21) of missing values (NAs) that derived from the fact that the training points corresponding to those NAs were just outside the shapefile that delimited the ROI.

Cross-Validation Setup

For this particular analysis, an appropriate K-fold setup had to be developed: due to the tendency for spatial auto correlation (the tendency for areas or sites that are close together to have similar values), all the pixels belonging to a given polygon should remain together in the same fold, otherwise the accuracy might be overestimated, as there would be very similar pixels in the training and testing folds. In other words, the cross-validation should be grouped by the grouping factor polygon_id. Additionally, to account for the unbalanced scenario, the K-fold should also be stratified. Thus, a stratified-grouped-K-Fold is the necessary approach. Since such option is not available on Scikit-Learn, it was implemented manually, based on an approach found in Kaggle [38]. This way, each fold will have all the pixels of a given polygon and will be stratified according to Class_ID. Table 6 shows an example of such fold. Perfectly stratified



(a)

(b)

(c)

(d)

Figure 15: Average reflectance of each S2 band per class, in each of the monthly composites (a-d).

folds are not viable most of the times because polygons have different sizes and therefore a different number of pixels. This approach makes the best possible splits considering these constraints.

Table 6: Example of one grouped stratified cross-validation fold.

	Forest	**Cashew**	**Spr Veg**	**Mangrove**	**Other**	**Water**
Full training set	31.17%	12.10%	7.74%	10.06%	14.55%	24.38%
Development set - fold 1	32.84%	12.76%	7.98%	9.54%	15.32%	21.55%
Validation set - fold 1	21.36%	8.28%	6.39%	13.05%	10.05%	40.87%

This is the cross-validation setup used throughout this dissertation in multiple occasions, and an adaptation was also used to divide the data into training (70%) and test sets (30%), ensuring the same restrictions (grouped and stratified).

Scaling

Since one of the classification algorithms that will be tested is scale-sensitive, the data was standardized using `scikit-learn`'s StandardScaler.

Variable Selection

To obtain some insight on the appropriate number of variables to select, scikit-learn's rfecv was used to decide the number of variables to keep. This algorithm is based on choosing a wrapper algorithm which must contain a feature importance attribute that will ultimately be used to figure out the appropriate number of features. Given this, a RF classifier was picked as a wrapper algorithm. The ideal number of features is obtained by recursive feature elimination which is made using cross-validation. Given the imbalanced nature of the data, the balanced accuracy was used as the scoring metric. Figure 16 shows the result of this step and Table 7 the selected variables. Several observations can be made regarding the variable selection. First, the sum average statistic of the GLCM matrix is clearly very important for the classification, as it is present across every month and for almost every band. This variable is a weighted average not of the frequency of the pixel value itself in the neighborhood but of the frequency of its occurrence in combination with a certain neighbor pixel value, ultimately being one of many possible measures of contrast. The higher the sum_avg, the bigger the contrast in the image.

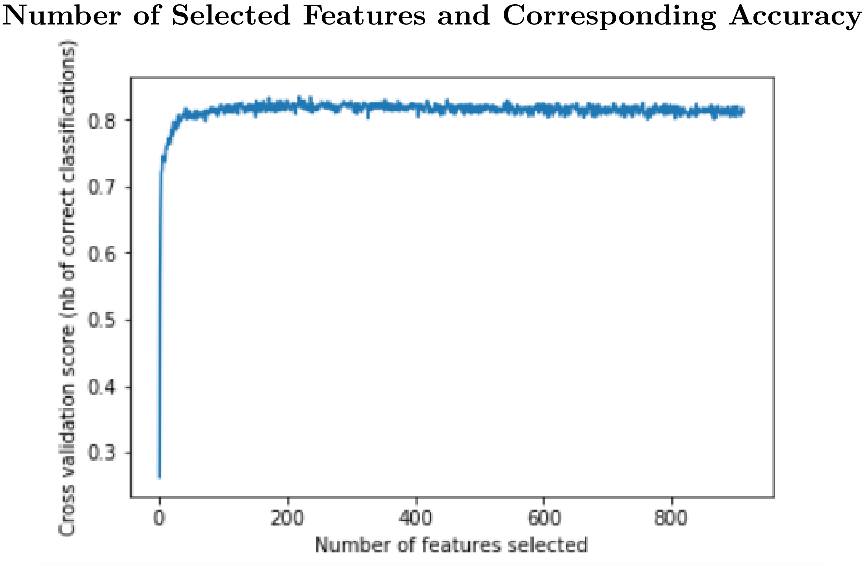


Figure 16: Number of selected features and corresponding accuracy.

In Figure 15, the separability of the classes seemed to be higher in April, the end of the dry season. The choice of variables is coherent with this observation, as the number of selected variables corresponding to April is significantly superior to that of the other months, meaning that there is more useful information for separating the classes in this month. Only in April GLCM-originating variables other than s_avg are selected, namely diss, inertia, contrast and dvar, which different ways of measuring texture in an image [36, 37]. Note also that for every month and against the expectation, the variable selection algorithm selects bands 01 (costal aerosol) and 09 (water vapor).

Table 7: List of the selected features.

Variables	**Jan/Feb/Mar**	**April**
B1	▨	▨
B1_savg	▨	▨
B2	▨	▨
B2_savg	▨	▨
B2_diss		▨
B2_inertia		▨
B3	▨	▨
B3_savg	▨	▨
B3_contrast		▨
B4	▨	▨
B4_savg	▨	▨
B4_diss		▨
B4_dvar		▨
B4_inertia		▨
B5	▨	▨
B5_savg	▨	▨
B6	▨	▨
B6_savg	▨	▨
B7	▨	▨
B7_savg	▨	▨
B8	▨	▨
B8_savg	▨	▨
B8_contrast		▨
B8_contrast		▨
B8_diss		▨
B8_dvar		▨
B8_inertia		▨
B8_var		▨
B9	▨	▨
B9_savg	▨	▨
B11	▨	▨
B11_savg	▨	▨
B12	▨	▨
B12_savg	▨	▨

4.3.2 Preliminary Classification

After the pre-processing steps, both a RF and a SVM were trained in the training set of the data. All default hyperparameters were kept, except for the number of trees in the RF that was set to 100. The default kernel for the SVM is the radial basis function. For assessing the accuracy of the algorithms, the implemented stratified grouped 10-fold cross-validation was used in the training set and both the balanced accuracy and the f1-score of the "Cashew" class were taken into account. The SVM seems to perform better not only for the "Cashew" class but also in general (Table 8).

Table 8: Cross-validation scores of the RF and SVM algorithms in the training set

	RF	**SVM**
F1 class cashew	83.99	86.31
Balanced Accuracy	91.00	92.72

4.3.3 Hyperparameter optimization

In order to improve the results and to be sure that the SVM was the appropriate choice of algorithm, the parameters of both algorithms were optimized using `hyperopt`. Instead of choosing the accuracy or the balanced accuracy as a scoring metric, the f1-score of the "Cashew" class was picked. This way, the algorithm will favor this metric over the global accuracy, meaning that the resulting parameters might result in a lower accuracy because they are optimized for a good performance in the "Cashew" class. The search spaces for every experiment presented in this dissertation were the following:

Random Forest

- Number of estimators: ranging from 10 to 490, with a step of size 10.

Support Vector Machine

- Cost: Log uniform search space ranging from -3 to 1 $[e^{-1}, e^{3}]$.
- Gamma: Log uniform search space ranging from -10 to 0 $[e^{-8}, e^{3}]$.

The algorithm ran for 100 iterations in the full training set. Even though the search spaces are quite large, it is expected that the algorithm makes more trials in the region where higher accuracy is attained and therefore that this does not end up being prejudicial to the search for optimal parameters. The following figure displays the accuracy for each value of the evaluated parameters. The SVM with optimized parameters resulted in a "Cashew" f1-score and balanced accuracy of 86.95 and 92.63 respectively, versus 84.24 and 90.94 using the RF optimal parameters (220 estimators). Given both the f1-score and balanced accuracy results the models, the SVM seemed to be the appropriate choice of model for the generation of the final land cover maps. The fact that the obtained values are very close to those obtained using the default parameters is not unexpected as the optimized parameters are very close to the default parameters (2.14 Cost and 0.03 Gamma vs 1 and 0.01). In addition to resulting in worse results, the optimization process for obtaining the ideal number of estimators for the RF does not seem to highlight a specific region of the search space and rather looks quite random, indicating that the optimal number of features obtained in the process can be highly variable. On the other hand, the performance of this algorithm seems to be less affected by the choice of parameters, as its performance suffers much more subtle changes upon trying out different parameters when compared with the SVM

Random Forest

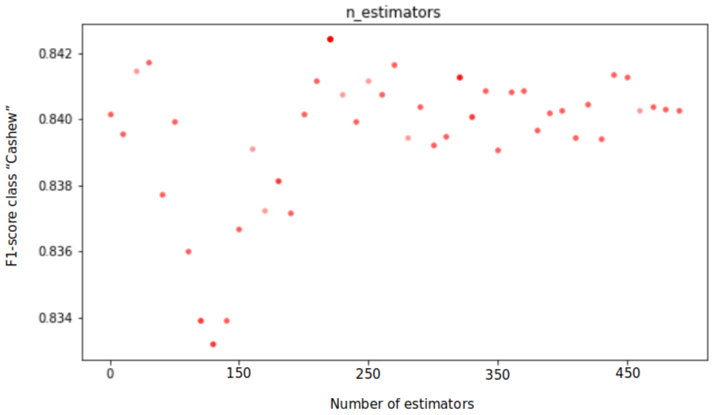
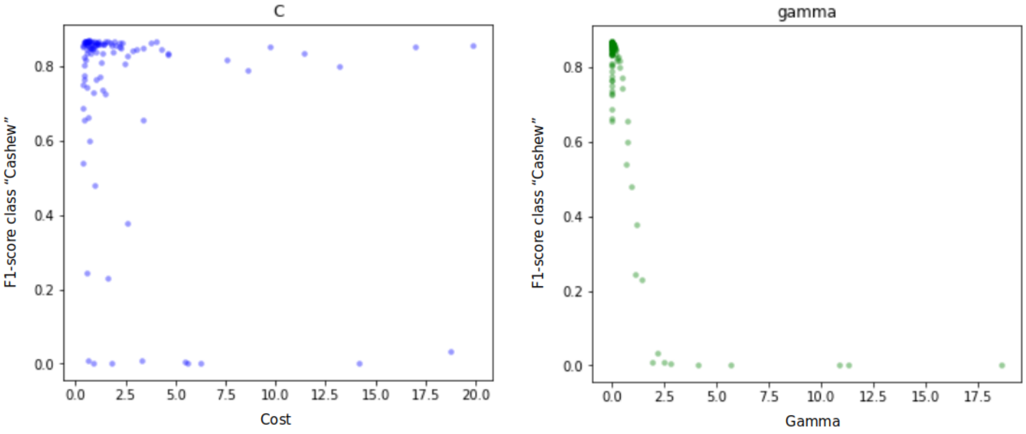


Figure 17: Optimization of the Random Forest's hyperparameters (section 4.3).

SVM



(a) (b)

Figure 18: Optimization of the Support Vector Machine's hyperparameters Cost (a) and gamma (b).

Considering these results, the SVM was the algorithm chosen to produce all the maps presented in this dissertation. Nevertheless, the RF results prior to optimization will always be depicted for informational purposes (Figures 17 and 18).

4.3.4 Test Set Classification and Land Cover Map

The model resulting from the previous step was applied to each pixel of the corresponding multi-spectral image. This image is also comprised of the same 100 bands used for training the model and classifying its pixels resulted in the following map (Figure 19).

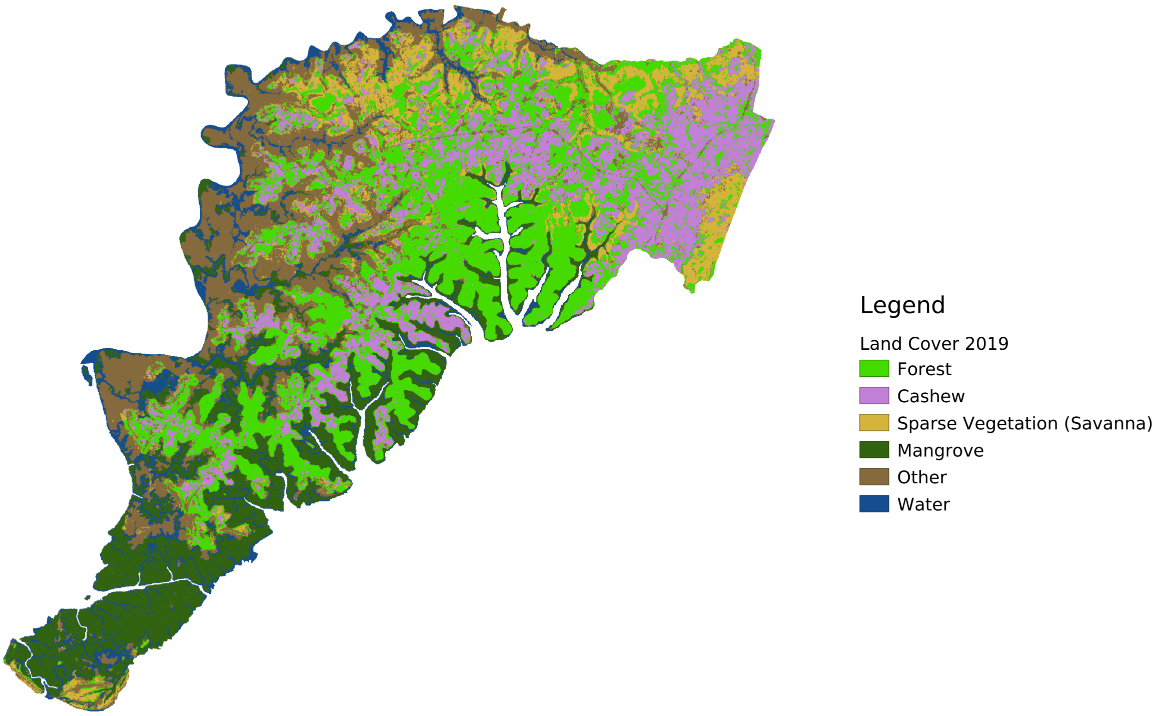


Figure 19: Land cover map resulting of the application of the classifier to all the 2019 pixels in the ROI.

Below in Tables 9 and 10 are the confusion matrix and other statistics regarding the application of the developed model to the test set. For assessing the accuracy of maps, two confusion-matrix derived statistics are usually presented additionally to the standard ones: the producer's and the user's accuracy [39]. The producer's accuracy is complementary to the omission error and is the accuracy of the map from the perspective of its producer, meaning that it corresponds to how often ground truth is accurately represented in the map. The user's accuracy is complementary to the commission error and depicts the user's perspective, as it represents how often a class in the map will be present in reality. These metrics are the same as precision and recall, respectively.

Table 9: Confusion Matrix corresponding to the performance on the test data regarding the algorithm used to produce Figure 19.

.

	Forest	Cashew	Sparse Veg.	Mangrove	Other	Water	**User's Accuracy**
Forest	2363	97	22	0	6	1	94.94
Cashew	182	754	1	0	18	0	78.95
Sparse Veg.	48	3	531	0	21	0	88.06
Mangrove	1	0	0	725	1	6	98.91
Other	12	21	51	0	1017	51	88.28
Water	0	0	0	2	83	1103	92.85
Producer's Accuracy	90.68	86.17	87.77	99.72	88.74	95.00	

Table 10: F1-score corresponding to the performance on the test data regarding the algorithm used to produce Figure 19.

	f1-score	Support
Forest	92.76	2489
Cashew	82.40	955
Sparse Veg.	87.91	603
Mangrove	99.32	733
Other	88.51	1152
Water	93.91	1674
Accuracy		91.19
Balanced Accuracy		90.33

The f1-score is the harmonic mean between the user's and producer's accuracy, and is a good way to take both values into account.

Although with a lower f1-score than that of the other land cover classes, the performance of the algorithm regarding the "Cashew" class it acceptable, reaching a value of 82%. To illustrate the functioning of the algorithm, in Figure 20, high resolution true-color images of two different portions of the ROI and their corresponding classification generated by the model are represented. Even though it might not be very clear for the inexperienced user, the classifier is correctly identifying the cashew orchards and labeling them as such, even in regions where the vegetation is still short and sparse. Additionally, the classifier is also identifying forest regions correctly and a portion of bare ground and a small road as "Other", which is also correct.

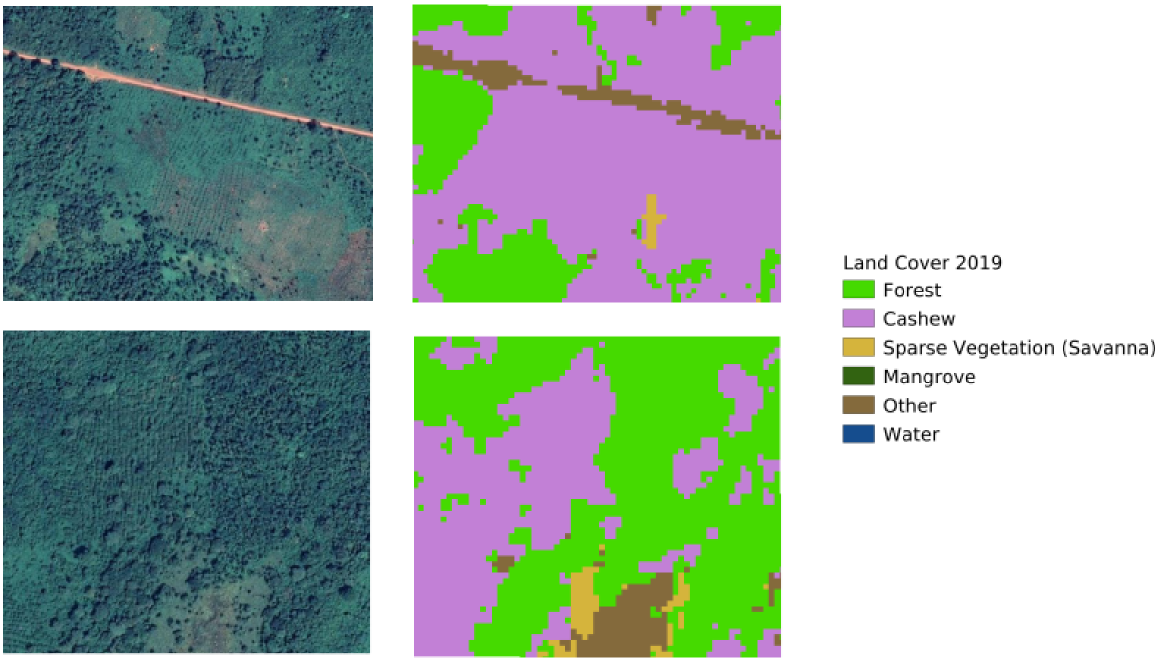
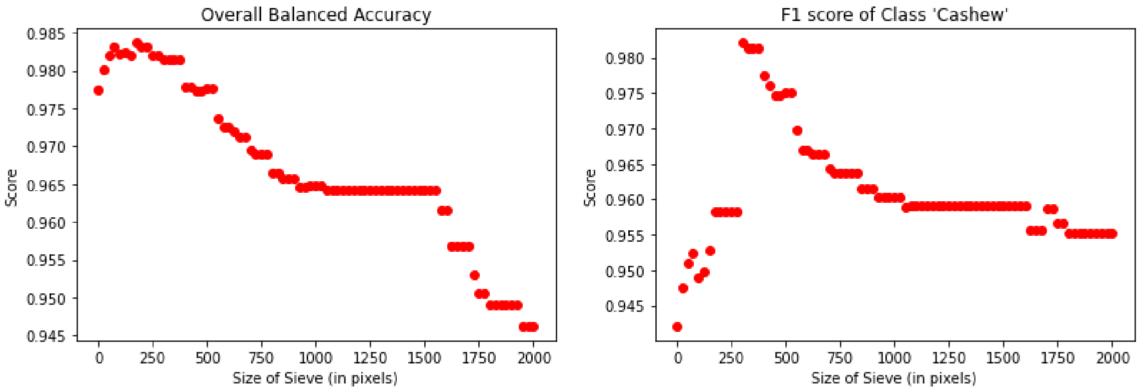


Figure 20: Two small areas inside the ROI and their corresponding classification.

4.3.5 Map post processing

The map shown in section 4.3.4 presents several single dispersed pixels and very small raster polygons that most likely correspond to noise, rather than to true regions of a given land cover type. GDAL's `sieve` algorithm is a good post-processing option in these situation, as it removes raster polygons (agglomerates of pixels of the same class) smaller than a given threshold, and replaces their value with that of the largest neighboring polygon, very often resulting in an increase in accuracy.



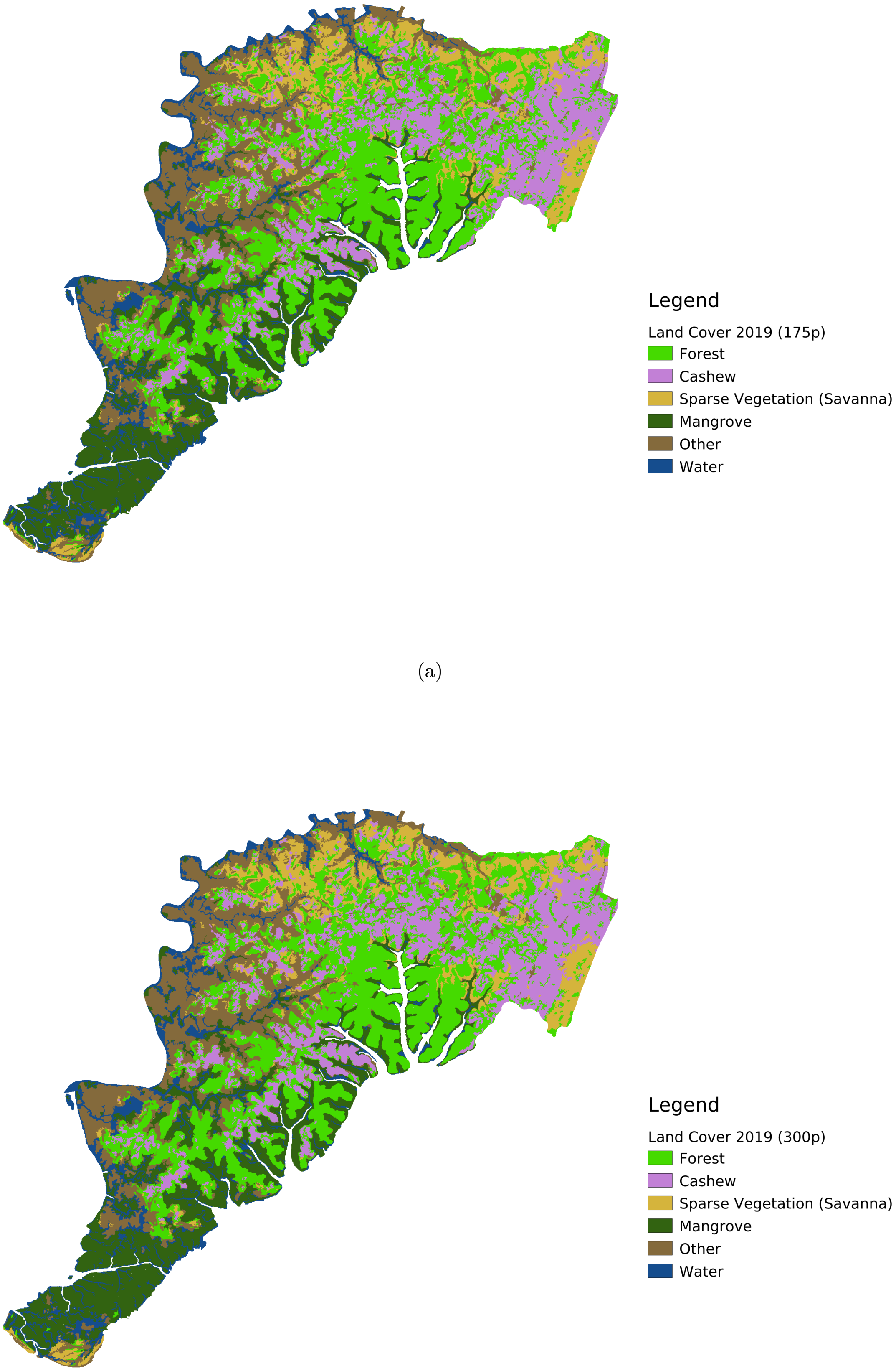
(a)

(b)

Figure 21: Balanced accuracy (a) and f1-score of class "Cashew" (b) for the tested sieve thresholds.

The choice of threshold is usually defined by the end user, as it can widely depend on the goal of each project. FAO's definition of forest states that for a given area to be considered as Forest it should have at least 0.5 ha (minimum mapping unit), so a sieve should have a threshold of at least that size (which corresponds to 50 sentinel pixels). However, larger thresholds can result in a higher accuracy. Ultimately, it is a trade-off between the amount of noise in the data (which very commonly deteriorates model performance) and spatial resolution. To illustrate this, sieves with a size ranging from 0 to 2000 pixels (0 to 20 ha) were applied to the map above and the resulting scoring metrics plotted (Figure 21). The graphs on Figure 21, used to decide on the threshold size, were developed using the training data.

The threshold that results in the highest accuracy is 175 pixels, which corresponds to 1.75 ha. Larger polygons result in lower accuracy, likely due to the fact that certain "true" polygons, rather than noise, are removed by this filter. For the "Cashew" specifically, the ideal threshold is considerably higher (300 pixels). However, using a sieve corresponding to the maximum f1-score of cashew will not only severely deteriorate the spatial resolution of the map, but also affect the accuracy of the remaining classes. The maps obtained using both maxima are represented below in Figure 22, and the corresponding performance in Tables 11 to 14 was assessed in the testing data of each sieved map.



(a)

(b)

Figure 22: Maps corresponding to a 175p (a) and 300p (b) sieve of the map represented in fig 19.

Sieve 175p

Table 11: Confusion Matrix corresponding to the performance on the test data regarding the algorithm used to produce Figure 22 a.

	Forest	Cashew	Sparse Veg.	Mangrove	Other	Water	**User's Accuracy**
Forest	2363	60	39	0	26	1	94.94
Cashew	163	778	5	0	9	0	81.47
Sparse Veg.	52	0	550	0	1	0	91.21
Mangrove	1	0	0	726	0	6	99.05
Other	12	13	35	0	1041	51	90.36
Water	9	0	0	3	73	1103	92.85
Producer's Accuracy	90.88	91.42	87.44	99.59	90.52	95.00	

Table 12: F1-score corresponding to the performance on the test data regarding the algorithm used to produce Figure 22 a.

	f1-score	Support
Forest	92.87	2489
Cashew	86.16	955
Sparse Veg.	89.29	603
Mangrove	99.32	733
Other	90.44	1152
Water	93.91	1674
Accuracy		92.15
Balanced Accuracy		91.64

Sieve 300p

Table 13: Confusion Matrix corresponding to the performance on the test data regarding the algorithm used to produce Figure 22 b.

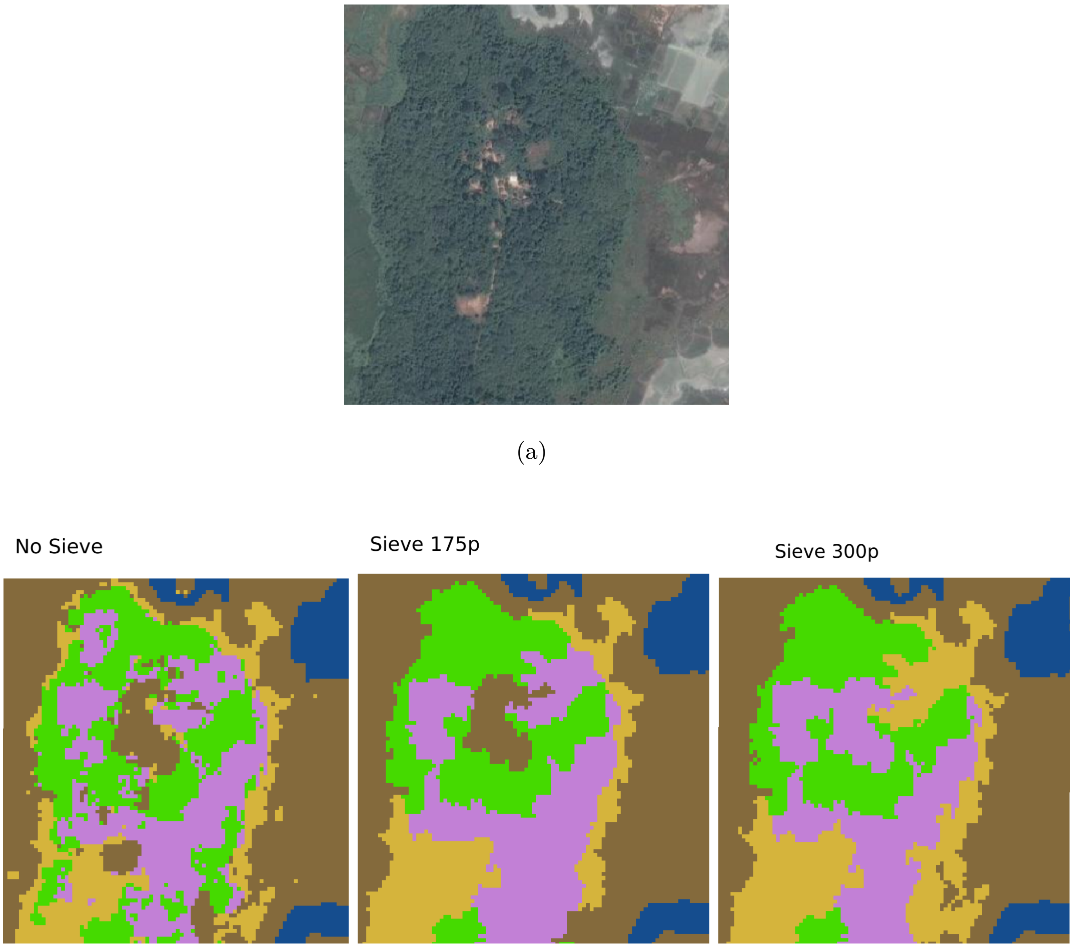
.

	Forest	Cashew	Sparse Veg.	Mangrove	Other	Water	**User's Accuracy**
Forest	2382	20	39	0	47	1	95.70
Cashew	157	784	5	0	9	0	82.09
Sparse Veg.	52	0	550	0	1	0	91.21
Mangrove	1	0	0	726	0	6	99.05
Other	12	13	48	0	1028	51	89.24
Water	9	0	0	3	73	1103	92.85
Producer's Accuracy	91.16	95.96	85.67	99.59	88.77	95.00	

Table 14: F1-score corresponding to the performance on the test data regarding the algorithm used to produce Figure 22 b.

	f1-score	Support
Forest	93.38	2489
Cashew	88.49	955
Sparse Veg.	88.35	603
Mangrove	99.32	733
Other	89.00	1152
Water	93.91	1674
Accuracy		92.32
Balanced Accuracy		91.69

As the scale of the maps might be too coarse to get a proper understanding of the sieve's effect, Figure 23 compares the effect of these two sizes of sieve with the original map for a small sub-area of the ROI. In this zoom-in, the effect of the sieve is clear: raster polygons smaller than the threshold are being replaced with a new value that corresponds to the value of the largest neighboring polygon. For the 300p sieve (Figure 23 d), the small village in the middle of the area disappears, being entirely classified as Forest.



(a)

(b)

(c)

(d)

Figure 23: Zoom-in of a small area inside the ROI (a) in the maps without sieve (b), with a 175p sieve (c) and a 300p sieve (d).

4.4 Classification of Years Without Ground Truth Data (2020)

As mentioned in the objectives, one of the goals of this work is to explore the possibility of expanding this system so it can produce valid land cover maps for years without land truth data, a process known as classification extension [40]. Retraining the model with new ground truth data for each new satellite image from other years instead of performing classifier extension would probably be more accurate and straightforward. However, only 2019 ground truth data is available, and ground truth data is only valid for the year of its collection: even though land cover changes happen slowly and progressively, alterations in the training sites can occur after (or before) their labelling by experts and therefore their label might not remain unchanged throughout the years. Due to this reason, it is not recommended to retrain a new land cover model for a new year using ground truth data obtained for another year, meaning that retraining the model implies the collection of new (or at least updated) data. This is the reason why it is important to attempt to perform classification extension, not only because in this case there is no other data available but also because otherwise automation will never be attained, as the system would require that new ground truth data was permanently acquired and that the model was constantly retrained.

4.4.1 Classification extension problem

Classification extension is an alternative to retraining the model for each new year, which eliminates the need of more data (thereby reducing cost) and allows a more automated classification process. However, it poses some well-known difficulties: this approach is rarely as accurate as retraining the model because it is much more sensitive to the radiometric variability present in the satellite data that results from differences in atmospheric (weather) and land surface (appearance) conditions [41]. In other words, the reflective properties of a given area (and thus the corresponding pixels in the satellite imagery) can present very high variability between years, meaning that even when considering the same area and season, the interannual variability of the images can be very high. Therefore, a model that is trained for a given image will very like present poor performance when applied to another image without being re-trained; this is known as the signature extension problem [42].

The compositing criteria can be determinant when dealing with this problem: using the NDVI maximum as a compositing criteria can render very good results, as it maximizes the spectral signal of the vegetation. However, since it selects a maximum value, it presents a very larger variability between years, and therefore would have made the task of making this classifier capable of dealing with other years even harder. The median is a much more stable, robust metric and, therefore, a better option for trying to make the classifier transposable.

To illustrate the signature extension problem, the model trained in the previous section for the year 2019 was applied in a composite regarding the year 2020 (Figure 24 and Tables 15-16). There is not an ideal way to assess the accuracy for the year 2020, since no ground truth data are available for this year. Since 2020 is only 1 year apart from the year for which ground truth is available and land cover changes very slowly, an assumption regarding the immutability of these polygons between 2019 and 2020 was made, in order to make it possible to evaluate the accuracy and other metrics. It is clear from both the performance metrics and the resulting map that there is an over-representation of the "Cashew" class and an under-representation of the "Forest" class when trying to perform classification extension using the model developed in section 4.3, as the extent of cashew seems to grow too drastically from 2019 to 2020. As seen during the data exploratory analysis, the spectral signature of these two classes is very similar and therefore their separability is small, and training in 2019 and testing in 2020 does not work well. Since the goal of these maps would be to have a reliable assessment of the expansion of cashew orchards in the region, this approach is not indicated as it does not represent a realistic increase in the extent of cashew. The amount of confusion with the "Sparse Vegetation" class also increases significantly. Therefore, performing classification extension to 2020 with the algorithm trained in section 4.3 for 2019 is not a good approach.

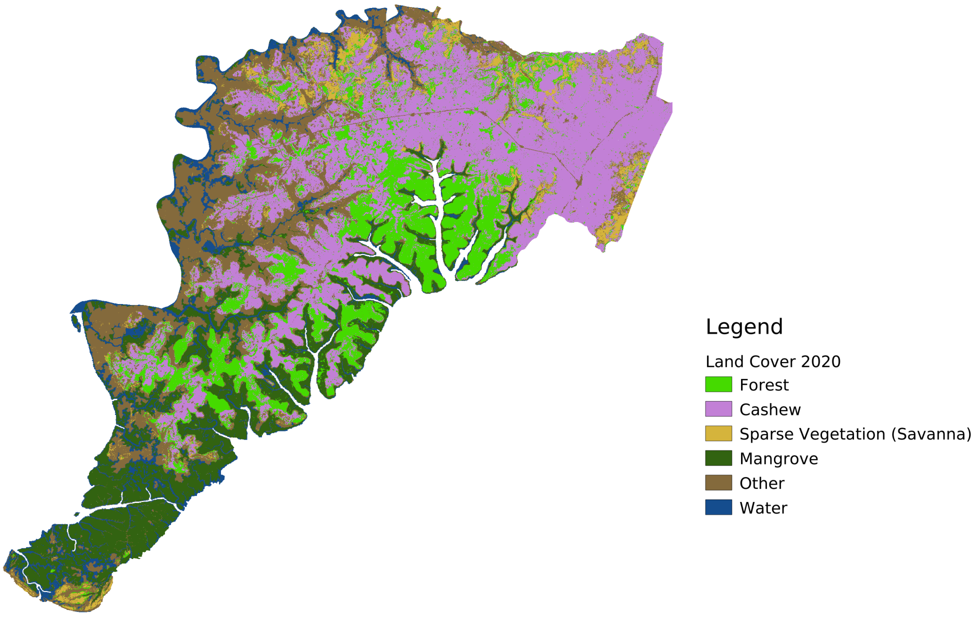


Figure 24: Map resulting of the application of the classifier to all the 2020 pixels in the ROI.

Table 15: Confusion Matrix corresponding to the performance on the test data regarding the algorithm used to produce Figure 24.

.

	Forest	Cashew	Sparse Veg.	Mangrove	Other	Water	**User's Accuracy**
Forest	1682	784	16	0	7	0	67.58
Cashew	29	917	1	0	7	0	96.12
Sparse Veg.	37	129	423	0	14	0	70.15
Mangrove	13	0	0	708	7	5	96.59
Other	3	86	91	32	922	18	80.03
Water	0	0	0	39	83	1067	89.74
Producer's Accuracy	95.35	47.86	79.66	90.89	88.65	97.89	

Table 16: F1-score corresponding to the performance on the test data regarding the algorithm used to produce Figure 24.

	f1-score	Support
Forest	79.10	2489
Cashew	63.90	955
Sparse Veg.	74.60	603
Mangrove	93.65	733
Other	84.12	1152
Water	93.64	1674
Accuracy		80.32
Balanced Accuracy		83.36

4.4.2 Single Composite Approach

The data design used in the previous sections implies that each pixel is characterized by 100 variables. When the goal is to classify the image in which the algorithm was trained, this helps the classifier to distinguish between classes as it makes them more specific. However, this setup poses some problems regarding the extension of the system for other years without retraining. The large feature space ends up being too specific when considering the inevitable shift in the distribution of values that occurs between different images of different years, due to both different land and atmospheric conditions that are the cause of the radiometric variability. This way, in order to create a more generalist classifier, able of classifying not only the year in which it was trained but also able to extrapolate to other images, a more generalist setup is needed. The possibility of considering one single image composite that represents the full season (instead of four, one per month) is now explored. Since it does not provide as much information as the previous setup, this approach is expected to produce a classifier that is not as accurate as the previous one but that is at the same time able to extrapolate better when classifying other years. This section describes an intermediate step necessary to attain what is described in the next section.

This simplified version has as a starting point the same setup used for data cleaning (table 2): there is only one set of bands that combines information for all the months of the 2019 dry season, which corresponds to 11 predictor variables, as band 10 was also discarded for the same reasons as before. The GLCM-derived variables were once again included, resulting in 18 additional bands per each pre-existing band. Since the composite is comprised of 11 bands, the final feature space will be comprised of 19 x 11 = 209 predictors. The spectral signature is very similar to those presented for the monthly composites (Figure 25), which is not a surprise as it represents a summary of the information contained in those images.

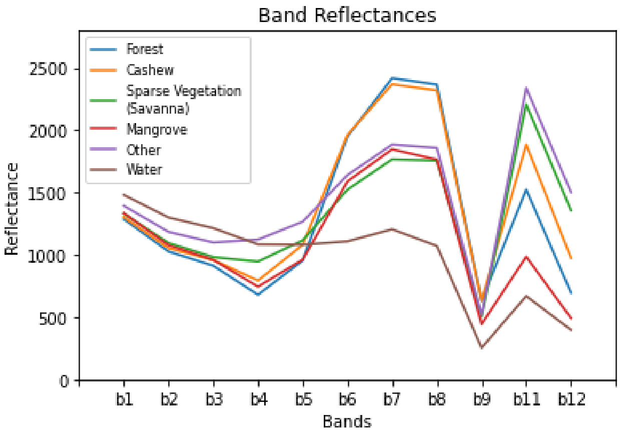

Band Reflectances

Figure 25: Reflectance of each S2 band per class, in a composite representing the full dry season.

This is considered to be an intermediate step needed to understand the next section, which describes the complete proposed approach for dealing with classification extension. Because of this and also because most of the steps are similar to those described throughout section 4.3, a summarized version of the pipeline is here presented, and most of the corresponding charts and tables are in the supplementary material section.

4.4.2.1 Data Pre-Processing

The same pre-processing steps from section 4.3 were taken. Given that the set of predictors is now different, so are the variables resulting from the selection process (fig. 26 and Table 17). The sum_avg statistic of the GLCM is once again very relevant for this new classification problem, as it is selected for almost every band.

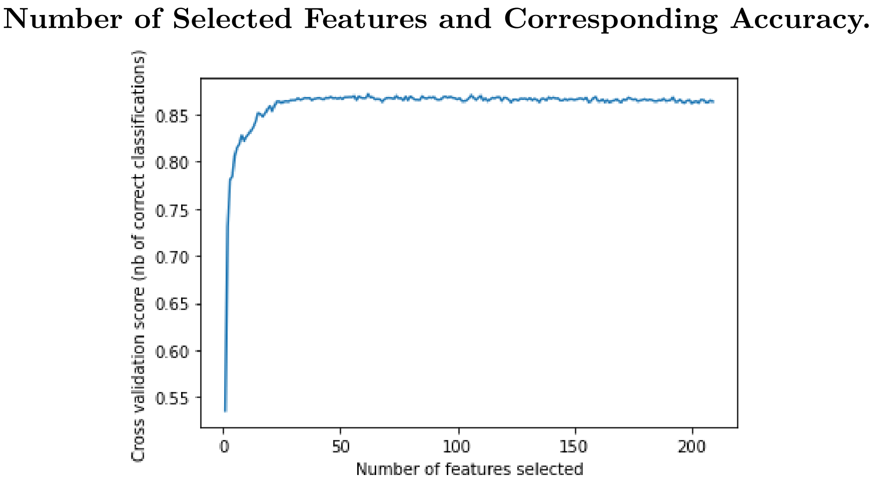

Number of Selected Features and Corresponding Accuracy.

Figure 26: Number of variables selected using rfecv.

Table 17: List of the selected variables.

Variables	
B1	B6
B1_savg	B6_savg
B2	B7
B2_savg	B7_savg
B3	B8
B3_savg	B8_contrast
B3_inertia	B8_savg
B4	B9
B4_savg	B9_savg
B4_dvar	B11
B5	B11_savg
B5_savg	B12
	B12_savg

4.4.2.2 Preliminary Classification

Exactly in the same conditions described in section 4.3, the following results (Table 18) were obtained when using this setup:

	RF	**SVM**
F1 class cashew	75.71	80.39
Balanced Accuracy	86.42	87.28

Table 18: Cross-validation scores of the RF and SVM.

4.4.2.3 Hyperparameter optimization

As the SVM is the algorithm to be used in the classification maps due to what was observed in section 4.3, only the SVM was optimized using the same parameters, search space and number of trials. The best Cost and Gamma parameters were 10.61 and 0.03 respectively resulting in 82.45 "Cashew" f1-score and 87.99 balanced accuracy, respectively (Figure S1). In this case, the improvement in performance attained with hyperparameter optimization is larger than before.

4.4.2.4 Test Set Classification

The test portion of the data was tested for both 2019 and 2020 and the results are presented in Tables S3 to S4. As expected, with this approach the results for the vegetation classes are not as good as with the previous setup, as the dynamics within the season is not being depicted. The mangrove is an exception to this pattern. For the remaining classes that do not represent vegetation, there is not much of a difference and in some cases their accuracy is even higher with this setup. This is probably due to the fact that these classes do not present variability between the months (for example, a village looks the same regardless of the time of the year) and therefore including variability is actually including noise, which worsens their results. The mangrove also follows this pattern as it's appearance is very consistent throughout time. Because of this combination of factors, the overall accuracy and balanced accuracy are similar between the approaches. In addition to the improvement of the f1-score of class "Cashew", the fact that the user's and producer's accuracy is more even than before is also very important, as it gives a more realistic view of the evolution of the cashew scenario. For the year 2020, the performance is better with this approach. This results in a significant improvement in the accuracy for the classification of this year.

4.4.3 Incorporating several years of data into training

The variability between years clearly still poses a problem, even when adopting a more general approach as the one in the previous section. Another measure that can help in the classifier extension process is to try to include the interannual variability into the training process. For that, the algorithm can be trained in several years of data simultaneously, in order to be prepared to face the variability that exists between the years. Similarly to what was made in section 4.4.1 to be able to access the performance of the algorithm in 2020, an additional immutability assumption will be made here regarding the year 2018 in order to incorporate two years of data into the model (2019 and 2018) and then test it in 2020. A small improvement of performance is expected, since interannual variability is being incorporated in the classifier. However, this is just a proof of concept as it would take more years of data to fully prepare the classifier to face all the possible interannual variability. Each pixel is now present in the data twice, one time regarding 2018 and another relative to 2019 (table 19).

Table 19: Schematic representation of the setup used throughout this section.

	B1	B2	B3	B4	B5	B6	B7	B8	B9	B11	B12	Poly_ID	Class_ID
P1_2018													
P2_2018													
P..._2018													
P1_2019													
P2_2019													
P..._2019													

The spectral signature graph is similar to the one presented in Figure 25, apart from the difference that it contains twice as much data. This makes sense, as essentially there are two copies of each pixel that will only exhibit slight variations.

4.4.3.1 Data Pre-Processing

The same number of features from section 4.4.2 was considered; although there is more data now, the additional data represents almost replicas (only very slight variations) of pixels already present, thus allowing more variables to be included in the model would most likely introduce noise and deteriorate model performance. The set of selected predictors is almost exactly the same which is not unexpected due to these reasons) (table 20).

Table 20: List of the selected variables.

Variables	
B1	B6
B1_savg	B6_savg
B2	B7
B2_savg	B7_savg
B3	B8
B3_savg	B8_contrast
B3_contrast	B8_savg
B4	B9
B4_savg	B9_savg
B4_inertia	B11
B5	B11_savg
B5_savg	B12
	B12_savg

4.4.3.2 Preliminary Classification

Exactly in the same conditions described in section 4.3, the following results (Table 21) were obtained when using this setup.

Table 21: Cross-validation scores of the RF and SVM.

	RF	**SVM**
F1 class cashew	75.90	80.02
Balanced Accuracy	86.62	87.12

4.4.3.3 Hyperparameter optimization

With this setup, the optimal choice of Cost and Gamma was 12.95 and 0.12 resulting in 82.14 "Cashew" f1-score and 87.50 balanced accuracy, respectively (Figure S2). Again, the improvement in performance attained with hyperparameter optimization is larger than in section 4.3.

4.4.3.4 Test Set Classification, Post Processing and Land Cover Maps

To allow for a fair comparison between the maps, a sieve of the same size was applied to both the years. Based on results shown in Figure S3, a sieve of size 125 was chosen for demonstration purposes. The resulting maps and statistics are below in Figurel 27 and Tables 22 to 25. These statistics represent unexpected results, as the performance deteriorates when incorporating another year of data into the training set. However, since using only one year of data was considered to be an intermediate step with the final goal of reaching this, the final maps concerning an "extensible" algorithm were produced using this approach. These outcomes will be further discussed in the conclusions section.

2019

Table 22: Confusion Matrix corresponding to the performance on the test data regarding the algorithm used to produce Figure 27 a.

	Forest	Cashew	Sparse Veg.	Mangrove	Other	Water	**User's Accuracy**
Forest	2440	25	5	0	19	0	98.03
Cashew	305	639	11	0	0	0	66.91
Sparse Veg.	46	0	542	0	15	0	89.88
Mangrove	3	0	0	724	1	5	98.77
Other	9	53	78	0	960	52	83.33
Water	0	0	0	2	87	1099	92.51
Producer's Accuracy	87.05	89.12	85.22	99.72	88.72	95.07	

Table 23: F1-score corresponding to the performance on the test data regarding the algorithm used to produce Figure 27 a.

	f1-score	Support
Forest	92.21	2489
Cashew	76.44	955
Sparse Veg.	87.49	603
Mangrove	99.25	733
Other	85.94	1152
Water	93.77	1674
Accuracy		89.94
Balanced Accuracy		88.24

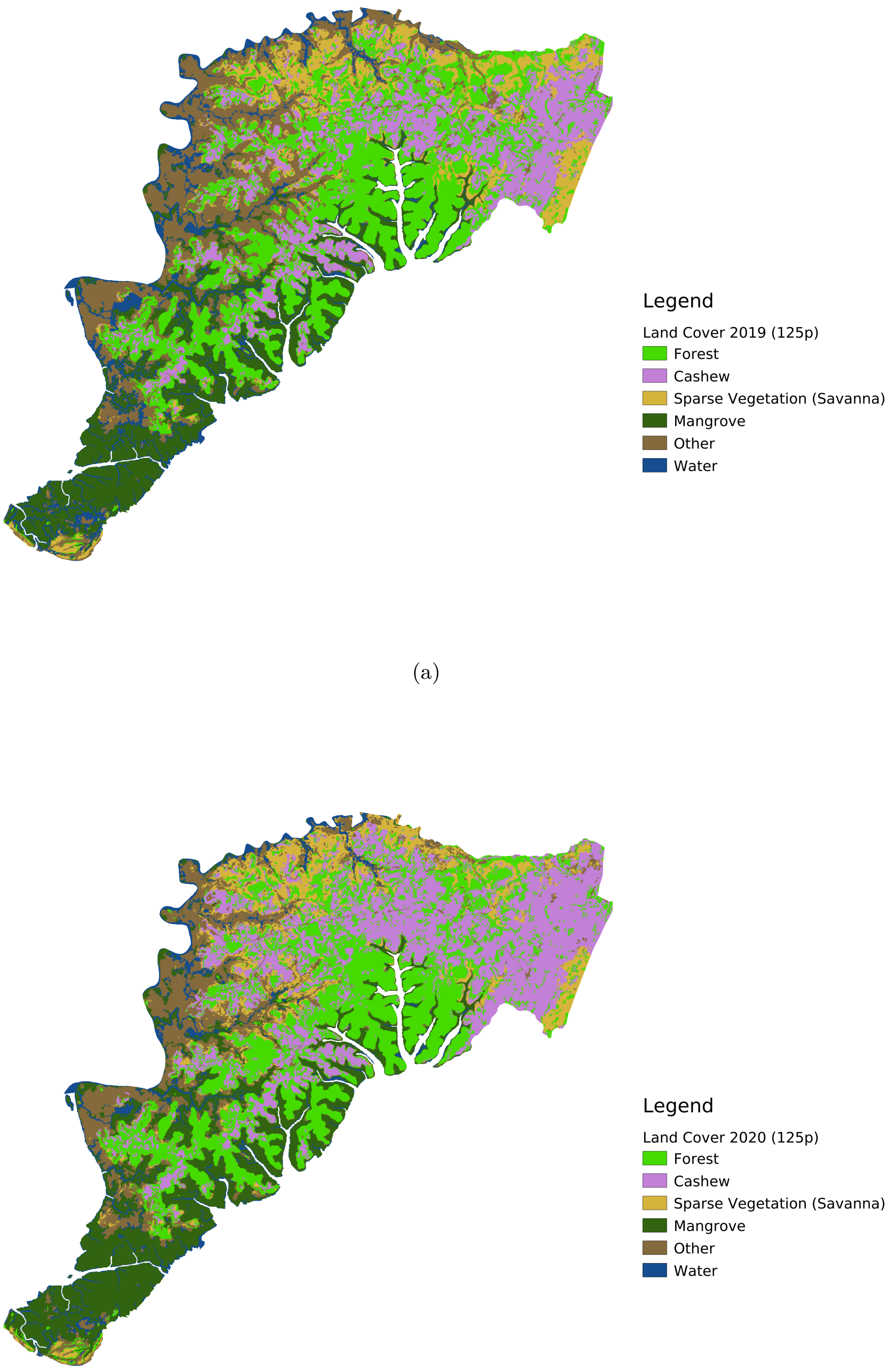
2020

Table 24: Confusion Matrix corresponding to the performance on the test data regarding the algorithm used to produce Figure 27 b.

	Forest	Cashew	Sparse Veg.	Mangrove	Other	Water	**User's Accuracy**
Forest	2354	124	11	0	0	0	94.58
Cashew	308	644	0	0	3	0	67.43
Sparse Veg.	57	119	397	0	30	0	65.84
Mangrove	30	0	0	700	3	0	95.50
Other	25	104	233	0	739	51	64.15
Water	0	0	0	83	104	1001	84.26
Producer's Accuracy	84.86	64.98	61.93	89.40	84.07	95.15	

Table 25: F1-score corresponding to the performance on the test data regarding the algorith mused to produce Figure 27 b.

	f1-score	Support
Forest	89.45	2489
Cashew	66.19	955
Sparse Veg.	63.83	603
Mangrove	92.35	733
Other	72.77	1152
Water	89.38	1674
Accuracy		81.95
Balanced Accuracy		78.03



(a)

(b)

Figure 27: Maps corresponding to the application of the algorithm described in this section and the posterior application of a 125p sieve. The non sieved corresponding map is not shown.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

Recap

Through this work, the possibility of monitoring cashew orchards in Guinea-Bissau using remote sensing techniques and machine learning models was explored. The first goal was to accurately map cashew orchards in the year for which ground truth data is available (2019 in this study) by training a model in small delimited labeled regions of the image and then applying the trained classifier to all the pixels of the ROI. The second and last goal was to explore methodologies that allowed the classification of images other than those in which the model was trained, without retraining the model. This is called classification extension. Addressing this second goal is the first step for building an automated system capable of monitoring cashew orchards without the need of constant retraining. This capability would circumvent the need for acquiring costly additional ground truth data for every new classification needed in up-to-date satellite monitoring of land cover change. The second task is very hard, as the inter-annual variability in both the atmospheric (weather) and ground surface conditions (state of the land targets, such as dryness) causes variations in the signal reaching the satellite between images from different years, even if they are retrieved at approximately the same time of the year. This is reflected in the data as a shift in the distribution of the predictor variables, which will ultimately result in the deterioration of the performance of the trained classifier once applied to new images.

Classification Years with Ground Truth Data (2019)

This work indicates that when ground truth land coverage data is available for a given year, a classifier can be successfully trained using those data to produce accurate land cover maps regarding that year. Even though the performance for the "Cashew" class is lower than that of the other classes, with the approach detailed in section 4.3 satisfactory results can be obtained even if there is some difficulty in distinguishing the spectral behaviours of cashew orchards, forests and sparse vegetation. The performance of both an SVM and a RF was assessed, as these are the state-of-the-art algorithms (together with ANNs) for this type of problem. The SVM performs better than the RF, and hyperparameter optimization has not proven to be very useful in this case, resulting in only a very slight improvement of the scoring metrics. This was not surprising, as the optimized hyperparameters are very close to the default parameters.

Classification of Years Without Ground Truth Data (2020)

As expected from what was described in the beginning of this chapter, the classifier designed for the first goal performs very poorly when used to classify images corresponding to other years of data (2020 in this study), even when considering the exact same area and time of the year (which is the case). To tackle this problem, a more generalist setup was developed. This new setup contains only a summary of the spectral information used in the previous approach compressed in a smaller number of variables, making it less specific but at the same time more capable of generalizing and thus classifying images that the model has never seen with a better performance than that of the classifier previously described.

Finally, an improvement of this generalist approach was carried out: another year of image data (2018) was included into the model as an attempt to incorporate the interannual variability into the training process. For this, an immutability assumption was made: the 2019 ground truth land cover data was considered to be valid for 2018, as these dates are only one year apart from each other. This study shows that incorporating another year of data into the training set slightly deteriorates the performance of the model when compared to the generalist approach using a single year of data. However, the results between these two steps are not directly comparable as the results of the generalist approach that uses only one year of image data were not used to produce any maps and thus not to subject post processing. Thus, this deterioration in performance may be circumstantial and caused by several factors: first, both the ground truth data for 2018 and 2020 are considered to be exactly the same as the original 2019 data, which is unlikely to be the case. In addition, using only two years of image data for training the model is most likely not enough to make a more robust model and might even be adding noise when trying to classify 2020. This is the reason why, even though a new variable selection step was made when incorporating another year of image data into the model, the same number of variables was considered. Otherwise, an additional number of variables could actually introduce noise rather than providing useful information, since the added data is a close replica of the initial 2019 image data. In retrospective, maybe not only the number of variables but also the features themselves (the bands and texture metrics) should have also been kept constant between these steps.

If more and more data regarding several years were to be incorporated into the model, its extension capability would likely improve significantly and its performance could be similar to the performance obtained by retraining the classifier each time new ground truth data for another year are available.

In the experiments regarding this topic, the SVM performs better than the RF again. Hyperparameter optimization of the SVM is in this case more useful because the increase in model performance is higher than before, and the optimized parameters are more distinct from the default.

Map post processing

Post processing the maps by applying a sieve algorithm (an algorithm that removes agglomerates of pixels of the same class smaller than a given size) results in an increase in the performance metrics, as a lot of noise ends up being removed from the maps. However, the final decision regarding the size of the sieve depends on the end user, as sieving ultimately degrades the spatial resolution of the maps. Figures 22 and S3 suggests that a threshold in the 100-200p range results in the highest value of balanced accuracy, which might be a better choice than choosing a larger sieve that maximizes the Cashew F1-score, as the performance regarding the class Cashew is almost equal between the two sieves and with that choice the performance regarding the remaining classes is not degraded. Additionally, the smaller the sieve, the lower the loss of the spatial resolution of the maps.

The sieves also allow to draw conclusions regarding the spatial resolution used in this study: since higher accuracy can be attained using a lower spatial resolution, then is it most likely not worth it to re-sample the bands to a 10m resolution. Given what is observed in this study, using a 20m resolution seems to be sufficient and it will both reduce the amount of data (however not in this specific case, as the points were sampled using a 25m grid) and most likely eliminate noise from the data.

Variable selection

Regarding variable selection, some interesting remarks can also be made. One surprising observation is the fact that bands 01 and 09, which are not meant to be used in remote sensing of the land (they target atmospheric conditions), are consistently selected by the variable selection algorithm, which ultimately means that they are being useful for the classification process. In addition, textural information (which is derived from spectral data) is clearly useful for the classification, as the variable selection algorithm consistently select variables representing statistics derived from the GLCM matrix. At a certain stage of their development, cashew orchards present a very visible row pattern (figure 10 b), which is probably being taken into account in these textural variables and ultimately helps during the classification process. Additionally, the sum average variable, which is one of many ways of measuring contrast in an image seems to be the most determinant feature regarding texture. In section 3.3, more variables regarding April are selected than those of any other month of the dry season (Table 7), meaning that out of the four months being considered (January to April), April is the most useful for separating between the classes. This is likely due to the fact that in April, the end of the dry season, most herbaceous species present in the ground below cashew orchards and forests dry out, enhancing the contrast between tree canopies and the bare ground underneath. This can ultimately can make the signal that reaches the satellite less noisy and enhance the textural patterns of the cashew orchards, thus facilitating the classification process. In addition, some of the species present in the forest are deciduous and therefore shed during the dry season, which might also contribute in helping to distinguish between cashew and forest. Finally, the fact that the herbaceous plants end up drying might also enhance the cashew orchard's row pattern.

Final remarks

Considering this work as a starting point, an automated (or semi-automated) cashew orchard detection and monitoring tool capable of classifying years that were never seen by the model (*i.e.*, that does not need constant retraining) can be developed using a more generalist setup than the one described initially for classifying the images of the year for which ground truth land cover data is available. Such an approach relies on classification extension and in incorporating data from many different years into the training process, in order to account for the interannual variability. Ground truth data regarding more years than those used in this study (only two years of data were used for training) should be incorporated in the training process in order to develop a classifier fully capable of dealing with interannual variability. Ideally, after enough data from many years is incorporated into the training process, the need for more ground truth data will eventually stop or at least decrease (as models should always be updated with recent data, hence the "semi-automated" designation) once enough variability has already been accounted for. Even though the results in section 4.4 are not very good, the same strategy using more data would probably allow a more robust system to be developed, and therefore comparing the area of cashew plantations throughout the years in an automated way would be possible.

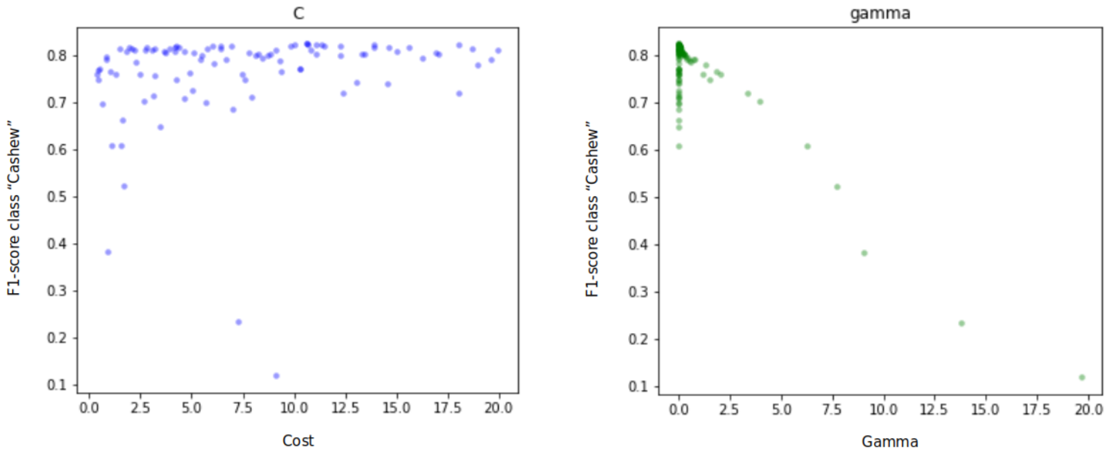
This study in a small region of the country serves as a viability test to prove that monitoring of this situation is in fact possible, contributing for the evaluation of the country's food, climate, and ecosystem sustainability outlook. It is also important to mention that the work developed in this dissertation can be adapted to other classes of land cover by adjusting some of it's components (variables, parameters or maybe even the algorithms).

5.2 Future Work

In the future, several improvements can be made. First, more ground truth data can be purchased (buying high resolution satellite images) or gathered (through field trips) and incorporated in the training process, in order to make the models more robust to interannual variability and therefore improve the system developed throughout section 4.4. Second, as mentioned in the background chapter, patch-based classification using deep learning is now commonly used and known for rendering very good results. With appropriate training data, this approach could be explored. Since this approach is known for taking advantage of the patterns present inside each patch, it would probably work very well for detecting cashew orchards. Another approach worth exploring to detect the orchards would be to use one vs all classifiers such as `scikit-learn`'s one class SVM or the `maxent` algorithm, one of the most common examples of a presence-background algorithm used in biology [43]. Unlike the current approach which works in a presence-absence framework that contains many classes, one class approaches work in a presence-background binary setup, where one class corresponds to the presence object of interest (cashew in this case) and another class that can contain everything else including cashew, hence the name background.

Appendix A

Supplementary Figures and Tables



(a)

(b)

Figure S1: Optimization of the Support Vector Machine's hyperparameters Cost (a) and gamma (b) from section 4.4.2

Table S1: Confusion Matrix corresponding to the performance on the test data regarding the algorithm from section 4.4.2 (2019), in which maps were not produced.

	Forest	Cashew	Sparse Veg.	Mangrove	Other	Water	**User's Accuracy**
Forest	2373	90	16	0	9	1	95.34
Cashew	289	654	8	0	4	0	68.48
Sparse Veg.	41	5	512	0	45	0	84.91
Mangrove	2	0	0	724	1	6	98.77
Other	4	24	81	0	992	51	86.11
Water	0	0	0	1	79	1108	93.27
Producer's Accuracy	87.60	84.61	82.98	99.86	87.79	95.03	

Table S2: F1-score corresponding to the performance on the test data regarding the algorithm from section 4.4.2 (2019), in which maps were not produced.

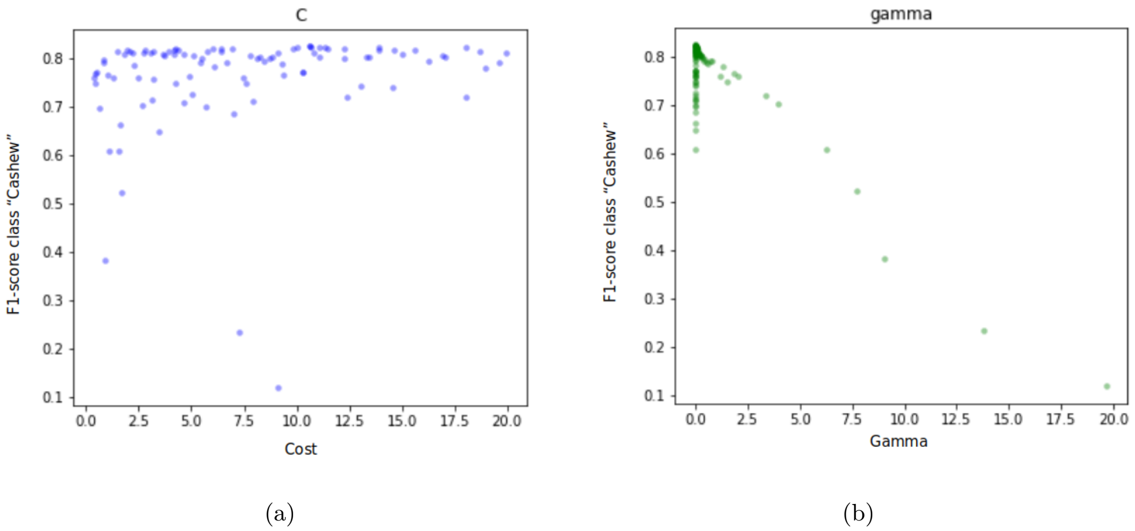
	f1-score	Support
Forest	91.30	2489
Cashew	75.69	955
Sparse Veg.	83.93	603
Mangrove	99.31	733
Other	86.94	1152
Water	94.14	1674
Accuracy		89.37
Balanced Accuracy		87.81

Table S3: Confusion Matrix corresponding to the performance on the test data regarding the algorithm from section 4.4.2 (2020), in which maps were not produced.

	Forest	Cashew	Sparse Veg.	Mangrove	Other	Water	**User's Accuracy**
Forest	2324	151	11	0	3	0	93.37
Cashew	316	634	0	0	5	0	66.39
Sparse Veg.	86	18	490	0	9	0	81.26
Mangrove	28	0	0	693	10	2	94.54
Other	36	32	332	4	701	47	60.85
Water	0	0	0	61	86	1041	87.63
Producer's Accuracy	83.30	75.93	58.82	91.42	86.12	95.50	

Table S4: F1-score corresponding to the performance on the test data regarding the algorithm from section 4.4.2 (2020), in which maps were not produced.

	f1-score	Support
Forest	88.05	2489
Cashew	70.84	955
Sparse Veg.	68.25	603
Mangrove	92.96	733
Other	71.31	1152
Water	91.40	1674
Accuracy		82.63
Balanced Accuracy		80.67



(a)

(b)

Figure S2: Optimization of the Support Vector Machine's hyperparameters Cost [a] and gamma [b] from section 4.4.3

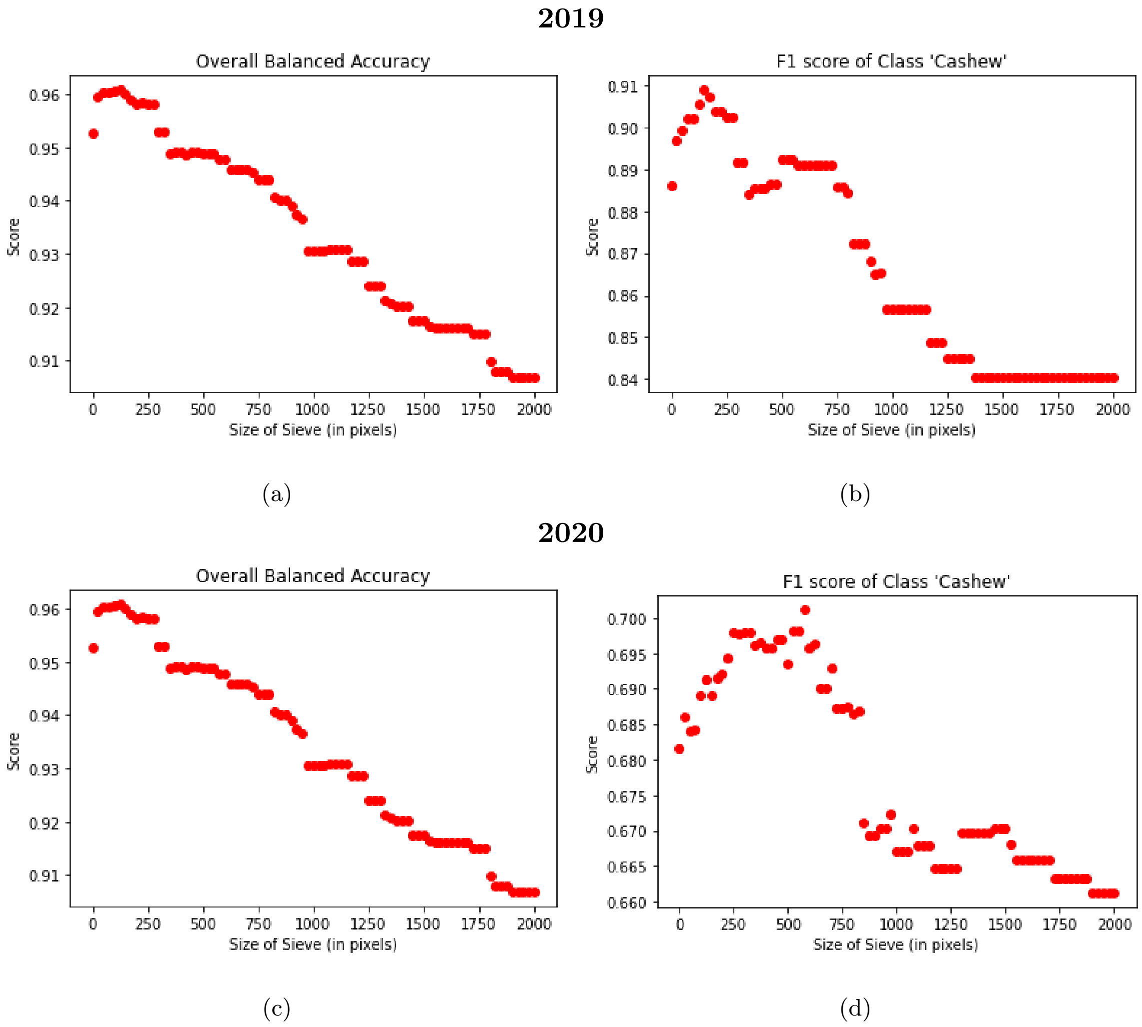


Figure S3: Balanced accuracy [a and c] and f1-score of class "Cashew" [b and d] for the tested sieve thresholds in both 2019 [a-b] and 2020 [c-d].

Bibliography

[1] "UNFCCC - United Nations Framework Convention on Climate Change | Knowledge for policy." [Online]. Available: https://ec.europa.eu/knowledge4policy/organisation/unfccc-united-nations-framework-convention-climate-change{_}en

[2] "Food and Agriculture Organization of the United Nations." [Online]. Available: http://www.fao.org/home/en/

[3] "Sentinel-2 - Missions - Sentinel Online." [Online]. Available: https://sentinel.esa.int/web/sentinel/missions/sentinel-2

[4] "Sustainable Development Goals .:. Sustainable Development Knowledge Platform." [Online]. Available: https://sustainabledevelopment.un.org/?menu=1300

[5] "Forests, desertification and biodiversity United Nations Sustainable Development." [Online]. Available: https://www.un.org/sustainabledevelopment/biodiversity/

[6] "Zero Hunger United Nations Sustainable Development." [Online]. Available: https://www.un.org/sustainabledevelopment/hunger/

[7] "Climate Change United Nations Sustainable Development." [Online]. Available: https://www.un.org/sustainabledevelopment/climate-change/

[8] "Sustainable consumption and production United Nations Sustainable Development." [Online]. Available: https://www.un.org/sustainabledevelopment/sustainable-consumption-production/

[9] "Land Use, Land Cover, and Trends in Guinea-Bissau | West Africa." [Online]. Available: https://eros.usgs.gov/westafrica/land-cover/land-use-land-cover-and-trends-guinea-bissau

[10] "WFD: Guinea-Bissau forests still at risk despite good protection system | UNIOGBIS." [Online]. Available: https://uniogbis.unmissions.org/en/wfd-guinea-bissau-forests-still-risk-despite-good-protection-system

[11] M. J. Vasconcelos, A. I. Cabral, J. B. Melo, T. R. Pearson, H. d. A. Pereira, V. Cassamá, and T. Yudelman, "Can blue carbon contribute to clean development in West-Africa? The case of

Guinea-Bissau," *Mitigation and Adaptation Strategies for Global Change*, vol. 20, no. 8, pp. 1361–1383, 2015.

[12] The Prepublic of Guinea Bissau- The State's General Office of the Environment, "Strategy and National Action Plan for the Biodiversity - 2015-2020," Tech. Rep. July, 2015. [Online]. Available: https://www.cbd.int/doc/world/gw/gw-nbsap-v2-en.pdf

[13] L. Catarino, Y. Menezes, and R. Sardinha, "Cashew cultivation in Guinea-Bissau risks and challenges of the success of a cash crop," pp. 459–467, sep 2015.

[14] "Cashew nut central to Guinea-Bissau economy: a blessing or a curse? | UNIOGBIS." [Online]. Available: https://uniogbis.unmissions.org/en/cashew-nut-central-guinea-bissau-economy-blessing-or-curse

[15] "REDD+ -Reducing Emissions from Deforestation and Forest Degradationă|ăFood and Agriculture Organization of the United Nations." [Online]. Available: http://www.fao.org/redd/initiatives/un-redd/en/

[16] C. Lopes, A. Leite, and M. J. Vasconcelos, "Open-access cloud resources contribute to mainstream REDD+: The case of Mozambique," *Land Use Policy*, vol. 82, pp. 48–60, mar 2019.

[17] O. Justus, L. Kirimi, and M. Mathenge, "Effects of climate variability and change on agricultural production:the case of small scale farmers in kenya," 01 2016.

[18] J. Sousa, A. L. Luz, F. N. Sousa, M. Cassama, A. Dabo, F. Dafa, and M. Bivar Abrantes, "Cashew Orchards Conserve the Potential for Forest Recovery," *Agroecology and Sustainable Food Systems*, vol. 39, no. 2, pp. 134–154, feb 2015.

[19] J. R. Jensen, *Remote sensing of the environment: an earth resource perspective second edition*, 2014, vol. 1.

[20] "U.s. geological survey, 2016, landsatearth observation satellites (ver. 1.2, april 2020): U.s. geological survey fact sheet 20153081, 4 p.,." [Online]. Available: https://doi.org/10.3133/fs20153081

[21] "User Guides - Sentinel-2 MSI - Overview - Sentinel Online." [Online]. Available: https://sentinel.esa.int/web/sentinel/user-guides/sentinel-2-msi/overview

[22] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google earth engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, 2017. [Online]. Available: https://doi.org/10.1016/j.rse.2017.06.031

[23] GDAL/OGR contributors, *GDAL/OGR Geospatial Data Abstraction software Library*, Open Source Geospatial Foundation, 2020. [Online]. Available: https://gdal.org

[24] S. Gillies *et al.*, “Rasterio: geospatial raster i/o for Python programmers,” Mapbox, 2013–. [Online]. Available: https://github.com/mapbox/rasterio

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine Learning in Python ,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] QGIS Development Team, *QGIS Geographic Information System*, Open Source Geospatial Foundation, 2009. [Online]. Available: http://qgis.org

[27] F. Pirotti, F. Sunar, and M. Piragnolo, “Benchmark of machine learning methods for classification of a Sentinel-2 image,” *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, vol. 41, no. June, pp. 335–340, 2016.

[28] S. Talukdar, P. Singha, S. Mahato, Shahfahad, S. Pal, Y.-A. Liou, and A. Rahman, “Land-Use Land-Cover Classification by Machine Learning Classifiers for Satellite ObservationsA Review,” *Remote Sensing*, vol. 12, no. 7, p. 1135, apr 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/7/1135

[29] P. Helber, B. Bischke, A. Dengel, and D. Borth, “Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, jul 2019.

[30] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org

[31] P. T. Inc. (2015) Collaborative data science. Montreal, QC. [Online]. Available: https://plot.ly

[32] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, oct 2001. [Online]. Available: https://link.springer.com/article/10.1023/A:1010933404324

[33] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” p. 144152, 1992. [Online]. Available: https://doi.org/10.1145/130385.130401

[34] J. Bergstra, D. Yamins, and D. D. Cox, “Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures,” *30th International Conference on Machine Learning, ICML 2013*, no. PART 1, pp. 115–123, 2013.

[35] “HOME | RSeT.” [Online]. Available: https://www.rset.eu/

[36] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural Features for Image Classification,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. 3, no. 6, pp. 610–621, 1973.

[37] R. W. Conners, M. M. Trivedi, and C. A. Harlow, "Segmentation of a high-resolution urban scene using texture operators (Sunnyvale, California)." *Computer Vision, Graphics, & Image Processing*, vol. 25, no. 3, pp. 273–310, 1984.

[38] "Stratified Group k-Fold Cross-Validation | Kaggle." [Online]. Available: https://www.kaggle.com/jakubwasikowski/stratified-group-k-fold-cross-validation

[39] FAO, "Map accuracy assessment and area estimation : a practical guide," p. 69, 2016. [Online]. Available: http://www.fao.org/3/a-i5601e.pdf

[40] C. P. Giri, *Remote sensing of land use and land cover : principles and applications.* CRC Press, 2012.

[41] I. Olthof, C. Butson, and R. Fraser, "Signature extension through space for northern landcover classification: A comparison of radiometric correction methods," *Remote Sensing of Environment*, vol. 95, no. 3, pp. 290–302, apr 2005.

[42] C. B. Chittineni, "Signature extension in remote sensing," *Pattern Recognition*, vol. 12, no. 4, pp. 243–249, jan 1980.

[43] S. Phillips, "A Brief Tutorial on Maxent in Species Distribution Modeling for Educators and Practitioners," *Lessons in Conservation*, vol. 3, pp. 107–135, 2010. [Online]. Available: http://ncep.amnh.org/linc