Semi-Supervised Learning Applied to Soil Nutrient Estimation Using Earth Observation Data

2nd Cycle Integrated Project in Computer Science and Engineering

Rafael António Fernandes e Costa Marques Candeias

Supervisors:
Dr. Amâncio Lucas de Sousa Pereira

MCSE
Instituto Superior Técnico - Lisboa
2022/2023

Acronyms

B Boron. 6, 9, 12, 13
C Carbon. 5, 6, 9, 13, 27
CO_2 Carbon dioxide. 5, 9
Ca Calcium. 6, 9–11, 13, 24
Cl Chlorine. 6, 9, 12, 13
Cu Copper. 6, 9, 11, 13
Fe Iron. 6, 9, 12, 13
H Hydrogen. 6, 9, 10, 13
H_2O Dihydrogen monoxide. 9
K Potassium. 6, 9–11, 13
Mg Magnesium. 6, 9, 11, 13
Mn Manganese. 6, 9, 11, 13
Mo Molybdenum. 6, 9, 12, 13
N Nitrogen. 6, 9, 10, 12, 13
NH_4^+ Ammonium. 10
NO_3^- Inorganic nitrate. 10
O Oxygen. 5, 6, 9, 13
P Phosphorus. 6, 9–13
S Sulfur. 6, 9, 11–13
Zn Zinc. 6, 9, 11, 13
AEC Anion Exchange Capacity. 9
AfSIS Africa Soil Information Service. 20
AI Artificial Intelligence. 6–8, 16, 17, 20, 22, 24, 27
ALOS The Advanced Land Observing Satellite. 14
BLUE Blue. 16, 21
CEC Cation Exchange Capacity. 9, 10
CNSA Chinese National Space Administration. 13, 14
DAICHI Advanced Land Observing Satellite-2. 14
DL Deep Learning. 21
DTM Digital Terrain Model. 20
EMS Electromagnetic Spectrum. 15
EOS Earth Observation Satellites. 6, 7, 13–15, 19, 20, 22, 23, 27
ESA European Space Agency. 13, 14
ESDAC European Soil Data Centre. 19
EVI Enhanced Vegetation Index. 16
FAO Food and Agriculture Organization. 5
FN False Negative. 18, 25
FP False Positive. 18, 25
FY-3 Fengyun-3. 14
GCOM Global Change Observation Mission. 14
GDP Gross Domestic Product. 5

GF-3 Gaofen-3. 14
GGOS Greenhouse Gases Observing Satellite. 14
GI Green Index. 21
GO Geostationary Orbit. 13
GREEN Green. 16
GSW Global Surface Water. 20
GVI Green Vegetation Index. 21
IFDC International Fertilizer Development Center. 20
ISRIC International Soil Reference and Information Centre. 20
ISRO Indian Space Research Organisation. 13, 14
ISS International Space Station. 12
JAXA Japan Aerospatial Exploration Agency. 13, 14
LEO Low Observation Orbit. 13
LUCAS Land Use and Coverage Area frame Survey. 13, 20
MAE Mean Absolute Error. 25
MCARI Modified Chlorophyll Absorption in Reflectance Index. 17
MEO Medium Observation Orbit. 13
MIR Middle-Infra-Red. 17
ML Machine Learning. 6–8, 17–26
MLR Multiple Linear Regression. 21
MSI Multispectral Instruments. 14, 15
NASA National Aeronautics and Space Administration. 8, 13, 14, 20
NCSS National Cooperative Soil Survey. 20
NDVI Normalized Difference Vegetation Index. 16
NIR Near-Infra-Red. 16, 17, 20, 21
OWID Our World in Data. 5
PCA Principal Component Analysis. 17
RED Red. 16, 17, 21
RF Random Forest. 21, 25
RGB Red-Green-Blue. 15, 16, 20
RL Reinforcement Learning. 17
RMSE Root Mean Square Error. 20, 25
S1 Sentinel-1. 13
S2 Sentinel-2. 13, 20
S3 Sentinel-3. 13
S5P Sentinel-5P. 13
S6 Sentinel-6. 13
SAVI Soil Adjusted Vegetation Index. 16, 17
SL Supervised Learning. 7, 17, 20, 21, 25, 26
SMOTE Synthetic Minority Over-sampling Technique. 24
SSL Semi-Supervised Learning. 7, 8, 17, 21, 24, 25, 27
SSO Sun Synchronous Orbit. 13
SVI Spectral Vegetation Indices. 6, 7, 9, 15–17, 21, 22

SVM Support Vector Machine. 21
SWIR Short Wave Infra-Red. 20
TAMASA Taking Maize Agronomy to Scale in Africa. 20
TN True Negative. 18, 25
TP True Positive. 18, 25
UL Unsupervised Learning. 17
ZY-3A Zi Yuan-3A. 14

1 Introduction

1.1 Context

Before interplanetary exploration, soil was described as an "unconsolidated mineral or organic material on the immediate surface of the earth that serves as a natural medium for the growth of land plants."[1]. Harold van Es, in 2017, concluded that this definition was outdated and contained some inconsistencies, such as its exclusivity to planet Earth, ignoring soils with some degree of consolidation, and some others. Therefore, they published a paper with "A new definition of SOIL" concluding that it is "The layer(s) of generally loose mineral and/or organic material that is affected by physical, chemical, and/or biological processes at or near the planetary surface and usually hold liquids, gases, and biota and support plants."[1]

Soil is a crucial pillar to our economy, society, and Earth's environment, as it supplies materials for human usage, provides food for livestock, fuel, fiber for clothing, and retains C emissions. Actually, in 2012, 19% of the global population was engaged in farming activities, and even though it only represented 2.8% of global income, it reflected 18% of the middle and low-income countries' GDP [2]. Despite this, it is estimated that by 2050 human population will grow by a third [3]. Due to this unprecedented rise, we are already experiencing soil over-exploitation to keep up with the never-ending demand for food and materials. Consequently, the world's total area of fertile soil is being lost to new habitation and other infrastructures[1]. Thus, as time goes by, humanity requires more soil, but has fewer amounts to explore.

Methods, like over-using fertilizers and pesticides, focus on getting the most out of the soil with little concern for the land's health. In fact, according to the FAO, "generating 3cm of topsoil takes 1,000 years" and "12m hectares of topsoil are lost every year". Therefore, if intensive farming continues by 2070, arable and productive topsoil will cease to exist [2]. However, some articles refute such claims, stating that "only 60 harvests left" is "overblown", but that we undoubtedly "should not detract from the fact that soil erosion is a problem." [3]. While this is a debatable and open topic, OWID firmly declares that "6% of soils are estimated to have a lifespan of fewer than 100 years. (...) half have a lifespan greater than 1000 years, and one-third have over 5000 years."[4].

Even though soil degradation by over-exploitation is an open topic, it is also the Earth's layer where flora proliferates. All living beings in this flora group play a major role in the environment as they perform photosynthesis, which is a vital process in a plant's life where it inhales CO_2, produces energy and exhales O. From 1982 to 2020, "global carbon dioxide concentrations in the atmosphere grew about 17%, from 360 parts per million (ppm) to 420 ppm"[5], and if we do not reduce our global emissions, "the typical weather patterns will change, some animal species will likely disappear (...) force 100 million people into extreme poverty by 2030" and it will "result in the deaths of over 250,000 people globally and annually".[6]

As a result, without an accurate evaluation of the soil's properties, such intensive techniques,

[1] Overpopulation Environment, https://tinyurl.com/4n2szetn
[2] Soil Degradation, https://tinyurl.com/2p9e9bnf
[3] 60 years left, https://tinyurl.com/267kz2js
[4] OWID 60 left, https://tinyurl.com/5xyrveue
[5] Plants climate change, https://tinyurl.com/y3j8dd6d
[6] Global warming, https://tinyurl.com/5cp764bf

like applying insecticides excessively, and overusing fertilizers, contribute to soil erosion. Maintaining this course of action will undoubtedly lead us to a downfall, where soil transforms to dry land, higher percentages of people starve, and, since soil is a natural solution to reduce green house gases, the unstoppable arrival of climate change.

1.2 Motivation

To protect our soils, it is necessary to adopt better agricultural practices. Accurately knowing the fertility state and health of the land prevents owners from over-dosing them. These predictions are based on the amount of nutrients in the soil that contribute for its fertility. There are sixteen elements that make a soil fertile: C, H, O, N, P, K, Ca, Mg, S, B, Cu, Cl, Fe, Mn, Mo, and Zn. Some of them are minerals, while others non-minerals, and all are required in different quantities. However, the mere presence of these elements is far from enough to extract the correct amount of chemicals to add. Other soil properties, such as the land's ph level, limit the amount of minerals that the plant's roots can absorb. Therefore, knowing the soil's properties is also required to adopt greener agricultural practices.

There are several methods to estimate these soil characteristics and the number of nutrients that they possess. However, soil measurements have been primarily made by conventional laboratory analysis - a reliable but complex process. Interestingly, it is also one of the oldest ways to perform these evaluations, which reflects the saturation in soil analysis investigation. As we can see, in 1970 there were already papers related to chemical soil analysis [4], and in 2008, in the United States of America, "Bray P1[7] was the most frequently used method for P extraction" [5]. This method consists of performing chemical analysis of soil pieces in a lab. Such analysis requires access to a laboratory which is not feasible for most farmers, especially those with limited resources. So, despite its accuracy, it is a time-consuming, expensive, and destructive method that is not available to everyone.

Some of the gases and chemical reactions during photosynthesis respond to wavelengths in the infrared spectrum and can be used to evaluate the plants' well-being. Thus, SVI are formulas from the reflection spectrum that estimate flora properties and conditions. With the rising number of EOS and the continuous advance in drone technology, the relevance that SVIs can have on estimating soil keeps growing. So, unlike laboratory analysis, "soil sensing approaches such as spectroscopic techniques, (...) offer the opportunity to measure various soil chemical, physical and biological parameters in a fast and inexpensive way" [6].

AI is one of the current most studied and explored areas in computer science, where algorithms are created to imitate the capacity of human brains. Nowadays, it relies on mathematical-statistical approaches that, given a set of data, train a model to predict behaviours - ML. This area of AI has been subject to several studies in the context of remote sensing. For example, the authors in [7] developed and evaluated a method for predicting soil nutrients using ML. Others also applied ML with remote sensing techniques to estimate soil indicators, like the following articles: [8], [9], and [10].

Unfortunately, most ML algorithms thrive with extensive labeled data sets, and in this field, the

[7]Bray P1, https://tinyurl.com/2wz9ypm9

percentage of unlabeled data is far bigger than labeled ones. As we can observe from the following paper [11], where the author clearly states that there are not enough labels for SL models to train. Such lack comes from the political resistance of countries to share satellite images of their soil. Moreover, sending a team to extract and analyse soil in several points of land is an expensive and time-consuming process that countries do not consider. In fact, in [12], the author had to create a labeled data set for running soil nutrient estimation ML experiments. However, these unlabeled data sets must not be discarded, as some ML algorithms have been created to take advantage of unlabeled data given only a few labeled examples.

These algorithms belong to an area of AI which is called SSL. It's goal "is to understand how combining labeled and unlabeled data may change the learning behavior and design algorithms that take advantage of such a combination." [13]. Thus, as this topic lacks on labeled data, exploring SSL methods naturally seems to be the path to follow. Interestingly, published researches on SSL for soil nutrient estimation are extremely scarce.

1.3 Objectives and Contributions

The main objective of this thesis is to conduct an extensive study on the application of SSL to estimate soil nutrients from satellite data. To achieve such goal, the investigation will be split into four crucial steps: 1) understanding soil fertility, 2) discovering or creating high-quality data sets, 3) exploring SVI formulas, and finally, 4) finding the best SSL algorithm to solve the problem in question.

As it was stated before, fertility corresponds to the amount of thirteen atoms in the soil and some of the properties it possesses. In order to properly evaluate the algorithm's outputs, understand the data sets that are found or created, and interpret the SVI formulas to apply them to the ML algorithms, one most be able to perceive the condition of the soil by reading the quantity and property values from the land. So, to gain such knowledge, reading articles and papers related to soil fertility will be the first step.

Following it, high-quality satellite data sets must be found or created. EOS images exist in abundance, however discovering large datasets with soil nutrient quantities has been proved to be a difficult task, as there are a few organizations that supply such information. Such lack forces investigators to create their own datasets, like the author from [12] did.

Before diving into the ML algorithms that might be useful to achieve our main objective, one must expand it's knowledge of SVI. Such approach is required in order to select the appropriate formulas to feed into the AI agent, and also to discard those that are unnecessary. Like the first step, an in-depth reading of several articles will be crucial to open someone's eyes to additional information and state-of-the-art theories. There are several articles regarding the topic, like the following [14], and [15].

Lastly, selecting the SSL algorithm, or the combination of algorithms, that is better suited for the problem, will require an in-depth investigation of ML and state-of-the-art. Several articles will be read to understand which algorithms are more compatible with SVI, such as [7] and [16]. After collecting a series of SSL algorithms, tests will be conducted to select the best, or the group of bests. Then, the chosen algorithms will be implemented and integrated into the TerraSenseTK, which is a open-source toolkit for developing remote sensing ML experiments.

Overall, the investigation represents a major contribution to the AI community, since SSL methods are not widely explored in soil evaluation through satellite images. In fact, due to the lack of labeled data, it is clear that such approach should be followed. Moreover, it also contributes to the society, as it might reduce the risk of soil over-exploration and its consequences.

As a second objective, it is intended to integrate the software created into the TerraSenseTK [12], which is a python written toolkit based on remote sensing and ML, created to reduce the issues that common soil sampling methods arise. Such unification also provides to the ML community, by feeding it with SSL investigation for remote soil sampling.

During the TerraSenseTK thesis [12], the author brought up the lack of toolkits in this area of study as an issue. Additionally, he also referenced a statement from NASA presented in a ML workshop in 2020, that stated: "challenges were grouped into three major categories: 1) training data, 2) algorithms and models, and 3) tools and analytic frameworks", and underlined that the second challenge is the most crucial, as there is a "limited number of open-source software and ML frameworks to develop, evaluate and share ML models". Such constraint limits the investigation, as it overburdens the comparison of results across other studies. Therefore, another obstacle comes with our investigation: the lack of baseline implementations of ML models that allow for benchmarks with other ML models and even across data sets.

However, the presented toolkit has some limitations. It misses SSL and is purely based on classic ML. Therefore, this investigation is also contributing to TerraSenseTK, and increasing the amount of ML models that are open to the public.

1.4 Document Outline

The remaining of this document is organized as follows: The Background, provides information about all the important topics crucial to understand the work and methodology proposed in this thesis. In Related Works, we report relevant researches that were found in the literature. The Research Methodology provides an in-depth description of the methodology that will be followed. The Work Schedule presents in a time-table the chronological order of steps planned to be executed during the investigation process. Finally, Conclusion, summarizes the main motivations to pursue this work and some of gaps that were identified.

2 Background

This section provides background on the topics that will be investigated in this thesis. This includes an overview of soil fertility and nutrients, acquisition/interpretation of satellite data, and ML.

2.1 Soil fertility

As it was stated in the previous section, to assess the ML algorithms' behaviors, and to interpret data sets, one must understand what conditions the soil has to verify to be blooming and fertile.

Additionally, comprehending the plant's processes, such as photosynthesis, proves to be useful to perceive how and which SVI serve our purpose. Such topics are clarified in the following sections.

2.1.1 Cation and Anion Exchange

CEC is a measure that represents the capacity of soil to hold on to cations (positively charged ions). In another words, it is the number of anions (negatively charged ions) that a soil possesses. Since anions and cations are strongly attracted to each other, soils with higher values of CEC can capture more quantities of cations [17]. On the other hand, AEC represents the amount of cations present in the soil, which can be used to measure the capacity for it to retain anions. They are commonly measured in units of milliequivalents per 100 grams of soil (meq/100 g) [17].

According to the [18] paper, these measures are affected by the soil's pH levels. It states that "at low pH, more positive charge exists due to higher H+ on mineral edges", and that "as pH increases, H+ concentration decreases, which rises negative charges". Therefore, soils with high pH levels have higher CEC and soils with low pH possess increased AEC.

Therefore, it is extremely important to know the correct pH value of the soil before adding fertilizers, as adding too much in soils with little CEC results in leaching and environmental pollution. On the other hand, applying too few nutrients to the soil with high CEC leads to mineral deficiencies [17].

2.1.2 Soil nutrients for fertility

A crop's healthy growth requires sixteen natural elements, mineral, and non-mineral. C, H, and O, represent the only non-mineral elements that the flora requests. The other thirteen elements "are taken up by plants only in mineral form from the soil or must be added as fertilizers"[19]. N, P, K, Ca, Mg and S are all macro-nutrients, although the first three are referred to as primary elements since they are required in higher quantities than the other three, the secondary elements. The remaining micro-nutrients consist of B, Cu, Cl, Fe, Mn, Mo, and Zn. These elements occur in very small amounts in soils and plants, but their role is equally important as the primary or secondary nutrients.

Macro-nutrients (primary elements): Plants predominantly acquire C and O by inhaling CO_2 from the atmosphere, as is stated in the following article: "Plants obtain all the Oxygen and Carbon they need from the air" [19]. Despite this, organic matter is rich in C in mineral form and possesses some levels of liquid H_2O. So, bearing in mind that soil humus is made "of highly decomposed residues of plant and animal remains," they can also incorporate these elements from the land. According to [8], any plant thrives with 340ppm of CO_2, and "as its levels are raised by 1,000 ppm, photosynthesis increases proportionately, resulting in more sugars and carbohydrates available for plant growth". However, in another article [18], the author defends that C and O values in the plant must be around 450000ppm.

According to the [19] article, H is gained through liquid H_2O and H+. It is absorbed through the plant's roots, and it is found on the soil humus. H serves a great value to the herb's well-being,

[8] Carbon dioxide in greenhouses, https://tinyurl.com/2xz846za

as it contributes for photosynthesis by driving the electron transport chain, reinforces its structure, is required to create amino-acids, and supports plant respiration[9]. Following the [18] investigation, plants should contain 60000ppm of H.

N is the most versatile element required for soil fertility, as it can be used in organic and inorganic form, as a solution and as a gas, as well as a cation and an anion [19]. Plants capture this element primarily through $NO_3{}^-$ and $NH_4{}^+$ ions embedded in the soil solution. However, N is not a natural constituent of rocks or minerals. Rather, the natural state of N is as N_2 gas in the atmosphere [19]. This element has a tremendous impact on the plant's health, and one must be careful to maintain it in adequate quantities. As the previously mentioned article shows, "a good supply of N is associated with vigorous growth and a deep green color.". Yet, when missing, the flora turns "stunted and yellow." Additionally, its excess "causes plants to remain in a vegetative growth stage and delay initiation of flowering or fruiting, resulting in lowered yields of some crops" [19]. As specified by the Colorado State University[10], lands should contain 40ppm of N. Contrarily, the article [18] states that plants should possess 15000ppm of N.

P is the succeeding primary element. It is found in the soil and is mainly extracted from a mineral entitled apatite, but can also be found in other sources like "decaying plant and animal residues, humus, and microorganisms"[19]. It is absorbed through the plant's roots as one of two different anions, $HPO_4{}^-2$ or $H_2PO_4{}^-$, being the first present in soils containing "pH values greater than 7.0", while the second one "with pH between 4.3 and 7.0"[19]. This inorganic substance "stimulates young root development and earlier fruiting" by regulating energy and plant growth. Therefore, its absence leads to plant stunt and an "abnormally dark green color"[19]. As claimed by the [18] article, P quantities should be 2000ppm.

Lastly, in the category of primary elements, there is K. Plants incorporate this mineral in the form of the cation K+ through the molecule K_2O. Unlike the others, this cation does not have a main purpose. Instead, it serves to aid several parts of the plant. Around "60 enzymes require the presence of K, with higher concentrations found in the active growing points and immature seeds". It also takes part "in photosynthesis, in carbohydrate transport, in water regulation, and in protein synthesis"[19]. Thus, a plant with the right amounts of K is invulnerable to diseases, has finer growth, and is tolerant to droughts[19]. Nevertheless, when absent, plants become stunted and develop poor root systems. Such lack can be perceived through the "bronzing near the edges of lower leaves" and eventual death[19]. Hence, maintaining a proper amount of K is crucial for crop yield. As stated by the [18] investigation, K concentrations in the plant must be around 10000ppm.

Macro-nutrients (secondary elements): Regarding the secondary elements, Ca is a mineral involved in cell growth that enhances the uptake of Nitrate. It is a crucial element for the plant's health, as it contributes to root proliferation. Consequently, its absence brings an underdeveloped root system, poor plant quality, and less fruit. According to the Midweastern Bioag[11], the amount of Ca present in the soil depends on the CEC. For a CEC of 5, Ca should be 700ppm. If it were 10, it should have 1500ppm. With 15, it should be 2250ppm, with 20, 3000ppm, and with 25,

[9] Hydrogen - Crop Nutrients, https://tinyurl.com/39w89mrk
[10] CSU, https://tinyurl.com/yck3hctx
[11] MWBA, https://tinyurl.com/ufys6vh3

4000ppm. On another article, [18], the authors defend that plants should possess 5000ppm of Ca in their system.

The next secondary element is Mg. It is absorbed as the cation Mg^{2+}. Like with K, this element does not have a main purpose. Yet, it is responsible for activating "a number of enzymes and plays a role in protein synthesis, and phosphorus reactions" [19]. Without suitable quantities, plants are bound to suffer from "interveinal chlorosis on the lower leaves," and leaf edges gain a thin hint of red or purple [19]. 2000ppm is the advised concentration of Mg that should be incorporated in a plant [18].

Finally, the last secondary element, and to summarize the macro-nutrients, is the S. It is assimilated as an ion after the molecule SO_4^{2-} goes through a biological oxidation process. This mineral regulates plant growth and takes part in the "synthesis of chlorophyll and in photosynthesis reaction." Without it, leaves turn pale green or even yellow, and its dimensions are kept short [19]. To avoid such consequences, as claimed by the New Jersey Agricultural Experiment Station[12], one must ensure that the soil has more than 15ppm of S. In another article [18], it is proclaimed that 2000ppm is the healthiest amount of S a plant can have.

Micro-nutrients: Mn is the most commonly found micro-nutrient, being reported that its concentration in soils worldwide is 4000ppm. Despite this, most of it is unreachable. Hence, it is not measured by its quantity but by the soil's easiness of releasing it. This availability depends on organic matter, and soil moisture and increases with pH levels below 5.5 [19]. It is an important element since plants deprived of it will suffer from "interveinal chlorosis with dark-green veins," and its "leaves develop brown speckling and bronzing" [19]. To prevent such disastrous conditions, soil Mn should be kept between 20ppm and 200ppm, although these limits oscillate with the crop's type [19]. Such affirmation is confirmed by another article, which defends that 50ppm is the concentration [18].

Zn is the worldwide most common nutrient in deficient levels. Its availability is greatly affected by soil pH, being higher in acidic soils, and does not dwell with significant amounts of organic matter. Moreover, its uptake is depressed in the presence of excess levels of P. Therefore, Zn concentration is considerably scarce in "highly leached, acid sandy soils such as those found in many coastal regions" [19]. Different plants require distinct proportions of Zn, making it difficult to establish a critical boundary for healthy concentrations. Despite this, plants with Zn contents below 20ppm are unhealthy, being the normal ranges between 25ppm and 150ppm [19]. Contrarily, another article [18] defends concentrations of 20ppm.

Cu is another micro-nutrient that plays a major role in the plant's well-being. With a worldwide average soil concentration of 30ppm, it "normally ranges from 1 to 50 ppm" [19]. It has more presence on soils developed from fine-grained sedimentary rocks (shales and clays) than those originating from coarse-grained materials, such as sands and sandstone. Its availability is reduced at pH levels bigger or equal to 7, and high for pH levels below 5 [19]. This element should also be supervised, as its excess harms the flora. Even though the proper amounts depend on the crop type, Cu toxicity has been reported with values between 150ppm to 300ppm [19]. In the following paper [18], the authors claim that 6ppm is the required amount of Cu a plant should present.

[12]NJAES, https://tinyurl.com/4398asdr

Fe is the most abundant element in the world. However, its amplitude varies with the soil's pH values, reaching a minimum between 6.5 and 8. Like several other minerals, it contributes to the activation of enzyme systems. Its deficiency affects chlorophyll production, which is easily perceived, as it stunts the plant's vegetable growth and, in severe cases, can turn them completely white. Therefore, it is crucial to control the amount of Fe in the soil, and in most plant species, consequences emerge with concentrations below 10ppm to 80ppm [19]. According to the [18] paper, plants must present Fe concentrations of 100ppm.

B global average is around 10ppm. With a soil concentration range from 2ppm to 100ppm, it is scarcer in lands derived from acid igneous soils, freshwater sedimentary deposits, and rough lands low in organic matter. Consequently, it is highly concentrated in soils with rich humus, and its uptake is higher as pH becomes shorter [19]. This mineral is responsible for forming the conductive tissue that transports nutrients, and, when missing, the plant stunts, acquires bushy foliage and its fruits shed excessively [19]. Thus, it is critical that B soil concentration is kept below 100ppm [19]. Such claim goes along with the [18] observation, which declares that B concentrations should be kept around 20ppm.

The second to last micro-nutrient investigated was Cl, which is the most abundant of the halogens. In fact, the amount of Cl found in the soil ranges from 50ppm to 500ppm. According to [19], the soil should contain around 250ppm to maintain the crop's well-being. If it is found to be missing, plants will exhibit "wilting of leaf tips, progressive chlorosis of leaves, followed by bronzing, and finally leaf necrosis. In severe cases, plants fail to form fruit". On the contrary, the authors from [18] state that Cl levels should be around 100ppm.

The final micro-nutrient, and to sum up all elements that take part in soil fertility, is Mo. It aids the plant's enzyme system, and without it, flora suffers from leaf chlorosis, which resembles the consequences of N deficiency. Such lack is often followed by marginal curling, wilting and necrosis [19]. Despite its contribution, Mo is the nutrient required in the smallest amount, making the range between deficiency and sufficiency very narrow. In fact, plant tissue usually has between 0.8ppm and 5ppm, and consequences only appear below 0.5ppm. In the United States, soil contains between 3ppm to 15ppm of Mo, having different concentrations depending on the soil's material origin, and its pH value. According to the [19] article, igneous soils include "0.9 - 7 ppm; shale, 5 - 90 ppm; black shale, up to 300ppm; limestone, sandstone and dolomite, 3 - 30 ppm.", and "deficiencies often occur in strongly acidic soils". Moreover, the total Mo content in the soil is not a good indicator of plant response, as it is arduous to uptake. In fact, its amount is also correlated with the presence of P, which increases its availability, and S, which has the opposite effect.

As stated before, all these elements must be present in the right amount to contribute to soil fertility, and consequently the plant's well-being. Table 1 summarizes what was previously written.

2.2 Satellite Data Acquisition and Interpretation

2.2.1 Satellite Concepts

An orbit is a regular, repeating path that one object in space takes around another one. The object in orbit is called a satellite, and it can be classified as natural or artificial. Natural satellites are created by nature, like our moon, whereas artificial ones are made by humankind, like the ISS.

Table 1: Optimal Nutrient Quantities for Soil Fertility in ppm.

Nutrient	Concentration	Nutrient	Concentration	Nutrient	Concentration
C	340 - 450000	N	40 - 15000	Mn	20 - 200
H	60000	P	2000	Zn	20 - 150
O	450000	K	10000	Cu	6 - 150
		Ca	700 - 5000	Fe	80 - 100
		Mg	2000	B	20 - 100
		S	15 - 2000	Cl	100 - 250
				Mo	0.8 - 5

For our investigation, we will consider satellites as man-made objects orbiting around Earth. They will be referenced as EOS from now on[13]

There are four types of orbits, GO, LEO, MEO, and SSO. They differ in the object's distance to our planet, which ultimately means different velocities and revisiting time - Time to complete one orbit. GO satellites are 37.015 km above Earth and are constantly overlooking the same place. Therefore, they have a precise 24-hour revising time. On the other hand, LEO are much closer, placed between 161 to 322 km of Earth, and take 90 minutes to complete one orbit. Such proximity to the Earth's crust allows them to capture high-precision images [14]. MEO are similar to LEO, simply slightly further (2000km to 35000km). Lastly, SSO are LEO satellites that orbit the Earth at a 90°angle of the Equator. Its characteristics potentiate crossing geographic locations every trip at the same time. For example, it cruises over the same town at 5 pm [12]. Since SSO space capsules are close enough to Earth to extract high-quality data, plus they are sun-synchronous, this research will focus on gathering images with SSO.

2.2.2 Space Missions

Nowadays, there are a variety of organizations that possess orbiting EOS, such as the ESA, NASA, JAXA, ISRO, and CNSA.

The ESA is a space exploration agency from the European Union. The LUCAS Copernicus program is one of their programs and aims to provide detailed and up-to-date information about land use and land cover in the European Union, including information about soil nutrients. The program uses a combination of 5 constellations, S1[15], S2[16], S3[17], S5P[18] and S6[19].

NASA has also conducted some Earth observation missions. Terra, active from 1999 to 2022, was a satellite that carried a suite of instruments used to study the Earth's land, oceans, and

[13]What is an orbit - NASA, https://tinyurl.com/2yahvhau
[14]What is an orbit - NASA, https://tinyurl.com/2yahvhau
[15]S1, https://tinyurl.com/srsaxc8m
[16]S2, https://tinyurl.com/nfx9d9sf
[17]S3, https://tinyurl.com/7y3kjbmu
[18]S5P, https://tinyurl.com/29cz7jtr
[19]S6, https://tinyurl.com/3krum2nc

atmosphere [20]. Aqua, from 2002, with the purpose of investigating our world's water cycle[21]. The oldest mission, Landsat[22], first deployed in 1972, also had the purpose of studying the Earth's land cover and land use, as well as monitoring natural disasters, land use changes, and resource management. Nine satellites took part in this mission; however, only two are active. Landsat 8[23] and Landsat 9[24]. Suomi is another NASA mission of MSI with the purpose of studying Earth's crust, atmosphere, and oceans[25].

The Japanese organization JAXA also has some EOS, although some are inactive. They are the GGOS, the ALOS, the DAICHI, and the GCOM. All these missions have collected or are still extracting data from the Earth's land surface, including topography, land cover, and vegetation[26].

ISRO launched three satellite missions that can be used to estimate soil nutrients, as they consist of space crafts equipped with MSI. These missions are the RESOURCESAT-2A[27], the OCEANSAT[28], and the INSAT[29]. Bearing in mind that the RESOURCESAT-2A is the only constellation with Polar Sun Synchronous satellites exploring the Earth's crust, it is the most important one. But, since the OCEANSAT also presents a Polar Sun Synchronous orbit, it might also be helpful for the investigation.

Regarding the CNSA, even though they own some EOS missions, such as the ZY-3A[30] for Earth's mapping, the FY-3[31] for moisture, and the GF-3[32] for oceanic supervision, due to political matters, it is not expected that they supply us with relevant information,

It is interesting to notice that these organizations have plans to launch other improved EOS for crop observations, which reflects the importance of the risks mentioned in Section 1.1. One of these up coming missions is the Chime mission [33], and the Sentinel-4 and Sentinel-5, [34], all from ESA. Moreover, some satellites referenced above are terminated, but still worth mentioning, as their data might still be available and serve to aid our cause.

2.2.3 Data Acquisition

For the investigation to be successful, it is crucial that we extract data from satellites, but this process requires some knowledge that will be introduced in this sub-section.

Electromagnetic radiation is energy that travels through space at the speed of light. These radiations vary in wavelength and frequency. The higher the energy it transmits, the shorter its

[20] Terra, https://tinyurl.com/4rbmucd5
[21] Aqua, https://tinyurl.com/2rp8z96j
[22] Landsat, https://tinyurl.com/36rzevw8
[23] Landsat 8, https://tinyurl.com/mtx2wh96
[24] Landsat 9, https://tinyurl.com/2p9bjw8x
[25] Suomi, https://tinyurl.com/2p8fxke7
[26] JAXA, https://tinyurl.com/22vddw6s
[27] RESOURCESAT-2A, https://tinyurl.com/yxt3j982
[28] OCEANSAT, https://tinyurl.com/2p88bkhb
[29] INSAT, https://tinyurl.com/mutyzhh6
[30] ZY-3A, https://tinyurl.com/yxxd47tp
[31] FY-3, https://tinyurl.com/mpmxncra
[32] GF-3, https://tinyurl.com/yuez4w7s
[33] Chime, https://tinyurl.com/2p9b983e
[34] Sentienl-4/-5, https://tinyurl.com/4bua7jan

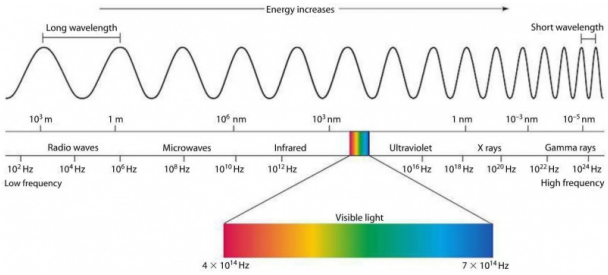

Figure 1: EMS with visible light highlighted

wavelength and the higher its frequency. The EMS is the range of all types of electromagnetic radiation. It is represented as a line and interpreted from left to right with increasing frequency and decreasing wavelength 1. Therefore, it is divided into regions conforming to the ray's properties. Radio waves, with the longest wavelengths, are used for communication, microwaves to cook food, visible light, the only region humans can capture, X-rays for medical imaging, ultraviolet to perform sterilizations, and gamma rays, for cancer treatments [12].

Remote sensing refers to the process of collecting information about an object using sensors without establishing physical contact. It can be divided into two types, active and passive. Active remote sensing involves transmitting a signal and measuring the response returned from the signal sent. A medical X-ray scan is one of the numerous examples of active remote sensing. On the contrary, passive remote sensing measures the electromagnetic radiation naturally emitted or reflected by the object under study. Our eyes, or a camera, are examples of passive remote sensing instruments.[12]

MSI are sensors or devices designed to measure electromagnetic radiation in multiple wavelengths or spectral bands. Bearing that dissimilar rays reflect different energy values, one can extract more information about the soil using MSI. Thus, EOS that possess such capabilities have higher importance to our investigation [12].

Despite this, visualizing data extracted from a MSI becomes a difficult task for the human eye. We are only capable of interpreting images from the visible region of the EMS, which can be translated to three color intensity values. All the images that humans perceive are composed of pixels, and every pixel has three color values: red, green, and blue. This translation is called the RGB spectrum and can summarize every color we observe. However, the EOS equipped with MSI might gather information in 12 different bands and become unreadable. Therefore, these hyperspectral images pass through a process that transforms a multi-dimensional image to RGB. This process is called false color, where different wavelengths of electromagnetic radiation are assigned to divergent colors to be perceived by the human eye [12]. The following image shows false color being applied to an aerial picture of Baltimore 2.

2.2.4 Spectral Vegetation Indices

SVI are mathematical formulas that provide information about the health, biomass, and other characteristics of vegetation. They correspond to the aggregation of values from different reflected wavelength rays. Several SVIs have been developed, each with its own specific purpose and set of

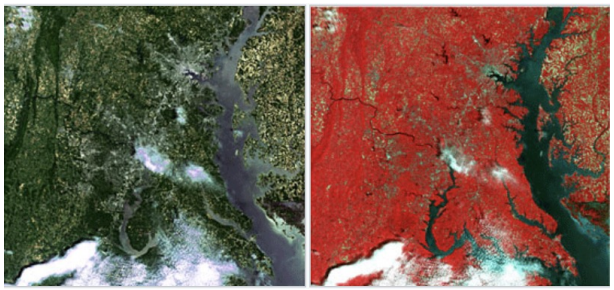

Figure 2: Landsat satellite images showing Chesapeake Bay and the city of Baltimore

assumptions. Bearing in mind that the more features a data set possess, the harder it becomes for the AI agent to compute outputs, all indices uncorrelated with soil nutrients or properties estimation are discarded from this investigation.

However, before rushing into the commonly used SVI for soil health estimation, one must understand how bands react to plant and soil properties. Only after can the formulas be perceived.

Plants absorb light in the visible spectrum (RED, GREEN, and BLUE) for photosynthesis, while they reflect light in the NIR spectrum. So, if a passive remote sensor for the RGB spectrum, that is pointing to a plant, receives low values, then one can conclude that the plant does not produce enough energy to maintain good health. However, if it were a NIR sensor, then it would mean the opposite. Moreover, when capturing information regarding lands or crops, one must take into consideration the specie of flora, as different plants require distinct amounts of energy to maintain the same health.

NDVI uses the reflected values of RED and NIR wavelengths to estimate vegetation density. It is sensitive to active photosynthetic compounds and is, therefore, a popular method used to measure the productivity of vegetation or "greenness". It has also been used to estimate soil nutrient content, such as nitrogen, phosphorus, and potassium. It is calculated according to the following formula: $\frac{NIR - RED}{NIR + RED}$. It ranges between -1 and 1, and values above 0.2 are considered indicators of vegetation [20]. However, as mentioned before, a "good" value depends on the crop in question.

EVI is similar to NDVI, but it uses blue, red, and NIR wavelengths to account for variations in atmospheric conditions and is more sensitive to areas containing dense vegetation[35]. It is calculated based on this formula, which incorporates an L value to adjust for canopy background, and C values as coefficients for atmospheric resistance: $G * (\frac{NIR - RED}{NIR + C1 * RED – C2 * BLUE + L})$. It also ranges between -1 and 1, with higher values meaning denser vegetation.

SAVI is also similar to NDVI, but it uses a correction factor to account for soil brightness in areas where vegetative cover is low. It is calculated as a ratio between the RED and NIR values

[35]EVI, https://www.usgs.gov/landsat-missions/landsat-enhanced-vegetation-index

with a soil brightness correction factor L. The factor is usually 0.5 as it accommodates most land cover types[36]. $\frac{NIR - RED}{NIR + RED + L} * (1 + L)$

MCARI evaluates the use of vegetation indices for crop nutrition mapping [21].It ranges from -1 to 1, with higher values indicating higher chlorophyll content. MCARI uses the MIR and is equal to $\frac{NIR - RED}{NIR + RED - MIR}$

These are just a few of the available and tested SVIs for soil nutrient estimation. As mentioned in Objectives and Contributions, part of the investigation will focus on discovering indices that contribute to the AI accuracy. Thus, some of the presented above might not be required, e.g., for being too correlated, and others might come to use.

2.3 Machine Learning

2.3.1 Basic Concepts

ML consists of a set o techniques for teaching computers to learn and make decisions autonomously, i.e., without being programmed to do so. It is a subset of AI, where computers learn from data and make predictions or decisions for a given input. They are taught by learning algorithms and statistical models that create patterns based on the dataset. For example, a ML agent feed with a large dataset of images labeled as "cat" or "dog", would then learn to classify new images based on patterns and features extracted from the training data. Thus, given an input image, it would respond with the class that it thinks to be the closest to.

2.3.2 Categories of Machine Learning Algorithms

There are several different types of ML, including SL, UL, SSL, and RL. In SL, the algorithm is trained on labeled data, which means that the data includes both input data and the corresponding correct output, such as the Logistic regression model[37]. On the other hand, with UL, the algorithm is not given any labeled data and must discover patterns and relationships in the data on its own. The PCA[38] model is one of the many examples of UL ML algorithms. In between these two types of algorithms, lays the SSL. It is a combination of supervised and unsupervised learning, as the algorithm trains the model with some labeled data and some unlabeled data. Self-training[39] is one of the SSL algorithms. This can be more efficient than either supervised or unsupervised learning alone because the algorithm can learn from both labeled and unlabeled data. RL is a type of ML where an agent updates its decision weights by interacting with the environment. The Q-learning[40] algorithm is one of the examples.

[36]SAVI, https://www.usgs.gov/landsat-missions/landsat-soil-adjusted-vegetation-index
[37]Logistic regression, https://tinyurl.com/34jejr56
[38]PCA, https://tinyurl.com/dd33u8c9
[39]Self-training, https://tinyurl.com/227xxyuf
[40]Q-learning, https://tinyurl.com/muuppmrr

2.3.3 Model Evaluation

Despite their divergences, they can all suffer from overfitting, which is a characteristic of ML models that were heavily trained on the data, and as a result, can not generalize unseen data. They are easily recognized, as they score high accuracy values on the training data, but act poorly with test datasets. Overfitting can occur for a variety of reasons. Some common causes are having short training datasets, and the model's complexity not matching the dataset's difficulty. To prevent such undesired events, it is common to apply regularization, which constrains the model's complexity, or cross-validation, to ensure that the model is not overfitting the training data, and to extend the training dataset[41].

Part of the ML model creation involves training and testing it on the dataset. Training a ML model involves using a dataset to learn the parameters of the input data, which enables it to make predictions on new, unseen data. On the other hand, testing a ML model involves evaluating the performance of the trained model on a separate dataset that was not used during the training process. During this phase, the model predicts the labels of the data, which are then compared to the real labels, and accounted as a TP, TN, FP, or a FN. Considering the case where a model was trained to identify cats in images, a TP corresponds to the model receiving a cat image and classifying it as such. A TN would be the model predicting "not a cat" when fed a dog picture. On the other hand, an FP would be the model stating that a non-cat image is a cat and a FN means that the model predicted a cat image to be a non-cat.

2.4 Data science

Data science is a field that involves using scientific methods, processes, and systems to extract knowledge and insights from structured and unstructured data. It involves the use of a wide range of techniques and tools from fields such as computer science, statistics, and domain expertise to analyze and interpret data. Data scientists are responsible for collecting, storing, and processing data, as well as designing and building algorithms and models to extract insights and knowledge from the data.

2.4.1 Data Profiling

Data profiling is the process of examining the characteristics of a dataset in order to understand its content, structure, and quality. It is an important step in the data preparation process for data science projects, as it helps to identify any issues or abnormalities in the data that may need to be addressed before it can be used for modeling. There are several different aspects of a dataset that can be profiled, including its structure, data types, missing values, and statistical properties.

Secondly, it is an important part of data preparation because it helps to ensure that the data is clean, accurate, and ready for analysis. It can also help to identify patterns and trends in the data that may not be immediately apparent and can provide valuable insights for data scientists as they design and build models[42].

[41] Overfitting, https://tinyurl.com/58dmv5be
[42] Data profiling, https://tinyurl.com/mpzb3uf6

2.4.2 Data Distribution and Data Balancing

The data distribution of a dataset refers to the frequency or probability of different values occurring within the data. Understanding the distribution of a dataset can be important for a variety of reasons, including identifying patterns and trends in the data, understanding the characteristics of the data, and selecting appropriate statistical techniques for analysis[43].

Data balancing refers to the process of adjusting the distribution of classes within a dataset so that each class is equally represented. This can be important when building ML models, as imbalanced datasets can lead to models that are biased towards the majority class, which can lead to poor performance for the minority class. Balancing a dataset can be an important step in the data preparation process for ML models. It is important to carefully consider the trade-offs between balancing the dataset and preserving the original distribution of the data[44].

2.4.3 Feature Selection

Feature selection is another crucial technique for datasets, as it is a method that may optimize the ML agent by reducing the number of variables in a dataset. Several methods are conducted to discover which features must be discarded, and variable correlation is one of them[45].

Correlation refers to the relationship between two variables and how they change together and ranges from -1 to 1. A positive correlation means that as one variable increases, the other variable also increases, while a negative correlation means that as one variable increases, the other decreases[46]. Understanding the correlation between features is crucial, as variables with high correlation transmit the same information. Therefore, one would remove one of these variables, as it is only increasing the dataset's size.

3 Related Works

This section provides an overview of related works. This includes an overview of the following topics: datasets for soil remote soil sensing, machine learning applications in soil sensing, and libraries/toolkits that are available for performing machine learning research with satellite data.

3.1 Datasets

In section Space Missions, a number of EOS expeditions were presented as potential hyperspectral aerial sources of soil images, which diminishes the task's difficulty. In fact, some authors used data imagery from those missions as training and testing datasets. For example, in the [12] thesis, the author took advantage of the ESDAC LUCAS Copernicus dataset[47], and in another paper[22], the investigation took advantage of the LANDSAT-7[48]. The first one scored less than

[43] Data distribution, https://tinyurl.com/3xntawpc
[44] Data balancing, https://tinyurl.com/28h7tdxs
[45] Feature selection reason, https://tinyurl.com/y7wb8a4d
[46] Correlation, https://tinyurl.com/2m9cja2e
[47] LUCAS Copernicus, https://tinyurl.com/4mt38my8
[48] LANDSAT-7, https://tinyurl.com/mwrrzbh7

10% error, and the second one had a RMSE ranging from 0.91 to 0.41. However, as their algorithms were based in SL, they could only use labeled data, which was not abundant, resulting in poorly trained models.

As the application of ML in soil sensing grows, so does the number of organizations that create high-quality datasets for training AI models. Some of them are put together from EOS missions, such as the Agriculture-Vision [23] dataset[49], and the BigEarthNet [24] dataset[50]. The first one only aggregates images from RGB and NIR channels, which might not satisfy our needs, whereas the second possesses values from the 12 different bands on the S2 satellite. Nevertheless, these datasets do not have any information regarding the number of soil nutrients. Thus, none can be used to gather labeled data, but BigEarthNet might serve to aid by supplying band values from broad crops in Europe.

The NCSS[51] Soil Characterization Database[52] is a comprehensive soil laboratory database built and maintained by the Kellogg Soil Survey Laboratory. It contains information about soil characteristics, including nutrient levels, for various locations across the United States. Moreover, it was used to train a model that predicts soil bulk density with a Random-Forest algorithm [25], where it scored an interesting RMSE of $0.13 g.cm^{-3}$.

In another article [26], the authors map soil nutrients and properties throughout the African continent with a two-scale ensemble ML model. To achieve their goal, the researchers created a dataset from a compilation of existing data, which consists of more than 100.000 soil sites. Some of the most important datasets that it includes are from the AfSIS I and II[53], the IFDC[54], the ISRIC Africa soil profile[55], the LandPKS mobile app [27], and finally, from the TAMASA[56] and AfricaRice [28]. Besides providing the quantity of each nutrient, the authors also incorporated other relevant features, including the Sentinel-2 bands B02, B04, B8A, B09, B10, B11, and B12 values[57], surface type from the DTM[58] and the NASA DEM 30 m resolution product[59], NIR and SWIR values from the Global Forest Change project[60], and ultimately surface water levels from the GSW[61]. Despite this, as the author stated, "not all soil nutrients were available at all sampling locations", which raises problems that can be overcome with missing value imputation. The paper concludes underlying the dataset's unbalance nature, as it "heavily under-represents tropical jungles or similar remote areas".

Im summary, taking into consideration the LUCAS Copernicus program, mentioned in Space Missions, the NCSS and the combination of datasets used in [26], we believe that we possess suffi-

[49] Agriculture-vision, https://tinyurl.com/yc76mb6p
[50] BigEarthNet, https://tinyurl.com/2p88evma
[51] NCSS, https://tinyurl.com/36t8xrjc
[52] NCSS-database, https://tinyurl.com/37ujj49m
[53] AfSIS, https://tinyurl.com/2aac7axa
[54] IFDC, https://ifdc.org
[55] ISRIC, https://tinyurl.com/2u7h69p6
[56] TAMASA, https://tinyurl.com/y23y2pmf
[57] Sentinel-e bands, https://tinyurl.com/336wc3nw
[58] DTM, https://tinyurl.com/94ackmmp
[59] DEM, https://tinyurl.com/5n7hmu7v
[60] Global Forest Change data, https://tinyurl.com/y8wsuwsb
[61] GSW, https://tinyurl.com/28expzc9

cient labeled and unlabeled data to develop a high-quality dataset for this research. Nevertheless, any new findings during the thesis that have the potential to enhance such dataset, will be taken into consideration.

3.2 Machine Learning and Soil Sensing

3.2.1 Learning Features

As mentioned in Datasets, the article [26] execution process begins by feeding a soil location into a RF model and outputs features of interest, which are then used to train a second model, that predicts the soil nutrient quantities. During the investigation, the author concluded that the most correlated soil/environmental parameters were "soil pH (...), soil organic carbon (...), and clay content (...)", and also praised the importance of rainfall (SM2RAIN)[62], bio-climatic, and land surface temperature (MODIS[63]) values. Consequently, focusing on "Landsat products and the Sentinel-2 bands B02 (BLUE), B04 (RED), B8A (NIR), B09 (Water vapour), B11 (SWIR1) and B12 (SWIR2)".

The article mentioned above fed band values into the ML models. However, we intend to make use of SVI as they are easily calculated and purposely created to interpret the health and productivity of vegetation, as it was verified in the following article [29], where the author used several SVI to estimate eggplant yield with ML. It is concluded that GI and GVI were the most important SVI. Such observation is coherent, as these indexes estimate chlorophyll content in the plant, which is a strong indication of plant growth.

3.2.2 Semi-Supervised Learning Algorithms

Unfortunately, to our knowledge, there are no papers where the authors applied SSL models to predict soil nutrient quantities. In fact, nearly all investigated the behavior of SL models, such as SVM, RF, MLR, and either diverged on the Earth's soil location to estimate, or the properties to predict. It is true that such lack elevates the investigation's relevance, however, its nonexistence hardens a final evaluation by comparison. Consequently, other papers with SSL investigations were considered in this review.

In the following paper, [30], the authors reinforce that DL algorithms have surpassed state-of-the-art to classify high resolution remote sensing images. However, underline their necessity for labeled data, which is scarce in this field. Thus, they create a SSL framework which combines DL features, and self-label technique. Impressively, after being experimented in 4 different datasets with a small number of labels, it scored an average accuracy of 92%. Additionally, another paper [31] reinforced the SSL strength in little labeled datasets. The investigation aimed to classify Sea-Ice on polar satellite images, where they created a SSL Generative Adversarial Network. After its implementation, they compared it with a SL model and concluded that "SSL improves the overall accuracy achieved by the SL approach by at least 5% in configurations with less than 100 labeled samples" In summary, both papers clearly state that, when labels are scarce, SSL performs better.

[62]SM2RAIN, https://tinyurl.com/ycydusz6
[63]MODIS, https://tinyurl.com/37w66ye6

3.3 Libraries & Toolkits

In software engineering, libraries are a collection of modules that provide additional functionality for a software application. These modules can include pre-written classes, functions, and variables that can be easily imported and used in a program, allowing developers to save time and avoid writing repetitive code. Some of the most useful libraries for the ML community are the PyTorch[64] library, which is "an open source ML framework" that is optimized for deep learning using GPUs and CPUs; the scikit-learn[65], that has an extensive number of useful algorithms for data analysis; and TensorFlow[66], used for implementing ML models.

On the other hand, toolkits refer to a collection of tools or modules that are used to perform specific tasks. There are many toolkits in Python that can be used with satellite imagery, however, none is related to soil nutrient estimation, which is an interesting truth due to its utility to the agricultural community and consequent impact on the world. In fact, our investigation group conducted extensive research on the matter and wrote a paper [12] which aimed to create an AI Python toolkit to conduct replicable ML experiments in remote soil sensing. In it, the author mentioned two other toolkits, Eo-learn[67] and Solaris[68]. The first one is an effort of the Sentinel Hub company to provide a set of libraries to facilitate and quicken the prototyping of complex EO workflows. Solaris aims to bridge the gap between ML and EOS. It proposes a unified data format, prepares the data in a standardized manner, trains computer vision models, generates a prediction on EOS data using common DL frameworks, and performs calculations using relevant metrics.

3.4 Summary

As we can see, there are some organizations that can supply data on soil properties and other interesting band values from locations across the world through EOS data. Nevertheless, to our knowledge, none is able to provide a single dataset with a reasonable size containing labeled and unlabeled data, forcing the creation of a new dataset. Moreover, from the section Machine Learning and Soil Sensing, we can estimate which soil properties and SVI values might be more relevant to train our models, and we can expect that the developed algorithms are better suited to reach our goals. Finally, in the toolkit section, we raised the value of this investigation as it will contribute to the only soil nutrient estimation toolkit that is presently available.

4 Research Methodology

This section presents an overview of the methodology that will be followed during the thesis development. It is divided in the following topics: the data that will be used and how it will be

[64] PyTorch, https://pytorch.org
[65] scikit-learn, https://scikit-learn.org/stable/
[66] TensorFlow, https://www.tensorflow.org
[67] eo-leaern, https://eo-learn.readthedocs.io/
[68] Solaris, https://solaris.readthedocs.io/

prepared, the process of developing and evaluating machine learning algorithms, and finally the integration of this thesis's results in the TerraSenseTK toolkit.

4.1 Data

4.1.1 Dataset Development

As it was mentioned before, possessing a high-quality and secure dataset from a reliable organization, containing all the crucial features that we underlined in Datasets, would facilitate our investigation process. Unfortunately, to the best of our knowledge, such dataset does not exist, thus the necessity to add the extra step of creating one. To do so, we will extract all the datasets previously mentioned, and join them in a single file or combination of files that share the same structure. We do not expect any inconveniences during this phase, as the data extraction seems straightforward and there are many simple ways to convert data into several files (e.g., CSV) with Python programming. Despite this, it is important to retain that our search for datasets will never cease to end, as more quality data only perfects ML models. Thus, if some are found during the investigation, they will be adjoined to our dataset and reported on the upcoming thesis.

Subsequently, we will assess every value in the dataset in order to examine if they respect the limitations and expectations discovered in Soil nutrients for fertility and Spectral Vegetation Indices. To do so, we will create Python scripts to conduct data profiling operations and, when executed, display useful information. At the moment, we do not see any reason to discover beyond the average, median, minimum, maximum, data type, data dimension, and data distribution, which can all be easily extracted using Python, with the pandas[69] library. Like the previous step, we also do not anticipate major obstacles to creating such files and displaying the intended data.

4.1.2 Dataset Preparation for Machine Learning

After conducting some data profiling techniques, we will create Python scripts to overcome any inconsistencies found and maintain a high-quality dataset. It is needless to describe these techniques as we do not have any guarantees that we will require them. Nevertheless, it is expected for some values to be missing, as the datasets combined do not possess the same features, and since EOS remote sensing commonly has failures or bad readings due to natural or technical inconveniences. To surpass these disruptions, one might remove the example with the missing value, or insert a new value from the mean, median, or even feature distribution. All of these options come with advantages and disadvantages, which will be thoroughly explored and tested.

During our research, and throughout this paper, we have concluded that there are lesser datasets including soil nutrient quantities (labels), than EOS datasets with band values (features). Consequently, the dataset created on the first step will undoubtedly be unbalanced, as it possesses little labeled data, and considerable amounts of unlabeled data. Such condition limits the model's performance, since it may not have enough examples of the minority class to learn from. This can lead to overfitting the majority class and not accurately classifying the minority. To overcome this hurdle, one must balance the dataset, being the most common techniques undersampling,

[69] pandas, https://pandas.pydata.org

oversampling and synthetic data generation. The first one consists of removing examples of the majority class until it equals the minority number. Contrary, oversampling is the duplication of examples of the minority class to equal the number of the majority. Finally, the last one involves creating new instances of the minority class by interpolation or extrapolation from existing instances, and SMOTE is one of these common approaches. It is important to know that all of these techniques have advantages and disadvantages, therefore, three datasets will be originated from the application of the three balancing techniques to the recently created dataset, and be tested in every ML model evaluated.

Furthermore, we will address the investigation with two divergent solutions. On one, we will interpret the research as a regression problem, but on another, we will adapt our methodology to treat it as a classification problem. In fact, recalling the table Optimal Nutrient Quantities for Soil Fertility in ppm., for some elements, there are no perfect quantities, but an interval of them, and for the other nutrients, soils do not suffer if they deviate -1 or +1 from the quantity displayed. Such observation means that we can transform the prediction into a classification problem, where we delimit intervals of quantities, and labeled them depending on the intensity of their effects on the soil. Moreover, there is little research done on balancing techniques with regression models [12], and it is easier to predict a label than a continuous value. To prepare our dataset for a classifier model, we will run Python scripts that transform the quantity values into labels that represent their effect on the soil. For example, bearing in mind that Ca needs to be between 700 ppm and 5000 ppm, any values in the dataset outside of these boundaries would be transformed into a 0, and the rest into a 1.

4.2 Model Development and Evaluation

4.2.1 Model Development

Following a careful creation of the dataset, we will shift our goal to the selection and preparation of the ML model. To do so, we will start by implementing some SSL algorithms mentioned in Semi-Supervised Learning Algorithms, as they scored high values in investigations related to ours. We will implement these models using the Python programming language to ensure compatibility with the TerraSenseTK. Such task is not expected to be arduous, as the language is frequently used for AI and contains several libraries with developed ML tools, like the previously mentioned in Libraries & Toolkits.

The training process allows models to learn the underlying patterns and relationships in the data, which they can then use to make predictions on new inputs. Without training, a model would not have any knowledge of the problem it is trying to solve and would not be able to make accurate predictions. Therefore, after writing the models, we will execute one of the many methods to train and test them. The most simple and frequent practice to train SSL models is the Self-training method, where the model is first trained on the labeled data, and then used to label a portion of the unlabeled data. The newly labeled data is then added to the training set, and the process is repeated until all of the unlabeled data has been labeled. Additionally, we will also apply ensemble methods, such as bagging, to increase the number of models to test, as it ensures the model selection.

4.2.2 Model Selection

By the time we reach this phase, we will have an exact number of $3 * N$ SSL trained models, where 3 represents the datasets created by the balancing techniques, and N the number of implemented models. Nevertheless, some will undoubtedly perform better than others, and to discover which ones, we will develop and implement a methodology to compare their performances. It is critical to consider that different types of models require divergent evaluation processes, therefore, we will apply two evaluation processes.

To assess the classifiers, we will rely on performance metrics such as accuracy, precision and recall, which are calculated respectively:

$$\frac{TP + TN}{TP + TN + FP + FN} \quad \frac{TP}{TP + FP} \quad \frac{TP}{TP + FN}$$

Likewise, to evaluate the regression ML models, we will calculate their RMSE, and the MAE, which measure the distance from the prediction to the label. They are calculated as following:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Subsequently, we will conduct a thorough assessment of each model, in order to select the finest one, or the combination of bests. Such selection will bear in mind the evaluation scores, overfiting status, execution time, architecture, and other characteristics that might seem relevant.

4.2.3 Model Benchmark

The previously mentioned evaluations can demonstrate how the models perform, but fail to show us if they are better than SL algorithms. Therefore, to perform this comparison, we will develop and evaluate some SL models. To do so, we will extract some common algorithms from previously mentioned Python libraries, that, during our research, were found to be related to our topic. For example, the RF model used in [12]. Subsequently, they will go through the same steps that the SSL models went, including the same data manipulation and evaluation metrics. Conclusively, we will be able to compare the best SL models with the finest SSL, and clearly state which one performs better in our conditions.

4.3 Integration in the TerraSenseTK

In this section, the reader may find how the process of code integration into the TerraSenseTK toolkit will be performed. After an initial exploration of the toolkit, we envision that the scripts developed in the scope of this thesis upgrade the existing toolkit modules, namely, dataset, algorithms and performance evaluation. To be more precise, we will integrate the created dataset in the datasets module, append the ML models into the algorithms module, and finally the evaluation scripts in the performance evaluation module.

5 Work Schedule

In this chapter the reader may visualize, in a time-table format, how the investigation process will undergo. Each step is directly correlated to the subsections described in Research Methodology.

For the first step, where we extract data from other organizations, and join them to create a dataset, we expect to take 1-2 months. We presume such period due to the difficulty one may have to get in contact with these organizations, if we require any clarification. Secondly, step Dataset Preparation for Machine Learning, can only begin when the previous step is completed, since we require data to apply data preparation methods. We presume to take 1 to three months, as the information extracted from the datasets might reveal too many inconsistencies that require its entire removal, and consequent necessity to be substituted.

Following this comes Model Development, which can only begin after the conclusion of the previous step, as the implemented models require a fully prepared dataset to train. We estimate to take one to three months, as code implementations come always with programming obstacles, and we intend to write a considerable number of algorithms. As the Model Development is undergo, we will start to work on step Model Benchmark. We decided this approach since both steps are meant to create ML models, and thus, might be done at the same time. We do not expect to take as much time as the previous step, as most of the SL algorithms are already developed in the TerraSenseTK algorithms module. Model Selection will only begin after the conclusion of the two previously mentioned steps, as to select a model we require all the implementations to be completed. Such task is not expected to take long, as Python is extremely adapted to compute performance evaluations.

Integration in the TerraSenseTK will start after the model selection, as we will only append the best models into the toolkit, which can only be pointed after the assessment of all models. As the TerraSenseTK is structurally well-organized, we do not estimate more than 2 months. To terminate, for the Thesis writing step, we presume to take the last 4 months of our work schedule. Naturally, we can only report all the investigation events and conclusions after its completion. However, as the research is rich and full of context, we will start to address it as the previous steps are coming to an end.

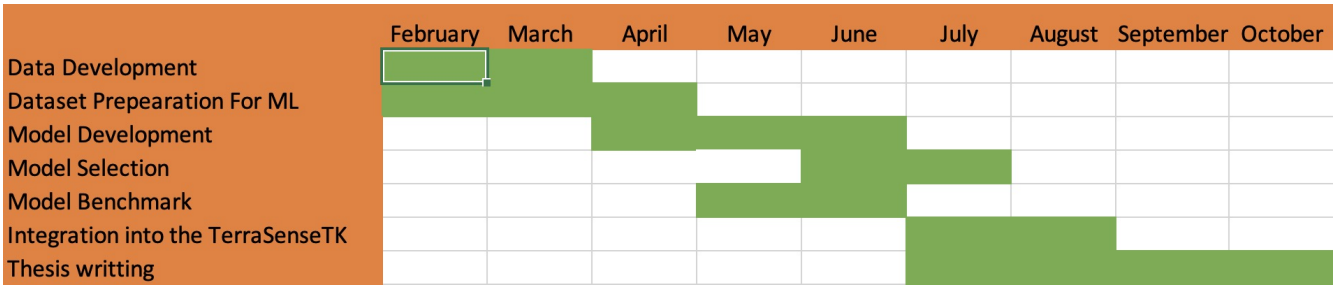

Figure 3: Work Schedule.

6 Conclusion

The never ending growth of the world population, and the rising of C atoms in the Earth's atmosphere keeps increasing our necessity for arable soil. Thus, to keep up with the escalating society's demand, soils are being over explored. Such actions might lead us into a world where fertile soils become rarer. Bearing in mind that laboratory analysis is an expensive and time consuming method, and represents the most common practice to estimate soil properties, having a fast and cheap way of analysing soils would prevent overdosing them, and ultimately, overcome these obstacles. To do so, we will improve the TerraSenseTK toolkit by updating it with all the investigation's conclusions. Additionally, our research is based on SSL algorithms, which, to the best of our knowledge, have not been explored in soil nutrient estimation through EOS data. Therefore, we believe that our investigation highly contributes to our society, AI community, and strives for a better tomorrow.

Furthermore, throughout our research, we have encountered several obstacles that difficult our procedure, being the more exigent the lack of a high-quality dataset with EOS band values and soil nutrient quantities, and the non existence of SSL investigations to predict ground nutrition amounts. To overcome such hurdles, we will create a dataset by joining all the other datasets found and described in section Datasets, and experiment on all SSL modules mentioned.

References

[1] Harold Van Es. "A new definition of soil". In: *Csa News* 62.10 (2017), pp. 20–21.

[2] Julian M. Alston and Philip G. Pardey. "Agriculture in the Global Economy". In: *Journal of Economic Perspectives* 28.1 (Feb. 2014), pp. 121–46. DOI: 10.1257/jep.28.1.121.

[3] Nikos Alexandratos and Jelle Bruinsma. "World agriculture towards 2030/2050: the 2012 revision". In: (2012).

[4] J Dewis, F Freitas, et al. "Physical and chemical methods of soil and water analysis." In: *FAO soils Bulletin* 10 (1970).

[5] J. Benton Jones Jr. "Soil testing in the united states". In: *Communications in Soil Science and Plant Analysis* 4.4 (1973), pp. 307–322. DOI: 10.1080/00103627309366451.

[6] Else K Bünemann et al. "Soil quality–A critical review". In: *Soil Biology and Biochemistry* 120 (2018), pp. 105–125.

[7] Janez Trontelj ml and Olga Chambers. "Machine Learning Strategy for Soil Nutrients Prediction Using Spectroscopic Method". In: *Sensors* 21.12 (2021), p. 4208.

[8] Freddy A. Diaz-Gonzalez et al. "Machine learning and remote sensing techniques applied to estimate soil indicators – Review". In: *Ecological Indicators* 135 (2022), p. 108517. ISSN: 1470-160X. DOI: https://doi.org/10.1016/j.ecolind.2021.108517.

[9] José Padarian, Budiman Minasny, and Alex B McBratney. "Machine learning and soil sciences: A review aided by machine learning tools". In: *Soil* 6.1 (2020), pp. 35–52.

[10] Sanjay Motia and SRN Reddy. "Exploration of machine learning methods for prediction and assessment of soil properties for agricultural soil management: a quantitative evaluation". In: *Journal of Physics: Conference Series*. Vol. 1950. 1. IOP Publishing. 2021, p. 012037.

[11] Lei Zhang et al. "A self-training semi-supervised machine learning method for predictive mapping of soil classes with limited sample data". In: *Geoderma* 384 (2021), p. 114809. ISSN: 0016-7061. DOI: https://doi.org/10.1016/j.geoderma.2020.114809.

[12] Manuel Pereira. "TerraSenseTK: A Toolkit for Remote Soil Nutrient Estimation". MSc Thesis. Funchal, Portugal: Univesidade da Madeira, Nov. 2022.

[13] Xiaojin Zhu and Andrew B Goldberg. "Introduction to semi-supervised learning". In: *Synthesis lectures on artificial intelligence and machine learning* 3.1 (2009), pp. 1–130.

[14] I Pôças et al. "Remote sensing for estimating and mapping single and basal crop coefficientes: A review on spectral vegetation indices approaches". In: *Agricultural Water Management* 233 (2020), p. 106081.

[15] Abdou Bannari et al. "A review of vegetation indices". In: *Remote sensing reviews* 13.1-2 (1995), pp. 95–120.

[16] M. Pérez-Ortiz et al. "A semi-supervised system for weed mapping in sunflower crops using unmanned aerial vehicles and a crop row detection method". In: *Applied Soft Computing* 37 (2015), pp. 533–544. ISSN: 1568-4946. DOI: https://doi.org/10.1016/j.asoc.2015.08.027.

[17] Malcolm E Sumner and William P Miller. "Cation exchange capacity and exchange coefficients". In: *Methods of soil analysis: Part 3 Chemical methods* 5 (1996), pp. 1201–1229.

[18] John L Havlin. "Soil: Fertility and nutrient management". In: *Landscape and land capacity*. CRC Press, 2020, pp. 251–265.

[19] Steven C Hodges. "Soil fertility basics". In: *Soil Science Extension, North Carolina State University* (2010).

[20] Mert Dedeoğlu et al. "Assessment of the vegetation indices on Sentinel-2A images for predicting the soil productivity potential in Bursa, Turkey". In: *Environmental Monitoring and Assessment* 192.1 (2020), pp. 1–16.

[21] Alireza Sharifi. "Remotely sensed vegetation indices for crop nutrition mapping". In: *Journal of the Science of Food and Agriculture* 100.14 (2020), pp. 5191–5196.

[22] Arman Naderi et al. "Assessment of spatial distribution of soil heavy metals using ANN-GA, MSLR and satellite imagery". In: *Environmental monitoring and assessment* 189.5 (2017), pp. 1–16.

[23] Mang Tik Chiu et al. "Agriculture-vision: A large aerial image database for agricultural pattern analysis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 2828–2838.

[24] Gencer Sumbul et al. "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding". In: *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2019, pp. 5901–5904.

[25] Amanda Ramcharan et al. "A soil bulk density pedotransfer function based on machine learning: A case study with the NCSS soil characterization database". In: *Soil Science Society of America Journal* 81.6 (2017), pp. 1279–1287.

[26] Tomislav Hengl et al. "African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning". In: *Scientific Reports* 11.1 (2021), pp. 1–18.

[27] Jeffrey E Herrick et al. "The global Land-Potential Knowledge System (LandPKS): Supporting evidence-based, site-specific land use and management through cloud computing, mobile applications, and crowdsourcing". In: *Journal of Soil and Water Conservation* 68.1 (2013), 5A–12A.

[28] Jean-Martial Johnson et al. "Near-infrared, mid-infrared or combined diffuse reflectance spectroscopy for assessing soil fertility in rice fields in sub-Saharan Africa". In: *Geoderma* 354 (2019), p. 113840.

[29] Sevda Taşan et al. "Estimation of eggplant yield with machine learning methods using spectral vegetation indices". In: *Computers and Electronics in Agriculture* 202 (2022), p. 107367.

[30] Wei Han et al. "A semi-supervised generative framework with deep learning features for high-resolution remote sensing image scene classification". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 145 (2018), pp. 23–43.

[31] Francesco Staccone. “Deep learning for sea-ice classification on synthetic aperture radar (SAR) images in earth observation. Classification using semi-supervised generative adversarial networks on partially labeled data”. In: (2020).