U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

Monitoring Greenhouses with Satellite Images and Machine Learning

Pedro Miguel Pereira Cardoso

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

2022

MSc

2.º CICLO

FCUP
2022

U.PORTO

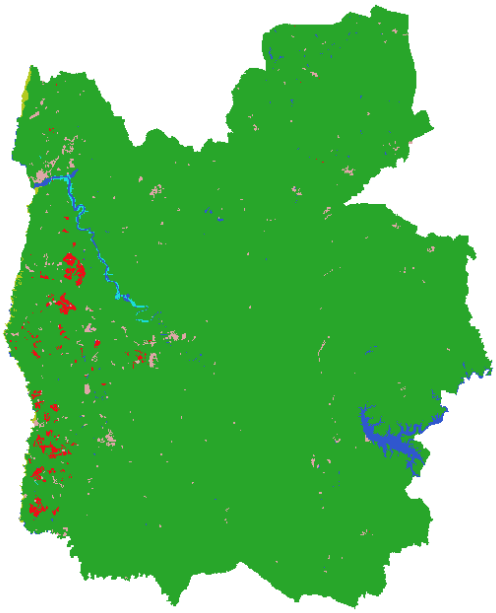
Monitoring Greenhouses with Satellite Images and Machine Learning

Pedro Miguel Pereira Cardoso

FC

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc



Monitoring Greenhouses with Satellite Images and Machine Learning

Pedro Miguel Pereira Cardoso
Engenharia de Redes e Sistemas Informáticos
Departamento de Ciência dos Computadores
2022

Orientador
João Pedro Pedroso, Professor Associado, FCUP

Coorientador
Manuel Campagnolo, Professor Associado, ISA/UL

Supervisora
Sofia Pereira, FEUP


U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Declaração de Honra

Eu, Pedro Miguel Pereira Cardoso, inscrito no Mestrado em Engenharia de Redes e Sistemas Informáticos da Faculdade de Ciências da Universidade do Porto declaro, nos termos do disposto na alínea a) do artigo 14.º do Código Ético de Conduta Académica da U.Porto, que o conteúdo da presente dissertação/relatório de estágio/projeto Monitoring Greenhouses with Satellite Images and Machine Learning reflete as perspetivas, o trabalho de investigação e as minhas interpretações no momento da sua entrega.

Ao entregar esta dissertação/relatório de estágio/projeto Monitoring Greenhouses with Satellite Images and Machine Learning, declaro, ainda, que a mesma é resultado do meu próprio trabalho de investigação e contém contributos que não foram utilizados previamente noutros trabalhos apresentados a esta ou outra instituição.

Mais declaro que todas as referências a outros autores respeitam escrupulosamente as regras da atribuição, encontrando-se devidamente citadas no corpo do texto e identificadas na secção de referências bibliográficas. Não são divulgados na presente dissertação/relatório de estágio/projeto Monitoring Greenhouses with Satellite Images and Machine Learning quaisquer conteúdos cuja reprodução esteja vedada por direitos de autor.

Tenho consciência de que a prática de plágio e auto-plágio constitui um ilícito académico.

Pedro Miguel Pereira Cardoso

Porto, 30/09/2022

Abstract

Agriculture has changed dramatically in the last few decades with the expansion of the use of plastic agricultural greenhouses. As this type of structures is low cost, compared to traditional greenhouses made of glass, their adoption by farmers was incredibly fast. This form of cultivation presents much higher levels of productivity in relation to traditional plantations and allows the growth of crops that are not native to the region, through the controlled simulation of an ideal habitat. Despite all these advantages, there is a need to monitor the expansion of areas occupied by plastic greenhouses, especially in areas with current or future water problems.This is the case of a vast area of Portugal, which will be used as a case study, where the risk of desertification and water scarcity is a possibility. This dissertation intends, through remote sensing and machine learning techniques, to develop tools and methods capable of identifying greenhouses and controlling their evolution. Throughout this work, promising results were obtained, allowing us to conclude that the generated system is able to classify plastic greenhouses in satellite images throughout the years. The created tool can help specialists in decision making tasks by providing a cheap and quick way of monitoring plastic greenhouses.

Resumo

A agricultura alterou-se drasticamente nas últimas décadas com a expansão do uso de estufas agrícolas de plástico. Como este tipo de estruturas é de baixo custo, em comparação com as estufas tradicionais feitas de vidro, permitiu que sua adoção por agricultores fosse incrivelmente rápida. Esta forma de cultivo apresenta níveis de produtividade muito maiores em relação a plantações ao ar livre e a possibilidade de cultivo de culturas não nativas à região através da simulação controlada de um habitat ideial. Apesar de todas estas vantagens há a necessidade de monitorizar a expansão das áreas ocupadas por estufas de plástico, principalmente em zonas com problemas hídricos atuais ou futuros, como é o caso de uma vasta área do território de Portugal, que será usado como caso de estudo, onde o risco de desertificação e escassez de água é uma possibilidade. Esta dissertação pretende através de técnicas de deteção remota e machine learning desenvolver ferramentas e métodos capazes de identificar estufas através de imagens de satélite e controlar a sua evolução. Ao longo deste trabalho foram obtidos resultados promissores que nos permitem concluir que o sistema gerado consegue classificar estufas de plástico em imagens de satélite ao longo dos anos. A ferramenta criada pode ajudar os especialistas na tomada de decisão, fornecendo uma forma barata e rápida de monitorizar estufas plásticas.

Agradecimentos

Primeiramente quero agradecer ao meu orientador Prof. João Pedro Pedroso por me ter proporcionado a oportunidade de realizar esta tese sobre um tema que me cativou bastante e pela ajuda ao longo da mesma. Quero também agradecer ao coorientador Prof. Manuel Campagnolo pela enorme partilha de conhecimentos que permitiram uma melhoria constante ao longo do desenvolvimento deste trabalho. Também agradecer à minha supervisora Sofia Pereira pela disponibilidade para me auxiliar sempre que encontrava um obstáculo. Muito obrigado pela vossa ajuda, não podia ter pedido um grupo melhor de pessoas para me guiar.

Quero também agradecer a todos os meus amigos, ao grupo que me acompanhou ao longo destes anos de faculdade, ao Daniel e ao seu pânico com datas de entrega, ao Teixeira que têm o dom de teletransporte, ao Rodrigo e à sua palavra mítica, ao Duarte pela companhia a tomar cafés, ao Tiago pelos seus ensinamentos, ao Naldo pelo apoio constante e ao Augusto por toda a sua sabedoria.

Também agradecer à Irmandade, ao Miguel por me ensinar o que é um histocárdio, ao Chico que passa maior parte do tempo em Leiria, ao Lino por me perguntar se já vi Watchmen e ao António pelos memes.

Agradecimento especial à minha família (mãe, pai e Bia) por me aturarem ao longo destes anos e me apoiarem sempre.

Por fim, agradecer à Sara por ser incrível e inspiradora, que pela sua companhia faz com que o final desta fase da minha vida seja ainda mais especial.

Obrigado a todos

Dedico a todos que me ajudaram

Contents

List of Tables

List of Figures

Acronyms

API	Application Programming Interface
COS	Cartografia de Uso e Ocupação do Solo
DT	Decision Tree
ESA	European Space Agency
GEE	Google Earth Engine
MSI	Multi-spectral instrument
NDVI	Normalized Difference Vegetation Index
NIR	Near-Infrared
NUTS	Nomenclature des unités territoriales statistiques
OOB	Out-of-bag
PG	Plastic Greenhouse
PGI	Plastic Greenhouse Index
RF	Random Forest
RPGI	Retrogressive Plastic Greenhouse Index
S2	Sentinel-2
SDG	Sustainable Development Goals
SWIR	Short Wave Infrared Spectral Range
VNIR	Visible/Near Infrared

Chapter 1

Introduction

1.1 Monitoring greenhouses in Portugal

For Portugal, with a temperate climate and relatively low cost of operation compared to other Western European countries, growing crops in greenhouses can become a very profitable business. Despite being a small country, the country's territory has differentiating characteristics between regions. In terms of population, the North and Center present a higher population density than the South. In terms of climate, the South is warmer compared to other regions throughout the year. These characteristics allow us to have in perspective how greenhouses can be used in different areas of the Portuguese mainland. For example, in the North of the country, in regions with high population density, greenhouses are usually small structures due to the lack of available area and urban encroachment. This pattern highlights one of the advantages of using greenhouses, which is their small area of occupation compared to the same production requirements if carried out in the open [1]. This may lessen the deforestation of areas for agricultural purposes and also allows greenhouses to be closer to population centers, reducing transport costs.

However, problems can arise further south where high temperatures, relatively flat terrain, low population density and consecutively larger natural areas allow for the existence of huge farms where greenhouses occupy large areas. The underlying problems are immense: First, the water level of the dams in these areas is much lower than similar areas without greenhouses. Second, the lack of available workers in the local population leads to the use of migrants, which may lead to human rights violations, potentiated by illegal immigration and poor housing and working conditions for these seasonal workers. Finally, covers of greenhouses result in the waste of degraded plastic [2].

Current tools do not allow us to have a concrete notion of the extent of the greenhouse area. The last Cartografia de Uso e Ocupação do Solo (COS) carried out in Portugal, a cartography map with a defined minimum unit (1 ha) based on photo interpretation and with 80 classes, does not contain a class for agricultural greenhouses, which limits the available ground truth data [3]. Furthermore, the COS is updated every 3 years at best, which is insufficient from a managing

perspective, therefore, the need to obtain this information through remote sensing arises.

The Copernicus Sentinel-2 mission launched by European Space Agency (ESA) consists of the use of satellites to monitor the Earth's surface, obtaining new multispectral data approximately every 5 days. This information is made freely available, and it is through it that we will obtain the data to feed a machine learning model. In this way, it will be possible to simply automate the process of identifying greenhouses and provide information in digital form that would otherwise have to be obtained manually and at higher costs.

1.2 Objectives

The main objective of this work is to develop, through machine learning, a tool that allows the classification of each pixel of an image into land use classes, while making the classifier flexible enough to work with the data of years after or before the year in which it was trained, without compromising the solution quality.

In addition to the main objective, it is intended to explore methods that allow to implement this tool in an accessible and fast way in other geographic scenarios.

1.3 Chapters Organization

The organization of this thesis is as follows: In the present chapter (**chapter 1**), a brief introduction to the motivation and objectives of this work is made. **Chapter 2** discusses the necessity of this work and some concepts to comprehend the next chapters Then, **chapter 3** presents the current state of the art, previous methodologies used in similar works and how this dissertation diverges from them. **Chapter 4** represents the main body of this work and describes each task implemented in the methodology. **chapter 5** presents and discusses the results obtained following the methods of the previous chapter. Finally, **chapter 6** concludes this dissertation with a critic analysis of the work done and discuss improvements that can be done in the future.

Chapter 2

Background

In this chapter, we will present the fundamentals to interpret the main concepts of this dissertation. We start by explaining the motivations behind this work and the impact it can have in long term sustainability of the planet's resources. In the following section, the basic concepts of remote sensing, satellite imagery, spectral channels and land cover are presented. To close the chapter, information regarding the software used is provided.

2.1 United Nations : Sustainable Development Goals

Figure 2.1: United Nations - Sustainable Development Goals.

In 2015, the United Nations decided to set goals to create a more prosperous and sustainable future. These goals, called the Sustainable Development Goals (SDG), aim to ensure that by 2030 all people in the world can enjoy peace and progress [4]. There are a total of 17 interconnected SDGs that influence one another (represented in Figure 2.1). This dissertation intends to help in

the implementation, directly or indirectly, of 4 of them, in order to understand how they correlate with this dissertation a brief discussion about their impact in the study case of Portugal follows. The identified SDGs are:

- SDG No. 6 – **Clean Water and Sanitation**

  Portugal is in a precarious situation when it comes to water resources and desertification. Hotter and drier periods during the year have become increasingly prevalent and prolonged, threatening to become a constant [5]. In 2022, Portugal is experiencing the most severe drought in the last 1200 years, jeopardizing the availability of water for agriculture, tourism and other activities that use water intensively[6].

  Monitoring greenhouses is important because they have lower water consumption by kilogram of product when compared to conventional agriculture, as they are able to retain water that is lost by evapotranspiration in more conventional settings [1]. On the other hand, greenhouses rely on intensive practices that in the long term can have irreversible damages at an environmental, human and economic level.

  According to this SDG, by 2030, we must increase the efficient use of water in all sectors and ensure the sustainability of these water sources.

- SDG No. 8 – **Decent Work and Economic Growth**

  Greenhouses, despite being increasingly optimized, still require a significant amount of manpower. Considering that Portugal has an increasingly aging population and that working in greenhouses is not a well-paid job, there is an enormous difficulty in finding available workers in the local populations.

  One of the alternatives found by the producers was the recruitment of foreign workers from poor countries. Through an application for temporary residence these workers can stay in Portuguese territory as long as they wish as long as they have a contract with a company and pay taxes, with the final objective of obtaining the Portuguese passport after 5 years of work, this will allow them to work freely throughout the European Union. However, this dream of obtaining a European passport can quickly turn into a nightmare, with the emergence of temporary contract companies that take advantage of the economic situation of these migrants and, through promises of good working conditions, traffic and exploit them on Portuguese soil [7].

  With the threat of withdrawing their contracts and, consequently, migrants losing the right to stay in Portugal and fulfill the 5 years required to obtain a passport, they are subject to precarious situations with long working hours under high temperatures in addition to accumulating debts to traffickers who facilitated their entry into the European Union [8].

  The detection of greenhouses can be a useful tool to identify possible cases where there may be abuses, for example, if a certain company with a large area of greenhouses declares a number of workers much lower than expected, some type of labor exploitation may be taking place. Remote sensing of greenhouses makes it possible to more easily protect these

seasonal agricultural workers, who suffer from the lack of surveillance by responsible public institutions [9].

- SDG No. 12 – **Responsible Consumption and Production**

  The objective of this SDG is to achieve sustainable economic growth without compromising standards that allow for sustainable production and consumption of resources. Plastic greenhouses provide the necessary capacities for a more sustainable production of food, being a closed environment where it is possible to have a more efficient agricultural production and consequently to use more effectively available natural resources like irrigation water. However, the reverse situation can also happen if certain measures are not adopted. Some issues of concern are the use and management of chemical products and recycling waste resulting from these productions, more specifically the huge amounts of plastic used in the cover of greenhouses. Moreover, greenhouse production requires more energy per unit of product, and therefore should be limited to environments where renewable energy is readily available [1].

- SDG No. 15 – **Life on Land**

  With biodiversity at risk and the loss of hectares of forest every year, it is necessary to act in order to preserve natural areas that are vital to combating environmental change. Once again, greenhouses can be a good ally or enemy depending on their use. If properly used, greenhouses can easily be located in urban areas or occupy smaller areas than if the same production were outdoors, allowing natural areas to be conserved.

  However, if there is no monitoring of the occupied areas, producers may uncontrollably expand their greenhouse area to zones previously occupied by natural habitats, ignoring the environmental damage caused. Thus, satellite images are a viable tool to detect this type of situations and alert the competent authorities.

2.2 Remote Sensing

Remote Sensing is the action of obtaining information about a particular object or area without physical contact with the target. For remote sensing to be possible, it is necessary that there is a body that emits electromagnetic radiation. For so-called optical remote sensing, which is the type considered in this dissertation, the source of radiation is the sun, whose light's reflection over the atmosphere and Earth's surface can be captured by sensors aboard Earth observation satellites [10]. The data collected by those sensors can later be used for monitoring geographic areas (Figure 2.2).

2.2.1 Copernicus Sentinel-2 Mission

The Copernicus Sentinel-2 mission consists of the use of two satellites in polar orbits, phased at 180º to each other, in order to monitor changes in the Earth's surface. With a frequency of

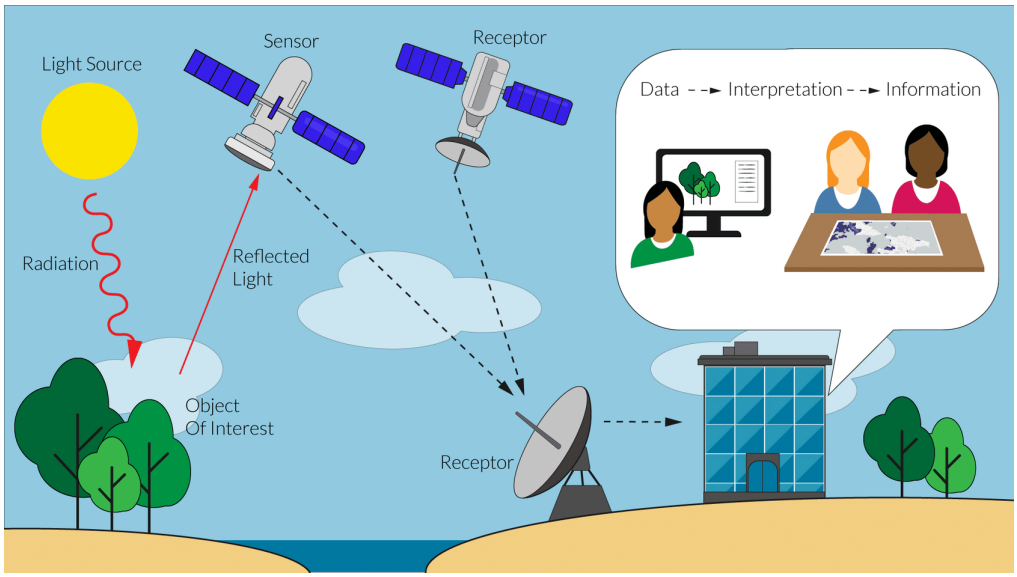


Figure 2.2: Remote sensing. Image from © The Nature Conservancy.

10 days at the equator, the use of two satellites makes it possible to revisit the same location every 5 days; this relatively small-time interval is an essential advantage in detecting short-term changes [11].

Each satellite has a Multi-spectral instrument (MSI) sensor that captures up to 13 spectral channels in the Visible/Near Infrared (VNIR) and Short Wave Infrared Spectral Range (SWIR).

Sentinel-2 Bands	Central Wavelength (µm)	Resolution (m)
Band 1 - Coastal aerosol	0.443	60
Band 2 - Blue	0.490	10
Band 3 - Green	0.560	10
Band 4 - Red	0.665	10
Band 5 - Vegetation Red Edge	0.705	20
Band 6 - Vegetation Red Edge	0.740	20
Band 7 - Vegetation Red Edge	0.783	20
Band 8 - NIR	0.842	10
Band 8A - Vegetation Red Edge	0.865	20
Band 9 - Water vapour	0.945	60
Band 10 - SWIR - Cirrus	1.375	60
Band 11 - SWIR	1.610	20
Band 12 - SWIR	2.190	20

Figure 2.3: Spectral bands and resolutions of Sentinel-2 MSI sensor. Taken from : https://www.researchgate.net/figure/Spectral-bands-and-resolutions-of-Sentinel-2-MSI-sensor_tbl3_325209585.

To obtain better results, we should aim to acquire the most correct and precise data possible. In the case of satellite images, a high quantity of pixels per area correlates to a bigger amount of data to collect, which leads to a more realistic portrayal of the region of interest. By analyzing Figure 2.3 we are able to observe that not all bands have the same resolution: the

ones with smaller resolutions should be preferred over the rest, as long as their wavelength is adequate for the problem at hand. Therefore, we choose as initial variables the band with the greatest spatial resolution, which are band 2 (blue), band 3 (green), band 4 (red) and band 8 (Near-Infrared (NIR)).

In addition to having the lowest spatial resolution among all the bands (10 meters), these bands are very useful for remote sensing of agriculture areas because they can be combined among themselves to produced indices that can also be added as variables.

2.2.2 Spectral Indices

In order to obtain better results, three new features that highlight or aid in the detection of plastic greenhouses were added by combining the previous four selected features.

- NDVI

  Figure 2.4 represents how Normalized Difference Vegetation Index (NDVI) measures the "greenness" of the land cover, which correlates to the amount of vegetation and its health, by calculating the normalized difference between NIR (heavily reflected by vegetation) and red light (absorbed by vegetation). Because of this, NDVI is one of the most commonly used indices used by analysts in remote sensing.

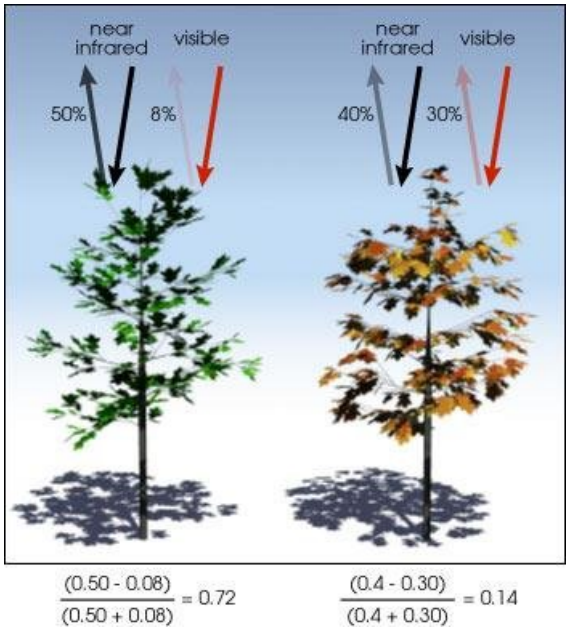


Figure 2.4: Behavior of radiation in vegetation. Taken from https://earthobservatory.nasa.gov/features/MeasuringVegetation/measuring_vegetation_2.php.

  NDVI values range between -1.0 and 1.0, where negative values are often water or clouds and values close to zero are rocks or bare soil. Positive values represent vegetation; while smaller positive values (like 0.2) can correspond to shrubs or meadows, high values (0.6 to

0.8) indicate healthy forests or croplands [12].

$$\text{NDVI} = \frac{\text{NIR} - \text{red}}{\text{NIR} + \text{red}} \tag{2.1}$$

- RPGI

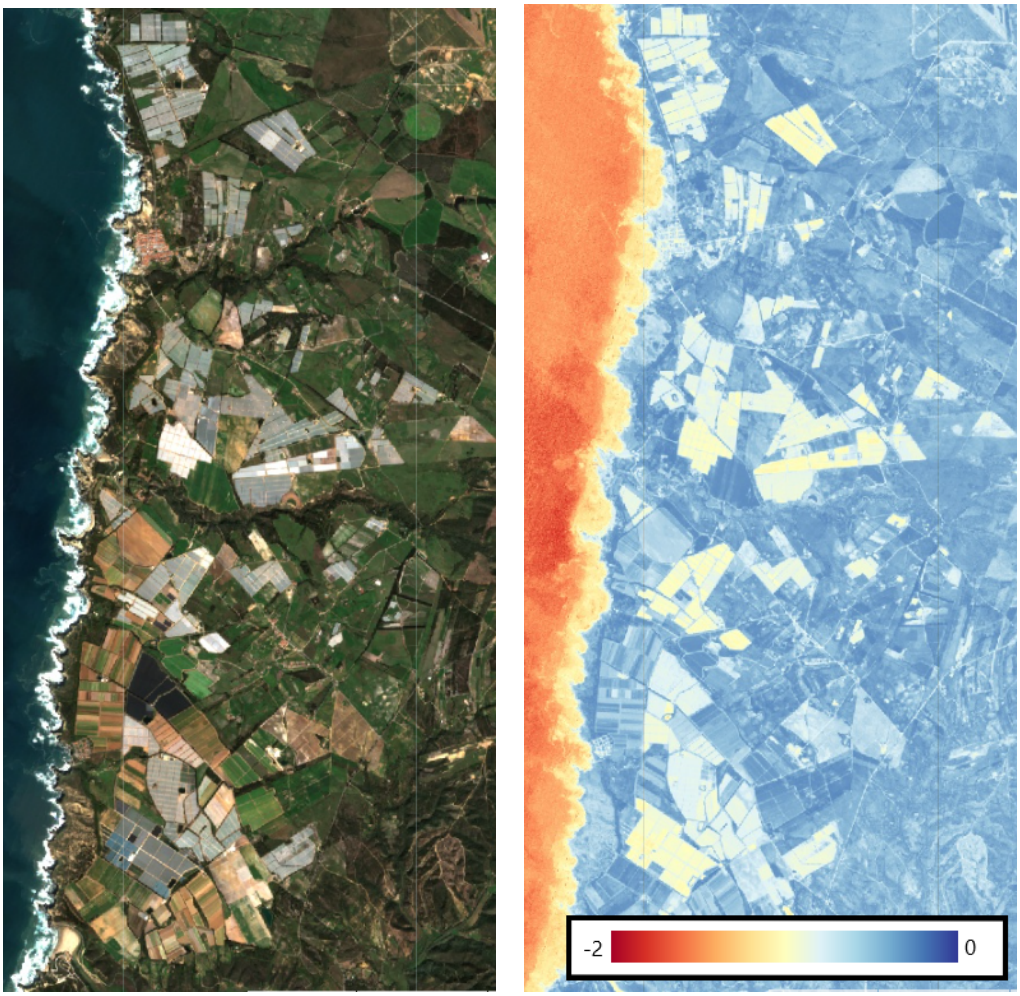
  Yang et al. [13] developed a new spectral index, Retrogressive Plastic Greenhouse Index (RPGI), with the goal of improving upon previous indexes used to detect plastic greenhouses. The new index enhances the spectral information of blue, green and NIR which are the B2, B3 and B8 bands of Sentinel-2 (S2) respectively. To calculate the values for each pixel, equation 2.2 is used.

$$\text{RPGI} = \frac{\text{blue}}{1 - \text{mean}(\text{blue} + \text{green} + \text{NIR})} \tag{2.2}$$

  The produced maps magnify the spectral values for the plastic greenhouse areas and help to discriminate them from other features, like croplands, bare soil, and man-made surfaces [13]. Figure 2.5 presents an example of how plastic greenhouses are highlighted using the index with the yellow color in image (b) of the Figure. The resulting values in this region varied from -2 to 0, with the smaller values being water bodies and the higher values representing other land surfaces that were not plastic greenhouses.

  Although RPGI highlights plastic greenhouses in most scenarios, it has some issues. In Figure 2.6 we have three black circles that represent different areas that have similar values of RPGI, we may notice that greenhouses and shallow bodies of water like rivers and coastlines don't differ much from each other in these maps.

  The other issue is the fact that this index does not perform well in summer months when greenhouses have a white cover instead of the usual dark tone. Because of these factors, RPGI cannot be deployed alone to detect greenhouses and other bands have to be utilized to help us surpass the mentioned restrictions.



(a) True color map (b) RPGI map

Figure 2.5: Maps of an area in the Alentejo Litoral Region; (a) True color image in February 2020 (b) RPGI: yellow objects on land are greenhouses.

(a) True color map (b) RPGI map

Figure 2.6: Maps showing a limitation of RPGI; (a) True color map; (b) RPGI map where the left circle indicates a coastline, the central circle indicates a greenhouse and the right circle indicates a water reservoir.

- Albedo

  With the RPGI, greenhouses often occupied an intermediate position in the range of values between water bodies (low values) and dense vegetation (high values). Problems can occur in regions where a transition between soil and water takes place, this is the case of coastal areas where waves clash against sand or rocks. In these conditions, the RPGI values are very similar to those of greenhouses. Due to the fact that most greenhouses in Portugal were located near the coast, this was an issue that need to be fixed. Therefore, we have considered a new variable, which estimates the surface albedo. In addition to reducing confusion with bright land cover, like coastal areas, albedo can also help to discriminate greenhouses from dark objects like roads or deep water.

  Albedo is the measure of the diffuse reflection of solar radiation out of the total solar radiation and measured on a scale from 0, corresponding to a black body that absorbs all incident radiation, to 1, corresponding to a body that reflects all incident radiation. The clash of sea waves, unlike greenhouses, often produced high reflectance and would have high values of albedo, as a result it would be possible to differentiate between them.

  To obtain albedo using S2 bands, we adapt the work done by Bonafoni and Sekertekin [14] but we only use the the 4 S2 10 m bands to estimate the albedo. As a result, and after re-scaling the coefficients, we obtain the weights of table 2.1.

  Table 2.1: Adapted Weighting Coefficients

Band	Weight
B2	0.2668
B3	0.1467
B4	0.1852
B8	0.4204

  Resulting in equation 2.3:

$$\text{Albedo} = \text{blue} \cdot 0.2668 + \text{green} \cdot 0.1467 + \text{red} \cdot 0.1852 + \text{NIR} \cdot 0.4204 \tag{2.3}$$

  We can observe in Figure 2.7 how this index performs by observing that coastlines in image [b] are far less distinguishable from greenhouses that in image [c] where the new albedo index is used. We can also notice that previous areas discussed previously in the RPGI section Figure 2.6 no longer are unrecognizable from greenhouses.

2.2.3 Land Cover and Land Use

It is also important to distinguish between land cover and land use, while land use documents how people are using the land or identify the land as, land cover uses the physical aspects of

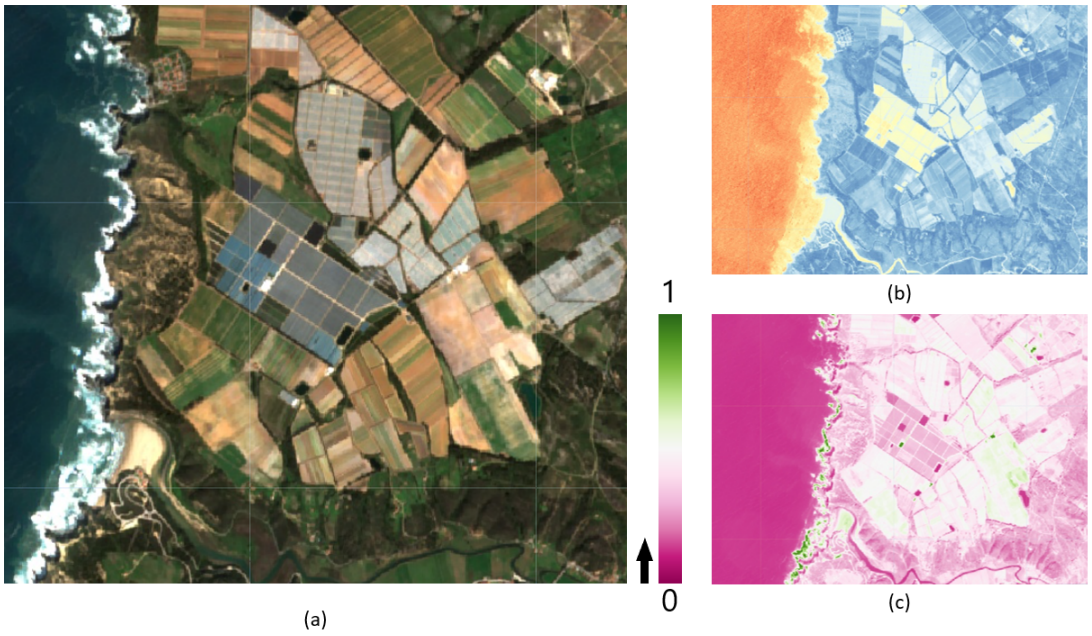


Figure 2.7: Maps showing the advantages of the Albedo; (a) Satellite Image; (b) RPGI map; (c) Albedo map that shows the zones indicated in Figure 2.6 being differentiated.

the land obtained using aerial imagery (captured by satellites in our case) to document how the region is really covered. A case scenario can be one particular area being document in land use as a forest, but the land cover indicates that now it is instead a low vegetation area, maybe because of a recent wildfire or deforestation.

For this work, we will produce land cover maps, but land use maps will still be useful because they allow us to filter out areas where plastic greenhouse occurrence is marginal while also providing us with ground truth data. For Portugal that map will be the Cartografia de Uso e Ocupação do Solo (COS).

2.2.3.1 COS - Cartografia de Uso e Ocupação do Solo

COS is a land use map created in and for Portugal with a minimum cartographic unit of 1 ha and a minimum distance between lines of 20 meters. There are five reference years for which these maps were produced (1995, 2007, 2010, 2015 and 2018). The COS nomenclature has undergone a reformulation for the production of COS2018. The new nomenclature now includes 83 classes, 35 more classes than the previous COS2015, with readjusted levels of disaggregation, however there is no specific class for greenhouses [3].

2.3 Software

2.3.1 Google Earth Engine

Most of the work carried out in this dissertation was produced in Google Earth Engine (GEE), a geospatial processing service powered by Google's cloud platform. GEE allows access to a huge array of open-source databases with satellite images in a quick and simple manner, that could be used in its integrated JavaScript-based development environment [15].

The GEE Application Programming Interface (API) was enough to do most of the tasks necessary for the implementation of this work, providing functions and classification algorithms that would be essential for the main body of this dissertation. However, there are several limitations that forced us to use other tools or software to obtain satisfactory results.

2.3.2 Colab

Colaboratory, or "Colab" for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing access free of charge to computing resources including GPUs [16].

Due to the fact that both Colab and GEE are a product of Google, official documented information about how to interconnect both is widely available. Colab proved to be important to overcome the limitations of GEE, using Python libraries such as Pandas, widely used for data science/data analysis and machine learning tasks, and TensorFlow that provides a collection of workflows to develop and train models resulting in an improvement of the overall work.

2.3.3 QGIS

QGIS is a free, open-source, cross-platform geographic information system (GIS) software that enables the visualization, edition, and analysis of geospatial data [17].

This application is helpful for tasks in which visualization of produced images is necessary, by being far quicker than GEE the user experience is much better, with the absence of delays when observing images. Most of the produced images present in this dissertation were captured using this tool.

Chapter 3

State of the art

Remote sensing is often used to obtain land use maps at different temporal rates and spatial resolutions. Previous studies showed potential in the identification of Plastic Greenhouse (PG) with Landsat satellite images. Yang et al. [13] designed a new spectral index, Plastic Greenhouse Index (PGI) and an improved version of it, Retrogressive Plastic Greenhouse Index (RPGI), specialized in highlighting and mapping PG using Landsat. However, a better alternative to Landsat imagery was required, since Landsat has at best a 30m resolution. This level of detail can be used for remote sensing medium/large scale greenhouse plantations, but it is not optimal for places like Portugal, where a more precise approach is required, since many greenhouses are scattered in peri-urban areas, are small in scale and spread out rather than concentrated in remote areas. Using Landstat-8, Hasituya et al. [18] showed that improvements in the performance for mapping plastic film covered farmland can be achieved using multi-temporal features rather than single temporal features. In 2021, Sun et al. [19] improves over previous works by using Sentinel-2 (S2) data with precise 10 m resolution imagery and also using two temporal S2 images to map plastic greenhouses in China, gathering more and better data, because of S2's better resolution and smaller revisit time of a given area (10 days) compared to Landsat 8 (16 days).

In this work, we further investigate the use of multi-temporal features and S2 band combinations for mapping small and large scale PG in Portugal, while also create the possibility of a year-to-year basis classification without the need of future updates to the classifier and provide a flexible framework able to be adapted for other locations.

Chapter 4

Methods

In this chapter, the methodology of this work will be described and each step of the resolution of the problem and adopted solutions will be discussed. We start by presenting a diagram that will delineate the main stages of this work and how they interact with each other. Then, by following the flow of the diagram, each topic will be discussed.

The workflow implemented is summarized in diagram 4.1 that divides the overall work in two main steps. In (step 1) we do the data preparation, first by processing the available Sentinel-2 (S2) collected data and merging it with the Cartografia de Uso e Ocupação do Solo (COS) 2018 map information to remove data related to artificial landscape, bare soil and bodies of water. Then we do the data curation by structuring and integrating the collected data (this is where the indexes first show up) and we finalize this step by selecting the features to be used.

In Step(2) or classification phase, we train the classifier by using the reference data collected by photo interpretation combined with the data of Step(1). After training, we analyze the feature importance, remove redundant features and apply the Sieve algorithm to obtain the final result (the greenhouse map) that can be overlapped with reference data for accuracy assessment and, finally, inference.

The resulting product is a plastic greenhouse detection map for each of the Portuguese regions defined in Nomenclature des unités territoriales statistiques (NUTS) III, except for the classes of COS designated above, i.e. artificial landscape, bare soil and bodies of water.

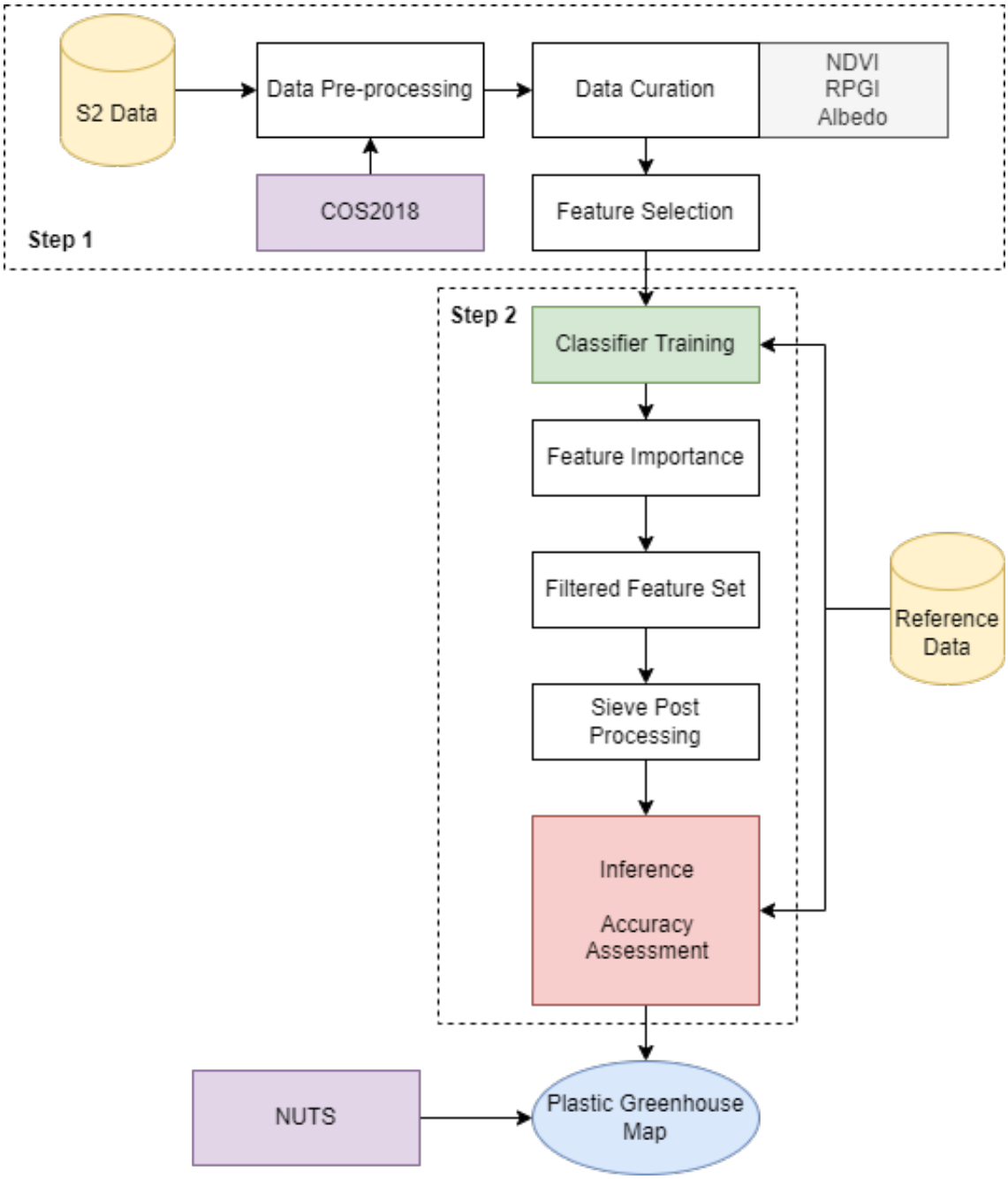


Figure 4.1: Workflow of this work.

4.1 Ground Truth Data Collection

In order to train the classifier and test its accuracy, it is necessary to manually obtain and label the collected data. However, due to the absence of existing ground truth data gathered specifically for greenhouses, which would be the optimal option, and the fact that the COS has an area unit of 1 ha, which is not good enough to identify greenhouses, all data collection and identification was done by photo interpretation of high resolution images available in Google Earth Pro. Due to the high resolution of Google Earth Pro, greenhouses can be identified with very high accuracy [20].

Figure 4.2: Greenhouse in Alentejo Litoral Region Used for data collection.

Greenhouses can have very different spectral signatures. This is due to three main factors: geographic location, type of agricultural crop and type of cover (covering plastics with different colors or even being transparent). In addition, those types of physical characteristics are not constant over time, since a greenhouse does not need to produce the same agricultural crop throughout the year, or it can even change its cover because of weather conditions or because certain crops require extra coverage during parts of their growing cycle. Thus, there is wide variation of spectral and temporal patterns in greenhouses during intra- and inter-years periods.

However, the varying signal pattern of greenhouses can be beneficial to create a greenhouse detection tool, since the temporal pattern of the signal for greenhouses tend to be less stable than many land cover types with a similar spectral signature. For instance, artificial structures (non-greenhouses) typically have constant spectral signatures along the year, which allows us to distinguish plastic greenhouses from those at some point of the year (Figure 4.3 and 4.4).

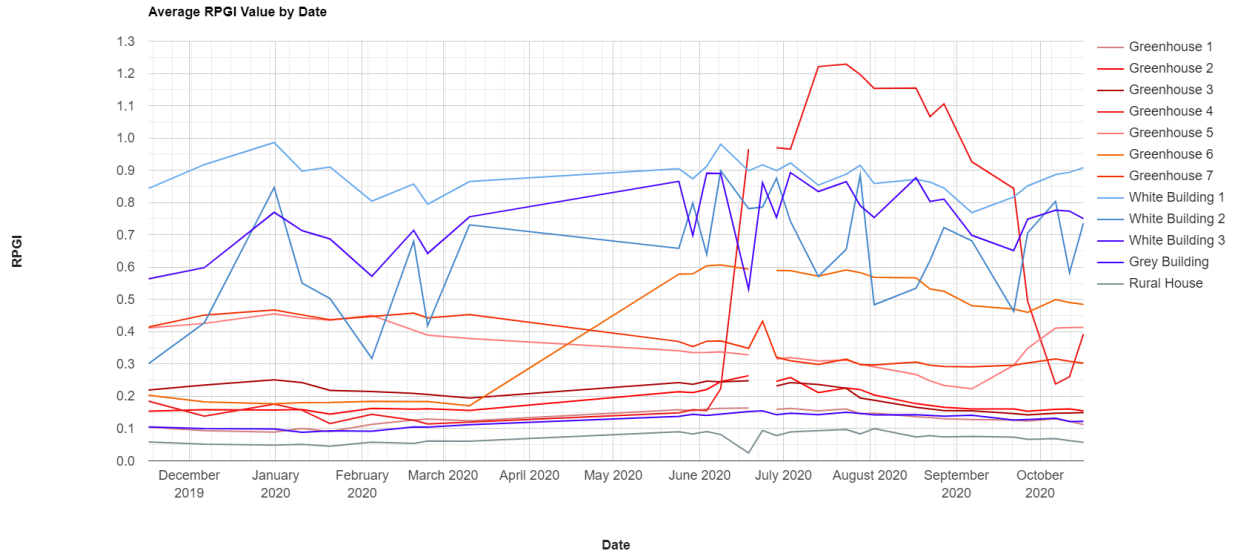


Figure 4.3: RPGI spectral signature.

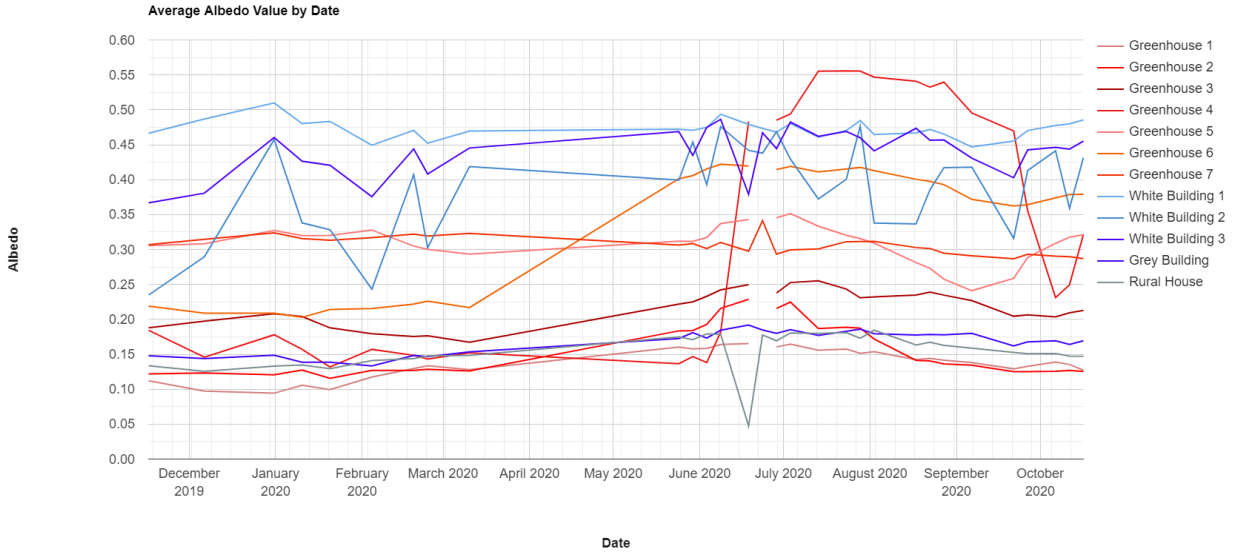


Figure 4.4: Albedo spectral signature.

Establishing the set of classes is a major component of the classification problem. For each class of interest, it might be useful to define sub-classes with homogeneous signatures. Since we haven't found clear clusters of greenhouse types, we considered a unique greenhouse class. However, it is useful to consider several non-greenhouse classes, even though we do not aim at mapping precisely those individual classes. The non-greenhouse classes we consider in this work are artificial (buildings, roads and parking lots), soil (agricultural land, forests, pastures and other natural landscapes) and water bodies. The distribution of collected data for the 4 classes can be observed in Figure 4.5.

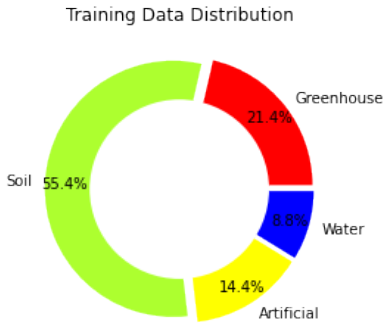


Figure 4.5: Distribution of collected training data.

To collect data, we digitized polygons that label the captured area in one of the four classes. Due to S2 satellite images having a resolution of 10 m, special care was taken to always leave a few meters between the polygon perimeter and the greenhouse's real perimeter when drawing the polygons, in order to avoid border pixels between greenhouse and non-greenhouse being misidentified in Sentinel-2 images. Examples are available in Figures 4.6 and 4.7.

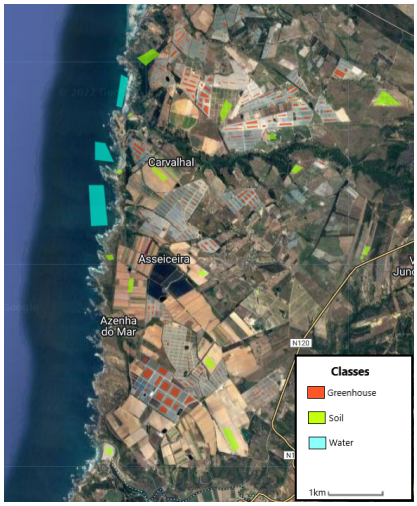


Figure 4.6: Collection of data in the Alentejo Litoral region.

To obtain a more diversified set of reference data, 5 areas with distinct characteristics were analyzed. Specifically, data was collected in the north of the Porto Metropolitan Area where small sized plastic greenhouses are dominant, in the Oeste region because of the occurrence of

Figure 4.7: Collection of data north of Porto Metropolitan Area.

large scale greenhouses existing next to suburban areas and in Alentejo Litoral and Algarve regions where the largest greenhouses complexes are located and temperatures are warmer. Data was also collected in the city of Lisbon, mainly for buildings and other artificial infrastructures 4.8.

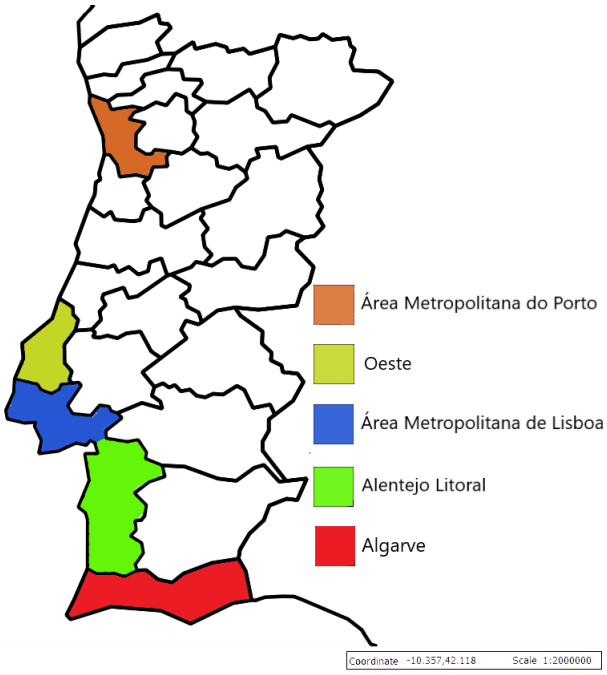


Figure 4.8: Regions of Portugal in each data was collected.

In addition, data collection was divided in each region in 2 sub-regions, one sub-region for the training data and other for the testing data (example in Figure 4.9), this way pixels from the same or close polygons wouldn't appear in both training and testing data, the classification would be less influenced and provide a more realistic accuracy assessment.

Each element of the reference set is labelled with the year when the occurrence is registered.

Figure 4.9: Division between training and testing region.

The year was identified with manual visualization of temporal sequences of S2 images, in addition to Google Earth Pro very high resolution imagery.

4.2 Satellite Data Collection

Through Google Earth Engine (GEE) it was possible to obtain S2 images from the period 2018 to 2021, the images are selected by choosing the median of the set of all images captured during a certain time span, usually one or two months.

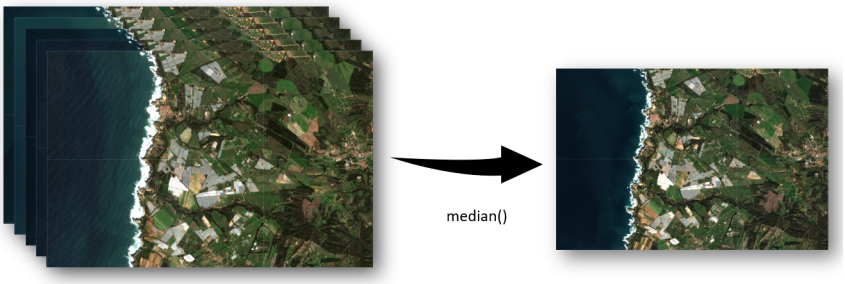
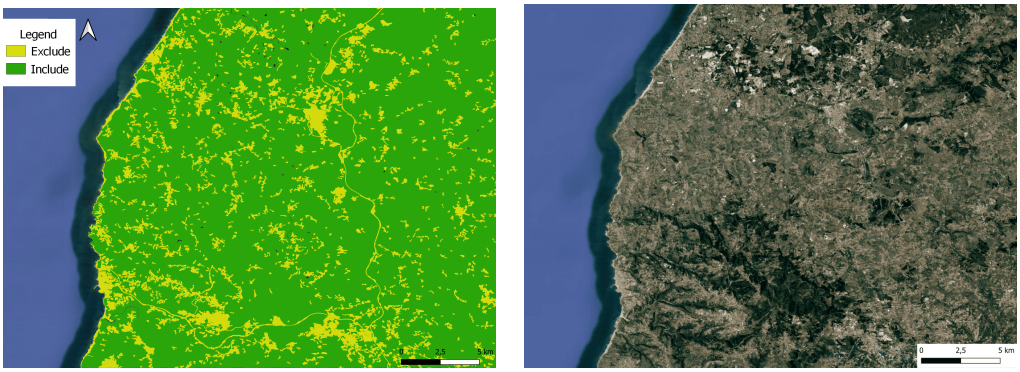


Figure 4.10: Representation of the selection of an image in an image collection.

As mentioned at the beginning of this chapter, the COS map was used to remove from consideration artificial landscape, bare soil and bodies of water, since those land use classes are not very relevant for the problem at hand (see example in Figure 4.11). By excluding these pixels, we are able to streamline the classification process by reducing the amount of data to classify while also avoiding misclassification of areas that were marginal for greenhouses.



(a) Areas to be excluded or included (b) Satellite image

Figure 4.11: Example of how COS is used to filter a region; Classification will only be made in the green areas.

4.2.1 Cloud Filter

One problem to address is the fact that not all the images captured by the satellites are viable, often the weather conditions are not ideal to obtain reliable data. If images were acquired on heavily cloudy days, the frames obtained contain no valuable information (getting information from clouds and not from the desired terrestrial surface) and will consume resources unnecessarily. Accordingly, to avoid obtaining incorrect data, GEE allows filtering the obtained set of images, removing images where the percentage of pixels that are identified as clouds exceeds a certain threshold.

For Portugal, during most part of the year the value of the threshold was able to be 20% or less, despite that for cloudier periods of the year such as the winter months, in order to be able to obtain at least one image, it was necessary to increase this value to 30%, bigger values could cause issues because the higher the threshold value is, the greater the loss of accuracy will be due to more collected data being incorrect.

4.2.2 Selecting temporal compositing periods

Instead of classifying areas based on a single temporal image and the respective variables (B2, B3, B4, B8, NDVI, RPGI, Albedo) obtained at a given time, we considered multiple images acquired at different times of the year and all their variables combined into a larger dataset, increasing the total of features used:

$$TotalFeatures = 7 \times Periods$$

For example, if we used an image per month during a year, we would obtain

$$7 \times 12$$

variables, this huge increase in available data would in theory improve the accuracy of the classifier.

However, we were not able to have observations for each month, due to lack of cloud-free images in certain regions of Portugal.

It was needed to decide which time interval was going to be considered for temporal compositing, which also depended on the agricultural calendar in Portugal. The agricultural year starts in November of year n-1 and ends in October of year n [21]. In addition, we divided the agricultural calendar in two seasons, the first season between November and April and the second between May and October.

After deciding on the initial and ending time of the 12-month period, the next step was testing different frequencies to obtain the best possible results. Initial 3 possibilities were evaluated, monthly, bimonthly and quarterly imagery, resulting in a total of 84, 42 and 28 features respectively. Exploratory results proved that a larger number of features (i.e. shorter compositing periods) produced a better performance, meaning that the monthly imagery was the best option. However, another factor needed to be taken in account which was related to the possibility of one period lacking available images due to them being corrupted by clouds, this problem was specially aggravated during the first season of the agricultural calendar (November-April) resulting in lack of data for these cloudier periods.

We tested two distinct options. The first one preserved the monthly frequency, but required an additional processing step to overcome the lack of data for some winter months over certain regions of Portugal. The second option used longer compositing periods to ensure that at least one clear observation per pixel was available for each period.

The first option was implemented by computing the average of the previous and following month to obtain an estimated value for the month missing data. However, images of areas in GEE can be the combination of multiple tiles, and the cloud filter tool tests each individual tile separately. As a result, if one tile passes the filter, the overall image is showed as valid. However, if another tile in the same composite fails the filtering stage the final image will have a gap but will still be considered as valid, resulting in an incomplete image. The other possibility would be masking every single pixel identified as cloud or shadow for each image. In addition to the increased computational requirements, this would not guarantee that missing values would not occur in the temporal composite, since some regions have a highly persistent cloud cover.

As a result, the first possible solution wasn't feasible, leading to the second option of extending the compositing interval for cloudy periods. During the sunnier months a monthly imagery was selected, while from November until April a bimonthly criterion was adopted (Nov-Dec, Jan-Feb, Mar-Apr), resulting in a total of 9 temporal compositing periods and 63 features. This option appeared to be the best compromise, maintaining the best possible performance without being unable to classified certain areas because of the lack of available data.

4.3 Algorithm

4.3.1 Random Forests

Random Forest (RF) is a machine learning algorithm developed and trademarked by Leo Breiman and Adele Cutler [22] inspired by earlier work by Zhang and Ma [23], which by combining the output of multiple Decision Tree (DT) achieve a single outcome. In our case, we will use this algorithm for a classification task, so the output will be the class selected by most trees and by doing this RF improves upon DT by avoiding overfitting of the training set.

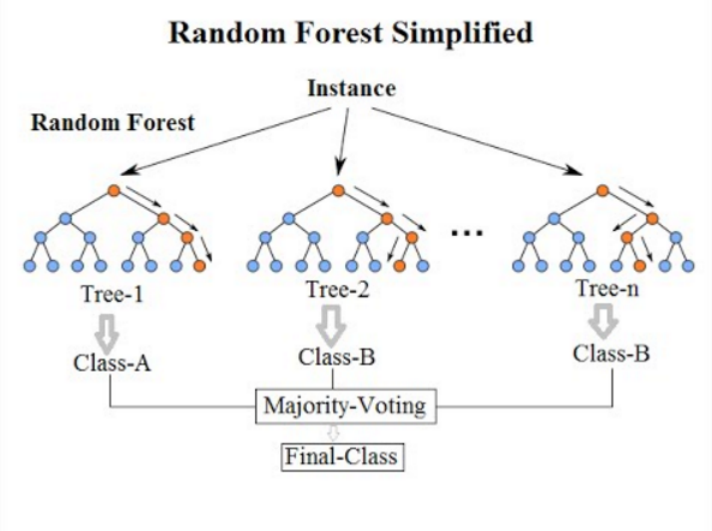


Figure 4.12: Random Forest Simplified. Taken from https://en.wikipedia.org/wiki/Random_forest.

RF can be appealing from a computational standpoint by being fast to train and to predict and depending on few tuning parameters. From a statistical perspective, RF is also appealing by having measures of variable importance.

4.3.1.1 Hyperparameters

One hyperparameter that can be configured in a RF classifier is the number of trees. This parameter that highly affects the overall performance of the classifier. Figure 4.13 shows that an early increase of trees, until 50, results in an exponential growth of Out-of-bag (OOB) accuracy, but its increase stabilizes after that. Due to modern computational capabilities permitting us to be less concerned about machine resources, we decided to use a number of trees larger than 150 trees.

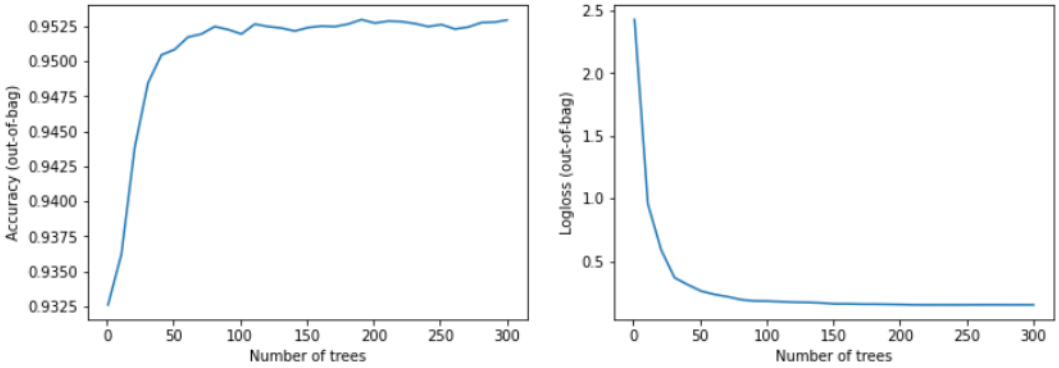


Figure 4.13: Effects on performance by number of trees in the random forest model.

4.4 Variable Selection

Using a larger number of variables can be penalizing since more data need to be stored, and the time spent in the classification process rises. In addition to this, sometimes new variables are not beneficial to the algorithm, by being redundant.

We need to measure how important each variable is and remove those that have a negative impact. For that task we used Colab/Python and the TensorFlow framework, by repeating the classifier training and using the concept of Mean Decrease in Accuracy, which is basically a method of computing the feature importance on permuted OOB samples to remove counterproductive features. After the removal, we repeated the process because the value of mean decrease in accuracy will be affected by having a new set of features and new unfavorable features may be detected. We kept repeating this task until either all features had a positive impact in the classifier or the removal of the latest set of features did not translate in an increase of accuracy [24].

Table 4.1: Original Feature Set

Nov/Dec	Jan/Feb	Mar/Apr	May	June	July	August	September	October
B2	B2	B2	B2	B2	B2	B2	B2	B2
B3	B3	B3	B3	B3	B3	B3	B3	B3
B4	B4	B4	B4	B4	B4	B4	B4	B4
B8	B8	B8	B8	B8	B8	B8	B8	B8
NDVI	NDVI	NDVI	NDVI	NDVI	NDVI	NDVI	NDVI	NDVI
RPGI	RPGI	RPGI	RPGI	RPGI	RPGI	RPGI	RPGI	RPGI
Albedo	Albedo	Albedo	Albedo	Albedo	Albedo	Albedo	Albedo	Albedo

Variable Importance: MEAN DECREASE IN ACCURACY

1. rpgi_1st : 0.000391
2. rpgi_2nd : 0.000301
3. B2_jul : 0.000211
4. B2_1st : 0.000180
5. ndvi_may : 0.000180
6. rpgi_sep : 0.000150
7. B2_3rd : 0.000150
8. B2_aug : 0.000150
9. B2_sep : 0.000120
10. rpgi_oct : 0.000120
11. rpgi_jul : 0.000120
12. B8_jul : 0.000090
13. B8_sep : 0.000090
14. ndvi_1st : 0.000090
15. B4_sep : 0.000090
16. alb_jul : 0.000090
17. ndvi_aug : 0.000090
18. ndvi_jul : 0.000090
19. B4_aug : 0.000090
20. ndvi_oct : 0.000090
21. B3_sep : 0.000090
22. rpgi_jun : 0.000090
23. B2_2nd : 0.000090
24. B2_may : 0.000090
25. ndvi_jun : 0.000060
26. rpgi_3rd : 0.000060
27. alb_sep : 0.000060
28. ndvi_3rd : 0.000060
29. alb_aug : 0.000060
30. B8_2nd : 0.000060
31. B3_1st : 0.000060
32. B4_may : 0.000060
33. B8_oct : 0.000030
34. B2_oct : 0.000030
35. B3_jul : 0.000030
36. B3_may : 0.000030
37. rpgi_aug : 0.000030
38. ndvi_sep : 0.000030
39. ndvi_2nd : 0.000030
40. B4_jun : 0.000030
41. alb_jun : 0.000030
42. B4_3rd : 0.000030
43. B8_may : 0.000000
44. B3_jun : 0.000000
45. rpgi_may : 0.000000
46. B8_jun : 0.000000
47. B3_oct : 0.000000
48. B8_aug : 0.000000
49. B8_3rd : 0.000000
50. B8_1st : 0.000000
51. B3_2nd : 0.000000
52. B4_jul : 0.000000
53. alb_oct : 0.000000
54. B4_oct : 0.000000
55. alb_3rd : 0.000000
56. B4_1st : 0.000000
57. alb_2nd : 0.000000
58. alb_1st : 0.000000
59. **B2_jun : -0.000030**
60. **B4_2nd : -0.000030**
61. **B3_3rd : -0.000030**
62. **alb_may : -0.000030**
63. **B3_aug : -0.000060**

Table 4.2: Filtered Feature Set

Nov/Dec	Jan/Feb	Mar/Apr	May	June	July	August	September	October
B2	B2	B2	B2		B2	B2	B2	B2
B3	B3		B3	B3	B3		B3	B3
B4		B4	B4	B4	B4	B4	B4	B4
B8	B8	B8	B8	B8	B8	B8	B8	B8
NDVI	NDVI	NDVI	NDVI	NDVI	NDVI	NDVI	NDVI	NDVI
RPGI	RPGI	RPGI	RPGI	RPGI	RPGI	RPGI	RPGI	RPGI
Albedo	Albedo	Albedo		Albedo	Albedo	Albedo	Albedo	Albedo

By observation of the list of feature importance and table 4.2 that represents the new set of features to be used, it is possible to conclude that no particular period, band or index was irrelevant and that almost all features are beneficial to the classifier. Nonetheless, we notice that the B2 band and the Retrogressive Plastic Greenhouse Index (RPGI) dominated the Top 10 in features importance.

4.5 Post processing

After the classification algorithm produces the map where each pixel is labelled in one of the four available classes, we reduced the numbers of classes to just 2, "greenhouses" and "non-greenhouses", by relabelling the classes of soil, artificial and water in the new class named non-greenhouse. The final map therefore represents a binary classification of the Portuguese territory except for the areas labelled by COS as artificial landscape, bare soil and bodies of water, as discussed at the beginning of this chapter. To be able to map the entire territory, we added to our binary map those areas that were initially removed.

4.5.1 Removing Noise with Sieve algorithm

Greenhouses are frequently relative large structures, which means that single pixels or small raster polygons classified as greenhouses in the produced map are likely to be noise rather than authentic greenhouses. In order to refine the produced image, a post-processing algorithm called Sieve was applied. Basically, the sieve algorithm identifies connected patches of pixels (an area composed of neighbor pixels of the same class in an eight connectedness pattern where pixels are connected horizontally, vertically, and diagonally). Then, those with a number of pixels smaller than a predefined threshold are relabeled with the label of the largest neighbor patch.

Since each pixel represents a 10 × 10 m square or the equivalent to 0.01 hectare, a small connected patch is typically irrelevant, either because it is just noise, or because it would

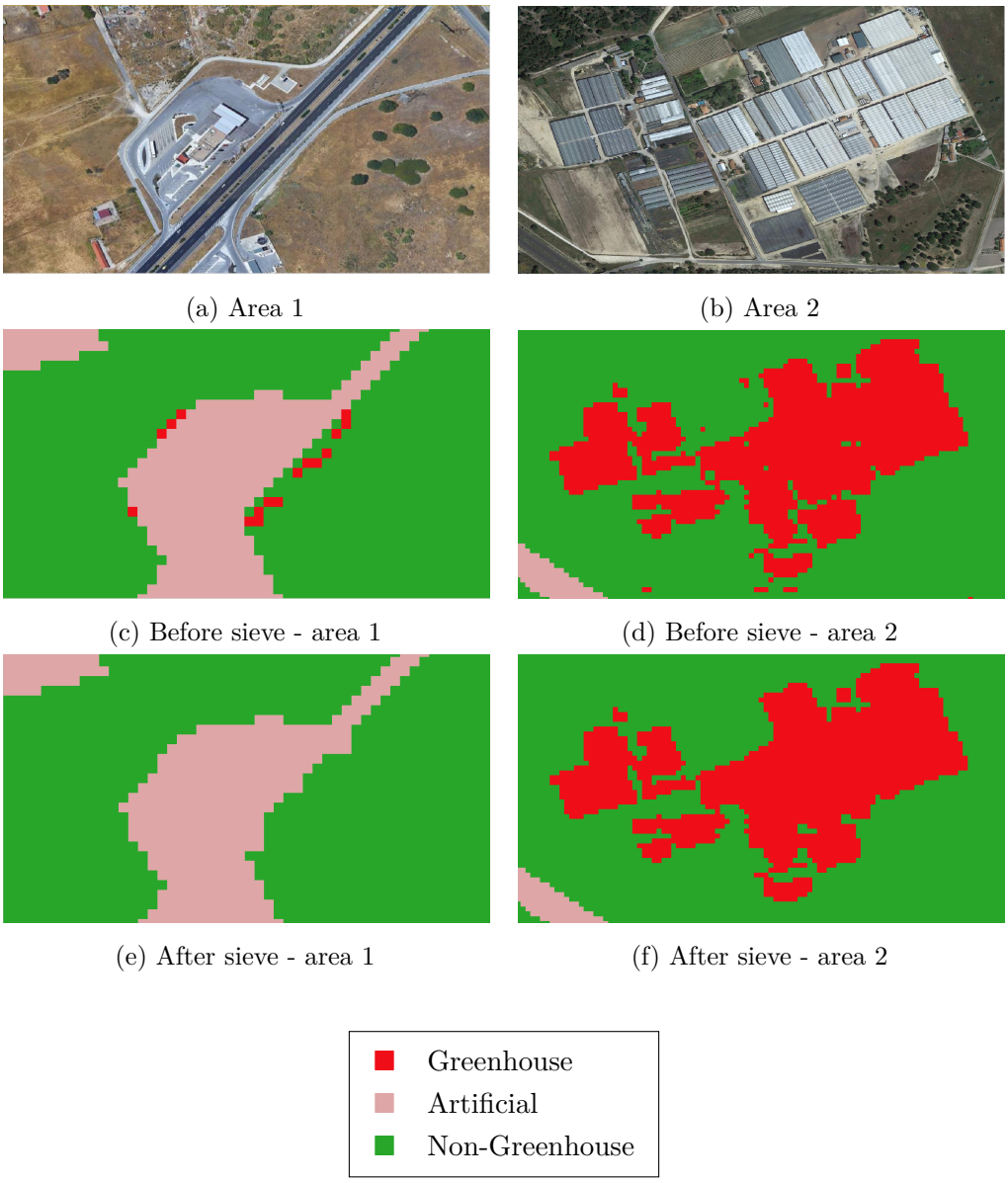


Figure 4.14: Maps of before and after sieve(6 pixels) in two different locations.

correspond to a very small structure. Figure 4.14 illustrates the removal of misclassified pixels that otherwise would stay in the final maps. Another advantage of applying the sieve post-processing algorithm is that it gives greenhouses more consistent shapes, as also illustrated in Figure 4.14 with the transformation of image (d) to image (f).

4.6 Accuracy Assessment

In Section 4.1, the procedure to collect reference data, also called "ground truth", was described. Each reference polygon belongs to either the "greenhouse" class or to one of the 3 "non-greenhouse" classes that were considered: artificial, soil (agricultural land, forests, pastures and other natural landscapes) and water bodies.

To train the classification algorithm, all 4 classes were used, since the map obtained after classification has those 4 labels. Moreover, only 40% of the reference polygons were used for training. The remaining 60% reference polygons were used to test the accuracy of the final map. All pixels from each reference polygon were only used either for test or training to avoid bias from spatial autocorrelation. Moreover, each of the 4 regions where reference polygons were collected (for both training and test) was split into two separate sub-regions, one for training and the other for test, to further reduce correlation between samples.

Since the final map, which is determined after (1) classification, (2) aggregation of the three non-greenhouse classes and (3) sieve post-processing, has only two classes (greenhouses and non-greenhouses), test polygons were separated into just those two classes for accuracy assessment.

For each test region, a contingency table was computed using all pixels from the test polygons, to compare the ground truth label (greenhouse or non-greenhouse) with the map label (also greenhouse or non-greenhouse). The resulting 2×2 accuracy matrices are used to estimate the accuracy of the binary map.

In order to do that, we computed the **producer's accuracy** (precision) and the **user's accuracy** (recall). The producer's accuracy specifies the percentage of pixels of the "greenhouse" class that were correctly classified by the map. The user's accuracy is the percentage of pixels mapped as "greenhouses" that in reality belong to that class [25].

Chapter 5

Results and discussion

In this chapter, we will present and discuss the produced maps, the performance of the classification model and the calculated plastic greenhouse areas in multiple regions.

5.1 Produced Maps

A more global result of the classifier's output can be seen in Figure 5.1, where the land cover map for the entire country is shown. This map is not ideal to have a precise view of plastic greenhouses, but allows the reader to grasp the scale and the main location of the areas occupied by the greenhouses. In Figure 5.2, small-scale examples of the resulting classification map are shown.

5.1.1 Regional Maps

To better visually understand the extension of the areas occupied by plastic greenhouses, it's important to focus on some regions of interest, such as Alentejo Litoral in Figure 5.3 and Oeste in Figure 5.4. The results also show how the areas evolved between the agricultural years of 2019 and 2021, with the creation of new plastic greenhouses and in rare cases their removal.

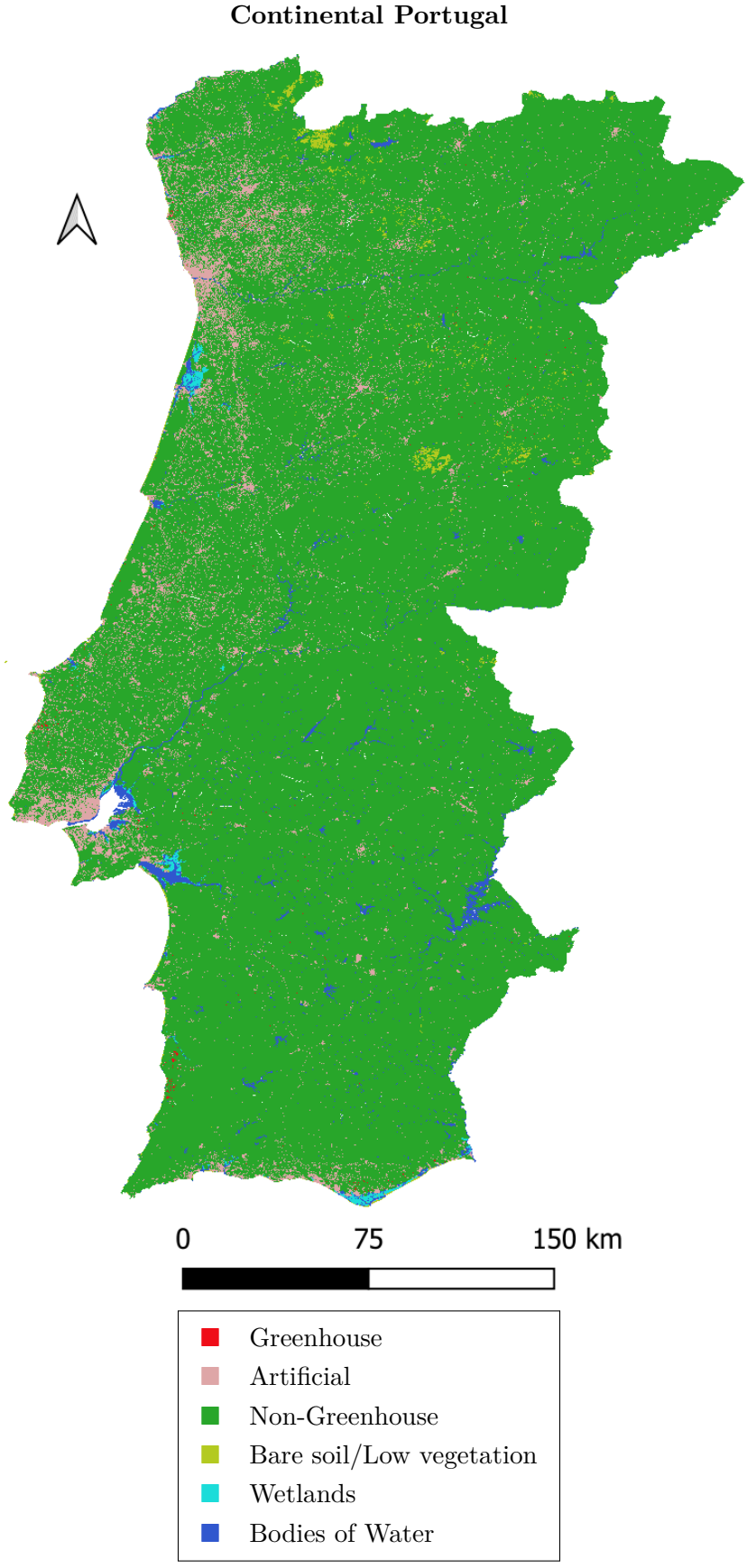


Figure 5.1: Produced map of continental Portugal in 2020.

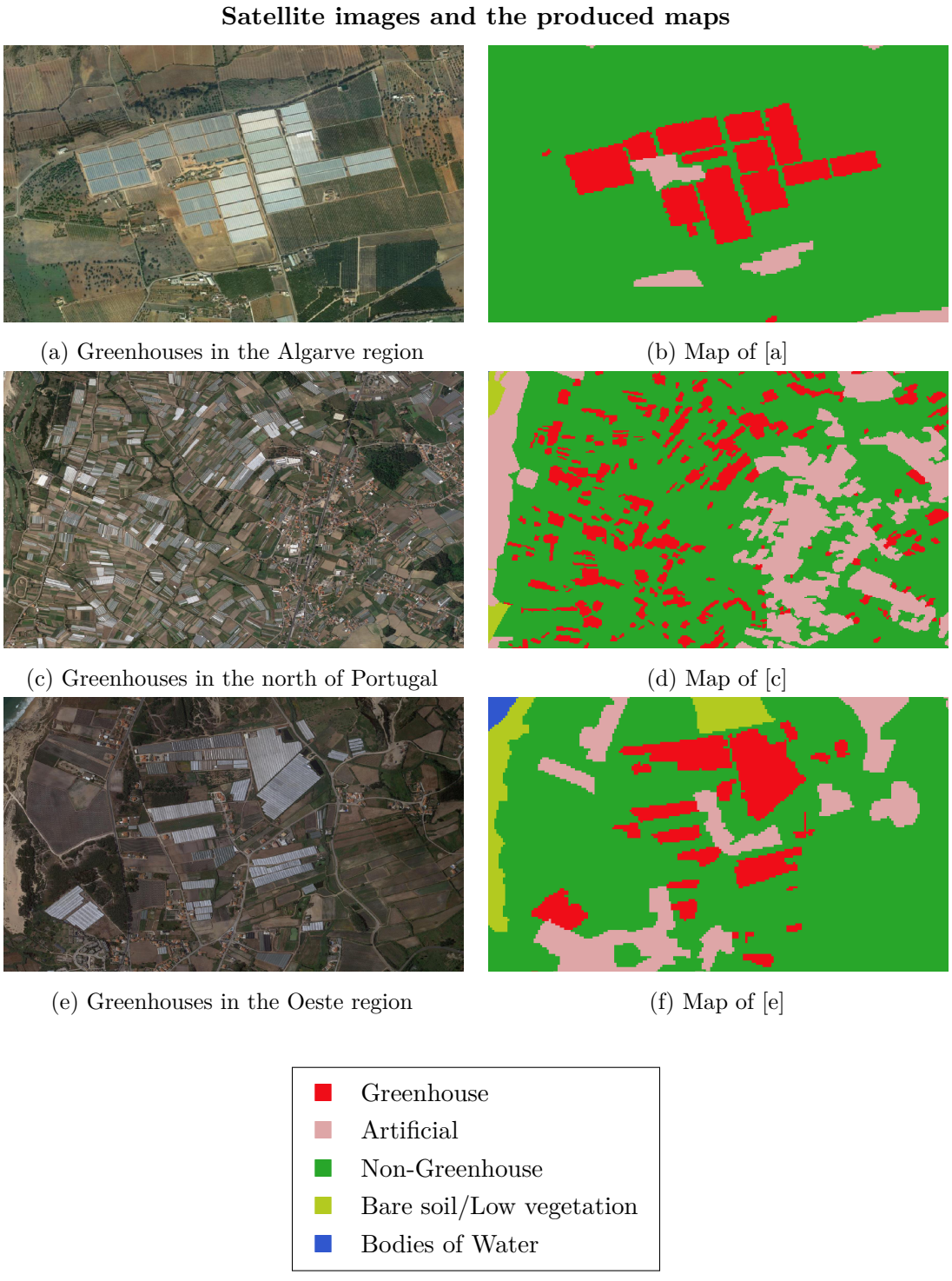
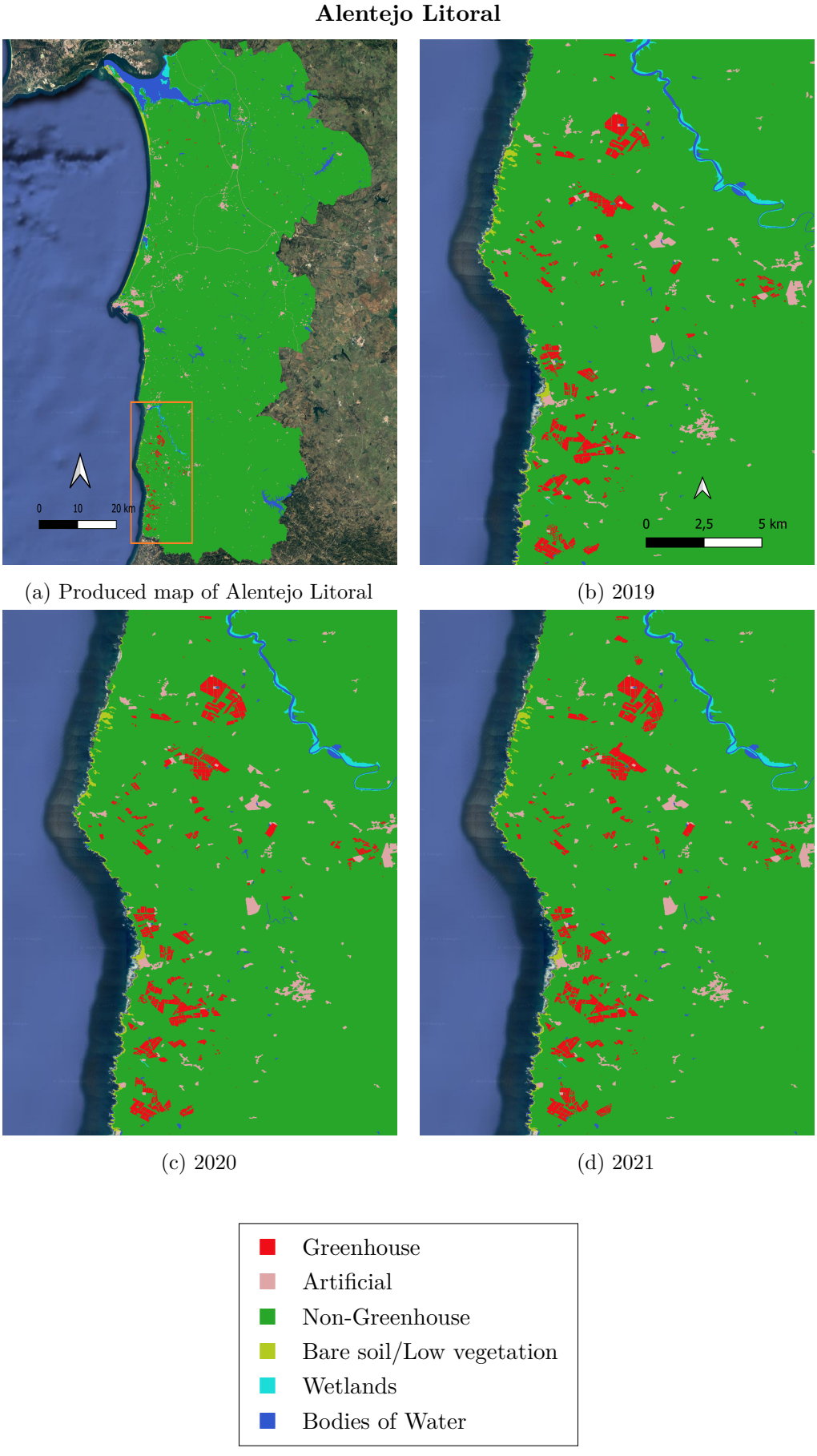


Figure 5.2: Local examples of produced maps.



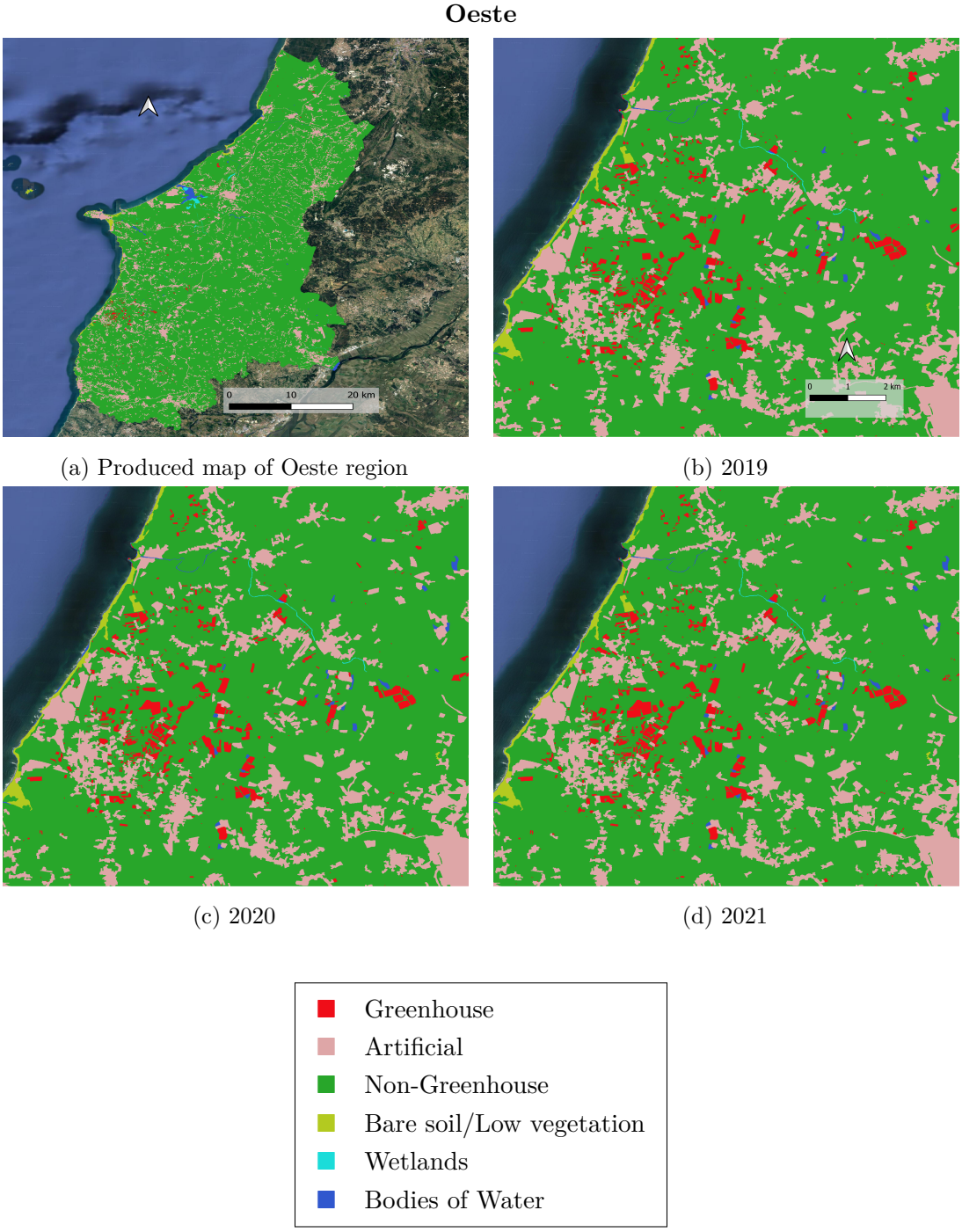
(a) Produced map of Alentejo Litoral

(b) 2019

(c) 2020

(d) 2021

Figure 5.3: Map showing the expansion of plastic greenhouses in Alentejo Litoral.



(a) Produced map of Oeste region

(b) 2019

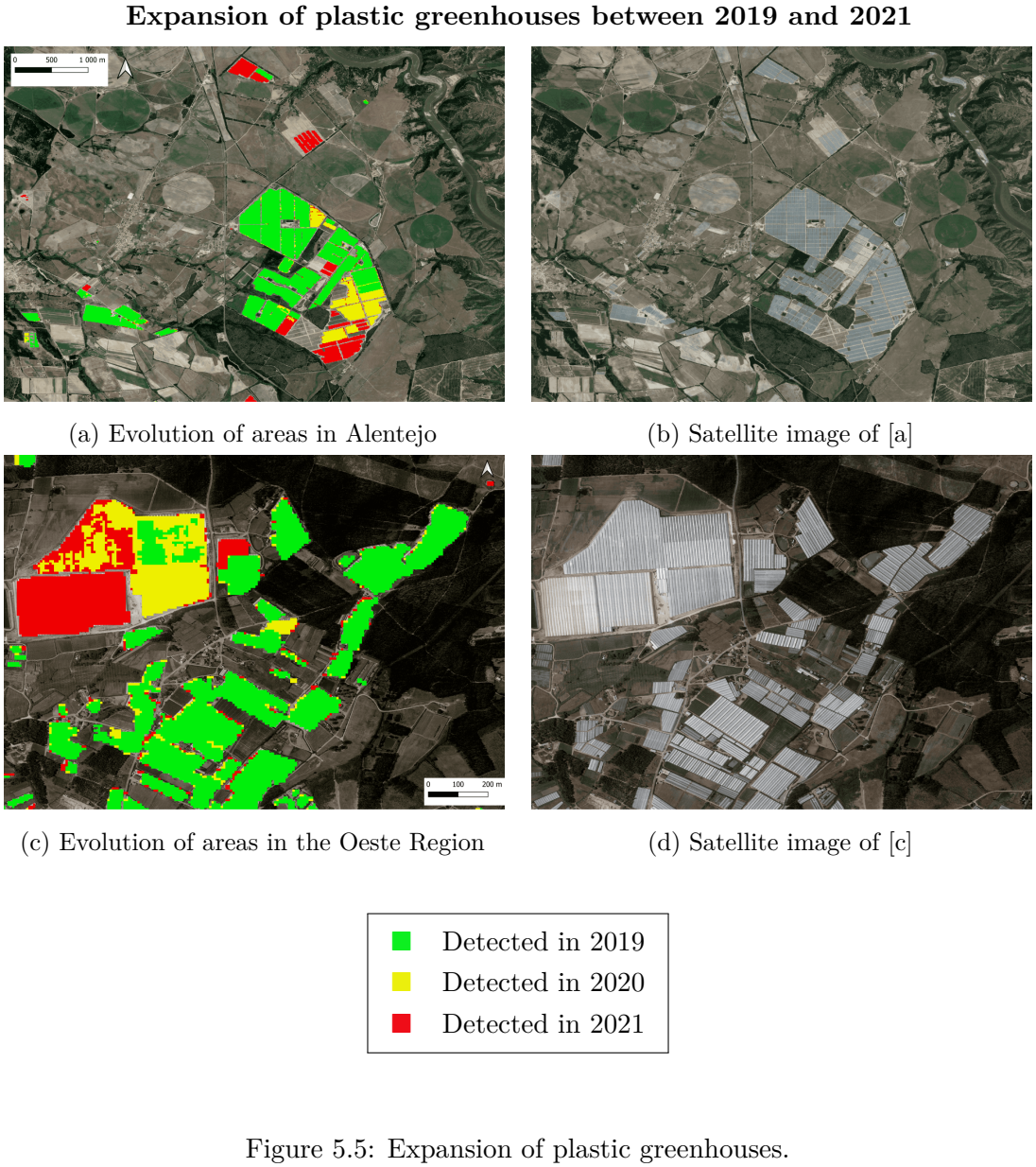
(c) 2020

(d) 2021

Figure 5.4: Map showing the expansion of plastic greenhouses in the Oeste region.

5.1.2 Greenhouse Land Cover Dynamics

In Figure 5.5, examples of the expansion of plastic greenhouses between 2019 and 2021 are presented.



(a) Evolution of areas in Alentejo

(b) Satellite image of [a]

(c) Evolution of areas in the Oeste Region

(d) Satellite image of [c]

Figure 5.5: Expansion of plastic greenhouses.

5.2 Accuracy Assessment

As discussed in the previous chapter, the accuracy assessment will be presented and measured using accuracy matrices, user's and producer's accuracy.

5.2.1 Accuracy Assessment for the year in which the classifier was trained (2020)

Due to the algorithm being trained with only data from 2020, this is the year when the best results should be achieved and the confusion matrices in Tables 5.1 for Oeste, 5.2 for Algarve, 5.3 for Porto/Cávado regions and 5.4 for Alentejo Litoral show precisely that, the results for these regions varied but both user and producer's accuracy remain above 95% in all cases.

Table 5.1: Confusion Matrix for Oeste in 2020

	Greenhouses	Non-Greenhouses	**User's Accuracy**
Greenhouses	3724	31	99.17%
Non-Greenhouses	23	4557	99.50%
Producer's Accuracy	99.39%	99.32%	

Table 5.2: Confusion Matrix for Algarve in 2020

	Greenhouses	Non-Greenhouses	**User's Accuracy**
Greenhouses	2996	5	97.91%
Non-Greenhouses	123	3043	97.04%
Producer's Accuracy	96.06%	99.84%	

Table 5.3: Confusion Matrix for Porto and Cávado Regions in 2020

	Greenhouses	Non-Greenhouses	**User's Accuracy**
Greenhouses	552	6	98.92%
Non-Greenhouses	2	735	99.73%
Producer's Accuracy	99.41%	99.38%	

5.2.2 Accuracy Assessment for the previous (2019) and following year (2021)

The other goal of this dissertation is to test how the classifier behaves in years for which it wasn't trained. For this task, the accuracy is measure in the same region (Alentejo Litoral) for

Table 5.4: Confusion Matrix for Alentejo Litoral in 2020

	Greenhouses	Non-Greenhouses	**User's Accuracy**
Greenhouses	6933	58	99.17%
Non-Greenhouses	9	7994	99.89%
Producer's Accuracy	99.87%	99.28%	

different years. By comparing the results of Table 5.5 for the previous year (2019), Table 5.6 for the following year (2021) and the already discussed Table 5.4 for the training year (2020), the conclusions are the expected. The training year has better results than the other periods, however the results are just slightly lower for 2019 and 2021, leading to the conclusion that the classifier can be used over the years to monitor plastic greenhouses.

Table 5.5: Confusion Matrix for Alentejo Litoral in 2019

	Greenhouses	Non-Greenhouses	**User's Accuracy**
Greenhouses	7490	409	94.82%
Non-Greenhouses	7	9488	99.93%
Producer's Accuracy	99.91%	95.87%	

Table 5.6: Confusion Matrix for Alentejo Litoral in 2021

	Greenhouses	Non-Greenhouses	**User's Accuracy**
Greenhouses	6395	279	95.81%
Non-Greenhouses	8	9046	99.91%
Producer's Accuracy	99.87%	97.01%	

5.2.3 Overall Accuracy

Finally, according to Table 5.7, the overall accuracy for the confusion matrices are promising in all regions and years, boosting the confidence in the use of remote sensing for greenhouse detection.

Table 5.7: Confusion Matrix Accuracy

	Accuracy
Alentejo Litoral 2019	97.61%
Alentejo Litoral 2020	99.55%
Alentejo Litoral 2021	98.17%
Oeste 2020	99.35%
Algarve 2020	97.96%
Porto/Cávado 2020	99.38%

5.3 Statistical results

In this section, the area of occupation of plastic greenhouses, calculated according to the developed classifier, will be discussed for several subregions of Portugal. The choice of subregions was made visually by selecting the areas with the largest incidence, according to Figure 5.1, the selected subregions can be seen in Figure 5.6.

It's important to understand that accuracy varies slightly between years, so in areas like the Norte subregion (Table 5.8), where the evolution was rather small, the decreased of 1.2% can be a statistical error. However, this may also be explained by the fact that plastic greenhouses, due to their flexible and cheap nature, can be built or dismantled based on market necessities or trends, resulting in temporary slight decrease or increase in the area of occupation.

Table 5.8: Evolution of plastic greenhouse areas between 2019 and 2021

	Odemira	Oeste	Algarve	Tejo	Norte
2019	1047 ha	664 ha	560 ha	302 ha	267 ha
2020	1187 ha	672 ha	591 ha	298 ha	275 ha
2021	1253 ha	703 ha	596 ha	322 ha	264 ha
Difference (ha)	+206 ha	+39 ha	+36 ha	+20 ha	-3 ha
Difference (%)	+ 19.7%	+ 5.9%	+ 6.4%	+ 6.6%	- 1.2%

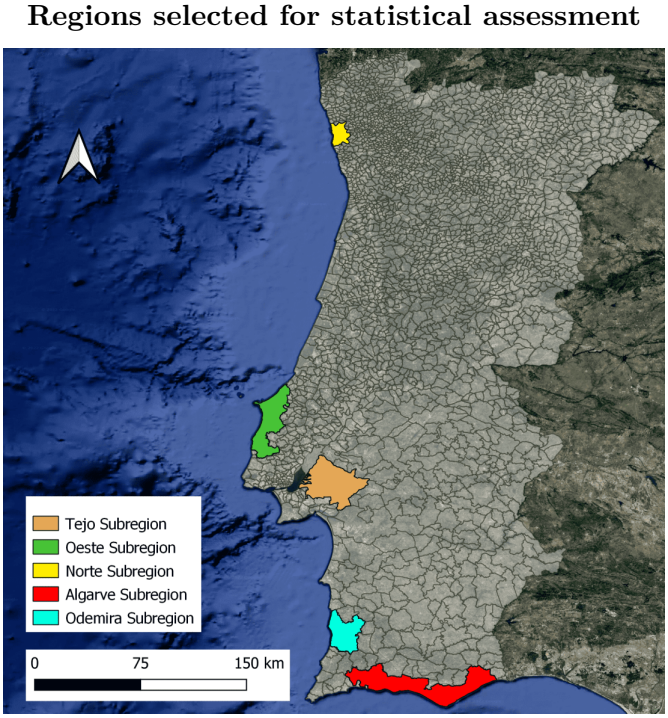


Figure 5.6: Regions selected for statistical assessment.

As previous observed, greenhouses did expand in area in most of the studied regions during the 3-year period analyzed. The growth was not homogeneous, with Odemira, a subregion located at Alentejo Litoral, being the region with the largest increase. Satellite images for the year of 2022 also show that this expansion is not stopping, however as of the release of this dissertation, we still don't have the total of necessary data to measure it. We conclude that Odemira is the most relevant region in Portugal regarding the studied topic.

The Norte subregion is the most stable of the five subregions, with a decrease of 1.2% (Figure 5.2 images [c] and [d]). We may speculate that this is because of how inserted among urban areas the greenhouses are, leaving little space for expansion unlike other regions.

Chapter 6

Conclusions

This dissertation addressed the problem of monitoring plastic greenhouses using satellite imagery. This is a difficult problem: for example, patterns in plastic greenhouses physical and spectral characteristics tend to be inconsistent, coastal areas and artificial buildings are prone to be misclassified as greenhouses and filtering clouds in satellite images is tricky. Furthermore, it is difficult to maintain a balance between the quality and quantity of data that could be obtained in order to support the accuracy of the classifier. Most of these problems were mitigated, to the point that they didn't affect the main goals of this dissertation.

The proposed method accomplishes the initial objective of developing a tool that could help to detect and monitor greenhouses in Portugal for the year of 2020. The algorithm also proved to be able to successfully classify areas for years for which it wasn't trained, resulting in a simple and time-saving solution for analyzing the expansion and temporal evolution of greenhouses. However, without the assistance of land-use maps like Cartografia de Uso e Ocupação do Solo (COS), the results may have too much noise in urban environments, which may create a less than ideal ending result. This is due to the fact that although good and precise outcomes were achieved, a final manual check up of the produced maps is always necessary, in order to be sure of the presence of greenhouses. In an ideal situation, land use maps like COS would not be required. A larger set of training data for the classifier could prove to be a solution to this problem.

In terms of the results obtained, in southern Portugal the produced maps showed that the areas occupied by plastic greenhouses continue to grow: according to our calculations, 19.7% between the years of 2019 and 2021. This is consistent with reports that for the municipality of Odemira (in Alentejo Litoral), recent permissions allowed to expand the area used for intensive agriculture (which includes plastic greenhouses) up to 40% of the municipal area (4800ha) [26]. This municipality is the largest in Portugal and one of the driest; this expansion may jeopardize goal 6 ("Clean Water and Sanitation") of United Nations' "Sustainable Development Goals", by having an unsustainable use of the local water resources. Goal 15, "Life on Land", that aims to prevent the destruction of natural habitats, is also hurt by this continuous growth of used areas. Still, regarding the expansion of plastic greenhouses, maps like the ones in Figure 5.5 show that there is an expansion of pre-existent structures, rather than new areas of greenhouses

appearing. This may be due to a high profitability of this type of agriculture, explaining its continued expansion.

We hope that the tool developed in this thesis will be able to help to take the best decisions to reduce the impact of the climate change in the short time, by facilitating the detection of plastic greenhouse areas; experts may then decide whether they are beneficial or harmful to the environment.

6.1 Future Work

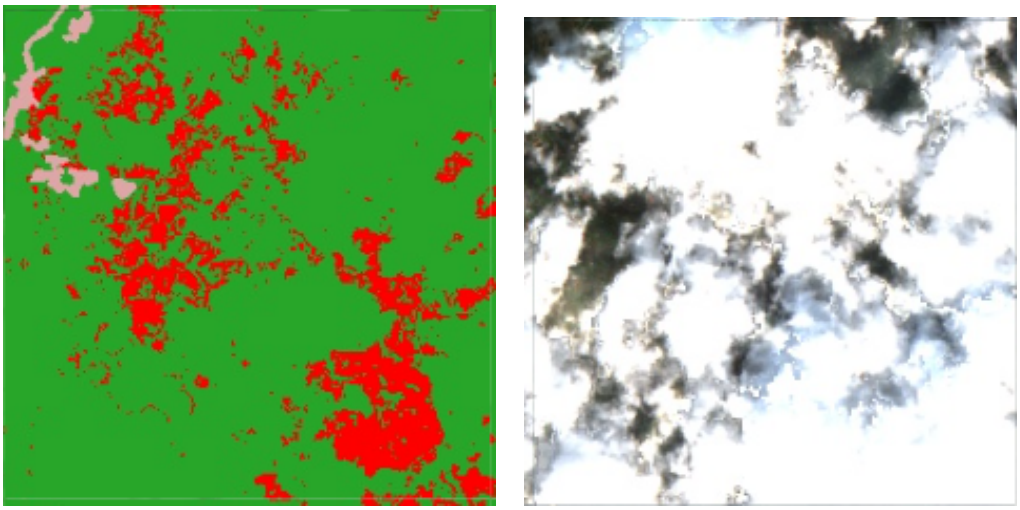
Images with clouds are a problem for the classification algorithm. This was partially mitigated, but it is still an issue. As a consequence, we were only able to produce maps for continental Portugal, because both the Autonomous Regions of the Azores and Madeira proved to be very cloudy, and it was not possible to gather data for certain periods of the year that are required by the classifier. For example, in the year of 2020, during the months of May in Madeira and June in the Azores, no images without clouds could be gathered, as shown in Figure 6.1; this makes the classification task very difficult.

(a) Açores, June 2020 (b) Madeira, May 2020

Figure 6.1: Satellite images of the island of São Miguel (a) and Madeira (b); white areas correspond to clouds.

Another improvement to be made is related to training data. Regions where plastic greenhouses abound were well documented and used to train the classifier; but this was not the case for all regions of Portugal. Neglected areas, often in the interior of the country, have different characteristics: they may be more mountainous, more cloudy and cooler in winter. This may explain why some natural areas are misclassified as greenhouses by our method (see Figure 6.2). These areas were largely neglected in the dataset; more data is necessary for improving accuracy.

If the need arises to recreate this work for a geographical area other than Portugal, or if there is a later need to update the dataset used, one of the most time-consuming tasks will be collecting and labeling data. In order to reduce the time spent and optimize the data collected (reducing the amount of redundant data), Active Learning can be used. This is a machine learning technique that, after initial data is collected and labeled and a classifier is trained, aims to reduce the amount of manual labeling done by the user for improving accuracy. This may be done interactively, with the user being asked to label data for which the algorithm had greater



(a) Misclassification of greenhouses

(b) Satellite image of produced map in December

Figure 6.2: Example of misclassification in the interior of Portugal.

uncertainty. After receiving a label, these new data will be introduced in the training set, and the procedure repeated [27]. In the case of the type of classifier used, Random Forests, the uncertainty can be measured by the number of trees that voted for the predicted class of a pixel; hence, the smaller the number of trees voting for the most voted class, the greater the uncertainty of the classifier regarding the prediction made. Unfortunately, Google Earth Engine (GEE) does not currently provide straightforward tools for measuring uncertainty (counting how many votes the predicted class had in each pixel); hence, tackling this has been left for future work.

Appendix A

Source Code

The code and workflow of this work are available in https://github.com/PedroMPCardoso/MonitoringGreenhousesPortugal. The repository has information that explains how to implement this work, provides the used data (training and testing datasets), provides the produced maps and presents the code created during this dissertation.

Bibliography

[1] Guilherme Lages Barbosa, Francisca Daiane Almeida Gadelha, Natalya Kublik, Alan Proctor, Lucas Reichelm, Emily Weissinger, Gregory M. Wohlleb, and Rolf U. Halden. Comparison of Land, Water, and Energy Requirements of Lettuce Grown Using Hydroponic vs. Conventional Agricultural Methods. *International Journal of Environmental Research and Public Health*, 12(6):6879–6891, June 2015. ISSN: 1660-4601. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/ijerph120606879.

[2] Paulo Barriga. No mundo secreto das estufas. July 2021.

[3] Carta de Uso e Ocupação do Solo - COS 2018 - RDF - Projeto Cross-Forest - Land Use Land Cover Map - COS 2018 - RDF - Cross-Forest Project (COS2018) - dados.gov.pt - Portal de dados abertos da Administração Pública, .

[4] Sustainable Development Goals | United Nations Development Programme, .

[5] Nathaniel Cresswell-Clay, Caroline C. Ummenhofer, Diana L. Thatcher, Alan D. Wanamaker, Rhawn F. Denniston, Yemane Asmerom, and Victor J. Polyak. Twentieth-century Azores High expansion unprecedented in the past 1,200 years. *Nature Geoscience*, pages 1–6, July 2022. ISSN: 1752-0908. Publisher: Nature Publishing Group. doi:10.1038/s41561-022-00971-w.

[6] Damian Carrington and Damian Carrington Environment editor. Spain and Portugal suffering driest climate for 1,200 years, research shows. *The Guardian*, July 2022. ISSN: 0261-3077.

[7] Rádio e Televisão de Portugal. Odemira. Manto de silêncio protege negócio das agências de trabalho temporário. *Odemira. Manto de silêncio protege negócio das agências de trabalho temporário.*

[8] Beatriz Ramalho da Silva and Corinne Redfern. Fruit pickers lured to Portugal by the dream of a 'raspberry passport'. *The Guardian*, January 2022. ISSN: 0261-3077.

[9] Bruna Alexandra Moreira Vieira. Precariedade na agricultura: a realidade dos trabalhadores agrícolas sazonais. July 2018. Accepted: 2019-02-06T18:35:22Z.

[10] Siamak Khorram, Frank H. Koch, Cynthia F. van der Wiele, and Stacy A. C. Nelson. *Remote Sensing*. Springer Science & Business Media, February 2012. ISBN: 978-1-4614-3103-9. Google-Books-ID: QpdQut59MPsC.

[11] Sentinel-2 - Missions - Sentinel Online - Sentinel Online.

[12] John Weier and David Herring. Measuring Vegetation (NDVI & EVI), August 2000. Publisher: NASA Earth Observatory.

[13] Dedi Yang, Jin Chen, Yuan Zhou, Xiang Chen, Xuehong Chen, and Xin Cao. Mapping plastic greenhouse with medium spatial resolution satellite data: Development of a new spectral index. *ISPRS Journal of Photogrammetry and Remote Sensing*, 128:47–60, June 2017. ISSN: 0924-2716. doi:10.1016/j.isprsjprs.2017.03.002.

[14] Stefania Bonafoni and Aliihsan Sekertekin. Albedo Retrieval From Sentinel-2 by New Narrow-to-Broadband Conversion Coefficients. *IEEE Geoscience and Remote Sensing Letters*, 17(9): 1618–1622, September 2020. ISSN: 1558-0571. Conference Name: IEEE Geoscience and Remote Sensing Letters. doi:10.1109/LGRS.2020.2967085.

[15] Google Earth Engine, .

[16] Google Colaboratory, .

[17] QGIS features, .

[18] Hasituya and Zhongxin Chen. Mapping Plastic-Mulched Farmland with Multi-Temporal Landsat-8 Data. *Remote Sensing*, 9(6):557, June 2017. ISSN: 2072-4292. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/rs9060557.

[19] Haoran Sun, Lei Wang, Rencai Lin, Zhen Zhang, and Baozhong Zhang. Mapping Plastic Greenhouses with Two-Temporal Sentinel-2 Images and 1D-CNN Deep Learning. *Remote Sensing*, 13(14):2820, January 2021. ISSN: 2072-4292. Number: 14 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/rs13142820.

[20] Introduction to Environmental Forensics. In Brian L. Murphy and Robert D. Morrison, editors, *Introduction to Environmental Forensics (Third Edition)*, pages 707–719. Academic Press, San Diego, January 2015. ISBN: 978-0-12-404696-2. doi:10.1016/B978-0-12-404696-2.18001-4.

[21] Agricultural year - Instituto Nacional de Estatística, December 2004.

[22] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN: 1573-0565. doi:10.1023/A:1010933404324.

[23] Cha Zhang and Yunqian Ma, editors. *Ensemble Machine Learning*. Springer US, Boston, MA, 2012. ISBN: 978-1-4419-9325-0 978-1-4419-9326-7. doi:10.1007/978-1-4419-9326-7.

[24] Hong Han, Xiaoling Guo, and Hua Yu. Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest. In *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pages 219–224, August 2016. ISSN: 2327-0594. doi:10.1109/ICSESS.2016.7883053.

[25] L.L.F. Janssen, Frans van der Wel, and F.J.M. Accuracy assessment of satellite derived land-cover data: A review. *Photogramm. Eng. Remote Sensing 60 (1994) 419-426.*, 60, April 1994.

[26] José Manuel Fernandes. Agricultura intensiva das estufas de frutos vermelhos no Parque Natural do Sudoeste Alentejano: Out of Control? *Observador*.

[27] Burr Settles. Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, June 2012. ISSN: 1939-4608. Publisher: Morgan & Claypool Publishers. doi:10.2200/S00429ED1V01Y201207AIM018.