Machine learning applied to Sb mineralizations in northern Portugal

Morgana Carvalho

Master in Geology

Department of Geosciences, Environment and Spatial Planning

Faculty of Sciences of University of Porto

2023

Supervisor

Alexandre Martins Campos de Lima, Associate Professor, Faculty of Sciences of University of Porto

Co-supervisors

Joana Cardoso Fernandes, First Level PhD Researcher, Faculty of Sciences of University of Porto

Ana Cláudia Moreira Teodoro, Associate professor with Habilitation, Faculty of Sciences of University of Porto


U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

A geologist once sought to know
The levels of antimony below
She tried soil samples and deep learning too
But the results just wouldn't come through

She measured and analysed with great care
But the antimony was simply not there
Perhaps it was hiding, or just out of reach
But the geologist couldn't find it on the beach

So she packed up her tools and went on her way
Hoping to find antimony another day
For the search for knowledge is never done
And there's always more discoveries to be won.

Sworn Statement

I, Morgana Carvalho, enrolled in the Master Degree of Geology at the Faculty of Sciences of the University of Porto hereby declare, in accordance with the provisions of paragraph a) of Article 14 of the Code of Ethical Conduct of the University of Porto, that the content of this dissertation reflects perspectives, research work and my own interpretations at the time of its submission. By submitting this dissertation, I also declare that it contains the results of my own research work and contributions that have not been previously submitted to this or any other institution. I further declare that all references to other authors fully comply with the rules of attribution and are referenced in the text by citation and identified in the bibliographic references section. This dissertation does not include any content whose reproduction is protected by copyright laws. I am aware that the practice of plagiarism and self-plagiarism constitute a form of academic offense.

Morgana Carvalho

September 30th, 2023

Agradecimentos

Agradeço primeiramente aos meus orientadores, Prof. Dr. Alexandre Lima, Dra. Joana Cardoso Fernandes e a Prof. Dra. Ana Cláudia Teodoro pelo conhecimento compartilhado e pela dedicação com que me orientaram. Em especial agradeço ao Prof. Dr. Alexandre Lima, pela confiança depositada e oportunidades que proporcionou dentro da faculdade que sem dúvida contribuíram imenso para meu desenvolvimento. À Dr. Joana Fernandes pela energia e entusiasmo com que compartilhou seu conhecimento e novas ideias sempre com muita disposição. À Prof. Dr. Ana Teodoro pela confiança e oportunidade de trabalhar em um contexto inovador nas geociências. bAgradeço também ao projeto AUREOLE, http://doi.org/10.54499/ERA-MIN/0005/2018 , que proporcionou a coleta e analise dos dados de campo.

A todos do espaço geologia, Giulia, Ana, Antônio, Filipa, Ricardo, Douglas, Ariane e Filipe, pelos momentos de descontração, pela paciência, pelo apoio. Valorizo muito a vossa presença o conhecimento construído no dia a dia.

Agradeço em especial à minha amiga Giulia, que esteve ao meu lado durante estes dois anos não só de mestrado, mas também de Portugal, aprendendo e ensinando, entre aventuras e desventuras.

Por fim, agradeço também a todos os meus professores que compartilharam seu amor pela geologia e pela ciência e que calçaram os alicerces que me motivam a seguir esse caminho com muita alegria.

Resumo

O antimónio (Sb) ganhou relevância como matéria-prima crítica dentro da União Europeia (UE) devido à sua importância estratégica em vários setores industriais, especialmente na indústria têxtil, como retardador de chamas e como componente de materiais semicondutores à base de Sb. Além disso, o Sb está emergindo como uma alternativa potencial para uso futuro em conjunção com o ânodo em baterias de iões de lítio, que são um elemento-chave na transição verde. Este estudo concentrou-se na avaliação das mineralizações de Sb através da assinatura espectral dos solos utilizando algoritmos de aprendizado de máquina. O objetivo desta dissertação foi explorar a viabilidade de identificar mineralizações de Sb usando técnicas de sensoriamento remoto não invasivas, reduzindo assim o tempo e o custo geralmente associados às análises geoquímicas tradicionais. A pesquisa utiliza dados de solo previamente coletados no projeto AUREOLE, que teve como objetivo avaliar a distribuição de Sb nas regiões de Ribeiro da Serra e Tapada, no norte de Portugal. Além disso, o estudo avalia a utilidade do *portable X-ray Fluorescence Analyzer* (pXRF) na medição direta das concentrações de Sb em amostras de solo bruto. Algoritmos de aprendizado de máquina, incluindo Redes Neurais Convolutivas (CNN), *Random Forest* (RF) e Regressão de Mínimos Quadrados Parciais (PLSR), foram empregados para analisar os dados. Enquanto RF e regressão PLSR apresentaram resultados insatisfatórios, o modelo CNN mostrou resultados mais promissores, apesar dos desafios relacionados ao sobreajuste do modelo. Embora previsões precisas para as concentrações de Sb não tenham sido alcançadas, este estudo fornece insights valiosos sobre estratégias potenciais para pesquisas futuras neste campo.

Palavras-chave: Antimónio, mineralização de Sb, aprendizado de máquina, sensoriamento remoto, análise de solo, matérias-primas críticas.

Abstract

Antimony (Sb) has gained significance as a critical raw material within the European Union (EU) due to its strategic importance in various industrial sectors, particularly in the textile industry for flame retardants and as a component of Sb-based semiconductor materials. Moreover, Sb is emerging as a potential alternative for use in lithium-ion batteries, a key element in the green transition. This study focused on evaluating Sb mineralizations in a selected area in northern Portugal through the spectral signature of soils using machine learning algorithms. The objective of this dissertation is to explore the feasibility of identifying Sb mineralization using non-invasive remote sensing techniques, thereby reducing the time and cost typically associated with traditional geochemical analyses. The research leverages soil data previously collected in the AUREOLE project, which aimed to assess Sb distribution in the Ribeiro da Serra and Tapada regions of Northern Portugal. Additionally, the study evaluates the utility of a portable X-ray Fluorescence Analyzer (pXRF) in directly measuring Sb concentrations in raw soil samples. Machine learning algorithms, including Convolutional Neural Networks (CNN), Random Forest (RF), and Partial Least Squares Regression (PLSR), were employed to analyze the data. While RF and PLSR regression yielded unsatisfactory results, the CNN model exhibited more promising outcomes despite encountering challenges related to overfitting. Although accurate predictions for Sb concentrations were not achieved, this study provides valuable insights into potential strategies for future research in this field.

Keywords: Antimony, Sb mineralization, machine learning, remote sensing, soil analysis, critical raw materials.

Table of Contents

List of Tables:

No table of figures entries found.

List of Figures:

Acronyms

Au - Gold

ANN - Artificial Neural Network

CACNN - Convolutional Autoencoder

Cd - Cadmium

CNN – Convolutional Neural Network

Cr - Chromium

CRMs - Critical Raw Materials

Cu - Copper

DCZ - Douro Shear Zone

EU - European Union

Fe - Iron

Fe_2O_3 - Ferric oxide

FWHM - Full-Width Half-Maximum

Hg - Mercury

ICP-MS -Inductively Coupled Plasma Mass Spectrometry Analysis

IDW -Inverse Distance Weighting

LVs - Optimal Latent Variable Numbers

ML - Machine Learning

MLR - Multiple Linear Regression Analysis

Mn - Manganese

MSE – Mean Square Error

Ni - Nickel

NIR - Near Infrared

OM - Organic Matter

P - Phosphorus

Pb - Lead

PCA - Principal Component Analysis

PLS - Partial Least Squares

PLSR - Partial Least Squares Regression

pXRF - X-ray Fluorescence Analyzer

R^2 - Coefficient of Determination

RF - Random Forest

RFR - Random Forest Regression

RMSE - Root Mean Square Error

RPD -Performance-to-Deviation Ratio

S - Sulfur

Sb – Antimony

SG - Savitzky-Golay method

SiO2 - Silicon dioxide

SOM - Soil Organic Matter

STFT - Short-Time Fourier Transform

SVM - Support Vector Machine

SWIR - Short-Wave Infrared

Ti - Titanium

TiO_2 - Titanium dioxide

V - Vanadium

US – United States

VIS -Visible

VNIR - Visible and Near-Infrared

Zn - Zinc

1 Introduction

Antimony (Sb) is currently considered a critical raw material to the European Union (EU), being strategic to its economy, in a scenario where the global market of Sb is dominated by China. Sb has its application mostly in the textile industry as flame retardants and as Sb-based semiconductor materials. Sb is also being studied as an alternative for use in lithium-ion batteries, which have importance for the green transition. In this context, is important to develop knowledge about the Sb resources in Portugal.

The objective of this dissertation is to verify the possibility of identifying Sb mineralizations in a selected area in northern Portugal through the spectral signature of the soils by applying machine learning algorithms.

Several authors (Guo et al. , 2022; Rodrígues-Peéres et al.; 2021; Zhou et al., 2021; Pyo et al., 2020; Cheng et al., 2019; Nanni Demattê, 2006; Kemper & Sommer, 2002) refer to the major advantage of developing machine learning (ML) methods to quantify and qualify heavy metals in soils is the possibility of analysing large-scale zones faster, avoiding the high cost and time demanded that is implicit in the traditional approach of geochemical analysis. Remote sensing is a non-invasive technique capable of identifying targets for mineral exploration reducing costs and avoiding environmental impacts.

The soils for this study were obtained and analyzed previously in the scope of the AUREOLE project, with the objective of verifying and characterising the Sb distribution in Ribeiro da Serra and Tapada, Northern Portugal, and also included the evaluation of the pXRF (portable X-ray Fluorescence Analyzer) tool in measuring Sb directly from the raw soil samples.

The ML algorithms tested in this dissertation were CNN (convolutional neural networks), RF (Randon Forest) and PLSR (Partial Least Squares Regression). RF and PLSR regression could not provide satisfactory results. The CNN model shows more promising results, but the overfitting problem couldn’t be avoided. Despite not having achieved good results for the Sb predictions, this study provides insights about which strategies could be incorporated into future studies.

2 Geographic Setting and Historical Context

The former Sb-Au mining concessions of Ribeiro da Serra and Tapada are in northern Portugal, located in the Gondomar municipality, at Serra das Flores, approximately 30 km from the city of Porto (Figure 1). The vertices coordinates and coordinate system are in the table 1:

Table 1: Vertices coordinate of the study area.

	WGS 1984 UTM Zone 29N	
	E	N
NE	547 625.50	4 548 197.27
SE	547 625.50	4 546 915.45
SO	546 557.79	4 546 912.60
NO	546 557.79	4 548 202.98

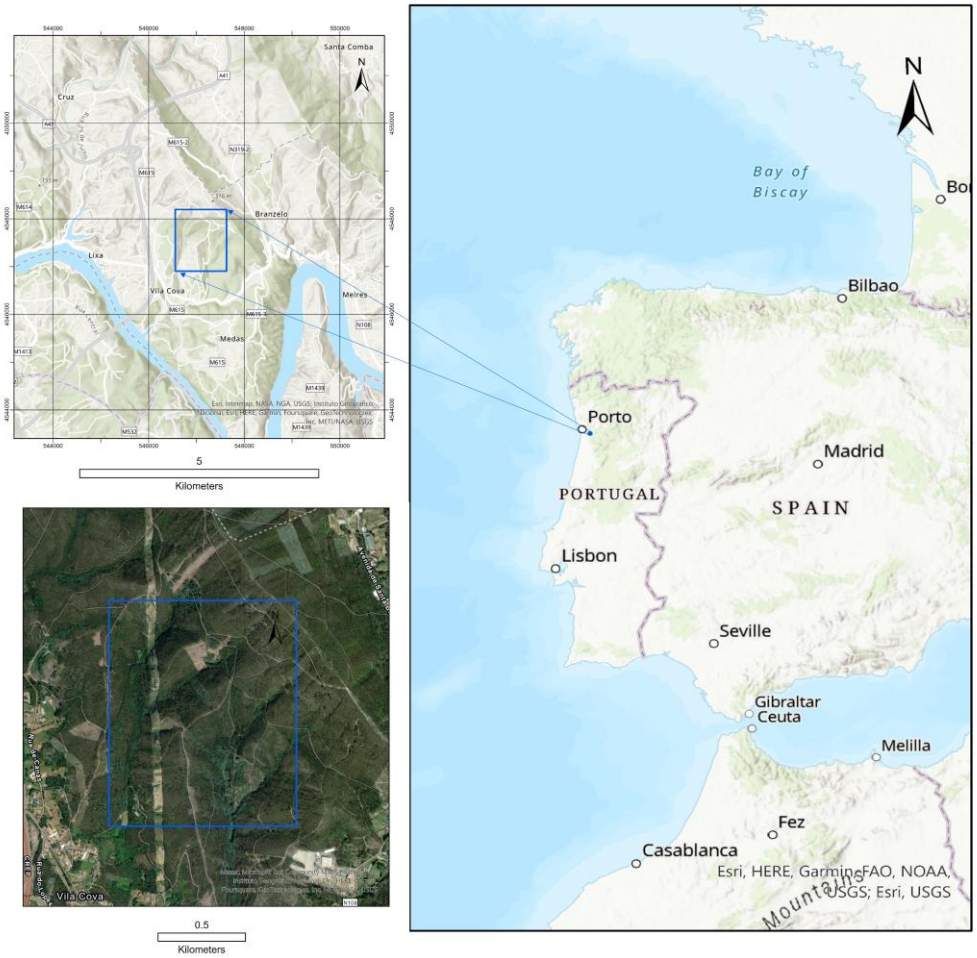

Figure 1: Location of study area in Portugal.

Sb was mostly exploited in Portugal in the late 19th century, having considerable importance in the districts of Valongo and Gondomar (Dúrico-Beirão district), with several underground mines. The Tapada and Ribeira da Serra mines (Figure 2) were opened in 1880 and 1881 respectively, producing annually thousands of tons of antimonite concentrates for exportation. Those mines are situated in the Valongo Anticline, a regional structure with several Au-As, Sb-Au, Sn-W and Pb-Zn-Ag deposits and a long history of exploration since the Roman era.

Figure 1: Ribeiro da Serra Infrastructures in the late 19th century and nowadays (Frutuoso, 2018).

The stibine was manually concentrated and by mechanical processes in hydroclassifiers and, at the end of the century, the Sb produced by Portugal had global importance, influencing the prices in the London stock exchange. In the first years of the 20th century, the European market experienced a crisis caused by the competition with the Asiatic Countries, leading to the closure of Portuguese mines. During the Second World War, there was an increase in mining activity, and since the 1960s some prospecting campaigns and reconnaissance studies have taken place (Couto, 1993). Both former concessions produced ~1 200 T of antimony and 2 T of gold, partially recovered (Couto et al., 1990, Moura, Velho, 2012.), while in the Dúrico-Beirão district, the most important mines produced about 12.000 t of Sb (Neiva et al, 2008).

3 Geological setting

The study area is encompassed within map sheet 13-B (Castelo de Paiva) at a scale of 1:50,000 and is located in a section of the Valongo Anticline, situated within the Iberian Central Zone.

In the study area, the lithostratigraphic succession consists of rocks with very low-grade metamorphism from the Cambrian/Pre-Cambrian (pre-Ordovician), Ordovician, and Carboniferous periods, composed of schists and some quartzites, and the Upper Carboniferous comprising breccias, conglomerates, and quartzites. The lithologies vary from east to west, with ages corresponding to Lower Ordovician, Middle and Upper Ordovician, Lower Carboniferous, and pre-Ordovician.

The Au-As, Sb-Au, Sn-W and Pb-Zn-Ag deposits in the Valongo Anticline are situated in the Iberian Massif, within a Proterozoic and/or Paleozoic unit that constitutes the occidental side of the Central Iberian Zone. Mineralizations in the form of veins are preferably located within the anticline's flanks, filling transverse fractures within this structure.

3.1 Valongo Anticline

The Valongo Anticline is a Variscan structure with a N150-N160° (NNW-SSE) orientation, dipping towards the NW. The anticline is asymmetric, with the reverse flank (northeastern side) gently inclined (around 40°) and the normal flank (southwestern side) sub-vertical to slightly inclined. The reverse flank is affected by the Douro Shear Zone, characterized by sinistral movement (Couto & Roger, 2017). This reverse flank extends for about 20 km from Valongo to the east of Castelo de Paiva, where it is cut by Variscan granites. The Sb deposits are in the western flank of the Valongo anticline, in the Dúrico-Beirão Mining District (Figure 3).

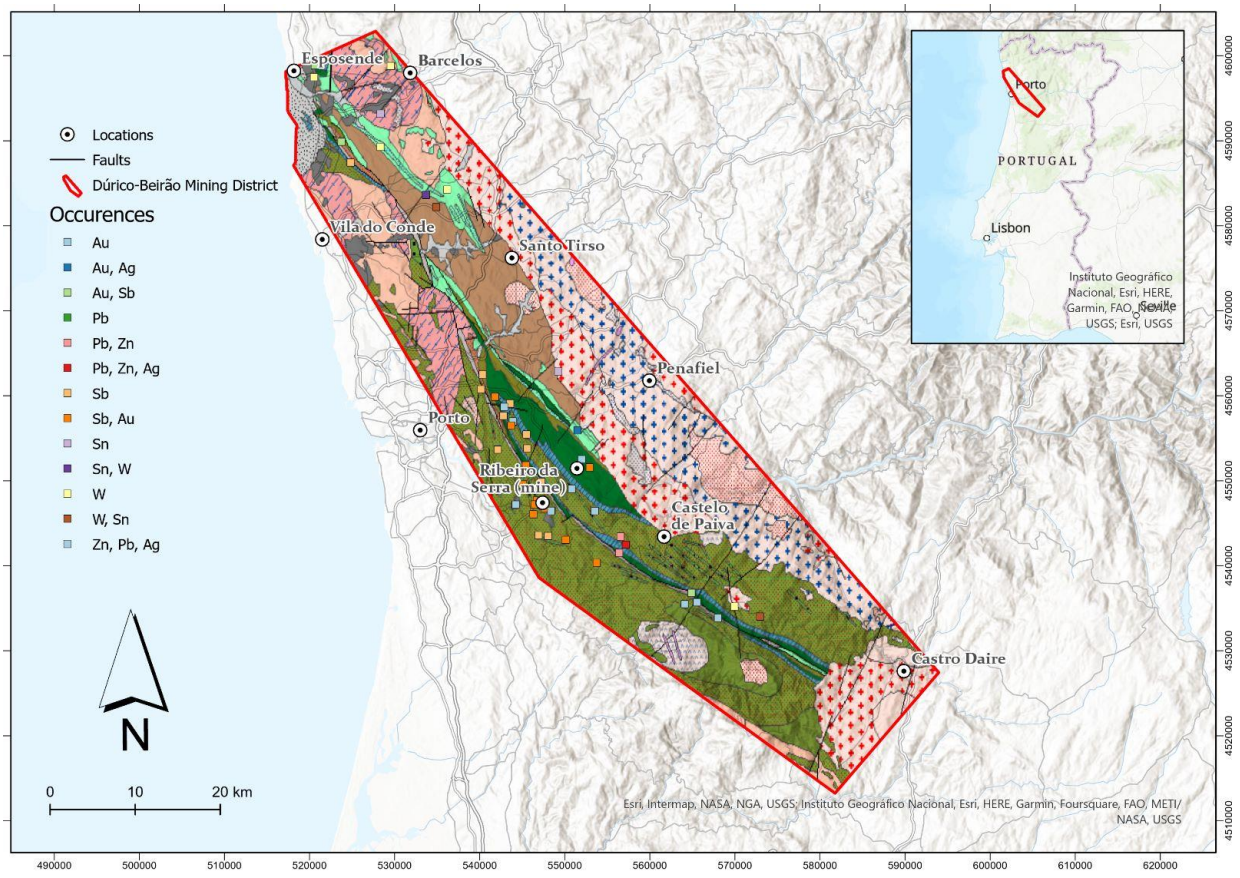

Figure 3: Location of the Dúrico-Beirão Mining District and mineral occurrences according to SIORMINP (Sistema de Informação de Ocorrências e Recursos Minerais Portugueses) (Patinha et al., 2020).

3.2 Lithologies

The target Sb-Au deposits are hosted in country metasedimentary rocks from the Cambrian/Precambrian Schist-Greywacke Complex (pre-Ordovician), also known as Dúrico Beirão supergroup, and breccias from the lower Carboniferous (Couto, 1993). The lithologies of the study area vary from West to East, ranging from the pre-Ordovician schists and greywackes to lower Carboniferous breccias, conglomerates, quartzites, black schists and lydites, as well as Ordovician schists with some quartzites (Figure 4).

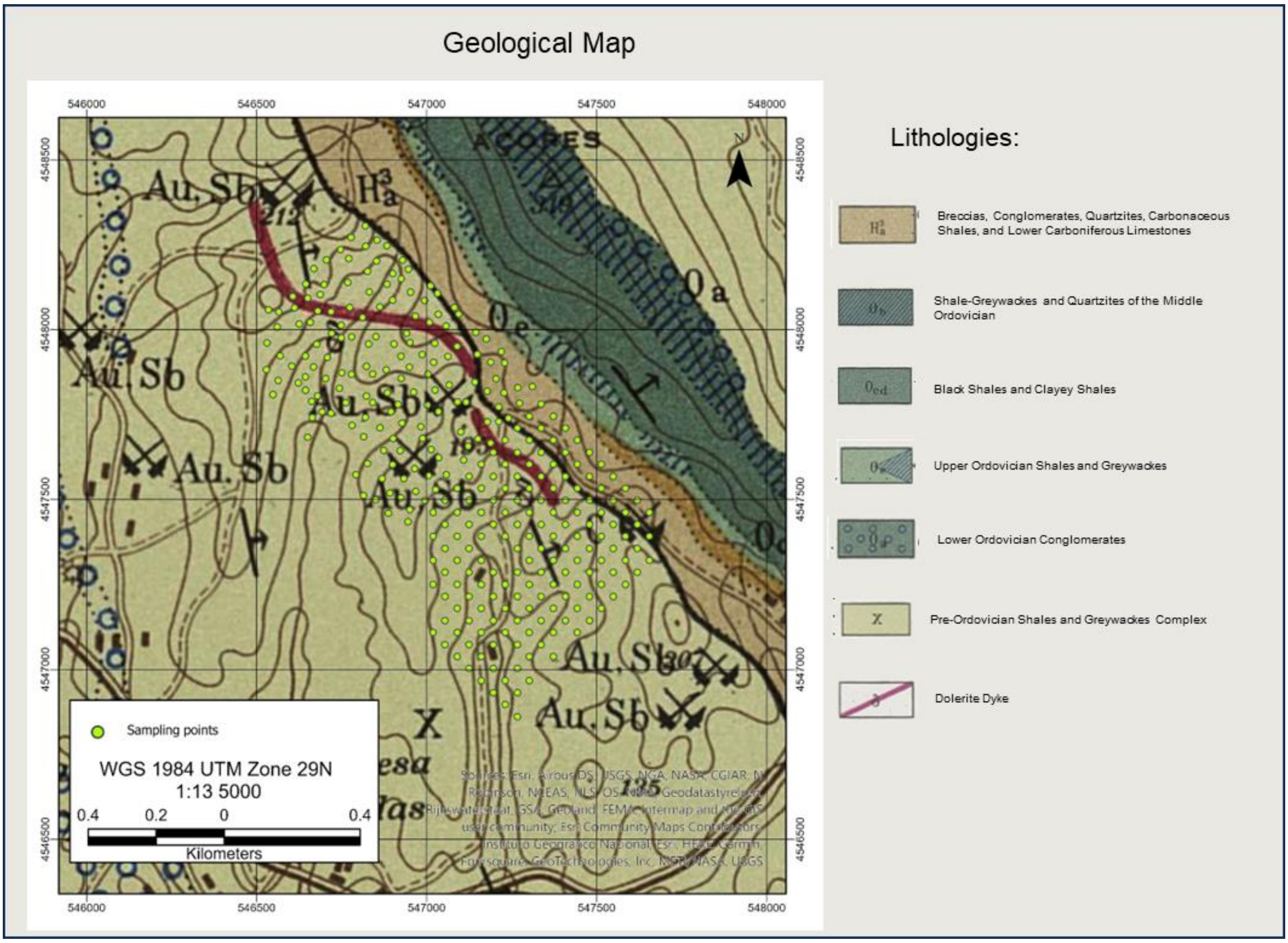

Figure 4: The sampling points and lithologies in the study area.

3.2.1 Breccias, Conglomerates, Quartzites, Carbonaceous Shales, and Lower Carboniferous Limestones.

The Carboniferous units are unconformably overlain on the Schist-Greywacke Complex. The Carboniferous formation consists of conglomerate breccias composed of pebbles of quartzites measuring 1 to 5 cm in diameter. Its cement is gray, siliceous-clayey, characteristic of elements derived from older neighbouring rocks like those of the Schist-Greywacke Complex. This conglomerate is likely related to deposits in a primitive basin, representing a slope formation (Medeiros, 1964).

The conglomerate displays well-stratified beds with a weathered layer formed by oxides. Its dip varies between 60° and 75° to the NE. The conglomerates can be coarse and predominantly composed of quartz.

3.2.2 Upper Ordovician Shales and Greywackes

The greywackes are slightly schistose, have light tones, are micaceous, and are often cut by thin veins of milky quartz. These rocks can also frequently be found in shades of grey or reddish colours. The boundary between the upper and middle Ordovician is subtle due to the lack of paleontological markers in the study area and the frequent alternation between light and dark-toned shales.

3.2.3 Black Shales and Clayey Shales + Shale-Greywackes and Quartzites of the Middle Ordovician

These are fine, dark shales. They form a northwest-southeast oriented band following the earlier formation recorded in map sheet 13-B across its entire extent. These shales have orientations ranging between N140°-165° and inclinations varying between 50°-80°NE.

3.2.3 Quartzites Intercalated with Shales of the Lower Ordovician

More competent rocks shape the ridges of the anticline's flanks in the Serra das Flores and Serra de Santa Justa. These are compact quartzites, sometimes crossed by thin veins of quartz. They alternate with layers of competent shales in dark grey tones. These layers are oriented N030°-N040° and dip 45°-85°NE according to Medeiros (1964).

3.2.4 Lower Ordovician Conglomerates

These conglomerates are found at the base of the Ordovician, exhibiting variable thickness, and almost always accompany the quartzites intercalated with shales of the Lower Ordovician, but in discordance with the pre-Ordovician Shale-Greywacke Complex. According to Medeiros (1964), these conglomerates, which are also found on the anticline's underside, are associated with the beginning of the transgressive movement during the Ordovician. They consist of rounded quartz pebbles with varying dimensions, sometimes reaching up to 5 cm in diameter.

3.2.5 Pre-Ordovician Shales and Greywackes Complex

The fine and clayey shales alternate with fine greywackes, displaying various tones including red, grey, green, and yellow. The shales are both micaceous and quartzose. From a structural perspective, joints and fractures in the shale can be observed, oriented in the same direction as secondary regional fractures, N040°, caused by the Douro Shear Zone (DCZ) (Medeiros, 1964).

3.2.6 Dolerite Dyke

Dolerites, also known as diabases, are hypabissal rocks of basaltic composition, characterized by medium-sized grains and an ophitic texture (with euhedral and anhedral crystals). They are frequently found in the form of dikes and sills (Le Maiter et al. 2003).

The dolerite in the study zone is highly altered. It is hosted in the Shale-Greywacke Complex. Its thickness can vary between 1 and 10 meters. According to Medeiros (1964), other dolerite dykes can be found near old Sb mining sites.

3.2.7 Sb and Pb-Sb Quartz Veins

The sedimentary rocks that host the veins are mainly Silurian schists. The genesis of those deposits is associated with granitic intrusions, or due to fluid mixing of CO_2-rich metamorphic fluids by surface-derived H_2O–NaCl fluids (Neiva et al, 2008). Carvalho (1969) referred that the veins that occur in the Schist-Greywacke Complex and derived metamorphic series, the Au-Sb deposits are mostly contained in the schists. These deposits have Sb with low volume, are discontinuous and have a hydrothermal nature as a result of the Variscan granites. The dominant directions of the country rocks go from N to NW dipping to W. The most productive veins occur in the EW direction dipping to N (Tapada) and NS dipping to W (Ribeiro da Serra). Ferreira (1970) reports that the Sb is present as stibnite, berthierite, and their oxides and hydroxides alteration forms. In addition, pyrite, sulfosalt of antimony, silver (pyrargyrite?), gold, and quartz are other minerals in the zone.

4 Antimony

The purpose of this chapter is to give an overview of the historical uses of Sb, its application nowadays, and its main producers at the global level, besides the main types of deposits in which Sb occurs. Also, to describe the Sb mineralization in northern Portugal and then briefly frame the Sb exploitation in Portugal's history.

4.1 Definition and uses

The element Sb is included in group 15 (VA) of the periodic table, located in the second long period of the table between Sn and Te. It is classified as a non-metal or metalloid and may exhibit a valence of +5, +3, 0, or -3, with metallic characteristics in the trivalent state. Sb has two stable isotopes, both abundant, with masses of 121 (57.25%) and 123 (42.75%) (Li, 2011).

Sb and its mineral sulphides are reported to have been used by humans at least since 4000 B.C. One of its reported uses in more ancient times is as a main ingredient of a black paste, named kohl, used for colouring eyebrows and lining eyes by Egyptians and others in early biblical times. An ornamental vase found at Tello, Chaldea, is reported to be cast Sb and date of 4000 B.C. The given name to the metal, stibium, is attributed to Pliny the Elder (50 AD), while “antimonium” is reported to be referred to by an Arabian alchemist, Geber, living in the eighth century (Butterman & Carlin, 2004) (Li, 2011).
A scientific treatise about the element Sb was written by Nicolas Lemery (1645–1715), containing results of his investigations about the proprieties and different preparations of mineral Sb, which was believed to be an important component in the alchemical lore, actuating as a magnet for extracting mercury, a key component for making the Philosopher Stone (Wisniak, 2005).
Nowadays the main uses of Sb in the EU remain in flame retardants, lead-acid batteries, lead alloys, catalysts, and stabilisers for plastics, and in the glass and ceramic industry (Fig. 5) (Latunussa et al., 2020).

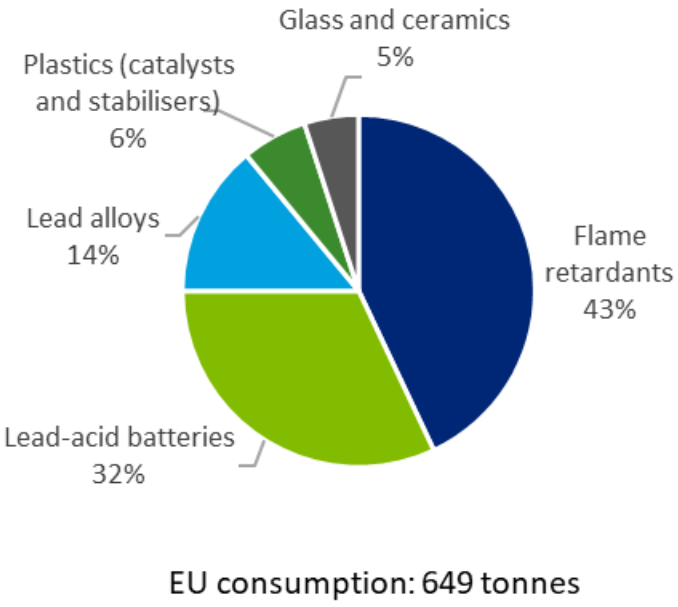

Figure 5: End uses of Sb in the EU (2012-2016) (Eurostat, 2019a, in Latunussa et al., 2020**).**

Currently, in the context of the growing demand for electric vehicles, with the urge to mitigate the use of fossil fuels, Sb is being studied as a promising alternative anode for use in lithium–ion batteries (Moolayadukkam et al., 2022). Today graphite is mainly employed as an anode in lithium-ion batteries and sodium-ion batteries, although Sb is also considered due to its structure, with a potential for a much better electrical conductor (He et al., 2018)

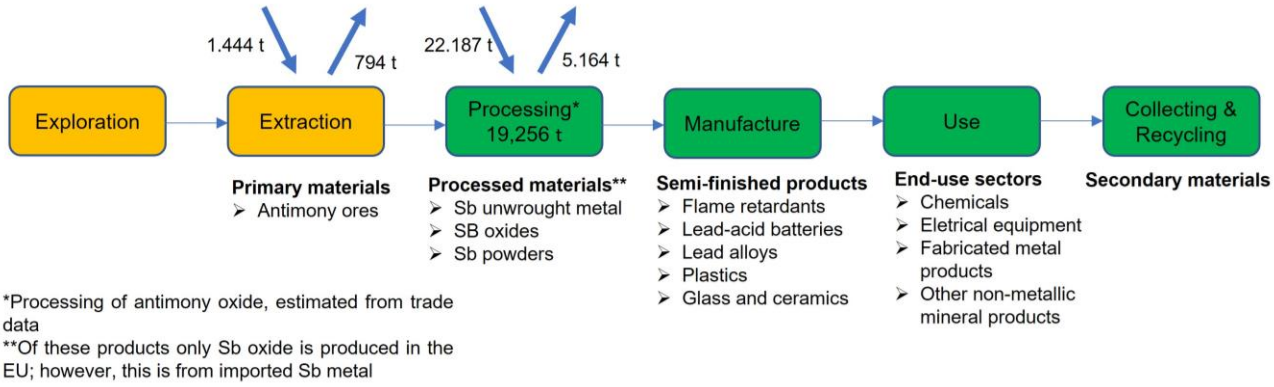

Figure 6: Simplified value chain for Sb for the EU, averaged over 2012-2016 (Latunussa et al., 2020).

Sb is contained in the EU's list of Critical Raw Materials (CRMs), a list containing materials, mostly minerals, that are considered strategic to the EU's economy and have high supply risk due to geopolitical, geographical, and geological factors, or others. The list is updated every three years (2011, 2014, 2017, 2020) to keep up with the rapid evolutions of the industry. In 2020, 30 minerals or groups of minerals were on the list. Sb is also recognized as a critical raw material/mineral in the US, Japan and Australia.

For the period 2012-2016, the EU was a major global producer of Sb oxide, but the production of Sb oxide depends on unwrought Sb, that comes from importation (see the simplified value chain for Sb in Fig. 6). Most of the Sb imported, as Sb metal and Sb oxides, originated from China, the annual import was 17,650 tonnes between 2012-2016, corresponding to 40% of the total EU sourcing (Latunussa et al, 2020) (Fig. 7).

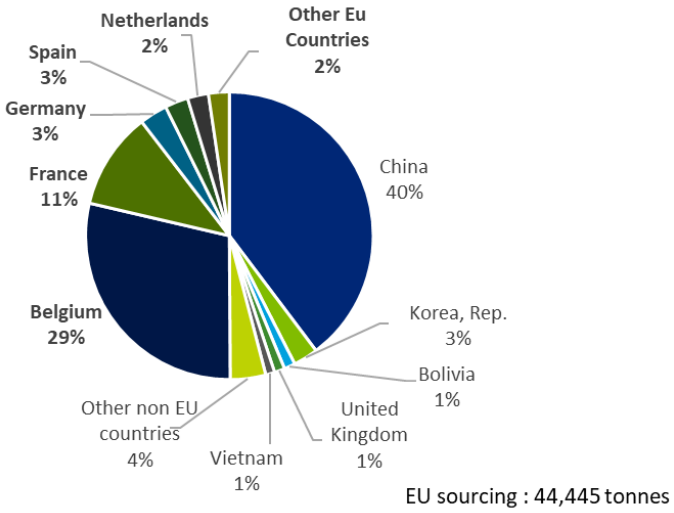

Figure 7: EU sourcing of unwrought Sb and Sb oxides (processing stage) (annual average for 2012-2016) (Latunussa et al., 2020).

5 State of the art

Zhou et al. (2021) analysed the heavy metal occurrence of Mn, Cu, Zn, Pb, Cr, Ni, related to soil organic matter and clay minerals, and iron-manganese oxides. They applied transformation processes as first derivative, second derivative, inverse-log, continuum removal, and multiple scattering corrections. They used Partial Least Squares (PLS), Support Vector Machine (SVM), and RF to build an inversion model to predict the soil heavy metal content, obtaining the best results with RF.

Rodríguez-Pérez et al. (2021) used a portable spectroradiometer (350–2500 nm) and PLSR methods to predict soil proprieties from soil samples from four Vineyards from central and north-western Spain. The preprocessing methodology involved several steps to enhance the spectral signatures for the analysis of soil properties: Outlier Identification, Spectral Averaging, Wavelength Grouping, and Scatter Correction. The wavelength grouping involved categorizing the data into four distinct spectral subsets:

- Visible (VIS) range: 350–700 nm
- Near-Infrared (NIR) range: 701–1000 nm
- Short-Wave Infrared (SWIR) range: 1001–2500 nm
- Full range: 350–2500 nm

For validation, they used the coefficient of determination (R^2), and root mean square error (RMSE) and the best performance was obtained for the pH, electrical conductivity, and phosphorous.

Pyo et al. (2020) used a convolutional neural network (CNN) to estimate As, Cu and Pb soil samples were taken from a mining area located in the Geum River watershed of South Korea. They used CNN with convolutional autoencoder (CACNN) and CNN and conventional MLmodels such as artificial neural network (ANN) and random forest regression (RFR). Dimensionality reduction was applied to the soil reflectance as a convolutional autoencoder for the CNN model and has principal component analysis (PCA) for ANN and RFR. They excluded regions from the spectra with poor signal-to-noise ratios in those regions. They obtained the best results with CACNN and CNN models, however, concluded that conventional ML and deep learning methods also show reasonable soil heavy metal estimation accuracy using soil reflectance spectra.

Nanni & Demattê. (2006) used soils of an agricultural zone of Rafard County, São Paulo State, Brazil, to compare remoting sensing techniques. They used a TM-Landsat 5 image and soil measurements obtained in the laboratory with a spectroradiometer sensor. They

applied convolution to the laboratory spectral measurements to simulate wavelength ranges corresponding to satellite sensor bands. To make statistics easier they choose twenty-two specific wavelengths (or ranges of wavelengths) to use instead of the entire spectrum. They were able to find high R^2 for most of the variables through multiple regression equations by estimating soil attributes.

Cheng et al. (2019) examined the feasibility of using soil reflectance spectra to estimate the concentrations of Cd, Pb, As, Cr, Cu and Zn in a suburban area of Wuhan City, Hubei Province, China. The concentration of the metals Pb, As, Cr, Cu and Zn in the soil samples were determined by Inductively coupled plasma atomic emission spectroscopy (ICP-AES). They removed the wavelengths at intervals of 350–399 nm and 2401–2500 nm to reduce random noise. They observed how different preprocessing treatments interfere with the results of the prediction model. They utilized the performance-to-deviation ratio (RPD), which the other authors didn't use, and the coefficient of determination (R^2), and RMSE, which are the common methods of evaluation. They also applied statistical analysis to ascertain the estimation mechanism, focusing on relationships between soil reflectance spectra and concentrations of Soil Organic Matter (SOM), Fe, and heavy metals (Cd, Pb, As, Cr, Cu, and Zn). They concluded that employing Savitzky-Golay spectral pre-treatment yielded favourable PLSR models, and more studies are needed to establish internal relationships between heavy metal concentrations and other spectrally active elements like clay minerals.

Kemper & Sommer (2002) aimed to estimate the special distribution of heavy metals using reflectance spectroscopy. They collected samples from an area in Spain flooded with sludge that resulted from a break of a mine tailings dam in Spain. This accident resulted in a very contrasting mineralogy between the background soil and the contaminated zone. They used XRF to estimate the major and trace elements. The used a spectroradiometer covering the range between 350 and 2500 nm. The preprocessing was made following the steps:

1. Spectral processing was conducted using IDL/ENVI software.

2. Spectra were degraded from 1 nm to 5, 10, and 20 nm using a Gaussian model that considered the band's centre and full-width half-maximum (FWHM). This resampling reduced wavelengths from 2151 to 108, smoothing spectra and mitigating over-fitting.

3. The effectiveness of spectral channel degradation was demonstrated in predicting various soil properties.

4. Multiple data treatments were applied to enhance spectral features.

5. Reflectance data were transformed into absorption (log 1/R, where R is reflectance) to address scattering effects, a common technique in chemometrics.

6. First- and second-order derivatives of absorption spectra were calculated using the Savitzky-Golay method.

7. Reflectance data were standardized, ensuring all spectra had a mean reflectance of 0 and a standard deviation of 1. This standardization facilitated data comparison across intensity scales and enhanced absorption features and curve shapes.

They used multiple linear regression analysis (MLR) and an ANN as an approach to achieve the prediction. They had similar results with the different ML methods. They could not achieve significant results for Cd, Cu, and Zn but they could predict well As, Fe, Hg, Pb, S, and Sb.

Guo et al. (2022) used remote sensing images of the GaoFen-5 (GF-5) satellite and applied normal pre-processing, radiometric calibration, atmospheric correction, and also applied mixed pixel decomposition. The spectral data processing included SG smoothing and continuous wavelet transform (CWT) of GF-5 image data and spectral reflectance extraction. They used RF, combined with geological statistical models, including ordinary kriging (Ordinary Kriging) and inverse distance weighting (IDW), in predicting Zn and Ni concentrations at open pit coal mine located in the Ordos Plateau, Inner Mongolian, China, to obtain the optimal inversion model for Zn and Ni and clarify the distribution features of those elements in the study area. They were able to infer the heavy metal concentration indirectly, based on the relationship between heavy metals and soil criteria and clay.

The soil preparation, spectral window, preprocessing methods, ML algorithms chosen and elements analysed for each author are summarized in Table 2.

Table 2: Summary of previous research and soil sample preparation for detecting different metals.

Author	Soil preparation	Spectral window	Preprocessing	ML methods	elements
Zhou et al. (2021)	Dried, 0,25 mm sieve	400 nm-2450 nm.	First derivative, second derivative, inverse-log, continuum removal, multiple scattering correction	PLS, SVM, RF,	Mn, Cu, Zn, Pb, Cr, Ni
Rodríguez-Pérez et al (2021)	Air-dried and spread in black soil cores	350 nm. – 2500 nm.	Outlier Identification, Spectral Averaging, Wavelength Grouping, Scatter Correction	PLSR	Mn, Ca, K, P, N
Kemper & Sommer (2002)	Each sample was measured four times and averaged afterward. Samples were air-dried and sieved through a 2 mm sieve prior to measurement.	350 nm. – 2500 nm.	Spectral Degradation, Reduction of Wavelengths, Data Enhancement, Reflectance to Absorption, First- and second-order derivatives of absorption spectra calculated using Savitzky-Golay method, Data Standardization	MLR, ANN	As, Cd, Cu, Fe, Hg, Pb, S, Sb, and Zn
Guo et al. (2022)	Air dried, grounded, 0.7 mm sieved	350–2500 nm	SG, CWT	RF	Zn and Ni
Pyo et al. (2020)	Pretreatment by microwave acid digestion. For each sample, reflectance spectra of a total of 20 points were measured with equidistant intervals.	749 nm - 2400 nm.	Noise bands removal, spectral smoothing, second-order differential transformation, PCA.	CNN	As, Cu and Pb.
Cheng et al. (2019)	Each soil sample was uniformly tiled in a petri dish and was successively scanned 10 times	350 nm. – 2500 nm.	SOM, LVs, Savitzky-Golay method, MSC and SNV, Absorption, first derivative	PLSR	Cd, Pb, As, Cr, Cu and Zn
Nanni & Demattê. (2006)	Pure spectral data of the samples Using a laboratory sensor	400 nm. – 2500 nm.	Convolution, ranges of wavelengths	Multiple regression equations	SiO_2, Fe_2O_3, and TiO_2

Visible, near, and shortwave infrared (VIS-NIR-SWIR); least squares (PLS) method, support vector machine (SVM) method and random forest (RF) model were used to build an inversion model based on the characteristic band. Optimal Latent Variable Numbers (LVs) Vary.

6 Remote sensing principles

In a general definition, remote sensing is the collection of information about certain objects without physical contact with them, employing electromagnetic energy- light, heat, and radio waves- to detect and measure target characteristics (Sabins, 1997). The first use of remote sense was in photography, an image recorded on light-sensitive film by a sensor (camera), using the photon energy (visible to near visible spectrum) reflected or radiated from objects (Kulkarni, 1986). The first aerial photograph from the Earth's surface that is recorded dates from 1858, made by the French balloonist and photographer from a hot air balloon. With the photograph's improvements, in 1882 kites started to be used for aerial photography. In 1903, a breast-mounted camera for carrier pigeons was developed, and the pigeons were used for aerial reconnaissance (Fig.8). The first aerial photograph taken from an airplane date in 1909, and during and post the First World War this technology developed fast (Professional Aerial Photographers Association).

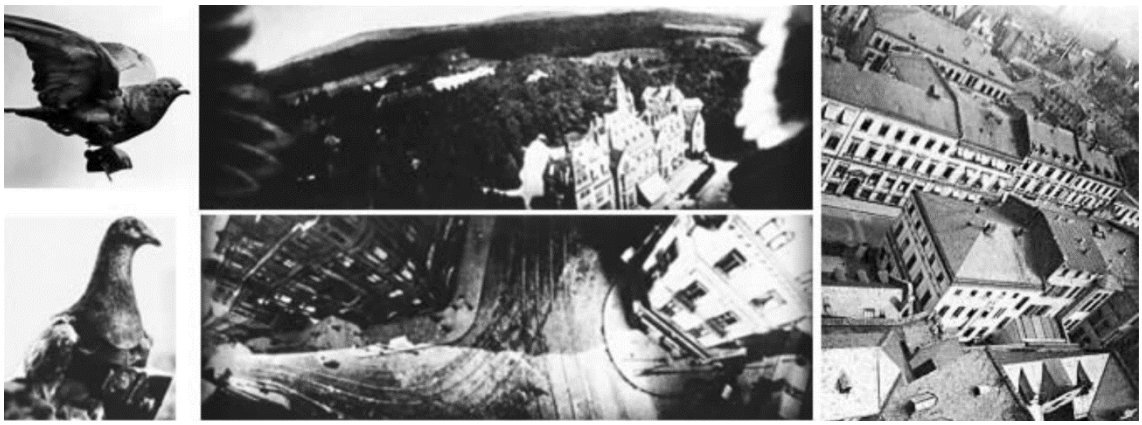

Figure 8: (left) Neubranner's pigeon-mounted camera. (Centre and right) Aerial photographs taken on pigeon photo flights. Note the wingtips showing in the centre top image of the castle. (Professional Aerial Photographers Association)

Today we use different sensors and techniques to acquire remote sensing data, applying passive and/or active systems. A passive remote system is capable of recording the energy naturally reflected from objects, while an active system uses its own source of energy applied to an object to measure the energy that is returned. In both passive and active remote sensing systems, a continuum part of the electromagnetic energy is

measured, the electromagnetic spectrum, that ranges from nanometres to kilometres in wavelength (Sabins, 1997; Lillesand et al., 2015), (Fig. 9).

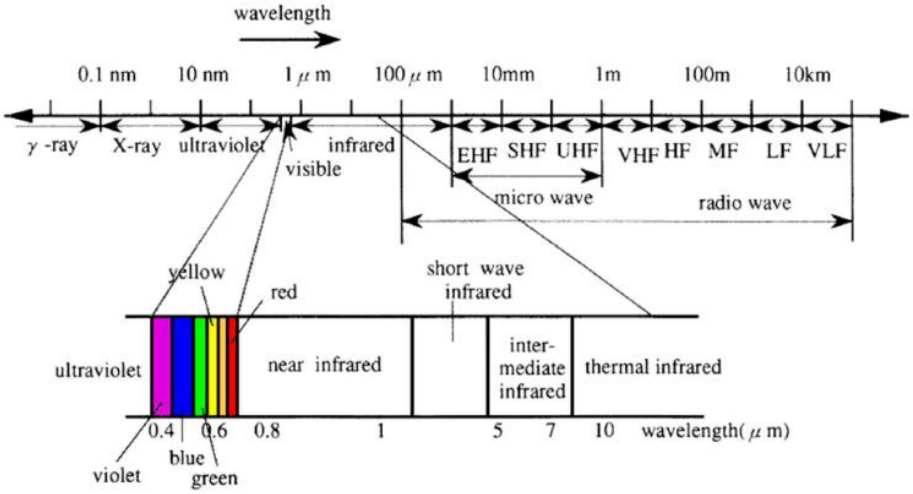

Figure 9: The electromagnetic spectrum. Adapted from: JARS, 1993

The region of the electromagnetic spectrum that can be used by remote sensing is the interval that is not absorbed by the atmosphere (detail in Fig. 9) and is called the atmospheric window. Table 2, from Sabins (1997), shows the distinct bands of the electromagnetic spectrum.

6.2 Visible and near-infrared reflectance spectroscopy and Soil Reflectance

Reflectance spectroscopy offers a means to extract multiple soil properties, both direct and indirect, as well as metal contents. The abundant data generated by soil spectroscopy, whether in the form of point measurements or images, necessitates the implementation of data-modelling procedures.

Materials may reflect or absorb electromagnetic radiation at varying wavelengths, governed by factors such as surface absorption, emissivity, and reflectance characteristics. The spectral range employed for soil reflectance analysis encompasses the Visible and Near-Infrared to Short-Wave Infrared (VNIR-SWIR) region, spanning from 400 to 2,500 nm. This range is further divided into two sub-ranges: VNIR (400 to 1,100 nm) and SWIR (1,100 to 2,500 nm). The interactions between light and matter are intricately tied to the wavelength. Though pure metals don't exhibit absorption within the

VNIR-SWIR region, their presence can be indirectly detected through associations with organic matter (OM), interactions with compounds like hydroxides, sulphides, carbonates, or oxides that manifest detectable properties, or adsorption to light-absorbing clays (Schwartz, & Eshel & Ben-Dor, 2011).

7 Machine Learning

One of the first practical works that used ML was developed by Arthur Samuel in the early 1950s, the author defined the area of ML as a field of study that gives computers the ability to learn without having been programmed to do so (Widerhold and McCarthy, 1992).

Today ML can be conceptualized as a data analysis method that automates the creation of analytical models. It is a branch of artificial intelligence that allows computers to learn from data, identify patterns and make decisions with minimal human intervention (Ayodele, 2010).

In geology, ML can be used for various tasks such as the prediction of mineral deposits in the form of prospectivity mapping or modelling (Farahbakhsh et al., 2023; Senanayake et al, 2023; Abedini et al. 2023; Shi et al, 2023), deposit classification (Nogueira & Maia, 2023) seismic data interpretation concerning reservoir (Okamoto et al., 2023; Jackson et al., 2023; Kumar & Sain, 2023) and earthquakes (Wenhuan et al., 2023). Satellite images, geochemical, geophysical, and structural data are the most common inputs using for the predictions. A supervised learning algorithm, for example, can be trained using labelled data of mineral deposits to predict the likelihood of finding a mineral deposit in a given area. ML models can automatically improve performance with experience. As the algorithm is used, it can be updated with new information, resulting in improved accuracy. Furthermore, those algorithms can handle and process large amounts of data and this can be particularly useful in cases where manual analysis of data is impractical or impossible. This can be a valuable tool for mineral exploration companies, as it can save time and resources by identifying potential areas of interest.

One of the challenges of using ML in geology is the availability of labelled data. In many cases, the outcome of a geological process is not known, making it difficult to train a supervised learning algorithm. In these cases, unsupervised learning algorithms can be used to discover hidden patterns or relationships in the data. Karpatne et al (2018) listed several challenges that geoscientists can face when trying to apply machine learning. The challenges listed were divided into three groups:

- Inherent challenges of geoscience processes: The amorphous boundaries of objects, as the patterns of geoscience objects that can exist in multiple scales and be continuous in space-time, manifesting as waves, flows, and coherent structures in different phases of matter; The spatiotemporal structure where geoscience phenomena occur, causing to variables a structural dependence, unless a discontinuity exists; The high dimensionality of variables, given the complexity of the Earth system, with multiple layers, implying in thousands or millions of dimensions; The interest in rare phenomena, frequently extreme events, such as vulcanism and extreme weather events as cyclones and flash floods are interest of study due the losses it can cause. This kind of event, considered rare, is often difficult to model and characterize because of the limited number of data samples from those rare class events.
- Geoscience data collection challenges: The multi-resolution data, given the heterogeneity of data obtained from different sources, such as satellite sensors, *in-situ* measurements and model-based simulation, and the various scales at which processes occur, it is important to develop algorithms that can identify patterns at this multiple resolutions without upsampling the data sets; The noise, incompleteness an uncertainty in data, that comes with the limited nature of types of equipment and sensors used for measuring the variables. Also, some variables can only be inferred from other observations or simulations.
- Paucity of samples and ground truth, in detail: small sample size, that is limited for the nature of the phenomenon being measured or for the limitation of possibilities in the data collection. Advances in ML methods are needed to deal with this large number of physical variables associated with a limited number of samples; The paucity of ground truth is a problem related to the difficulty of obtaining high-quality measurements of several geoscience variables, that can only be taken by overly expensive or time-consuming techniques.

Despite its challenges, the use of ML in geology is growing rapidly. Its ability to process substantial amounts of data and identify patterns that may be missed by human analysis makes it a valuable tool for mineral exploration, resource discovery, environmental monitoring, and natural hazard prediction.

7.1 Machine Learning for Regression

ML algorithms can be divided in two categories according if they are trained with human supervision or not.

1. Supervised learning is a type of ML where the algorithm is trained on labelled data, where the outcome or correct answer is already known. This task of learning a function that maps an input to an output based on the example input-output pairs (Mahesh, 2020). The algorithm learns to make predictions or decisions by finding patterns in the data that are associated with the known outcomes. For example, a model to classify minerals based on Spectroscopy Images (Flesche et all, 2000). The algorithm would be trained on a dataset of rock samples that have been previously analysed and labelled with their mineral compositions.
2. Unsupervised learning, otherwise, is a type of ML where the algorithm is not given labelled data. Instead, it is used to discover hidden patterns or relationships in the data. For example, an algorithm to cluster tight carbonates (Glover et all, 2022). The algorithm analyses the carbonates and groups them together based on similarities, without being given any prior knowledge of the mineral compositions.

In this work, supervised ML algorithms are employed, as the data contain labels, the Sb content, that are expected to be the outcome. It is usual in supervised learning to have *classification* problems and *regression* problems. As Géron (2017) explains, in a ML classification problem, there will be presented classes as labels, and the predictions are expected to be included in the presented classes, and in a ML regression problem, continuous numbers are expected to be predicted. In this work, the aim is to predict an Sb content that is given as a number, not a class, for this reason, it is called a regression problem.

8 Methodology

8.1 Soil Sampling Campaigns - Ribeiro da Serra e Tapada

The present dissertation benefited from the soil samples collected in the scope of the AUREOLE project- ERA-MIN/0005/2018 (http://doi.org/10.54499/ERA-MIN/0005/2018). The soil sampling campaigns were carried out covering the area where the underground works took place and the surrounding areas in order to test the mineralization distribution, identify possible new mineralized structures, soil contamination distribution and test portable X-ray fluorescence analyser (pXRF) as a useful exploration tool. The first soil sampling campaign was carried out on the Ribeiro da Serra antimony-gold mine zone in 2021. This campaign was done with the following objectives:

For this soil sampling campaign, a total of 174 sample points were planned on a grid of 50x50 meters angled at 45º. From these 174 planned points, 17 had to be cancelled due to the inaccessibility of the terrain, which leaves a total of 157 samples collected (Figure 10). Some of the points also had to be slightly relocated to avoid unpaved roads, tracks, infrastructures, and streamlines.

The Tapada Mine zone is adjacent to Ribeiro da Serra, and the soil sampling campaign conducted in the zone followed the same grid format and orientation as the Ribeiro da Serra (50x50 meters and 45º). This second soil sampling took place during July and August 2022 and had the objective of complementing the studies in the western limb of the Valongo anticline. For the Tapada region, from the total of 180 proposed points, 152 were collected (Figure 17). Some points could not be collected due to the difficult access, the high slope nature of the terrain or the extremely dense and thorny vegetation (known locally as “silvas”).

In both areas, for each collected sample it was recorded the coordinates and the associated parameters: colour, depth, soil horizon, and terrain. Colour was described using acronyms, according to Table 3:

Table 3- Acronyms used for description in the parameter “colour”.

LBRW	Light Brown
BRW	Brown
DBRW	Dark Brown
GRY	Grey
ORA	Orange

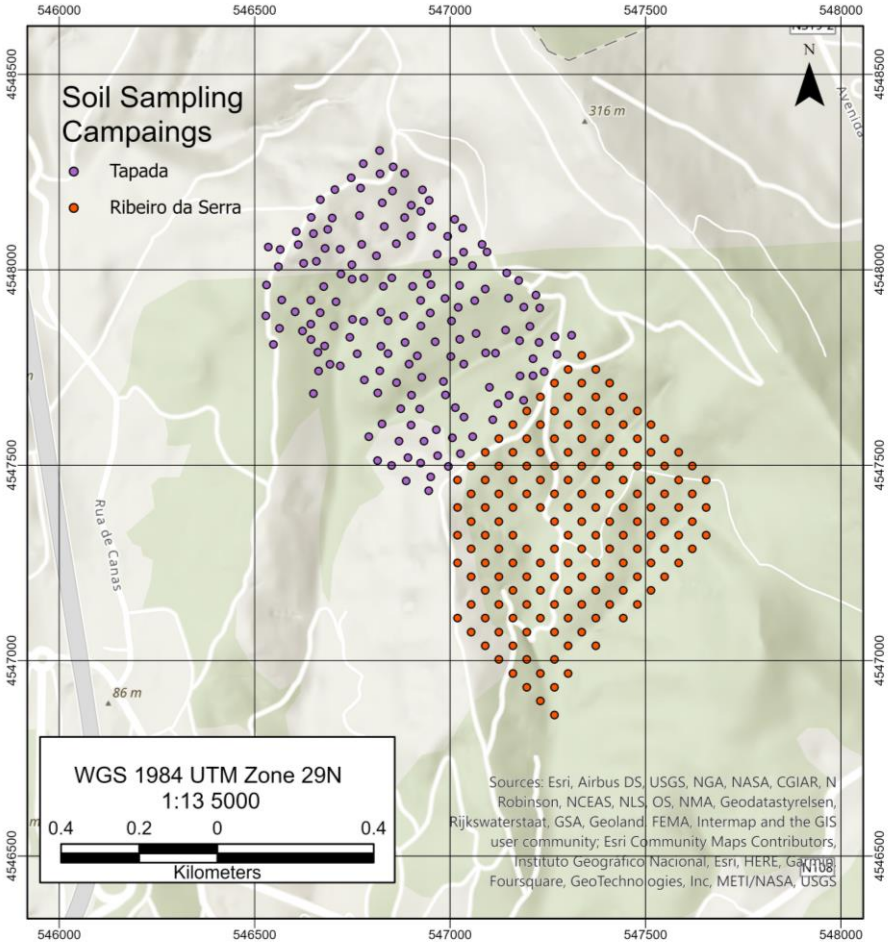

Figure 10: Location of collected soil samples for Ribeiro da Serra and Tapada.

The samples were collected from the B and C horizons, ranging from 10 to 30 cm in depth. Rainy days and wet soils were avoided, and the same plastic bags were used for every sample to maintain the same conditions of analysis.

8.2 Geochemical analysis

8.2.1 XRF analysis

For the geochemical analysis, each sample was analyzed with the Bruker S1 Titan X-ray fluorescence analyzer (Figure 11). The samples were analysed directly from the collection bags (polyethylene bags). Each sample was analyzed on three different points with a 90-second beam. Values below the detection limit of the equipment were considered zero.

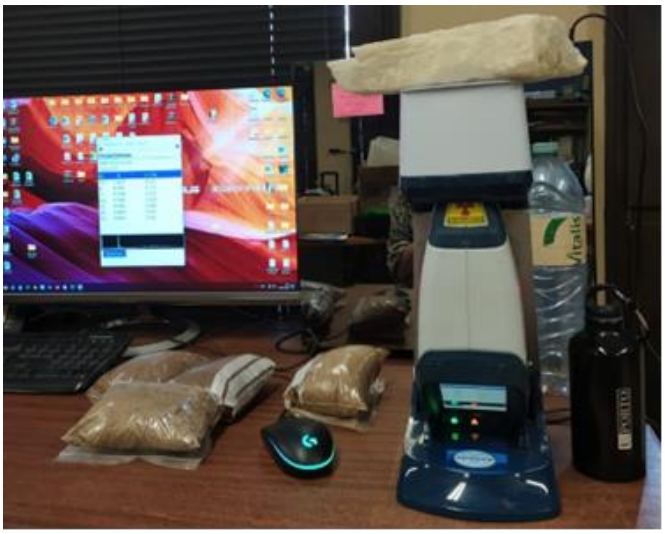

Figure 11: Bruker S1 Titan X-ray Fluorescence Analyzer used for the geochemical analysis of the samples.

The multi-element average values of the three analyses were calculated. Descriptive statistics are presented for the selected elements, V, FE, Ni, Sb, As, Ti, Pb, Mn, and Zn, for the two zones and separately for the Ribeiro da Serra and Tapada zones.

8.2.2 ICP-MS analysis

From the totality of samples, 54 samples from Ribeiro da Serra and 53 samples from Tapada were selected to be analysed in an accredited laboratory (Bureau Veritas) by Inductively coupled plasma mass spectrometry analysis (ICP-MS).

The preparation for the analyses involved drying the selected samples in a dryer, at a temperature of 55 °C for at least 48 hours. They were quartered and 1\8 of the sample was ground in a mechanical agate mill for 10 minutes. Later, the samples were manually ground in mortar with a pestle, until all the contents passed through a 200 µm sieve. The sifted content was stored in plastic bags with the sample identification and sent for laboratory analysis to Bureau Veritas laboratory in Vancouver (Canada) with ICP-MS. multi-acid digestion. $^{+2}$

8.2.3 Test the pXRF as a useful exploration tool for Sb exploration

The evaluation of the pXRF results was made considering the uncertainty associated with the pXRF analyser, the combined standard uncertainty, u_c, was calculated using equation 1, where SD is the standard deviation of the repeated measuments and u_{FP} is the maximum uncertainty reported by the manufacture's FP algorithm (Guimarães et al., 2016).

Equation 1: combined standard uncertainty

$$u_c = \sqrt{SD^2 + (u_{FP})^2}$$

To obtain evaluation metrics such as the RMSE and R^2 for comparing the XRF results with the ICP-MS results, the calculation was implemented in Python (Annex A) using the librarys pandas, numpy and matplotlib.pyplot in the Jupyter Notebook, an interative web-based environment used for programming, data analysis and visualization in Python language. The comparison between the Factor Loads obtained by R-mode Factor Analysis (Klovan, 1975) and the PCA scores is also employed to evaluate the results obtained (Lemière et al., 2020).

The PCA Scores were obtained using the sklearn-learn library, and the Factor Scores were obtained using the factor_analyzer Python module, both executed using Jupyter noterbook. Before obtaining the Factor Scores the adequacy of the data was evaluated by using Bartlett's test of sphericity and by the Kaiser-Meyer-Olkin Test. The number of factors was chosen considering the eigenvalues with values greater than one.

8.3 Soil reflectance measuring

Different soil preparations were used by authors when looking for heavy metals in soils (Table 2). In this study the raw soil was used, it was previously dried, and gravel and pieces of plants were removed.

The equipment ASD FieldSpec 4 Hi-Res NG spectroradiometer (Figure 12) was used to collect the spectral data. The proceeding included "heating" the equipment for 30 minutes before starting to measure. The spectroradiometer has three sensors, one VNIR and two SWIR sensors. The time of "heating" is necessary to ensure that the three sensors are at the same temperature. Normalization with a perfect albidum (white plate) also is needed every time a new measurement project is started and every two hours of work. The software used was Indo Pro (Cardoso-Fernandes et al., 2021).

Figure 12: Spectroradiometer used to collect the spectral data.

The soils were spread in a watch glass (Figure 13) above a black surface and five different points of the sample were measured, resulting in five spectra per sample. Each spectrum collected is a result of an average of several measurements. For this work an average of 5 measurements was used, each measurement resulted in 40 scans.

Figure 13: Soil spread in the watch glass and represented disposition of the measurements taken.

8.4 Spectral preprocessing

The first approach for using the spectral data was to try to obtain results with the minimum preprocessing needed. The logic in this is that for practical use, the less preprocessing, the easier it will be to use the implemented solution with new data. As the models could not handle the data to obtain good results, spectral preprocessing techniques were employed to remove noise and irrelevant information. Besides finding the outliers through PCA analysis, the steps followed in the preprocessing until the generation of the file to be used for training the models are summarized in Figure 14.

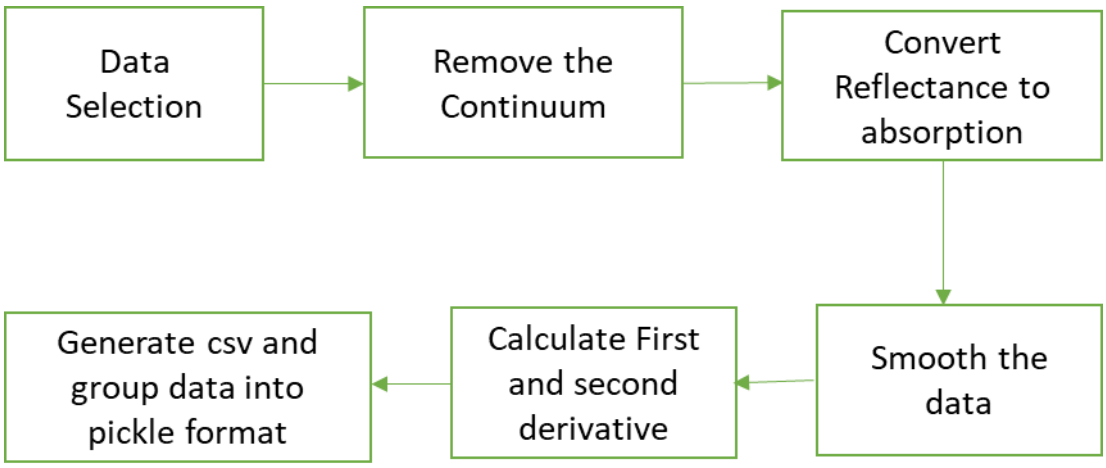

Figure 14: General preprocessing steps from the data collection until the compilation of data for using in the machine learning model.

8.4.1 Continuum removal using convex hull

Continuum removal is a technique that allows the extraction of characteristic absorption bands on the reflectance spectrum curves (Zhou et al., 2021). The convex hull forms a polygon connecting the outermost points within the sample while ensuring that all the internal angles of this polygon are less than 180 degrees, forming the smallest convex shape that encloses all the points in a given set (Baíllo & Chacón, 2021). The continuum removed data was obtained using a Python script (Cardoso-Fernandes et al., 2021), courtesy of Joana Cardoso-Fernandes available in the suplementary material pubished.

8.4.2 Reflectance to absorption, Smoothing the data and calculate the First and Second derivatives

The absorption was calculated from the data, after the continuum removal, using a function implemented in Python, as mentioned in Kemper & Sommer (2002), based on reflectance using the logarithmic relationship.

After the conversion to absorption, the data was smoothed by applying the Savitzky-Golay filter from the savgol_filter function in the scipy.signal module in Python. To smooth the signal data, the Savitzky-Golay filter calculates a polynomial fit of each window based on polynomial degree and window size. The obtained smoothed data is used to calculate the First and Second derivatives by applying the np.gradient function from the NumPy library. Those preprocessing steps were executed in a Python script in the Jupyter Notebook (Annex B) and, as a result, the data is saved as CSV and in pickle format.

8.4. 3 Downsampling and upsampling

When there is an unbalance between different classes in the data set, upsampling or downsampling one of the classes can be a solution. The approach chosen was downsampling by using the ICP-MS data, which was more balanced. By doing this, also it was possible to make tests including elements that don’t show very accurate measured values by pXRF, such as Pb, As and Fe.

8.4.4 Convert waveform to spectrogram

This step was not always applied in the RF or the PLSR models. It consisted of the application of the Short-Time Fourier Transform (STFT) to the equal_length tensor using tf.signal.stft. A new dimension is added to the spectrogram tensor using spectrogram [..., tf.newaxis]. This dimension is added to make the spectrogram suitable as input data with convolution layers in the CNN. The script used for this transformation was executed in the Jupyter Notebook and the details on the libraries and modules used is in Annex B.

8. 5 Machine Learning Algorithms

After the preprocessing steps, for each test, it was necessary to follow a few more steps to get the data matched with the labels (when changing the samples used for the test) before training the models and also for testing the effect of the conversion to spectrogram. The general steps followed until starting to train the models and validate are in Figure 15:

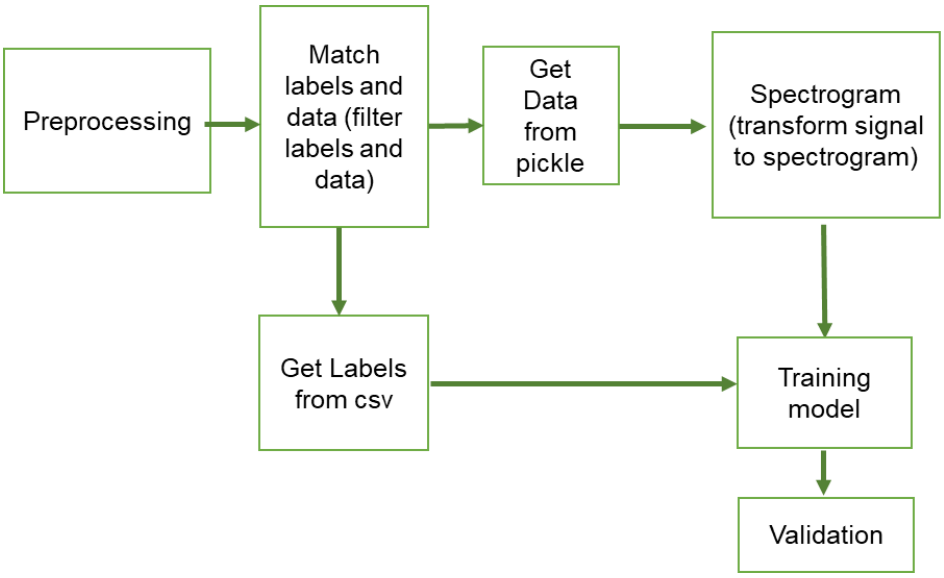

Figure 15: Steps followed to provide the labels and spectral data set to the machine learning model.

8.5.1 Partial Least Squares Regression regression

PLSR finds extensive utility in the realms of VNIR laboratory spectroscopy and hyperspectral remote sensing. It is a versatile statistical technique addressed to:

- Co-linearity Management: In spectroscopic datasets, especially in the VNIR, reflectance values at different wavelengths often exhibit strong co-linearity. PLSR effectively handles this issue by extracting underlying patterns and relationships among these correlated variables.
- High-Dimensional Data: PLSR is particularly advantageous when dealing with datasets where the number of predictors (wavelengths) exceeds the number of samples (comprising the database [X, Y]). This situation frequently arises in laboratory VNIR spectroscopy, where spectrometers can capture reflectance across a vast spectrum, often spanning more than 2,000 wavelengths (Gomez & Lagacherie, 2016).

PLSR operates by projecting the descriptive variables (wavelengths) in X and the dependent variables in Y onto a lower-dimensional space defined by a set of orthogonal vectors known as latent variables. The key objective is to maximize the covariance between the two datasets, X and Y, while simultaneously reducing the dimensionality of the data. PLSR can find essential relationships between the spectral information (X) and the target properties (Y) (Gomez & Lagacherie, 2016).

8.5.2 Random Forest

RF is an ensemble method based on decision trees, which can be applied to both classification and regression tasks. Decision trees are models that use a tree-like structure to make decisions by recursively splitting data based on certain rules. In classification, the Gini index is commonly used to determine the optimal splits, while in regression, as in this study, the goal is to minimize the sum of squared deviations from the mean (de Santana et al., 2018).

RF consists of an ensemble of decision trees organized as a set of structured classifiers or trees denoted as {T1(Θ), T2(Θ), … TB(Θ)}, where B represents the number of classifiers, and Θ is the set of bootstrap samples drawn with replacement from the original training dataset. Each bootstrap sample Θ = {Θ1, Θ2, … ΘB} has dimensions of q × p, where q is approximately two-thirds of the total number of samples (m), and p (or mtry) is the number of randomly selected variables used at each node to construct the ensemble of trees. The default values for mtry are specific to the problem type, such as n/3 for regression and a fixed value for classification. The key tuning parameters are:

- mtry (p): The number of input variables randomly selected for each bootstrap, varying from 1 to n (the total number of variables).
- Number of Trees: It's important to choose a sufficiently large number of trees for a stable model. According to de Santana et al. (2018), typically, around 500 trees are considered sufficient, but using more trees won't significantly change results statistically.
- Minimum Node Size: Sets the minimum size of nodes where no further splits will be attempted. The default values are 5 for regression and 1 for classification.

RF offers several advantages, including the ability to handle datasets with missing values, robustness to noise and outliers due to bootstrapping and variable randomization, and the ability to assess variable importance and identify outliers using proximity matrices (de Santana et al., 2018). The application of RF and PLSR was made in the Jupyter Notebook, and the details on the libraries and modules used can be found in Annex C.

8.5.3 Convolutional neural networks

In recent years, there has been a growing trend in utilizing deep learning algorithms for extracting valuable features from soil spectral data to predict soil constituent content.

CNNs are well-established in various domains, including object detection, image classification, and spectral analysis (Yang et al., 2020).

CNNs consist of three primary types of layers: convolutional layers, pooling layers, and fully connected layers:

1. Convolutional Layer: In this layer, input features are convolved with learnable kernels, generating various output feature maps. Each kernel has a fixed length and slides over the input feature map with a stride of 1, performing convolution with local regions where the kernel overlaps the input feature map. Non-linear activation functions are then applied to the convolution results to produce output feature maps. This layer's key advantage lies in its ability to learn local patterns while significantly reducing the number of model parameters through weight sharing.
2. Pooling Layer: After convolution, the pooling layer is employed to progressively reduce the spatial size of the output feature maps generated by the convolutional layer.
3. Fully Connected Layer: The fully connected layer comprises neurons that connect to all activations extracted from the convolutional and pooling layers. This layer plays a crucial role in generating the final output results.

By leveraging sparse local connections and weight sharing, CNNs have proven to be effective in learning and extracting local and abstract features from raw spectral data. By stacking multiple convolutional and pooling layers, the CNN model can efficiently capture intricate patterns within the data, making it well-suited for soil content prediction tasks (Yang et al., 2020).

In the present work, in order to deal with limited computational capacity, a MobiletNet model was used (Howard et al., 2017). MobileNets are a class of highly efficient CNN models, built upon a streamlined architecture that leverages depth wise separable convolutions, being a deep neural network with significantly reduced computational demand. The script used in this work was executed in the Jupyter notebook and the details about the libraires and modules is in the Annex D.

9 Results and discussion

9.1 Test portable X-Ray Fluorescence analyser (pXRF) as a useful exploration tool.

The descriptive statistics for the selected elements, Sb, As, Ti, Pb, Mn, Zn and Fe obtained for Ribeiro da Serra and Tapada by pXRF and ICP-MS are in Table 4 and Table 5.

Table 4: Descriptive Statistics for selected elements in Ribeiro da Serra.

	Ribeiro da Serra													
	XRF analisys (ppm)							ICP-MS analisys (ppm)						
	Sb	As	Ti	Pb	Mn	Zn	Fe	Sb	As	Ti	Pb	Mn	Zn	Fe
Mean	2160.46	183.31	0.3	88.23	72.8	26.28	36060.04	660.38	152.97	0.28	55.61	58.15	30.07	35161.11
Std. Deviation	5971.45	437.64	0.13	225.16	107.54	27.26	10888.2	1273.48	279.16	0.15	86	68.89	14.92	9739.15
Skewness	3.65	3.58	0.44	3.79	3.03	2.56	0.98	2.03	2.98	4	3.14	3.96	2.15	0.32
Std. Error of Skewness	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32
Kurtosis	13.79	12.53	0.5	14.07	12.2	7.28	1.9	2.57	9.3	20.93	10.28	19.03	6.88	0.73
Std. Error of Kurtosis	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64
Minimum	0	11.67	0.05	3.67	0	6.33	15390.67	9.5	14.8	0.13	12.26	9	11.1	11500
Maximum	31247.33	2152	0.68	1113.67	616.33	148	73976	4001	1431.2	1.14	449.03	442	95.3	62800

Table 5: Descriptive Statistics for selected elements in the Tapada Zone.

	Tapada													
	XRF analisys (ppm)							ICP-MS analisys (ppm)						
	Sb	As	Ti	Pb	Mn	Zn	Fe	Sb	As	Ti	Pb	Mn	Zn	Fe
Mean	519.17	47.94	2777.34	26.95	86	24.52	34702.14	394.62	103.23	3339.43	55.37	112.94	44.85	39466.04
Std. Deviation	2822.02	124.68	1188.79	97.76	188.15	15.89	11137.25	917.02	194.32	1815.28	146.13	164.01	26.85	10470.91
Skewness	10.26	9.12	1.11	11.06	4.21	2.8	0.2	3.44	4.23	3.47	6.25	3.19	1.91	0.96
Std. Error of Skewness	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.33	0.33	0.33	0.33	0.33	0.33	0.33
Kurtosis	113.76	94.36	5.74	129.57	20.46	10.97	-0.02	11.24	20.77	13.96	41.61	11.21	4.38	2.98
Std. Error of Kurtosis	0.39	0.39	0.39	0.39	0.39	0.39	0.39	0.64	0.64	0.64	0.64	0.64	0.64	0.64
Minimum	0	6	138	0	0	3.33	3818.33	6.3	16.4	1450	9.26	13	17.8	20000
Maximum	32716.67	1403.67	9515.67	1181.33	1328	115.67	68743	4001	1208.1	12000	1040.14	844	146	79300

Concerning the Sb values obtained by XRF, the mean values for Sb are considerably higher in Ribeiro da Serra than in the Tapada. Although the maximum values are similar for both zones, the higher values in Tapada have a strong association with outliers, which is indicated by the Kurtosis and Skewness. The same happens to the higher values for As and Pb in the Tapada zone. For the ICP-MS analysis, there was a limit of detection for Sb of 4000 ppm. The samples above this limit were discarded from the data set used for evaluating the pXRF analyser. The parameters coefficient of determination (R^2), RMSE and RMSE for different concentrations of the elements were obtained by comparing the XRF results with the results obtained by ICP-MS analysis and considering the uncertainty (Table 6).

Compared with the samples that were sent to the ICP-MS analysis, Sb shows an R^2 of 0.9 for the Ribeiro da Serra samples and 0.88 for the Tapada samples.

Table 6: Evaluation of the results obtained by pXRF compared to ICP-MS results.

Ribeiro da Serra						
	R^2	RMSE	RMSE (0- 50 ppm)	RMSE (50- 100 ppm)	RMSE (100- 1000 ppm)	RMSE (>1000ppm)
Sb	0.89	164.65	0.00	0.00	68.22	502.85
As	0.38	225.61	10.20	19.99	330.35	1150.12
Ti	0.73	0.11	-	-	-	0.11
Pb	-0.03	95.83	10.99	10.07	112.31	407.13
Zn	0.24	19.40	13.59	37.49	44.63	-
Mn	0.52	47.98	5.63	13.90	93.30	-
Fe	0.98	0.40	-	-	-	0.40
Tapada						
	R^2	RMSE	RMSE (0- 50 ppm)	RMSE (50- 100 ppm)	RMSE (100- 1000 ppm)	RMSE (>1000 ppm)
Sb	0.87	54.82	0.11	64.40	38.92	124.61
As	0.89	65.67	4.99	8.50	53.26	-
Ti	0.13	0.13	-	-	-	0.13
Pb	0.99	9.41	9.15	7.28	-	-
Zn	0.75	0.87	14.80	34.72	-	-
Mn	0.95	37.84	5.83	-	73.76	-
Fe	0.98	0.48	-	-	-	0.48

It's important to notice that some elements, such as As, Pb, Zn, and Mn show significantly different R^2 for the measurements from Ribeira da Serra and Tapada. In the case of As and Pb, it happens because Ribeira da Serra has higher concentrations for those elements, and the pXRF accuracy lowers as the concentration of those elements increases. It is expressed in the higher values for RMSE observed for concentrations of As and Pb that are higher than 1000 ppm. It's also worth noticing that when the values are very low, they may have a good R^2 as a consequence of the relatively high value of uncertainty. In a scenario where a statistical analysis will be conducted, it is worth evaluating how much uncertainty is acceptable. Figure 16, Figure 17, and Figure 18 show, logarithmic scale, the pXRF and ICP-MS values for Sb, As and Pb, respectively.

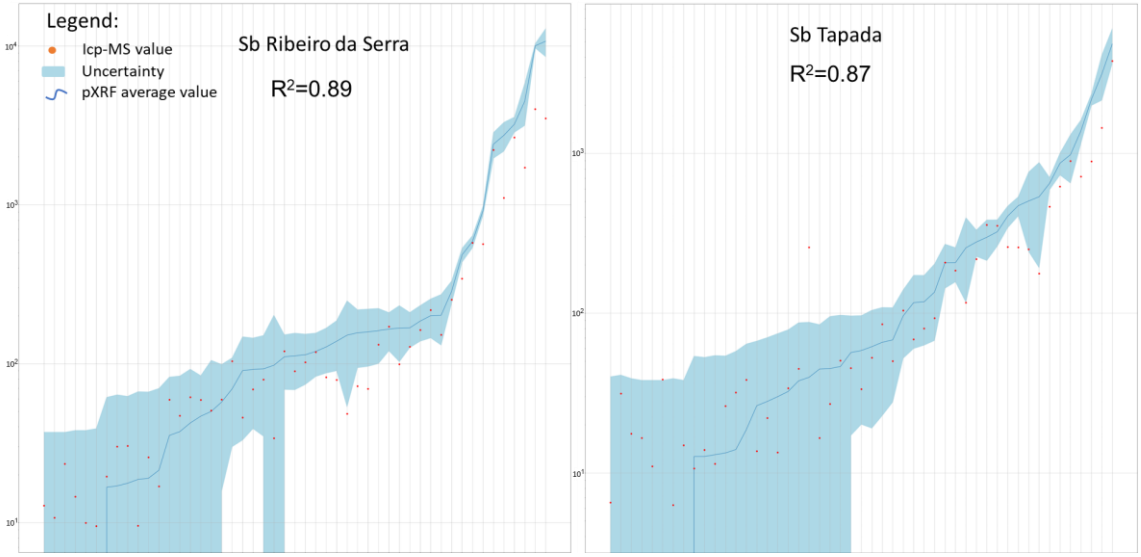

Figure 16: Comparison between antimony (Sb) values acquired through pXRF and ICP-MS, along with the associated uncertainty from pXRF. The x-axis represents samples arranged in ascending order based on their concentration of Sb.

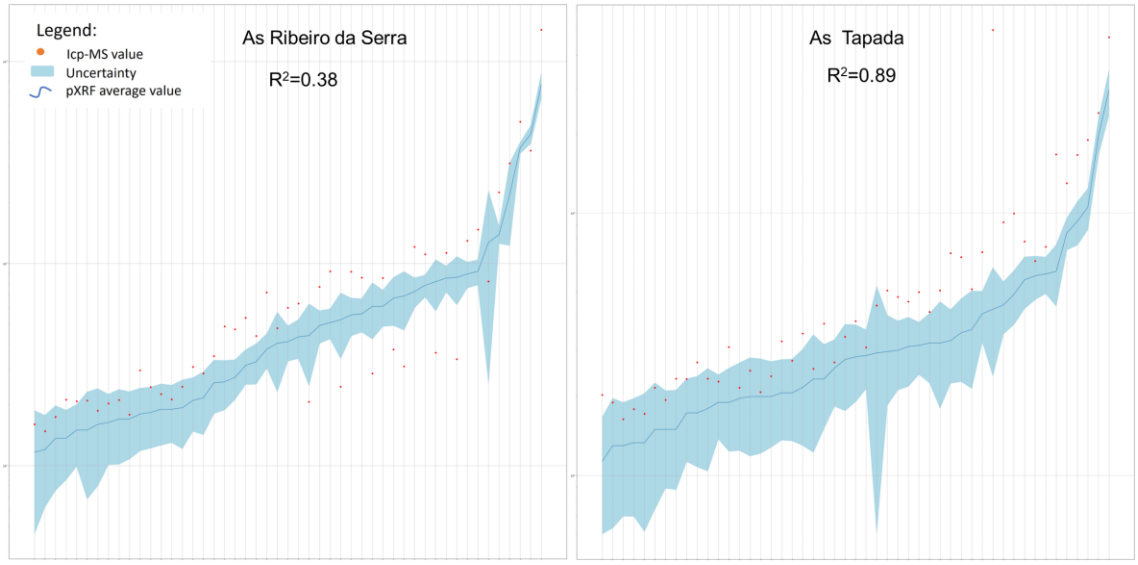

Figure 17: Comparison between antimony (Sb) values acquired through pXRF and ICP-MS, along with the associated uncertainty from pXRF. The x-axis represents samples arranged in ascending order based on their concentration of As.

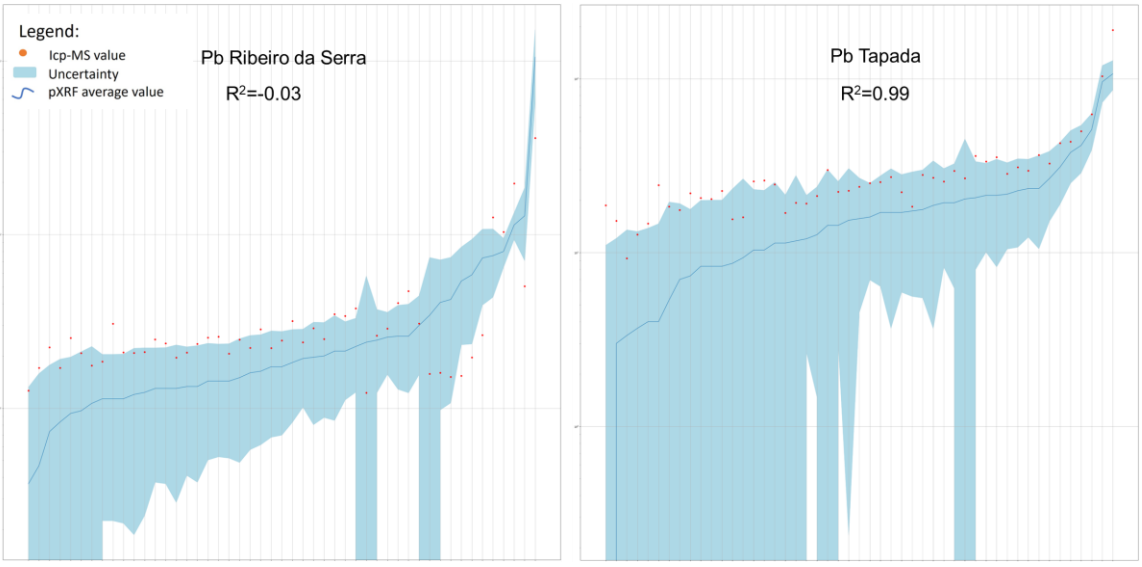

Figure 18: Comparison between antimony (Sb) values acquired through pXRF and ICP-MS, along with the associated uncertainty from pXRF. The x-axis represents samples arranged in ascending order based on their concentration of Pb.

The Factor analysis and PCA help investigate the association of elements. They allow us to perceive if the association between the elements remains with the variation in the measurements. The results after normalization of the data for the PCA analysis are in Table 7 and the results for the Factor analysis are in Table 8.

For the factor analysis (Table 7), we see that for Ribeiro da Serra, Pb (the element with the worse R^2), is a load in a different factor when the two analyses are compared. For Tapada, the elements exhibit slightly different loads but remain in the same factors.

Table 7: Factor analysis.

Ribeiro da Serra						Tapada					
XRF			ICP-MS			XRF			ICP-MS		
	F1	F2		F1	F2		F1	F2		F1	F2
Sb	0.98	-0.04	Sb	0.93	-0.11	Sb	1.00	0.03	Sb	0.84	0.05
As	0.95	0.18	As	0.73	0.16	As	0.97	0.05	As	0.95	0.03
Ti	-0.33	-0.16	Ti	-0.08	0.34	Ti	-0.23	0.56	Ti	-0.22	0.51
Pb	0.55	0.68	Pb	0.80	0.18	Pb	0.98	0.06	Pb	0.93	0.02
Zn	0.26	0.94	Zn	0.43	0.67	Zn	0.39	0.67	Zn	0.27	0.90
Mn	0.35	0.86	Mn	0.58	0.74	Mn	0.03	0.66	Mn	0.09	0.83
Fe	-0.06	0.30	Fe	0.54	0.34	Fe	0.11	0.72	Fe	0.03	0.66

The PCA analysis (Table 8) resulted in components that show different associations between elements than seen in the factor analysis. We see there are variations in the components for Ribeiro da Serra, and for Tapada the association between elements remains the same and reflects what was found in the factor analysis.

Table 8: PCA (principal component analysis).

Ribeiro da Serra						Tapada					
XRF			ICP-MS			XRF			ICP-MS		
	PC1	PC2		PC1	PC2		PC1	PC2		PC1	PC2
Fe	0.21	-0.61	Fe	0.72	0.00	Fe	0.36	0.73	Fe	0.47	0.64
Sb	0.68	0.64	Sb	0.77	-0.38	Sb	0.95	-0.28	Sb	0.78	-0.46
As	0.82	0.48	As	0.78	-0.10	As	0.94	-0.26	As	0.81	-0.49
Ti	-0.45	-0.25	Ti	0.10	0.88	Ti	-0.07	0.76	Ti	0.12	0.71
Pb	0.91	-0.14	Pb	0.83	-0.28	Pb	0.95	-0.25	Pb	0.79	-0.51
Mn	0.85	-0.36	Mn	0.84	0.31	Mn	0.27	0.72	Mn	0.57	0.66
Zn	0.82	-0.46	Zn	0.73	0.36	Zn	0.64	0.53	Zn	0.72	0.56

For the elements highlighted in this study, pXRF demonstrates some limitations in providing accurate values when high contents of As and Pb are present in the soil samples, as found in Ribeiro da Serra.

How much the uncertainty varies in an interval is an important aspect to be considered when using pXRF measurements to evaluate low concentrations of elements, such as concentrations that can indicate contaminations. In a scenario where a statistical

analysis will be conducted, it is worth evaluating how much uncertainty is acceptable and the urgency that the results are needed. In scenarios when the time frame allows, it may be interesting to consider additional steps for the sampling preparation, as confection of pressed pellets, grinding and others (Goff et al., 2020) can extenuate this problem.

The evaluation of the pXRF proves its efficiency in the exploration of Sb, providing fast and satisfactory results with no need for previous preparation of the samples. However, to use correlations of geochemical data between values of several elements obtained by pXRF, it is important to consider that for some elements the results cannot be accurate. Sampling preparations that can improve the results obtained by pXRF, such as preparation of pressed pellets, grinding and others (Goff et al., 2020) can extenuate this problem.

9.2 Geochemical analysis

A R- mode factor analysis (Table 9) was conducted to understand the relations between selected elements, Sb, As, Ti, Pb, Zn, Mn, Ni, Fe, and V. The minimum distance from tailings and veins and the sampling points was included to verify if the distance from these features can be associated with anomalies of Sb, Pb and As. Varimax rotation was applied. Varimax components are unipolar, each defining an independent cluster of related variables. On varimax components, most elements have loadings close to zero values, while related values common for a group of samples have high loadings (Hongyu, K. 2018). Also, the factor scores were calculated (DiStefano et al., 2009). In general, an association between Sb, Pb, As and the tailings is verified.

Table 9: Factor analysis results for Ribeira da Serra e Tapada.

	F1	F2	F3
tailings	0.15	0.40	0.21
veins	-0.13	0.15	0.52
Sb	0.03	0.88	-0.01
As	0.09	0.83	0.03
Ti	0.32	-0.15	0.50
Pb	0.17	0.77	-0.21
Zn	0.94	0.19	0.01
Mn	0.82	0.15	0.19
Fe	0.93	0.20	-0.05
Ni	0.73	0.05	0.26
V	0.28	-0.08	0.44

Factor analysis explains with three factors 96.1% of the data variance. The Factor Scores help to understand the spatial relationship between elements and their distribution. Factor 1 (Figure 16) shows the relationship between the elements Zn, Fe, Mn and Ni, these elements are not associated with the distance of tailings or veins.

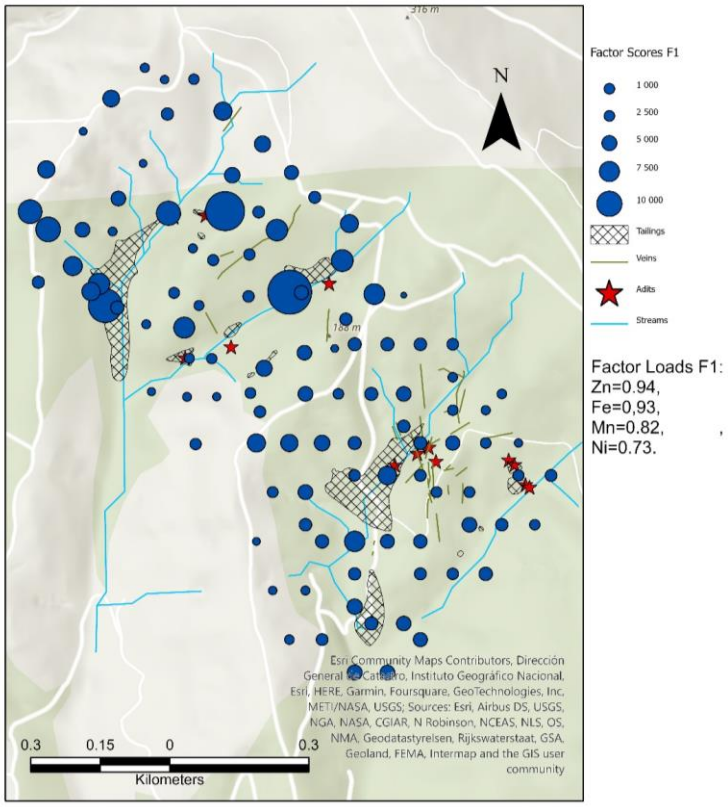

Figure 16: Factor 1 scores and the corresponding factor loads for Ribeiro da Serra and Tapada.

Although there are some high scores coinciding with tailing zones. Their presence can be attributed to the regional geology and the process of soil formation. The distribution of the elements of Factor 1 is in Figure 17. The elements Zn, Mn an Ni show a very similar distribution along the Ribeiro da Serra and Tapada zones, Fe is the exception and has higher values in the Tapada zone, showing a similar distribution with the other elements of Factor 1.

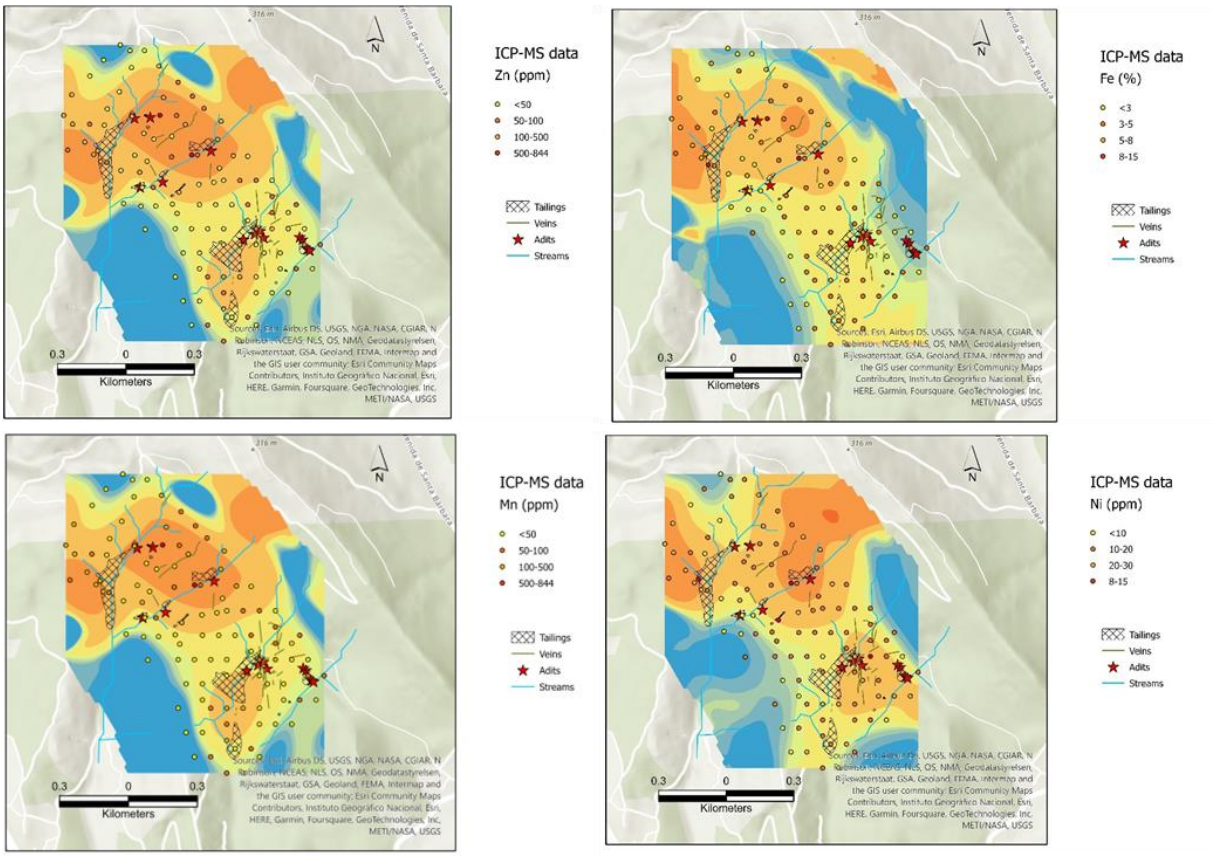

Figure 17: Kernel interpolation for the elements belonging to Factor 1, Zn, Mn, Fe and Ni.

Factor 2 (Figure 18) is compound for Sb, As, Pb and tailings, we can see the tailings influenced the high concentration of those elements. The Sb highest values (Figure 19) are mostly associated with the mining tailings, in both zones Ribeiro da Serra and Tapada. The Sb high values also are found associated with the known veins in the area, equally with the As. There is one sampling point, indicated by the black arrow, in Figure 19, that indicates a zone relatively high topographically, with no influence of tailings or known veins or water lines. Thus, the area associated can be considered as an interest zone. The same zone has a slightly high anomaly in Pb and is anomalously high in As.

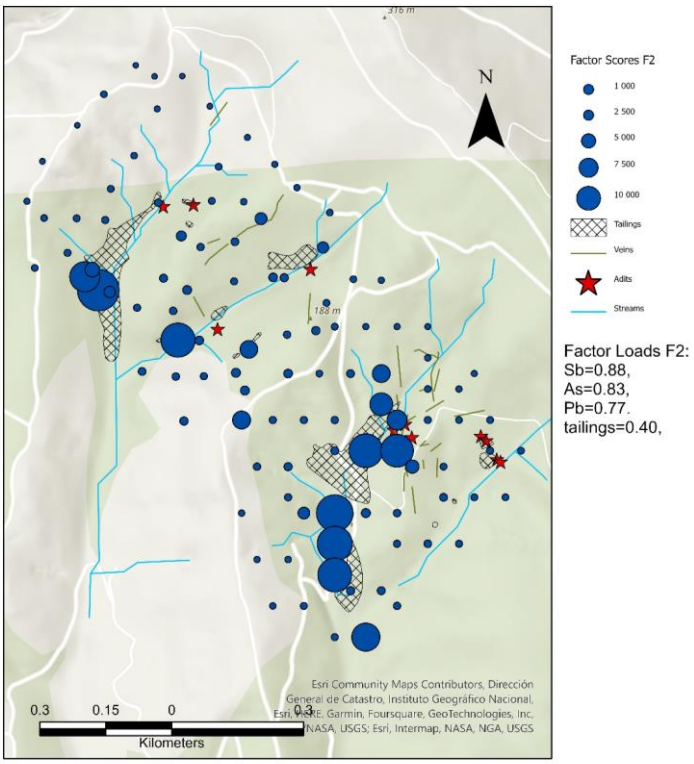

Figure 18: Factor 2 scores and the corresponding factor loads for Ribeiro da Serra and Tapada.

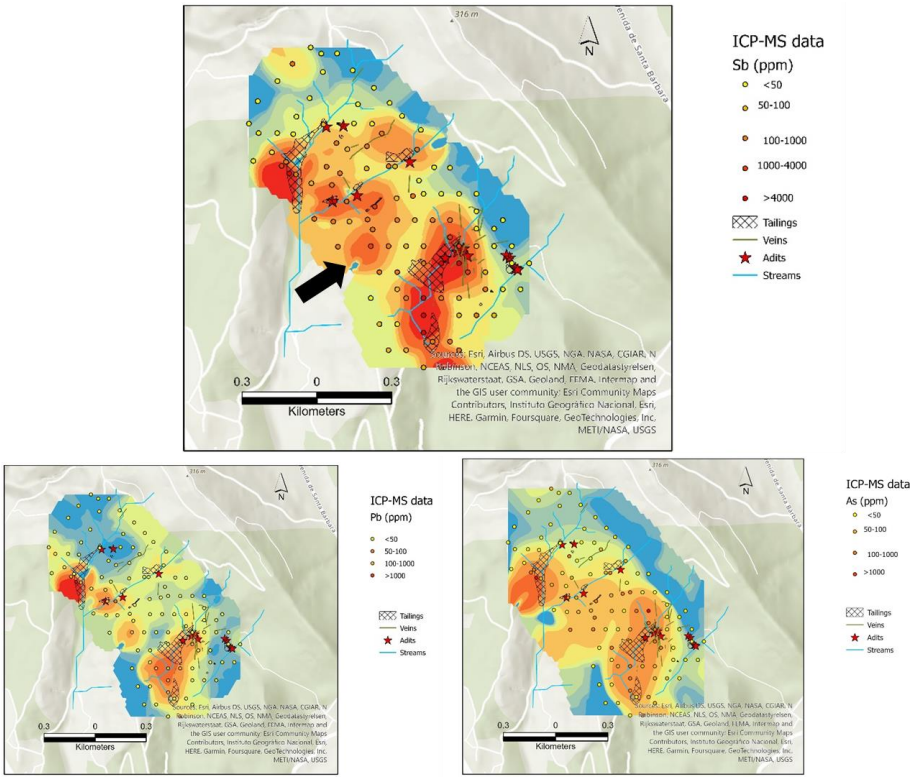

Figure 19: Kernel interpolation for the elements belonging to Factor 2, Sb, Pb and As

Factor 3 is compound for the feature veins, and the elements Ti and V. The higher scores in this factor are seen to follow a trend NW-SE (Figure 20). This pattern suggests that the presence of higher concentrations of Ti and V (Figure 21) can be attributed to the proximity to a dolerite dyke that is placed in the zone and follows the same NW-SE orientation. Also, the veins are found associated with this geological feature in the scale of the soil sampling campaign executed.

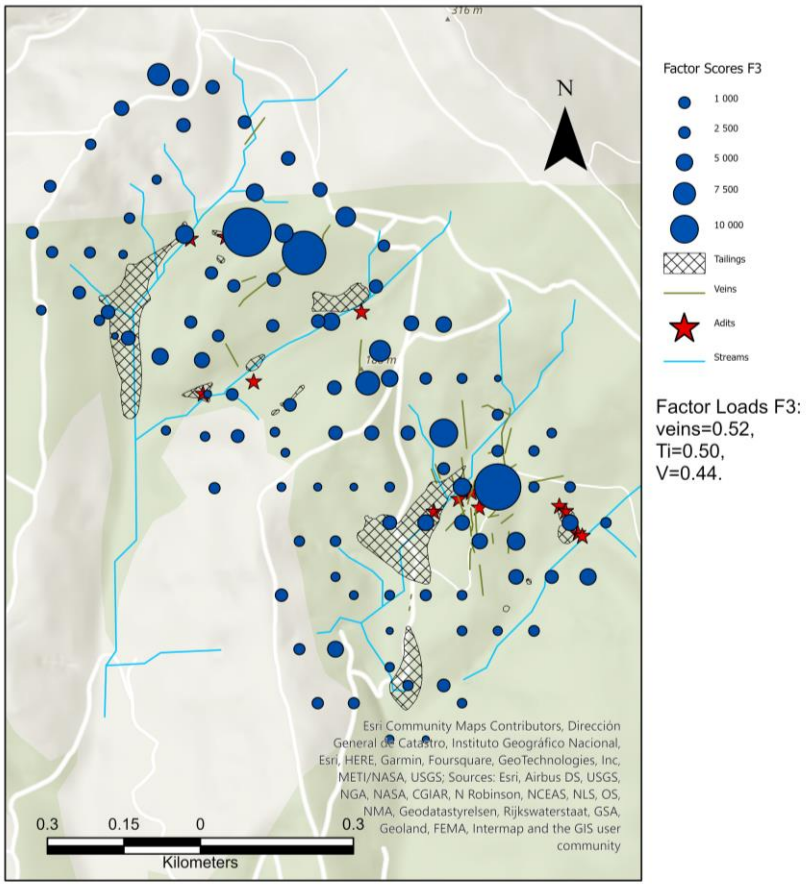

Figure 20: Factor 3 scores and the corresponding factor loads for Ribeiro da Serra and Tapada.

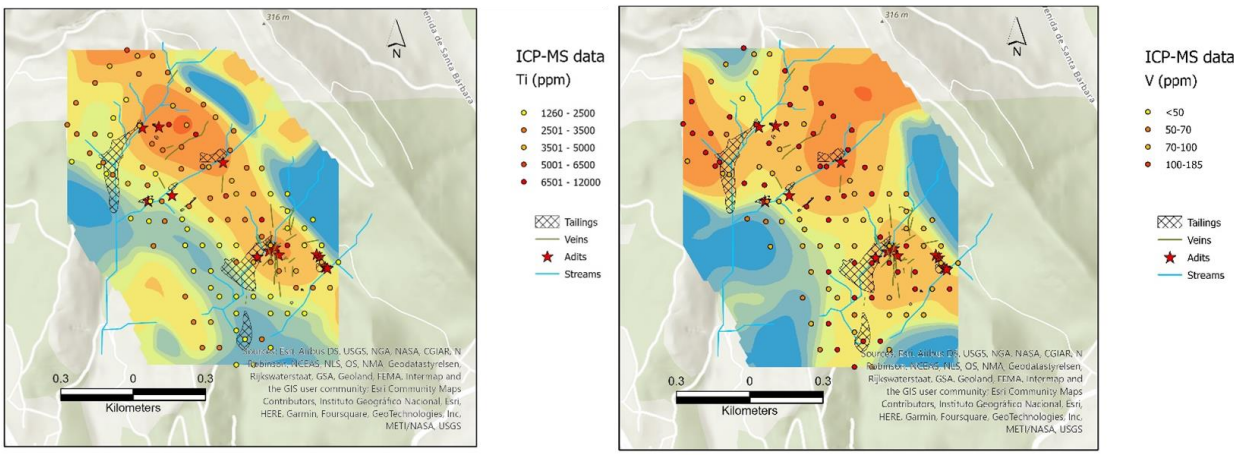

Figure 21: Kernel interpolation for the elements belonging to Factor 3, Ti and V.

9.3 Spectra preprocessing results

9.3.1 Convex hull

The application of the hull quotient procedure removes the continuum and normalizes the spectra, enhancing the absorption features (Cardoso-Fernandes et al., 2021) (Figure 30).

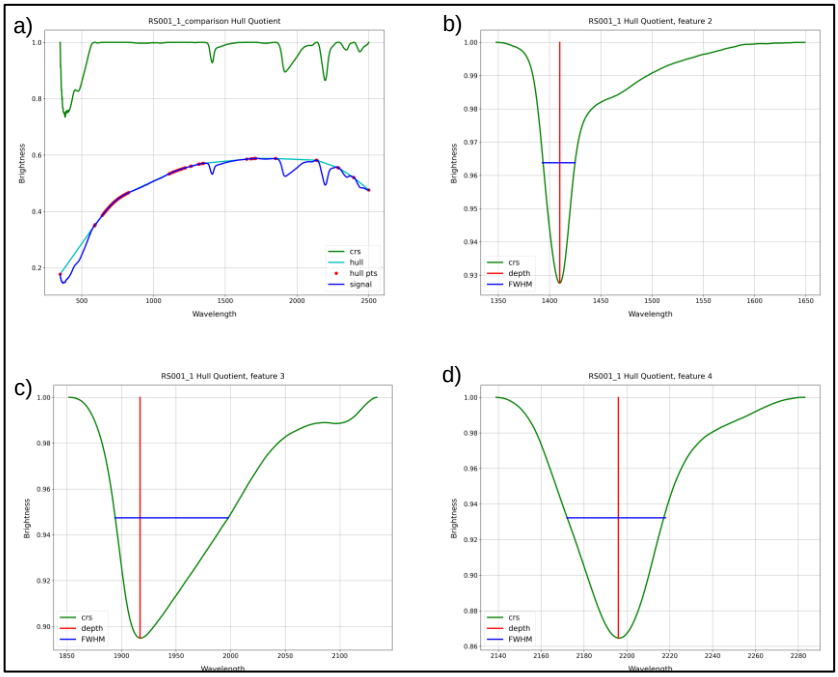

Figure 25: a) Convex hull fitting to the original signal and the respective continuum removed spectrum (crs); (b–d) automatic extraction of the main absorption features using the *pysptools* library (Cardoso-Fernandes et al., 2021). pts–points; FWHM–full width at half maximum.

9.3.2 Savitzky-Golay filter, First and second derivatives

After converting the reflectance to absorption, the signal was smoothed and then the first and second derivatives were calculated (Figure 31). It was observed that the first wavelengths contain excessive noise that cannot be smoothed enough with the function.

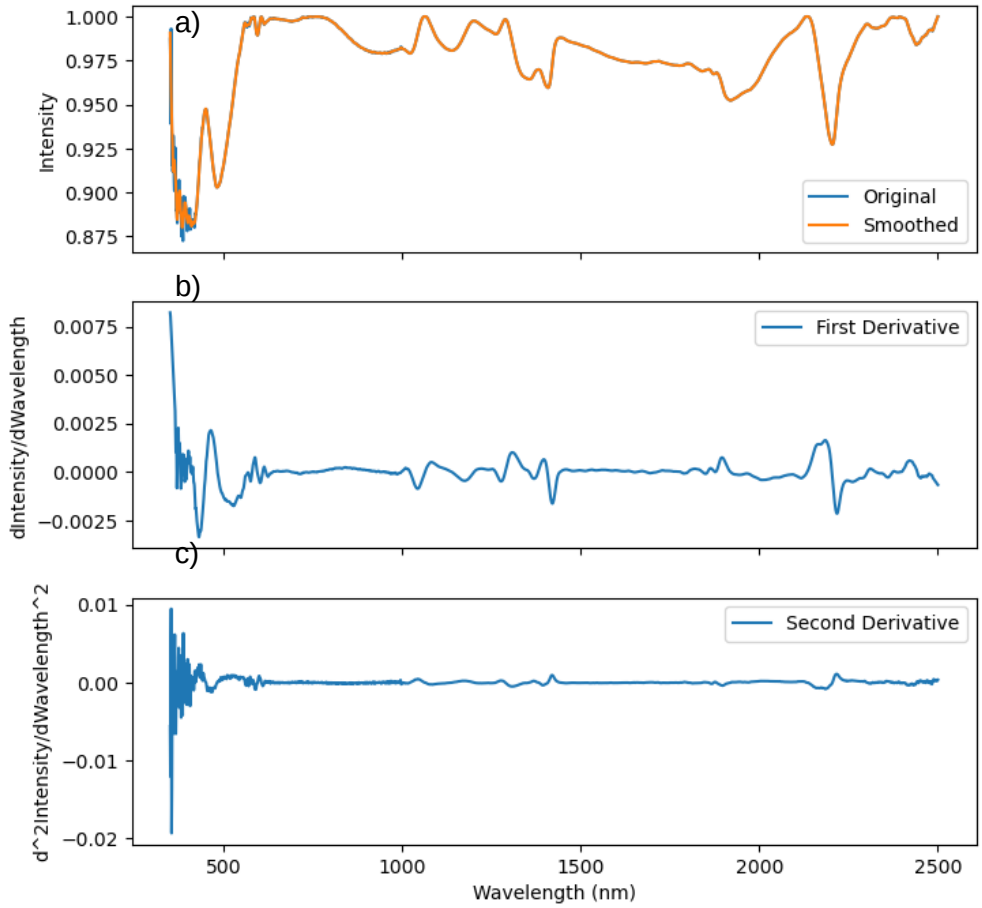

Figure 26: a) Smoothed signal overlapping the original signal for comparison, b) first derivative of the signal, c) second derivative of the signal.

It was decided to remove the portion between 350 and 399 nm of the spectra and it resulted in a slightly better performance of all the models.

9.3.3 Spectrogram transformation

In the PLSR and RF models, the tests were performed using the waveforms and the spectrograms. There is no significant difference in the results for the present data attributed to this choice. For the CNN model, only spectrograms were used. Figure 27 shows the transformation applied to the first derivative of the sample RS001 spectra and Figure 28 shows the transformation for the second derivative of the same sample.

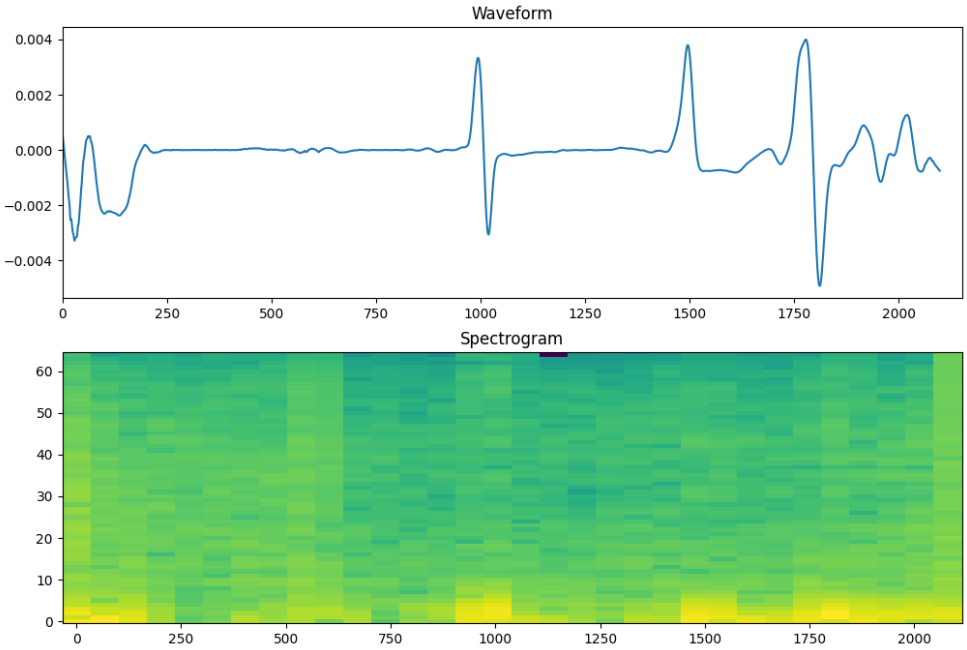

Figure 27: First derivative of a spectrum converted to spectrogram; the signal is from the sample RS001.

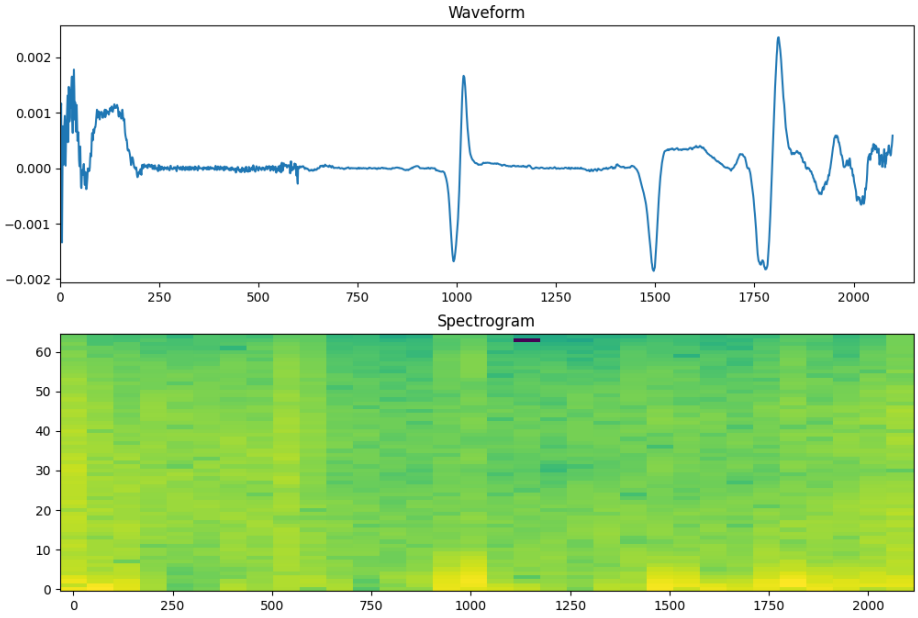

Figure 28: Second derivative of a spectrum converted to spectrogram; the signal is from the sample RS001.

9.3.4 Dowsampling effect and other effects choosing samples

The high values of Sb (>1000 ppm) constituted only 10% of the totality of the samples, intermediate values (>100 and <1000 ppm) and low values (>100 ppm) were relatively balanced. For this reason, it was preferable to use only the samples that were sent to

ICP-MS analysis, as they don't just have more balance between the high, intermediate and low values, but also other elements, that cannot be measured with accuracy by the pXRF, could be used for training the models. Downsampling had a positive effect on the predictions, also removing the outliers, i.e., the very high values of Sb.

Training the models using classes (high, intermediate, and low) didn't improve the results in any way. Likewise, separating the samples based on the colour of the soil or lithology didn't result in any improvement in any model, as the data is unbalanced in those characteristics.

9.4 Spectral test results

The best results obtained after the different preprocessing tests are presented for the discussion.

9.4.1 Convolutional neural networks

Different selections of elements and wavelengths were tested to find a combination that would better suit a solution for the problem. MSE is used as a metric for evaluation. Table 10 details the wavelength interval chosen for every test presented, the elements selected, the preprocessing steps and the training epochs.

Table 10: Wavelength, elements, preprocessing steps, and epochs for the test presented.1 Continuum removal, 2 reflectance to absorption, 3 SG, 4 first derivative, 5 second derivative.

Test	Wavelength	Elements	Preprocessing					Epochs
			1	2	3	4	5	
T3a	350-2400 (nm)	Sb, As	X	X	X	X		200
T3a	350-2400 (nm)	Sb, As	X	X	X	X		1000
T3b	350-2400 (nm)	Sb, Fe	X	X	X	X		500
T3c	350-2400 (nm)	Sb	X	X	X	X		500
T4a	400-2400 (nm)	Sb	X	X	X	X		500
T5a	400-2400 (nm)	Sb	X	X	X		X	500
T4a	400-2400 (nm)	Sb, As	X	X	X	X		500
T5a	400-2400 (nm)	Sb, As	X	X	X		X	500
T6a	400-1400 (nm)	Sb	X	X	X	X		500
T6b	400-1400 (nm)	Sb, As	X	X	X	X		500
T6b	400-1400 (nm)	Sb, As	X	X	X	X		1000
T6c	400-1400 (nm)	Sb, AS, Fe	X	X	X	X		500
T6c	400-1400 (nm)	Sb, AS, Fe	X	X	X	X		1000

Test T3a was conducted using the elements Sb and As, and the training was made with the number of epochs set to 200 and then to 1000 (Figure 29). There is no significant improvement by training the model with 1000 epochs, so it was decided to train the model using 500 epochs. Also, the training error decreases significantly fast in the very first epochs and the model goes into overfitting. Test T3b was performed with the same wavelengths and the first derivative of the spectral data but choosing the elements Sb and Fe and test T3c used the same parameters and only the Sb as labels (Figure 30).

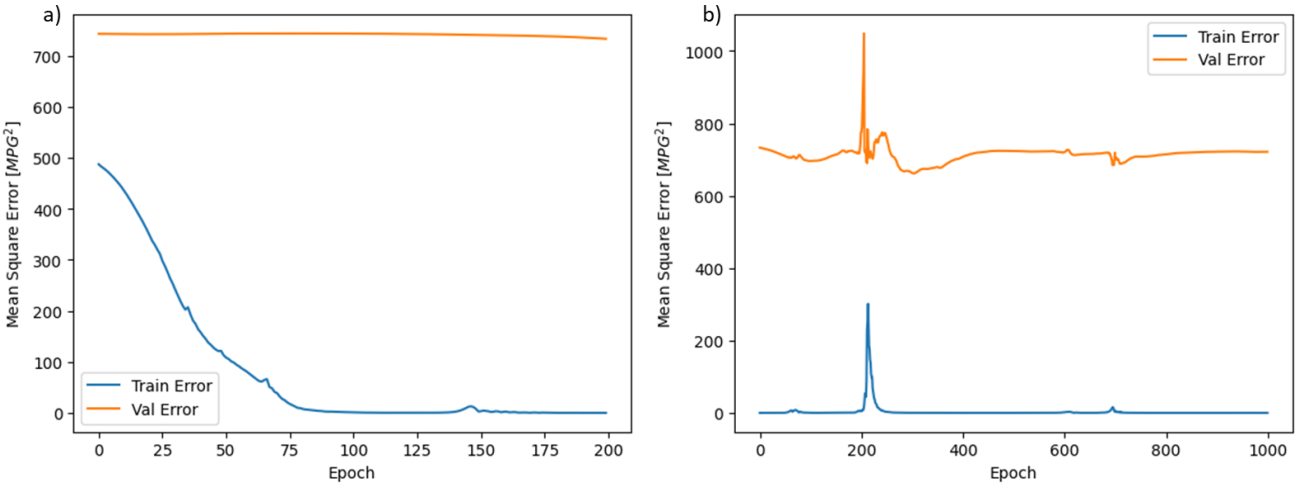

Figure 29: Test evaluating the effect of different epochs on the prediction (train error and validation error). The elements Sb and As were used as labels.

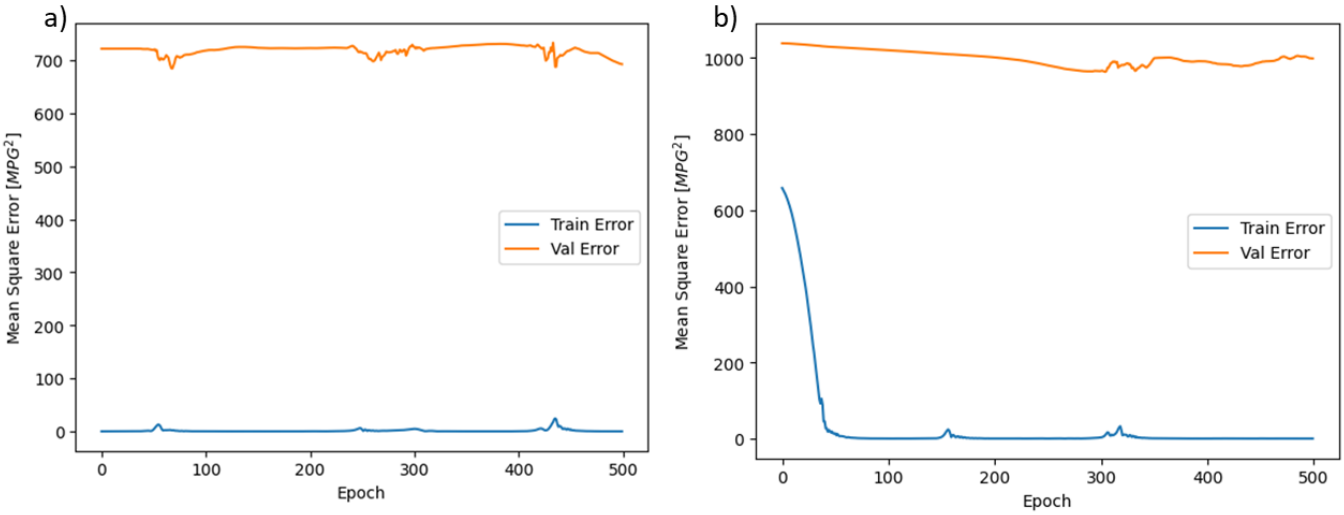

Figure 30: a) Test T3b with T3 parameter performed using the elements Sb and Fe and b) test T3c with T3 parameter performed using only the element Sb.

For the test T4 and T5, the wavelengths from 350 to 399 nm were eliminated and the range between 400 to 2400 nm was used. Test T4 used the first derivative of the data

and test T5 used the second derivative of the data. Test T4a and T5a were performed using only Sb as the label (Figure 31).

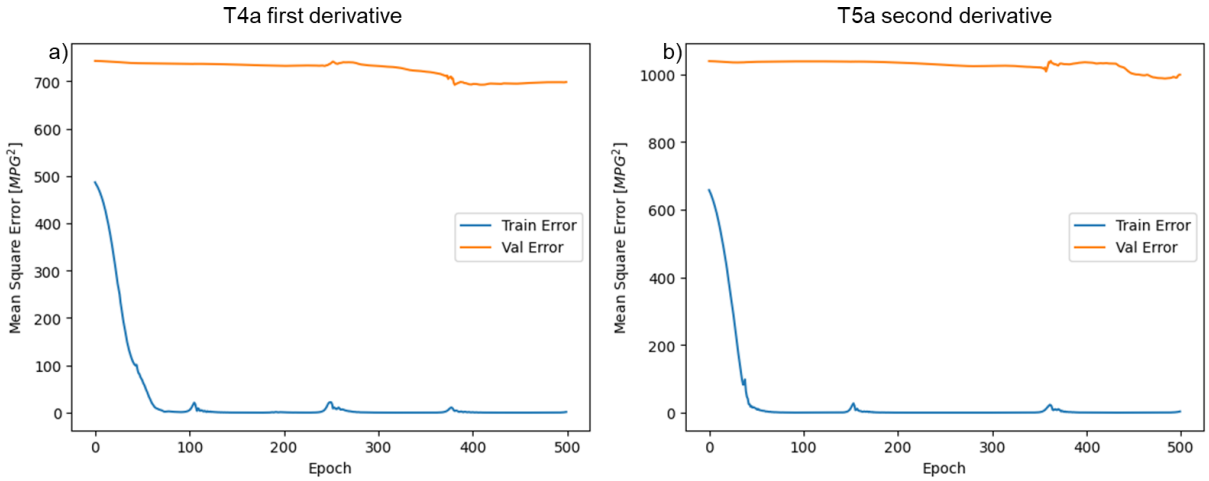

Figure 31: a) Test T4a using the element Sb and the first derivative of the data. b) using only the element Sb and the second derivative of the data respectively.

The predicted result for Sb values from the test T4a, which has a lower MSE, is depicted in Figure 32. Some predicted values are close to the 1:1 regression line (circle), but some predictions that fall far from the measured value make the R^2 low.

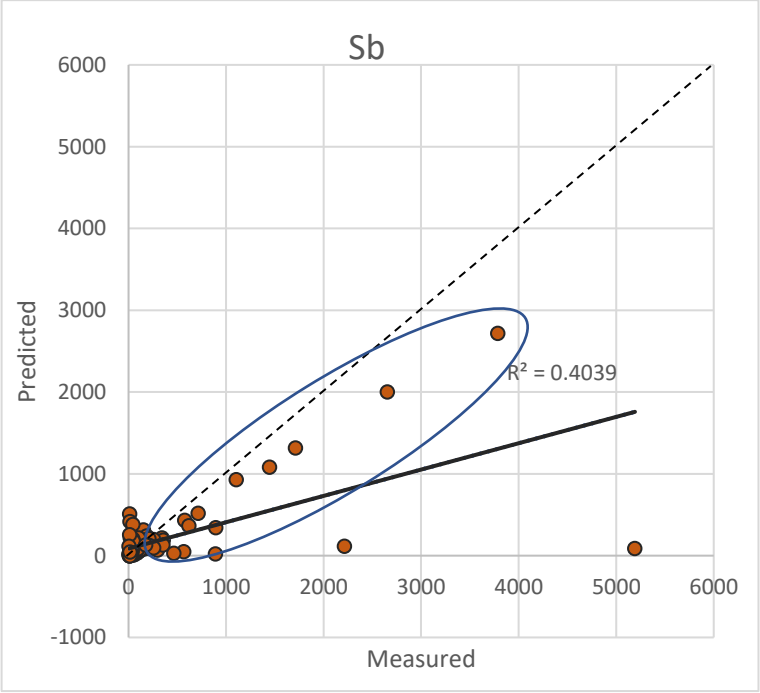

Figure 32: R^2 obtained for the Sb measured and predicted in the test T4a.

The tests T4b and T5b use the first derivative of the data and the second derivative of the data and the elements Sb and As (Figure 33). The predicted result for the T4b is in Figure 34.

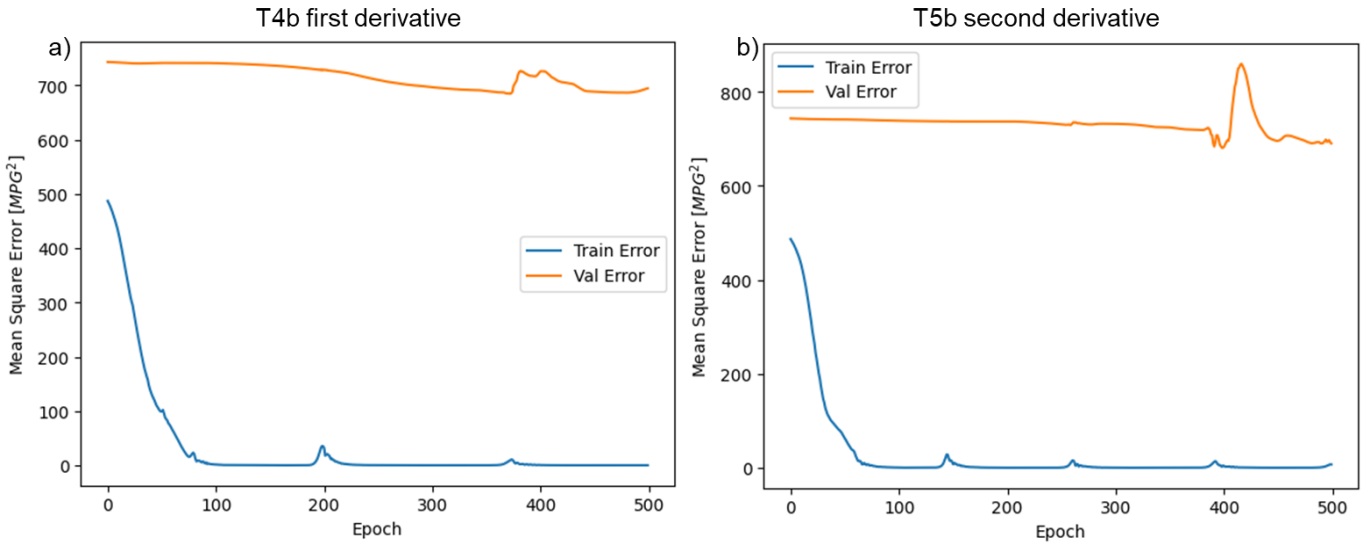

Figure 33: a) Test T4a using the elements Sb and As and the first derivative of the data. b) test T5a, performed using the elements Sb and As and the second derivative of the data.

The test T6 was run using wavelengths between 400 to 1400 nm in an attempt to reduce the noise and overfitting, but the expected effect could not be achieved. Test T6a used the first derivative of the spectral data and just Sb as the label (Figure 40).

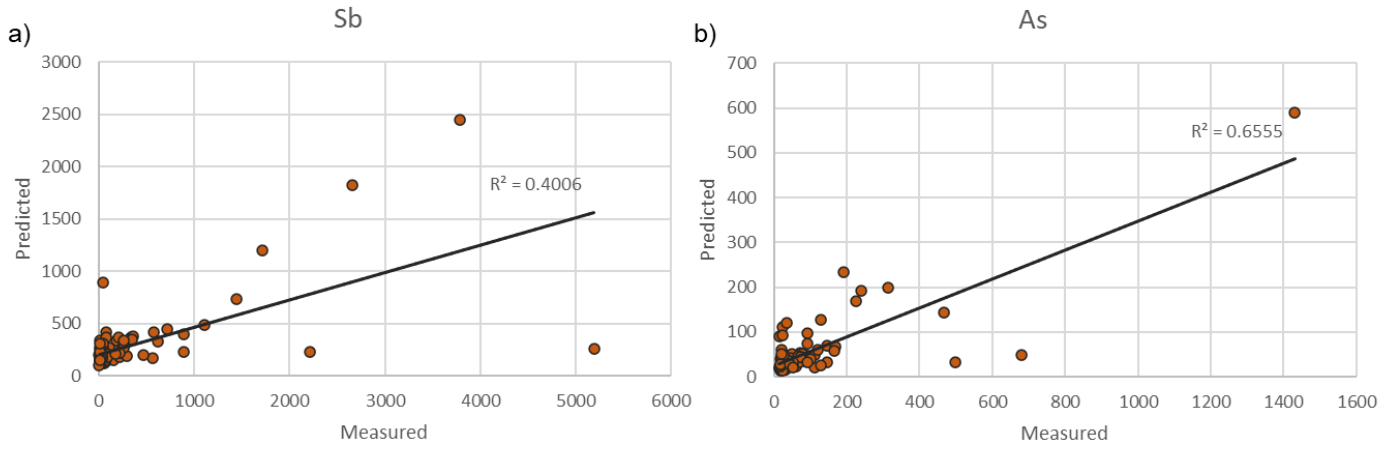

Figure 34: Results from the test T4b. a) R^2 obtained for the Sb measured and predicted. b) R^2 obtained for the As measured and predicted

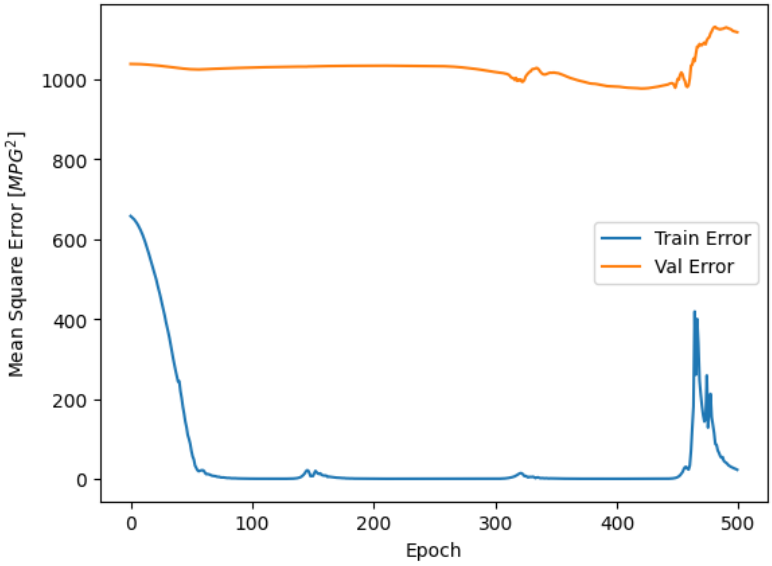

Figure 35: Tests T6a performed using the element Sb and using the first derivative of the data.

Test T6b was trained with Sb and As as labels (Figure 36), while test T6c has Sb, As and Fe as labels (Figure 37), varying between 500 and 1000 epochs. It can be seen that the reduction of wavelengths didn't improve the model and some instability is observed especially with the configuration using the three elements as labels.

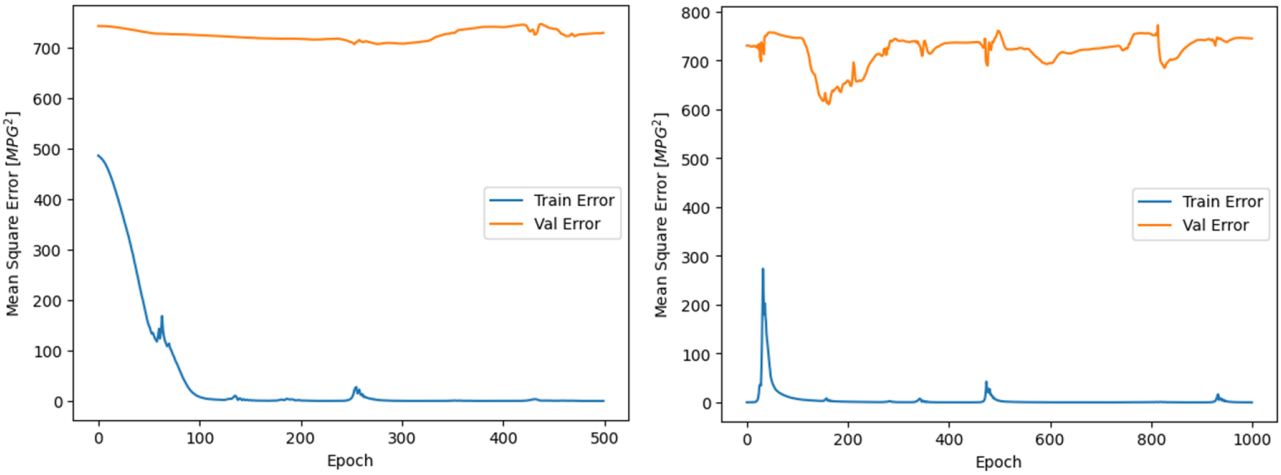

Figure 36: Test T6b performed using the elements Sb and As in 500 and 1000 epochs.

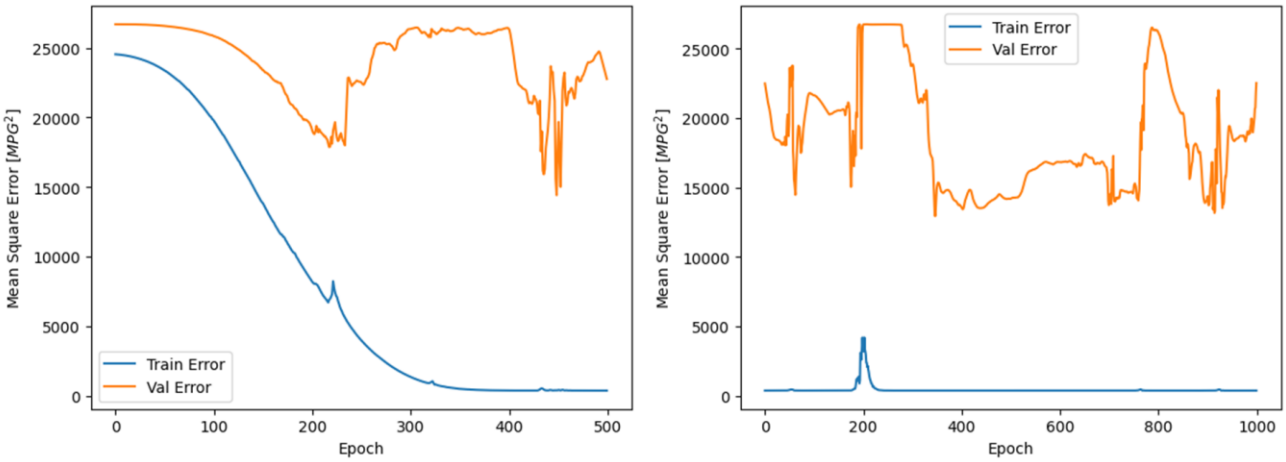

Figure 37: Test T6b performed using the elements Sb, As and Fe in 500 and 1000 epochs.

The performance of the CNN model could be improved by adding the element As in the training data, removing the first wavelengths that contain noise that cannot be smoothed and applying the diverse preprocessing steps to the signal. avoiding the overfitting in the model was not accomplished. Kemper & Sommer (2020) to resolve the overfitting issue, used a methodology to degrade the spectra taking into account the band center and the full-width half-maximum. The band center refers to the central wavelength or position of a spectral feature or band of interest. This approach was not possible to be applied in the current study, because the band center for the target element is unknown. Wu (2007) found the correlation with total Fe, active and residual, was a major predictive mechanism for heavy metals in soils. Also, organic matter and clay have a correlation. The soil analysis in the current study didn't include those proprieties, which can be a way to obtain better results.

9.4.2 Random Forest and Partial Least Squares Regression

Tests with random forest included the utilization of the first derivative of the waveforms directly and the spectrograms generated from the waveforms, while the convolutional neural networks only used the spectrograms. The tests selected to show are those using the first derivative of the data after SG smoothing, continuum removal and conversion of reflectance to absorption, and using the wavelengths between 400 to 2400 nm. Like in the CNN, using the interval from 400 to 1200 didn't improve the performance, but reduced it. However, removing the interval from 350 to 399 did improve the models. Table 11 synthetizes results obtained for RF and PLSR regression. Other test combinations had similar results and were omitted.

Table 11: Wavelength, elements, preprocessing steps, and epochs for the tests presented.

		Random Forest			PLSR regression		
		Waveforms					
Test	Labels	RMSE	MSE	R2	RMSE	MSE	R2
T4a	Sb	959.18	920017.05	-0.15	1020.85	1042127.09	-0.15
T4b	Sb, As	685.73	470230.31	-0.15	732.51	536569.27	-0.09
T4c	Sb, As, Fe	13159.07	173161221.13	-0.09	13358.11	178439106.93	-0.04
		Spectrograms					
T4a2	Sb	944.04	891216.97	-0.04	1020.85	1042127.09	-0.15
T4b2	Sb, As	683.68	467423.61	-0.15	732.51	536569.27	-0.09
T4b3	Sb, As	680.11	462552.55	0.07	-	-	-
T4a3	Sb	952.76	907761.07	0.001	-	-	-

1 Continuum removal, 2 reflectance to absorption, 3 SG, 4 first derivative, 5 second derivative.

Test T4b3 and T4a3 repeat tests T4b2 and Ta2, respectively, using 50 trees and 500 trees. For the model using the two elements (Figure 38), there is some improvement, but it is still far from good predictions. For the model using just the Sb, there is no improvement in the prediction.

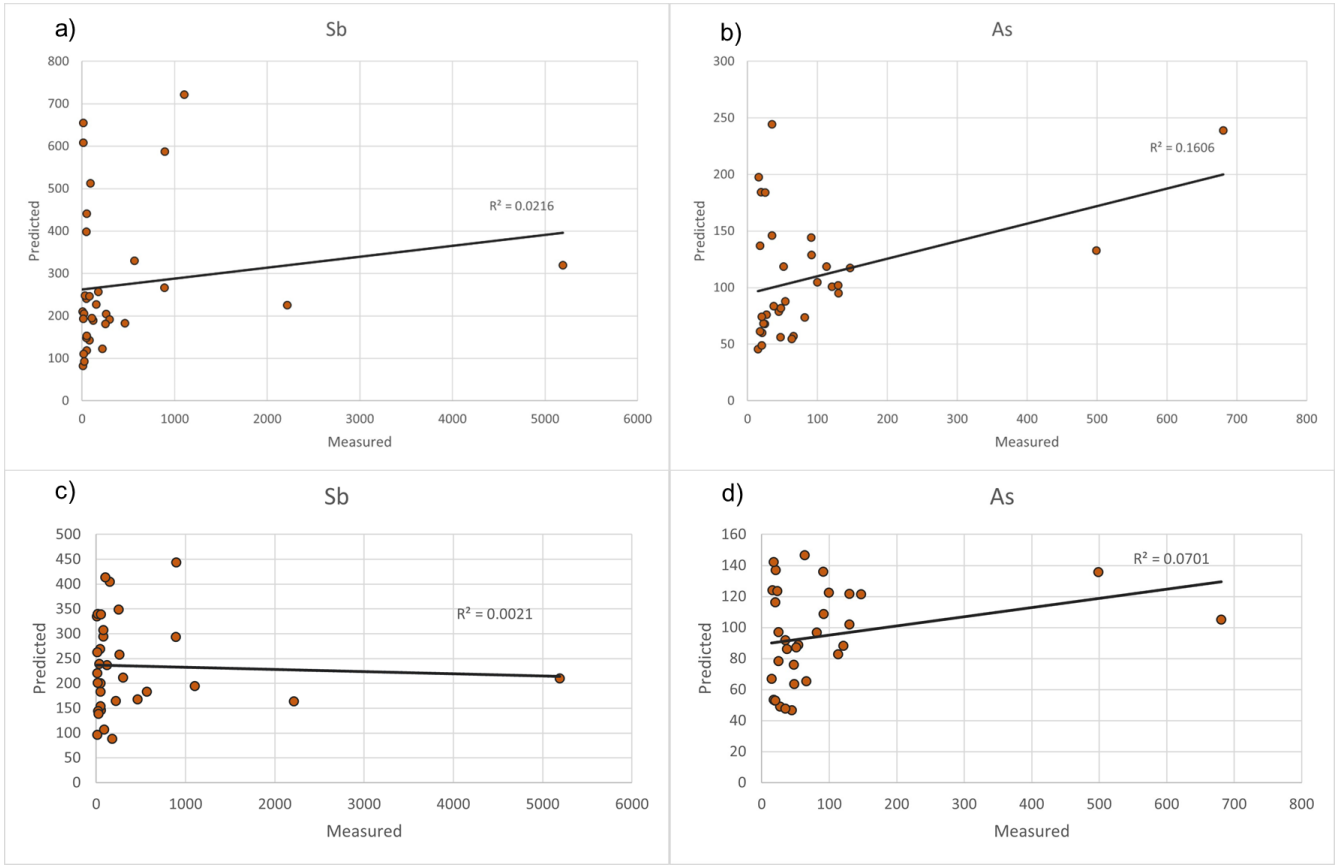

Figure 38: a) and b) test T4b3 using 500 trees. c) and d) test T4b2 using 50 trees.

10 Conclusions

The geochemical distribution of the elements and results obtained by the analysis of the soil samples, and interpretation of the factor analysis indicate that the high concentrations of Sb found in the soils in Ribeiro da Serra and Tapada can be mostly attributed to the influence of tailings left by the mining works in the last century. However, there is one sampling point where the Sb anomaly doesn't seem to be associated with the tailings and also is not associated with known veins or having influence of water lines, being a potential interest zone.

The evaluation of the pXRF proves its efficiency in the exploration of Sb, providing fast and satisfactory results with no need for previous preparation of the samples. However, to use correlations of geochemical data between values of several elements obtained by pXRF, it is important to consider that for some elements the results cannot be accurate.

Between the spectral pretreatments, a more balanced data set, by downsampling, had a positive effect, highlighting the importance of a balanced data set. Also, the application of SG and the first derivative had positive effects.

Concerning ML algorithm training, the CNN model performed better than the RF and PLSR regression, but avoiding the overfitting in the model was not accomplished.

Also, the soil sampling campaigns that were executed and provided the soil samples for this study didn't focus on this kind of problem, but on the Sb distribution. Thus, the possibility of the collected data not being representative of the problem, as many of the soil samples with Sb high values are from tailings, cannot be discarded. Another sampling methodology, focusing on the soils near the known Sb veins and in the progressive contents of Sb in the soils associated with the veins could be more appropriate for this study and can work as a solution for the overfitting in the CNN model. And, also, obtaining organic matter and clay data from the soils can be a different approach that can help to better understand the features in the spectral signature of soils containing Sb.

11 References

Abedini, M., Ziaii, M., Timkin, T., & Pour, A. B. (2023). Machine Learning (ML)-Based Copper Mineralization Prospectivity Mapping (MPM) Using Mining Geochemistry Method and Remote Sensing Satellite Data. Remote Sensing, 15(15), 3708.

Acharya, T., & Ray, A. K. (2005). Image processing: principles and applications. John Wiley & Sons.

Ayodele, T. O. (2010). Machine learning overview. New Advances in Machine Learning, 2, 9-18.

Baíllo, A., & Chacón, J. E. (2021). Statistical outline of animal home ranges: an application of set estimation. In Handbook of Statistics (Vol. 44, pp. 3-37). Elsevier

Butterman, W. C., & Carlin Jr, J. F. (2004). Mineral commodity profiles: Sb (No. 2003-19).

Cardoso-Fernandes, J., Silva, J., Dias, F., Lima, A., Teodoro, A. C., Barrès, O., ... & Ribeiro, M. A. (2021). Tools for remote exploration: A lithium (li) dedicated spectral library of the Fregeneda–almendra aplite–pegmatite field. Data, 6(3), 33.

Cardoso-Fernandes, J. M. (2018). Remote Sensing Applied to the Bajoca Litiniferous Area (Master dissertation, Universidade do Porto).

Carta Geológica de Portugal 1: 200.000. Serviços Geológicos de Portugal. Lisboa. Folha 2 (Pereira et al, 2006).

Carvalho, A. D. de. (1969). Minas de Antimónio e Ouro de Gondomar. In Publicações, Estudos, Notas e Trabalhos: Vol. XIX.

Caté, A., Perozzi, L., Gloaguen, E., & Blouin, M. (2017). Machine learning as a tool for geologists. The Leading Edge, 36(3), 215-219.

Cheng, H., Shen, R., Chen, Y., Wan, Q., Shi, T., Wang, J., ... & Li, X. (2019). Estimating heavy metal concentrations in suburban soils with reflectance spectroscopy. Geoderma, 336, 59-67.

Couto, H. (1993). As mineralizações de Sb-Au da região Dúrico-Beirã [Faculdade de Ciências da Universidade do Porto]. http://hdl.handle.net/10216/10678

Couto, H., Roger, G., Moëlo, Y., & Bril, H. (1990). Le district à antimoine-or Dúrico-Beirão (Portugal): évolution paragénétique et géochimique; implications métallogéniques. Mineralium Deposita, 25(S1), S69–S81. https://doi.org/10.1007/BF00205252

de Santana, F. B., de Souza, A. M., & Poppi, R. J. (2018). Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 191, 454-462.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. Practical assessment, research, and evaluation, 14(1), 20.

Farahbakhsh, E., Maughan, J., & Müller, R. D. (2023). Prospectivity modelling of critical mineral deposits using a generative adversarial network with oversampling and positive-unlabelled bagging. Ore Geology Reviews, 105665.

Ferreira, M. R. P. V. (1970). Relatório sobre a Mina de Antimónio e Ouro de Alto do Sobrido.

Flesche, H., Nielsen, A. A., & Larsen, R. (2000). Supervised mineral classification with semiautomatic training and validation set generation in scanning electron microscope energy dispersive spectroscopy images of thin sections. Mathematical Geology, 32(3), 337-366.

Frutuoso, R. (2018). Soil Sampling Campaign Report Ribeiro da Serra Mine.

Géron, A. (2017). Hands-On machine learning with Scikit-Learn and TensorFlow: concepts, tools. and Techniques to Build Intelligent Systems. nd.

Glover, P. W., Mohammed-Sajed, O. K., Akyüz, C., Lorinczi, P., & Collier, R. (2022). Clustering of facies in tight carbonates using machine learning. Marine and Petroleum Geology, 144, 105828.

Goff, K., Schaetzl, R. J., Chakraborty, S., Weindorf, D. C., Kasmerchak, C., & Bettis III, E. A. (2020). Impact of sample preparation methods for characterizing the geochemistry of soils and sediments by portable X-ray fluorescence. Soil Science Society of America Journal, 84(1), 131-143.

Gomez, C., & Lagacherie, P. (2016). Mapping of primary soil properties using optical visible and near infrared (Vis-NIR) remote sensing. In Land surface remote sensing in agriculture and forest (pp. 1-35). Elsevier.

Gonzales, R. C., & Wintz, P. (1987). Digital image processing. Addison-Wesley Longman Publishing Co., Inc..

Gumiel, P., & Arribas, A. (1987). Sb deposits in the Iberian Peninsula. Economic Geology, 82(6), 1453-1463.

Guimarães, D., Praamsma, M. L., & Parsons, P. J. (2016). Evaluation of a new optic-enabled portable XRF instrument for measuring toxic metals/metalloids in consumer goods and cultural products. Spectrochimica acta. Part B, Atomic spectroscopy, 122, 192.

Guo, B., Guo, X., Zhang, B., Suo, L., Bai, H., & Luo, P. (2022). Using a Two-Stage Scheme to Map Toxic Metal Distributions Based on GF-5 Satellite Hyperspectral Images at a Northern Chinese Opencast Coal Mine. Remote Sensing, 14(22), 5804.

He, J., Wei, Y., Zhai, T., & Li, H. (2018). Sb-based materials as promising anodes for rechargeable lithium-ion and sodium-ion batteries. Materials Chemistry Frontiers, 2(3), 437-455.

Hongyu, K. (2018). Análise Fatorial Exploratória: resumo teórico, aplicação e interpretação. E&S Engineering and Science, 7(4), 88-103.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Jackson, E., Bedle, H., & Ha, T. (2023). Assessment of spectral attributes in identifying gas hydrates in seismic data from the Pegasus Basin, offshore New Zealand. Marine Geophysical Research, 44(4), 21.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., & Kumar, V. (2018). Machine learning for the geosciences: Challenges and opportunities. IEEE Transactions on Knowledge and Data Engineering, 31(8), 1544-1554.

Kemper, T., & Sommer, S. (2002). Estimate of heavy metal contamination in soils after a mining accident using reflectance spectroscopy. Environmental science & technology, 36(12), 2742-2747.

Klovan, J.E. (1975). R- and Q-Mode Factor Analysis. In: McCammon, R.B. (eds) Concepts in Geostatistics. Springer Study Edition. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-85976-2_2

Kulkarni, A. D. (1986). Digital processing of remotely sensed data. In Advances in electronics and electron physics (Vol. 66, pp. 309-368). Academic Press.

Kumar, P. C., & Sain, K. (2023). Machine learning elucidates the anatomy of buried carbonate reef from seismic reflection data. Artificial Intelligence in Geosciences, 4, 59-67.

Latunussa, C. EL, Georgitzikis, K., de Matos, C. T., Grohol, M., Eynard, U., Wittmer, D., Mancini, L., Unguru, M., Pavel, C., Carrara, S., Mathieux, F., Pennington, D., Blengini G. A. European Commission, Study on the EU's list of Critical Raw Materials (2020), Factsheets on Critical Raw Materials

Lemière, B., Melleton, J., Auger, P., Derycke, V., Gloaguen, E., Bouat, L., ... & Middleton, M. (2020). pXRF measurements on soil samples for the exploration of an antimony deposit: Example from the Vendean Antimony District (France). Minerals, 10(8), 724.

Li, T. (2011). Sb and Sb Alloys. In Kirk-Othmer Encyclopedia of Chemical Technology, (Ed.). https://doi.org/10.1002/0471238961.011420091209.a01.pub3

Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386.

Lillesand, T., Kiefer, R. W., & Chipman, J. (2015). Remote sensing and image interpretation. John Wiley & Sons.
Manzo, C., Salvini, R., Guastaldi, E., Nicolardi, V., & Protano, G. (2013). Reflectance spectral analyses for the assessment of environmental pollution in the geothermal site of Mt. Amiata (Italy). Atmospheric environment, 79, 650-665.

Medeiros, A. C. de. (1964). Notícia Explicativa da Folha 13-B - Castelo de Paiva.

Moolayadukkam, S., Bopaiah, K. A., Parakkandy, P. K., & Thomas, S. (2022). Sb (Sb)-Based Anodes for Lithium–Ion Batteries: Recent Advances. Condensed Matter, 7(1), 27.

Moura, A., & Velho, J. L. (2012). Recursos geológicos de Portugal.

Nanni, M. R., & Demattê, J. A. M. (2006). Spectral reflectance methodology in comparison to traditional soil analysis. Soil science society of America journal, 70(2), 393-407.

Neiva, A. M., Andráš, P., & Ramos, J. M. F. (2008). Sb quartz and Sb–gold quartz veins from northern Portugal. Ore Geology Reviews, 34(4), 533-546.

Nogueira, P., & Maia, M. (2023). Magnetite Talks: Testing Machine Learning Models to Untangle Ore Deposit Classification—A Case Study in the Ossa-Morena Zone (Portugal, SW Iberia). Minerals, 13(8), 1009.

Okamoto, K., Mukuhira, Y., Darisma, D., Asanuma, H., & Moriya, H. (2024). Machine learning automatic picker for geothermal microseismicity analysis for practical procedure to reveal fine reservoir structures. Geothermics, 116, 102832.

Patinha, P... [et.al.] - SIORMINP: Plataforma de gestão e disponibilização de dados harmonizados dos Recursos Minerais Portugueses. In: Comunicações Geológicas, 2020, Vol. 107, nº Especial I, p. 61-63

Professional Aerial Photographers Association, "History of aerial photography", access on 29/01/2023: https://professionalaerialphotographers.com/content.aspx?page_id=22&club_id=808138&module_id=158950

Pyo, J., Hong, S. M., Kwon, Y. S., Kim, M. S., & Cho, K. H. (2020). Estimation of heavy metals using deep neural network with visible and infrared spectroscopy of soil. Science of the Total Environment, 741, 140162.

Quesada, C. (1991). Geological constraints on the Paleozoic tectonic evolution of tectonostratigraphic terranes in the Iberian Massif. Tectonophysics, 185(3-4), 225-245.

Rodríguez-Fernández, L. R., López-Olmedo, F., Oliveira, J. T., Medialdea, T., Terrinha, P., Matas, J., ... & Nozal, F. (2015). Mapa Geológico de España y Portugal a escala 1: 1.000. 000. Instituto Geológico y Minero de España, Laboratório Nacional de Energía e Geología de Portugal: Madrid, Spain.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. J. O. G. R. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews, 71, 804-818.

Sabins, F. F. (1997). Remote Sensing—Principles and Interpretation. 3rd Edition, W.H. Freeman, New York, NY

Senanayake, I. P., Kiem, A. S., Hancock, G. R., Metelka, V., Folkes, C. B., Blevin, P. L., & Budd, A. R. (2023). A Spatial Data-Driven Approach for Mineral Prospectivity Mapping. Remote Sensing, 15(16), 4074.

Schwartz, Guy & Eshel, Gil & Ben-Dor, Eyal. (2011). Reflectance Spectroscopy as a Tool for Monitoring Contaminated Soils. 10.5772/23661.

Shi, Z., Zuo, R., & Zhou, B. (2023). Deep Reinforcement Learning for Mineral Prospectivity Mapping. Mathematical Geosciences, 1-25.

Wenhuan Kuang, Jie Zhang, Wei Zhang, A novel deep-learning image condition for locating earthquake, Geophysical Journal International, Volume 235, Issue 3, December 2023, Pages 2168–2178, https://doi.org/10.1093/gji/ggad350

Wisniak, J. (2005). Nicolas Lémery. Revista CENIC. Ciencias Químicas, 36(2), 123-130.

Wu, Y., Chen, J., Ji, J., Gong, P., Liao, Q., Tian, Q., & Ma, H. (2007). A mechanism study of reflectance spectroscopy for investigating heavy metals in soils. Soil Science Society of America Journal, 71(3), 918-926.

Yang, J., Wang, X., Wang, R., & Wang, H. (2020). Combination of convolutional neural networks and recurrent neural networks for predicting soil properties using Vis–NIR spectroscopy. Geoderma, 380, 114616.

Zhou, W., Yang, H., Xie, L., Li, H., Huang, L., Zhao, Y., & Yue, T. (2021). Hyperspectral inversion of soil heavy metals in Three-River Source Region based on random forest model. Catena, 202, 105222.

Annexes

Annex A: Uncertainty

```
# coding: utf-8

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
get_ipython().run_line_magic('matplotlib', 'inline')
plt.figure(facecolor='white')
plt.rcParams["figure.figsize"] = (20,20)

_df_origin = pd.read_csv("XRF_TAPADA.csv", delimiter=";", on_bad_lines='skip')

_df_origin["Fe"]


_df_origin.shape


# start from second and skip every 3 lines

df_origin = _df_origin[1::]

df_origin
df_origin.columns

df_clean = df_origin.drop(["Rh", "Pd", "Ag", "In", "Pt", "Au", "Hg"], axis=1)

# fill the column #samples"

aux = ""
for index, row in df_clean.iterrows():
    if not pd.isnull(df_clean.loc[index,'sample']):
        aux = df_clean.loc[index,'sample']
    else:
        df_clean.loc[index,'sample'] = aux

cols = [c for c in df_clean.columns if "Err" not in c][1:]
cols_err = [c for c in df_clean.columns if "Err" in c]

map_type = dict([(c, "float") for c in cols])
map_type_err = dict([(c, "float") for c in cols_err])

df_clean = df_clean.astype(map_type)
df_clean = df_clean.astype(map_type_err)

out = {}
for c in cols:
    name_in = c + " incerteza"
    name_mean = c + " media"
    name_mean_sup = c + " superior"
    name_mean_inf = c + " inferior"

    out["sample"] = []
    out[name_in] = []
    out[name_mean] = []
    out[name_mean_sup] = []
    out[name_mean_inf] = []

len(out["sample"])

df_clean.groupby(['sample'])
```

```
61. 
62. 
63. df_out = pd.DataFrame()
64. for key, grp in df_clean.groupby(['sample']):
65. 
66.     out["sample"].append(key[0])
67.     for c in cols:
68.         c_erro = [ce for ce in cols_err if ce.split(" ")[0] == c]
69.         std = np.std(grp[c].astype(float))
70.         err_max = grp[c_erro].max()
71.         incerteza = np.sqrt(std**2 + err_max ** 2)
72.         incerteza_value = incerteza.values[0]
73. 
74.         name = c + " incerteza"
75.         out[name].append(incerteza_value)
76. 
77.         mean = np.mean(grp[c])
78.         name = c + " media"
79.         out[name].append(mean)
80. 
81.         name = c + " superior"
82.         out[name].append(mean + incerteza_value)
83. 
84.         name = c + " inferior"
85.         out[name].append(mean - incerteza_value)
86. df_incerteza = pd.DataFrame(data=out)
87. df_incerteza.to_csv("incertezas_tapada.csv")
88. df_incerteza
89. 
90. # ## Save partial
91. df_incerteza.columns
92. columns = ["Co incerteza","Co media","Co superior", "Co inferior"] # From above list
93. df_partial = df_incerteza.loc[:, columns]
94. df_partial.to_csv("partial_incertezas.csv")
95. 
96. df_icp = pd.read_csv("Tapada_ICPMS.csv", delimiter=";", on_bad_lines='skip')
97. 
98. 
99. columns_pct = df_icp.columns[np.where(df_icp.values[0] == "%")]
100. columns_ppm = df_icp.columns[np.where(df_icp.values[0] != "%")]
101. 
102. elements = df_icp.columns[1:]
103. elements
104. 
105. elements_incerteza = df_incerteza.columns
106. elements_incerteza
107. 
108. out_el = {}
109. 
110. for e in elements:
111.     out_el[e] = []
112.     for ei in elements_incerteza:
113.         if ei.startswith(e):
114.             out_el[e].append(ei)
115. 
116. out_el = {k:out_el[k] for k,v in out_el.items() if len(v) != 0}
117. 
118. 
119. del out_el["P"]
120. 
121. out_el
122. 
123. df_out = pd.merge(df_icp, df_incerteza, on="sample", how='outer') # All samples
124. df_out = pd.merge(df_icp, df_incerteza, on="sample", how='inner') # Only icp samples
125. df_out.fillna(0)
126. 
```

```
elem = "Fe"
v = ["{} media".format(elem), "{} superior".format(elem), "{} inferior".format(elem), "sample"]
v.append(elem)

_subset = df_out.loc[:, v][1:]
subset = _subset.sort_values(by=[v[0]])
x = subset["sample"]
#subset
plt.rcParams["figure.figsize"] = (40,40)

#https://matplotlib.org/stable/gallery/color/named_colors.html (name of the colors)
#plt.yscale('log')  #comment for normal scale
plt.xticks(rotation=90, ha='right')

plt.grid()

plt.plot(x, subset[v[1]], c="lightblue", zorder=1)
plt.plot(x, subset[v[2]], c="lightblue", zorder=1)
plt.plot(x, subset[v[0]], zorder=2)
plt.scatter(x, subset[elem].fillna(0).astype('float32'), c="red", zorder=4)
plt.fill_between(x, subset[v[1]], subset[v[2]], color="lightblue")

```

Annex B: Reflectance to absorption and SG

```
# coding: utf-8
import numpy as np
from scipy.signal import savgol_filter
import matplotlib.pyplot as plt
import pandas as pd
import pickle
import csv
import math

experiment_name="T6"
input_csv = r"C:\Users\espectra_soils\data\T6\clean_csv.csv" #receves the clean csv
output_pickle= f"../data/{experiment_name}/spectrum_degraded.pickle"
df = pd.read_csv(input_csv, delimiter=";", skiprows=[1])

# invert reflectance to absortion

#  A = -log10(R)
def _calculate_absorbance(lst_reflectance):
    result = []
    for reflectance in lst_reflectance:
        reflectance = float(reflectance)  # Convert reflectance to a float
        if reflectance == 0:
             continue # Handle zero reflectance case

        result.append(math.log (1 / reflectance))
    return result

    absortion_converted    =    pd.concat([df["Column1"],    df["Wvl"],    df.iloc[:, 2:].apply(_calculate_absorbance)], axis=1)
           absortion_converted.to_csv(f'../data/{experiment_name}/absorbance_data.csv', index=False)

df

absortion_converted

input_SG = f"../data/{experiment_name}/absorbance_data.csv"
iSG = pd.read_csv(input_SG)

iSG

# save data as csv and as pickle

window_size = 31  # Number of data points to use for the polynomial fit
poly_order = 2  # Order of the polynomial
def calcula_smooth(sample):
    window_size = 31  # Number of data points to use for the polynomial fit
    poly_order = 2  # Order of the polynomial
    return savgol_filter(sample, window_size, poly_order)

    smooth_data     =     pd.concat([iSG["Column1"],     iSG["Wvl"],     iSG.iloc[:, 2:].apply(calcula_smooth)], axis=1)

# Save smoothed_data as a CSV file
smooth_data.to_csv(f'../data/{experiment_name}/smoothed_data.csv', index=False)

# # Save smoothed_data as a pickle file
# with open(f'../data/{experiment_name}/smoothed_data.pickle', 'wb') as handle:
#     pickle.dump(smooth_data, handle, protocol=pickle.HIGHEST_PROTOCOL)

input_Smda = f"../data/{experiment_name}/smoothed_data.csv"
iS = pd.read_csv(input_Smda)
iS
```

```
61.
62. # # Load the pickle file (check if pickle works)
63. # with open('../data/smoothed_data.pickle', 'rb') as handle:
64. #     loaded_data = pickle.load(handle)
65.
66. # # Print or inspect the loaded data
67. # print(loaded_data)
68.
69. # Calculate first derivative
70.
71. def calculate_derivative(data):
72.     deriv_order=1
73.     first_derivative = np.gradient(data, deriv_order, axis=0)
74.     return first_derivative
75.
76.     first_derivative     =     pd.concat([iS["Column1"],     iS["Wvl"],     iS.iloc[:,
2:].apply(calculate_derivative)], axis=1)
77.
78. first_derivative
79.
80.        first_derivative.to_csv(f'../data/{experiment_name}/first_derivative_data.csv',
index=False)
81. # # Save first derivative as a pickle file
82. # with open(f'../data/{experiment_name}/first_derivative.pickle', 'wb') as handle:
83. #     pickle.dump(first_derivative, handle, protocol=pickle.HIGHEST_PROTOCOL)
84.
85. # Calculate second derivative
86. def calculate_segunda_derivative(data):
87.     deriv_order=2
88.     second_derivative = np.gradient(data, deriv_order, axis=0)
89.     return second_derivative
90.
91.     second_derivative     =     pd.concat([df["Column1"],     df["Wvl"],     df.iloc[:,
2:].apply(calculate_segunda_derivative)], axis=1)
92. second_derivative
93.
94. # Save the second derivative as a CSV file
95.      second_derivative.to_csv(f'../data/{experiment_name}/second_derivative_data.csv',
index=False)
96. # # Save first derivative as a pickle file
97. # with open(f'../data/{experiment_name}/second_derivative.pickle', 'wb') as handle:
98. #     pickle.dump(second_derivative, handle, protocol=pickle.HIGHEST_PROTOCOL)
99.
100. #view data
101. first_derivative = r"C:\Users\espectra_soils\data\T6\first_derivative_data.csv"
102. second_derivative = r"C:\Users\espectra_soils\data\T6\second_derivative_data.csv"
103. fd = pd.read_csv(first_derivative)
104. sd = pd.read_csv(second_derivative)
105.
106. sd
107.
108. # Plot the original and smoothed data, and the first and second order derivatives
109. wavelengths = df["Wvl"]
110. spectral_data =df["TP179_1"]
111. first_derivative= fd["TP179_1"]
112. second_derivative= sd["TP179_1"]
113. fig, axs = plt.subplots(3, 1, figsize=(32, 32), sharex=True)
114. axs[0].plot(wavelengths, spectral_data, label='Original')
115.   axs[0].plot(wavelengths,   savgol_filter(spectral_data,   window_size,   poly_order),
label='Smoothed')
116. axs[0].set_ylabel('Intensity')
117. axs[0].legend()
118. axs[1].plot(wavelengths, first_derivative, label='First Derivative')
119. axs[1].set_ylabel('dIntensity/dWavelength')
120. axs[1].legend()
121. axs[2].plot(wavelengths, second_derivative, label='Second Derivative')
```

```
axs[2].set_xlabel('Wavelength (nm)')
axs[2].set_ylabel('d^2Intensity/dWavelength^2')
axs[2].legend()
plt.show()
```

Annex C: Conversion to Spectrogram

```
# coding: utf-8
import pandas as pd
import tensorflow as tf
import numpy as np
import matplotlib.pyplot as plt
from scipy.io import wavfile
from scipy import signal
import pickle

# Conversion to Spectrogram
from tensorflow.keras import datasets, layers, models
import matplotlib.pyplot as plt
get_ipython().run_line_magic('matplotlib', 'inline')

# select file
experiment_name="T6"
input_file = f'../data/{experiment_name}/sample_matched.pickle'
output_file = f'../data/{experiment_name}/spectogram.pickle'

# Load data (deserialize)
with open(input_file, 'rb') as handle:
    data_dict = pickle.load(handle)

#df = np.load("../data/processed_clean.npy")

#df.shape

# # Spectogram normalization

def normalize_spectrogram(spec):
    """
    Normalize a spectrogram using min-max scaling.
    :param spec: Input spectrogram.
    :return: Normalized spectrogram.
    """
    # Compute the minimum and maximum values of the spectrogram.
    spec_min = np.min(spec)
    spec_max = np.max(spec)

    # Scale the spectrogram to the range [0, 1].
    spec_norm = (spec - spec_min) / (spec_max - spec_min)

    return spec_norm


def get_spectrogram(waveform):
    # Zero-padding for an audio waveform with less than 16,000 samples.
    input_len = 2151
    waveform = waveform[:input_len]
    zero_padding = tf.zeros(
      [2151] - tf.shape(waveform),
      dtype=tf.float32)
    # Cast the waveform tensors' dtype to float32.
    waveform = tf.cast(waveform, dtype=tf.float32)
    # Concatenate the waveform with `zero_padding`, which ensures all audio
    # clips are of the same length.
    equal_length = tf.concat([waveform, zero_padding], 0)
    # Convert the waveform to a spectrogram via a STFT.
```

```
    spectrogram = tf.signal.stft(
      equal_length, frame_length=128, frame_step=64)
    # Obtain the magnitude of the STFT.
    spectrogram = tf.abs(spectrogram)
    # Add a `channels` dimension, so that thbe spectrogram can be used
    # as image-like input data with convolution layers (which expect
    # shape (`batch_size`, `height`, `width`, `channels`).
    spectrogram = spectrogram[..., tf.newaxis]
    return spectrogram

def plot_spectrogram(spectrogram, ax):
    if len(spectrogram.shape) > 2:
        assert len(spectrogram.shape) == 3
    spectrogram = np.squeeze(spectrogram, axis=-1)


    # Add an epsilon to avoid taking a log of zero.
    log_spec = np.log(spectrogram.T + np.finfo(float).eps)
    height = log_spec.shape[0]
    width = log_spec.shape[1]
    X = np.linspace(0, np.size(spectrogram), num=width, dtype=int)
    Y = range(height)
    ax.pcolormesh(X, Y, log_spec)


def draw_spec(waveform):
    spectrogram = get_spectrogram(waveform)
    fig, axes = plt.subplots(2, figsize=(12, 8))
    timescale = np.arange(waveform.shape[0])
    axes[0].plot(timescale, waveform)
    axes[0].set_title('Waveform')
    axes[0].set_xlim([0, 2151])

    spec = normalize_spectrogram(spectrogram.numpy())

    plot_spectrogram(spec, axes[1])
    axes[1].set_title('Spectrogram')
    plt.show()

data = data_dict["RS001"].iloc[:, 1]
draw_spec(data)

def map_dict_wave_spec(data):
    o_aux = {}
    for sample, data in data.items():
        aux = []
        for c in data.columns:
            freq, time, spec = signal.spectrogram(data[c])
            aux.append(normalize_spectrogram(spec))
        o_aux[sample] = np.array(aux)
    return o_aux

def map_wave_spec(data):
    o_aux = []
    for d in data:
        aux = []
        for l in d:
            freq, time, spec = signal.spectrogram(l)
            aux.append(spec)
        o_aux.append(aux)
    return np.array(o_aux)

specs = map_dict_wave_spec(data_dict)
specs["RS001"].shape

#tf.squeeze(tf.transpose(specs, [0, 2, 3, 1])).shape
```

```
125.
126. #respec = tf.squeeze(tf.transpose(specs, [0, 2, 3, 1]))
127.
128. #norm_spec = normalize_spectrogram(specs)
129.
130. # np.save("../data/spectograms_scipy.npy", specs)
131.
132. with open(output_file, 'wb') as handle:
133.     pickle.dump(specs, handle, protocol=pickle.HIGHEST_PROTOCOL)
134.
```

Annex D: Random Forest and Partial Least Squares Regression

```
  1. # coding: utf-8
  2.
  3. import pandas as pd
  4. import tensorflow as tf
  5. import numpy as np
  6. import os
  7. import datetime
  8. import pickle
  9.
 10. from sklearn.model_selection import train_test_split
 11.
 12. from tensorflow.keras import datasets, layers, models
 13. import matplotlib.pyplot as plt
 14. get_ipython().run_line_magic('matplotlib', 'inline')
 15.
 16. # INPUT DATA
 17.
 18. experiment_name="T7"
 19. input_data= f'../data/{experiment_name}/sample_matched.pickle' # gets spectral data and
labels (targets) for waveforms
labels (targets)for spectrograms
 20.  #input_data=  f'../data/{experiment_name}/spectogram.pickle'  #gets  spectral  data  and
labels (targets) for waveforms
 21. input_label= f"../data/{experiment_name}/targets_full.csv"
 22.
 23. output_file = f"../data/{experiment_name}/random_forest-result-{experiment_name}.csv"
 24.
 25. # labels and date quantities must be the same
 26.
 27. #with open('../data/data.pickle', 'rb') as handle:
 28. #    waves_dict = pickle.load(handle)
 29. with open(input_data, 'rb') as handle:
 30.     specs_dict = pickle.load(handle)
 31.
 32. labels = pd.read_csv(input_label)
 33.
 34. labels
 35.
 36. # _labels = labels[["amostra", "Sb_ppm", "As_ppm", "Pb_ppm"]]
 37. _labels = labels[["amostra", "As_ppm"]]
 38. specs = []
 39. target = []
 40. for r in _labels.iterrows():
 41. #     label = [r[1][1], r[1][2], r[1][3]]    # define what to use
 42.     label = [r[1][1]]    # define what to use
 43.     spectograms = specs_dict[r[1][0]]
 44.    #spectograms = tf.squeeze(tf.transpose(spectograms, [1, 2, 0])) # Descomentar quando
usar espectograma
 45.     #spectograms = tf.squeeze(tf.transpose(spectograms, [1, 0])) # Descomentar quando
usar espectograma
 46.     target.append(label)
```

```
47.     specs.append(spectograms)
48. specs = np.array(specs)
49. target = np.array(target)
50. 
51. specs = specs.reshape(specs.shape[0], specs.shape[1] * specs.shape[2] ) # for spectogram
52. #specs = specs.reshape(specs.shape[0], specs.shape[1] * specs.shape[2])  # for waveform
53. 
54. specs.shape, target.shape
55. 
56. 
57. train_examples, test_examples, train_labels, test_labels = train_test_split(specs,
target, test_size=0.33, random_state=42)
58. 
59. # ### random forest starts
60. 
61. from sklearn.ensemble import RandomForestRegressor
62. 
63. # Create random forest regressor object
64. print(" Create random forest regressor object")
65. rf = RandomForestRegressor(n_estimators=500, max_depth=20, random_state=42)
66. 
67. print(" Train the model on the training data")
68. 
69. # Train the model on the training data
70. rf.fit(train_examples, train_labels)
71. 
72. 
73. from sklearn.metrics import mean_squared_error, r2_score
74. y_pred = rf.predict(test_examples)
75. mse = mean_squared_error(test_labels, y_pred)
76. rmse = np.sqrt(mse)
77. r2 = r2_score(test_labels, y_pred)
78. 
79. print("Random Forest RMSE:", rmse)
80. print("Random Forest MSE:", mse)
81. print("Coefficient of Determination (R^2):", r2)
82. 
83. 
84. y_pred
85. 
86. 
87. _labels.columns[1:]
88. 
89. 
90. df_dict = {}
91. _test_labels = list(zip(*test_labels))
92. #_test_pred = list(zip(*y_pred))  # more than two labels
93. _test_pred = [y_pred]    # one label
94. for idx, column in enumerate(_labels.columns[1:]):
95.     df_dict[column] = _test_labels[idx]
96.     key_pred = f"{column}_pred"
97.     df_dict[key_pred] = _test_pred[idx]
98. 
99. df = pd.DataFrame(df_dict)
100. 
101. df
102. 
103. df.to_csv(output_file)
104. 
105. # # PLSRegression Starts
106. 
107. from sklearn.cross_decomposition import PLSRegression
108. from sklearn.metrics import mean_squared_error, r2_score
109. from sklearn.model_selection import train_test_split
110. from sklearn.preprocessing import StandardScaler
111. 
```

```
# Standardize the predictor variables
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(train_examples)
X_test_scaled = scaler.transform(test_examples)

# Initialize and fit the PLSR model
n_components = 3  # Number of components/latent variables
plsr = PLSRegression(n_components=n_components)
plsr.fit(X_train_scaled, train_labels)
# Make predictions on the test set
y_pred = plsr.predict(X_test_scaled)
# Model evaluation
mse = mean_squared_error(test_labels, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(test_labels, y_pred)
# Print evaluation metrics
print("Mean Squared Error (MSE):", mse)
print("Root Mean Squared Error (RMSE):", rmse)
print("Coefficient of Determination (R^2):", r2)
```

Annex E: Convolutional Neural Networks

```
# coding: utf-8

import pandas as pd
import tensorflow as tf
import numpy as np
import os
import datetime
import pickle

from sklearn.model_selection import train_test_split

from tensorflow.keras import datasets, layers, models
import matplotlib.pyplot as plt
get_ipython().run_line_magic('matplotlib', 'inline')

experiment_name="T6c"
input_data= f'../data/{experiment_name}/spectogram.pickle' # gets spectral data and labels (targets) for waveforms
input_label= f"../data/{experiment_name}/targets_full.csv"

#with open('../data/data.pickle', 'rb') as handle:
#    waves_dict = pickle.load(handle)
with open(input_data, 'rb') as handle:
    specs_dict = pickle.load(handle)

labels = pd.read_csv(input_label)

# ## Generate labels

_labels = labels[["sample", "Sb_ppm"]]
specs = []
target = []
for r in _labels.iterrows():
    label = [r[1][1]] # define, see ex.
    spectograms = specs_dict[r[1][0]]
    respec = tf.squeeze(tf.transpose(spectograms, [1, 2, 0]))
    target.append(label)
    specs.append(respec)
specs = np.array(specs)
target = np.array(target)

exemple:
r[1] = [amostra, as, sb]
r[1][0] = amostra
r[1][1] = as
r[1][2] = sb
# In[100]:

specs.shape

# ## data split between train and test

# test_ size (%)
#

train_examples, test_examples, train_labels, test_labels = train_test_split(specs, target, test_size=0.33, random_state=42)

#convert data in array from np and to tensor format
train_dataset = tf.data.Dataset.from_tensor_slices((train_examples, train_labels))
test_dataset = tf.data.Dataset.from_tensor_slices((test_examples, test_labels))

AUTOTUNE = tf.data.AUTOTUNE
```

```
train_dataset = train_dataset.prefetch(buffer_size=AUTOTUNE)

BATCH_SIZE = 64


train_dataset = train_dataset.batch(BATCH_SIZE)

train_examples.shape

# # Model definition (Mobilenetv2)

#convolution, batch normalization, activation function (relu)
def _conv_block(inputs, filters, kernel, strides):
    """Convolution Block
    This function defines a 2D convolution operation with BN and relu6.
    # Arguments
        inputs: Tensor, input tensor of conv layer.
        filters: Integer, the dimensionality of the output space.
        kernel: An integer or tuple/list of 2 integers, specifying the
            width and height of the 2D convolution window.
        strides: An integer or tuple/list of 2 integers,
            specifying the strides of the convolution along the width and height.
            Can be a single integer to specify the same value for
            all spatial dimensions.
    # Returns
        Output tensor.
    """

   x = tf.keras.layers.Conv2D(filters, kernel, padding='same', strides=strides)(inputs)
    x = tf.keras.layers.BatchNormalization()(x)
    return tf.nn.relu6(x)

def _bottleneck(inputs, filters, kernel, t, s, r=False):
    """Bottleneck
    This function defines a basic bottleneck structure.
    # Arguments
        inputs: Tensor, input tensor of conv layer.
        filters: Integer, the dimensionality of the output space.
        kernel: An integer or tuple/list of 2 integers, specifying the
            width and height of the 2D convolution window.
        t: Integer, expansion factor.
            t is always applied to the input size.
        s: An integer or tuple/list of 2 integers,specifying the strides
            of the convolution along the width and height.Can be a single
            integer to specify the same value for all spatial dimensions.
        r: Boolean, Whether to use the residuals.
    # Returns
        Output tensor.
    """

    tchannel = inputs.shape[-1] * t

    x = _conv_block(inputs, tchannel, (1, 1), (1, 1))

      x = tf.keras.layers.DepthwiseConv2D(kernel, strides=(s, s), depth_multiplier=1,
padding='same')(x)
    x = tf.keras.layers.BatchNormalization()(x)
    x = tf.nn.relu6(x)

    x = tf.keras.layers.Conv2D(filters, (1, 1), strides=(1, 1), padding='same')(x)
    x = tf.keras.layers.BatchNormalization()(x)

    if r:
        x = tf.keras.layers.add([x, inputs])
    return x

```

```
def _inverted_residual_block(inputs, filters, kernel, t, strides, n):
    """Inverted Residual Block
    This function defines a sequence of 1 or more identical layers.
    # Arguments
        inputs: Tensor, input tensor of conv layer.
        filters: Integer, the dimensionality of the output space.
        kernel: An integer or tuple/list of 2 integers, specifying the
            width and height of the 2D convolution window.
        t: Integer, expansion factor.
            t is always applied to the input size.
        s: An integer or tuple/list of 2 integers,specifying the strides
            of the convolution along the width and height.Can be a single
            integer to specify the same value for all spatial dimensions.
        n: Integer, layer repeat times.
    # Returns
        Output tensor.
    """

    x = _bottleneck(inputs, filters, kernel, t, strides)

    for i in range(1, n):
        x = _bottleneck(x, filters, kernel, t, 1, True)

    return x

def inverted_residual_block(x, expand=64, squeeze=16):
    m = Conv2D(expand, (1,1), activation='relu')(x)
    m = DepthwiseConv2D((3,3), activation='relu')(m)
    m = Conv2D(squeeze, (1,1), activation='relu')(m)
    return Add()([m, x])

def inverted_linear_residual_block(x, expand=64, squeeze=16):
    m = Conv2D(expand, (1,1), activation='relu')(x)
    m = DepthwiseConv2D((3,3),  activation='relu')(m)
    m = Conv2D(squeeze, (1,1))(m)
    return Add()([m, x])

def bottleneck_block(x, expand=64, squeeze=16):
    m = tf.keras.layers.Conv2D(expand, (1,1), padding="same")(x)
    m = tf.keras.layers.BatchNormalization()(m)
    m = tf.keras.layers.Activation('relu6')(m)
    m = tf.keras.layers.DepthwiseConv2D((3,3), padding="same")(m)
    m = tf.keras.layers.BatchNormalization()(m)
    m = tf.keras.layers.Activation('relu6')(m)
    m = tf.keras.layers.Conv2D(squeeze, (1,1), padding="same")(m)
    m = tf.keras.layers.BatchNormalization()(m)
    return tf.keras.layers.Add()([m, x])
def simple_cnn(input_shape, n_output):
    inputs = tf.keras.layers.Input(input_shape)
    x = tf.keras.layers.Conv2D(64, (3,3), padding="same")(inputs)
    x = tf.keras.layers.Conv2D(32, (3,3), padding="same")(x)
    x =  tf.keras.layers.BatchNormalization()(x)
    x = tf.keras.layers.Flatten()(x)
    x = tf.keras.layers.Dense(64, activation='relu')(x)
    outputs = tf.keras.layers.Dense(n_output)(x)
    return tf.keras.Model(inputs, outputs)

# ### Best result 3 blocks CNN
def simple_cnn_3_blocks(input_shape, n_output):
    inputs = tf.keras.layers.Input(input_shape)

    x =  _conv_block(inputs, 32, (3, 3), strides=(2, 2))
    x = _inverted_residual_block(x, 16, (3, 3), t=1, strides=1, n=1)
    x = _inverted_residual_block(x, 24, (3, 3), t=6, strides=2, n=2)
    x = _inverted_residual_block(x, 32, (3, 3), t=6, strides=2, n=3)

```

```
193.     x = tf.keras.layers.Flatten()(x)
194.     x = tf.keras.layers.Dense(64, activation='relu')(x)
195.     outputs = tf.keras.layers.Dense(n_output)(x)
196.
197.     return tf.keras.Model(inputs, outputs)
198.
199.      #input_shape      =      (train_examples.shape[1],      train_examples.shape[2],
train_examples.shape[3])
200. def simple_cnn_7_blocks(input_shape, n_output):
201.     inputs = tf.keras.layers.Input(input_shape)
202.
203.     x = _conv_block(inputs, 32, (3, 3), strides=(2, 2))
204.
205.
206.     x = _inverted_residual_block(x, 16, (3, 3), t=1, strides=1, n=1)
207.     x = _inverted_residual_block(x, 24, (3, 3), t=6, strides=2, n=2)
208.     x = _inverted_residual_block(x, 32, (3, 3), t=6, strides=2, n=3)
209.     x = _inverted_residual_block(x, 64, (3, 3), t=6, strides=2, n=4)
210.     x = _inverted_residual_block(x, 96, (3, 3), t=6, strides=1, n=3)
211.     x = _inverted_residual_block(x, 160, (3, 3), t=6, strides=2, n=3)
212.     x = _inverted_residual_block(x, 320, (3, 3), t=6, strides=1, n=1)
213.
214.     x = _conv_block(x, 1280, (1, 1), strides=(1, 1))
215.     x = tf.keras.layers.GlobalAveragePooling2D()(x)
216.     x = tf.keras.layers.Reshape((1, 1, 1280))(x)
217.     x = tf.keras.layers.Conv2D(1, (1, 1), padding='same')(x)
218.     x = tf.keras.layers.Flatten()(x)
219.     x = tf.keras.layers.Dense(64, activation='relu')(x)
220.     outputs = tf.keras.layers.Dense(n_output)(x)
221.     return tf.keras.Model(inputs, outputs)
222.
223. # select the CNN model
224.
225.      input_shape      =      (train_examples.shape[1],      train_examples.shape[2],
train_examples.shape[3])
226. n_output= _labels.shape[1] - 1
227. #model = simple_cnn(input_shape, n_output)
228. model = simple_cnn_3_blocks(input_shape, n_output)
229. # model = simple_cnn_7_blocks(input_shape, n_output)
230.
231. len (model.layers) #number of layers
232. model.summary() #see layers
233.
234. # ## define loss function
235. optimizer = tf.keras.optimizers.Adam(0.01)
236.
237.                       model.compile(loss='mse',                       optimizer="adam",
metrics=[tf.keras.metrics.RootMeanSquaredError()])
238.
239. #save the loss and metric by epoch
240.
241. now = datetime.datetime.now().strftime("%Y%m%d-%H%M%S")
242. name_experiment = f"{experiment_name}-{now}"
243.
244. logdir = os.path.join("../logs1", name_experiment)
245. tensorboard_callback = tf.keras.callbacks.TensorBoard(logdir)
246.
247.                                  checkpoint_callback                                    =
tf.keras.callbacks.ModelCheckpoint("../checkpoints/custom_cnn_simple",       save_freq=500,
monitor="val_root_mean_squared_error")
248.
249.
250. # # START TRAINING
251.
252.   history   =   model.fit(train_dataset,   epochs=500,   validation_data=(test_examples,
test_labels), callbacks=[tensorboard_callback]) #
```

```
253.
254. model.save(f"../saved_model/{name_experiment}")
255.
256. #eval_data = specs[0:100]
257. #eval_label = labels[["As_ppm", "Sb_ppm"]][0:100]
258.
259. #result = model(eval_data).numpy()
260.
261. #list(zip(result, eval_label.values))
262.
263. #from sklearn.metrics import mean_squared_error
264. #mean_squared_error(eval_label, result)
265.
266. history.history
267.
268. def plot_history(history):
269.     hist = pd.DataFrame(history.history)
270.     hist['epoch'] = history.epoch
271.
272.     plt.figure()
273.     plt.xlabel('Epoch')
274.     plt.ylabel('Mean Square Error [$MPG^2$]')
275.     plt.plot(hist['epoch'], hist['root_mean_squared_error'],
276.            label='Train Error')
277.     plt.plot(hist['epoch'], hist['val_root_mean_squared_error'],
278.            label = 'Val Error')
279.
280.     plt.legend()
281.     plt.show()
282. plot_history(history)
283.
```

Mineral Deposits Studies Group 2023 *Abstract volume*

Identifying source of Sb, As and Pb anomalies at Ribeiro da Serra, Portugal, using factor analysis and features spatially associated

Carvalho, M.[1*], Resta, G.[1] Lima, A.[1,2]

[1]Department of Geosciences, Environment and Spatial Plannings, Faculty of Science of the University of Porto-Rua do Campo Alegre, s/n, 4169-007 Porto, *Portugal-morgana.carvalho@fc.up.pt
[2]Institute of Earth Sciences (ICT)-Polo do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal

Ribeiro da Serra Mine, an abandoned mining complex from the late 19th century exploited antimony and gold, belongs of a regional structure of Valongo Anticline, with several Au-As, Sb-Au, Sn-W and PbZn-Ag deposits and a long history of exploitation since at least the Roman era. Today the antimony, for instance, is considered a critical raw material for the European Union.

A soil geochemical survey was undertaken in the zone to understand the mineralisation distribution, identify possible new mineralized structures and verify soil contamination distribution. The soil analysis method used was the inductively coupled plasma mass spectrometry (ICP-MS).
Since the area had previous exploitations, there was a necessity to understand if high concentrations of target elements are due anthropogenic contamination or natural occurrences.
A factor analysis verified the relation between selected elements, Sb, As, Ti, Pb, Zn, Mn, and certain features - wastepiles, adits, veins, and streams – and to identify the source of anomalies of Sb, As and Pb. The near distance between the sampling points and the features wastepiles, adits, veins and streams were used in the analysis.

The factor analysis explains 97.79% of the variance for the selected data through three factors. Combined with the interpretation of the outputs, associating the geology and topography the method helped in the understanding the source of the elements and its distribution. Although is clear that the wastepiles have influence in some high Sb and As values, acting as source of contamination, some high values are associated with known mineralized veins and maybe undiscovered veins.

Resumos do XI CNG 2023, Coimbra, 16-20 de julho de 2023

Evaluation of pXRF as an exploration tool in soil analysis to detect antimony mineralizations: Case of study in Ribeiro da Serra and Tapada- Northern Portugal

Avaliação da pXRF como ferramenta de prospeção em análises de solo para deteção de mineralizações de antimónio: Caso de estudo em Ribeiro da Serra e Tapada- Norte de Portugal

M. Carvalho (1,2), G. Resta (1,2) A. Carvalho (1,2), R. Frutuoso (2) and A. Lima (1,2)

(1) Departamento de Geociências, Ambiente e Ordenamento do Território, Faculdade de Ciências, Universidade do Porto, rua do Campo Alegre s/n, Porto 4169-007, Portugal. Margana.carvalho@fc.up.pt

(2) Instituto de Ciências da Terra, Pólo Porto, rua do Campo Alegre 687, Porto 4169-007, Portugal.

Sumário: *Fluorescência de Raios-X portátil (pXRF) é uma ferramenta que possibilita análises elementares qualitativas e semi-quantitativas em tempo real. Neste trabalho foi avaliada a exactidão do método para a análise quantitativa de antimónio nas antigas zonas mineiras de Ribeiro da Serra e Tapada, utilizando amostras de solo. Os resultados obtidos foram favoráveis quanto à capacidade de identificação e quantificação deste elemento.*

Palavras-chave: *Fluorescência de raio X portátil, mineralizações de antimónio, análise de solos, Ribeiro da Serra, Tapada*

Key words: Portable X-Ray Fluorescence, antimony mineralization, soil analysis, Ribeiro da Serra, Tapada

Portable X-ray Fluorescence (pXRF) analysis in exploration allows to reduce costs and sample preparation steps that are associated with other analyses, besides allowing to evaluate the results in real time during field work.

In this study pXRF was tested in the encompassing zone of Ribeiro da Serra and Tapada, two abandoned mining complexes from the late 19th century were stibine was exploited. This antimony deposit is located in the western flank of the Valongo anticline, in the Dúrico-Beirão Mining District (Fig. 1).

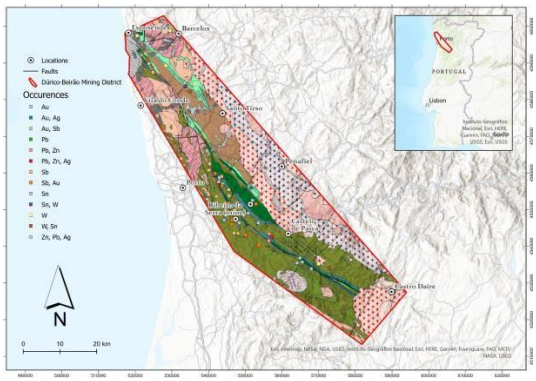

Fig. 1. The Dúrico-Beirão Mining District with its multiple deposits occurrences (Figure elaborated for Resta, G., unpublished).

A soil sampling campaign was performed using a 50x50 m grid over the target areas, with an angle of 45°. The angle of the soil lines considered the regional Douro Shear Zone, the Valongo anticlinal and the direction exploited veins rich in Sb-Au. The soil samples were collected preferentially from B or C horizons. Rainy days and wet soils were avoided. In the total, 309 samples were collected, 157 from Ribeiro da Serra and 152 from Tapada (Fig.2).

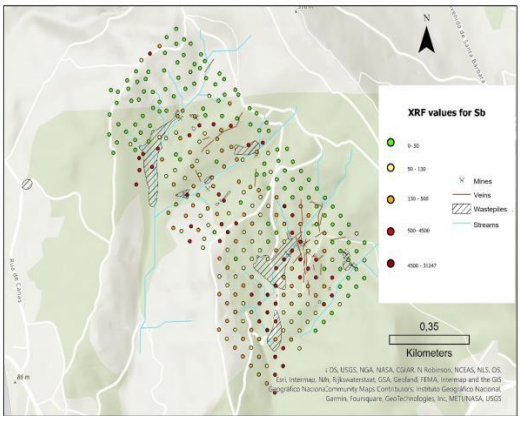

Fig. 2. Location where the soil samples were collected and the pXRF measurements for antimony (Sb).

The samples were analysed directly from the collection bags (polyethylene bags) with the equipment Bruker S1 Titan X-ray fluorescence analyser. Each sample was

analysed on three different points with a 90 second beam. To compare the results, 1/3 of the samples collected was sent to analysis by inductively coupled plasma mass spectrometry (ICP-MS).

The evaluation of the results was made considering the uncertainty associated to the pXRF analyser, calculated using the combined standard uncertainty.

Compared with the samples that were sent to the ICP-MS analysis, antimony shows a coefficient of determination (R^2) of 0,9 for the Ribeiro da Serra samples and 0,88 for the Tapada samples. Correct strong correlation between Sb and other elements also can be obtained with the pXRF data, as Factor analysis loads and PCA scores.

Considering the uncertainty associated with the pXRF measurements, it was found a very effective tool into measure since low to high values of antimony, being useful for testing the mineralization distribution, identify possible new mineralized structures and even soil contamination.

Acknowledgements: This work is supported by national funding awarded by FCT - Foundation for Science and Technology, I.P., projects UIDB/04683/2020 and UIDP/04683/2020.