POLITECNICO DI TORINO

Master's Degree in Data Science and Engineering

Master's Degree Thesis

Land Cover and Crop Type Classification Using Machine Learning Techniques on Satellite Multispectral Data

Supervisors

Prof. FABRIZIO LAMBERTI

Prof. LIA MORRA

Candidate

ALBERTO MARIA FALLETTA

JULY 2021

Abstract

As the global population grows in the near future, is expected to reach the number of 9.7 billion by 2050, agriculture will become more important than ever. Worldwide, in fact, the food production industry is already under severe pressure. On one hand, the change in climatic patterns caused by global warming triggers mutations in ecosystems with the development of new plant diseases and stronger pests, with significant impacts on production. On the other, the public opinion is day by day less in favor of unsustainable production methods as the food sector is responsible for intensive exploitation practices and about one-third of global greenhouse gas emissions. The industry will have to inevitably reinvent itself in order to increase its productive capacity and be able to meet the expected growing demand, lowering costs and reducing environmental impacts. One of the most promising innovations, whose aim is to guide agriculture in this transition, is the application of satellite data to map and monitor every key aspect of crop growth maximizing yields, reducing water wastes, reducing fertilization pollution, lowering crop management costs and much more. On this path, the aim of this study is to examine the possibility of obtaining land cover and crop type information by means of multispectral data provided by Sentinel-2 mission of the European Copernicus program, which are fundamental knowledge in order to identify characteristics of a selected area allowing its management in a way completely tailored on its specifics. In order to do that, it is proposed the creation of two datasets on which to train and evaluate models, one for the task of land cover classification, one for the task of crop type classification, by computing time series of spectral indices associated with points included in a third dataset from Eurostat called LUCAS used for the labels.

"Man must rise above the Earth, to the top of the atmosphere and beyond, for only thus will he fully understand the world in which he lives."
Plato, Phaedo, IV century BC

Table of Contents

List of Tables

List of Figures

Acronyms

CIR
Color infrared

EO
Earth Observation

ESA
European Space Agency

MSI
Multispectral Instrument

MSS
Multispectral Scanner

NASA
National Aeronautics and Space Administration

NIR
Near-Infrared

SWIR
Short-wave Infrared

USGS
U.S. Geological Survey

VNIR
Visible and Near-infrared

Chapter 1

Introduction

Once solved the challenges posed by the storage and computational power, fundamental requirements for processing and managing an ever-increasing amount of data, both the worlds of academic research and industry research have experimented applying machine learning models to all possible sectors in order to extract knowledge from data and consequently solve problems, detect patterns, increase efficiencies, manage costs, identify new market opportunities and boost market advantages.

Nowadays data science techniques have broad and comprehensive application areas ranging from health to finance, marketing, process automation, energy production and many more. Remote sensing or, more specifically, Earth observation (EO) makes no exception. Data produced in these sectors are, in fact, extremely valuable not only due to the high costs suffered by companies and institutions on whose shoulders weights the burden of producing satellites and putting them into orbit, but also since these data are soaked in important knowledge necessary to analyze the impacts of climate change, monitor ocean temperatures, detect land change and, as for the scope of this thesis, monitor and improve agriculture-related processes.

1.1 Remote Sensing and Earth Observation

Remote sensing, which is often related to as Earth observation (EO) when its focus is the blue planet, is the discipline based on the acquisition of information about objects or phenomena without making physical contact with them [1]. Given the definition it does not come as a surprise this discipline began with photography. Cameras, the first remote sensing devices to be developed and perfected, enabled, in fact, revolutions in many fields, among them the most relevant being science and art with landscape and naturalistic photography first and later, once shortened

exposure times, portrait and aerial photography.

Aerial photography, an early form of EO, was first conceived by the French photographer and balloonist Gaspard-Félix Tournachon, best known as Nadar, who patented the idea of using photographs captured from a height for map-making and surveying and was able to take the very first picture over Paris in 1858 [2]. Slowly, progress in technology and the awareness of the dangers associated with balloons, made it possible to take cameras into the skies in other safer ways, using kites, pigeons and rockets, but it was not until war times that the discipline had its major improvements. As many technologies do, in fact, aerial photography too benefited from unprecedented public funding as reconnaissance aircraft were equipped with cameras to record enemy positions, movements and defenses.

The military, in fact, has always driven technological advancements essential to achieve victory, and while during the first World War the economic effort focused mainly on improving the hardware of cameras in order to easily and efficiently allocate them on aircraft, one of the most important improvements, traceable to World War II as an aid in camouflage detection, was undoubtedly the origin of non-photographic films, not much for the step ahead in technology but mostly for the idea behind it, idea that, more than any other, pointed out the direction for the future development of the discipline. The main concept was, in fact, to broaden the operability spectrum of the camera, having the possibility of acquiring not only panchromatic images (obtained as a combination of the information from the visible wavelengths of blue, green and red resulting in a single band formed by the total light energy in the visible spectrum) but also Color-infrared (CIR) imagery (obtained using a portion of the electromagnetic spectrum known as near-infrared (NIR) ranging from 0.70 µm to 1.0 µm, just beyond the wavelengths associated to the red color), all made possible by improvements of radar (radio detection and ranging), thermal infra-red detection, and sonar (sound navigation ranging) systems.

Up until this point, EO had been a synonym of aerial photography but it all changed with satellites and the space race started in 1957 between the two Cold War rivals, USSR and USA, with the first-ever artificial satellite to be put into orbit, Sputnik 1. Once again identified as strategic from a military point of view, the discipline was pushed by funding in research for space dominance and bellicose advantage. Satellites for scientific purposes benefited too from the industry advancements, proving to be particularly useful in meteorology and specifically allowing scientist to obtain and study images depicting complete cloud systems captured from way higher than what could have previously been captured by aircraft flying just above the clouds. The first weather satellite, Vanguard 2, designed to measure cloud cover for the first 19 days in orbit, was launched on 17$^{\text{th}}$ February 1959, but due to a poor axis of rotation and its elliptical orbit, it was not able to collect a notable amount of useful data and was soon followed by a more successful

mission, TIROS-1 [3].

William Pecora, Director of the United States Geological Survey (USGS, U.S. government agency whose objective is to study the natural resources of the country, its landscape and the natural hazards associated with it) in 1965 was one of the main advocates of the idea of a civilian satellite to conduct scientific and exploratory studies of the Earth's surface to gather information about the planet's natural resources, receiving however strong oppositions by several entities for different reasons. While the ones thought aerial photography from high-altitude aircraft would have been a more responsible approach, the others thought it would have posed risks to national defense with further concerns about photographing foreign countries without permission. In 1970, however, NASA finally received approval to develop the Earth Resources Technology Satellite later known as Landsat-1 and launched it on July 23rd 1972 becoming the first-ever satellite designed specifically to study and to monitor the Earth's surface, capturing over 300.000 multispectral images, thanks to its Multispectral Scanner (MSS), before its termination in January 1978 [4].

Many years and missions have passed since the launch of the first Landsat satellite. Nowadays the program is the longest-running one and it has no intention to stop any time soon with a scheduled launch of its 9th satellite in September 2021. More than 5.6 million acquisitions later, sensed from Landsat-1 to Landsat-8, there is no doubt that the ambitions of the program have largely been met and exceeded. Such legacy of success has not only shaped remote sensing as a whole but has also strongly influenced in many ways following missions and programs such as Copernicus, one of the most relevant ways being the free and open data policy, enabling factor for this very study. Prior to 2008, the costs for a Landsat Multispectral Scanner (MSS) image varied from $20 (1972–1978) to $200 (1979–1982), increased from approximately $3000 to $4000 for a Landsat Thematic Mapper image (1983–1998), and was $600 for an Enhanced Thematic MapperPlus (ETM+) image (1999–2008). In 2008, however, the adoption by USGS of the aforementioned free data policy resulted in a substantial increase in the use of previously costly Landsat images and remarkable benefits for scientific studies, researches and discoveries guided by the knowledge extracted from large numbers of such data. The results were not only the wide adoption, proved by a twenty-fold increase of annual data downloads in 2017 with respect to 2009, but more importantly the increase by more than four-fold of the annual number of publications (considering papers with "Landsat" in the title or abstract in 2017 with respect to 1983) producing knowledge and boosting innovation and employment. The high intrinsic value of Landsat images to users and stakeholders can be summarized by a survey revealing that U.S. users have gained $1.8 billion USD in benefits from the 2.38 million images they downloaded prior to the survey, while the National Geospatial Advisory Committee estimated an economic benefit of Landsat data for the year 2011 as $1.70 billion for U.S. users

plus $400million for international users in sixteen economic sectors [5].

Moreover, Landsat has influenced the industry by proving the reliability and versatility of the Multispectral Scanner that has, considerably, paved the way for subsequent multispectral and hyperspectral sensors until this day, leading to the more recent definition of remote sensing as the use of electromagnetic energy to measure physical properties of distant objects [6]. Since this discipline has moved to the domain of spectroscopy, it is, therefore, important to understand what spectroscopy is and a few of its basic principles.

1.2 Spectroscopy

The year was 1666, when a young Isaac Newton, using a simple instrument made of a small aperture to define a beam of light, a lens to collimate it, a glass prism to disperse it and a screen to display the result, showed that white light from the Sun could be dispersed into a continuous series of colors, describing this phenomenon with the word "spectrum" [7].

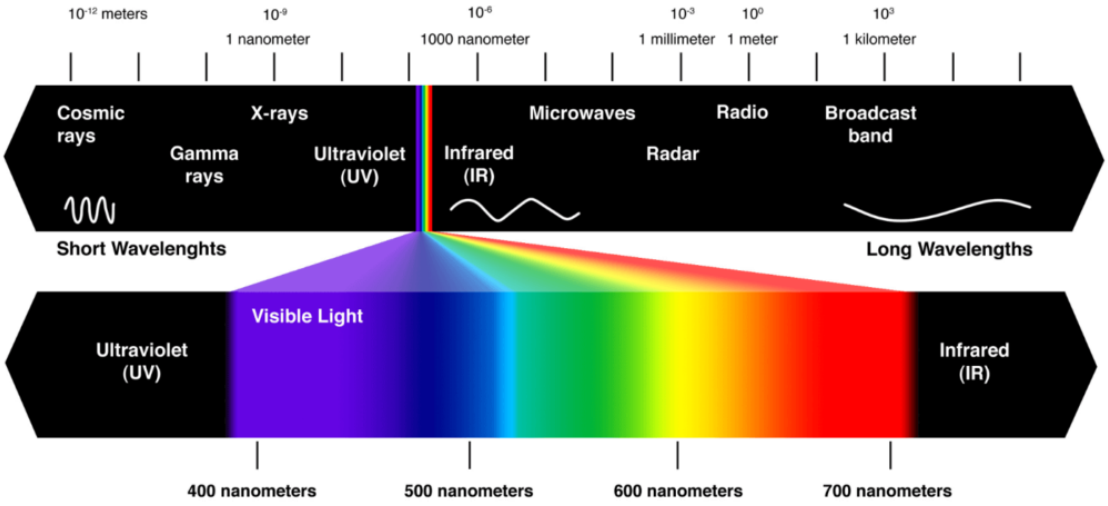

Figure 1.1: Electromagnetic spectrum as we know it today.

That moment was undoubtedly the beginning of spectroscopy but since then many steps ahead have been made starting from the ones taken by the Scottish physicist Thomas Melvill, who had been observing and studying the resultant colors obtained by dropping various salts into a flame and recording the spectra resulted when a slit of light from the flame was passed through a prism and projected on a surface. He found that each substance had not a continuous spectrum but rather a unique set of lines, later called emission lines, in certain sections of the spectrum [8]. During the first half of the 19^{th} century, many scientists have been working in the

field of emission and absorption spectra for both celestial and earthly sources until Joseph von Fraunhofer stretched the spectrum from the Sun, to reveal over 600 dark absorption lines. On studying other stellar spectra and spectra of reflected sunlight, Fraunhofer deduced that each star had a unique set of such lines, and they were actually a function of the star itself. Since then, once explained the link between spectral response and matter, many outstanding discoveries have been made, from the identification of new elements such as cesium and rubidium by Robert Bunsen and Gustav Kirchhoff in 1860 to the deduction, a few years later, of the elemental composition of stars and planets light-years away from the Earth.

Nowadays spectroscopy is defined as the study of the interaction between matter and electromagnetic radiation as a function of the wavelength or frequency of the radiation [9]. This discipline is a fundamental exploratory tool in many fields such as physics, chemistry and astronomy, allowing the composition, physical structure and electronic structure of matter to be investigated from atomic scales up to astronomical distances. Light, in fact, carries much information about the material which it interacts with and since different materials interact differently with light, it is possible to use light to understand what a given target is made of. This is especially possible because matter is made of atoms with a unique structure of a nucleus surrounded by electrons orbiting at different energy levels. Only the light with the exact energy required to go between energy levels can be absorbed and no others. Then, when electrons fall down to lower orbits they release as much energy as the difference between the levels in the form of light, explaining why different atoms emit different colors of light. All elements, in fact, absorb and emit specific wavelengths of light that correspond to those energy levels. It is called absorption spectrum the spectrum of light transmitted through a substance, showing dark lines or bands where light has been absorbed by atoms, while it is called emission spectrum the one made by electrons falling down energy levels. This is exactly what Thomas Melvill was experimenting with, at the first steps of spectroscopy using excited gasses heated by a heat source. Heating, in fact, moves the electrons up in energy levels and when they fall back down the results are bright, colored spikes due to the release of light at precise wavelengths [10].

Such interaction between matter and electromagnetic radiation is, usually, analyzed by means of a tool called spectrometer whose basic function is to take in light, collimate it, break it into its spectral components thanks to a diffraction grating, digitize the signal as a function of wavelength by means of a detector, and display it through a computer. Multispectral scanners and multispectral instruments equipped on satellites work very similarly but are a little more complex. For example the Multispectral Instrument of Sentinel-2 mission by Copernicus, which will be later covered, accepts the light reflected up from Earth and its atmosphere, collects it by a three-mirror telescope (M1, M2 and M3 in figure 1.3) and focuses it via a beam-splitter onto two Focal Plane Assemblies, one for the

visible and near-infrared (VNIR) wavelengths and one for the SWIR wavelengths, where it then finds two distinct arrays of 12 detectors mounted on each focal plane in a staggered configuration to cover the entire field of view.

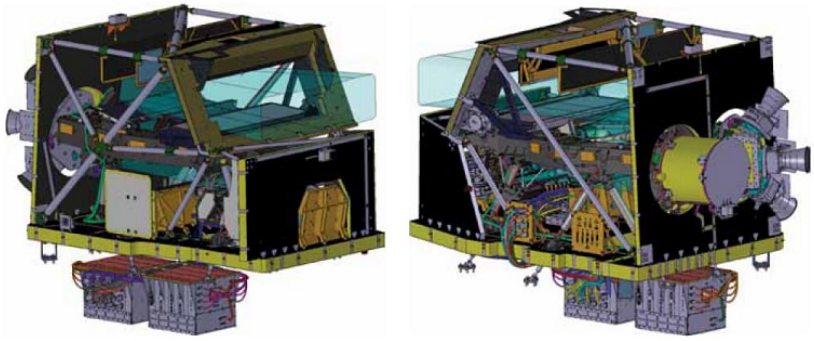

Figure 1.2: Sentinel-2 Multispectral Instrument produced by Airbus Defence and Space). Source: ESA

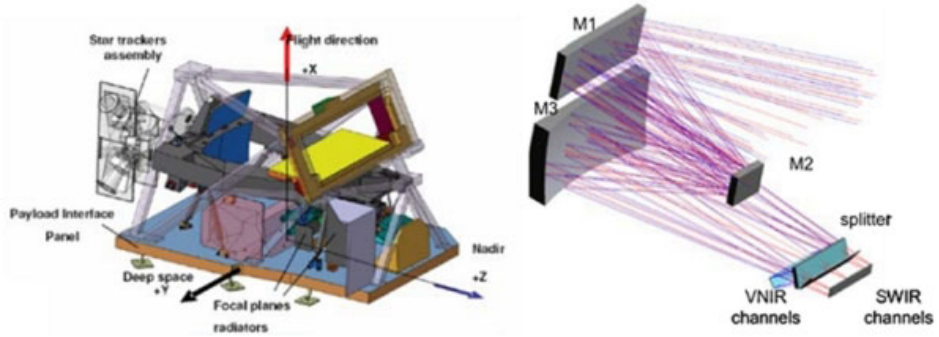

Figure 1.3: Sentinel-2 Multispectral Instrument Internal Configuration. Source: ESA

Analyzing the spectral response registered by a spectrometer carried by a satellite in orbit however, comes with an additional set of challenges unknown to spectroscopy on Earth where the discipline is performed in controlled environments and distances do not affect measurements. When electromagnetic radiations coming from the Sun travel through the atmosphere, in fact, they may interact with particles along their paths giving place to several phenomena, among which:

- Absorption: it happens when the radiation energy is converted into excitation energy of the molecules it interacts with. Ozone, carbon dioxide, and water vapor are the three main radiation-absorbing atmospheric constituents. Ozone absorbs ultraviolet radiation, carbon dioxide absorbs radiation in the far

infrared portion of the spectrum trapping heat inside the atmosphere, while water vapour absorbs much of the incoming longwave infrared and shortwave microwave radiation. Since these gases absorb electromagnetic energy in these very specific regions of the spectrum, by comparing the characteristics of the two most common radiation sources (the Sun and the Earth) along with the combination of wavelengths absorption of the aforementioned atmospheric constituents, it is possible to define the wavelengths that can be used most effectively for remote sensing. The regions, shown in figure 1.4, containing groups of these wavelengths, found in the visible, near-infrared, thermal infrared and microwave, are known as the Atmospheric Transmission Windows [11].

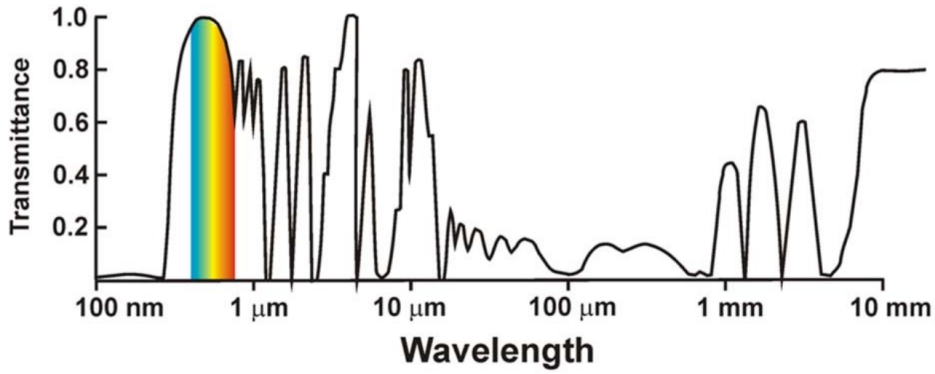

Figure 1.4: transmittance of electromagnetic radiation across the spectrum. Values close to 1, represent 100% transmittance, indicating the all radiation is able to pass through the atmosphere at the given wavelength.

- Scattering: it happens when the radiation is redirected from its original path redistributing the energy of the incident beam to all directions. There exist three types of scattering: the first, called Rayleigh scattering, occurs when particles are very small compared to the wavelength of the radiation. These could be particles such as nitrogen or oxygen molecules, causing shorter wavelengths of energy to be scattered much more than longer wavelengths. This type of scattering is the dominant scattering mechanism in the upper atmosphere, so much that the sky appears blue during the day because of this phenomenon. As sunlight passes through the atmosphere, in fact, the shorter wavelengths of the visible spectrum are scattered more than the longer ones. At sunrise and sunset the light has to travel further through the atmosphere with respect to midday and the scattering of the shorter wavelengths is more complete, leaving a greater proportion of the longer wavelengths to penetrate

the atmosphere, resulting in the red coloration of the sky. The second, called Mie scattering, occurs, mostly in the lower portions of the atmosphere, when the particles are just about the same size as the wavelength of the radiation, such as dust, pollen, smoke and water vapour and tends to affect longer wavelengths. The final scattering effect, called non-selective scattering, occurs when the particles are much larger than the wavelength of the radiation, such as water droplets and large dust particles. Non-selective scattering gets its name from the fact that all wavelengths are scattered about equally. This type of scattering causes fog and clouds to appear white since blue, green, and red light are all scattered in the same way. Scattering causes degradation in the final product in the form of an hazy appearance of the image or in the form of a blur of the targets due to spreading of the reflected radiations and resulting in a reduced resolution image [12]. Another effect, related to scattering, causing the degradation of the sensed image occurs when the light from targets outside the field of view of the sensor is scattered into its field of view. This effect is known as the adjacency effect and it is particularly evident near a boundary between two regions of different brightness resulting in an increase of brightness of the darker region and a reduced brightness of the brighter region.

The amount and combination of the two phenomena depend on several factors including the wavelength of the radiation, the abundance of particles or gases, and the length of the path that the radiation has to travel through the atmosphere.

Among other challenges posed by the distance, there are the facts that:

- Earth surface materials are known to be generally non-Lambertian in nature, which means, they do not reflect the incoming radiation equally in all directions, and tend to be anisotropic (exhibit reflectance directionality) [13][14]. The degree of anisotropy depends on the spectral and directional nature of the radiation and on the properties of the surface itself and more specifically on its density and arrangement (surface structure), which in turn introduce shadows under clear skies with varying illumination zenith and azimuth angles and transmittance and absorption properties of the surface. In addition, the measured reflectance will vary depending on the view, the illumination and the solar zenith angles of the surface under clear sky conditions [15][16]. The non-Lambertian property of Earth surface materials is a limitation especially during the calibration phases of the sensors. Calibration, in fact, requires a near-Lambertian surface on which the sensed reflectance should be spectrally flat with change in time.

- The output signal is not only affected by the presence of the atmosphere, and all the aforementioned phenomena, but it also depends on the sensor carried

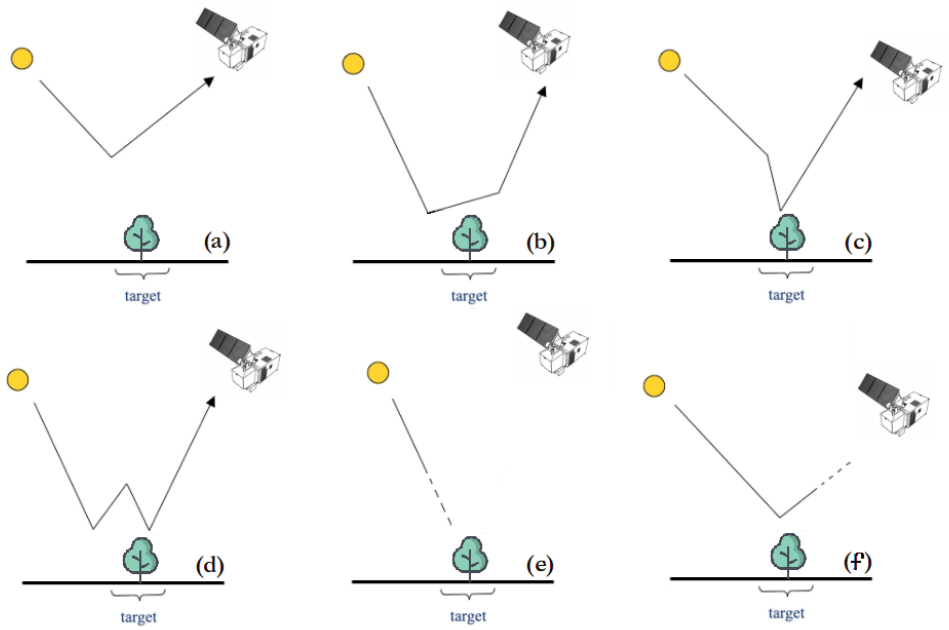

Figure 1.5: (a) Atmospheric scattering, (b) adjacency effect, (c) energy transmitted and diffused from the atmosphere to the target, (d) multiple reflections and scattering, (e, f) absorption

by the satellite itself and more specifically on its Point Spread Function (PSF), a function of the sensor's optical properties and detector's properties. As for the optical properties of the sensor, aberrations and misalignments may cause spectral non-uniformity of the sensed image, called "smile effect" (resulting in a non-uniform spectral response given by an actual uniform surface) or spatial misregistrations commonly known as the "keystone effect" (resulting in black pixels in given areas of the image), while for the detector, it might happen, for elements of the array which it is made of, to have a slightly different gain with respect to one another causing striping effect (resulting in literal stripes in the image).

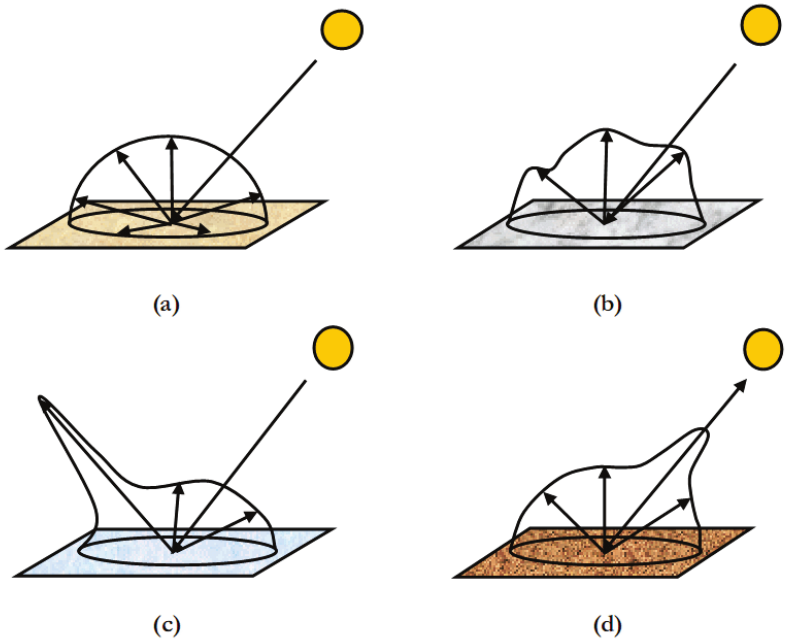

Figure 1.6: Four examples of surface reflectance: (a) Lambertian reflectance (b), non-Lambertian (directional) reflectance (c), specular (mirror-like) reflectance, (d) retro-reflection peak (hotspot). Source: [17]

1.3 Data

1.3.1 Data Source

The data in use in this study come from an Earth observation program called Copernicus, managed by the European Commission in partnership with the European Space Agency (ESA), and more specifically from missions Sentinel-2. Copernicus's objective is to provide data to policymakers who need information to develop environmental legislation and policies, to public authorities in order for them to take critical decisions in emergency situations, such as a natural disaster or a humanitarian crisis, as well as to companies and to the general public for economic purposes due to jobs creation and innovation boost. The program achieves such objective via a collection of services fueled by data gathered from in situ systems supervised by European Environment Agency and EU Member States, by means of ground stations, which deliver data acquired by a multitude of sensors on the ground, at sea or in the air, as well as from data gathered from dedicated satellites belonging to missions grouped under the name of Sentinels, satellite constellations that have been growing in number year after year since the launch of the first, Sentinel-1A in 2014. Copernicus also transforms this wealth of satellite and in situ data into value-added ready to use products by processing and analyzing the data, examining patterns to create better forecasts and creating maps from imagery and identifying features and anomalies and extracting statistical information, resulting in new business opportunities for companies [18].

Sentinel-2

Sentinel-2 mission's aim is to provide global acquisitions of high-resolution and high revisit frequency multispectral images. Differently from color images consisting in the representation of light reflected only from the portion of the spectrum associated to the wavelengths of red, green and blue, multispectral images are simply image data that represent the spectral response in a greater number of determined wavelengths, with the sequence of such values called spectral signature. Sentinel-2 is composed of a constellation of two satellites manufactured by a consortium led by Airbus Defence and Space. The satellites, polar-orbiting phased at 180° to each other in the same Sun-synchronous orbit at a mean altitude of 786 km, allowing them to achieve a high revisit time (10 days at the equator with one satellite, and 5 days with 2 satellites under cloud-free conditions), carry a MultiSpectral Instrument (MSI), produced by Astrium SAS (France), that samples, by passively collecting the sunlight reflected from the Earth, four bands at 10 m (meaning each pixel of the sensed image covers an area of 10 m x 10 m), six bands at 20 m and three bands at 60 m spatial resolution, for a total of 13 bands. As shown in Figure 1.3, the incoming light beam is split and focused onto two separate

focal planes within the instrument, one for VNIR bands and one for Short Wave Infra-Red (SWIR) bands. The spectral separation of each band into individual wavelengths is accomplished by stripe filters mounted on top of the detectors. A shutter mechanism prevents the instrument from direct illumination by the Sun in orbit and avoids contamination during launch. The same mechanism functions as a calibration device by collecting the sunlight after reflection by a diffuser.

Band	Number	Central wavelength (nm)	Bandwidth (nm)	Res
Coastal	1	442.7	21	60m
Blue	2	492.4	66	10m
Green	3	559.8	36	10m
Red	4	664.6	31	10m
Red Edge 1	5	704.1	15	20m
Red Edge 2	6	740.5	15	20m
Red Edge 3	7	782.8	20	20m
NIR 1	8	832.8	106	10m
NIR 2	8A	864.7	21	20m
Water Vapour	9	945.1	20	60m
SWIR 1	10	1373.5	31	60m
SWIR 2	11	1613.7	91	20m
SWIR 3	12	2202.4	175	20m

Table 1.1: Spectral bands of Sentinel-2A sensor

1.3.2 Data Applications

The precious data gathered by the aforementioned mission have a crucial significance in shaping the future of the world. Apart from agriculture, the following are few of the main field of applications of such data.

Biodiversity and environmental protection

Human activities are causing deterioration to ecosystems and put enormous strain on the environment that progressively degrades under the weight of pollution, urbanization, and global warming as demonstrated by an ever-increasing decline in biodiversity. For these reasons, satellite data, providing, with unprecedented frequency, information useful to monitor the environment in its entirety via land, atmosphere and ocean parameters, allowing the tracking of vegetation health, chlorophyll content estimations, oceanic currents, oceanic temperature and more,

are part of a higher project that is the support of the European Union environmental policies whose aim is the preservation of the natural environment, essential to have clean water and air, maintain soils, regulate the climate, recycle nutrients and provide mankind with food. The environment, however, can only be protected if these policies are properly implemented. These data are, therefore, important means of awareness for local and regional authorities, decisive players in environmental protection being responsible for rule-making, enforcement and undertaking investments, boosting the implementation of EU environmental standards and sustainable growth.

Climate, water and energy

Water and energy, both strictly linked to climate, are two key aspects of life in an ever-increasing demand. Climate change affects the availability of the two resources in multiple ways and for this reason, effective adaptation measures need to be taken to reduce exposure and vulnerability to shortages. Satellite data provide authoritative, quality-assured information to help understanding climate change and guide the development of policies addressing mitigation and adaptation measures avoiding the solutions for the needs in one area to produce unintended outcomes in another, with unexpected broader economic, environmental, and security consequences. Such measures are sustained thanks to the monitoring of inland water basins and snow/glaciers, by performance forecasting for renewable energy sources such as solar, wind and hydro-power, and finally ocean surface temperature, ocean surface height and more.

Territorial Management and urban planning

In Europe, over two thirds of the population lives in urban areas, using around 80% of the energy and generate up to 85% of the GDP. Geospatial information regarding land use and land cover, urban growth, urban green areas and urban heat islands is key in order to manage such areas and guide them towards sustainable development by integrating different scales of cities and human settlements, making sure that supplies and demands between urban and rural areas are smoothly flowing and territories are connected and ensuring that citizens' private and social living is balanced, planning infrastructure and services that facilitate trade and productivity safeguarding the environment and social public places.

Civil protection

Floods, landslides, earthquakes, wildfires, volcanic eruptions and other disasters can occur at any moment in time and not only cause economic and environmental damage but more importantly, threaten lives. Satellite data can be used to organize

the response to emergencies in the immediate aftermath of a disaster, improving preparedness through mapping risk-prone areas and providing early warnings related to specific types of events such as floods or wildfires, but it can also be used for post-disaster assistance, rehabilitation and reconstruction, or even before disasters take place organizing prevention with monitoring and alerting functions for some types of disasters such as volcanic eruptions.

Transports, Civil infrastructure and safety

Countries need efficient transport systems and reliable infrastructures if they are to prosper and provide a decent standard of living for their populations, and ensuring passenger safety is a priority for public authorities. Satellite data allow improved planning and management of civil infrastructure and the prevention of future damages through information on the topography and on instabilities of the terrain surface that may arise due to subsidence, sliding or underground natural or human-induced activities (such as public utility works). Such security is also been provided over sea settings, forecasting oceanic currents and estimating sea ice concentrations and drift.

Public Health

Protecting and improving the health of people has been a priority like never before since the rise of the Covid-19 pandemic in early 2020. Public health can be achieved not only through healthcare and assistance but also through research for new drugs and preventing and responding to infectious diseases or other risk factors. In this frame, satellite data provide useful information to support public health policies, especially in relation to air quality and respiratory diseases, as poor air quality continues to prematurely claim the lives of millions of people every year. It is possible to use these data, in fact, to track the range of trace gases that affect air quality such as carbon monoxide, nitrogen dioxide and ozone, forecasting air pollutants, greenhouse gases and small particles such as dust, smoke and pollen, ozone concentrations and UV radiation harmful for skin and eyes. It is furthermore possible to exploit these data to design cooler, more comfortable cities by delineating urban areas affected by severe heatwaves or identify toxic algal blooms that could potentially hit coastal areas and affect human activities such as bathing and fish farming. Finally, these data can also support the identification of areas prone to the emergence and spread of vector-borne epidemics, such as malaria, which greatly depend on environmental factors such as water, sanitation, food or air quality.

1.4 Scope of the Thesis

Among all areas of application of satellite data above, agriculture and the food production industry are surely among the most important as well as reference sectors for this study. These sectors are increasingly subject to various threats linked to anthropogenic pressure such as global warming and intensive exploitation practices. As the population increases and climatic patterns change, in fact, so does the spatial distribution of ecological zones, ecosystems, plant diseases and pests, with significant impacts on food production. These fields are not only crucial since they form the basis of food supply but also since they constitute relevant economic sectors. Satellite data not only help to monitor the health status of crops allowing for sustainable food production by reducing water waste, fertilizer waste and maximizing yields but can also support the setup of more efficient and environment-friendly agricultural practices for public authorities, farmers and other companies (such as insurance companies) alike replacing, for example, on-farm checks for determining governmental subsidies amount or insurance costs.

The scope of the thesis is to extract knowledge from Sentinel-2 multispectral data and translate it into the information of land cover and crop type, important insights making it possible to remotely identify many characteristics of the constituents element of a selected area allowing its management in a way completely tailored on its specifics. Land cover, although often confused with land use, refers to the identification of the physical material covering a given portion of Earth's surface, such as artificial material, cropland, woodland, shrubland, grassland, bare soil, water body and more. Its study and identification constitute fundamental information not only in planning, management and monitoring programs at local, regional and national levels providing a better understanding of land utilization aspects and guiding the formation of policies and programs required for development planning but also for change detection analysis and thematic mapping which result in environmental assessments especially crucial given the pace and extent of land cover change across the globe and worldwide concern for issues such as global warming [19][20].

Crop classification is the first step toward crop mapping and monitoring, activities that play and will play a fundamental role as mankind will progress into a new way of making agriculture. As previously stated the agricultural sector is crucial from many points of view, being strongly connected to food security, to the economy, politics and the environment. For this reason, even minor innovations in both processes and technologies linked to the sector might have a huge impact on societies, providing food for billions of people and saving billions of dollars optimizing operations and cutting costs. Crop monitoring allows the possibility of forecasting crop yields that, at a governmental level, is essential for determining how much food can be stored or exported and for assessing food losses along the supply

chain, while at a local level, provides useful information in various decision-making processes for managing resources with precision, limiting water wastes, limiting chemical pesticides and fertilizers, moving toward a maximized and sustainable production able to satisfy the demand of the 7.9 billion people of today as well as the 9.7 billion people expected by 2050.

1.5 Spectral Indices

In order to achieve the goals mentioned above, it is important to understand the distributions of both land cover classes as well as vegetation types and their biophysical and structural properties in relation to spatial and temporal variations. To do that one useful conceptual tool is represented by spectral indices. While it is true that both the tasks of land cover and crop type classification could be carried out using the raw values of the multispectral bands, however, spectral indices are preferred since these are quantities able to enclose higher-level information. Spectral indices are, in fact, spectral transformations of two or more multispectral bands designed to enhance the contribution of specific compounds or features allowing reliable spatial and temporal comparisons of Earth surface areas and relative land cover material. Features that can be extracted using spectral indices range from vegetation (highlighting, among many aspects, photosynthetic activity), to geologic and artificial features (identifying for example high contrasts between nearby areas), or related to burned areas, snow-covered areas, and many others. Being such indices a simple transformation of spectral bands, they are computed directly without the need for any assumption regarding land cover class, soil type, or climatic conditions [21].

The computation of such indices is possible since the nature of multispectral data, as previously stated, is simply quantitative. As for color pictures, where each pixel has three values representing brightness in each of the three spectral wavelengths of red, green and blue (RGB), the same happens for multispectral data representing, in fact, the physical measurement of reflectance response to electromagnetic radiation for each of the available wavelength windows, called spectral bands, in the multispectral instrument. The resulting data is, therefore, very similar to a colored picture where, however, each pixel has more than three channels, specifically one for each spectral band. Analyzing multispectral data by means of calculated indices instead of using directly spectral bands not only has the advantage of emphasizing specific features or phenomena within remotely sensed imagery and extracting, therefore, meaningful information, but also reducing the dimensionality of multispectral data, resulting in overall easier interpretations. The most common mathematical formulas used for computing indices are ratios and

normalized differences:

$$Index = \frac{Bx}{By}$$

$$Index = \frac{Bx - By}{Bx + By}$$

These types of formulas are very useful to enhance spectral features and minimize as much as possible the effects of illumination and more importantly shadows.

It is fair to point out, however, that there is no general mathematical expression from which to derive all spectral indices due to the complex reproducibility of the task given by instrumentation, platforms, and resolutions. For this reason, ad hoc formulas have been developed and empirically tested against a variety of applications according to specific mathematical expressions that combine visible light radiation to obtain proxy quantifications of the measure of interest [22].

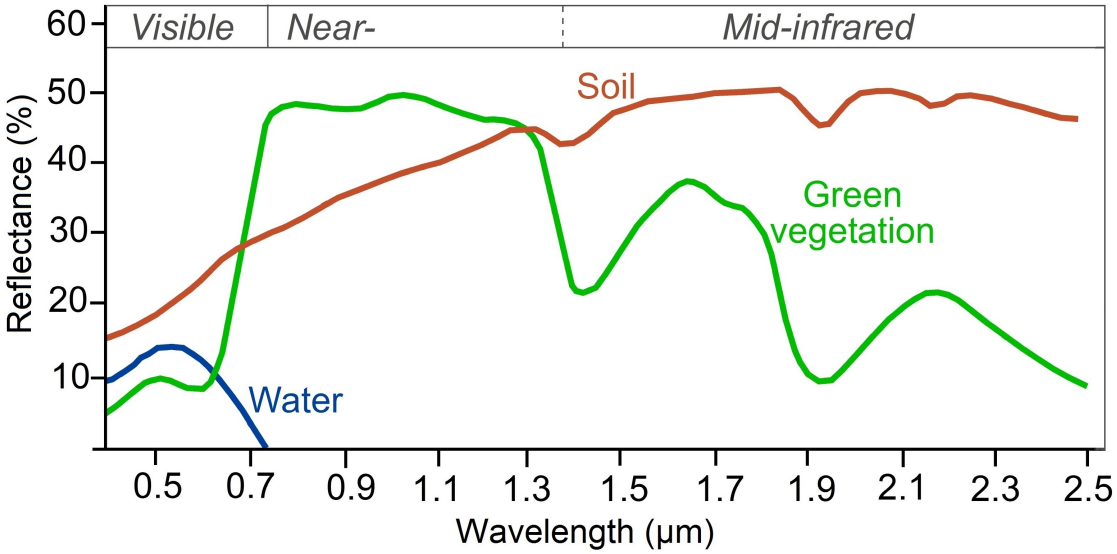

Figure 1.7: Spectral signatures. Source: [23]

To make an example it is possible to use one of the most famous vegetation indices called Normalized Difference Vegetation Index (NDVI), quantity ranging from -1 to 1 used to assess greenness and health of biomass as well as to distinguish between vegetation and other types of surface, consisting basically in a normalized difference NIR/RED that is $NDVI = (NIR - RED)/(NIR + RED)$. By considering the vegetation reflectance shown in Fig 1.7 it is possible to see that the response is low in the red region (around 0.7 micrometers) but high in the NIR. Based on this, the result of the index for a densely vegetated area would be two positive values for both the numerator and the denominator, leading to an overall NDVI value close to 1, while the same formula computed over bare soil, where red reflects about equally

to NIR, the numerator will have a value close to 0 and, therefore, the overall result would be around 0, while, finally, water surfaces reflect higher in the red than NIR, leading to an NDVI value close to -1.

Vegetation properties measured with spectral indices can generally be divided into three main categories: structure, biochemistry and physiology. Structure indices have the objective of measuring properties such as fractional cover, green biomass, leaf area index (LAI), biomass senescence, and fraction absorbed photosynthetically-active radiation (FAPAR). The objective of biochemistry indices is to measure biochemical properties including water, pigments that among the most important ones count chlorophyll (the most critical plant pigment due to its fundamental role in photosynthesis and primary production), anthocyanins (plant pigments that increase in response to environmental stress and play a role in minimizing photo-inhibition) and carotenoids (pigment aiding in the process of light-harvesting for photosynthesis and protect chlorophyll from photo-oxidation via the reversible conversion of the xanthophyll violaxanthin to zeaxanthin). Also among biochemistry indices, some are particularly useful in measuring nitrogen, important components of many light-absorbing compounds in the visible to SWIR range. Finally, physiology indices, whose objective is to measure stress-induced changes in xanthophyll cycle pigments, chlorophyll content, fluorescence and leaf moisture, make especially use of red edge wavelengths. Changes in leaf physiology and stress, in fact, impact the position and shape of the red edge, shifting it toward either shorter wavelengths (blue shift) or longer wavelengths (red shift). Blue shifts have been observed in response to heavy metal stress in plants while red shifts typically occur during chlorophyll development and nutrient stress [24].

Limitations

While someone might be tempted to use spectral indices as a universal tool for any remote sensing application, it is fair to point out that even though their potential is in fact considerable, so are their limitations. To start, differently from what one may think, there is no unique signature for a given surface especially in the case of vegetation. This happens for many reasons among which the interference of soil on leaf reflectance, approximate atmospheric corrections and more. In the case of vegetation, for example, the signatures vary across latitudes, plant phenology, plant pathology and even internal factors of the plant like water content or other parameters for the same crop type [25]. This causes the impossibility of building a universal spectral signature dataset on which to train models and deploy solutions effortlessly, but instead, as it will happen in this study, the only viable option for a dataset is to build it enforcing location constraints. A further consequence of the variable nature of spectral signatures is the impossibility of making assumptions by analyzing the value of a given area in a single moment in time. What is, in fact,

a normal value for an area in one location might not be the same for an area with the same land cover in a second location. For this reason, assumptions can only be made by analyzing the time series of the given index over the area of interest. Finally, most of the spectral indices have been firstly theorized for broad-band systems and have been later approximated with multispectral equivalents.

Chapter 2

Dataset Creation and Models

As previously stated, there is no unique signature for a given surface but instead, the response varies with change in many factors. In the case of vegetation, the response may change as location, phenology, pathology and other internal factors of the plant, such as water content, change [25]. This explains the lack of a universal spectral signature dataset on which to train models in a supervised way. There is no certainty, in fact, that using a dataset built with data coming from a specific location would give accurate results when used for another one. For this reason, two paths could be undertaken, the first, consisting in reaching a classification of land cover and crop type by means, in an early phase, of unsupervised techniques such as clustering, while the second, consisting in building from scratch a dataset associated to the specific area of interest. Although it seems like there is room to ponder pros and cons and to weigh every possibility, what looks like a choice unfortunately is not really one. In both cases, in fact, there would be the need for validating the results and assess the quality of processing and models. This would mean that in both cases there would be the need to have labeled data to test on, and since an effort has to be made in order to gather and label data, then it is reasonable to make a slightly bigger effort to collect more data in order to build a dataset and be able to train, validate and test models on it.

2.1 Land use/cover area frame statistical survey

In order to build the dataset for both the tasks of land cover and crop type classification, the fundamental information required is the one associating a location, whether identified by a polygon containing the area having coordinates points as vertices or by just a coordinates point in a specified coordinate system, to the type

of cultivation, in case of agricultural field or type of surface for everything else. Unfortunately making such mapping in person would mean investing not only a non-negligible amount of time planning a way to sample the region of interest in order for the dataset to be as representative as possible, and consequently wandering around with a GPS device, but also investing economic resources in terms of fuel and probably in term of an expert able to distinguish between crops. This would be unfeasible in this preliminary phase where the interest is to obtain a proof of concept, having in mind the considerable amount of instances required for such a dataset.

There is, therefore, the need to link the geolocation of an area to the relative surface cover without having to personally inspect the area. One possible solution could be retrieving these pieces of information from archives of local authorities by means of the Land-parcel identification system (LPIS) which is a governmental system linking land use to each parcel identified by a unique number for a given country. Retrieving documents from local authorities is however a long process, particularly so if the digitalization level of the country of interest lags behind. The real solution is, therefore, provided by the European Union, and specifically by the European Statistical Office (Eurostat), which is a department located in Luxembourg responsible to provide statistical information to EU institutions, in the form of the Land use/cover area frame statistical survey (LUCAS) of 2018. LUCAS, initially developed across a limited number of EU Member States to provide early crop estimates for the European Commission, launched as a pilot in 2001 following Decision 1445/2000/EC of 22nd, May 2000 on the application of aerial-survey and remote-sensing techniques to agricultural statistics, little by little established itself over time as a key tool for policymakers and statisticians due to both the increasing amounts of data as well as their growing variety. Just 5 years after the first survey, in 2006, in fact, the focus of the sampling methodology shifted from an agricultural land survey to a broader land cover, land use and landscape survey, allowing for more extensive studies and statistics. The next step for LUCAS was to expand its geographical coverage, reaching up to 23 of the then 27 EU Member States (Bulgaria, Cyprus, Malta and Romania were not covered) in 2009 and completing it in 2012 reaching all 27 Member States. Nowadays LUCAS is defined as an "in situ" survey program (data are, in fact, also gathered through direct observations made by surveyors on the ground as well as sensors) that extends over the whole EU's territory and whose objective is to build a consistent framework for coherent sampling plan, classifications and data collection processes to provide harmonized and unbiased statistics on land cover and land use for agriculture, environment and landscapes in the European Union, useful for the definition and evaluation of common European agricultural, environment and sustainable development policies, as well as ground evidence for satellite images calibration.

LUCAS is composed of two sampling phases, the first, called Master or Frame,

consists in the remotely interpretation and assignment of a land cover class among a pool of predefined classes to 1.1 million points forming the intersections of a 2 km spaced virtual grid covering the whole of the EU's territory, while the second phase consists in the selection of a stratified sub-sample of points from the first phase points by means of an iterative algorithm in order to minimize sampling errors, on which to proceed for on-ground assessment carried out personally by a surveyor. Excluded from the second phase subset are points above 1500 meters or far from the road network, therefore considered inaccessible, in order to limit the cost of data collection effort. The bias for the exclusion of such points from the field assessment phase is, however, compensated by "in office" interpretation and classification from images with the further help of regression models also taking into account data from previous surveys.

Land Cover Class	Code
Artificial land	A
Cropland	B
Woodland	C
Shrubland	D
Grassland	E
Bareland	F
Water	G
Wetlands	H

Table 2.1: LUCAS land cover classes

The survey design has been fine-tuned in several aspects over the years, going from seven original land cover categories (arable land, permanent crops, grassland, wooded areas and shrubland, bare land, artificial land, and water) to the current eight, made to be comparable with other statistical standards such as EU's farm structure survey (FSS) and many more, listed in table 2.1, and from 273.500 points (of which 67.000 points interpreted "in office") visited by 750 field surveyors in 2015 to 337.854 points (of which 99.777 points interpreted "in office") in 2018 [26].

Furthermore, a fundamental fact for the task of crop classification, each land cover category is made of several classes and subclasses of which it is possible to see the ones associated with Cropland land cover category in table 2.2. It is important to notice that points might be labeled BX1 and BX2, two classes that do not take place among the ones in table 2.2. Points are, in fact, assigned to such classes when the crop associated to the given LUCAS point is not recognizable from the "in-office" image classification, specifically BX1 is associated with "Temporary crops" covering the classes from B11 to B55 and BX2 to "Permanent crops" covering the classes

from B71 to B84.

2.2 Spectral Indices Selection

Once found the information associating coordinates to types of surface, what is left in order to build the dataset is firstly the determination of spatial and temporal boundaries and finally the selection of spectral indices derived from the spectral bands to populate the dataset's feature space, and therefore describe the behavior of the points, for both the tasks of land cover and crop type classification. Since the study is linked to possible commercial use for Italian customers the study area is set to Italy, which contains 33442 points from LUCAS 2018 usable for land cover classification, 8418 of which belongs to the land cover class of Cropland and can therefore be used for crop type classification by means of their subclasses. The temporal scope, following the study [27] (which will be later resumed to implement the neural network it proposes) consists of one picture per month from July 2018 to July 2019 (both included), excluding the months of November, December and January, for a total of 10 Sentinel-2 products for a given area. More specifically Sentinel-2 has a 290 km field of view when capturing its products that are then projected onto a UTM grid and made available publicly on 100x100 km^2 tiles. Italy is covered by roughly 50 Sentinel-2 UTM tiles ("roughly" because tiles with land content much lower than sea content have been discarded to optimize the process. Including, in fact, for example, the tile 33STV covering Lampedusa, would have meant requesting, downloading and processing more than 50 GB of Sentinel-2 products just to obtain a fistful of additional LUCAS points). From these 50 tiles, have been requested and downloaded from Copernicus SciHub, for each of the aforementioned months, the product with lower cloud coverage of the month, where cloud coverage is a parameter of the percentage of the product obscured by clouds (since the wavelengths of Sentinel-2 are not able to go through clouds), for a total of 500 Sentinel-2 products corresponding to around 650 GB. Finally, the dataset is populated with the values that the spectral indices, described below, assume in the given months for each LUCAS point. With the bands provided by Sentinel-2, it is possible to compute up to 200+ spectral indices, in this study, however, the ones in use are the ones that more frequently appear in papers and researches and therefore the ones having a more solid background.

Family	Crop Class	Code
Cereals (B1X)	Common wheat	B11
	Durum wheat	B12
	Barley	B13
	Rye	B14
	Oats	B15
	Maize	B16
	Rice	B17
	Triticale	B18
	Other cereals	B19
Root Crops (B2X)	Potatoes	B21
	Sugar beet	B22
	Other root crops	B23
Non-Permanent Industrial Crops (B3X)	Sunflower	B31
	Rape and turnip rape	B32
	Soya	B33
	Cotton	B34
	Other fibre and oleaginous crops	B35
	Tobacco	B36
	Other non-permanent industrial crops	B37
Dry Pulses, Vegetables, Flowers (B4X)	Dry pulses	B41
	Tomatoes	B42
	Other fresh vegetables	B43
	Floriculture and ornamental plants	B44
	Strawberries	B45
Fodder Crops (B5X)	Clovers	B51
	Lucerne	B52
	Other leguminous and mixtures for fodder	B53
	Mixed cereals for fodder	B54
	Temporary grasslands	B55
Permanent Crops: Fruit Trees (B7X)	Apple fruit	B71
	Pear fruit	B72
	Cherry fruit	B73
	Nuts trees	B74
	Other fruit trees and berries	B75
	Oranges	B76
	Other citrus fruit	B77
Other Permanent Crops (B8X)	Olive groves	B81
	Vineyards	B82
	Nurseries	B83
	Permanent industrial crops	B84

Table 2.2: LUCAS crop classes

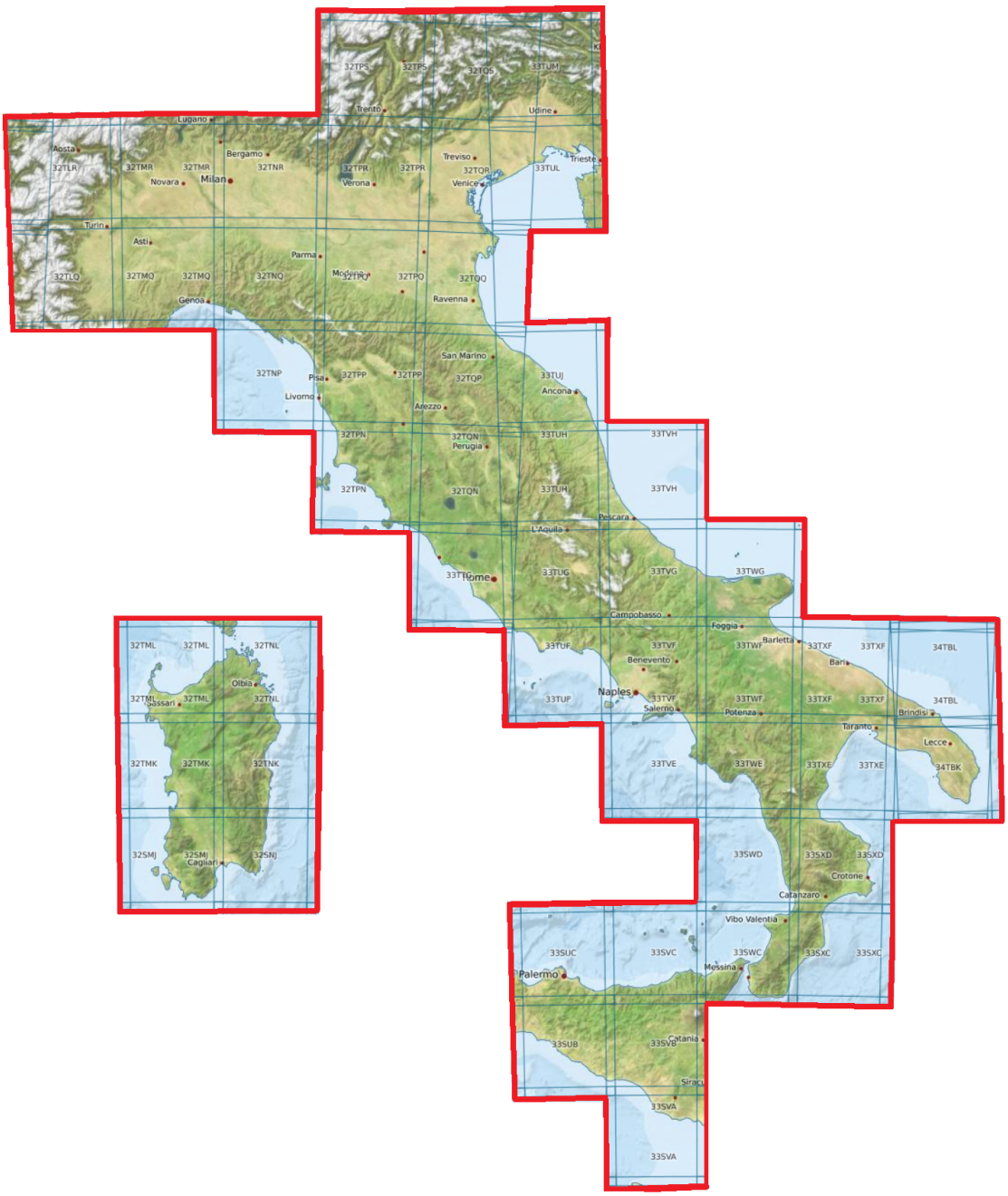

Figure 2.1: Tiles considered for the dataset

2.2.1 Land Cover Indices

Forest Index (FI)

$$FI = (\frac{NIR - RED - L}{NIR + RED}) * (\frac{c_1 - NIR}{c_2 + GREEN})$$

FI, proposed in [28] is designed to highlight forest vegetation by assuming positive and high values for pixels associated with forests and low values for pixels associated with non-forest areas. As forest is a kind of vegetation, it is generally easy to distinguish it from non-vegetation surfaces with the help of any vegetation index, while it is rather difficult to distinguish forest from non-forest vegetation, however, FI solves both the two aspects. To identify forests, non-vegetation is firstly highlighted by a kind of vegetation index in the form of the multiplicand, then, according to the spectral difference introduced by the multiplier, forest is discriminated from non-forest vegetation. The idea behind the multiplier is based on the observation that the reflectance of forest is usually lower than other vegetation in the visible and shortwave infrared bands. L is a soil adjustment parameter as found in many other vegetation indices, empirically set to 0.01, while c_1 and c_2 are empirical parameters used to scale the function, empirically set to 1 and 0.1, respectively.

Modified Normalized Difference Water Index (MNDWI)

$$MNDWI = \frac{GREEN - SWIR}{GREEN + SWIR}$$

MNDWI is an NDWI [29] derived index and as such is suitable to monitor changes related to water content in water bodies. Differently from NDWI, which is sensitive to built-up land and can result in over-estimation of water bodies, however, the index is able to enhance open water features suppressing the influence from a background dominated by built-up land areas and vegetation [30]. Open water has, in fact, greater positive values in MNDWI with respect to NDWI as it absorbs more SWIR light than NIR light as used in the latter index. Built-up land has usually negative values in the SWIR, while soil and vegetation will still have negative values as soil reflects SWIR light more than NIR light and the vegetation reflects SWIR light still more than green light. Consequently, compared to NDWI, the contrast between water and built-up land of the MNDWI will be considerably enlarged thanks to increasing values of water feature and decreasing values of built-up land from positive down to negative.

Normalized Difference Built-Up Index (NDBI)

$$NDBI = \frac{SWIR - NIR}{SWIR + NIR}$$

NDBI is a suitable index to highlight bare soil, urban and built-up areas since, differently from vegetation where the reflection of NIR is higher than SWIR, there is an opposite response for the same bands for the aforementioned surfaces, that means higher reflectance in SWIR with respect to NIR [31].

Normalized Built-up Area Index (NBAI)

$$NBAI = \frac{SWIR3 - \frac{SWIR2}{GREEN}}{SWIR3 + \frac{SWIR2}{GREEN}}$$

The index proposed in [32] claims not only to perform better than NDBI in distinguishing between built-up areas with respect to vegetation and water, but also improving the accuracy in highlighting differences between built-up areas and bare soil.

2.2.2 Crop Classification Indices

Normalized Difference Vegetation Index (NDVI)

$$NDVI = \frac{NIR - RED}{NIR + RED}$$

As shown in the previous chapter, NDVI is one of the simplest and most used indices. Sensitive to the effects of foliage chlorophyll concentration, canopy leaf area, foliage clumping and canopy architecture, it measures the quantity and vigor of green vegetation and more specifically the overall amount of photosynthetic material, essential for the vital functions of the plant, comparing reflectance measurements from the near-infrared bandwidth, which has much greater penetration depth through the canopy, to the ones taken in the red window which is where chlorophyll absorbs photons to store into energy through photosynthesis. It is also correlated with the vegetation parameter of fractional absorption of photosynthetically active radiation (fAPAR) [33].

Modified Soil Adjusted Vegetation Index 2 (MSAVI2)

$$MSAVI2 = \frac{2 * NIR + 1 - (\sqrt{(2 * NIR + 1)^2 - 8 * (NIR - RED)})}{2}$$

MSAVI2 is a simpler version of the MSAVI proposed in [34], which is an improvement of the Soil Adjusted Vegetation Index (SAVI). While NDVI is highly sensitive to soil color, soil moisture and prone to saturation effects from high-density vegetation, SAVI is a much more stable index thanks to the suppression of the effects of soil pixels by means of a canopy background adjustment factor, L, which is a function of vegetation density and often requires prior knowledge of vegetation

amounts. MSAVI, however, not only reduces soil noise and increases the dynamic range of the vegetation signal but is also based on an inductive method that does not use a constant L value to highlight healthy vegetation [35].

Structure Intensive Pigment Index (SIPI)

$$SIPI = \frac{NIR - COASTAL}{NIR + RED}$$

SIPI, one of the biochemistry spectral indices related to vegetation stress and light use efficiency, provides a measure of the efficiency of the vital functions of the plant and its stress level, aspects related to carbon uptake efficiency and growth rate. The index takes advantage of relationships between different pigment types to assess the overall status of the vegetation. In particular, SIPI, whose increase indicate increased canopy stress, is designed to be sensitive to the ratio of bulk carotenoids (such as alpha-carotene and beta-carotene) to chlorophyll without being influenced by variations in canopy structure [36].

Carotenoid Reflectance Index (CRI)

$$CRI = \frac{1}{BLUE} - \frac{1}{GREEN}$$

As stated above, carotenoids are not only strongly related to the efficiency of light absorption processes in plants but are also fundamental in protecting plants from the harmful effects of too much light. Weak vegetation, in fact, contains higher concentrations of carotenoids. CRI comes, therefore, from the same category of SIPI as high CRI values, obtained exploiting reflectance measurements in the visible spectrum, mean high carotenoid concentration with respect to chlorophyll [37].

Normalized Difference Moisture Index (NDMI)

$$NDMI = \frac{NIR - SWIR}{NIR + SWIR}$$

NDMI is a spectral index that provides a measure of the amount of water contained in the foliage canopy, a parameter particularly important since higher water content indicates healthier vegetation that is likely to grow faster and be more fire-resistant. To do that the index uses the SWIR band, which, being negatively related to water content, is sensible to its variation at mesophyll (internal leaf structure) level in vegetation canopies, and the NIR band, which penetrates deeper and whose reflectance is affected by leaf internal structure and leaf dry matter content but not by water content. The combination of the two bands removes

variations induced by leaf internal structure and leaf dry matter content, improving the accuracy in retrieving the vegetation's total column water content [38].

Anthocyanin reflectance index (ARI)

$$ARI = \frac{1}{GREEN} - \frac{1}{R.EDGE}$$

Anthocyanins, water-soluble pigments abundant in newly forming leaves and those undergoing senescence, common in higher plants causing red, blue and purple coloration, provide valuable information about the physiological status of plants. They are considered indicators of various types of plant stresses as their concentration is higher in weakening vegetation. Anthocyanins reflectance is highest around 550nm which corresponds however to the wavelengths reflected by chlorophyll. To isolate anthocyanins, the 700nm spectral band only related to chlorophyll is subtracted. ARI, by means of reflectance measurements in the visible spectrum, is sensible to anthocyanins and it is, therefore, suitable to sense canopy changes in foliage via new growth or death and canopy stress [39].

Modified Chlorophyll Absorption Reflectance Index (MCARI)

$$MCARI = ((R.EDGE - RED) - 0.2 * (R.EDGE - GREEN)) * \frac{R.EDGE}{RED}$$

MCARI, one of several indices derived from Chlorophyll Absorption Reflectance Index (CARI) which measures the relative abundance of chlorophyll of a given plant, not only measures the depth of chlorophyll absorption and is very sensitive to variations in chlorophyll concentrations, but it also extends CARI being more resilient to the combined effects of illumination conditions, background reflectance from soil and other non-photosynthetic materials observed [40].

Canopy Chlorophyll Content Index (CCCI)

$$CCCI = \frac{\frac{NIR - R.EDGE}{NIR + R.EDGE}}{\frac{NIR - RED}{NIR - RED}}$$

Nitrogen is one of the most vital fertilizer components in agriculture as it directly affects the amount of chlorophyll in plants. Under the condition of nitrogen malnourishment the plant growth process is disturbed, chlorophyll development stops, and finally, the leaves begin to turn yellow. In order to survive the plant takes nitrogen from older leaves and transfers it to new ones, thus lower-level leaves show an indication of nitrogen starvation. CCCI, derived from the Normalized Difference Vegetation Index (NDVI) and Normalized Difference Red Edge (NDRE), analyzes the amount of chlorophyll in vegetation, thereby allowing detection of nitrogen starvation [41].

RedEdgeNDVI

$$RedEdgeNDVI = \frac{R.EDGE2 - R.EDGE1}{R.EDGE2 + R.EDGE1}$$

This index is a modification of the NDVI index introduced above, differing from the usage of bands along the red edge, instead of the main absorption and reflectance peaks, shifting its focus on the vegetation red edge to small changes in canopy foliage content, gap fraction, and senescence. It is strongly correlated to vegetation stress [42].

Red-Green Ratio (RGR)

$$RGR = \frac{RED}{GREEN}$$

RGR is a useful index as a measure of foliage development, leaf or flower production, stress. The ratio measures the relative expression of leaf redness caused by anthocyanin to the one of chlorophyll [43].

Red Edge Position (REP)

$$REP = 700 + 40 * \frac{(\frac{RED+R.EDGE3}{2} - R.EDGE1)}{(R.EDGE2 - R.EDGE1)}$$

The red edge position refers to the wavelength of the steepest slope within the range of 690nm to 740nm, where the common range for green vegetation is between 700nm and 730nm. Such position moves to longer wavelengths as chlorophyll concentration rises. REP is a reflectance measurement sensitive to changes in chlorophyll concentration that estimates the red edge position [44].

2.3 Models Selection and Training

The dataset created this way, with rows corresponding to LUCAS points and columns corresponding to the aforementioned indices computed for each of the 10 months, is tabular, reason why among the selected machine learning models, along with the well known Random Forest and SVM, there are models proved to be exceptionally performing on tabular data such as Extreme Gradient Boosting Machine and Light Gradient Boosting Machine. Additionally, deep learning approaches are taken into account both to test new concepts proposed in recent papers trying to fix the fame of neural networks of not being the preferred models when it comes to tabular data, as well as to manage and account for the temporal component of the dataset.

2.3.1 Random Forest

Random Forests is a supervised learning classification method, applied to decision tree model, based on the concept of bootstrap aggregation, a simple yet powerful ensemble method consisting in the combination of the predictions from multiple learners to achieve an accuracy higher than the one achieved by any of the single learners. Since decision trees are sensitive to the data they are trained on, all the trees trained on the same training data would end up being exactly equal in all aspects to each other, in order to avoid having equal trees which would make the aggregation useless, the trees are instead trained on different data, more specifically on different subsets of the training data drawn with replacement. Additionally, only a random subset of all the features, generally as big as the squared root of the total number of features, is considered to subdivide nodes in each decision tree. This way, although each tree may present a high variance with respect to a particular set of training data, overall the entire forest will have a lower variance thus achieving higher accuracy when predictions, obtained with a simple majority voting system among all trees, are finally made.

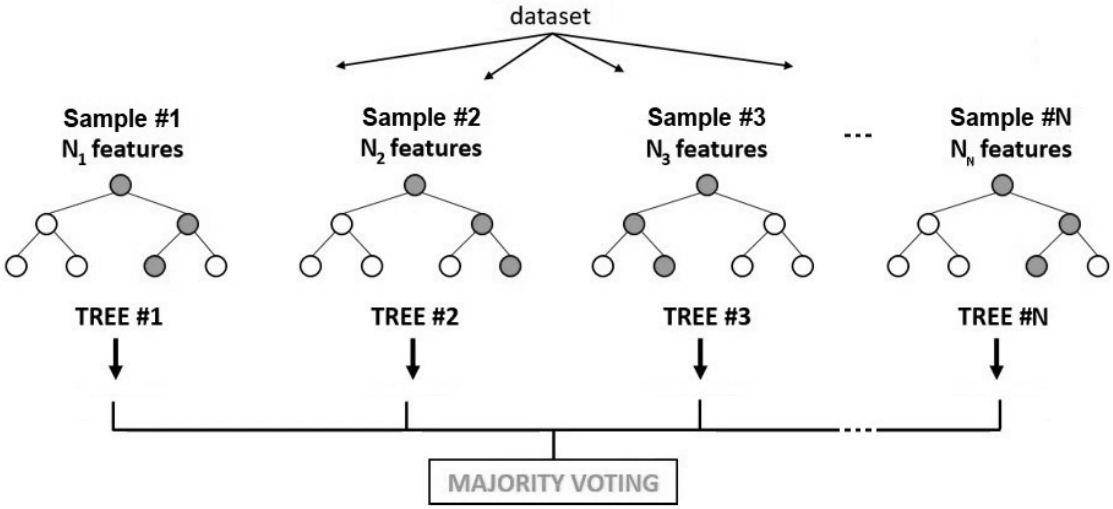

Figure 2.2: Random Forest

2.3.2 Support Vector Machine

SVM, popular in applications such as natural language processing, speech recognition and computer vision, is a supervised machine learning algorithm, proposed by Vapnik in 1963, based on the idea of finding hyperplanes that best divide the training data into classes, that means finding hyperplanes that maximize the distance between support vectors and the hyperplanes themselves. In the case of

a classification task with only two spatial dimensions, as it is possible to see in figure 2.3, for example, the hyperplane takes the form of a line that best divides data. In three dimensions the hyperplane would have the form of a plane and so on. Are called support vectors the data points closest to the hyperplane, according to which the optimal hyperplane is computed, removing or modifying these points would alter the position of the dividing hyperplane. The margin is defined as the distance between the closest support vectors of different classes to the hyperplane. Once found the best hyperplane in the training phase, when new data is submitted for prediction, the model decides the class based on the position of the data point in the space with respect to the hyperplane.

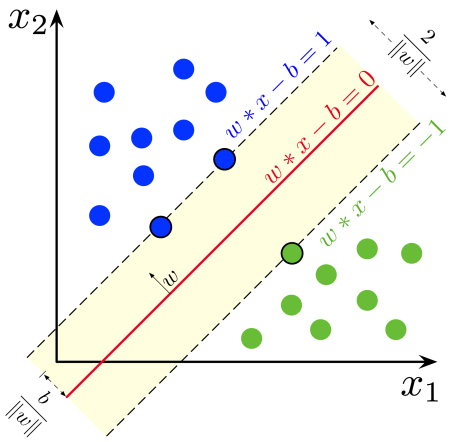

Figure 2.3: Example of hyperplane

The SVM described above does not only works when the classes are linearly separable and that is, for the case of binary classification, when exists a linear hyperplane such that all the data points of a class are on one side and all the data points from the other class are on the other side. If a dividing linear hyperplane does not exist in fact, SVM can allow for a certain degree of misclassification with the introduction of a slack variable or, by means of a non-linear mapping known as kernel trick, can transform the training data into a higher dimension to achieve separability. SVM works also for tasks of multi-class classification using the same principle described above breaking down the problem into multiple binary classification problems. This can happen by exploiting a One-to-One approach or a One-to-Rest approach. In the first approach, the algorithm looks for the hyperplanes dividing every combination of couples of classes, neglecting the points belonging to third classes, while in the second approach the algorithm looks for the hyperplanes dividing the points from each class from the union of points belonging to all the other classes. This means that in this case, each attempt of finding any hyperplane takes all points into account, dividing them into two groups, the first

group for given class points while the second group for all other points.

2.3.3 Extreme Gradient Boosting Machine

Extreme Gradient Boosting Machine, also known as XGBoost, is an implementation of the Gradient Boosting method that has been for many years the state of the art when it came to tabular structured data thanks to its more accurate approximations to find the best tree model. Among its peculiarities, there is the computation of the second-order gradients which consist of the partial secondary derivatives of the loss function and provide more information about the direction of the gradients to get to the minimum of the loss function, and the usage of regularization which improves the generalization of the model. One further advantage is that the training is fast and can be distributed between different machines. The XGBoost formation is an iterative procedure that calculates at each step the best possible subdivision for the k^{th} tree, listing all the possible structures still available at that point of the path. The first step of the XGBoost algorithm consists of performing an initial prediction with a default value and subsequently calculate residuals which are the difference between the predicted value and true values. After that, thanks to a similarity score and information gain, XGBoost builds a tree in the most accurate way by selecting the split with major gain computed as a function of the similarity scores of all leaves and roots from the trees derived by all possible splits of the training data. Starting from the found root, each leaf will be divided into additional leaf nodes until there is only one residue left or a specific depth is reached. Additionally, XGBoost uses a pruning technique, which consists of removing branches that make use of less important features, to reduce the complexity of the tree and consequently increase the accuracy and generalization capabilities of the trees. This is performed by subtracting a quantity γ to the gain value beginning from the bottom leaves up to the root, and removing the branches that start from nodes with negative gain value after the subtraction along with the node itself. Even if γ is set to zero, however, it is possible for the gain to be negative, in which case the node would be deleted according to the pruning technique anyway. Finally, the root node is deleted if and only if all the child nodes have negative gain values. The prediction is then adjourned, residuals are computed and the construction of a new tree, that will be combined with the existing ones by means of a quantity η learning rate, can then begin until the desired number of iterations is reached.

2.3.4 Light Gradient Boosting Machine

LightGBM is a fast, distributed and high-performance gradient-based framework implemented on top of decision tree algorithm, used in many machine learning tasks such as classification, regression and many others, released in January 2017

by Microsoft, whose aim is to maintain an accuracy comparable to the more famous XGBoost while reducing training time and memory occupation. Differently from XGBoost, which uses pre-sorting or histogram-based algorithms for computing the best splits, LightGBM uses instead a technique called Gradient-based One-Side Sampling (GOSS) to filter out data and find splits. While pre-sorting splitting consists of sorting data by feature value for each feature and using a linear scan to decide the best split along with that feature according to the gain quantity, and histogram-based algorithms consist of splitting all data into discrete bins and use these bins to find the split value, GOSS is faster than both methods in training time thanks to its use of the gradient as an indicator for the importance of instances. Gradient, in fact, represents the slope of the tangent of a loss function, so logically if the gradient of data points is large in some sense, these points are important for finding the optimal split point as they have higher error. GOSS keeps all the instances with large gradients and performs random sampling on the instances with small gradients, exploiting the assumption that samples with training instances with small gradients have smaller training error and it is already well-trained. Finally, in order to keep the same data distribution, when computing the gain, GOSS introduces a constant multiplier for the data instances with small gradients. Thus, GOSS achieves a good balance between reducing the number of data instances and keeping the accuracy high.

2.3.5 Deep Learning Models

Pixel R-CNN

In [27] is proposed a deep learning architecture called Pixel R-CNN, for land cover and crop classification, consisting of a Recurrent Neural Network (RNN) in combination with a Convolutional Neural Network (CNN) which join forces to first extract temporal correlations from time-series data, then to analyze and encapsulate patterns through convolutional filters. RNN is a powerful and robust type of neural network able to remember certain aspects of the received sequence data input, which is basically ordered data in which related things follow each other with the most famous being time-series data, which is just a series of data points that are listed in time order, allowing the achievement of more accurate predictions of what is coming next. Differently from feed-forward neural networks, where the information only moves in one direction, from input to output layer never flowing through the same node twice, and where therefore only the input at a given time is considered to obtain the output, making the architecture not ideal at predicting the next item of an input sequence, RNN is instead ideal for sequential data since when a prediction is to be made the network considers not only the current input but also what it has learned from the inputs it previously received. The RNN used in the paper is specifically a Long short-term memory

(LSTM) which is an extension of vanilla RNN, being able to remember inputs over a long period of time thanks to a computer-like memory allowing LSTM to read, write and delete information from its internal state. LSTM memory is much like a gated cell since it can decide whether to let new input in, delete the information because it is not important, or let it impact the output at the current timestep, based on the importance it assigns to it by means of weights, which are also learned by the algorithm during training time.

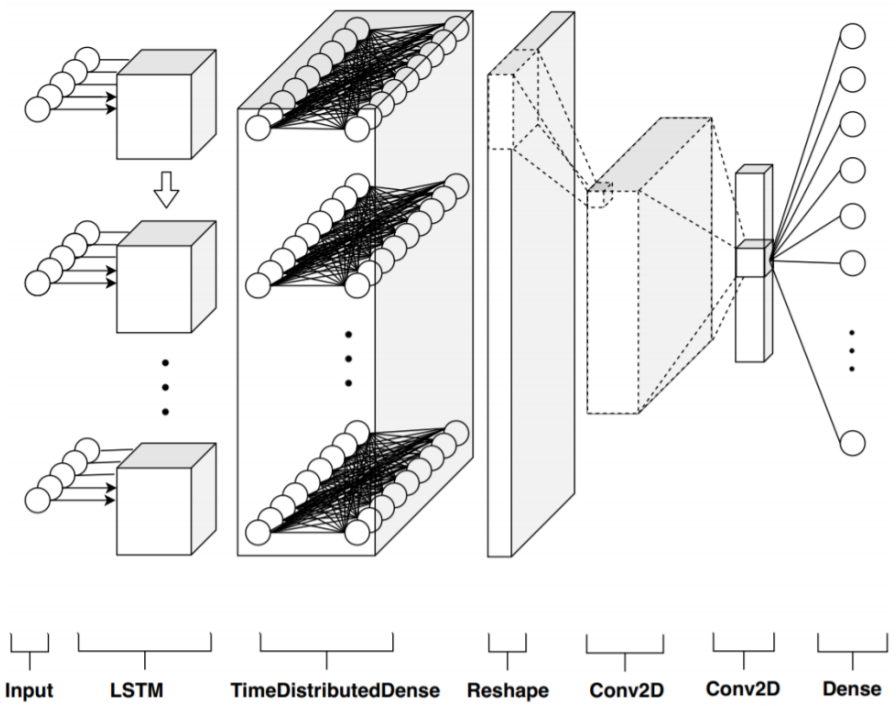

Figure 2.4: Pixel R-CNN Architecture

The three main tasks of Pixel R-CNN are:

- Temporal correlations extraction from multispectral temporal pixels exploiting a sequence-to-sequence recurrent neural network based on long short-term memory (LSTM) cells, followed by a time-distributed layer to compress and maintain a sequence structure, preserving multidimensionality exploiting temporal and spectral correlations simultaneously.
- Temporal pattern extraction where temporal sequences are processed by a subsequent cascade of convolutional filters, which in a hierarchical fashion, extracts essential features.
- Multiclass classification that maps the feature space with a probability distribution with K different probabilities, where K is equal to the number of

classes.

Neural Oblivious Decision Ensembles

Neural Oblivious Decision Ensembles (NODE) is an architecture proposed in [45] consisting of differentiable oblivious decision trees (ODT), which are regular trees of depth d constrained to use the same splitting feature and splitting threshold in all their internal nodes of the same depth, trained end-to-end by backpropagation. This constraint of using the same splitting feature and splitting threshold, on one hand, makes the ODTs weaker learners with respect to unconstrained decision trees, making them less prone to overfitting when used in an ensemble which is perfect for gradient boosting, while on the other, it allows the representation of ODTs as a table with 2^d entries, corresponding to all possible combinations of d splits which makes them very efficient during inference time since ODTs can compute d independent binary splits in parallel and return the appropriate table entry while unconstrained decision trees require evaluating d splits sequentially. One layer of the proposed architecture is composed of m differentiable oblivious decision trees of equal depth d which accept as input a common vector containing n numeric features. While in unconstrained decision trees, the feature choice to split a node by is deterministic, in ODTs, for differentiability reasons, to have a sparse feature selection for the split so the decision can be made on only a small number of features, α-entmax transformation proposed in [46] is used over a learnable feature selection matrix.

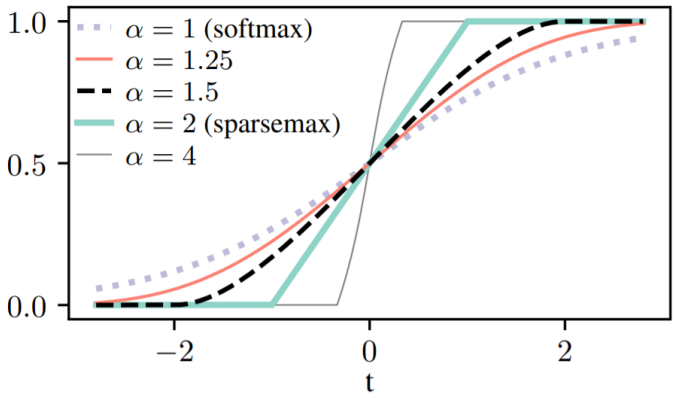

Figure 2.5: Illustration of entmax in the two-dimensional case

Additionally, to increase the learning capabilities of the model, it is possible to stack several NODE layers, each one on top of the other linking them with residual connections and giving to each layer as input features the concatenation of the input and the outputs of all previous layers, averaging all outputs for the final prediction.

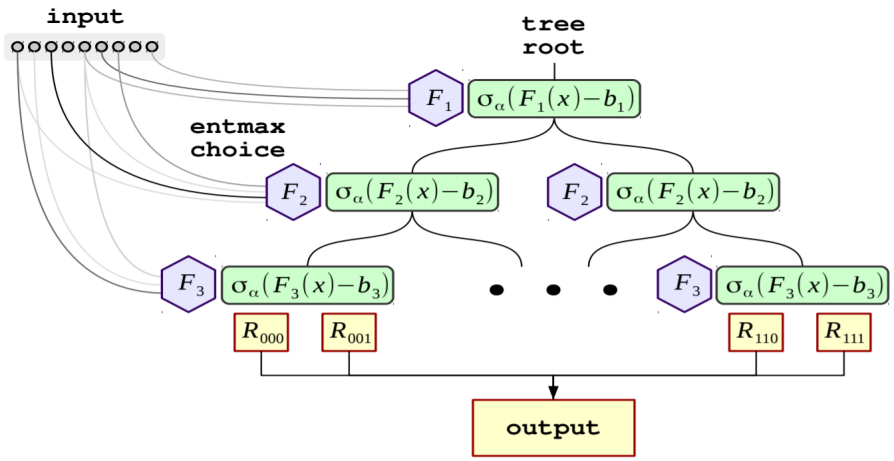

Figure 2.6: Single ODT inside the NODE layer. The splitting features and the splitting thresholds are shared across all the internal nodes of the same depth. The output is a sum of leaf responses scaled by the choice weights.

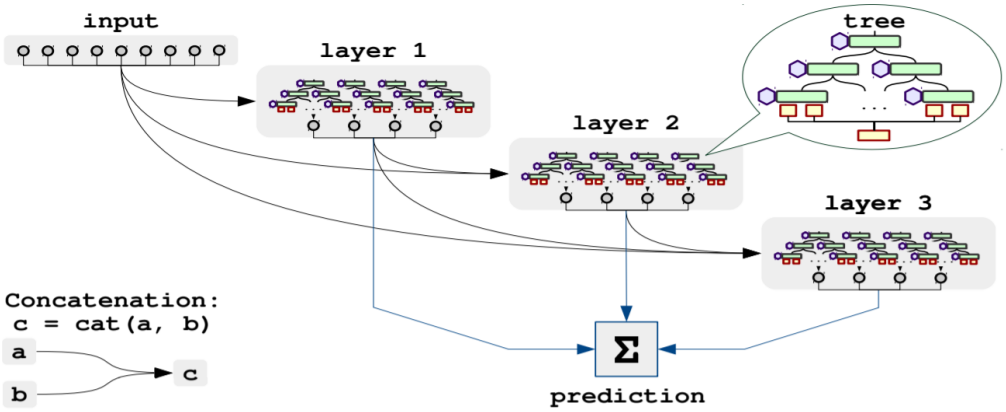

Figure 2.7: The NODE architecture, consisting of densely connected NODE layers. Each layer contains several trees whose outputs are concatenated and serve as input for the subsequent layer. The final prediction is obtained by averaging the outputs of all trees from all the layers

TabNet: Attentive Interpretable Tabular Learning

TabNet is a high-performance and interpretable deep learning architecture for tabular data, proposed in [47] by Arık and Pfister, that makes use of sequential attention to choose the features to be considered at each decision step, enabling interpretability and more efficient learning as the learning capacity is mostly used for salient features. The authors demonstrate that TabNet outperforms other models on a wide range of non-performance-saturated tabular datasets and yields

interpretable feature attributions.

Only raw numerical features are used, therefore, categorical features have to be mapped, in the specific case by means of trainable embeddings. TabNet's encoding is based on sequential multi-step processing with N_{steps} decision steps where each step receives the same D-dimensional features that are passed to all other steps. The i^{th} step inputs the processed information from the $(i-1)^{th}$ step to decide which features to use and outputs the processed feature representation to be aggregated into the overall decision.

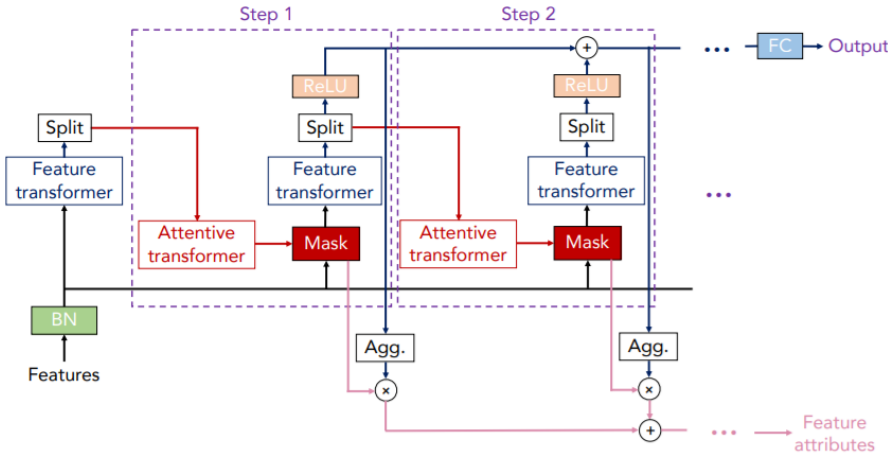

Figure 2.8: TabNet architecture for encoding tabular data

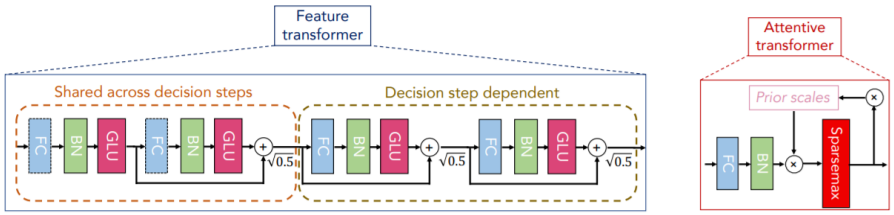

Figure 2.9: TabNet's feature transformer and attentive transformer

A learnable mask is obtained, by means of an attentive transformer, for a sparse selection of important features ensuring that the learning capacity of the given decision step is invested properly in relevant features only, making the model more parameter efficient. The filtered features are then processed using a feature transformer, consisting of layers that are shared across all decision steps as well as decision step-dependent layers for parameters efficiency and robust learning, and then split for the decision step output and information for the subsequent step. The feature transformer is therefore implemented as a concatenation of two shared layers and two decision step-dependent layers where each fully connected layer is

followed by batch normalization and the gated linear unit, eventually connected to a $\sqrt{0.5}$ normalized residual connection to stabilize learning by ensuring that the variance throughout the network does not change dramatically.

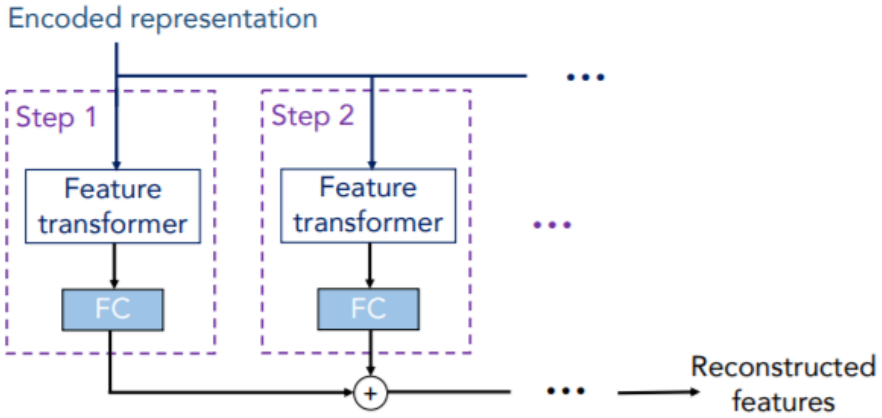

Figure 2.10: TabNet's decoder architecture

Chapter 3

Processing and Results

The datasets built following the methodology described in the previous chapter consists of 33442 points for the task of land cover classification, divided among classes as shown in table 3.1, while for the task of crop type classification the dataset is obtained starting from the one of land cover filtering all the points belonging to the Cropland class, for a total of 9543 points (BX1 and BX2 subclasses included), divided among classes as shown in table 3.2. At this stage, however, the datasets are not suitable to be given to models and proceed to the training and evaluation of performances. A processing phase is, in fact, required since the datasets present several aspects to be tackled in order to achieve the best possible results. There are, indeed, missing and "Inf" values resulting from indices computations and divisions between bands to be accounted for, outliers to be removed, the values are then to be normalized, the features selected and finally, as the two tables reporting the distributions of classes show, the dataset have to be balanced.

Land Cover Class	Code	Points
Woodland	C	13208
Cropland	B	9543
Grassland	E	5602
Artificial land	A	2827
Shrubland	D	1490
Bareland	F	417
Water	G	292
Wetlands	H	63

Table 3.1: Land cover dataset distribution

Family	Crop Class	Code	Points
Cereals (B1X)	Common wheat	B11	732
	Durum wheat	B12	1389
	Barley	B13	376
	Rye	B14	6
	Oats	B15	174
	Maize	B16	1079
	Rice	B17	126
	Triticale	B18	31
	Other cereals	B19	43
Root Crops (B2X)	Potatoes	B21	36
	Sugar beet	B22	25
	Other root crops	B23	29
Non-Permanent Industrial Crops (B3X)	Sunflower	B31	163
	Rape and turnip rape	B32	15
	Soya	B33	308
	Cotton	B34	0
	Other fibre and oleaginous crops	B35	14
	Tobacco	B36	17
	Other non-permanent industrial crops	B37	7
Dry Pulses, Vegetables, Flowers (B4X)	Dry pulses	B41	274
	Tomatoes	B42	87
	Other fresh vegetables	B43	208
	Floriculture and ornamental plants	B44	5
	Strawberries	B45	2
Fodder Crops (B5X)	Clovers	B51	95
	Lucerne	B52	754
	Other leguminous and mixtures for fodder	B53	226
	Mixed cereals for fodder	B54	429
	Temporary grasslands	B55	479
Permanent Crops: Fruit Trees (B7X)	Apple fruit	B71	28
	Pear fruit	B72	21
	Cherry fruit	B73	12
	Nuts trees	B74	164
	Other fruit trees and berries	B75	118
	Oranges	B76	23
	Other citrus fruit	B77	25
Other Permanent Crops (B8X)	Olive groves	B81	526
	Vineyards	B82	336
	Nurseries	B83	32
	Permanent industrial crops	B84	4
Not Recognisable	Temporary Crops	BX1	967
	Permanent Crops	BX2	158

Table 3.2: Crop dataset distribution

3.1 Processing

3.1.1 Imputation

While it is true that many machine learning algorithms are able to handle missing values, others (Random Forest for one), are not. For this reason, a solution is to be found in order to tackle the problem. One possible way is to drop all instances presenting at least a missing values among their features but this would lead not only to a waste of data but, in the case of this study, to the annihilation of the entire datasets since the missing values are 122634 in the land cover dataset, of which 33054 are in crop type classification dataset. These numbers may seem big but they actually represent around 1% of values in the dataset. Another possible solution could be replacing all missing values with the most common value or with zero but in this case, the action of replacing missing values with constant values would inevitably change relationships between points in the datasets leading to misclassifications. Zero values carry, in fact, specific information when it comes to spectral indices, in the case of NDVI, for example, it identifies the bare soil.

One possible way, that has been tried in several experiments in this study, has been the imputation of missing values thanks to the mean of the preceding and the following values. Differently, however, from computing the mean of preceding and following value on the same column feature, the mean is to be computed along a given row, since the row is a set of temporal spectral index sequences, paying attention not to mix several different indices together whether the missing value is the last of one or the first of the other.

The best accuracy scores, however, have been achieved by means of the method of iterative imputation according to which each feature is modeled as a function of the other features much like it happens in a regression task where missing values are imputed sequentially and iteratively to improve estimates, one after the other allowing for prior imputed values to be used as part of a model in the prediction of subsequent features. While different regression algorithms can be used to estimate the missing values LightGBM has been preferred for its performances with the number of iterations of the procedure set to 5.

3.1.2 Outliers & inconsistent points Removal

In statistics, is defined as an outlier an observation point that is distant from other observations, basically unusual values in the dataset, having a different underlying behavior than the rest of the data, able to distort statistical analyses due to the increase in variability they introduce, which lead to decreases in statistical capability. Machine learning algorithms are, in fact, sensitive to the range and distribution of attribute values, outliers can spoil and mislead the training process resulting in

longer training times, less accurate models and ultimately poorer results. In order to identify outliers, the visualization method of box plot has been used for all features separately for each class and index. Box plot is a method for graphically depicting groups of numerical data through their quartiles and indicating also variability outside the upper and lower ones allowing for the identification of outliers. Since the datasets are not particularly big, not all outliers identified by the plots have been removed but rather the most extreme ones.

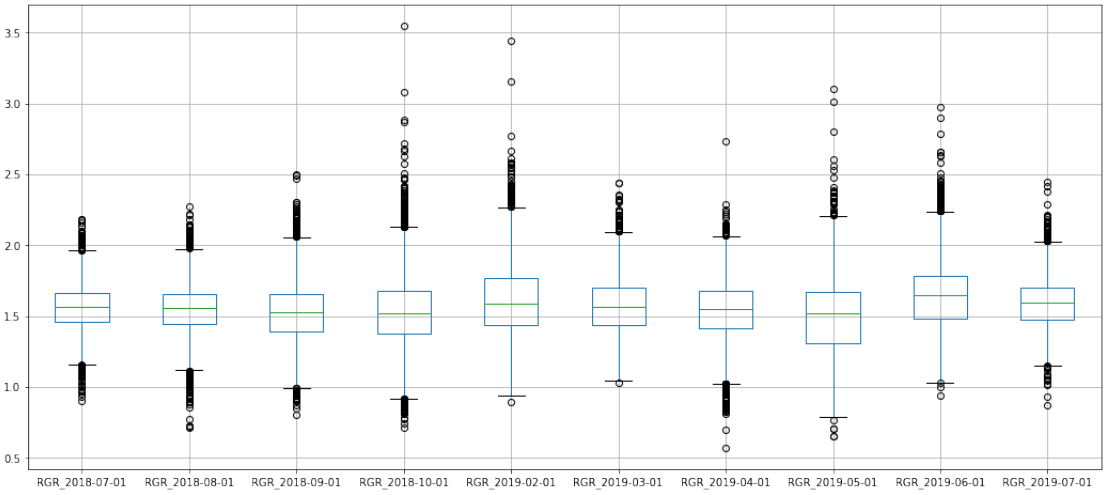

Figure 3.1: Box plot example for RGR index for Cereal Crops

Furthermore, a check for inconsistent points has been performed. Points have been in fact filtered out in case some of theirs values were outside the accepted range for the belonging class on the basis of the scientific literature of any given index. For this reason, for example, have been discarded vegetation points with NDVI values close to -1, which identifies water surfaces, as well as water surfaces with MNDWI values smaller than zero, and so on. Obviously, this removal has been rather mild, to resume the example of NDVI, it is proven, in fact, that the index for vegetation is supposed to assume values close to 1 while for bare soil surfaces is supposed to assume values close to 0, however, only vegetation points with values close to -1 have been removed since it is not infrequent for a crop field to assume 0 value, for instance right after the harvest, while it is not consistent for the field to assume water surface values (with the only exception for rice fields).

3.1.3 Normalization

Normalization is a data preparation technique whose goal is to change the values of numerical columns in a dataset to use a common scale, without distorting differences in the ranges of values, relationships among instances, general distribution or losing

information. As such it is also required for some algorithms to model data correctly since they look for trends in the data by comparing features of the instances and tend to give more importance to a feature with the increasing magnitude of values. The preferred and selected normalization method in the study is Z-Score, whose formula is:

$$Zscore = \frac{value - \mu}{\sigma}$$

where μ is the mean value of the given feature and σ its standard deviation of the feature. According to the normalization, if a value is exactly equal to the mean of all the values of the feature is normalized to 0, if it is below the mean is mapped to a negative number, and if it is above the mean is normalized to a positive number, with the size of those negative and positive numbers determined by the standard deviation of the original feature.

3.1.4 Feature Selection & Feature Engineering

In the development phase of a predictive model, it is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model. It is called feature selection the process in charge of this reduction. While it is true that there exist many effective statistical-based feature selection methods consisting in the evaluation of relationships between input variables and class variables, preferring features with the strongest relationship with the latter, the features relating to Band2, Band3, Band4, Band8, Band11 (corresponding to the three bands of the visible spectrum plus NIR and SWIR), NDVI, MNDWI, NDMI, NBAI, MCARI for the task of land cover classification and NDVI, SIPI, CRI550, NDMI, ARI, MCARI, RedEdgeNDVI, MSAVI, CCCI, RGR and REP for the task of crop classification have been empirically selected as being the ones achieving the best accuracy scores among many carried out experiments.

Additionally, some features have been engineered, namely the mean, median and standard deviation for each index and each point computed over the ten months values.

3.1.5 Class Groupings

One further important step is the one associated with the identification of the classes for both tasks. Not only, in fact, the number of classes is too high to efficiently and accurately train any predictive model, especially concerning crop type classification, with its 40 classes, but also classes are grouped according to the scientific belonging family, so families do not necessarily contain classes whose

points have a similar spectral signature and it might happen for classes belonging to two different scientific families to have a similar spectral signature. An example is "Rape and turnip rape" belonging to family "Non-Permanent Industrial Crops" having a spectral signature similar and comparable to the classes belonging to the "Root Crops" family. Additionally, families include classes that are an aggregation of many minor and rarer subclasses, this means that the spectral signatures of these classes are combinations of all the signatures of different subclasses within them, and this has to be accounted for. Furthermore, some classes have little to no points and have to be aggregated with some other classes. Finally, there are some families like "Fodder Crops" that contain a mix of classes that should belong to other families for the purpose of this study but are instead aggregated in such families on the basis of the usage, for example, "Other leguminous and mixtures for fodder" should be found in "Dry Pulses" while "Mixed cereals for fodder" should be found in "Cereals".

Concerning the task of land cover classification, the classes have been reorganized according to the considerations proposed in [48] and then once again modified on the basis of spectral signature analyses as well as tracking performances in several experiments finally obtaining classes shown in table 3.3.

Land Cover Class	Code	Points
Artificial land	A	1082
Cropland Seasonal	CS	6972
Cropland Perennial	CP	1394
Woodland	F	13167
Grassland	G	5362
Water Body	W	94
Wetlands	WL	64

Table 3.3: Reorganized Land Cover Classes

Where Artificial not only contains points belonging to LUCAS A class but also barren soil, rock and sand surfaces previously belonging to class "Bareland", making it more like an "Impermeable surface" class. LUCAS class B, used in the crop classification task, has been split into CS and CP according to whether the crop is seasonal like cereals or permanent like fruit trees. Finally, LUCAS Shrubland, which consists of small to medium vegetation surfaces like bushes, has been aggregated with Grassland.

Concerning the task of crop classification, the classes have been reorganized according to in-depth spectral signature analyses as well as empirically in several experiments finally obtaining classes shown in table 3.4.

Crop Class	Points
Cereals	3005
Dry Pulses	755
Floriculture	160
Fresh Vegetables	365
Fruit Trees	852
Maize	1023
Rice	114
Vineyard	327
Other	813

Table 3.4: Reorganized Crop Classification Classes

As a first thing the instances belonging to classes BX1 and BX2 have been dropped since they do not convey any useful information in terms of crop classification, they in fact represent respectively seasonal and permanent crops for which the LUCAS operator was unable to understand the exact class and family. Then, all cereals have been aggregated with the only exceptions being Maize and Rice that have clearly distinguishable spectral signatures. All Root Crops along with "Rape and turnip rape" belonging to the family of "Non-Permanent Industrial Crops" have been put in "Fresh Vegetables" along with classes from family "Dry Pulses, Vegetables, Flowers" that are "Tomatoes", "Other fresh vegetables" and "Strawberries". "Sunflowers" from "Non-Permanent Industrial Crops" and "Floriculture and ornamental plants" have formed the class of "Floriculture". Finally, "Other leguminous and mixtures for fodder" from "Fodder Crops" Family has been put in "Dry Pulses", "Mixed cereals for fodder" in "Cereals", all trees have been put in "Fruit Trees", "Nurseries" and "Permanent industrial crops" have been dropped respectively because not informative the first and with too few points the second, and all the remaining instances have been put into "Others".

3.1.6 Dataset Balance

At this point of the processing phase, after all the procedures described above what is left of the two datasets are 28136 points for the land surface classification dataset distributed among classes as shown in table 3.3, and 7414 points for the crop classification dataset distributed among classes as shown in table 3.4. As it is possible to see both the datasets are highly unbalanced, that means the classes are not represented approximately equally, therefore, balance is pursued by means of the last processing procedure before training models. Class balance

within a given dataset allows machine learning models to make more accurate and reliable predictions since with unbalanced data, classifiers are more sensitive to detecting the majority class and less sensitive to the minority class leading to a biased classification output. There are two possible ways to achieve balance, the first, Undersampling, consists in the removal of instances from the majority classes up until the number of instances is comparable among all the classes, the second, oversampling, consists of the synthetic creation of instances belonging to the minority classes up until the number of instances is comparable among all the classes.

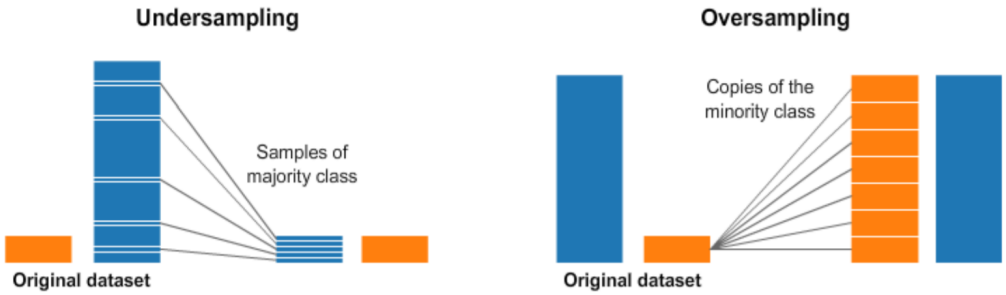

Figure 3.2: Undersampling and Oversampling

Since undersampling would lead to a huge information loss, especially considering the relatively small crop type classification dataset (it is, in fact, advisable only when the amount of data is so big, its processing constitute a too expensive computational cost), oversampling has been preferred by means of a technique called SMOTE proposed in [49]. SMOTE consists of the selection of a minority class instance 'A' at random and the finding of its k nearest minority class neighbors. A synthetic instance is then created by choosing one of the k nearest neighbors 'B' at random and connecting 'A' and 'B' to form a line segment in the feature space. The synthetic instances are generated as a convex combination of the two chosen instances 'A' and 'B'.

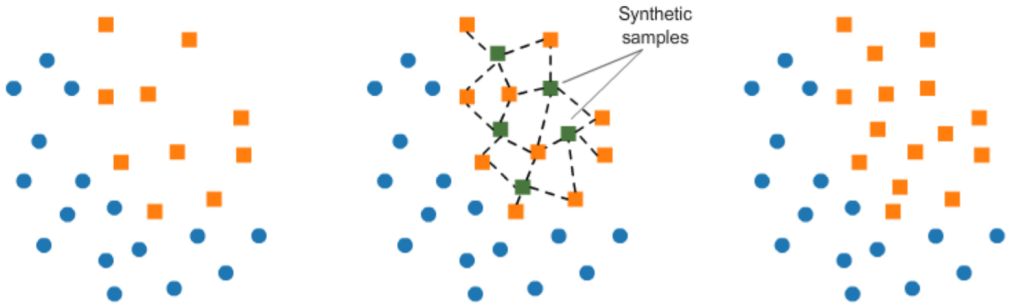

Figure 3.3: SMOTE

However, it is rather important to understand when to perform the dataset oversampling. As it is possible to see from figure 3.4, in fact, oversampling before cross-validation, would make the model be trained on instances that are the same as the ones used for validating the model voiding the purpose of the validation phase and leading to overfitting. It is, therefore, important to perform oversampling only on training data.

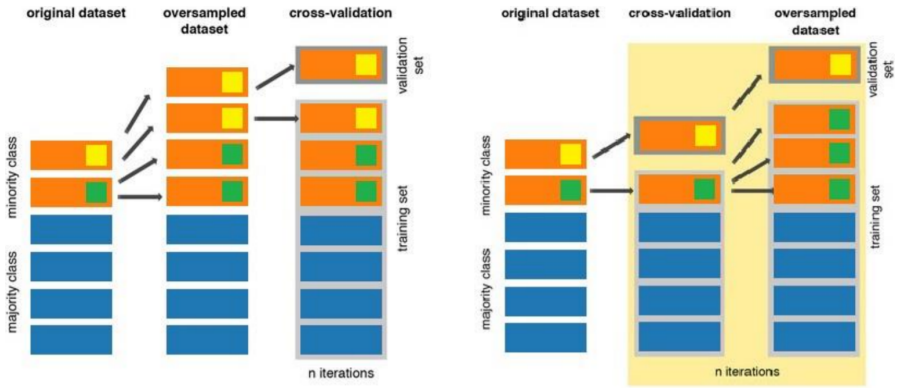

Figure 3.4: Wrong vs correct oversampling methodology

3.2 Results

The following results have been obtained by the models described in the second chapter, trained on the datasets built according to the methodology depicted in the same chapter on which have been applied the processing procedures illustrated in the current chapter. Furthermore, to tune hyperparameters in the best possible way, both Random Search first and Bayesian Optimization later, have been performed in combination with stratified 5-fold cross-validation, which is a technique for assessing how the results of a statistical analysis will generalize to an independent data set, mainly used in settings where the goal is prediction, according to which the original training and validation sets are partitioned into k = 5 equal-sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for tuning the model, and the remaining k - 1 subsamples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data. The k results can then be averaged to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once. "Stratified" means that each partition contains roughly the same proportions of class labels, important since the validation has to be performed on a subsample whose underlying distribution is as close as possible to the one of

the test sample and finally as close as possible to the whole dataset's underlying distribution. As for the hyperparameters tuning algorithms, differently from Grid Search, Random Search does not require an explicit set of possible values for each hyperparameter, but rather a statistical distribution for each hyperparameter from which values are sampled, while Bayesian optimization is a sequential model-based optimization algorithm that uses the results from the previous iteration to select the next hyperparameter value candidates, that means instead of blindly searching the hyperparameter space as it happens with Grid Search and Random Search, this method selects as next set of hyperparameters the one which will improve the model performance.

Generally, results follow the same trend for both the tasks of land cover classification and crop type classification, which means the models' rankings by accuracy are the same for both tasks with the only difference being a decrease of roughly 10% in overall accuracy (OA) from land cover classification to crop type classification. As expected the best models are the ones belonging to the gradient boosting family with LightGBM achieving 82.6% OA on land cover and 73.8% on crops followed by XGBoost achieving 82.0% OA on land cover and 72.6% on crops. Of the more simple models, Random Forest proves to be suitable to tackle the tasks achieving 81.2% OA on land cover and 70.7% on crops, while, differently from what claimed in [27] SVM, although being experimented with extensively and implemented with several kernels and thorough hyperparameters searches, has not been able to go past 46.8% OA on land cover and 40.5% on crop. Finally, deep learning models have proved to be worthy of the tasks reaching comparable results with respect to the shallow models mentioned above, with the best among them being NODE achieving 81.3% OA on land cover and 71.6% on crops, followed by TabNet achieving 80.6% OA on land cover and 67.7% on crops while Pixel R-CNN's claimed accuracy [27] has not been met as the model reached 77.7% OA on land cover and 64.1% on crops even though it is fair to point out that in [27] the number of data points was bigger since all the pixels of a given field containing a LUCAS point were taken and labeled with the same label of the LUCAS point, by means of a manual mapping of the geo-polygon of the field. This method performed on just one UTM tile leads to a dataset of many points belonging to a fraction of the classes analyzed in this study, therefore, making it possible to reach such high accuracy values.

Investigating results more deeply, as it is possible to see from figures 3.5 and 3.6 the OA is not equally distributed among the classes of the two tasks. Concerning land cover, in fact, while there are classes, like "Artificial, Rock and Bare soil", "Forest" and "Grassland and Shrubland", with very high accuracy score, others, like "Crop Permanent", which is often confused with other classes especially "Forest", and "Wetland", whose number of instances is too limited to have an effective impact on dataset balancing, have very low scores.

Concerning crop classification, on the other hand, the scores have less variance with the less performing classes being "Fresh Vegetables" which is most of the times confused with "Cereals", and "Vineyards" which are confused with "Fruit Trees".

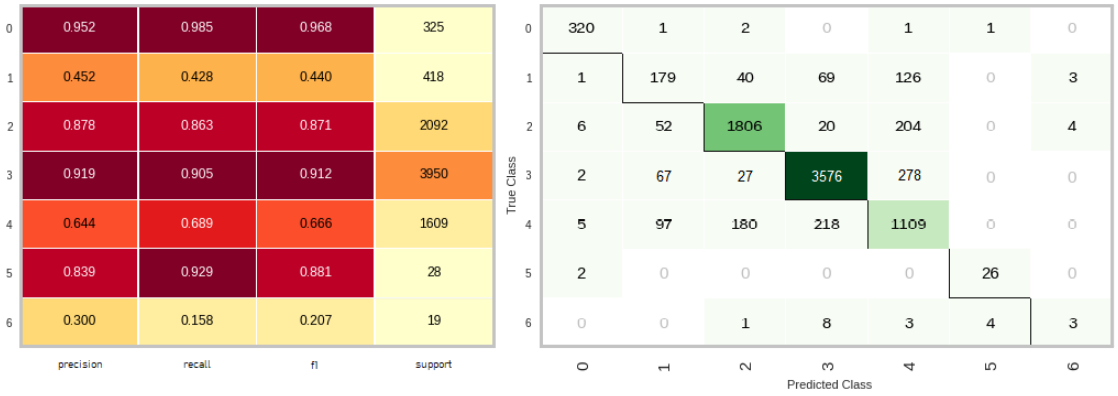

Figure 3.5: Land Cover Classification class report and confusion matrix. Artificial, Rock & Bare Soil: 0, Crop Permanent: 1, Crop Seasonal: 2, Forest: 3, Grassland and Shrubland: 4, Water body: 5, Wetland: 6

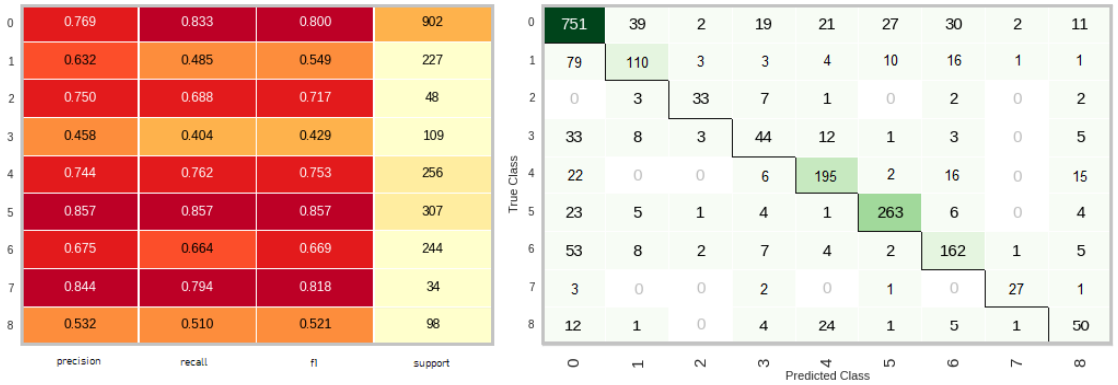

Figure 3.6: Crop Classification class report and confusion matrix. Cereal: 0, Dry Pulses: 1, Floriculture: 2, Fresh Vegetables: 3, Fruit Trees: 4, Maize: 5, Other: 6, Rice: 7, Vineyard: 8

Finally, the best performing model has been used to predict both the task on two random Italian areas, whose only constraint was to include as many different features as possible such as towns, crops, water bodies, forest, etc. To better explain the top image being a simple RGB satellite image of the area, the bottom left one being the predicted land cover and the bottom right one being the predicted crop type classification. One important thing to point out is that the crop classification has been carried out starting from the land cover classification, which means only

pixels associated with "Crop Permanent" and "Crop Seasonal" have been passed down to the crop classification model, this is why the bottom right image shows a further class called "Not Crop".

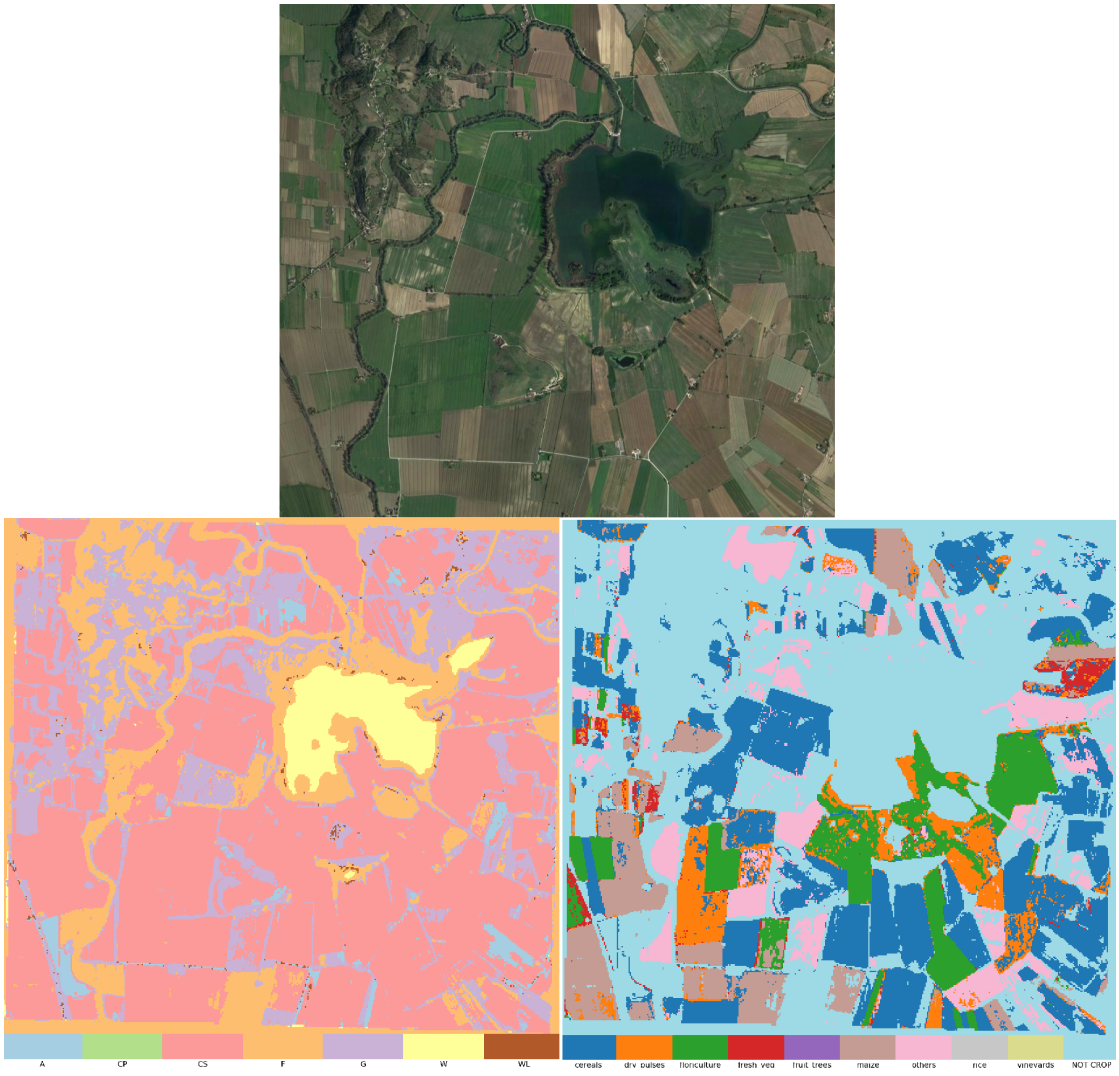

Figure 3.7: Best performing model applied on the area of Lago Ripasottile, RI

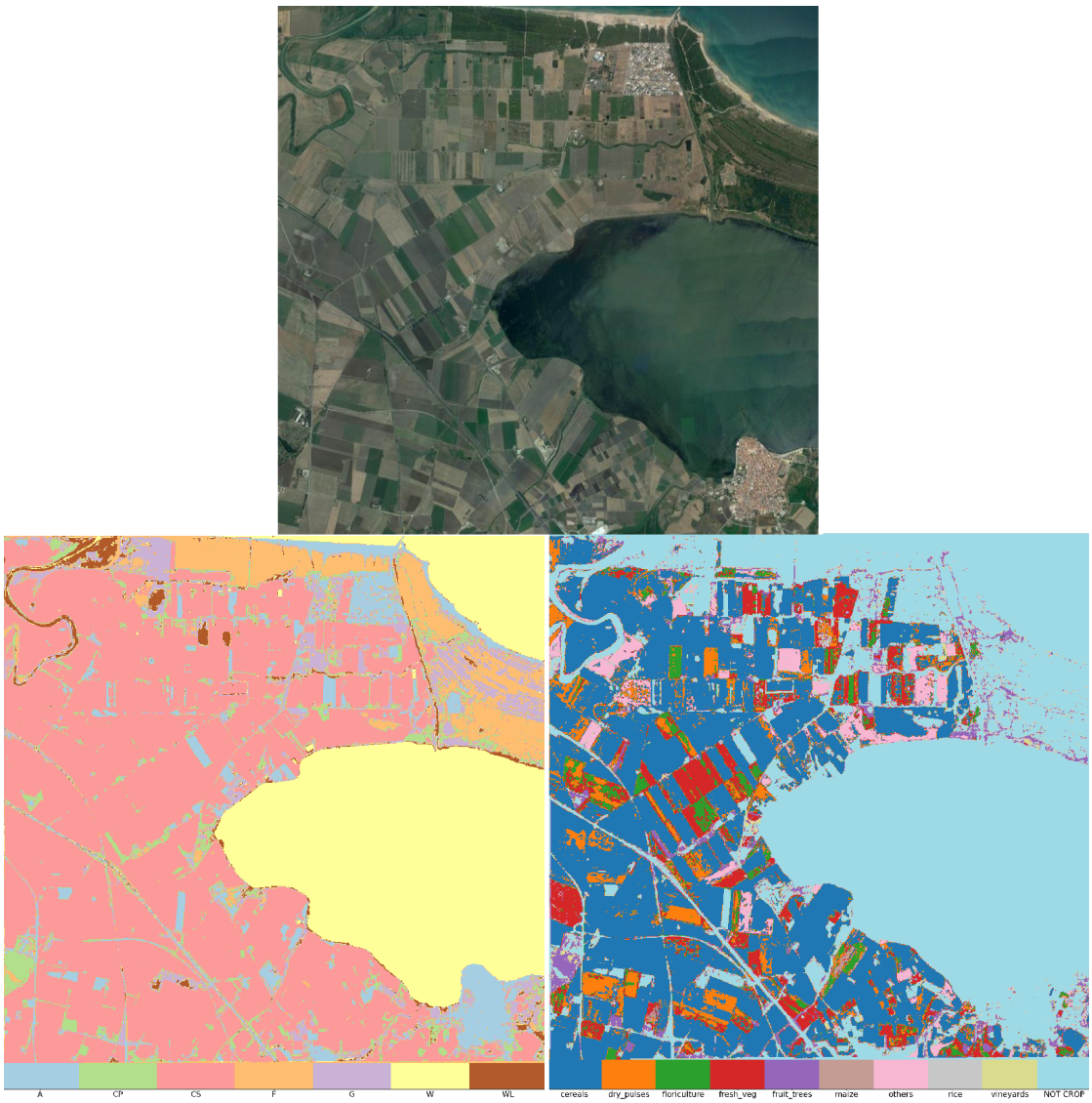

Figure 3.8: Best performing model applied on the area Lesina, FG

3.3 Conclusions

In this study has been examined the possibility of obtaining land cover classification and crop type classification relying on the only information carried by multispectral satellite data provided by Copernicus Sentinel-2 mission. To do so, two tabular datasets have been built starting from a dataset provided by Eurostat called LUCAS which has been used to obtain ground truth labels, in conjunction with a careful selection of spectral indices. Furthermore, both shallow and deep learning models have been used in order to obtain the best possible accuracy, of which LightGBM has been found to be the best in both tasks achieving an overall accuracy of 82.6% on land cover and 73.8% on crops which can be considered a good result considering the limited amount of points in the datasets. There are however many possible improvements that may boost the performances of the models used in the study on both tasks, many of which will certainly be pursued in the near future as technology advances in the sector. Such improvements are not much in relation to the models themselves, many of which still have learning capacity to be exploited, but rather in relation to the data that is fed to them.

Increasing Points with Border Detection

The first improvement would certainly concern the number of data. Both the datasets built as described in this study, unfortunately, count too few instances, especially for the crop type classification one, not only limiting the number of classes among which it is possible to classify since many classes have to be aggregated with others or have to be dropped due to the scarce number of instances, insufficient to train a model to be able to recognize them, but also precluding deep learning models, that need big datasets especially with the increasing number of parameters, to reach their full potential, leading them to overfit.

In order to increase the number of data without having to personally register geographical points and their land cover, given all the obstacles described in the previous chapter, the datasets could be filled not only with the exact LUCAS point but, by means of border detection models, with all the points within the same field of the given LUCAS point that is reasonable and safe to assume would belong to the same class, thus exceptionally increasing the number of data.

Increasing Temporality with Sen2Like

In this study, 10 Sentinel-2 multispectral products distributed over one year, as described in paragraph 2.2, have been considered. Increasing the number of products may lead to more accurate and distinguishable spectral signatures and therefore more accurate predictions. Sentinel-2, a constellation of the satellites 2A and 2B, is however limited by a 5 day revisit time that along with cloud coverage

translates to a small increase in the number of usable products. For this reason, it is possible to use a tool proposed in [50] called Sen2Like, whose objective is to harmonize Sentinel-2 and Landsat-8 multispectral products, which by nature are not comparable since they have not only different geometries (angles, orbits, product formats, etc.) but also different band resolutions. Sen2Like harmonization process improves significantly revisit time with the theoretical number of acquisitions of the virtual constellation made of Sentinel-2 and Landsat-8 (consisting of 95 products per year) being increased by as much as 30% with respect to Sentinel-2 only acquisitions (consisting of 73 products per year).

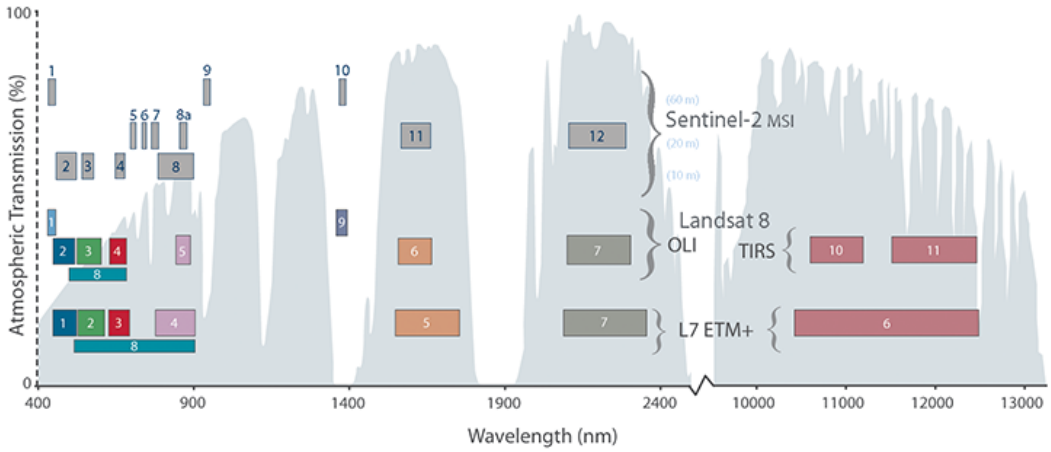

Figure 3.9: Specific placement of Sentinel-2 bands, as compared to Landsat-7 and 8 bands. Source: USGS

Increasing Band and Spatial Resolutions

Further improvements could concern a general enhancement of spectral imaging data in the two main directions of increasing the number of bands and consequently transitioning from multispectral products to hyperspectral ones or increasing the spatial resolutions of the bands. In the first case, since, differently from multispectral products which consist from 3 to 13 bands, hyperspectral products might count hundreds to thousands of narrower bands packed in the same spectral range, there might be not only benefits given by the number of spectral indices that were not computable with multispectral products and consequently a higher descriptive capability but also benefits given by the extreme accuracy of the computed values achieving a high level of distinctiveness even among spectral signatures of classes belonging to the same family.

In the second case the higher the resolution the more precise is the spectral value of the band. Sentinel-2 satellites provide several bands at different resolutions as

shown in table 1.1, where the resolution identifies the length of one side of a pixel in the given band. Having higher resolution bands would not only avoid obtaining a downgraded result when computing indices with two different resolutions but also would make it possible to filter out pixels that do not belong to the target, such as pixels associated with soil, instead of having huge pixels whose value is the mean of the responses of all the components of the associated area.

Sentinel-1 Integration

One last improvement consists in the enhancement of the descriptive capabilities with the integration of SAR data provided by another Copernicus mission: Sentinel-1. The mission is composed of a constellation of two satellites, 1A and 1B, created by an industrial consortium led by Thales Alenia Space Italy as prime contractor, along with Astrium Germany being responsible for the C-SAR payload, incorporating the central radar electronics sub-system developed by Astrium UK. The two satellites, sharing the same polar orbital plane with a 180° phasing difference, provide a continuous radar mapping of the Earth with an enhanced revisit frequency of 6 days. Each satellite is, in fact, able to map the global landmasses once every 12 days, in a single pass (ascending or descending). The mission includes a C-band, IEEE designation for a portion of the electromagnetic spectrum in the microwave range of frequencies ranging from 4 to 8 GHz, Synthetic Aperture Radar (SAR), radar type able to reach a resolution way higher than the one of a normal radar with equal antenna length. This happens because the satellite acquires not just one response from a given target on the ground but as many responses as long as the target is inside the illumination beam emitted by the antenna, the complex echo signals received during this time are then added coherently to obtain higher resolutions. The antenna is said to be "synthesized" with the synthetic aperture length being equal to the distance traveled by the satellite during the echo signals integration time. The fact that the system works coherently from end-to-end means that both the amplitude and the phase relationships between the complex transmitted and received signals are maintained throughout the whole process. This facilitates aperture synthesis as well as multi-pass radar interferometry using pairs of images taken over the same area at different times. Moreover, the satellite can collect several different images from the same series of pulses by using its antenna to receive specific polarisations simultaneously. Sentinel-1 is a phase-preserving dual polarization SAR system and can transmit a signal in either horizontal (H) or vertical (V) polarisation, and then receive in both H and V polarisations. The main advantage of operating in the C-band is that the wavelengths are not impeded by clouds or lack of illumination so images can, therefore, be acquired during day or night and under almost all weather conditions.

Bibliography

[1] Elachi. *Introduction to the Physics And Techniques of Remote Sensing*. Wiley-Interscience, 1987 (cit. on p. 1).

[2] Professional Aerial Photographers Association. *History of Aerial Photography*. URL: https://professionalaerialphotographers.com/content.aspx?page_id=22&club_id=808138&module_id=158950 (cit. on p. 2).

[3] Cracknell. *The development of remote sensing in the last 40 years*. International Journal of Remote Sensing, 2018 (cit. on p. 3).

[4] Baumann. *History of Remote Sensing, Satellite Imagery, Part II*. URL: http://employees.oneonta.edu/baumanpr/geosat2/rs%5C%20history%5C%20ii/rs-history-part-2.html (cit. on p. 3).

[5] Zhu et al. *Benefits of the free and open Landsat data policy*. Remote Sensing of Environment, 2019 (cit. on p. 4).

[6] Moore. *What is a picture worth? A history of remote sensing*. Reading, MA: Addison-Wesley, 1979 (cit. on p. 4).

[7] MIT Spectroscopy Laboratory. *The Era of Classical Spectroscopy*. URL: https://web.mit.edu/spectroscopy/history/history-classical.html (cit. on p. 4).

[8] Hunt. *History of Spectroscopy*. 2011 (cit. on p. 4).

[9] Kroto. *Molecular Rotation Spectra*. Wiley, 1975 (cit. on p. 5).

[10] NASA Hubble Space Telescope Website. *Spectroscopy: Reading the rainbow*. URL: https://hubblesite.org/contents/articles/spectroscopy-reading-the-rainbow (cit. on p. 5).

[11] Government of Canada. *Interactions with the Atmosphere*. URL: https://www.nrcan.gc.ca/maps-tools-publications/satellite-imagery-air-photos/remote-sensing-tutorials/introduction/interactions-atmosphere/14635 (cit. on p. 7).

[12] Sensing CRISP (Center for Remote Imaging and Processing. *Effects of Atmosphere*. URL: https://crisp.nus.edu.sg/~research/tutorial/atmoseff.htm#:~:text=Atmospheric%5C%20absorption%5C%20affects%5C%20mainly%5C%20the%5C%20visible%5C%20and%5C%20infrared%5C%20bands.&text=The%5C%20reflected%5C%20radiance%5C%20is%5C%20also,of%5C%20the%5C%20target%5C%20being%5C%20observed (cit. on p. 8).

[13] Diner et al. *Multi-angle Imaging SpectroRadiometer (MISR) instrument description and experiment overview*. IEEE Transactions on Geoscience and Remote Sensing, 1998 (cit. on p. 8).

[14] Abdou, Helmlinger, Conel, Bruegge, Pilorz, Martonchik, and Gaitley. *Ground measurements of surface BRF and HDRF using PARABOLA III*. Journal of Geophysical Research, 2001 (cit. on p. 8).

[15] Anderson and Milton. *Characterisation of the apparent reflectance of a concrete calibration surface over different time scales*. Institute of Geographic Sciences and Natural Resources Research, 2005 (cit. on p. 8).

[16] Abdou, Helmlinger, Conel, Bruegge, Pilorz, Martonchik, and Gaitley. *On the variability of the reflected radiation field due to differing distributions of the irradiation*. Remote Sensing of Environment, 1976 (cit. on p. 8).

[17] Disney. *Improved estimation of surface biophysical parameters through inversion of linear BRDF models*. PhD thesis, University College London, UK, 2001 (cit. on p. 10).

[18] Copernicus.eu. *Copernicus In Brief*. URL: https://web.archive.org/web/20180815041358/http://copernicus.eu/main/copernicus-brief (cit. on p. 11).

[19] Jones. *Importance of Land Cover and Biophysical Data in Landscape-Based Environmental Assessments*. URL: http://www.aag.org/galleries/nalcs/CH13.pdf (cit. on p. 15).

[20] Satpalda Geospatial Services. *Significance of Land Use / Land Cover (LULC) Maps*. URL: https://www.satpalda.com/blogs/significance-of-land-use-land-cover-lulc-maps (cit. on p. 15).

[21] Huete, Didan, Miura, Rodriguez, Gao, and Ferreira. *Overview of the radiometric and biophysical performance of the MODIS vegetation indices*. Remote Sensing of Environment, 2002 (cit. on p. 16).

[22] Xue and Su. *Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications*. Journal of Sensors, 2017 (cit. on p. 17).

[23] EUMeTrain. *Training Module on METOP AVHRR RGB Images*. URL: http://www.eumetrain.org/data/4/461/navmenu.php?tab=5%5C&page=2.0.0 (cit. on p. 17).

[24] Thenkabail, Lyon, and Huete. *Hyperspectral Indices and Image Classifications for Agriculture and Vegetation*. CRC press, 2018 (cit. on p. 18).

[25] Chang, Peng-Sen, and Shi-Rong. *A review of plant spectral reflectance response to water physiological changes*. Chinese Journal of Plant Ecology, 2016 (cit. on pp. 18, 20).

[26] Eurostat. *LUCAS - Land use and land cover survey*. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=LUCAS_-_Land_use_and_land_cover_survey#Defining_land_use.2C_land_cover_and_landscape (cit. on p. 22).

[27] Mazzia, Khali, and Chiaberge. *Improvement in Land Cover and Crop Classification based on Temporal Features Learning from Sentinel-2 Data Using Recurrent-Convolutional Neural Network (R-CNN)*. Applied Sciences, 2020 (cit. on pp. 23, 34, 49).

[28] Yea, Li, Chena, and Zhang. *A spectral index for highlighting forest cover from remotely sensed imagery*. Land Surface Remote Sensing II, 2014 (cit. on p. 26).

[29] McFeeters. *The use of Normalized Difference Water Index (NDWI) in the Delineation of Open Water Features*. International Journal of Remote Sensing, 1996 (cit. on p. 26).

[30] Xu. *Modification of Normalised Difference Water Index (NDWI) to Enhance Open Water Features in Remotely Sensed Imagery*. International Journal of Remote Sensing, 2006 (cit. on p. 26).

[31] Zha, Gao, and Ni. *Use of Normalized Difference Built-Up Index in Automatically Mapping Urban Areas from TM Imagery*. International Journal of Remote Sensing, 2003 (cit. on p. 27).

[32] Waqar, Mirza, Mumtaz, and Hussain. *Development of new indices for extraction of built-up area and bare soil from landsat*. Open Access Scientific Reports, 2012 (cit. on p. 27).

[33] Rouse, Haas, Schell, and Deering. *Monitoring Vegetation Systems in the Great Plains with ERTS*. Third ERTS Symposium, 1973 (cit. on p. 27).

[34] Qi, Chehbouni, Huete, Kerr, and Sorooshian. *A Modified Soil Adjusted Vegetation Index*. Remote Sensing of Environment, 1994 (cit. on p. 27).

[35] Huete. *A Soil-Adjusted Vegetation Index (SAVI)*. Remote Sensing of Environment, 1988 (cit. on p. 28).

[36] Penuelas, Baret, and Filella. *Semi-Empirical Indices to Assess Carotenoids / Chlorophyll-a Ratio from Leaf Spectral Reflectance*. Photosynthetica 31, 1995 (cit. on p. 28).

[37] Gitelson, Zur, Chivkunova, and Merzlyak. *Assessing Carotenoid Content in Plant Leaves with Reflectance Spectroscopy*. Photochemistry and Photobiology, 2002 (cit. on p. 28).

[38] Hardisky, Klemas, and Smart. *The Influence of Soil Salinity, Growth Form, and Leaf Moisture on the Spectral Radiance of Spartina alterniflora Canopies*. Photogrammetric Engineering and Remote Sensing, 1983 (cit. on p. 29).

[39] Gitelson, Merzlyak, and Chivkunova. *Optical Properties and Nondestructive Estimation of Anthocyanin Content in Plant Leaves*. Photochemistry and Photobiology, 2001 (cit. on p. 29).

[40] Daughtry. *Estimating Corn Leaf Chlorophyll Concentration from Leaf and Canopy Reflectance*. Remote Sensing Environment, 2000 (cit. on p. 29).

[41] Barnes et al. *Coincident detection of crop water stress, nitrogen status and canopy density using ground based multispectral data*. Proc. 5th Int. Conf. Precis Agric, 2000 (cit. on p. 29).

[42] Sims and Gamon. *Relationships Between Leaf Pigment Content and Spectral Reflectance Across a Wide Range of Species, Leaf Structures and Developmental Stages*. Remote Sensing Environment, 2002 (cit. on p. 30).

[43] Gamon and Surfus. *Assessing leaf pigment content and activity with a reflectometer*. New Phytologist, 1999 (cit. on p. 30).

[44] Curran, Windham, and Gholz. *Exploring the Relationship Between Reflectance Red Edge and Chlorophyll Concentration in Slash Pine Leaves*. Tree Physiology, 1995 (cit. on p. 30).

[45] Popov, Morozov, and Babenko. *Neural oblivious decision ensembles for deep learning on tabular data*. arXiv, 2019 (cit. on p. 36).

[46] Peters, Niculae, and Martins. *Sparse sequence-to-sequence models*. arXiv, 2019 (cit. on p. 36).

[47] Arik and Pfister. *Tabnet: Attentive interpretable tabular learning*. arXiv, 2019 (cit. on p. 37).

[48] Pflugmacher, Rabe, Peters, and Hostert. *Mapping pan-European land cover using Landsat spectral-temporal metrics and the European LUCAS survey*. Remote Sensing of Environment, 2019 (cit. on p. 45).

[49] Chawla, Bowyer, Hall, and Kegelmeyer. *SMOTE: Synthetic Minority Oversampling Technique*. Journal of Artificial Intelligence Research, 2002 (cit. on p. 47).

[50] Saunier, Louis, Debaecker, Beaton, Cadau, Boccia, and Gascon. *Sen2like, A Tool To Generate Sentinel-2 Harmonised Surface Reflectance Products - First Results with Landsat-8*. IGARSS, 2019 (cit. on p. 54).

Ringraziamenti

A conclusione di questo elaborato, desidero ringraziare tutte le persone il cui sostegno non solo ha reso possibile questo lavoro ma mi ha anche e soprattutto dato la forza per conquistare quello che un giorno pensavo essere un traguardo irraggiungibile.

Ringrazio il Politecnico di Torino per avermi dato la possibilità di iscrivermi al corso di Laurea Magistrale ed avermi introdotto ad una disciplina che velocemente e con entusiasmo è entrata a far parte delle mie passioni prima che delle mie skills lavorative. Allo stesso modo ringrazio il mio relatore Fabrizio Lamberti ed il mio correlatore Lia Morra per la disponibilità ed il contributo.

Ringrazio Data Reply per le risorse che mi ha messo a disposizione per portare a compimento le ricerche. In particolar modo ringrazio Ruggiero Giuseppe, Corradini Giovanni e Rontauroli Matteo per i consigli, gli insegnamenti, l'organizzazione e più in generale l'ottima accoglienza che sono stati in grado di offrirmi annullando la distanza fisica imposta dalla pandemia.

Ringrazio di cuore i miei genitori Antonino e Nicoletta e mio fratello Edoardo. Grazie per avermi sempre sostenuto e permesso di portare a termine gli studi universitari.

Ringrazio particolarmente Gloria, protagonista quanto me di questo piccolo successo. La vita è molto più semplice e bella quando si fa parte di un team così forte.

Ringrazio infine gli amici di lunga data che ci sono sempre stati, specialmente Renfe e i piccioni tutti, i compagni di Politecnico di prima ora, Ferra, Fra, Fede e Peppe,

così come i data scienziati del cluster che mi hanno accompagnato nell'ultimo miglio. Tra questi non ringrazierò mai abbastanza Marco per quei fondamentali consigli su APA che mi diede la sera prima dell'esame.

Vorrei infine ringraziare l'Unione Europea ed ESA per la gratuità e facilità di accesso dei preziosi dati utilizzati in questa ricerca, a dimostrazione del fatto che il progresso trovi terreno fertile nella condivisione di risorse e nell'unione piuttosto che nelle divisioni.