N° Ordre........../FHC/UMBB/2023

PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

M'HAMED BOUGARA UNIVERSITY OF BOUMERDES

Faculty of Hydrocarbons and Chemistry

Doctoral thesis

Presented by

ELBEGUE Aref Abderrahmane

Submitted in partial fulfillment of the requirements for the degree of doctor of geophysics

Field of study: Hydrocarbon Engineering

Option: Geophysics

Contribution of artificial intelligence to the geological mapping of the SILET region (Western Hoggar) using aero-geophysical and satellite data

Titre en français **:** *Contribution de l'intelligence artificielle à la cartographie géologique de la feuille SILET (Hoggar occidental) à partir des données aérogéophysiques et satellitaires*

Board of examiners:

Prof.	HAMOUDI	Mohamed	USTHB	Chairperson
Prof.	FERAHTIA	Jalal	UMBB	Examiner
Prof.	HAMAI	Lamine	CRAAG	Examiner
Dr.	YSBAA	Sadia	UMBB	Examiner
Prof	ALLEK	Karim	UMBB	Supervisor

2022/2023

Acknowledgment

In the light of this modest work, I would first like to thank ALLAH for giving us the strength, courage, will, and patience to complete this thesis. I would like to express my thanks to all those who have contributed to the development of this work, both directly or indirectly.

My sincere gratitude and appreciation go to my thesis advisor Prof. ALLEK Karim for accepting to supervise this study, for his help, guidance, advice, and for his expertise and availability. My best thanks also go to the members of the laboratory of the Earth's Physics.

I would like to express my sincere gratitude to Mr. Groun *for his invaluable assistance in processing the raw data for this project. His expertise and dedication were instrumental in ensuring the accuracy and completeness of the results.*

I extend my heartfelt thanks to all the members of the jury for accepting to examine this work.

I also wish to deeply thank my colleagues and friends for their support and moral support, as well as all the people with whom I shared my studies, especially during these years of doctoral studies.

Finally, I thank my family, my brothers, my sisters, and especially my parents who have always been by my side during this thesis and who have always encouraged me.

Abstract

Geological mapping is a fundamental task in the study of the Earth's crust, as it provides crucial insights into the structure, composition, and evolution of the planet's surface. Traditionally, geological mapping has relied on surface observations, geological drilling, and other time-consuming and expensive techniques. However, in recent years geophysical data has emerged as a valuable tool for enhancing geological mapping, allowing for more efficient and accurate characterization of the subsurface.

This thesis explores the application of machine learning techniques to geophysical and satellite data for geological mapping. Specifically, we focus on the integration of airborne magnetic and gamma ray spectrometry data with Landsat images of the Silet region located in central Hoggar. Our goal is to improve our understanding of the geology of this region and explore the effectiveness of machine learning algorithms in this context.

Our findings show that geophysical data can provide valuable information on the subsurface structure and lithology, which can help to refine geological interpretations and reduce uncertainty in geological maps. In particular, we demonstrate the importance of integrating geophysical data with geological observations, as well as the importance of high-quality data acquisition and processing. Additionally, we show that machine learning techniques can help to automate the interpretation of geophysical data and improve the accuracy of geological maps.

In our case study, we applied a range of machine learning algorithms, including random forests (RF), Deep neural networks DNN) and extreme gradient boosting (XGBoost). We demonstrate that these algorithms can effectively classify geophysical data into different lithological units and identify subsurface structures. Specifically, we show that the machine learning tool can distinguish different rock types and identify the boundaries between different rock units based on magnetic and gamma ray spectrometry data.

Overall, this thesis provides a comprehensive overview of the contribution of machine learning applied to geophysical data for geological mapping, and highlights the potential for the utility of these data to revolutionize our understanding of the Earth's crust.

Résume

La cartographie géologique est une tâche fondamentale dans l'étude de la croûte terrestre, car elle fournit des informations cruciales sur la structure, la composition et l'évolution de la surface de la planète. Traditionnellement, la cartographie géologique s'appuie sur des observations de surface, des forages géologiques et d'autres techniques coûteuses et chronophages. Cependant, ces dernières années, les données géophysiques ont émergé comme un outil précieux pour améliorer la cartographie géologique, permettant une caractérisation plus efficace et plus précise du sous-sol.

Cette thèse explore l'application des techniques d'apprentissage automatique aux données géophysiques et satellites pour la cartographie géologique. Plus précisément, nous nous concentrons sur l'intégration des données de magnétiques aéroportées et de spectrométrie gamma avec des images Landsat de la région de Silet située dans le Hoggar central. Notre objectif est d'améliorer notre compréhension de la géologie de cette région et d'explorer l'efficacité des algorithmes d'apprentissage automatique dans ce contexte.

Nos résultats montrent que les données géophysiques peuvent fournir des informations précieuses sur la structure et la lithologie du sous-sol, ce qui peut aider à affiner les interprétations géologiques et à réduire l'incertitude dans les cartes géologiques. En particulier, nous démontrons l'importance de l'intégration des données géophysiques avec les observations géologiques, ainsi que l'importance de l'acquisition et du traitement de données de haute qualité. De plus, nous montrons que les techniques d'apprentissage automatique peuvent aider à automatiser l'interprétation des données géophysiques et à améliorer la précision des cartes géologiques.

Dans notre étude de cas, nous avons appliqué une gamme d'algorithmes d'apprentissage automatique, notamment les forêts aléatoires (RF), les réseaux de neurones profonds (DNN) et le renforcement extrême du gradient (XGBoost). Nous démontrons que ces algorithmes peuvent efficacement classifier les données géophysiques en différentes unités lithologiques et identifier les structures du sous-sol. En particulier, nous montrons que l'outil d'apprentissage automatique peut distinguer différents types de roches et identifier les limites entre différentes unités géologiques en se basant sur les données magnétiques et de spectrométrie gamma.

Dans l'ensemble, cette thèse fournit un aperçu complet de la contribution de l'apprentissage automatique appliqué aux données géophysiques pour la cartographie géologique et met en évidence le potentiel de ces données qui révolutionne notre compréhension de la croûte terrestre.

الملخص

تعتبر عملية رسم الخرائط الجيولوجية مهمة أساسية في دراسة قشرة الأرض، حيث توفر رؤى حاسمة حول البنية والتركيب وتطور سطح الكوكب. وعادة ما تعتمد رسم الخرائط الجيولوجية على الملاحظات السطحية والحفريات الجيولوجية وتقنيات أخرى مكلفة وتستغرق وقتًا طويلاً. ومع ذلك، في السنوات الأخيرة، ظهرت البيانات الجيوفيزيائية كأداة قيمة لتحسين رسم الخرائط الجيولوجية، مما يسمح بتوصيف الطبقات الأرضية الأسفل بطريقة أكثر كفاءة ودقة.

تستكشف هذه الأطروحة تطبيق تقنيات التعلم الآلي على البيانات الجيوفيزيائية والأقمار الصناعية لرسم الخرائط الجيولوجية. وبالتحديد، نركز على دمج بيانات المغناطيسية وطيف الأشعة الغاما التي تجرى من الجو مع صور لاندسات لمنطقة سيليت الموجودة في وسط هوجار. هدفنا هو تحسين فهمنا للجيولوجيا في هذه المنطقة واستكشاف فعالية خوارزميات التعلم الآلي في هذا السياق.

تظهر نتائجنا أن بيانات الجيوفيزيائية يمكن أن توفر معلومات قيمة عن هيكل التربة الداخلية والتصنيف الصخري، والتي يمكن أن تساعد في تحسين التفسير الجيولوجي وتقليل عدم اليقين في الخرائط الجيولوجية. بالخصوص، نبرز أهمية دمج بيانات الجيوفيزيائية مع الملاحظات الجيولوجية، بالإضافة إلى أهمية اقتناء ومعالجة البيانات عالية الجودة. كما نبين أن تقنيات التعلم الآلي يمكن أن تساعد على تلقين بيانات الجيوفيزيائية وتحسين دقة الخرائط الجيولوجية.

في دراستنا، قمنا بتطبيق مجموعة من خوارزميات التعلم الآلي، بما في ذلك الغابات العشوائية (RF) والشبكات العصبية العميقة (DNN)و .extreme gradient boosting (XGBoost) نبرهن أن هذه الخوارزميات يمكن أن تصنف بيانات الجيوفيزيائية بفعالية إلى وحدات صخرية مختلفة وتحديد الهياكل الداخلية. ونبين على وجه الخصوص أن أداة التعلم الآلي يمكنها تمييز أنواع مختلفة من الصخور وتحديد الحدود بين وحدات الصخور المختلفة على أساس بيانات المغناطيسية وطيف الأشعة الغاما.

بشكل عام، تقدم هذه الرسالة نظرة شاملة على دور تقنيات التعلم الآلي المطبقة على بيانات الجيوفيزياء لرسم الخرائط الجيولوجية، وتبرز إمكانية الاستفادة من هذه البيانات لفهم مكونات القشرة الارضية.

Table of contents

List of figures

List of tables

Chapter 1 : General introduction

1.1. Introduction

Geological mapping is a term designed to describe all procedures by which the geologist/ geoscientist aims to prepare a special map that keeps a record of the distribution of the outcropped rocks belonging to different formations. Depending on the desired scale and the usage of the map, different methods are adopted to plot the geological phenomena on a 2D map, but in general, a geological map contains information about the historical evolution of bedrock as well as surficial formations, and to every formation, a distinct colour is used to set the various formations apart. Besides that, it also contains information about the linear features (e.g., faults, folds, and bedding) that can occur in nature. Since the geological maps are usually presented as 2D maps, special symbology (e.g., strike, dip and plunges) are developed to take into account not only the surficial extension of the geological formations but also their underground extension. Furthermore, besides identifying the lithological units, geological mapping also focuses on studying the disposition of the geological layers, their formation and their age in relation to each other. The geological discipline which is concerned with studying the geological layers is called “stratigraphy.” This discipline allows the identification of the stratigraphic units that consist of a group of lithologies with known disposition and age. These units accompany the geological map in the form of a stratigraphic column, which allows for the geological map to give information about the lithology types as well as the aspect of the chronological dating of the lithological units.

Traditionally, mapping an area of interest consisted only of rigorous fieldwork that starts from the reconnaissance expeditions to identifying and dating the observed lithologies and ends by assigning a distinct colour to every lithology. These mapping procedures are usually long and can span over a long time, and the final product is not unique since the geological interpretations (i.e., identifying the geological formations) are subjugated to the experience of the geologist (i.e., subjective interpretations).; however, nowadays, the integration of the geographic information systems (GIS) had a great advantage in assisting the traditional fieldwork. This system allows the incorporation of different data types, which are known to have an innate link with geology, at different scales throughout the mapping expedition. This facilitates tracking the limits of the geological formations more accurately, especially in hard-to-explore terrains. Usually, in traditional fieldwork, these geological limits are inferred based on the experience of the geologist, leading to a potentially biased interpretation, whereas a GIS environment offers a visual interactive window to draw the limits (semi-assisted interpretation), or more advanced technologies such as machine learning to learn and infer the limits (assisted interpretation).

Aero-geophysical data are useful in a variety of Earth science applications such as natural resources exploration and geological mapping. Applying Machine Learning Algorithms (MLA) to these data types, which are widely used in image analysis and statistical pattern recognition, can enhance preliminary geological mapping and interpretations.

1.2. Research scope and objectives

In this project, we would like to determine and expose the contribution and the limitation offered by an ancient aero-geophysical survey to the geological mapping in the Silet region (western/central Hoggar). Before using it for geological interpretation, the recorded data (Magnetic and spectrometry gamma data) must be prepared to maximize their coherence with geology. After that, various statistical and pattern recognition methods are used in consort with these data to extract meaningful and objective geological information. The resulting maps, called predictive maps, are confronted with the published geology map of Silet to assess the limitations and the contributions of the airborne geophysical data as well as the proposed method in improving the accuracy of the geological limits within the study area.

1.3. Organization of the dissertation

Following a brief introduction, the first chapter of this dissertation gives a theoretical overview of airborne geophysical surveys as well as some of the most used filters to enhance geophysical features and increase their agreement with geology.

In the second chapter, we describe the technical characteristics of the airborne surveys of Algerian territory as well as the processing chain that was applied before to data. After that, we go through our proposed processing chain that we have implemented to clean airborne geophysical data from the remaining noise and thus make the data suitable for modelling.

The third chapter starts by describing the general geology of the Hoggar, and then, it gives the geology description of the study area Silet-Tamenrrast. We conclude this chapter by reviewing the state-of-the-art applications of geological mapping, and also highlighting some shortfalls of the past works. These shortfalls are inspected in more detail in the following chapter.

The fourth chapter is split into two parts. The first part is concerned with discussing the theory behind machine learning algorithms with the main focus on supervised learning. By the end of this part, we demonstrate, through the use of a detailed practical example of a small learning problem, the implementation of a real supervised problem, including training and evaluating a machine learning classifier. In the second part of this chapter, we implement the extreme gradient boosting algorithm by applying it to the geological mapping of the Silet region. The results of the algorithm, which is used for the first time in a geological mapping context, are carefully discussed in the present chapter.

Chapter 2 : Airborne geophysical surveys and theoretical aspects of data filtering

2.1. Introduction

The term airborne geophysical survey is used to describe any survey that is conducted using fixed-wing aircraft or helicopters or even drones which are gradually being introduced in recent years. During these surveys, and depending on which method is used, the aircraft (helicopter or drone) carry a special measurement tool that can keep records of the physical properties of the rocks at the ground level or several kilometers of depth. These records can form the basis of geological mapping or natural resources exploration.

Some of the geophysical properties, which are measured during airborne surveys, are the earth's magnetic field, the naturally occurring gamma rays, and gravity measurements. The interpretation of the measured data gives a detailed first-pass map that provides an overview of the surveyed area. Usually, the data of these surveys are the target of several transformations and filters to enhance their quality, allowing the delivery of complete and thorough information about the study area.

In the present chapter, we are concerned with two airborne survey methods: the magnetic surveys and the gamma-ray spectrometric surveys. In the subsequent sections, we show the functioning theories behind these methods and also the parameters that are used to plan and control these surveys. In addition, we conclude this chapter by presenting the theory behind the most common filters that are applied to the data of these surveys

2.2. Airborne magnetic survey

2.2.1. History of the magnetic data usage

The first use of the geomagnetic field has a long history that extends back to the 12th century when it was first exploited by the Chinese for navigation purposes. Despite the early usages of the geomagnetic field, its characteristics were not explored until the 17th century. In 1600, "William Gilbert" published his work "de magneto", suggesting that the Earth is, in fact, a gigantic magnet. However, the origin of the Earth's field has remained inexplicable for another 300 years after Gilbert's proposal. It was also known early on that the field was not constant in time. Therefore, the secular variation has been well recorded, so that a very useful historical record of the variations in strength would be available, especially for scientific studies.

In 1838, "Carl Friedrich Gauss" used spherical harmonics and showed that the coefficients of the field expansion, which he determined by fitting the surface harmonics to the available magnetic data at that time, were almost identical to the coefficients for a field due to a magnetized sphere or to a dipole. He also demonstrated that the best-fitted field is if the dipole forms an angle of approximately 11° with Earth's rotation axis.

In the 1940s, a major leap in the understanding of the origin of the field came with the emergence of the "Dynamo theory". The theory that was proposed by "Walter Maurice Elsasser" and "Edward

Bullard.” explains the magnetic field as a self-exciting dynamo. In this dynamo mechanism, the motion of a conducting material (liquid iron) in the Earth’s outer core, which is driven by the convection and the rotation motion of the Earth, across a magnetic field generates an electric current. This current in turn generates a magnetic field that also interacts with the fluid motion to finally create a secondary magnetic field. Together, the two fields form the Earth’s magnetic field. This theory forms the basis for our current understanding of the origin of the magnetic field.

2.2.2. Principals of airborne magnetic survey

2.2.3. Earth’s magnetic field

The magnetic field has an internal and external source. Both sources exhibit a time dependence. Spherical harmonics can be used to account for both components. In a spherical coordinate system, it can be written as:

$$\begin{cases} a\sum_{n=1}^{\infty}\sum_{m=0}^{n}\left(\frac{a}{r}\right)^{n}\left[q_n^m(t)\cos m\varphi + S_n^m(t)\sin m\varphi\right] P_n^m(\cos\theta) & \text{External sources} \\ a\sum_{n=1}^{\infty}\sum_{m=0}^{n}\left(\frac{a}{r}\right)^{n+1}\left[g_n^m(t)\cos m\varphi + h_n^m(t)\sin m\varphi\right] P_n^m(\cos\theta) & \text{Internal sources} \end{cases} \qquad (2\text{-}1)$$

a is the mean radius of the earth. It is estimated to be 6371.2 Km; r is the distance of the measuring point from the center of the earth; θ is the geocentric colatitude; φ is the longitude; Pnm is the associated Legendre polynomial with a degree n and order m. gnm, hnm are the spherical harmonic coefficient. They are measured in nanotesla (nT). We calculate the total magnetic potential by summing the two equations. (Barraclough, 1987).

The International Geomagnetic Reference Field (IGRF) is calculated using the internal sources components in equation (2-1). It is an international mathematical model of the Earth’s magnetic field of internal origin (Macmillan & Finlay, 2011), and it is issued by the International Association of Geomagnetism and Aeronomy (IAGA). Every five years, this model undergoes a recalculation process until a definitive model is developed for the next 5 years. This definitive model is called the Definitive Geomagnetic Reference Field (DGRF).

The IGRF takes into account the internal components of the geomagnetic field that originated in the outer core. The rapid field fluctuations caused by the electric current systems in the magnetosphere and ionosphere and crustal fields are not included in the IGRF (Macmillan & Finlay, 2011)

In December 2019, the (IAGA) Division V Working Group (V-MOD) adopted the thirteenth generation of the IGRF (Alken et al. 2021).

2.2.3.1. Internal field

The Earth’s magnetic field forms a protective shield around the planet, called the magnetosphere. It is mainly produced within its interior, and it can be defined as the superposition of two components. They are “The core field” and “crustal field.”

- The core field is generated by the dynamo mechanism in the Earth’s outer core.
- The crustal field is generated by the magnetized rocks on the Earth’s crust.

2.2.3.2. External field

External fields originate in Earth’s magnetosphere and ionosphere. They are produced because of the constant interaction of the magnetosphere with the solar winds. The intensity of these fields is much weaker than that of the internal field, and unlike the internal field, the change of intensity for the external sources is rapid and unstable.

2.2.3.3. The components of the Earth’s magnetic field

The geomagnetic field is a vector field that is characterized by a direction and magnitude F at every point in space (Gunnarsdóttir, 2012). At a given point P in space, we can describe the geomagnetic field as:

$$\mathrm{F} = \sqrt{\mathrm{X}^2 + \mathrm{Y}^2 + \mathrm{Z}^2} \tag{2-2}$$

- F: magnitude
- X, Y, Z: North, east and downward components of the field vector

The x-axis is directed towards the geographic north, the y-axis is directed to the east and the z-axis is vertical and positive downwards into the Earth.

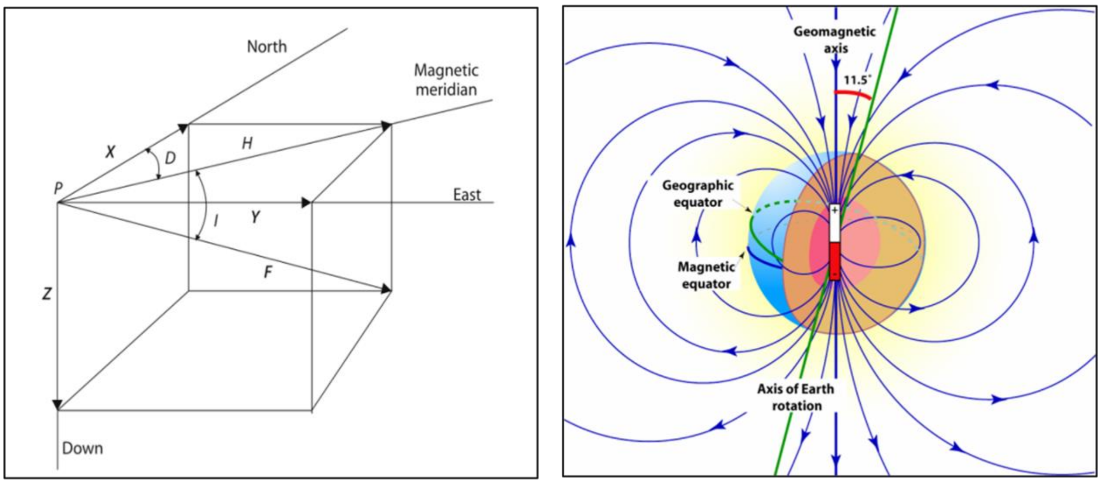


Figure 2-1 Earth magnetic field components; a) F is the total intensity of the field, H is the horizontal component and I and D are the inclination and declination respectively; b) Approximate position of the geomagnetic axis and the axis of Earth rotation. (Laurent Marescot, 2017)

Other characteristics of the magnetic vector field are Inclination (I) and Declination (D). The inclination is the angle between the vector field and its horizontal component. The declination is the angle between the x-axis and the component H. They are described as follows:

$$\begin{aligned} \mathrm{I} &= \arctan\frac{\mathrm{Z}}{\mathrm{H}} \\ \mathrm{D} &= \arctan\frac{\mathrm{Y}}{\mathrm{X}} \end{aligned} \tag{2-3}$$

The components X, Y, and Z can be written in terms of I and D as follows:

$$\mathrm{X} = \mathrm{H}\cos\mathrm{D}, \mathrm{Y} = \mathrm{H}\sin\mathrm{D}, \mathrm{Z} = \mathrm{F}\sin\mathrm{I}$$

The total intensity of the magnetic field varies on the surface of the Earth. It has a maximum value of 62000 nT around the poles, and it decreases the more we get closer to the equator. It has an intensity of 23000 nT around the equator.

2.2.3.4. Time variation

Geomagnetic field measurement and historically recorded data of the earth’s magnetic field show that the main magnetic field of the Earth changes with time. These changes include intensity changes as well as inclination and declination changes.

The rate of change can range from seconds to millions of years, and they can be periodic or completely random. The total field strength can range from a few nT to thousands of nT. According to the duration of these changes, they can be divided into two categories: long-term changes and short-term changes.

Long-term changes, also called “Secular variations,” are related to the internal field (i.e., the core field). Its duration is in the range of 5 years or longer. Short-term changes are related to the external field (current in the magnetosphere and ionosphere). They can be in the range of a few seconds or longer, but their duration rarely exceeds one year.

Long term changes

The existence of long-term changes was discovered in 1634 when the geomagnetism measurements showed that the declination of the earth’s magnetic field is not only a function of position but also a function of time. These changes are quite small, but still easily observable when looking at geomagnetic data that spans several years. Both the long-term changes and the main field come from the same source inside the earth. Therefore, long-term changes provide important information about the dynamics of the conductive fluid in the earth's outer core and the earth's magnetic field itself (Gunnarsdóttir, 2012)

Short term changes

Short-term changes are divided into regular changes and various irregular changes caused by the orbital motion and/or rotation of the Earth. The short-term changes in the geomagnetic field are related to changes in the external field, of which the sun is the most important factor (Love, 2008).

2.2.4. Airborne magnetic survey

The purpose of the magnetic survey is to study subsurface geology. This is carried out by studying Earth's magnetic field anomalies which are caused by the magnetic properties of the rocks below. Generally speaking, the magnetic content (susceptibility) of rocks depends on the type of rock and the environment in which it is located. Common causes of magnetic anomalies include dykes, faults, and lava flows. The anomalies of the Earth's magnetic field are caused by induced or remanent magnetism; the first one is induced by a magnetic field produced by the magnetization proprieties of ferrous bodies when they are placed under the effect of the geomagnetic field, and the second one is the residual magnetization in the ferrous bodies left by the orientation of the Earth's magnetic field at the time of body formation.

Airborne magnetic surveys do not include information about the direction of the anomaly field. However, if the intensity of the geomagnetic field is greater than the anomaly field, which is typically the case, the scaler value (Fobserved-FIGRF) is approximately equal in magnitude to the anomaly field ΔF projected in the direction of the geomagnetic field FIGRF (Figure 2-2). Therefore, the magnetic anomalies are recorded in the direction of the geomagnetic field (Reeves, 2006).

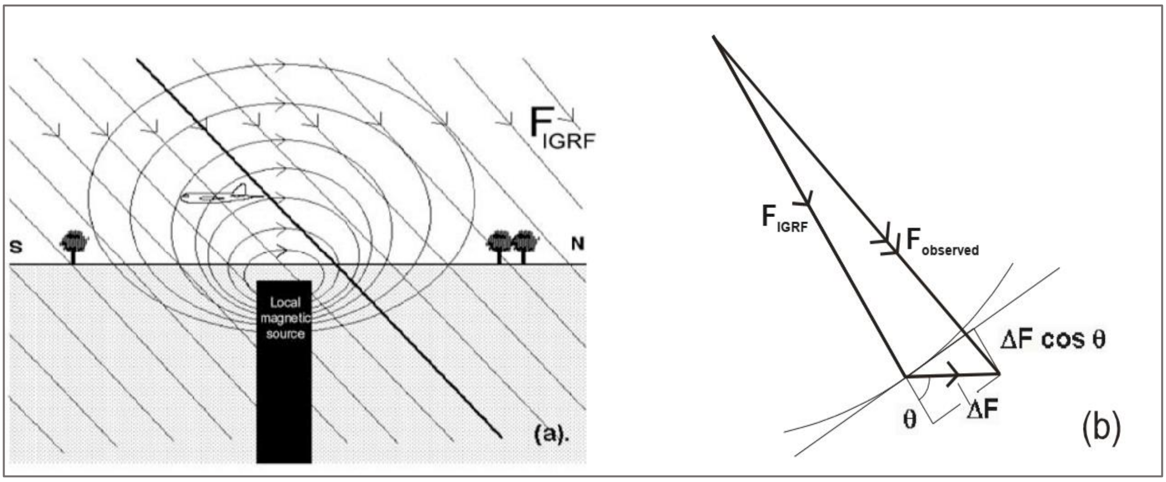


Figure 2-2 (a) the measured magnetic field is the vector sum of geomagnetic field and the anomaly field. (b) vector representation of the measured magnetic field (Reeves 2006)

2.2.5. Airborne magnetometry

The first use of magnetometers in aircraft dates back to 1936 when Soviet scientists developed a prototype for measuring the magnetic field. In 1943, during World War II, US Navy researchers designed a more sensitive magnetometer to detect submarines. These magnetometers only measured the intensity of the total geomagnetic field. The rapid development of magnetometers and the experience gained from using them in the war set airborne magnetometry into motion in the exploration industry in the late 40s (Reford & Sumner, 1964). Some of the widely used magnetometers are optical absorption magnetometers, proton precession magnetometers and flux-gate magnetometers.

2.2.6. Magnetometer installation

Aeromagnetic systems are installed on a variety of fixed-wing aircraft or helicopters. Usually, they are installed in a boom (stinger) attached rigidly to the airframe of the aircraft or helicopter. In some instances, however, a system of a towed-bird aerofoil can be used with a light helicopter. Both systems are hung below the aircraft to minimize the noise caused by the aircraft and the recording instruments. (Groune, 2019)

2.3. Airborne gamma-ray spectrometry survey

2.3.1. Introduction

Gamma-ray spectrometry is a geophysical method based on measuring naturally occurring gamma rays. At first, this method had the sole purpose of searching for uranium ores. In the present days, it is widely used in a multitude of fields such as uranium exploration, assessing the health risks caused by nuclear incidents, mapping ore deposits of valuable minerals, hydrocarbon exploration and geological mapping of the soil surface.

Gamma rays are produced from the atomic nuclear decaying of the radioactive isotopes. These radiations are electromagnetic radiation, and they are characterized by the shortest wavelength (highest energy) in the electromagnetic spectrum.

Historically, the first portable detectors were developed in the first decade of the twentieth century. Rigorous uranium exploration and the increase of measurement sensitivity due to the development of the scintillation detector during the 1940s led to the first airborne radiometric surveys in the USA, Canada and the former USSR in 1947 and Australia in 1951 (Seligman, 1992).

2.3.2. Types of radioactivity decay

Atoms are composed of an interior part, called the nucleus, surrounded by a group of negatively charged electrons. The nucleus is the whereabouts of uncharged neutrons and positively charged protons. Both particles are called nucleons. A nuclide is a term used to describe nucleons and their associated electrons. (Gilmore, 2008)

A radionuclide is a term used to define nuclides that have a surplus of energy. These radionuclides tend to be unstable which leads them to spontaneously decay to form more stable isotopes. The decaying process comes with the emission of certain particles or certain forms of electromagnetic energy, which is referred to as nuclear radiation. There are three kinds of natural radiation designated by the Greek letters α, β, and γ.

2.3.2.1. Alpha decay (α)

Alpha decay is a characteristic of a heavy nucleus with an atomic number Z>83. In this decay, the nucleus shoots out a large mass and positive charge particles in the form of a helium nucleus ${}_{2}^{4}He$. Due to the large mass of alpha particles, this type of radiation can only travel short distances in the order of 10-2 m in the air and 10-5 m in rocks.

2.3.2.2. Beta-decay (β)

Beta-decay occurs in a nucleus that is neutron-rich or neutron deficient. This decay is accompanied by the emission of either a positive beta particle or a negative beta particle. A beta particle is an electron and the positive beta particle is called a “positron or antielectron” (electron with positive electric charge,). In both processes, the number of nucleons “A” remains the same.

β- decay: this decay occurs in neutron-rich nuclei. Within the nucleus, the neutron transforms into a proton with the emission of an electron β- as well as an electron antineutrino. The energy of the decay is shared between the electron and the antineutrino.(Gilmore, 2008)

β+ decay: contrary to the β- decay, a proton is transformed into a neutron. This transformation is accompanied by the emission of a positron particle β+ as well as a neutrino particle. This decay is characteristic of neutron-deficient nuclei. Both β+ and β- decays have similar proprieties.

The particle issued from the β decay has more penetration power than the particles issued from alpha decay (about 100 times more than α rays). They can travel distances up to 8 m in the air and 1 cm in the water.(Groune, 2019)

Remarque: α and β decay also result in the emission of gamma radiations.

2.3.2.3. Gamma-ray (γ)

Gamma-ray is electromagnetic energy emitted by the nucleus of some radionuclides. These radionuclides tend to be “excited.” To return to a more stable state, these radionuclides (i.e., the ground state) emit radiations in the form of “Photons.”

Gamma-ray is characterized by a wavelength λ and a frequency υ. These characteristics are related according to the following formulas:

$$\begin{cases} \mathrm{E} = \mathrm{h}.\upsilon \\ \upsilon = {}^{\mathrm{c}}/_{\lambda} \end{cases} \tag{2-4}$$

Where: h (Plank constant) = 4.135×10^{-15} eV Hz^{-1};
c (Velocity of light) = 2.997926×10^{8} m s^{-1}.

Gamma rays have a lot of penetrating power that several inches of a dense material like lead, or even a few feet of concrete may be required to stop them. They can completely penetrate the human body.

2.3.3. Interaction of gamma rays with materials

When the incident gamma rays collide with materials, three phenomena can occur. The “Photoelectric absorption,” “Compton scattering” and “Pair production.”

Photoelectric absorption: this phenomenon happens when an incident gamma-ray collides with tightly bonded electrons. These electrons are usually located in the inner shell (i.e., K shell) of the atom. In the collision process, the incident gamma rays lose all their energy to the electron of the material. This causes the liberation of the electron as well as the disappearance of the photons. Part of the initial energy of the photons is employed to unbind the electron from its orbit, and the majority of the energy is converted to the liberated electron as kinetic energy. The photoelectric effect only occurs when the initial energy of the photons exceeds the binding energy of the electron. (Gilmore, 2008)

Photoelectric absorption is more important in heavier atoms such as uranium and lead. It is also the predominant effect of low-energy gamma rays.(G. Nelson, 1991)

Compton scattering: this effect takes place when the incident gamma-ray interacts with a free or loosely bonded electron which is usually located in the outer shell of the atom. The incident gamma-ray energy transfers part of its energy to the electron, which causes the liberation of the electron. Oppositely to the photoelectric absorption, this interaction produces two particles: a liberated electron, and a scattered gamma-ray. The direction of both particles depends on the amount of the exchanged energy between the two particles.

Pair production: for this interaction to happen, the energy of the incident photons must be at least equal to two times the rest energy of an electron, which equals 1022 keV. When these photons enter the electromagnetic field that is created by the nucleus of an atom, it splits into two particles: an electron and a positron. After losing its kinetic energy, the positron disappears by colliding with another electron. This phenomenon is referred to as “Annihilation.”

The annihilation releases two annihilation photons, and both photons have an energy of 511 keV(G. Nelson, 1991)

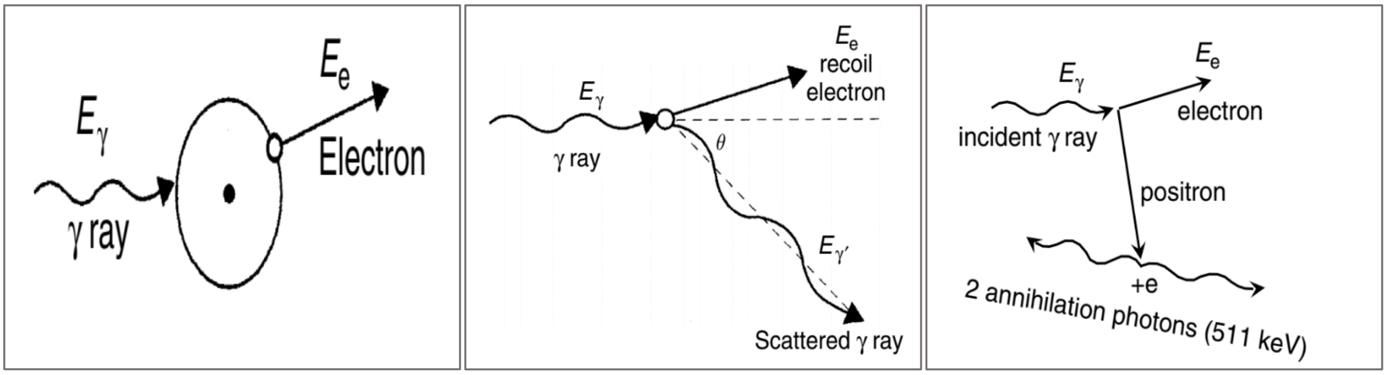


Figure 2-3 Different interactions of gamma rays with materials. From left to right, the photoelectric absorption; the scattering Compton phenomena; the pair production and the annihilation of the positron. (Gilmore 2008)

2.3.4. The decay equation

The radionuclides tend to be unstable. Therefore, they decay by emitting various types of radiation. The decaying process is governed by statistical law. It can be represented as follows:

$$\mathrm{dN} = \lambda \mathrm{N}.\mathrm{dt} \tag{2-5}$$

Where: N = the number of radioactive nuclei;
λ = it is the decay constant. It represents the probability of decay per nucleus per unit of time.

We can also define radioactivity or the decay rate as the representation of the number of disintegrations per unit of time. It is written by the following equation:

$$A = -{}^{dN}/_{dt} = \lambda N \qquad (2\text{-}6)$$

After integrating equation (2-5), we can obtain the more commonly used decay equation. It is written as follows:

$$N_t = N_0 e^{-\lambda t} \qquad (2\text{-}7)$$

Where: Nt = the number of radionuclides at time t;

N0 = the number of radionuclides at time t=0

Besides the decay constant λ, another parameter can be defined. It is the "half-time T1/2." Given an initial number N of a radionuclide, the half-time represents the necessary time for one-half of the initial quantity to disintegrate. It is written as follows:

$$T_{1/2} = \frac{\ln 2}{\lambda} \qquad (2\text{-}8)$$

Both λ and T1/2 are element-specific parameters.

2.3.5. Radioactive decay series

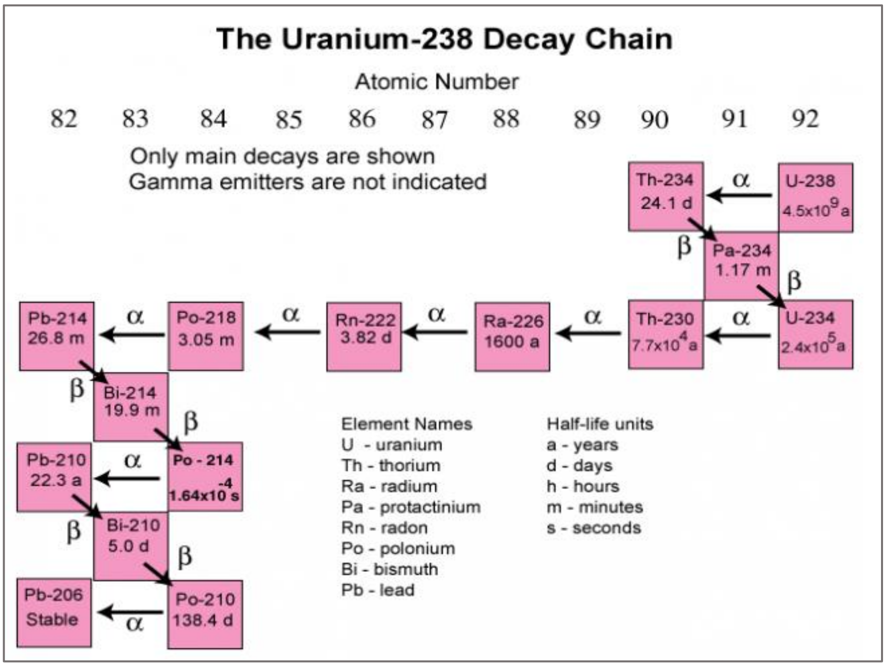


Figure 2-4 An example of a series radionuclide of the Uranium-238. (From https://pubs.usgs.gov/of/2004/1050/uranium.htm)

Radionuclides are in an unstable state that causes them to decay to produce more stable isotopes. Generally, most radionuclides need a single decay before reaching a stable state. However, this is not the case for some heavier radionuclides such as U238 and Th232. These radionuclides are called “series radionuclides.”

The series of radionuclides decay multiple times before reaching the stable form in view of the fact that the by-products (i.e., daughter atoms) of the first decay are also radioactive. Hence, they also tend to decay. The series of transformations and decays required for a radionuclide to reach a stable form are called the “decay chain.” In the decay chains, only the final products of a decay series are stable, and for some decays, the nucleus of the parent radionuclides transforms into the nucleus of a different chemical element.

Two factors control the number of daughter nuclei in a decay chain. The number of disintegrated nuclei of the parent nucleus, which can be determined by equation (2-7), and the number of the newly produced nuclei of the daughter nucleus, which are produced by the disintegration of the parent nuclei. Both quantities are controlled by the half-time of both nuclei.

Depending on the half-time of the parent nucleus, we can define different equilibrium stats. They are:

Secular equilibrium: this equilibrium is typical for the Thorium series and the Uranium series where the half-time of the parent nucleus is much longer than the half-life of the daughter nucleus. After n half-times of the daughter are passed, the radioactivity of both the parent and the daughter nuclei can be determined by: $\lambda_1 N_1 = \lambda_2 N_2 = \ldots \lambda_n N_n$ (2-9)

Transient equilibrium: This equilibrium occurs when the half-time of the parent nucleus is slightly longer than the daughter nucleus. When the equilibrium state is achieved, the radioactivity of the daughter nuclei is greater than the radioactivity of the parent nuclei. The radioactivity can be determined by:

$$A_2 = \frac{\lambda_2}{\lambda_2 - \lambda_1} A_1 \qquad (2\text{-}10)$$

No equilibrium: this case happens when the half-time of the parent nucleus is shorter than the half-time of the daughter nucleus.

2.3.6. Airborne gamma-ray spectrometry survey

2.3.6.1. Introduction

The purpose of the airborne gamma-ray spectrometry survey (AGRS) is to measure the abundance of naturally occurring elements. These are the “Potassium,” the “Thorium” and the “Uranium.” Besides these elements, many radioactive elements emit spontaneous radioactivity; however, the above-mentioned elements are the only elements that exhibit radioactivity with enough intensity that can be measured in AGRS surveys. (Seligman, 1992)

The potassium family has three naturally occurring isotopes. They are the K41, K40, and K39. Two of them, namely the K41 and K39 are stable and account for 93.25 and 6.73 percent of the earth's total potassium concentration respectively. The K40 is the only radioactive isotope of potassium and accounts for only 0.012 percent of the potassium concentration.

The K40 nucleus can decay, depending on the decay type, into Calcium or Argon nucleus. 89 percent of the time it decays into calcium by emitting a β- particle, and the remaining 11 percent, the K40 decays into Argon by absorbing a β- particle. This decay is accompanied by the emission of gamma rays at 1.46 MeV, which is the gamma-ray peak that is employed to detect the K40 in the AGRS.

Uranium is a trace element that exists in nature. It has various radioactive isotopes, of which the most abundant ones are U238 and U235. U238 accounts for 99.25 percent and U235 accounts for 0.75 percent of the uranium concentration in nature. During AGRS surveys, the detection of 1.765 MeV gamma-ray, which is a characteristic gamma-ray of one of the daughters of the uranium decay series (i.e., Bismuth Bi214), is used to estimate the uranium concentration.

Thorium exists in nature as one isotope. It is the Th232. Similar to the uranium, the estimation of the thorium concentration is carried out indirectly by detecting gamma rays at 2.615 MeV of the Thallium Tl208, one of the daughter nuclei in the thorium decay series (IAEA, 2003)

2.3.6.2. Gamma-ray estimation in surveys

In AGRS surveys, a spectrometer with a 256 (or 512) channel is commonly used. These channels keep a record of all radiation that spans over 0-3.0 MeV, which is the range of natural radiation.

Table 2-1 Standard windows for natural radionuclides.

Window name	*Isotope used*	*Gamma-ray energy (MeV)*	*Energy window (MeV)*
Potassium	K40	1.460	1.37 − 1.57
Uranium	Bi214	1.760	1.66 − 1.86
Thorium	Tl208	2.615	2.41 − 2.81
Total count	-	-	0.41 − 2.81
Cosmic	-	-	3.0 − ∞

Monitoring the counts of four spectral windows is commonly the standard method to acquire spectrometric data. These windows are centered around the value energy of gamma rays emitted from the disintegration of potassium as well as gamma-ray emitted from the decay series of uranium and thorium. The boundaries of the windows are issued by the International Atomic Energy Agency (IAEA). These windows are presented in Table 2-1.

The total count window and the cosmic window monitor the overall radioactivity and all radiations due to incidents that exceed 3 MeV respectively (Grasty & Minty, 1995).

2.3.6.3. Instrumentations employed in AGRS surveys

Detectors: materials called “Scintillator” is the main constituent of a detector in AGRS. These scintillators are made of special substances that tend to emit visible and invisible radiations when struck by electromagnetic radiations (e.g., gamma rays). This phenomenon is called “fluorescence.” The scintillator system is connected to a photomultiplier tube (PMT). This system absorbs the photons produced from the fluorescence of the scintillator and converts them into an electric signal via the photoelectric effect. The electric signal has an amplitude proportional to the intensity of the radiations that interacted with the scintillator (Seligman, 1992).

The most commonly used scintillator in AGRS is a crystal of sodium iodide activated with thallium NaI (TI). A system of at least two detector packages with a volume of 32.8 L where each package is composed of four 10.2 cm × 10.2 cm × 40.6 cm NaI (TI) crystals and each crystal is connected to its PMT system. These crystals are usually placed in a thermally insulated container(IAEA, 2003).

The detectors are calibrated to achieve better resolution. The resolution is expressed as the ratio of the full width of a photopeak divided by half-maximum amplitude (FWHM), and it is usually expressed as a percentage of the amplitude of the peak.

The 0.622 MeV photopeak, the characteristic gamma-ray energy of Cesium C137, and the 2.61 MeV photopeak, the characteristic gamma-ray energy of Thallium Tl, are used to calculate the resolution of the detectors. Most modern detectors can achieve a 10 percent and 7 percent resolution for Cesium and Thallium peaks respectively (Grasty & Minty, 1995).

Ancillary instruments: additional instruments include:

- A GPS navigation system. It is used to locate the exact position of the measurement point.
- An altimeter radar. It is used to measure the height of the aircraft.
- A barometer and thermometer. They are used to record the temperature and the pressure.
- Other geophysical instruments such as magnetometers can also be carried.

The height of the aircraft along with the temperature and pressure records are employed in the processing phase of the spectrometric data. They are used to eliminate the attenuation effect of the air between the ground surface and the detector.

The volume of the detector

This parameter is linked with airborne gamma-ray spectrometry surveys. It is decided according to the capacity of the aircraft. As a rule of thumb, a detector of 33 L to 50 L is used for a fixed-wing aircraft, and a smaller detector of 17 L to 33 L is chosen for helicopter surveys.

2.4. Airborne survey parameters

Careful attention to planning is essential to the final result and success of the airborne surveys. Attaining the required quality needs attention to some parameters of the survey, such as survey altitude, line spacing, line direction, and sampling interval. These parameters must be chosen so that no vital anomalies are missed, and the specified geological resolution is achieved(Isles & Rankin, 2013).

Survey outline

This parameter is important, especially in large-scale surveys. For this type of survey, a sketch map contains an outline of the boundaries for all surveys that were carried out in the area. The purpose of the sketch map is to facilitate the linking of several adjacent surveys at the data processing phase. The flight lines of adjacent surveys are usually extended beyond the boundary of the survey to allow a confident interpolation in the overlapped areas (Reeves, 2006).

Survey altitude (Ground clearance)

Common to all physical measures, the closer we are to the target the clearer it will appear on the map. Therefore, in airborne surveys, the lower the survey altitude the higher the obtained geological resolution. However, this is not always feasible, especially in areas that are characterized by rugged terrains. Irregular topographic surface causes giant variations in ground clearance, for the reason that the aircraft can’t precisely ‘drape’ over the topographic surface. Hence, a tolerance quantity is defined. The tolerance parameter is a threshold by which the aircraft may deviate from the nominal terrain clearance without rejecting the observed results. Flight lines that exceed this tolerance are required to be re-flown (Reeves 2006). To circumvent the problem encountered in the rugged terrain, a loose tolerance is usually pre-planned with help of the digital elevation model, such as SRTM, and DEM products. This would make matching altitude at the intersection of flight lines and tie lines an easy task. Unmatched altitude at the intersection points would harm the quality of the final data grid. The typical flight height adopted for most of the airborne surveys ranges from 100 m to 200 m for a fixed-wing aircraft. Helicopters can flow at an even lower altitude (Isles & Rankin, 2013).

Line spacing

Choosing an appropriate line spacing is vital when planning for a survey because setting a wide spacing lead to creating gaps in the final data grid. Even if it is possible to numerically interpolate across these gaps, it is impossible to say for certain that the observed patterns in the grid truly reflect geology. To determine the suitable line spacing, planners usually start by setting the desired geological resolution. With it comes also setting a tolerance parameter. The tolerance is selected in a way that the separation between 2 adjacent lines can exceed a factor, for example, 1.5, multiplied by the nominal separation. Surpassing this tolerance is only allowed for a limited distance. Going beyond this distance creates gaps in the data grid.

A set of empirical rules have been established to select the best line spacing:

- The grid data of the survey can only be zoomed in where 1 cm represents one flight line spacing (D. Boyd, 1967; D. M. Boyd & Isles, 2007). This means that a survey flown with 250 m line spacing will offer a grid down to a scale of 1:25000. Zooming further to a more detailed grid will diminish the correspondence of the magnetic data to real geology(Isles & Rankin, 2013).
- Splitting the line spacing into two will double the geological resolution.
- A good trade-off between the nature of the study being carried out and the cost of the survey should be taken into consideration to choose the optimal spacing. In reconnaissance geological surveys, a line spacing of 1 Km is chosen. In the absence of enough funding, wider line spacing of 2 Km can also be chosen. Another set of lines is usually flown perpendicularly to flight lines. They are called the control or traverse lines. They are flown with a line spacing about 5 times that of the line spacing.
- For studies that require a detailed flight line grid (e.g., uranium exploration), and assuming a low flight height (less than 100 m), a narrower spacing of up to 100 m can also be adopted.

Line direction

Usually, if the dominant geological strike of the study area is known, the flight lines are flown perpendicularly to that direction to maximize the signature of subsurface anomaly sources. Like that, the short wavelength across the strike can be sampled on every flight line. However, for reconnaissance surveys, direction N-S or W-E is often chosen.

At low field inclinations (i.e., Equatorial regions), magnetic bodies display small anomalies that tend to be stretched in an E-W direction. Therefore, the airborne magnetic surveys are flown in the N-S direction. This would better define the anomalies in these regions regardless of the dominant structural direction (Isles & Rankin, 2013; Reeves, 2006).

Tie lines

Tie lines, or control lines, are flown perpendicular to the flight lines. Their main utility is to offer an additional method to match the base level of the measured parameter across the flight lines. Tie line spacing is commonly taken as 10 times the flight line spacing. However, narrower tie line spacing is chosen in regions where more sensitivity is desired (e.g., sedimentary basins).

Sample interval

Modern magnetometers are capable of sampling 10 times per second. This sampling rate would provide enough measurement points that can be reliably interpolated. A lacking sampling rate would cause the appearance of artifacts (short-wavelength noise) in the grid data. According to the Nyquist criterion, the sampling frequency must be greater than twice the high-frequency component. In other words, the sampling rate must be less than half the short wavelengths.

During the airborne magnetic survey, the measurements are often made every second, which corresponds to an average measurement step of 55 m for a flight speed between 50 and 60 m/s (Groune, 2019). In an airborne gamma-ray spectrometry survey, the data is also acquired with a sampling interval of 1 s. Taking into account the velocity of the aircraft, which is usually about 198 Km/h, 60 to 70 percent of the count rate corresponds to an oval with a width twice the flying height and length twice the flying height plus traversed distance during the accumulation (Seligman, 1992).

2.5. Numerical filters

Filters in geophysical methods are related to separation or smoothing. When used as a smoothing tool, filters can be used to smooth the signal to minimize the short-wavelength noise, and when used as a separator, they consist of separating the desired signal from the undesired signal. The latter is due mainly to systematic error which includes observational errors, imperfect instrument calibration and environmental interference; random error is defined as unpredictable noise. It differs from one measurement point to another.

The filtering process is used to distinguish between two events as long as they differ in their characteristic. The common filtering method used to distinguish between signals is frequency filtering. This filtering procedure consists of using the “Fourier transformation.” This transformation converts a function of time to a function of frequency.

In geophysical methods, bringing the measured signal to the frequency domain, also called the “Fourier domain,” has many advantages which we can mention:

- The frequency components of the desired signal and noise can be separated. Therefore, after applying the Fourier transformation, different frequency-based filters can be used to amplify some frequencies and attenuate others. In practice, there is no such “Ideal filter” that can fully separate the two signals, thus, a quantity called “signal to noise SNR” is defined. The purpose of the filtering process is to improve this ratio as much as possible.

- The observed anomalies are usually the superposition of various geological bodies. These bodies tend to, depending on the geometry and depth, have distinguishable frequency components. Filters like low pass, high pass, and bandpass can be opted to separate them.

In reality, a discrete 2-dimensional version of the filter, called “2D Discrete Fast Fourier transform DFFT”, is used so that data is collected at discrete intervals. Moreover, by applying 2D DFFT, we can deal with the data as a function of wavenumber, or wavelength. By applying the DFFT, we are decomposing the signal into the sum of a finite number of sine and cosine function terms. They represent the real and imaginary coefficients of the transformed signal, and they are defined as:

$$\mathrm{Re}(k) = \sum_{x=0}^{N-1} f(x) \cos\left(\frac{2\pi xk}{N}\right) \tag{2-11}$$

$$\mathrm{Im}(k) = \sum_{x=0}^{N-1} f(x) \sin\left(\frac{2\pi xk}{N}\right) \tag{2-12}$$

Both terms can be used to calculate the power spectrum of the signal. It is calculated as follows:

$$|F(k)|^2 = (\mathrm{Re}(k)^2 + (i.\mathrm{Im}(k))^2)^2 \tag{2-13}$$

In Figure 2-5, we can see how calculating the power spectrum facilitated the separation between the different frequential components of the input signal. After separating the desired anomalies, the “Inverses Fourier transform” is used to bring back the data to the spatial domain.

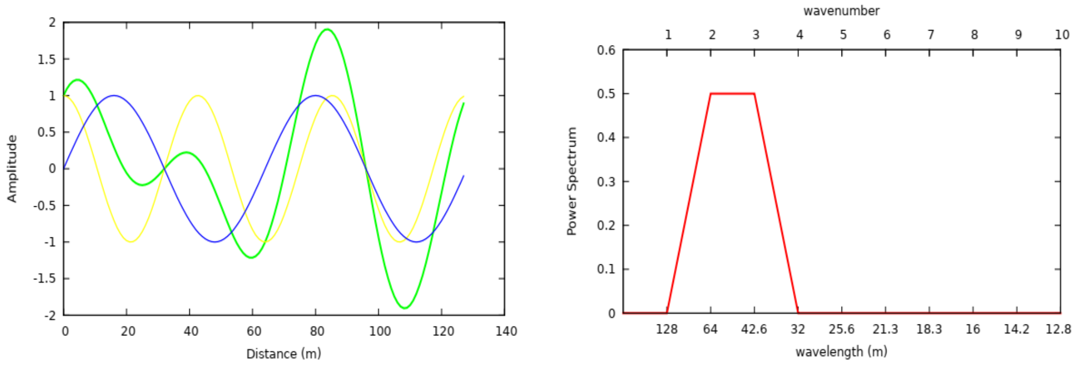


Figure 2-5 the usage of the power spectrum to distinguish between the frequential components of the green signal. In the left figure, the green signal is plotted in the spatial domain. It is defined as the superposition of the yellow and blue signals. Both signals have distinct picks in the wavenumber domain (right figure) that correspond to different wavenumber values. (From (Chuck & Laura, n.d.))

For the data acquired during airborne surveys, various filters are applied to extract particular information or enhance the quality of the data. For example, in airborne magnetic surveys, the filtering process can be used to separate the geomagnetic field from the anomaly field or transform the magnetic signature of the observed anomalies, while in airborne gamma-ray spectrometric surveys, it can be used to minimize the interference of the short-wavelength noise. Some of the common filters which are applied to the data of the airborne surveys are presented in the next section.

2.5.1. Reduction to pole transformation

The magnetic anomalies issued from bodies that are not located near the magnetic poles display an asymmetric magnetic signature. This problem occurs due to the effect of the orientation of the magnetization vector. This vector is oriented according to the inclination angle.

The purpose of this filter is to eliminate the effect of the inclination and transform the response of an arbitrary anomaly, located at any place on the earth, to the response of that particular anomaly if it were measured in a polar region. The vertical magnetization vector (i.e., I=90°) in these regions simplifies the observed anomalies.

This transformation renders the geological interpretation of the magnetic data easier, because the magnetic anomalies in the polar regions have a symmetric shape, and this would position the peak of these anomalies vertically to the causative magnetic body.

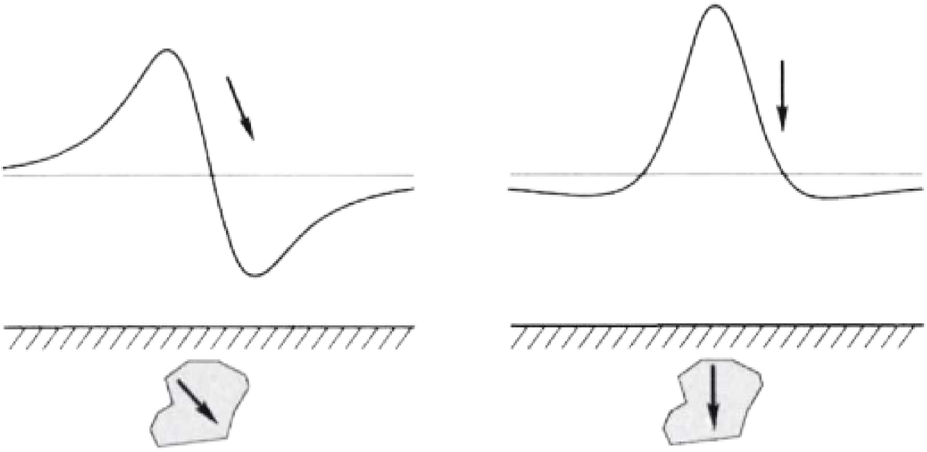


Figure 2-6 the magnetic signature of an arbitrary body. The image to the left is the observed anomaly in a mid-latitude region. It has an asymmetric shape. The image to the right is the same signature after applying the reduced to pole transformation. The pick of the anomaly is located vertically to the causative body. (Blakely, 1996)

The operator of the reduced to pole filter is as follows:

$$F[\Psi_{\text{pole}}] = \frac{1}{[\sin I_a - i.\cos I \cos(D+\lambda)]^2} \tag{2-14}$$

Where λ is the wavenumber direction. It is calculated as $\arctan {}^{k_x}/_{k_y}$; kx and ky are the wavenumbers in the x and y directions respectively; I and D are the geomagnetic inclination and declination.

It should be noted that this filter is only stable at high latitudes. In equatorial regions, where the latitude values are small, the filter grows increasingly unstable. Normally, latitude values less than 15° destabilize the operator of the filter. (Silva, 1986)

To circumvent this problem,(Macleod et al., 1993) introduced a correcting factor I_a in equation (2-14) to correct the amplitude of the transformed data. It is chosen based on the latitude of the study area.

Another problem with this filter is that the implementation of this filter in the frequential domain requires that the values of the inclination and declination remain invariable in the entire study area, and this is not the case in vast region studies where the inclination and declination values change in terms of spatial coordinates(Ansari & Alamdar, 2009).

2.5.2. Butterworth filter

This filter is a type of signal processing that is applied in the Fourier domain to smooth the input signal. It is designed to mimic the response of an ideal filter that has a response as flat as possible in the passband. This filter is also referred to as a *"maximally flat magnitude filter."*

Butterworth filter has a slow transition band (i.e., slow roll-off) that is centered around the cut-off wavenumber. This minimizes the ripple effect (Gibbs oscillation phenomena) in the vicinity of the cut-off wavenumber. The rate of transition is controlled by the degree of the filter. Increasing the degree offers a narrower roll-off, but this produces ripple oscillation. On the other hand, decreasing the degree gives smoother curves, but at the expense of a wider transition band. In Figure 2-7, we demonstrate how the degree of the filter controls the width of the transition band.

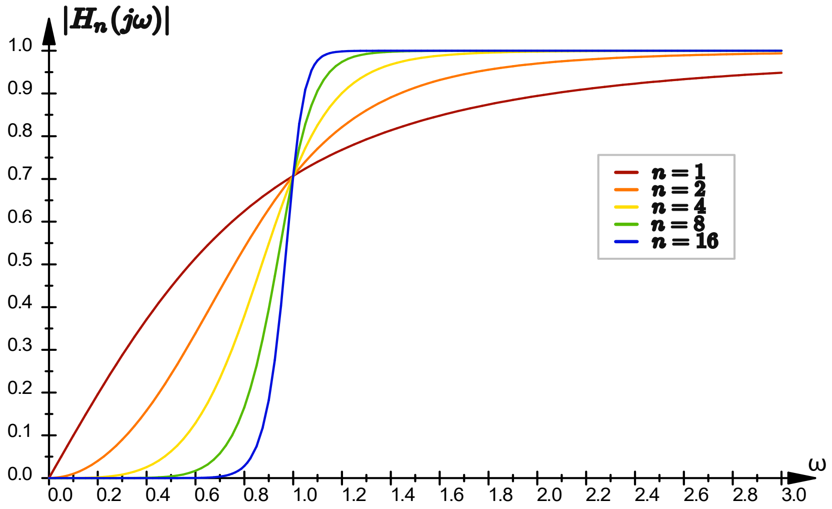


Figure 2-7 the response of a high-pass Butterworth filter calculated with different degrees n. Increasing the degree narrows the transition band between the passband and the stop band. (From (Pieter, 2021)).

Compared to other filters, it offers the smoothest curve in the passband as well as the stopband. However, due to the slower roll-off, it requires a higher degree to filter the same wavenumber band as other filters (Podder et al., 2014).

The Butterworth filter can be used to attenuate high wavenumber as well as low-wavenumber components. Equation (2-15) and equation (2-16) represent the operator of the filter when used as a low pass filter and high pass filter respectively.

$$L(K) = \frac{1}{\left[1+\left(\frac{k}{k_c}\right)^n\right]} \tag{2-15}$$

$$L(k) = \frac{\frac{k}{k_c}}{\left[1+\left(\frac{k}{k_c}\right)^n\right]} \tag{2-16}$$

2.5.3. Cosine directional filter

The directional filter is a filter that can be applied after transforming data with DFFT. It is used to remove noises oriented with a certain azimuth. These noise components are usually characterized by a linear geometry. The operator of the filter is as follows:

$$L(\theta) = \left|\cos\left(\alpha - \theta + \frac{\pi}{2}\right)^n\right| \tag{2-17}$$

$$L(\theta) = 1 - \left|\cos\left(\alpha - \theta + \frac{\pi}{2}\right)^n\right| \tag{2-18}$$

The degree *"n"* controls the roll-off rate, and α is the angle along which the noise is directed. The operator of the filter in equation (2-17) is employed to reject features in the direction of α, whereas the operator in equation (2-18) is used to reject all the features that are not in the direction of α.

The advantage of using this filter instead of the straight low/ high filter is that it minimizes the ripple effect associated with DFFT. A common use for this filter is to de-corrugate the poorly leveled magnetic data.

2.5.4. Pseudo gravity transformation

This filter can be applied in the Fourier domain. Given that a geological body has the same magnetization and density boundaries, and has a homogenous distribution of the magnetization vector and density, the magnetic potential issued from that particular body can be related to its gravimetric potential using the following operator:

$$V = -\frac{C_m}{\gamma}\frac{M}{\rho}\widehat{m}\nabla_p U \tag{2-19}$$

with : $g_m = \widehat{m}\nabla_p U$

ρ is the density; M and the intensity of the magnetization; $\hat{m}$ the direction of the magnetization; g_m is the gravity component in the direction $\hat{m}$; γ is the gravitation constant and C_m is a constant equal to: $10^{-7}\ \mathrm{Henry/m}$.

In the case when we are dealing with variable distribution of the density and magnetization vector, we can think of the body as an amalgamation of elementary bodies. The distribution of the magnetization and density can be considered uniform in these elementary bodies. By a simple summation, we can get the distribution of both parameters for the entire body (Blakely, 1996)

This transformation proves to facilitate the geological interpretation of a magnetic anomaly because determining the geometry and the location of the gravimetric anomalies are easier than those of the magnetic anomalies.

When we try to interpret magnetic anomalies, multiple factors should be taken into consideration such as the inclination of the magnetization vector and the orientation of the structure of the magnetic meridian. All these factors distort the shape of the magnetic anomalies and make it difficult to locate the causative bodies. On the contrary, the gravimetric anomalies are easier to interpret for the reason that they do not suffer from the distortion effect due to the variation of the inclination. Therefore, their spatial extension can be determined quite directly.

The anomalies calculated using this filter do not reflect a true gravimetric anomaly, because even after deducing its gravimetric signature, we still do not have information about the density of the feature. Moreover, this formula only takes account of magnetized masses in the body. There could be other parts that are not magnetized but can contribute to the gravimetric anomaly of the body (Baranov, 1957).

This filter is usually applied in the Fourier domain, and similarly to the reduced-to-pole transformation, it also has some limitations in the low-latitude regions. Another limitation of this filter is that it can amplify the long-wavelength anomalies and attenuate the short-wavelength anomalies. Therefore, if the study area is characterized by the presence of a long-wavelength noise, this transformation should be used with caution.

2.5.5. Apparent density and apparent susceptibility transformation

These transformations are idealized filters that assume a limited spatial extension and apply the operator in equations (2-20) and (2-21) to calculate the apparent density grid and the apparent susceptibility grid respectively. These transformations are usually applied to gridded data with a fixed grid cell size.

The apparent density operator takes the observed gravity field as input and calculates the apparent density. This filter is also called “Pseudo density”

The geometry model of the gravimetric anomaly is considered to have a fixed thickness *"t"* and varying density. The geometry model of the causative body is also assumed to be caused by a combination of vertical, square-ended prisms of infinite depth extent. The horizontal dimensions of which are equal to the input grid cell size.

The susceptibility operator is a compound filter that performs a reduction to the pole, downward continuation to the source depth, correction for the geometric effect, and division by the total magnetic field to calculate the susceptibility. This operator takes the same assumptions regarding the geometry of the causative body as in the density operator. In addition to that, it also assumes that the observed anomalies do not have a remanent magnetization.

$$L(r) = \frac{r}{2\pi(1-e^{-tr})} \tag{2-20}$$

G is the gravitational constant, and t is the thickness of the model.

$$L(r,\theta) = \frac{1}{2\pi F.H(r)\Gamma(\theta)K(r,\theta)} \tag{2-21}$$

Where:F = total magnetic field;

$H(r) = e^{-hr}$. It is the downward continuation to the causative body, and h is the depth to the causative body;

$\Gamma(\theta) = [\sin I_a + i\cos I.\cos(D-\theta)]^2$. It is reduced to pole operator. I and D are the inclination and declination of the magnetic field.

$K(r,\theta) = \frac{\sin(\arccos\theta).\sin(\arcsin\theta)}{\arccos\theta.\arcsin\theta}$. This operator is introduced to correct the geometric effect.

The most widely used term to describe these transformations is "*apparent density*" and "*apparent susceptibility*," as the calculations are based only on assumptions regarding the geometry. The calculated density and susceptibility are only an approximation to the real values (Hinze et al., 2010).

Chapter 3 : Processing the Airborne survey data

3.1. Introduction

From 1969 to 1974, the American company "AeroService Corp" was appointed to carry out regional aerogeophysical coverage of the Algerian territory to support geological mapping and oil and mining prospecting. These surveys covered the whole Algerian territory and were conducted in two stages, the first in 1969 on behalf of the oil company SONATRACH and the second took place from 1971 to 1974 on behalf of the mining company SONAREM.

3.1.1. Airborne surveys in Algeria

3.1.1.1. 1969 survey

This survey covered 20% of the Algerian territory, which equates to an area of 418.000 Km^2. This survey was issued by *"SONATRACH"* and had as a primary objective to explore hydrocarbon. Therefore, it was concentrated on the sedimentary basins in the central Sahara. In this survey, only magnetic data were recorded. This survey was carried out at different barometric altitudes that varied from 800 m to 1100 m, and for the survey scale, the company opted for a broad line spacing of 5 Km.

3.1.1.2. 1971-1974 survey

It took 35 months to complete this survey as it started at the beginning of 1971 and ended by the beginning of 1974. It covered the rest of Algerian territory and accounted for an area of 2.173.000 Km^2. The flight distance that covered this area was approximately 904.500 Km. This survey was issued on behalf of the "*SONAREM*". The survey had the main objectives of identifying areas with high mineralization potentials, and mapping the tectonic structure and determining the extent of the major tectonic units. In this survey, a magnetometer was equipped along with a gamma-ray detector to keep a record of both magnetic and gamma-ray signals.

Due to the vast surface covered by this survey, the national territory was split into multiple blocks and each block was flown with different survey specifications. Some blocks were flown with a line spacing density of (2 ×40) Km for the line and traverse lines respectively, and others were flown with a density of (5 ×25) Km. For other regions, the company opted for an even tighter density of (2×10) Km. Figure 3-1 illustrates the flown blocks and their specifications. All these blocks were flown with a nominal flight height of 150 m.

As explained in the previous paragraph, the line spacing density varied from one block to another. One reason for that is the magnetic anomalies in the northern parts of the nation and the sedimentary basins display a very low amplitude which makes them hard to capture. Another reason is, back then, these regions were believed to have an economical interest. Therefore, these blocs were flown with a tighter traverse line spacing equal to 5 times the line spacing. On other hand, the geological structure in the southern part of the Algerian territory, which is characterized by a very old crystalline basement, the magnetic anomalies display higher amplitudes than in the northern parts. These anomalies are easily captured, thus the surveys in these regions were flown with a broad traverse line spacing of 20 times the line spacing.

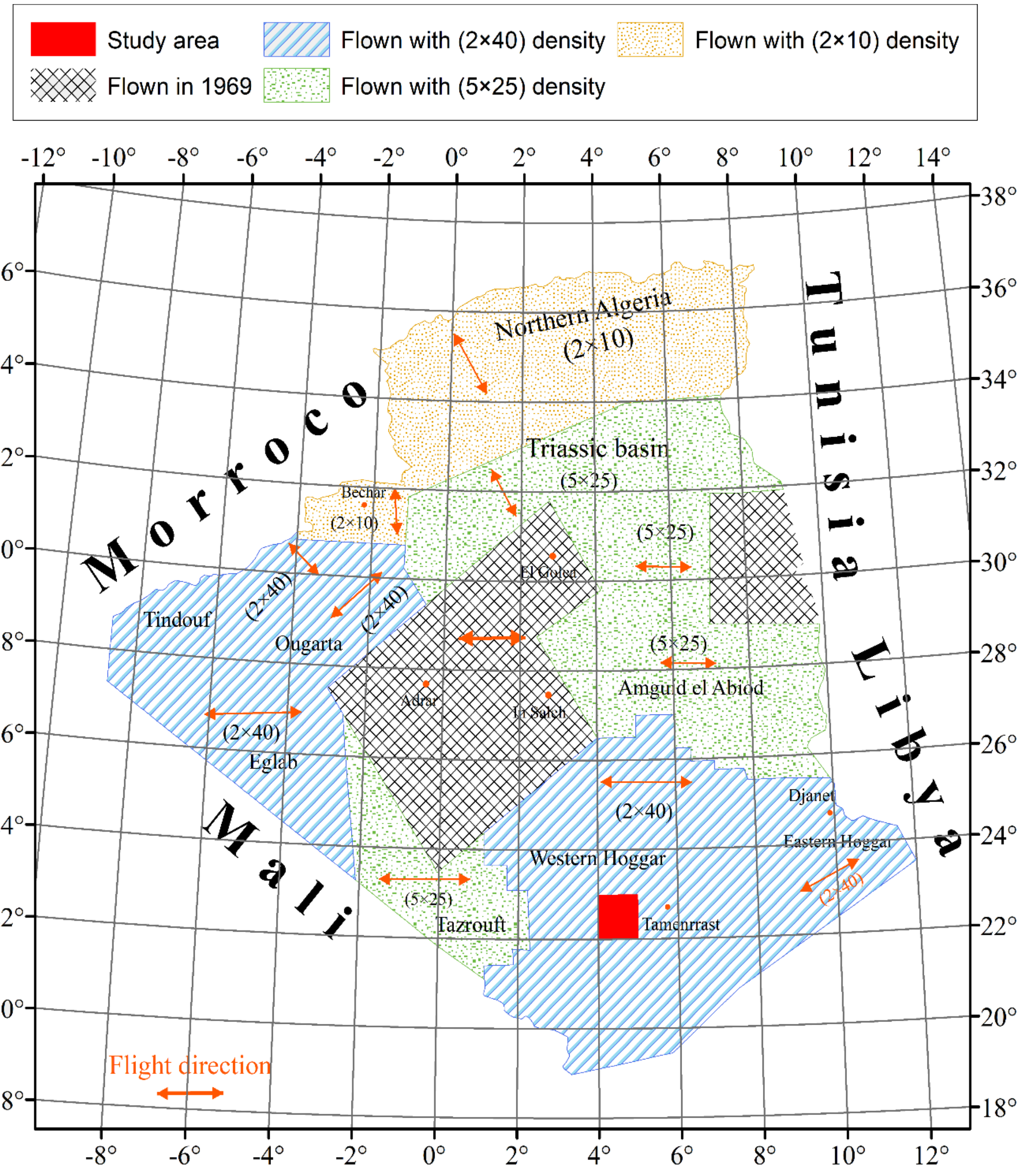


Figure 3-1 Flight line spacing densities. Study area is highlighted in red square (AeroService corporation 1975).

3.1.2. The Aircraft characteristics

The survey planning consisted of using three aircraft of two different models. A single aircraft model *"DC-3"*, and two aircraft models *"Aero-commander"*. The DC-3 aircraft had an accident that caused the company to replace it with an aircraft of the same model, and one of the aero-commander aircraft, after having its detector damaged, was withdrawn from the survey.

Both models underwent multiple modifications to minimize the effect of the residual magnetism. To do that, the interior of the aircraft was completely replaced by a non-magnetic material, and even the electrical installation was assured with special wiring to minimize the effect of the magnetic field induced by the electrical current.

Two installation systems were used to carry the magnetic detectors. The rigid boom *"Stinger"* installation and the towed bird system. For the DC-3 aircraft, they adopted the towed bird system where the detector was suspended 21 m below the aircraft, and for the aero commander, the detector was installed in a boom attached rigidly to the airframe of the aircraft. Both systems present a good strategy to minimize the noise due to the aircraft body and the recording instruments.

For the detector model, an optically pumped magnetometer was used for DC-3 aircraft and a Gulf fluxgate magnetometer for the aero commander (Allek, 2005).

3.1.2.1. Navigation system

A doppler Bendix navigation radar model *"A DRA 12"* was used in the aircraft to determine the expected navigation error when flying straight and to provide the exact location of the measurement points. These measurement points were recorded with a sampling distance of 46 m. Alongside the navigation system, a gyro-magnetic compass model *"Sperry C12"* was also attached to keep a record of the aircraft's orientation. This compass had a resolution of 1°.

3.1.2.2. Altimeter

An altimeter radar model *"Honeywell Minneapolis"* was used to monitor the elevation of the aircraft above surface level. It measures the altitude by sending a radio signal toward the surface. It can calculate elevation by measuring the time for the signal to bounce back. During the survey, the flight height was maintained at a nominal altitude of 500 ± 30 feet (approximately 150 ± 9 m).

3.1.2.3. Camera

A 35 mm continuous-strip camera, also called *"Sonna"*, was used to photograph the entire flight path. These cameras were designed for low-altitude, high-speed photography which made them suitable for this kind of survey. The camera keeps a record of the terrain by passing a film, synchronized to the aircraft movement, over a stationary slit. The slit is usually very small, and only a narrow strip of the ground below is recorded. As the aircraft moves in flight, the film continuously records an image of the terrain below, and the final image is an uninterrupted strip photograph of the path traveled by the aircraft.

The films recorded by the camera were used to determine the exact location of the flight line as well as to calculate the traveled distance by the aircraft. This was carried out by marking the film by the means of an automatic fiducial numbering system that simultaneously prints marks on the geophysical records. The fiducial number serves as a reference during the compilation of the data. The fiducial marks in the AeroService survey were spaced 1520 feet (460m) apart.

3.1.3. Acquisition and compilation systems

The air photo mosaics serve as a base map for planning the flight lines and compilation of the geophysical data. They are a series of aerial photographs put together in a way that the detail of one photograph matches the detail of all adjacent photographs. During the survey, a *"flight strip,"* which contains the intended flight lines that were marked and numbered by the camera, of the air photos is provided to the navigator. At the end of each flight, the flight lines are recovered by plotting the fiducial marks of the 35 mm films in the mosaic photo. Since the geophysical records are also marked by the same fiducial marks, they can be accurately positioned in the recovered flight line by simply matching the fiducial marks.

An acquisition system *"Lancer"* was used to sample all the recorded data including the geophysical records and other complementary records such as altitude and fiducial marks. These records were sorted and digitized by the means of the acquisition system and then saved on magnetic tapes that were, by turn, connected to an IBM computer #360 model 44. This computer was used to preprocess and compile the recorded data.

3.1.4. Magnetometers

Two magnetometer models were used to record the total magnetic intensity during the survey. The first model is an optically pumped magnetometer model *"Varian."* It was carried by the DC-3 aircraft and had an accuracy of 0.02 nT. The other magnetometer model was used for the aero-commander aircraft. It was the fluxgate *"Gulf"* magnetometer, and it had an accuracy of 0.5 nT. To record the diurnal changes, an optical absorption magnetometer operated by cesium gas was used.

3.1.5. Spectrometers

To keep a record of the gamma-ray counts in the four windows uranium, thorium, potassium and total count a detector with a crystal of sodium iodide activated with thallium was used. *"Horshow Hammer,"* a gamma ray detector, with different crystal volumes was used in both aircraft. In the DC-3 aircraft, a crystal of 800 pouces^3 (~13 L) was used, and in the aero-commander, the volume of the crystal was 600 pouces^3 (approximately 10 L).

3.2. Data acquisition

After receiving the data from the field, the geophysical data underwent a careful analysis. These analyses were carried out by a team of geophysicists who had the final objective to prepare a final interpretation document. This document also included recommendations to re-investigate areas that showed economic importance.

The following section explains the various preparations, preprocessing and interpretation steps that the recorded data underwent by AeroService company.

3.2.1. Preparation

Preparing data initiates by converting the magnetic records to *"nano Tesla."* The converted data are then inspected for erroneous values that are not in conformity with adjacent records. After that, non-linear filters are applied to eliminate and replace them with plausible values.

3.2.2. Lag correction

This correction removes the effect of distance between the detector location and the positioning sensor. This distance is called *"lag,"* and the correction is carried out by subtracting the lag (in fiducial) from the start fiducial of the data channel in the database. The value of the lag is a function of the installation system of the detector. The records measured from the DC-3 aircraft, which had the towed bird installation, suffered more from the lag effect compared to the records measured by the aero-commander (i.e., rigid boom installation).

Unfortunately, an examination of the data acquired over the western Hoggar, where our study area is located, shows that this correction was not taken into account or was poorly applied (Groune, 2019)

3.2.3. Positioning

The positioning starts by searching for the coordinates of the intersection of the flight lines with the traverse lines. The search method is carried out by assembling all the marked films of the traverse lines and the flight lines which they cut across. Like that, we can obtain the fiducial number, which was marked during the flight, and the coordinates of the intersection points by visually inspecting the intersection areas. After that, an interpolation program, which uses the fiducial number of the intersection points, is employed to calculate the coordinates of all the measurement points as well as the exact configuration of the flight path.

3.2.4. Leveling mise ties

The leveling of the magnetic data aims to minimize the difference between the measured values of the flight lines and the traverse lines at the intersection points. This difference is caused by the short temporal changes of the magnetic field, especially the diurnal changes. Other common problems that cause mise ties are poor positioning of the intersection points, the height difference between the flight and traverse lines, instrumental interference, and random noise.

The common procedure to eliminate the diurnal changes is to use a base station. These base stations are located at a fixed position near or inside the survey area. The records of these stations are then subtracted from the records of the aircraft, and the difference between the two records is taken as the magnitude of the diurnal changes. However, this procedure was not used in this survey and the records of the base stations were only employed to detect the magnetic storms. It is very important to determine when magnetic storms occur because they inflict sudden and drastic variations of the magnetic records. These uncontrollable variations reduce the quality of the recorded data; thus, flying operations during these storms are usually suspended.

To eliminate diurnal changes, AeroService company consisted of developing a low-order polynomial surface (2^{nd} or 3^{rd} order) by employing the least square approach. The coefficients of this surface are calculated in terms of the difference of the records at the intersection points. These coefficients are then calculated in a way that the surface would minimize the root mean square of the differences.

3.2.5. Subtracting the internal magnetic field IGRF

In the final report of the aero service company, it was mentioned that the internal magnetic field was subtracted by using a gradient that is estimated using the IGRF 1965. Further inspection of the magnetic tapes reveals that the values of the retrieved magnetic records were not only corrected by subtracting the IGRF 1965 but also were corrected by adding a constant value of 34000 nT to the magnetic data. This value represents a reference value that was used by the AeroService company as part of another contract with the Malian government (Allek, 2005).

3.2.6. Map tracing

Contour maps were created for both the magnetic data and spectrometric data. These maps were created by interpolating the data according to a fixed contour interval, and the final product was 1/100,000 and 1/200,000 scale contour maps.

For the magnetic data, the contour interval varies, depending on the magnetometer used and flown block, between 2 and 25 nT, and for spectrometric data, a fixed contour interval of 100 cps (count per second) was used for all blocks. In addition to the contour maps, 1/200,000 scale maps that plot the flight path of the surveys were also traced. The created maps were projected to the "*Universal Transverse Mercator"* UTM projection system according to *"Clarke 1880"* reference ellipsoid.

Various quantitative interpretation methods were used to locate the magnetic anomalies. They were carried out by applying the *"Werner deconvolution"*.The end products of this method were interpretive maps that reveal the depth, the susceptibility, and the rock type of the causative bodies related to the magnetic anomalies.

3.3. Quality control of Silet airborne data

3.3.1. Data reformatting

At the present, the data were retrieved from the magnetic tapes and saved in "*.DAT*" files that contain all flight lines of the surveyed blocks. To distinguish between the flown blocks, a special naming strategy was used for the DAT files. It consists of using an acronym of 2 letters to refer to the name of the region followed by 2 digits number that indicates the UTM zone. It also includes the number of the segment which is separated by an underscore character from the name and UTM zone.

Our study area, Silet, is located in the western Hoggar region, corresponding to the UTM 31 zone. It was flown as a part of the first segment of the survey in this area. Therefore, the DAT file that includes Silet data holds the name *"HW31_1."*

The raw format of the DAT files starts with the header of the flight line which holds the following information: line number, measurement points numbers, fiducial number, the start time of the flight line, finishing time of the flight line, the azimuth, flight number, date, and $X_{0,}$ and Y_0, the coordinates of the reference point. After the header information, the values from the geophysical data (magnetometer and spectrometer) are recorded in separate fields and the information relating to the position (relative X, relative Y, altitude) is multiplexed in a single field in a way that every measurement point has either relative X, relative Y or altitude record. A sample of the raw formatting is presented in Table 3-1.

Table 3-1 The raw formatting of the DAT files. It is a sample from the HW31_1 DAT file.

Line number	*Measurement points number*	*Fiducial number*	*Start time*	*Finish time*	*Azimuth*	*Flight number*	*Date*	*X_0*	*Y_0*
11780	*2546*	*147825*	*657*	*725*	*90*	*525*	*40571*	*102167*	*215959*
1702118	762	530	21	38	128	-	-	-	-
1702097	759	-29999	21	26	151	-	-	-	-
1702076	756	530	20	35	130	-	-	-	-

From the above table, we can see that the data at its current state needs to be reformatted. Firstly, the values of the magnetic records, in the first column, are the real values multiplied by 50. Secondly, the formatting of the time and date, which both are presented as a real number, is not adequate for both quantities. Thirdly, the fiducial number is only available for the first measurement point. Lastly, the coordinates and the altitude are multiplexed in one column which leads to a lack of the exact positioning of the measurement points.

A script that runs multiple functions was developed to solve the above-mentioned formatting problems. The first function starts by separating the relative coordinates and the altitude in the multiplexed column. After that, it uses a linear interpolation function to calculate the relative coordinates and the altitude for all measurement points. The function simultaneously converts the altitude records from foot to meter as well. The absolute coordinates are calculated by summing the coordinates of the reference points (X_0, Y_0) with the relative coordinates and multiplying the sum result by 10. The second function is used to transform the time and date into an adequate format. Subsequently, it produces the time column by assigning the start time to the first measurement point and iteratively adding 1 second to calculate the time for the next point. For the date column, the function just uses the same value for all measurement points. The last function is used to calculate the fiducial number. It takes the value of the fiducial in the header and increments it by 1 to calculate the next fiducial number for the next record. The final result of the script is presented in Table 3-2.

Table 3-2 The format of the same DAT file sample (HW31_1) after employing the developed script.

Line 11780

X_abs	*Y_abs*	*Altitude (m)*	*Fiducial*	*Azimuth*	*Mag (nT)*	*Tc*	*Th*	*U*	*K*	*Time*	*Date*
721633	2459590	161.5	147825	90	34042.4	762	21	38	128	6:57:00	1971/04/05
721680	2459590	161.5	147826	90	34041.9	759	21	26	151	6:57:01	1971/04/05
721727	2459590	161.5	147827	90	34041.5	756	20	35	130	6:57:02	1971/04/05

After completing all transformations, the script saves the final product to a (.XYZ) file. This extension is commonly used for the data imported to *"Oasis montaj"*. It is a computer software for analysing and modelling geophysical data.

3.3.2. Initial preparations of the Silet data

The absolute coordinates in the transformed files were projected according to the datum *"Clarke 1880"* reference ellipsoid. After obtaining the coordinates, we calculated the geographical coordinates (Longitude, Latitude) for all the measurement points. Subsequently, they are used to subset the data for our study area *"Silet"*.

Figure 3-2 shows the flight path of the data for the western Hoggar block (HW31_1). The flight path of Silet region is highlighted in red. The figure also shows that in the study area, a direction E-W was chosen for the flight lines and the tie lines were flown in a perpendicular direction N-S. With this in mind, we used the azimuth channel, which is a record of the direction of the lines, to separate the flight line from the tie-lines.

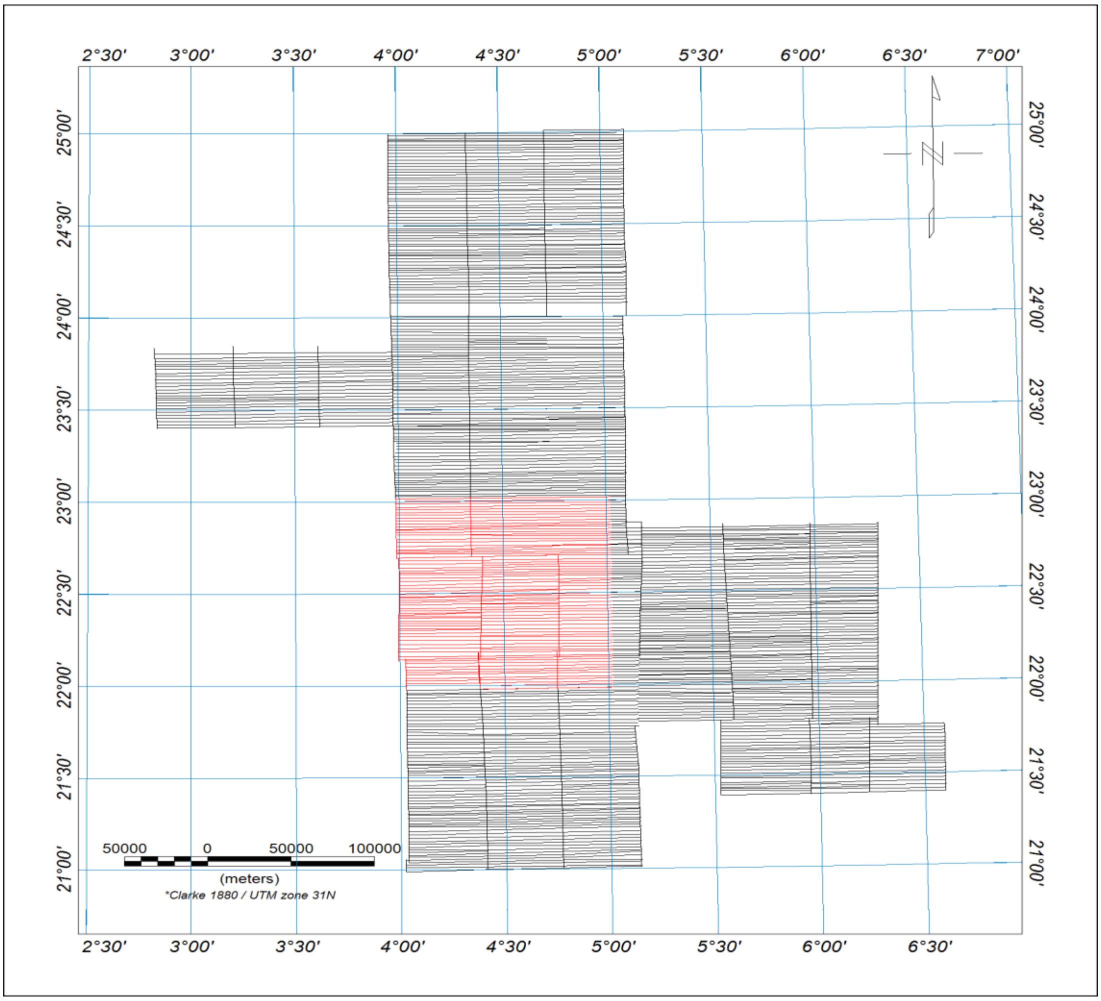


Figure 3-2 The flight path of the western Hoggar block (Hw31_1 XYZ file). The Silet region is highlighted in red.

The Silet area was flown with a flight line spacing of 2 Km and 40 Km tie-line spacing. It was covered using 81 flight lines and only 8 tie lines. The total flight distance equated to 7953,1 Km.

3.3.3. Evaluation of the quality of the data

The total number of measurement points in the Silet database is 169,197 points. The majority of them, 161,459 points, which account for approximately 95% of the data, were acquired in 1971 and a small proportion, 1667 points (~1%) of the data, was acquired in 1972. Moreover, the date records of the rest of the points, 6071 points (~4 %) of the data, were corrupted due to the poor preservation of the magnetic tapes; thus, no information about when they were acquired.

The data was inspected for faulty records. This includes the missing records as well as the inaccurate negative. Afterwards, a summary is plotted as a bar chart where the height of the bars indicates the number of faulty records. The data where no date information is present was not included in the figure below. However, it is also analysed for faulty records and the inspection showed that no faulty records were present.

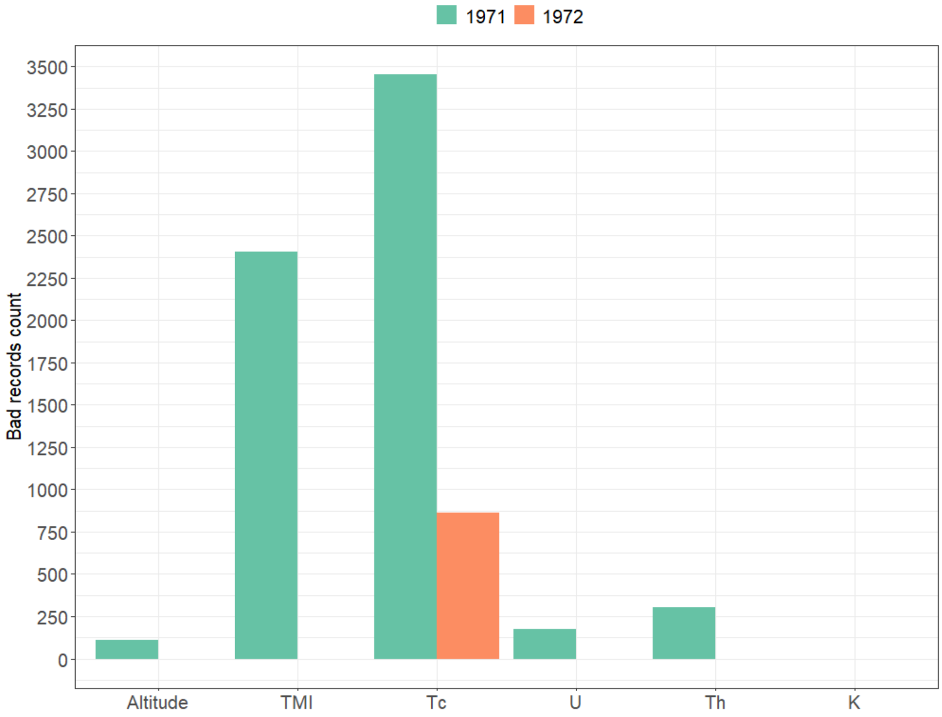


Figure 3-3 The count of the faulty records (missing values+ negative values) in the Silet database. They are organized according to the date channel.

As we can see in Figure 3-3, the faulty records for the 1971 data are concentrated in the total magnetic intensity *"TMI"* and gamma-ray total count *"Tc"* channels, where the height of the bar exceeds the value 3375 and the value 2375 for the two channels respectively. The figure also shows that aside from the potassium channel *"K"*, where no faulty records are present, the number of faulty records for the other channels is less than 350. For the 1972 data, and excluding the Tc channel where 875 faulty records are present, no faulty records are observed for the other channels.

The statistical analysis shows that the majority of the data of Silet is in good condition, and only a tiny proportion of ~4% is corrupted; therefore, we can proceed to evaluate the other specifications of the survey.

3.3.4. Evaluation of the flight height

The nominal flight height in Silet was set to 150 m. In this section, we try to evaluate how was the nominal flight height tolerated by performing a statistical study. The summary of that study is presented in Figure 3-4.

Figure 3-4 is a plot of the altitude of the measurement points across the flight lines, where the altitude values were split into 4 discrete ranges, and to each range, we assigned a colour that indicates which range a particular point belongs. In addition to that, we accompanied another chart, commonly called the *"Doughnut chart"*, to better discern the proportions of the data that falls within each range.

The value of nominal flight height falls in the range between 100 m and 180 m which is also the range that holds ~94% of the data. Thus, this range was statistically analysed by calculating the mean value, the standard deviation and the variation coefficient to determine which value our data is dispersed around and to compare it, afterward, with the nominal flight height.

The mean value for this range, which equates to ~153 m, and the small variation coefficient value of 0.06 indicate that the great part of our data is tightly dispersed around the value ~153 m; therefore, we can safely say that the nominal flight height in the Silet region was acceptably tolerated.

To explain the sudden increase of the flight height in some areas, where the altitude of the radar exceeded 250 m, we resorted to the digital elevation model (DEM) of the region. The DEM file showed that, in these areas, irregular variations of the altitude of the topographic surface impeded the aircraft from precisely draping over it according to the nominal height; consequently, the important increase of the aircraft's altitude while surveying in these areas.

Some altitude records are less than 100 m (some even less than 50 m) which in our opinion are not plausible values and they might be corrupted records due to the bad preservation of the magnetic tapes.

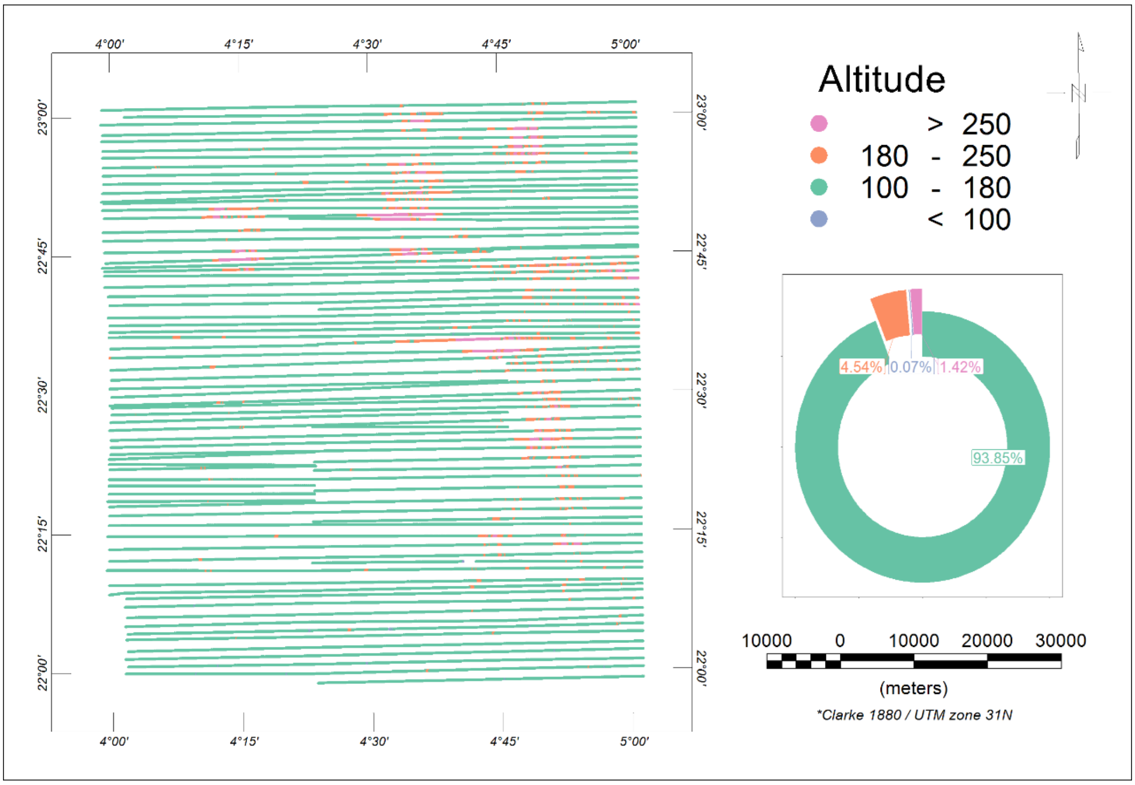


Figure 3-4 The radar altitude across the flight lines in the Silet region. The bulk data was acquired at an aircraft altitude that conforms with the nominal altitude.

3.3.5. Evaluation of the flight-lines deviation and their separation

In the Silet region, the flight lines were flown in east-west direction with a separation of 2 Km. Other specifications concerning the flight lines are the maximum separation and the deviation tolerance. The first is how much the maximum distance between two lines can reach, and the latter is an assessment of whether or not a certain line was being deviated from the planned *"ideal"* path.

In this section, we evaluate the flight line deviation from the ideal path by checking if the distance between a particular line and its ideal path exceeds 400 m over a distance of 15 Km, and we also evaluate if the flight lines separation exceeds the maximum separation, which was set to 2.5 Km.

An ideal flight path is defined, as far our study area is concerned, as having flight lines that are separated exactly by 2 Km and oriented exactly in the direction E-W. Unfortunately, the ideal flight path for our study area is missing. Therefore, we decided to build a synthetic ideal flight path to assess if these specifications were tolerated.

Flight lines in our study area are oriented in the direction E-W which implies that the coordinate Y across a flight line should remain fixed and only the coordinate X should change. In light of that, we started by extracting the coordinate limits (X_{min}, X_{max}) and (Y_{min}, Y_{max}) of every flight line, after that, we assigned, depending on the azimuth of the line, Y_{min} or Y_{max} as the fixed value for the Y coordinate. For the lines with an azimuth of "270", we assigned the Y_{max} as the fixed value due to its tendency to deviate to the left of the ideal path. For the lines with an azimuth of "90", we assigned the Y_{min} as the fixed value because they tend to deviate to the right of the ideal path. In Figure 3-5, we show a comparison between the real flight path (in red) and the corresponding ideal path of the same lines (in green).

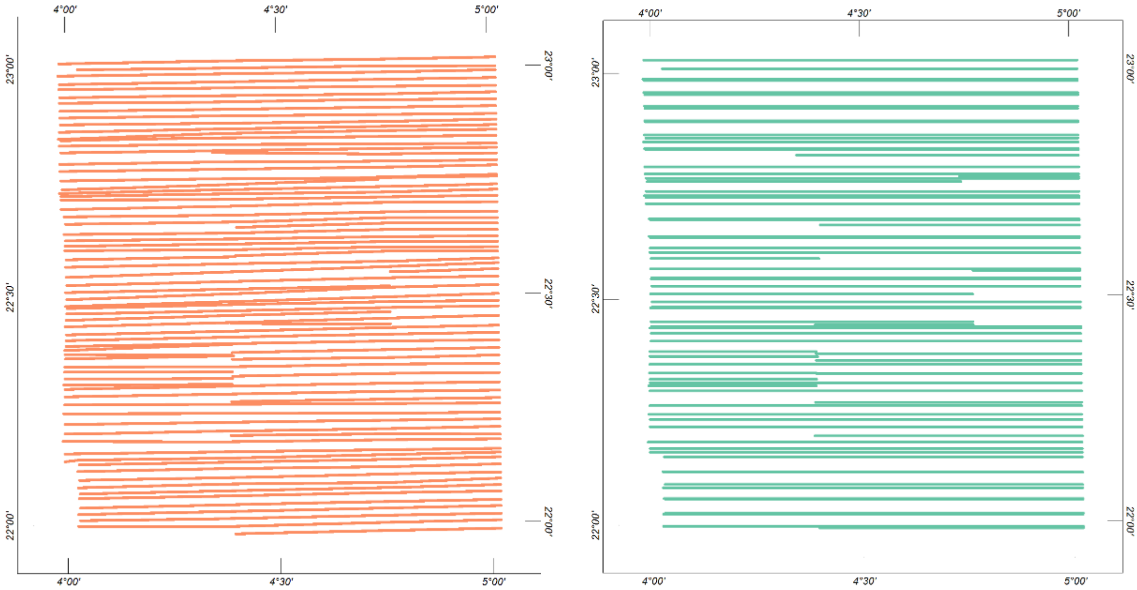


Figure 3-5 Comparison between the real (to the left), and the ideal flight paths (to the right).

Figure 3-6 is a plot of the flight path, on which, we plotted the segments that were flown beyond the deviation tolerance as well as the segments that were separated by more than the allowed maximum separation (i.e., 2.5 Km). The figure also includes a chart that shows the proportion of the segments where whether or not the two parameters were tolerated.

The analysis of the flight lines separation shows that it was highly tolerated, as the separation of the majority of the lines (95%) was less than the maximum allowed separation. Contradictory to the separation, the maximum allowed deviation was not tolerated to a great extent, since a large segment of the lines (65%) got past the allowed deviation.

We could not find the real reason why a substantial ratio of the flight segment bypassed the deviation tolerance. One possible reason for that is the nature of the survey that our study area was flown as a part of, which covered a vast area. Therefore, it was hard to maintain a fixed direction throughout the survey area. A more plausible reason would be that the deviation of flight-line was intended. As we can observe in Figure 3-5; unlike flights-lines where the ideal path was respected, by slightly deviating the flight path, the aircraft was able to cover the area with fewer flight lines and simultaneously decrease the observed gaps between the lines. This leads, consequently, to minimize the costs of the survey. The last reason is that the basis of our analysis is the synthetic ideal path we built which might not conform with the actual-ideal path that was planned back then. Thus, our analysis might not reflect the magnitude of the real deviation.

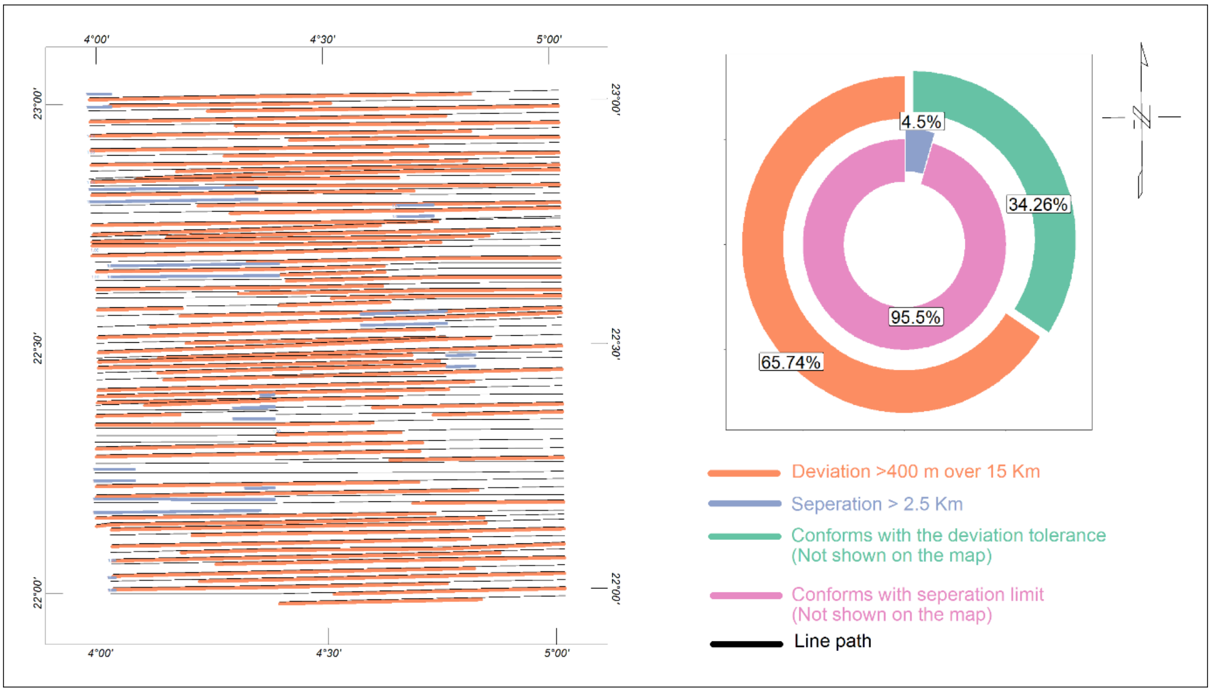


Figure 3-6 The flight segments where the separation and the deviation tolerances were respected, and the proportion of the segments that conform with the two parameters.

3.3.6. Conclusions

The quality control of the airborne data of the Silet region shows that the raw data, despite being part of an old survey, is preserved in a good condition and can be used for modern studies such as the one this dissertation is intended for. After removing the erroneous values and smoothing unusual values in the data using non-linear filters, we can reliably say that it is ready to undergo the processing scheme that is intended to improve the quality of the data and render it usable for geological interpretation. This scheme is further discussed in the following sections.

3.4. Gridding data

Until now, we have only dealt with the data statistically and abstractly and this is, usually, a good way to give insights into the quality of the data and its distribution. However, for interpretation purposes, this approach is lacking because abstract presentations are hard to precept and, therefore, hard to comprehend. To overcome this limitation, we changed how the data is presented by transforming the numeric data into a graphical presentation.

The graphical presentation is obtained by constructing a squared grid map whose nodes are established by the intersection of parallel lines of the latitude and the longitude lines. The parallel lines also divide the grid into columns and rows of small components called *“grid cells”*. These cells are equally sized and are filled with a unique value that was calculated by the interpolation of the nearest numeric data points. The grids are also color-filled which gives a visual attribute to the grid as the color corresponds to the value of the cell.

The purpose of gridding data is that it gives the congested numeric presentation an aesthetic visual context which renders the data more natural for the human mind to comprehend. Consequently, this makes it easier to identify the desired structures on a grid map without resorting to the numeric analyses that usually accompany the numeric databases. Another purpose to grid the data is that the filters, that are applied to enhance the quality of the data, are usually applied on the grid maps rather than on the database. This is because it is computationally efficient to do so and it is much easier to get a grab of the changes made by the filters.

There are a lot of algorithms for gridding data, and choosing one over the others all depends on the dispersion of the data points. In our study area, the data points are dispersed across parallel lines. For this reason, the *“Minimum curvature”* was used to grid our data.

3.4.1. Minimum curvature method

This interpolation method was described in (Briggs, 1974; Swain, 1976). The method starts by generating a coarse grid whose cell size is, usually, 8 times the cell size that the user wants. Then, it computes the values at the nodes by the usage of the *“Inverse Distance Average”* method. This method assumes that closer points are more related to the node values than the further points. Thus, it calculates the weighted average of all the data points within a specified search distance in a way that closer points have bigger weight values. In the case where there are no data points near the node, the average of all data is taken as the node value.

Following the calculation of the values of the nodes, an iterative method is used to fit the rest of the grid surface so that it matches the values of the data points as much as possible. This step is crucial in the gridding process because increasing the number of iterations will produce the smoothest possible surface that will better fit the data points but at the expense of increasing the processing time.

After an acceptable fit is achieved, the cells of the fitted grid are, by turn, divided into 2 and the same procedure is repeated to calculate the values of the nodes. The processes of calculating, fitting, and dividing are iterated until the cell size of the fitted grid matches the predefined cell size.

Minimum Curvature is a widely used gridding tool for potential field data because it generates the smoothest possible surface for a given data compared to other methods. As a general rule, the cell size of the grid is taken as ½ or ¼ flight spacing (Geosoft Inc, 2014).

3.4.2. Shaded map

The shaded map algorithm gives the ability to associate a grid map with an artificial illumination source. This helps to accentuate all the anomalies which are perpendicularly oriented to the illumination vector by way of granting them the texture of a real topographic surface. Depending on the desired anomalies, the interpreter can change the elevation and the azimuth of the illumination source because these two parameters control which direction of anomalies are highlighted and which is not.

3.5. Gridding the raw data of Silet

3.5.1. Airborne magnetic data

In Figure 3-7, we can see the shaded gridded map of the airborne magnetic data in the Silet region. Groune (2019) mentioned that the typical values for the observed magnetic field in western Hoggar, where our study area is located, should be around 36000 nT. However, in the figure below we can see that our data exhibit magnetic values that are dispersed around the value 34000 nT. These values, as we explained in section 3.2.5, is due to the corrections brought about by the AeroService company where they added a reference value of 34000 nT to the magnetic records (Allek, 2005).

A linear strong magnetic trend is observed near the eastern edge of our study area. This trend is related to regional contact *"4°50"* that passes through the region and is considered to be a natural deposit for minerals. This explains the strong values of the strong values of the magnetic anomalies.

The grid also shows the presence of a high-frequency component noise which appears as linear strips in the direction of the flight path. This noise is called the *"flight-line noise."* Due to the distortion effect introduced by this noise, this noise should be removed to ensure an accurate geological interpretation. In the subsequent sections, we demonstrate the most practiced methodology to remove this noise.

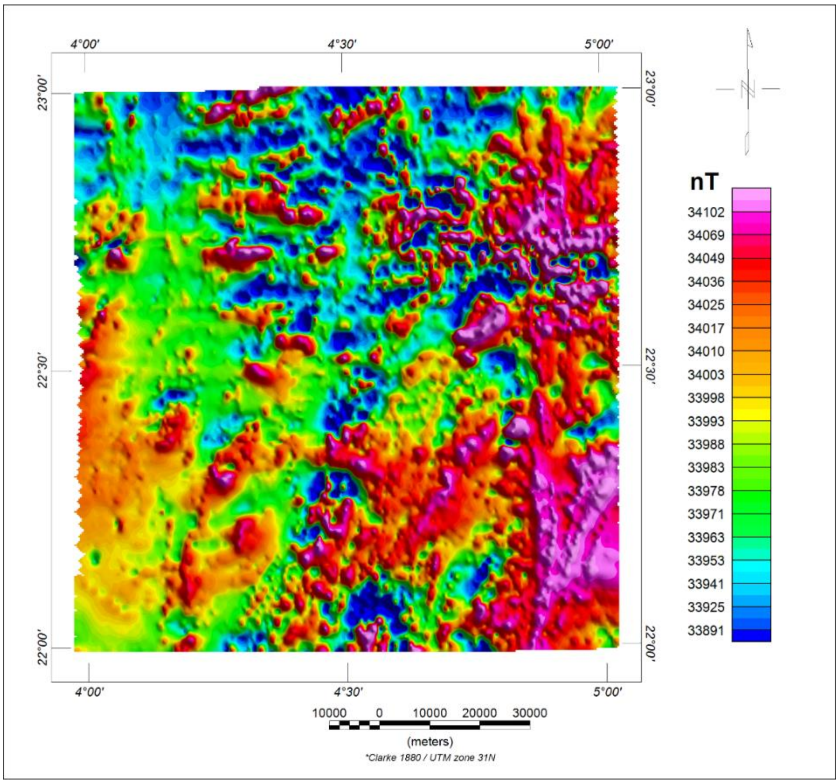


Figure 3-7 Shaded grid of the raw airborne magnetic data of Silet

3.5.2. Total count channel

The figure below is the shaded grid for the gamma-ray total count channel which is expressed in Count Per Second (cps). The figure shows that, contrarily to the grid of the magnetic data, the total count channel is greatly affected by the flight line noise. This noise is mainly due to the interference of the atmospheric radon layer between the ground surface and the detector as well as the fluctuation of the flight height.

The values of the total count grid range from a minimum value of 259 cps to a maximum value of 1212 cps, and the distribution of the data appear to be around the value 576 (i.e., mean value) cps with a spread-out (i.e., standard deviation) value of 87.

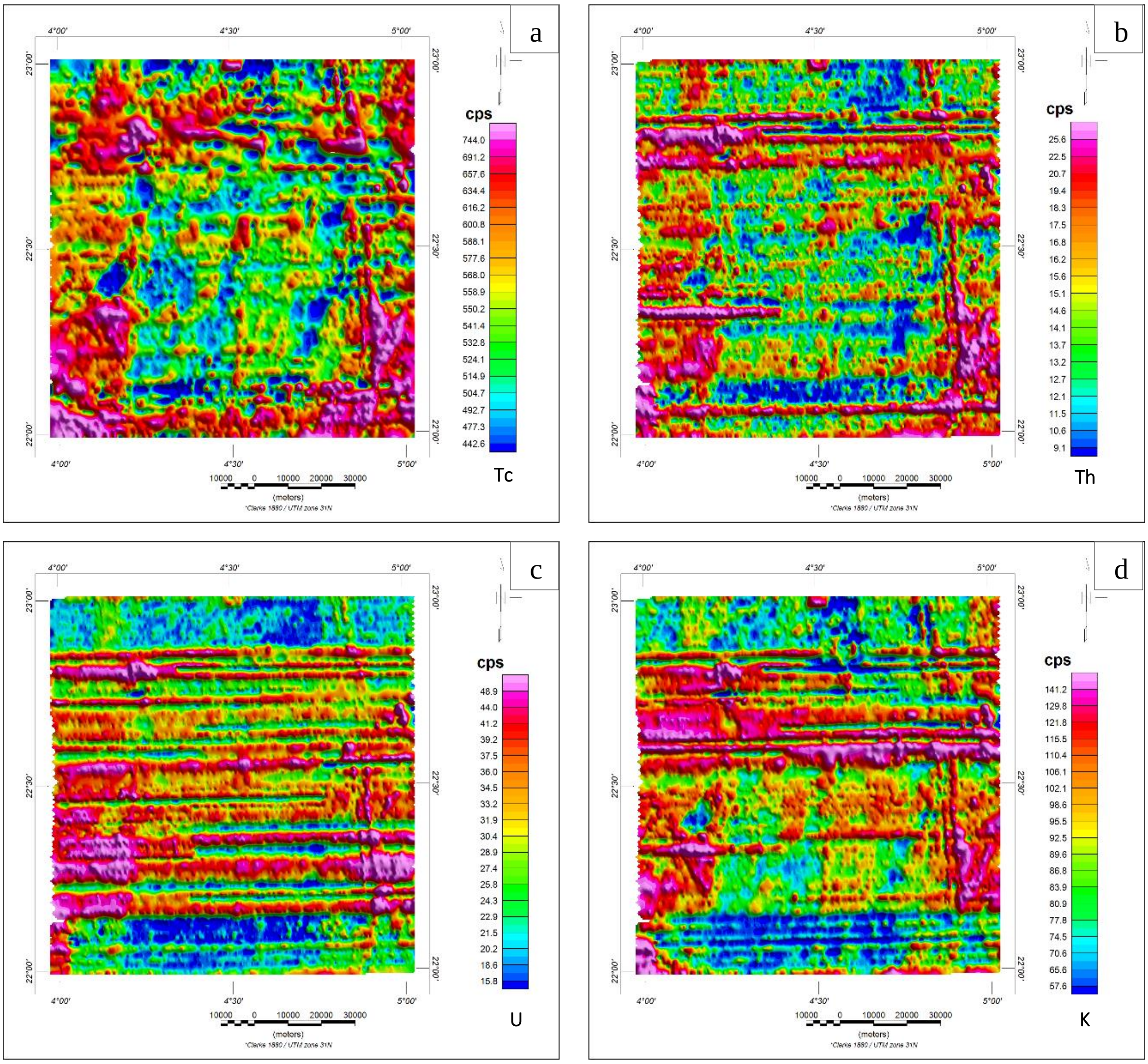


Figure 3-8 Shaded grids of the raw gamma ray spectrometry channels. a) Total count channel; b) Thorium channel; c) Uranium channel; d) Potassium channel.

Despite being affected by the flight line noise, the grid map is still able to locate some of the geological units which are characterized by big natural radioactivity values (>690 cps). These are the Taourirt granite, the Imezzarene granitoid complex in the southwestern area of the map, and the metamorphic sequences which are located in the vicinity of the contact “4°50”. They are observed in the southeastern parts of the area and span along the geological contact until the eastern middle area.

3.5.3. Thorium channel

Figure 3-8.b shows the shaded grid of the thorium channel. Similar to the total count channel, this channel is also affected by the flight noise, and it is observable as linear trends in the direction E-W. Values-wise, they range from a minimum value of 3 cps to a maximum value of 67 cps, and they are averaged at the value of 16 cps with a standard deviation of 5.

The flight noise greatly distorts the shape and the distribution of anomalies and gives them an elongated shape that would hinder giving them a plausible geological interpretation. Therefore, the grid in its current state cannot be used for interpretation purposes.

3.5.4. Uranium channel

Figure 3-8.c shows the shaded grid of the uranium channel. Among all the spectrometric channels, the uranium grid is the most affected by flight noise. This noise strongly affects the middle portion of the grid of the area. For the grid values, they range from a minimum value of 7 cps to a maximum value of 121 cps. The values of the mean and the degree of the spread-out of the grid values are valued as 31 cps and 10 respectively.

3.5.5. Potassium channel

In Figure 3-8.d, we can see the shaded grid of the potassium channel. It is also affected by flight noise, which is especially observable in the upper area of the grid. Statistically, the values of the potassium grid range from a minimum value of 13 cps to a maximum value of 288 cps, and the mean value is estimated to be 96 cps with a standard deviation value of 24.

3.6. Processing the airborne magnetic data

3.6.1. Subtracting the internal magnetic field IGRF

This step is applied to magnetic data, and it has the purpose of deducting the core field component of the geomagnetic field to calculate the anomaly field. As explained before, the core field can be approximated by the IGRF model. Thus, we can calculate the anomaly field by subtracting the IGRF

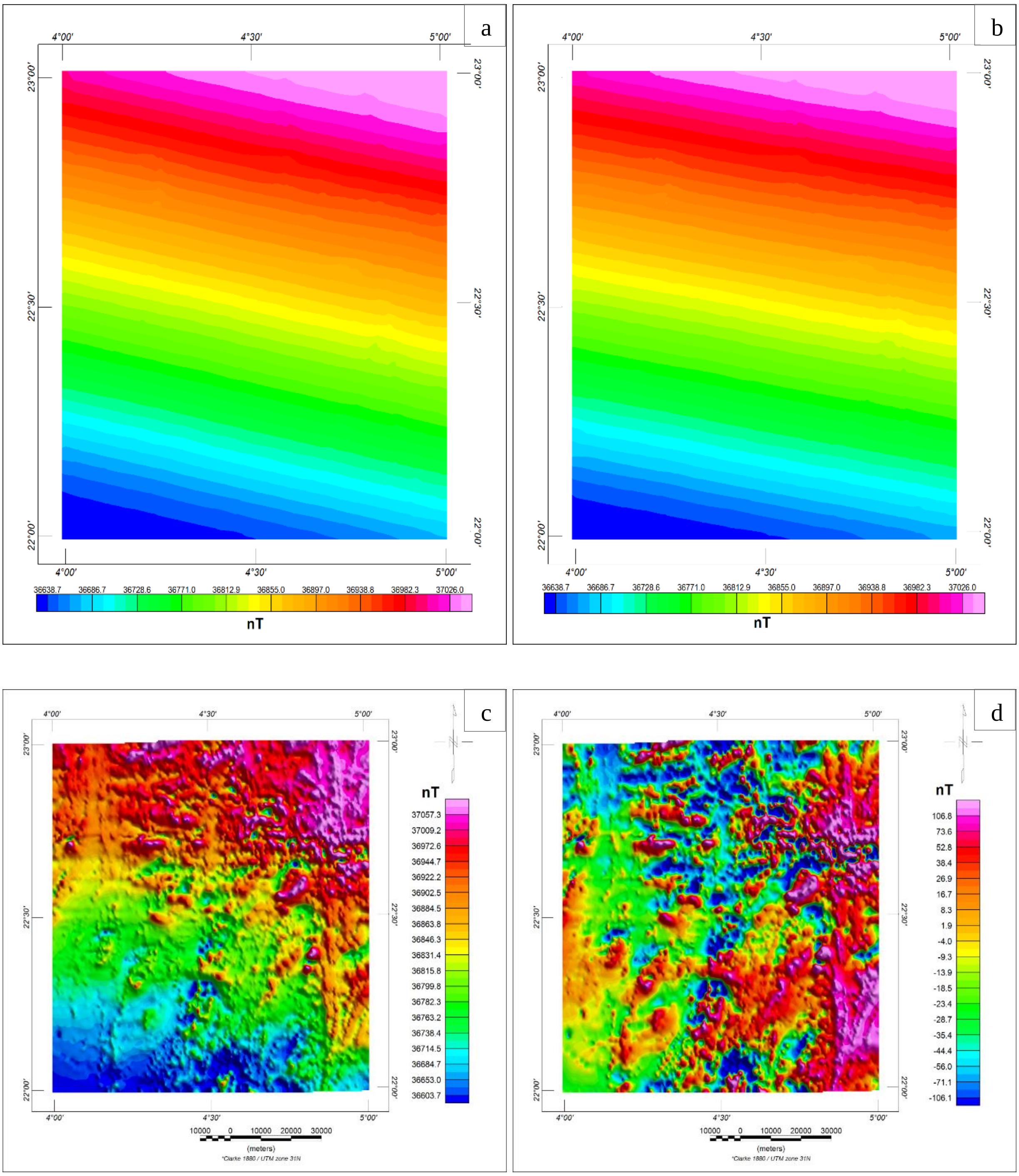


Figure 3-9 Shows the procedure to retrieve the raw airborne magnetic data and to remove the geomagnetic field. a) Calculated IGRF 1965; b) Calculated DGRF 1970; c) Obtained grid after removing the corrections carried out by AeroService; d) Anomaly field after subtracting the DGRF 1970.

model from the observed field. However, this simple approach cannot be applied directly to our data. As discussed by (Allek, 2005), the retrieved data do not reflect the true measured magnetic field at that time but rather represent the measured magnetic field after subtracting the IGRF 1965 and adding a constant value of 34000 nT.

The first step of calculating the anomaly field starts by retrieving the true measured magnetic field by reversing the corrections brought about by AeroService company. This was carried out by subtracting a value of 34000 nT and adding the IGRF1965 to the raw data (see Figure 3-9.a). The obtained grid after reversing the corrections of the AeroService company is shown in Figure 3-9.c. The next step of calculating the anomaly field consists of calculating the DGRF 70 (shown in Figure 3-9.b) and then subtracting its values from the retrieved magnetic field. The final product of this step is shown in Figure 3-9.d (Allek, 2005).

3.6.2. Levelling the airborne magnetic data

The levelling step targets the flight-line noise which, as previously demonstrated, appears as linear strips in the direction of the flight line. The following section presents the most used methods to minimize the effect of this noise.

3.6.2.1. Tie line levelling

The purpose of the leveling method is to minimize the difference between the readings of the flight and the traverse lines at the intersection points. This difference is caused by the short temporal changes of the magnetic field, poor positioning of the intersection points, the height difference between the flight and traverse lines, instrumental interference, and random noise. The most employed method to carry out this levelling approach is to develop a (2nd or 3rd order) polynomial surface whose coefficients are calculated in a way that would minimize the root mean square of the differences. Unfortunately, the limited number of traverse lines in our study area (only 8 tie lines grouped into 3 major traverse lines, see Figure 3-2) and the low coverage density for these lines prevent us from releveling the magnetic data in the Silet region.

3.6.2.2. Microlevelling

We finalize the processing of the magnetic data by applying a microlevelling filter or what is known as the decorrugation filter. This filter aims to remove any residual flight noise which persists after applying the previously mentioned corrections. Many algorithms were developed to carry out the microlevelling procedure. For our data, the Paterson, Grant and Watson (PGW) algorithm was chosen (Paterson et al., 2003) due to the inherent ability of the PGW to remove a significant portion of the residual noise, and at the same time, preserve the geological signal.

The PGW algorithm extracts a decorrugation noise channel by running a sixth-order high-pass Butterworth filter combined with a directional filter perpendicular to the flight line direction. This first process is necessary because it allows the extraction of the flight-line noise that contains some of the high-frequency components of the geological signal. The PGW algorithm then applies an amplitude limiting and low pass filter to the extracted noise to remove any information related to the geological signal. Finally, the algorithm finishes the filtering process by subtracting the noise channel from the original data. The final product of the last step is called a microlevelled data.

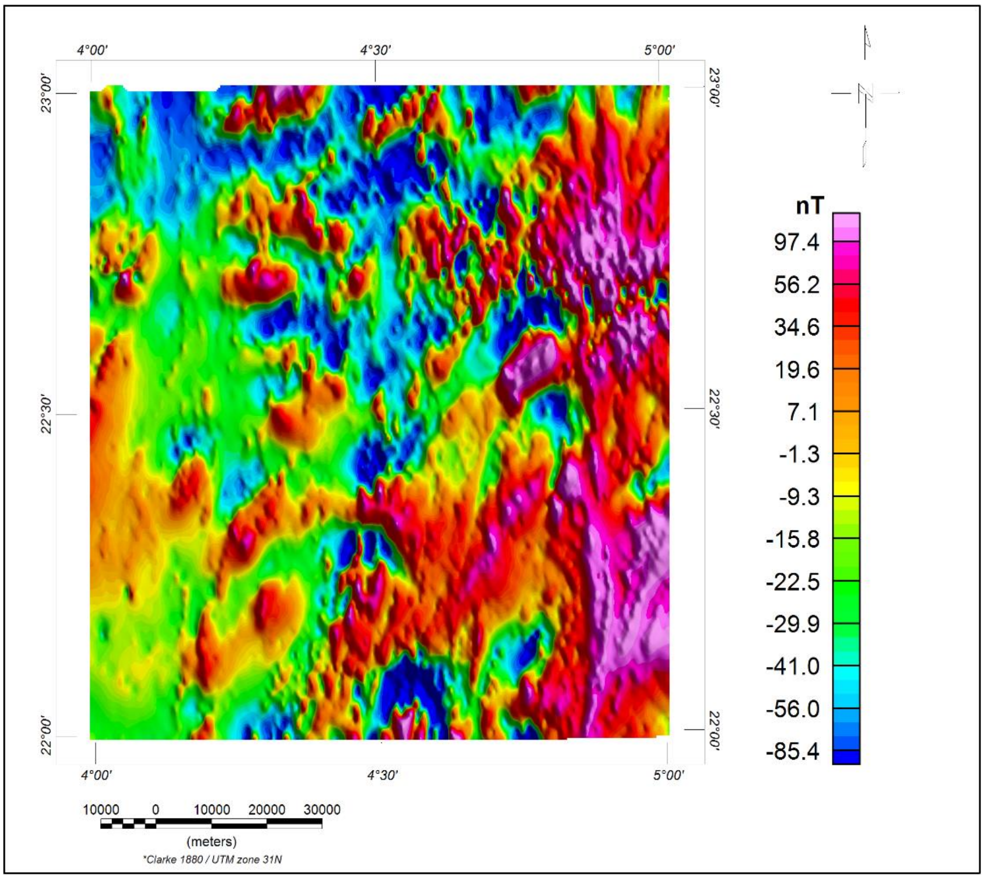


Figure 3-10 Shaded grid of the anomaly field after applying the PGW algorithm.

Oasis montaj offers an interactive window to allow a manual setting of the amplitude limiting and the low pass filter parameters. The default values are set to the standard deviation of the noise grid for the amplitude limiting filter, and a filter width equal to five times the filter line spacing for the low pass filter (Paterson et al., 2003). These values, however, did not seem to fully eliminate the persistent noise for our data, so a manual inspection of the grid noise was carried out to choose the optimal parameters.

Taking into consideration that the geological anomalies are characterized by higher amplitude and shorter wavelength than the flight line noise, we executed numerous tests to estimate the maximum amplitude and the minimum wavelength that characterizes the flight line noise. The final grid of the microlevelled data is presented in Figure 3-10.

From Figure 3-10, we can see that the PGW eliminated the majority of the flight line noise. This can be observable in Figure 3-9.d as linear trends in the flight line direction. The filter seems also to mildly distort the shape of the geological anomalies in the N-S direction.

Table 3-3 Comparison between the distribution of the aeromagnetic data before and after applying PGW.

	Minimum	*Mean*	*Median*	*Maximum*	*Standard deviation*
Before applying PGW	-569	-0.52	-7.94	615.26	60.16
After applying PGW	-303	-0.53	-7.93	350	50.66

To confirm if the algorithm affected the distribution of the airborne magnetic data, we performed a statistical analysis to ensure the preservation of the statistical distribution of the magnetic anomalies. The analysis is shown in Table 3-3. From the table above, we can see that the algorithm did not affect the distribution of the data as the value of the mean, the median and the standard deviation were practically preserved before and after applying the filter. However, the filter seems to eliminate some of the higher amplitude anomalies which were associated with the minimum and the maximum value.

3.6.3. Reduction to pole transformation (RTP)

As mentioned before, the magnetic anomaly due to a geological body that is not located in a polar region has an asymmetric shape. This distortion is introduced because of the orientation of the magnetization vector.

The purpose of the reduction-to-pole transformation is to eliminate the effect of the magnetization vector and transform the magnetic response of a particular geological body into the response of the same body if it were measured in a polar region. The vertical magnetization vector in the polar regions simplifies the magnetic anomalies and gives them a symmetric shape. This transformation also renders the geological interpretation of the magnetic data easier due to the symmetric shape of magnetic anomalies that would position the peak of these anomalies vertically to the causative body. Therefore, this transformation would reduce mislocation errors caused by the asymmetric shape of the original response.

To generate the grid of the reduced-to-pole, three pieces of information are required. They are the inclination, the declination and the correction factor I_a (Macleod et al., 1993). This correction factor is introduced to correct the amplitude of the transformed data.

The inclination and the declination channels are obtained along with the DGRF 70 channel. The values of these parameters change in terms of spatial coordinates (Ansari & Alamdar, 2009). Since our study area is not very large, we can safely assume that the inclination and the declination values remain practically invariable. Therefore, we used a single value (the average value) whether for the inclination or for the declination.

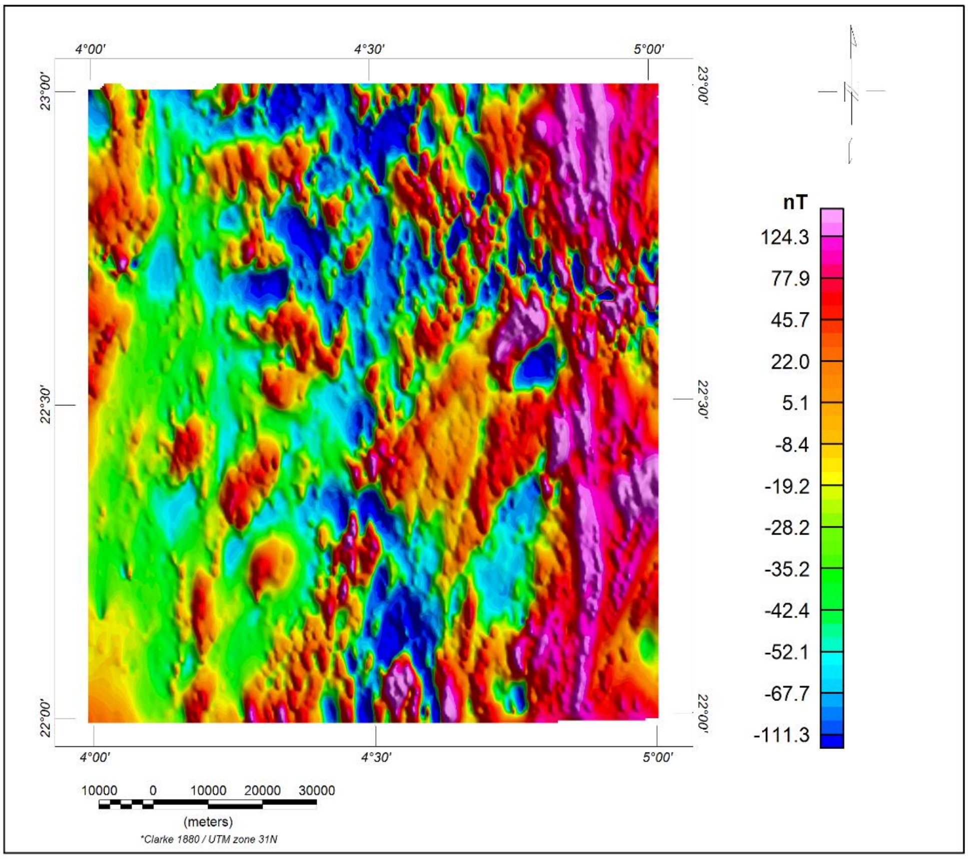


Figure 3-11 Shaded grid of the anomaly field after applying the reduction to pole transformation.

To estimate the value of the correction factor I_a, Bourans (2001) executed multiple tests in the Hoggar region. These tests conclude that a value of 50 would completely remove the amplification effect of the inclination. Therefore, we used a value of 50 as a correction factor to generate the reduced-to-pole grid for the study area.

As a conclusion, we used the respective values of 27.8, -4.2 and 50 for the inclination, the declination and the correction factor to apply the reduction to pole transformation. The outcome grid of the anomaly field after being reduced-to-pole is shown in Figure 3-11

From Figure 3-11, we can see that the magnetic anomalies are slightly shifted towards the north direction. As previously discussed, magnetic anomalies before applying the transformation of the reduction-to-pole would have an asymmetric shape, which would introduce location errors. As a result, we can say that the magnetic anomalies are shifted towards their exact location after applying the reduction-to-pole filter. The figure also reveals a strong linear magnetic trend that runs along the 4°50 meridian in a north-south direction. This trend is explained by the existence of the major fault that separates the western Hoggar from the central Hoggar. This major tectonic incident is a natural deposit for the basic/ultra-basic intrusions which are known to have a strong positive magnetic signature.

3.6.4. Apparent susceptibility

The apparent susceptibility filter combines a reduction-to-pole, downward continuation to the source depth and a coefficient to correct the geometric effect. The advantage of this filter is that it gives additional information about the magnetization of the geological structures. This may better help in distinguishing between the causative bodies. However, this advantage can only be brought about if the observed field conforms with the assumptions, we discussed in section 2.5.5.

Figure 3-12 shows the distribution of the apparent susceptibility in the region of Silet. It shows a strong linear trend that runs along the major fault 4°50. These strong susceptibility values are associated with the basic/ ultra-basic magmatic formations of the western Hoggar, such as migmatites and granulite facies.

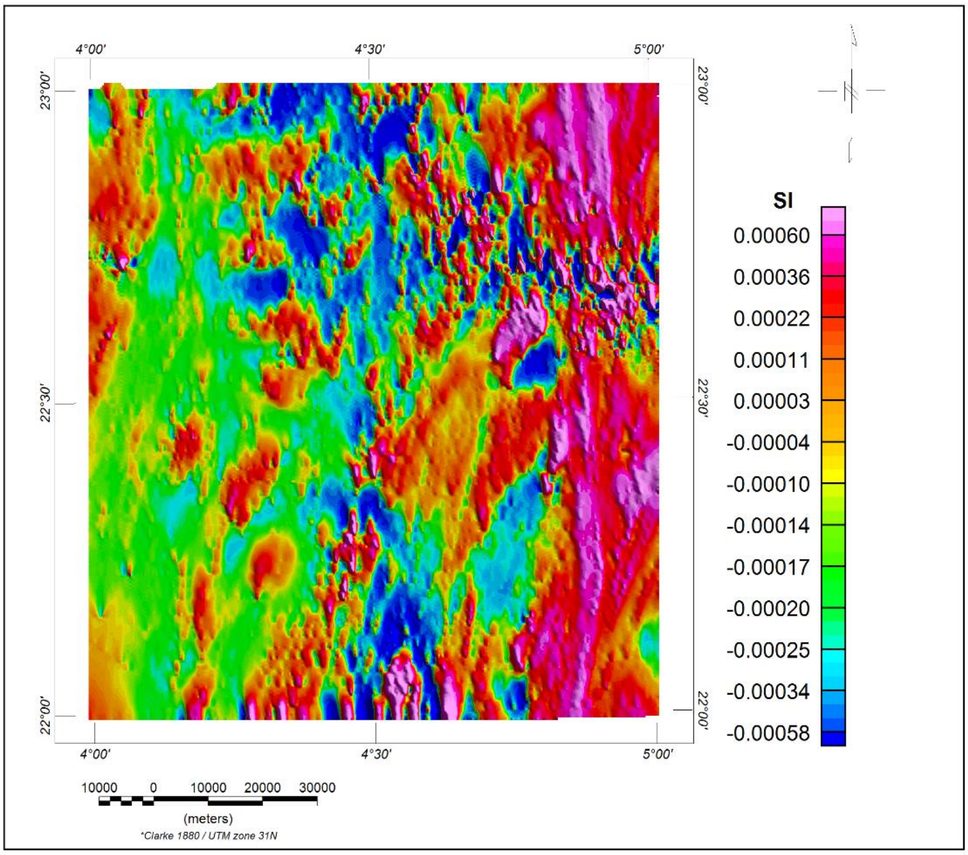


Figure 3-12 Shaded grid of the apparent susceptibility.

3.6.5. Pseudo-gravity and apparent density

Generating an apparent density grid requires as an input an observed gravity grid. Unfortunately, that was not available for our study area. However, to work around this problem, we used a transformation called "*Pseudo gravity*."

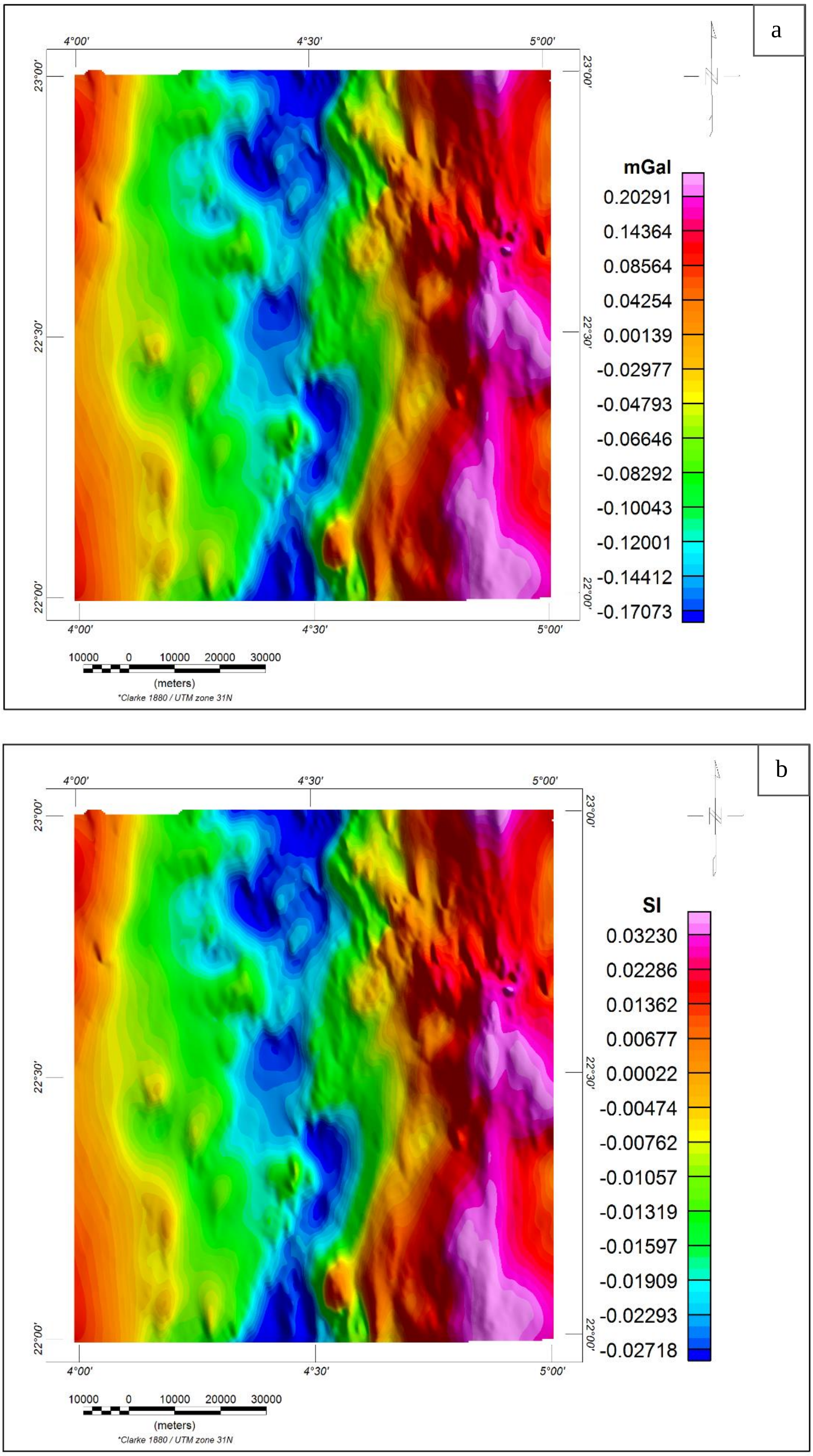


Figure 3-13 a) distribution of the pseudo gravity. b) distribution of the apparent density.

The pseudo gravity filter is used to calculate a gravity-like response that would be obtained if the body's magnetism were replaced with same density distribution using the *"Poisson"* equation. Considering that a geological body possesses a homogenous distribution of density and magnetization vector, the Poisson equation can transform the magnetic signature of that body into its gravity signature. Obtaining a pseudo-observed gravity distribution for our study area gives us the ability to produce the apparent density grid. The apparent density filter starts by assuming an earth model with fixed thickness and varying density. Then, it poses several assumptions concerning the geometry of causative bodies. It assumes that the response is generated from an assembly of multiple vertical prisms with infinite depth. These prisms have a horizontal, square-end surface with a dimension equal to the cell size of the input grid. Finally, the operator of the filter calculates an approximation of the density

Figure 3-13.a illustrates the distribution of the pseudo gravity values in the region of Silet. After applying the downward continuation filter to be close to the ground surface, we generated the gravity values while assuming while taking a density contrast of 1 g/cm^3, a magnetization of 0.5 gauss, inclination of 27.8 and declination of -4.2.Figure 3-13.b shows the distribution of the apparent density in the study area. It was produced while assuming: an earth's model with a fixed thickness of 150 m, a response due to a collection of prisms with a square horizontal surface of 350m and a background density contrast of 0 g/cm^3.

3.6.6. Contributions of the airborne magnetic data in the distinction between lithological units

In this section, we are going to discuss the potential of the airborne magnetic data products, namely the reduced-to-pole, the apparent susceptibility and the apparent density to discern between the lithological units in the region of Silet.

To achieve that, we first built a database that contains the RTP, apparent density and apparent susceptibility channels which are considered the numerical part of the data, and their corresponding lithology which is considered the categorical part of the data. After that, we grouped the data according to the lithology channel to be able to calculate the descriptive variables (i.e., minimum, 1st quartile, mean, median, 3rd quartile and maximum) for each lithological unit. This may help to verify if the sole usage of the aeromagnetic data can contribute to the classification of lithology. Finally, we plot results in the form of a boxplot to better understand the distribution of each numerical variable according to the lithology.

Since the numeric data are not expressed in the same measurement unit, we mean standardized them before calculating the descriptive variables to assure a more accurate comparison between the lithological units.

Figure 3-14 shows the distribution of the reduced-to-pole, apparent susceptibility and apparent density values for each lithological unit in the region of Silet. In this figure, to better explain the impact of the airborne magnetic data products, we employed 2 criteria to organize the lithological units in the x-axis. The first one, according to their class (i.e., metamorphic, igneous volcanic/ plutonic), and according to their rock family (i.e., granite, gabbro, diorite…. etc.). The second criterion is organizing the rocks of the same class and family from the oldest (left) to the youngest (right).

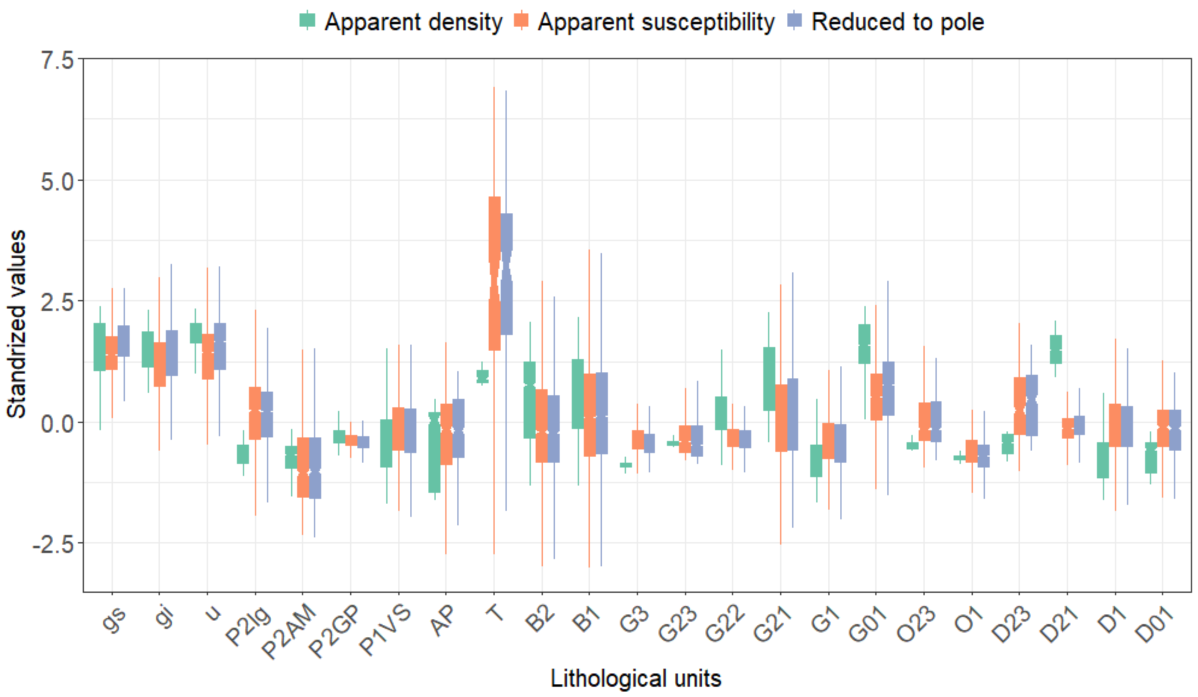


Figure 3-14 Values distributions of the reduced to pole and its products according to the lithological units in Silet region.

The granite family (labels start with the letter *"G"*) values of the RTP and the apparent susceptibility seem to vary in an overlapped range across all the granite types. On other hand, the values of the apparent density appear to vary volatilely and their distribution varies from one granite to another. Therefore, we can say that for this family of rocks in our study area, using the RTP and the apparent susceptibility may not be regarded as a valuable tool to discern between them. The apparent density, however, shows a promising tool that can help in the classification process of this family.

The diorite family and the gabbro family (labels start with the letter *"D"* and the letter *"O"* respectively) share the same range of values for the RTP; apparent susceptibility and apparent density. Hence, using the products of aeromagnetic data to classify these rock families may not accommodate any value.

For the remaining rock families, the airborne magnetic data are not able to create any pattern that might be used as a classification basis to distinguish between these rocks. The only exception to this is the *"Trachyte,"* an extrusive igneous rock. This rock, labeled as the letter *"T,"* seems to have a distinct magnetic signature, as it was characterized by the highest RTP and apparent susceptibility values among all other rocks. Thus, using only the products of the airborne magnetic data might facilitate setting this rock apart from other rocks.

3.7. Processing the airborne gamma-ray spectrometry data

AeroService team did not apply any form of processing to the airborne gamma-ray spectrometry data. Hence the data acquired is extremely noisy and require a careful and precise methodology to clean the undesirable signals.

In this section, we describe the procedures employed to process airborne gamma-ray spectrometry. These procedures have the purpose of removing any none geology-related signals that harm the quality of the data and decrease the accuracy of the geological interpretations. The procedures also allow the estimation of the ground concentrations of the radioelements.

3.7.1. Dead time correction

Dead time is defined as the time during which the spectrometer is not actively recording radioactive pulses. The reason for that is the detector needs some time to process one pulse, and while doing so, all other incoming pulses are automatically rejected. Dead time can also be defined as the difference between the sample accumulation time and the time when the detector is actively recording incoming pulses, called *“Live time.”*

The effect of this correction is usually small, but in areas with high radioactivity, it can be substantial. However, it is always required to carry out this correction because the pulses that are rejected during the dead time are considered as lost counts. This subsequently leads to the underestimation of the ground concentration of the radioelements.

The dead time effect can be corrected as follows:

$$N = \frac{n}{1-Ct} \tag{3-1}$$

Where: N = corrected count rate (counts/sec);

C = total count rate over all channels (counts/sec);

n = observed count rate (counts/sec);

t = the equipment dead time per pulse.

A typical dead time of most instruments falls in the range of 5-15 µs/pulse. In our case, taking into account the old model of the spectrometers used to survey the Hoggar region, we picked 15 µs/pulse to remove the effect of the dead time.

3.7.2. Background corrections

3.7.2.1. Cosmic and aircraft background

The cosmic background originates from the interaction of high-energy cosmic-ray particles with the air, the aircraft and the detector. The count rate of this background component can be determined using a four-channel detector, one of its channels keeps a record of all the incoming pulses with energy over 3 Mev.

The aircraft background originates from the incident radiations that are caused by the aircraft body and its instruments. It has a constant value and remains unchanged throughout the survey.

At a sufficiently high flight where the radon and terrestrial radiations are minimal, the cosmic and the aircraft background components in a given channel have a linear relation with the count rate in that channel (IAEA, 2003). This relation can be expressed as follows:

$$N = a + bC \tag{3-2}$$

Where: N = aircraft and cosmic background count rate in a given channel;

C = cosmic channel count rate;

a = aircraft background count rate for the same channel;

b = cosmic background in the channel normalized to unit count in the cosmic channel.

The aircraft background and the cosmic background of a given channel (i.e., a and b factors in equation 3-2) can be determined empirically. This can be done by flying a series of flights at different heights over a water surface. The recommended flights range from 1500 m to 3500 m at 300 m intervals (Grasty & Minty, 1995).

3.7.2.2. Radon background

The removal of the radon background is an important step in the processing of airborne gamma-ray spectrometry because it aims to remove the contribution of the atmospheric radon and its daughters from the various channels of the detector. The excess count rate caused by the radon can lead to the bad estimation of the ground concentrations of the natural radioelements, especially the uranium, as it shares the same daughter element with the radon, namely the bismuth Bi^{214}.

The most used method to estimate the radon background is by using an additional detector directed upwardly. This detector is partially shielded from terrestrial radiation to assure a better separation between the ground and the atmospheric radiation (IAEA, 2003).

The data from the upward-looking detector is used to estimate the background components of the atmospheric radon in the various channels of the downward detector. This can be achieved, firstly, by extracting a relation between the count rate of the radon in the upward and the downward detector. This can be done by performing a series of high-altitude flights over the sea, where the interference of terrestrial radiations is minimized. Under such conditions, the count rates observed in the detectors are mainly due to the background components, namely the cosmic, the aircraft and the radon background.

Lastly, after removing the aircraft and the cosmic components of the background by following the method discussed previously, and owing to the linear relationship between the count rate of the radon component in the uranium channel between the upward and the downward detector, the background radon for all channel can be determined, as the radon background specific to every channel in the downward detector are linearly related (Seligman, 1992).

In the final report of AeroService, we could not find any reference that indicates the usage of one of the above-mentioned method. Therefore, we concluded that the background corrections of the airborne gammy ray spectrometry data of the region of Silet were not implemented.

In their work, to solve the problem of the absence of any correction that targets the background components, Groune (2019) adopted a statistical methodology to estimate the contribution of the background to the data of Hoggar. The methodology in Groune's study consists of calculating a first-order trend of the raw data using the least square approach. Knowing the high level of background noise in the data of the Hoggar, the researcher estimated that 20 % of the trend can be taken as background noise.

Putting into consideration the satisfactory results obtained in their work, and because the region of Silet is located within the Hoggar shield, we decided to employ the same methodology to estimate and remove the background noise.

3.7.3. Stripping correction

In reality, the count rate of the channel of one of the radioelements Th, U and K can have interference from the channel of the two others. The proportions of the observed count rates of a given radioelement that are characteristic of another are called the *"Stripping ratio."*

The interference between the count rates of the channels happens for several reasons. For example, due to the Compton scattering, or the incomplete absorption of the high energy radiations, the photons issued from a pure source of Th can be detected in the lower energy channels of U and K. Contrarily, high energy photons originated from a pure source of U can be detected in the channel of the Th. When the photons are characterized by low energy, however, they can be detected in the low energy channel of K(Grasty & Minty, 1995).

To distinguish between the stripping ratios, a naming methodology was adopted. α, β and γ are used to denote the interference of the high energy channels in the low energy channels. α is the count rate of Th in the U channel, β is the count rate of Th in the K channel and γ is the count rate of the U in the K channel.

a, b and g are used to denote the interference of the low energy channels in the high energy channels, where a is the count rate of U in the Th channel, b is Th count rate of K in the Th channel and g is the count rate of the K in the U channel.

The stripping correction is applied as follows:

$$\begin{aligned} Th_{corr} &= \frac{Th_{obs}-aU_{obs}}{1-\alpha a} \\ U_{corr} &= \frac{U_{obs}-\alpha Th_{obs}}{1-\alpha a} \\ K_{corr} &= \frac{(1-\alpha a)K_{obs}+(\alpha\gamma-\beta)Th_{corr}+(\alpha\beta-\gamma)U_{corr}}{1-\alpha a} \end{aligned} \tag{3-3}$$

The ratios a, b and g have a small effect on the count rate, so they are usually neglected in the correction process (IAEA, 2003). Therefore, the corrections above can be rewritten as follows:

$$\begin{aligned} U_{corr} &= U_{obs} - \alpha Th_{obs} \\ K_{corr} &= K_{obs} - \beta Th_{corr} - \gamma U_{corr} \end{aligned} \tag{3-4}$$

The stripping ratios are obtained by employing special calibration pads. These pads are made of concrete and have a square dimension with known concentrations of Th, U and K. The calibration procedures require the usage of four pads, three of which are used to estimate the stripping ratio in the Th, U and K; the last one is used to estimate the background. Ideally, the calibration should be executed in the aircraft. If multiple detectors are used, each detector should be calibrated individually. The calibration ratios can be then calculated by averaging the results of both detectors (IAEA, 2003).

In airborne surveys, the stripping ratios should also be corrected from the survey height because their values increase with the altitude. In Table 3-4, we show how the stripping ratios increase in function of the altitude.

Table 3-4 The increase of the stripping ratio according to the altitude. (The increase of ratios a, b and g is small, and therefore can be neglected.)

Stripping ratio	*Increase per meter*
α	0.00049
β	0.00065
γ	0.00069

No reference to the calibration of the stripping ratio is mentioned by AeroService. Consequently, we decided to use stripping ratios estimated from a calibration procedure in which the same detector configuration (i.e. same crystal distribution) is used (Groune, 2019). The stripping ratios used in this study are:

$$\alpha = 0.45$$

$$\beta = 0.59$$

$$\gamma = 0.94$$

3.7.4. Height correction

Because of the non-uniformity of the topographic surface, the altitude of the aircraft keeps drifting away from the nominal ground clearance. Hence, it is crucial to bring all the heights to the nominal altitude. The count rate of the detector's channels can be corrected for the height effect using the following equation:

$$n = n_0 e^{\mu(H_{stp} - h)} \tag{3-5}$$

Where: n = corrected count rate of a given channel for the nominal height;

n_0 = observed count rate of a given channel after applying the background and stripping corrections;

μ = attenuation coefficient characteristic to each channel;

H_{stp} = height above ground level converted to equivalent height at STP conditions.

It should be noted that before applying the height correction, the flight height must be converted to the equivalent height standard temperature and pressure conditions (STP). The necessity to do this rises from the fact that both temperature and pressure affect the attenuation properties of the air (IAEA, 2003). The conversion can be executed using the following equation:

$$H_{stp} = \frac{273.15 \times P \times H_{obs}}{(T + 273.15) \times (101.325)} \tag{3-6}$$

Where: H_{stp} = equivalent height at the STP conditions;

H_{obs} = flight height above ground level;

T = air temperature;

P = air pressure.

To complete this conversion, the temperature and the pressure data should be available. Unfortunately, that was not the case for the survey in Silet. Therefore, knowing that the majority of the survey was carried out between 10 am and 12 am in late October of 1971, we decided to assign a fixed value of 25° for the temperature.

For the pressure data, they can be obtained by applying the following equation:

$$P = 101.325\, e^{-\frac{H}{8581}} \tag{3-7}$$

Where H is the barometric altitude. It is equal to the sum of the flight altitude plus the topography altitude. The latter was obtained from the digital elevation model (DEM).

As mentioned before, each channel has its appropriate absorption coefficient. Their values are:

$$\mu_k = 6.8617 \times 10^{-3}\ m^{-1}$$
$$\mu_U = 6.3726 \times 10^{-3}\ m^{-1}$$
$$\mu_{Th} = 5.2247 \times 10^{-3}\ m^{-1}$$

These values, however, were not used, and an average value of $\mu = 6.56 \times 10^{-3}\ m^{-1}$ was used instead. This decision comes following Groune (2019) in their work. Since we could not find any passage to the coefficients used by AeroService, we also decided to use the same mean value for the absorption coefficient for all the channels.

3.7.5. Applying the processing steps on the raw airborne data of Silet

To facilitate the processing of the airborne gamma-ray spectrometry, and following the same methodology used by (Groune, 2019), we grouped all the corrections we have discussed so far into four equations, each one designed to correct the Tc, Th, U and K channels. These equations are as follows:

$$\begin{aligned} Tc_{corr} &= \lambda\ (Tc_{obs} - BC_{Tc}) \\ Th_{corr} &= \lambda\ (Th_{obs} - BC_{Tc}) \\ U_{corr} &= \lambda[(U_{obs} - BC_U) - \alpha(Th_{obs} - BC_{Th}] \\ K_{corr} &= \lambda\ [(K_{obs} - BC_K) - \beta(Th_{obs} - BC_{Th}) - \gamma(U_{obs} - BC_U)] \end{aligned} \tag{3-8}$$

Where: BC_{Tc}, BC_{Th}, BC_U and BC_K are the background components of each channel;

Tc_{obs}, Th_{obs}, U_{obs}, K_{obs} are the raw count rate related to each channel;

Tc_{corr}, Th_{corr}, U_{corr}, K_{corr} are the raw count rate related to each channel;

$\alpha, \beta,$ and γ are stripping ratios;

λ is a term to take into consideration the height correction. It is equal to $e^{\mu(H_{stp}-h)}$.

All these corrections were applied to the gridded data, and the results of the processing scheme we followed thus far are shown in Figure 3-15.

The next step in the processing scheme should have been to convert the count rates of the Th, U and K to ground concentrations. But, as clearly demonstrated in Figure 3-15, the four channels are still heavily suffering from the flight noise. This substantial flight noise, which persisted even after applying the conventional processing methods, is due mainly to the approximate approaches we opted to process the data, especially the backgrounds. Our methods, although were able to approximate the background components, they are still just a statistical approach, and they cannot replace the real calibration methods.

Similarly, to the airborne magnetic data, the application of the PGW algorithm proves to be crucial for that airborne gamma-ray spectrometer; hence, we are going to apply the PGW algorithm to remove, as much as we can, the persistent flight noise.

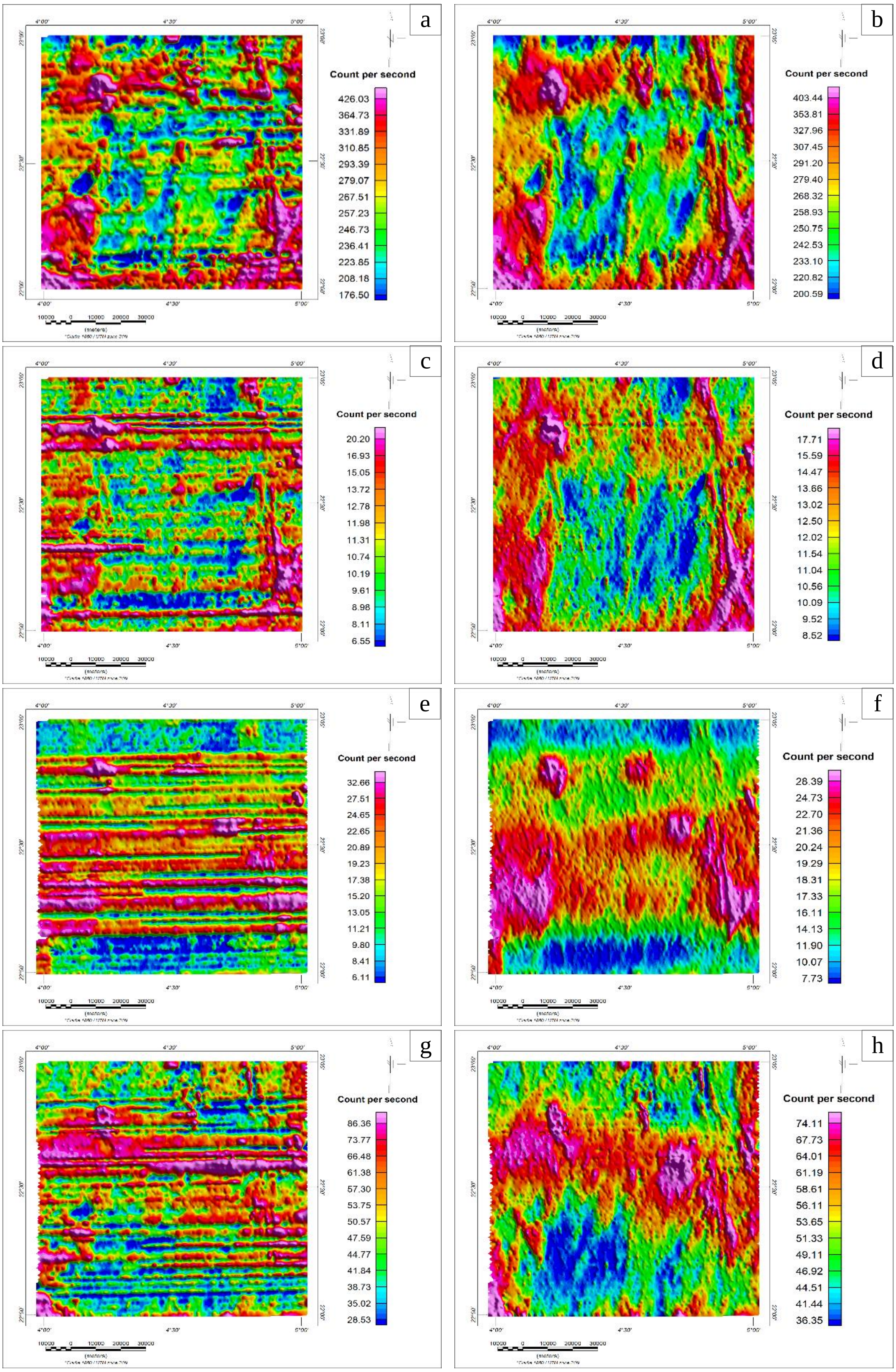


Figure 3-15 Shaded grids of the corrected count rates of the four channels before (left) applying the PGW, and after (right) applying it. a-b) total count channel before and after applying the PGW; c-d) thorium channel before and after applying the PGW; e-f) uranium channel before and after applying the PGW; g-h) potassium channel before and after applying the PGW.

In Figure 3-15, we also show the grids of the count rate of the four channels after applying the PGW algorithm. A considerable amount of the flight noise was eliminated. The grids of the airborne gamma-ray data, at their current state, can be used in the context of geological mapping. For example, the Tourirt granite can be easily identified using the grids of the total count, the thorium and the uranium channels due to their distinguished signature. The mylonite related to the Pan-African orogeny is another lithological unit that can be easily identified due to its signature in the total count channel and the thorium. It has a linear shape that runs along the 4°50 fault from the bottom left of the study area to the parallel 22°30. It should be noted, however, that a small amount of flight noise remained in the data. This is especially observable for the uranium and the potassium channels, presented respectively in Figure 3-15-f and Figure 3-15-h.

In Table 3-5, we compare the distribution of the airborne gamma-ray spectrometry data of the study area before and after applying PGW algorithm. This was achieved by comparing the minimum, the mean, the median, the standard deviation, and the maximum.

Table 3-5 Comparison of the distribution of the airborne gamma-ray spectrometry before and after applying PGW algorithm.

		Minimum	*Mean*	*Median*	*Maximum*	*Standard deviation*
Tc	Before PGW	20.97	284.46	269.35	1022.50	76.28
	After PGW	113.46	284.30	270.04	1037.25	62.30
Th	Before PGW	2.58	12.37	11.41	52.41	4.38
	After PGW	3.89	12.58	12.1	52.54	3.06
U	Before PGW	-1.94*	18.17	17.72	81.93	8.39
	After PGW	2.01	18.25	18.47	54.45	5.96
K	Before PGW	-0.88*	54.32	51.12	243.71	18.84
	After PGW	22.27	55.12	54.06	225.12	11.89

* The negative values in these channels were introduced because of the inadequate estimation of the background components.

The PGW algorithm seemed to eliminate some low amplitude anomalies, as the minimum value in all four channels was increased. The highest increase was observed in the total count as well as the potassium channel. On other hand, the algorithm had a less important effect on high amplitude anomalies. As seen in Table 3-5, excluding the uranium channel where a substantial decrease of the maximum value was observed, the maximum value for all other channels was practically the same. In addition, the mean, the median, and the standard deviation for all channels were nearly preserved after the filtering process. As a result, we can say that the effect of the PGW algorithm on the data was kept at a minimum, and the distribution of the data was, to a certain degree, conserved.

3.7.6. Evaluating the ground concentration

After finishing applying all airborne gamma ray spectrometry processing steps, the level of the flight noise was substantially attenuated. After that, the corrected data are then divided by a *"sensitivity"* coefficient to estimate the ground concentration. This gives the measured counts a direct geological relationship that is independent of the survey specifications, such as crystal volume, survey height as well as instruments used during the survey.

The sensitivity coefficients are usually determined by selecting a calibration range where the ground concentrations are already determined. After that, we survey the range by numerous flights at different altitudes, and by using the measured data from this survey, we can easily determine the sensitivity coefficients specific to each channel (IAEA, 2003). They can be calculated using the following formula:

$$S = \frac{N}{C} \tag{3-9}$$

Where: N = corrected count rate at nominal height;
C = average ground concentration;
S = sensitivity coefficient.

The ground concentrations of the calibration range are determined using a portable spectrometer, and to ensure that accurate coefficients are calculated, they should ideally be determined at the same time as the airborne survey. Estimating the coefficients within a few days of the airborne survey is also acceptable. (Seligman, 1992)

For our survey, the coefficients were not estimated using the traditional way, so we had to use an approximate formula designed to give an approximation of the ground concentrations. This formula was proposed by (Darnly, 1972), and it is written as follows:

$$\begin{aligned} eTh_{ppm} &= \frac{Th_{corr}}{F_{Th}} \\ eU_{ppm} &= \frac{U_{corr}}{F_U} \\ K_{\%} &= \frac{K_{corr}}{F_K} \end{aligned} \tag{3-10}$$

Where eTh_{ppm}, eU_{ppm} and $K_{\%}$ are the equivalent ground concentration of the radioelements, and F_{Th}, F_U, F_K are the sensitivity coefficients for each channel.

For a spectrometer with a given volume, we can calculate the coefficients using the following formula:

$$F = V.\mu \tag{3-11}$$

Where: V = crystal volume of the spectrometer;

μ = absorption coefficient specific to each channel.

The sensitivity coefficients used to convert the count rates to ground concentration for the thorium channel, uranium channel and potassium channel are 4.18, 5.1, and 5.49 respectively.

3.7.7. Calculating the ratios of the spectrometry channels

Calculating the ratios of the uranium, the thorium, and the potassium channels is a common procedure that is used to extract geological information from airborne data. Their advantage comes from the fact that the grid of a ratio does not suffer from environmental agents, such as soil moisture, vegetation, and topography. (IAEA, 2003)

The ratios which are known to assist in mapping geology are:

- The U/Th ratio is known to facilitate the identification of the fractionation zoning in felsic igneous rocks (Gabriel, 2007);
- The K/Th ratio proved to assist in the discrimination between fresh and weathered mafic bedrock (Dauth, 1997);
- The U/K can be used to delimit Precambrian rocks and sedimentary rocks.(M. A. S. Youssef & Elkhodary, 2013)

Knowing their utility, we calculated the ratios for the Silet region. They are shown in Figure 3-16.

3.7.8. Contributions of the airborne spectrometry data in the distinction between lithologies

To unveil the potential of the spectrometry data to map lithology, and following the same methodology as for airborne magnetic data, we show in Figure 3-17 the value distribution of the airborne spectrometric data relating to each lithological unit in the region of Silet. The sedimentary rocks were excluded from this study because they are geologically heterogenous which hinders the extraction of an observable geophysical signature for this rock class using our data.

For the diorite family, the label starts with the letter *"D."* We can see that we cannot use the values of the three channels and also their ratios to distinguish between the different diorite lithologies because they share an overlapped range of values. Thus, our data cannot provide any help in discerning between the rock of this family.

For the gabbro family, the label starts with the letter *"O."* We can see that using the data of the uranium and the potassium channel may help distinguish between the rocks of this family because the younger gabbro lithology, labeled as *"O23,"* shows a weaker potassium value and yet a stronger uranium value compared to the older lithology, labeled as *"O1."* Also, the ratios can also be used to

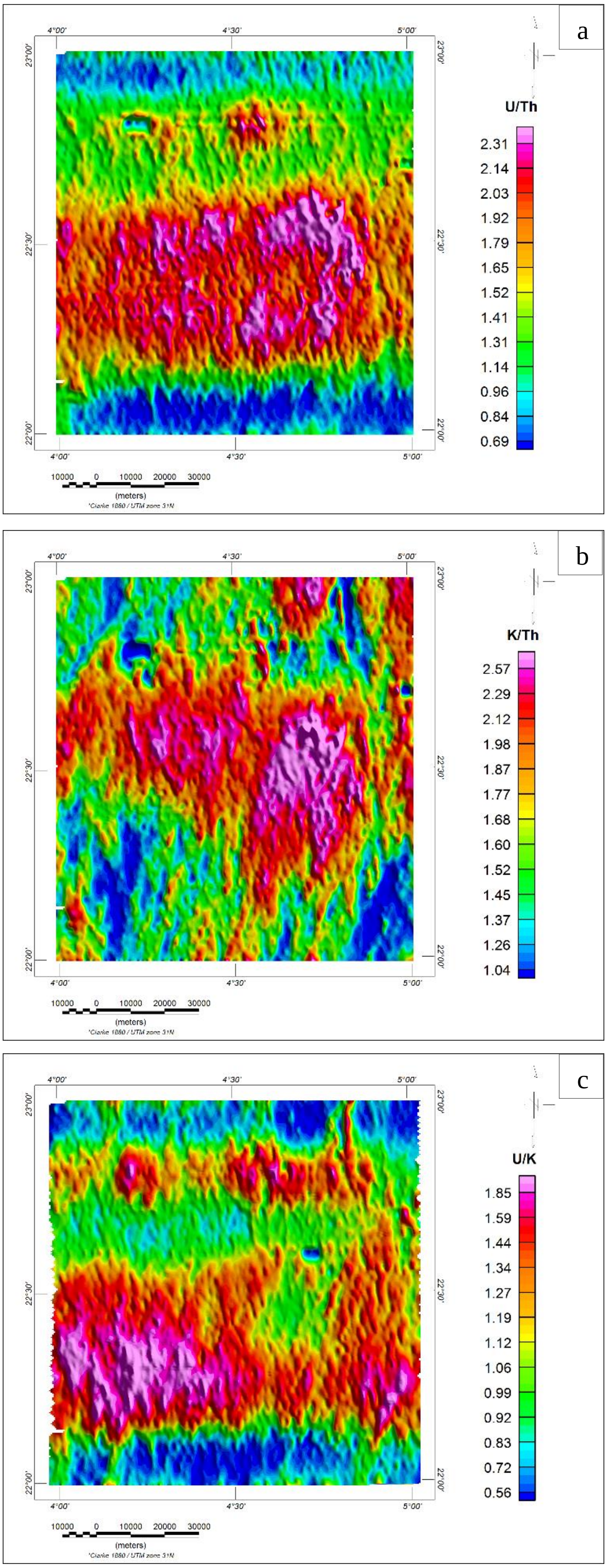


Figure 3-16 Ratios of the airborne gamma ray spectrometry in the region of Silet. a) uranium ground concentration over the thorium concentration; b) potassium ground concentration over the thorium concentration; c) uranium ground concentration over the potassium concentration.

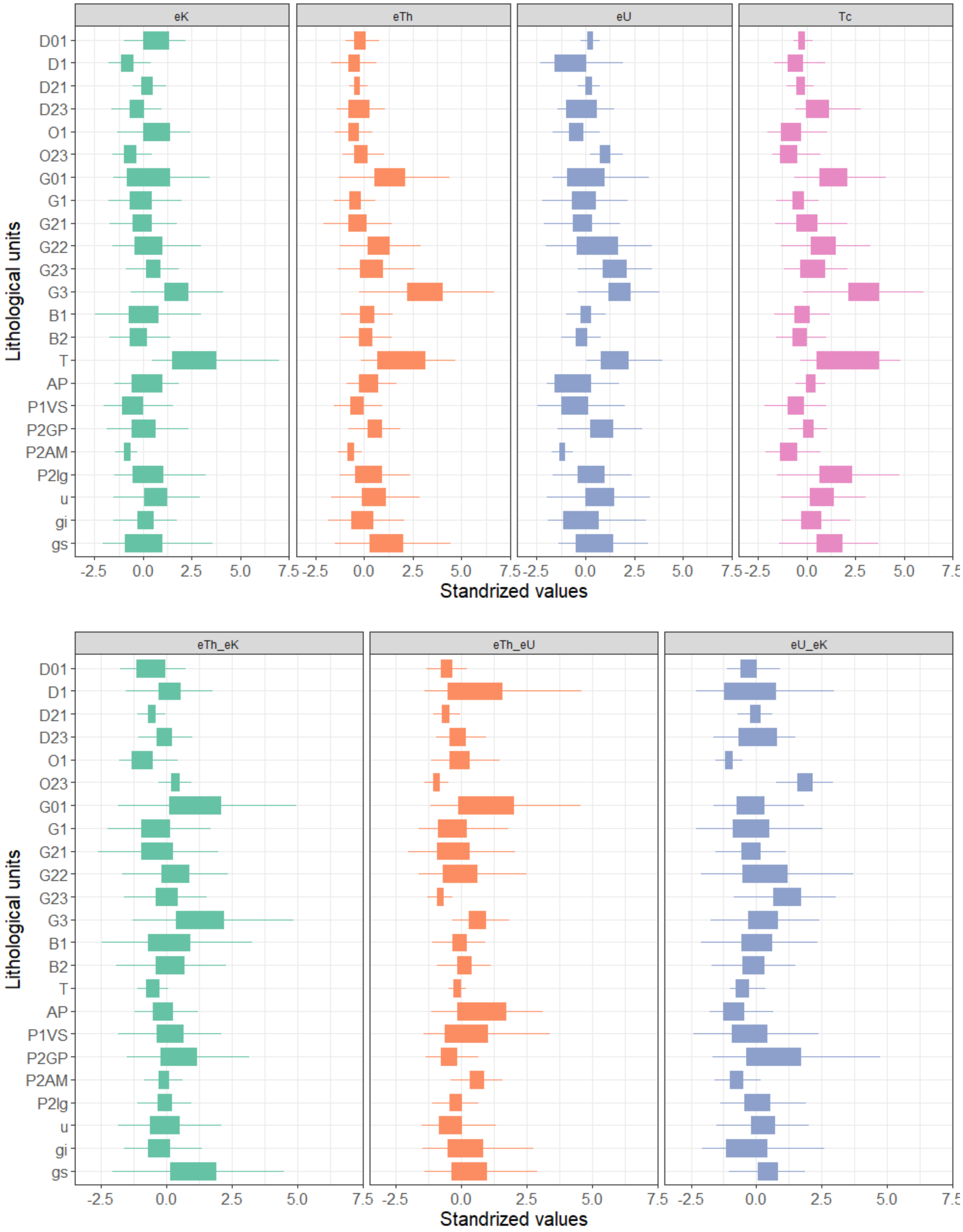


Figure 3-17 Values distributions of the airborne gamma ray spectrometry data according to the lithological units in Silet region.

set the rocks of this family apart, especially the U/K. Therefore, using gamma-ray spectrometry data should provide an additional tool to map this rock family.

For the granite family, the label starts with the letter *"G."* the distribution of the values for the rocks of this family shows no visible trend in the three channels that might be used to differentiate between them, and they appear to exhibit the same range of values. Likewise, the same observation can be said about the values of the ratios relating to this family of rocks. As a result of that, employing gamma-ray spectrometry data should yield no use in mapping the granite family.

The extrusive igneous lithologies, labeled as "*T", "B2"* and *"B1"*, and excluding the *"Trachyte"* which exhibits a strong and distinct signature in all four channels, the rocks of this class display a similar values distribution. Thus, our data may only help distinguish the Trachyte lithology. For the other basaltic lithologies, they may not prove to be useful to set them apart.

For the metamorphic lithologies, labeled as *"u"*, *"gs"* and *"gi"*, they also show the distribution of a similar value across all channels. Interestingly, the Th/K values for the *"gs"* show a strong and distinct signature compared to the other rocks, which should help distinguish them from other metamorphic lithologies.

3.7.9. Conclusion

So far, we only discussed the ability of airborne gamma-ray spectrometry to discern between the rocks of the same family. However, we did not include any indications about the utility of the data to differentiate between rocks, for example, of different classes (i.e., igneous, metamorphic or sedimentary). By visually observing the distribution of the data relating to rocks of the same class, we could not infer any patterns that might be used to say whether a certain signature in a certain channel can indicate the class of a certain rock. Therefore, the need for a powerful tool to deduce complex patterns in multi-variable data, such as the case for our data, is indispensable to objectively judge the utility as well as the limitation of the airborne spectrometry data in the context of geological mapping.

Chapter 4 : Geological setting and geological mapping applications

Part I Geological setting of the Silet region-Hoggar

4.1. The geological setting of the Tuareg shield

The Tuareg shield is a part of the Pan-African belt, which belongs to the Transaharan belt (Dallmeyer, 1990). It covers a surface of 500.000 Km^2, and it is composed of three areas. The western area spans in Mali and it is referred to as *"Aïr."* The central area, where our study area is situated, is named the *"Hoggar,"* and it is located in the Algerian Sahara. The western area spans in Niger, and it is called *"Iforas."*

The Hoggar shield, where the study area is located, was formed during the orogeny of the Pan-African (750-550 Ma) due to the collision of the west African craton and the Saharan metacraton (see Figure 4-1). It is mainly composed of Precambrian rocks, and most of its surface is covered by Paleozoic sediments. The entire shield is affected by major N-S shear zones, namely but not limited to the 4°50 and the 8°30 faults. These shear zones divide the shield on north-oriented blocs which have distinguished age, nature and evolution (Caby, 2003).

4.1.1. The geochronological evolution of the Hoggar shield and main lithostratigraphy units

Evaluating mineral ages shows that not only the majority of the rocks forming the Hoggar date between 600-500 Ma but also some lithological units date from Archean to the Paleoproterozoic (2000 - > 2700 Ma). According to Bertrand and Caby (1978), the age of the main lithostratigraphic

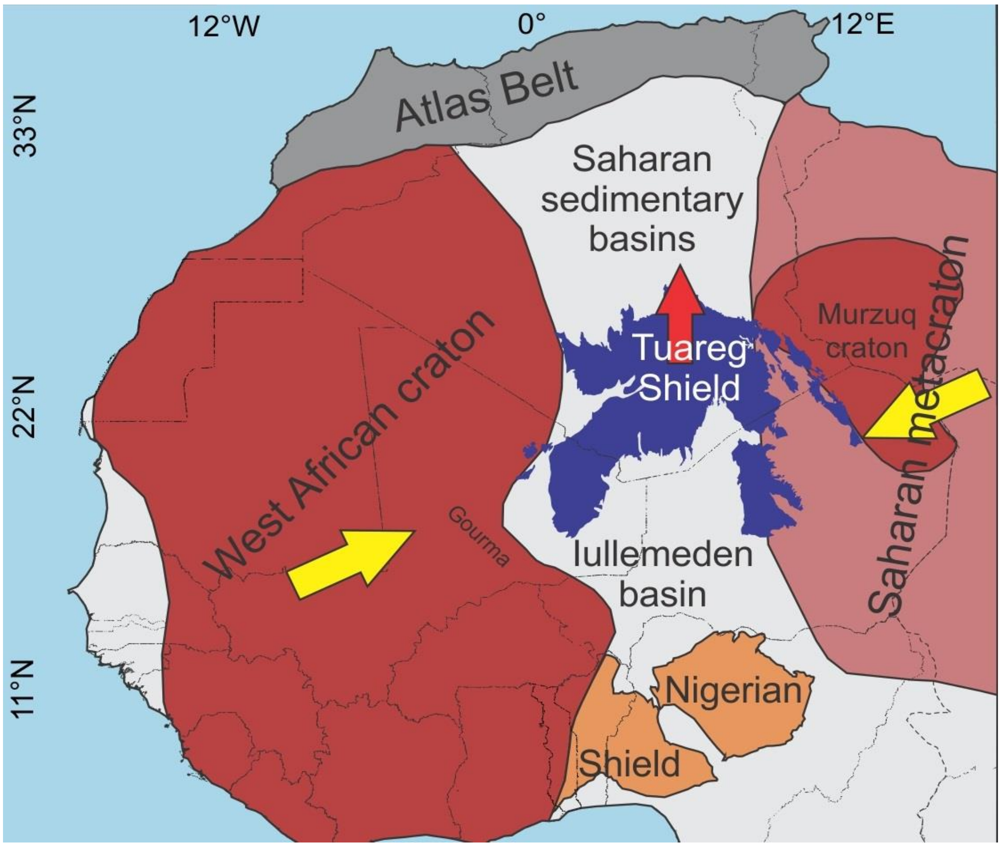


Figure 4-1 Map shows the main geological entities in northwest Africa. The Tuareg shield is squeezed between the west African craton and the Saharan metacraton. (Liégeois 2019)

units is the following

- Archean age units (older than 2700 Ma) are observed in Ouzzalian, the basement of the Oumelalen area and Arechchoum unit;
- Paleoproterozoic age units date back to 2000 Ma and are observed in Tassendjanet, Arechchoum and Oumelalen formations;
- Mesoproterozoic age units date back to 1000 Ma. They are observed in Aleksod, Ahnet quartzite and Toukmatine formations;
- Neoproterozoic age units date between (800 - 650 Ma). They are observed in the western Hoggar, the Pharusian II and Tiririne formations;
- Late Neoproterozoic and Cambrian age units. They are observed in the "Série pourprée" and" Série intermédiaire."

The orogenic events which affected the Hoggar shield are, from the oldest to youngest, the *"Eburnean,"* the *"Kibaran"* and the *"Pan-African."*

The Eburnean orogen is well defined in regions where Archean age units are observed, such as: In Ouzzal, Tassendjanet, Aleksod and Oumelalen formations. It is believed that it affected the Hoggar shield around $2000 \pm 100\ Ma$. The Kibaran event was not as well documented as the Eburnean event but it was defined in some of the units of the central Hoggar. These units showed that this event took effect around $1000 \pm 100\ Ma$. For the last event, the Pan-African affected the shield in two phases: the first phase and the second phase. The first phase, called the early Pan-African, affected the Hoggar around (750 – 660 Ma). This phase was well documented in the western Hoggar using the syn-tectonic granites. The second phase, named the late Pan-African, affected the Hoggar around (650 - 580 Ma). It was defined in the eastern Hoggar through the post-tectonic granites. (J. M. L. Bertrand & Caby, 1978; Black et al., 1994)

4.1.2. The old Hoggar subdivision

The evolution of the Hoggar shield and the multiple orogenic events lead to a complex structural history of folds, sedimentations and metamorphism of the shield's geology. Despite the different evolutions of the area, Bertrand and Caby (1978), were able to delimit three distinct blocks which form the old subdivision of the Hoggar by employing numerous criteria. Namely, and from west to east, these blocks are the Pharusian belt (western Hoggar), the polycyclic central Hoggar and the eastern Hoggar. The regional fault 4°50 separates the western and the central blocks of the shield while the 8°30 fault separates the central and the eastern blocks. In Figure 4-2, we show the old subdivision of the Hoggar shield.

The criteria used to define these blocks are:

- The lithology of the Neoproterozoic and its proportion;

- The types of folding and the metamorphism relating to the Pan-African orogeny;
- Existence or inexistence of the indications of the Kibaran event;
- The syn-kinematic granite age;
- Existence or inexistence of the molassic deposits.

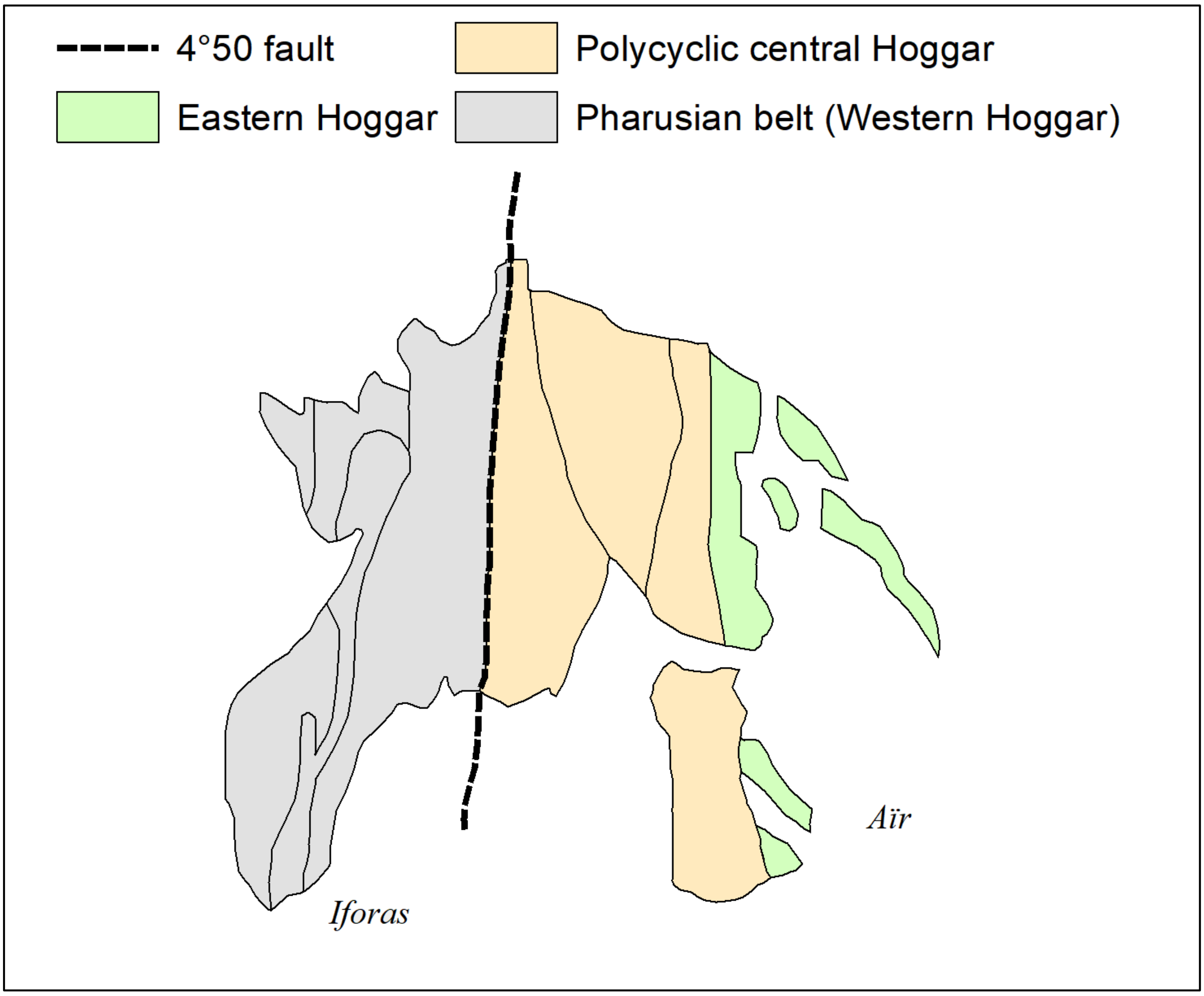


Figure 4-2 the old subdivision of the Hoggar. This subdivision was proposed by (J. M. L. Bertrand & Caby, 1978). The geological map was modified from(Liégeois et al., 2003) .

4.1.2.1. Eastern Hoggar

This block is composed of two pre-Pan-African structural units, namely the "Issalane" and the "Tafassasset-Djanet" units. The two units are separated by a narrow rectilinear belt of Neoproterozoic rocks. The belt is referred to as the *"Tiririne belt."*

4.1.2.1.1 The Issalane unit

The lithology of this unit is homogenous. It is composed of banded and veined granitic gneisses, a meta-sedimentary unit composed of green Cr-bearing quartzites, marbles and calc-silicate rocks, and pelitic gneisses with associated coarse-grained migmatites (J. M. L. Bertrand & Caby, 1978).

The Issalane unit witnessed a polyphase deformation and metamorphism that can be seen in the eastern margin of the unit. This deformation cycle dates back to before the deformation caused by the Tiririne belt, which dates to the Neoproterozoic.

4.1.2.1.2 The Tafassasset-Djanet unit

This unit is a part of the east Saharian craton, and in the same way as the Issalane unit, it is overlain by the Tiririne formation. The Tafassasset-Djanet unit is mostly composed of large batholiths of calc-alkaline granites, and to a less degree, it is composed of low-grade flysch-type deposits, marble calc-silicate sequence, various pre-tectonic volcanic and plutonic rocks, such as: alkaline gneiss, granites and ultrabasic rocks (J. M. L. Bertrand & Caby, 1978).

4.1.2.2. Polycyclic Hoggar (central block)

Lithologically, and observed especially in the western part of the block, the Pan-African lithologies constitute the most abundant formation in this block. The low-grade Neoproterozoic lithologies also exist, but they are less frequent as they only represent less than 10% of the entire block. The older Archean and eburnean rocks occurred as gneisses and high-grade schists. Moreover, the granulite facies in this block have undergone a polymetamorphic and poly-tectonic evolution. This evolution was caused by the Kibaran event and the deformations of different intensities relating to the Pan-African event.

The central block of the Hoggar has three distinguished polycyclic rock domains. They are the "*Aleksod domain,*" the "*Oumelalen-Temasint domain,*" and the "*Tefedest-Atakor domain."* The first domain, the Aleksod domain, is characterized by an Eburnean basement which is composed of augen gneisses and banded grey granodioritic gneisses dated at 1940 Ma and 2200 Ma respectively. Above the Eburnean basement, a high-grade complex of amphibolite and meta-sediments lies upon the basement with a structural disconformity. The second domain, the Oumelalen-Temasint domain, has a depositional history that can be used to describe most of the eastern part of central Hoggar. Similarly, to the Aleksod domain, it is characterized by an Eburnean basement which is formed mostly by meta-sediments and a granulite facies (Oumelalen Formation) which was metamorphosed at 2000 Ma. Two distinguished units are observed in this domain, namely the "*Toukmatine formation*" and the "*Tit n'afara unit.*" These units share the same age as the Aleksod unit. They consist of a monotonous sequence of alumina-rich schists with some quartzites, amphibolite, marbles and calc-silicate rocks, serpentinites, and locally layered alkaline orthogneisses. The last domain, the Tefedest-Atakor domain, and contrary to the other two domains, is characterized by abundant Pan-African granite intrusions and a pre-Pan-African relict and high-pressure mineral assemblages which include kyanite-bearing rocks and granulite facies in the Tamanrasset area. In the south of Tamanrasset, as the domain idens, flat-lying foliations prevail and may represent a non-reactivated zone unconformably overlain by residual basins of Neoproterozoic rocks (J. M. Bertrand, 1974).

The central block also has a monocyclic Neoproterozoic rock unit. They occur as either linear belts of grade schists and volcanic, such as in the regions of Arefsa, Serkout, and Temasint, or as small basins like in the regions of Laouni and In Ebeggui.

4.1.2.3. The Pharusian belt (western Hoggar)

The western and eastern limits of the Pharusian belt are defined by the west African craton from the west and the central Hoggar from the east. In this block, two domains with a similar evolution are observed, and the main component of these two domains are rock units that date between 1750 Ma and 600 Ma. These domains are separated by older granulitic facies which date at 2100 Ma. The granulitic facies is referred to as the *"In Ouzzal"* block.

The main lithostratigraphic units that define the western domain of the Pharusian belt are:

Mesoproterozoic units: they are considered the deepest tectonic level of the Pharusian belt. During their evolution, they have undergone a polyphase deformation due to metamorphism of upper to lower amphibolite facies grade.

Neoproterozoic shelf deposit units: these units are believed to be evolved from shelf deposits that were metamorphosed by various degrees. These units are observed as a cover, called *"Série à stromatolites"*, which overlay the Eburnean basement in Tassendjanet, or as a unit that rests concordantly upon the Ahnet quartzites.

Neoproterozoic magmatic cycle: the magmatic cycles were introduced by the intrusion of several magmatic rocks into the preexistent rocks. This includes, but is not limited to: the intrusion of basic to ultrabasic rocks into the shelf deposits of "Série à stromatolites", the intrusion of the pelitic and siltstone horizons of the gabbroic sills in the Tassendjanet nappe, the intrusion of the Ougda layered lopolith into the shelf deposits of "Série à stromatolites," the intrusion of the Tassendjanet area by the banded gabbro with cumulates and the intrusion of the dolomites at the top of the series in the same area by serpentinized ultrabasic rocks.

Late Neoproterozoic volcano-clastic/ volcanic rocks: these rocks are related to the upper Pharusian, which is defined in the central zone of the belt, and they mostly occur as metavolcanics, volcano-clastic and calc-alkaline rocks.

Molassic deposits: these deposits rest atop the Pharusian belt formations with unconformity, and they crop out as N-S graben in the Ouallen area and as residual basin from the western Hoggar where they were called "Série pourprée de l'Ahnet."

The eastern domain of the Pharusian belt, called central Pharusian by Lelubre (1952), is subdivided into two levels separated by a regional unconformity (J. M. L. Bertrand et al., 1966). They are the inferior Pharusian and the superior Pharusian. The first is an intercalation of volcano-sedimentary rocks which are composed of conglomerate, clastic rocks and volcanic rocks such as basalts, andesite, and a rare occurrence of rhyodacites. The second level, superior Pharusian, is also composed of the volcano-sedimentary rocks that are formed by the same sedimentary rocks and volcanic acidic rocks.

4.1.3. The new Hoggar subdivision

The new subdivision of the Hoggar shield, the current subdivision, was proposed by Black et al (1994), and it came following the subdivision model of the Aïr region (i.e. eastern Tuareg shield). The new model of the Hoggar describes the shield as a group of 23 displaced terranes. Each one of these terranes has a distinct lithology, metamorphism, magmatic and tectonic characteristics. These terranes are separated either by a sub-vertical strike-slip mega-shear zone or major thrust fronts.

Figure 4-3 shows the terranes which constitute the new subdivision of the Hoggar. They are: Djanet (Dj), Edembo (Ed), Aouzegueur (Ao), Barghot (Ba), Assodd-Issalane (As-ls), Tchilit (Tch), Tazat (Ta), Sérouénout (Se), Egéré-Aleksod (Eg-Al), Azrou-n-Fad (Az), Tefedest (Te), Iskel (Isk), In Teidini (It), Tin zaouatene (Za), Tirek (Tir), Ahnet (Ah), In Ouzzal (Ou), Iforas granulitic unit (Ugi), Tassendjanet (Tas), Kidal (Ki), Tilemsi (Til), Timétrine (Tim). In addition to the model proposed by (Black et al. 1994), two new additional terranes were proposed by Liègeois (2019). They are the Aouilène and Afara terranes.

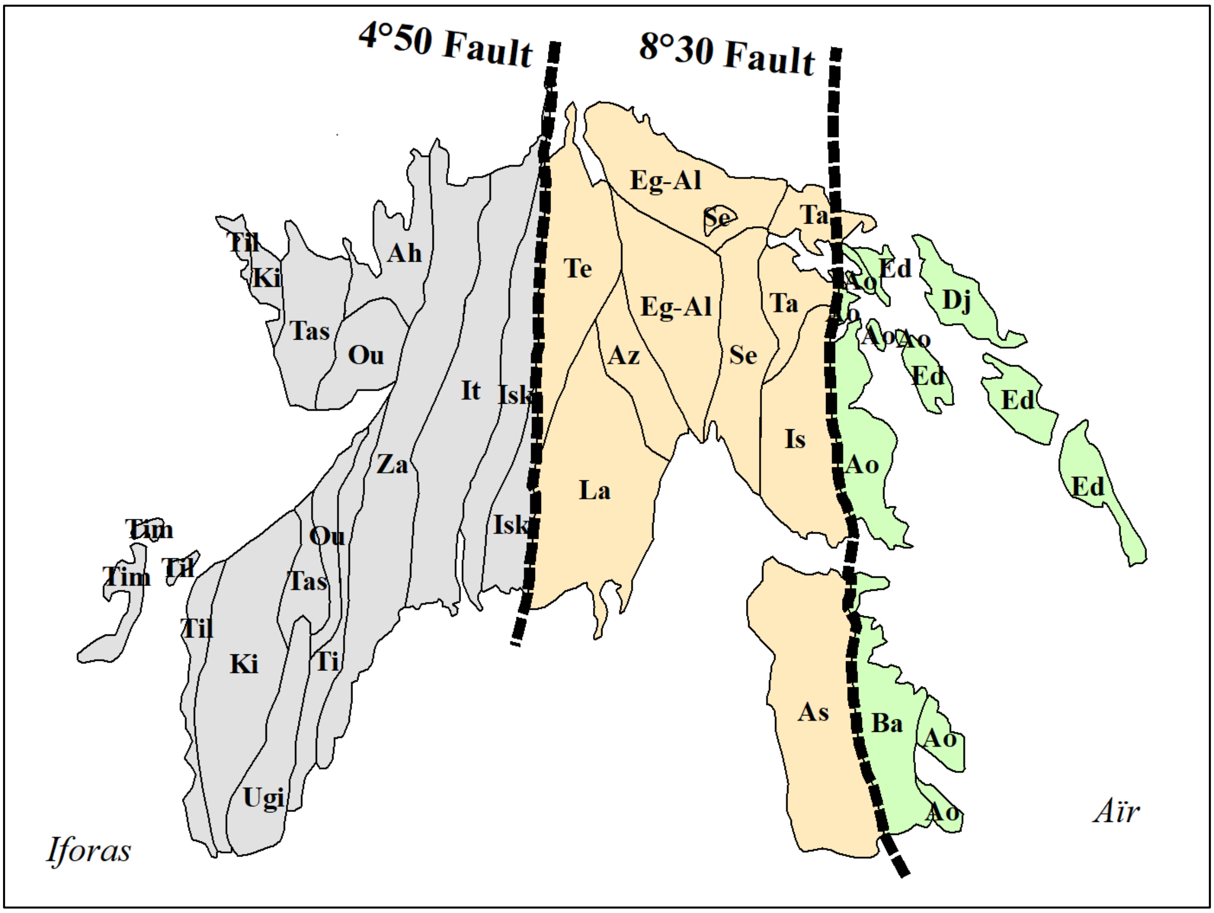


Figure 4-3 The limits of the terranes that constitute the new subdivision of the Hoggar. The color indicates the limits of the old subdivision. This map was modified from (Liégeois et al. 2003).

4.2. The geological setting of Silet

4.2.1. Introduction

The locality of Silet is located 130 km east of the city of Tamanrasset and 2000 km from the capital of Algeria. In this work, we are interested in mapping the area of Silet presented as a (1° × 1°) with a scale of 1/200.000 map. The area is limited by the latitudes 22° and 23° east and the longitudes 4° and 5° north. Historically, the area was the target of numerous geological expeditions that had the purpose of studying the geology of the area, such as the works of (Bechiri-benmerzoug, 2009; J. M. L. Bertrand et al., 1966; Boissonnas, 1973; Bouhkalfa, 2002; Chikhaoui, 1981; Dupont, 1987; Gravelle, 1969).

The geological mapping of the Silet area was issued by *"l'Office de la Recherche Géologique et Minière (O.R.G.M)"*. It was a part of the endeavours of the office to update the geological maps in the southern regions. The final map of the area was produced by Zeghouane and Hamis (2009). As a part of the work realized in this research project, we georeferenced and vectorized the map. The vectorized geological map is shown in Figure 4-4, and Table 4-1, summarizes the key description of the map.

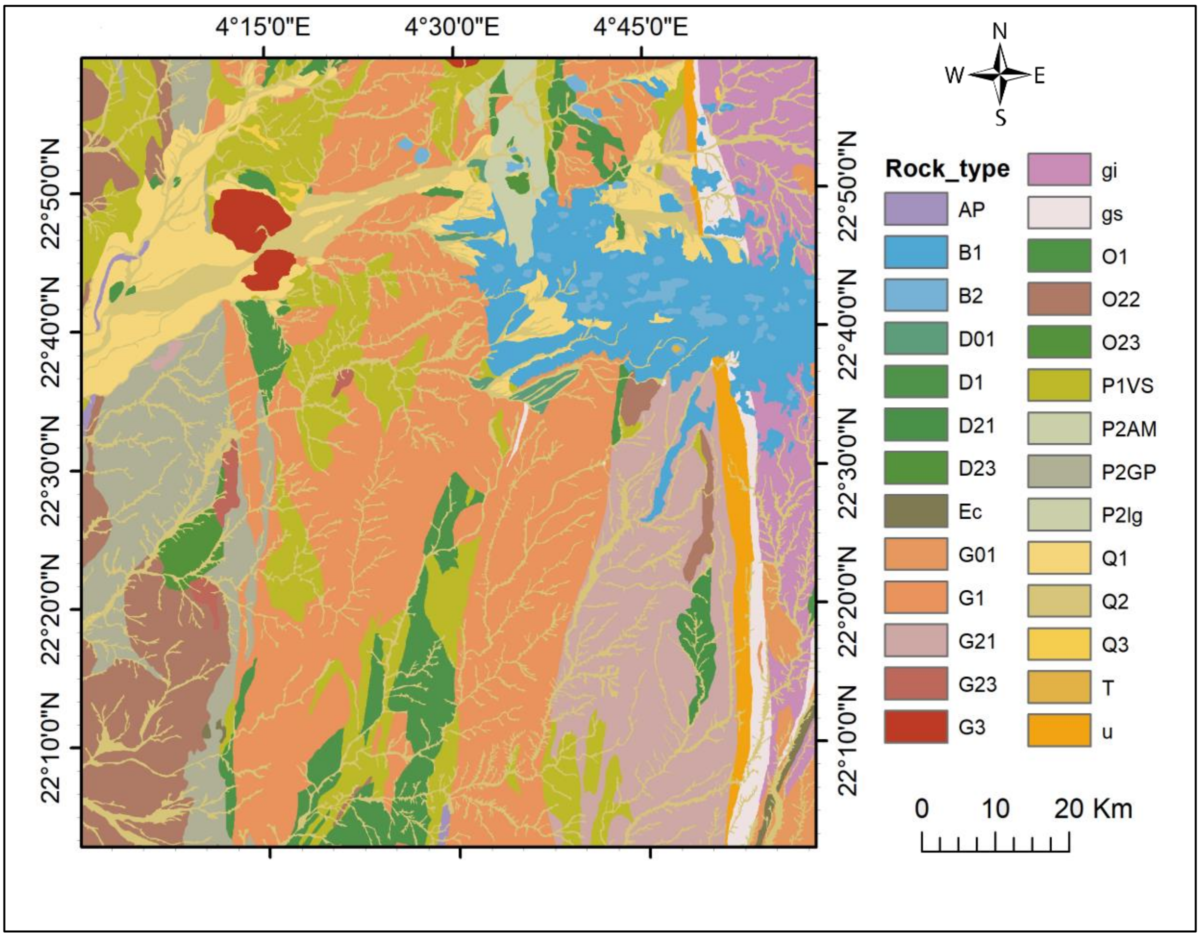


Figure 4-4 georeferenced map of Silet. It was vectorized while taking (Zeghouane & Hamis, 2009) as a reference map. Refer to Table 4-1 for key description.

Table 4-1 The description and the percentage of the surface occupied by the lithological units in Silet. Labels are an abbreviation used to denote the lithological unit. This table is based on the key of the map in (Zeghouane & Hamis, 2009).

	Lithological unit	*Description*	*Labels*	*Percentage of the surface (%)*
Quaternary units	Alluvium	Old and recent river deposits consist mainly of sand and gravel.	Q2	18.9
	Æolian deposits	Sands of dunes and small isolated ergs.	Q3	
	Peneplain deposits	Resulting from disaggregation of mountainous landforms, consists of very varied lithologies.	Q1	
Cenozoic volcanism	Trachyte and Phonolite	Basic and alkaline lavas; association of alkaline basalts-trachyte-phonolites.	T	8.16
	Recent cone-shaped Basaltic associated with pyroclastic	Heterogeneous lithology; Olivine basalts, andesites, trachyte, and phonolites as well as their pyroclastic products (lapilis and cinerites).	B2	
	Old basalts sequence	Homogeneous lavas which have an oolitic to porphyritic microlitic structure, and sometimes vitreous structure (volcanic glass).	B1	
EO-Cambrian units	Pan-African molasses	Acid lavas (felsite and white rhyolites) that are associated with breccias and conglomerates and sometimes stretched pebbles.	Ec	0.29
Western Hoggar units	Pan-African mylonite	Pan-African mylonite (strongly deformed rocks)	u	63.5
	Ighellochem volcanogenic series	Mainly volcanogenic with calc-alkaline affinity metamorphosed in greenschist-facies conditions; Basic, semi-basic (Dacite), and acidic lavas (Rhyolite) associated with pyroclastic (Basalt, Andesite).	P2lg	
	Amded volcano-sedimentary series	Detrital volcano-sedimentary series (sandstone, pelites, aleurolites, conglomerates, pudding, and schists) associated with acidic metavolcanics and marbers Versicolor rocks.	P2AM	
	Pelitic-sandstone series	Sandstone and pelites inter-bedded with rare neutral to acidic metavolcanite, shist, and marble.	P2GP	
	Timeslarsine volcano-sedimentary series	Basic to neutral meta-volcanites (meta-basalt, meta-andesite) with ultrabasic lenses associated with platform metasediments (marble, quartzites, jasper).	P1VS	
	Gneissic series	Dioritic and banded gneisses with rare quartzite.	AP	
	Taourirt granites	Alkali to sub-alkali granite, porphyroid to biotite.	G3	
	Calco-alkaline granitoids	Medium-grained non-porphyroid biotite calc-alkaline granite and granodiorite.	G23	
	Associated diorite facies	Fine-grained diorite, (Adrar, Ighellochem region).	D23	
	Gabbro	Gabbro associated with Iskel granites.	O23	
	Imezzarene calco-alkaline granitoids complex.	Medium to coarse-grained porphyritic calc-alkali granite and granodiorite.	G22	
	Isseimane river calco-alkaline granitoids complex.	Calc-alkaline granite and granodiorite oriented to biotite and rarely to medium-grained amphibole.	G21	
	Basic plutonic formation	Diorite associated with gabbro.	D21	
	Tin Tikadiouit and Taket Granitoids complexes	Medium to coarse-grained oriented biotite-amphibole granite and granodiorite.	G1	
	Diorite and quartz-diorite	Diorite and associated quarzitic diorite.	D1	
	Mafic-ultramafic formations	Gabbro, pyroxenic gabbro, and associated ultrabasic.	O1	
Eastern Hoggar units	Tinef gneissic series	Leptinites and fine leucocratic granito-gneisses with interbedded with versicolor marbles (yellow and bluish-white sandstone).	gs	9.19
	Arechchoum gneissic series	Highly evolved gneiss material dominated by calc-alkaline orthogneiss with biotite +/- amphibole as well as eye-shaped or migmatitic gneisses	gi	
	Anfeg granitoids	Syn-tectonic calc-alkaline granite with biotite +/- amphibole medium grain.	G01	
	Associated diorite facies	Plutonic rocks of dioritic facies.	D01	

4.2.2. Geology of Silet

Silet is separated by the regional faults 4°50 E into western and eastern branches. According to the old subdivision of the Hoggar shield, the eastern branch is part of the central Hoggar and its formations represent approximately 9.5% of the whole surface of the region. The western branch is part of the Pharusian belt (i.e., western Hoggar). It occupies most of the surface of the area (approximately 64%). The remaining surface is occupied by Cambrian units, Cenozoic volcanism units and Quaternary units. According to the new subdivision of the Hoggar, Silet is constituted of the pre-Neoproterozoic terranes Tefedest and Aouilène, and the juvenile Neoproterozoic terranes Iskel and In-Teidini. Above the old Proterozoic basement, lies a moderate cover of quaternary formations in the form of wadis, dunes, and Cenozoic volcanism formations. the latter formation is observed exclusively in Tahalra-tassetafet (northeast of Silet), which rests with a stratigraphic unconformity on the eroded Proterozoic basement.

Silet is located in the eastern domain of the Pharusian belt which led to the occurrence of formations of both the Pharusian I and the Pharusian II cycles. The volcano-sedimentary formations of the Pharusian I in Silet are observed in the Timeslarsine series and the Edjedjou Oued complex, whereas those of the Pharusian II are observed in the Ighellochem volcanic complex and the Amded series. The formations of the two cycles outcrop mostly in the northern part of the map. In the southern part; however, the granitoid formations have the most occurrence, while the formations of the Pharusian have less presence.

The plutonism in the region is observed by the existence of syn-kinematic plutonic rocks. These rocks are presented by the N-S elongated batholiths of Tin-Tikadiouit dated at 870 and 840 Ma (Caby et al., 1982); the quartz diorite of the Timeslarsine wadi dated at 868Ma; the Taklet batholith dated at 839Ma; post-tectonic granitic complexes of the "Taourirt" type such as Tin-Erit, Taharaït N'abror and Tioueïne dated at 525Ma (Paquette et al., 1998); the Ahambatou granodioritic unit dated at 650Ma; The tonalitic batholith of Tamtèq dated at 732Ma; the monzogranitic pluton of Silet dated at 650Ma; the Eheli batholith dated at 638Ma; Imezzarene granitoids dated at 583 Ma (J.-M. Bertrand et al., 1986); the undated Ijelhèk pluton and the undated Iharèdj batholith.(Bechiri-benmerzoug, 2009)

4.3. Geological mapping

Geological mapping is the process of preparing a special map that keeps a record of the distribution of the outcropped rocks belonging to different formations. Besides including information about the historical evolution of rocks, a geological map should also show information about linear features, such as faults, folds and bedding. Usually, these linear features are attributed to a special symbology, such as “the strike”, “the dip” and “the plunge” to indicate the orientation and the underground extension of rocks.

We can classify geology maps into geological reconnaissance maps, regional geological maps, detailed geological maps and specialized maps (Lisle et al., 2013).

Geological reconnaissance maps are usually produced at a scale of 1:250 000 and smaller. These maps are often based on geological interpretations extracted by remote sensing techniques or some airborne acquired data which have a geological affinity, including geophysical and geochemical data (Lisle et al., 2013). During the production of reconnaissance maps, and due to the vast surface covered by these maps, the in-situ confirmatory field work is limited. Thus, this type of maps only serves the purpose of giving an overview of the general distribution of the geology as well as the general structure in large areas.

After studying the general outline of the geology and producing reconnaissance maps, regional geological maps can be created after detailed geological studies. These maps are produced at a scale of 1:100 000 and larger. As a part of increasing the accuracy of the map, they should also be plotted on topographic base maps. Similar to the reconnaissance maps, any methods that can help in plotting geology including geophysics and satellite images or field works, like trial pitting, augering and drilling should be incorporated.

Detailed geological maps are any map made at 1:10 000 and larger. The need for such maps arises when trying to further investigate discoveries encountered in a smaller-scale mapping or to carry out a preliminary investigation in major engineering projects.

Specialized maps are made on a large scale. They are designed to map geological features in small areas with greater detail. This type of maps is used, for example, to map open pit mines, underground geological mines, and engineering site. The scale in which these maps is made ranges from 1:10 000 and can get up to 1:500.

4.3.1. Geological mapping of Algeria

To map the geology distribution of the national territory, the geological service agency "*l'agence du sevice geologique de l'algerie ASGA*" publishes geology maps at different scales. According to their published catalogue (ASGA, 2019), the most adopted scales to produce geology maps are:

Geology map at 1:2 100 000: this map gives an overview of the general distribution of the geology as well as the general structure of the whole national territory. The work on this map was issued on behalf of the ASGA (*"service géologique national"* at that time) in 2008 and was finished in 2015. The published map can be found on the website of the *"Banque Nationale de Données Géologiques."*

Geology maps at 1:2 000 000: the first published geology maps at this scale go back to 1952. This scale was adopted to produce two geology maps that cover north Africa. The first map covers the western part of the national territory and Morroco. The second map covers the eastern part of Algeria as well as Tunisia. In 1962, this scale was adopted again to produce the geology distribution maps of the central and western Sahara

Geology map at 1:1 000 000: to this date, the only available geology map at this scale was the one produced by *"Société Nationale de la Recherche Minière"* SONAREM in 1977. This map shows the geology distribution in the region of Hoggar. It was produced as a part of the mineral exploration surveys issued by the company.

Geology maps at 1:500 000: this scale was adopted early on by ASGA *("service de la carte géologique de l'Algerie"* at that time*)* to produce six geology maps between 1934-1944. These maps covered the northern and the southern parts of Algiers, the northern and the southern parts of Oran and the northern and the southern parts of Constantine. ASGA published the second edition of the same maps in 1952. In 1990, the agency published the third edition of these maps and initiated the program of covering the rest of the Algerian territory with geology maps at 1:500 000. This program holds the name *"la cartographie géologique au 1/500 000 émé (CGA-500)"*. It splits the Algerian territory into 43 rectangular blocks with dimensions of (300 × 200) km. As of now, 65% of the Algerian territory is covered by this program, since the geology maps of 28 blocks of the 43 blocks are produced.

Geology maps at 1:200 000: to achieve the endeavour of producing geology maps at this scale, the map of Algeria is split into 246 squared blocks with dimensions of (100×100) km. According to the website of the *"Banque Nationale de Données Géologiques,"* the geology maps of 59% of these blocks are produced.

Geology maps at 1:50 000: this scale was adopted to produce the first-ever geology map in Algeria. It was used to produce the geology map of Thenia-Boumerdes (Ex menville) in 1895. Nowadays, this scale is primarily used to produce geology maps for north Algeria. To achieve that, north Algeria is split into 327 rectangular blocks with dimensions of (30×20) km. Currently, this program produced 54% of the geology maps of northern Algeria.

4.3.2. Geological mapping methods

At the start of any geological expedition, and before starting to map the different formations in the area to be mapped, reconnaissance work should be done first. This work includes getting initial impressions about the rock-types present, the general structure, locating areas with good exposure and locating accessible routes and zones. After that, the work schedule of the expedition can be planned (Lisle et al., 2013). To produce a geology map, (Greenly & Williams, 1930) described three methods. These are: following contacts, traversing and outcropping mapping.

4.3.2.1. Following contacts

A formation is a geological unit composed of a suite of rocks, and together, these rocks possess distinct characteristics that mark them off from other formations. During the mapping procedure, the geologist tries to follow the limits of the different formations and draw them on a map. These limits are referred to as *"contacts,"* and depending on the geologist and also on the scale of the map, the contacts between the formations may change.

The easiest way to determine the contacts between formations is to visually follow them on the ground. This can only be achieved in well-outcropped and homogenous areas. However, in the opposite case, the contacts can be hard to follow, so in this case, the contacts are either inferred using contour structures or using aerial photographs. Topography variations in these photographs can indicate the position of the contact even when covered by superficial lithologies.

4.3.2.2. Traversing

This method is an alternative way to follow contacts. It consists of planning a daily route along which the geology is plotted, and this route is selected in a way that crosses the general geological trends in the mapped area. This method is usually used in reconnaissance work. In such cases, the route is planned as a group of wide-spacing parallel lines, and any encountered geological features or contacts are extrapolated across the route lines. Besides the reconnaissance work, this method can also be used to create highly detailed maps. This can be achieved in well-outcropped areas and in areas where the geological structures are not complex. In these conditions, the spacing between the lines of the planned route is decreased to achieve the desired resolution.

4.3.2.3. Exposure mapping

This method is the most used when mapping at scales of 1:10 000 or larger. Usually, the occupied surface by the exposure is coloured using a colour designed for this formation. For some geologists, to ensure that the coloured map can be used over long periods without being worn off or having the colouring to be blurred, they tend to draw the limit of the exposures by a solid line, and after that, the solid line is inked in green which is referred to as the *"green line mapping."* (Greenly & Williams, 1930).

4.3.3. Modern geological mapping techniques

In modern days, and due to the immense availability and ready-to-use remotely sensed data, traditional fieldwork is now assisted through the utilization of computer-based mapping. Integrating different types of data such as: satellite data, geophysical data and aerial photos is now possible owing to the introduction of geographic information systems (GIS). The GIS allows the manipulation of different data types at different scales and uses them to extract geological interpretations. Consequently, it produces geological predictive maps. This can be achieved by either a visual interpretation of the compiled remote sensing data which is usually provided as digitized/gridded maps or by integrating computer-assisted techniques like machine learning. (R.J. et al., 2012)

Part II State of the art of geological mapping

4.1. Geological mapping and state-of-the-art application

This section presents a collection of published and peer-reviewed papers which focuses on the different utilizations of different types of remote sensing data such as satellite data and geophysical data. First, we provide an overview of studies that uses remote sensing data mainly for mineral prospecting and environmental usages. Then, we narrow the subject of the studies to when remote sensing data were used for geological mapping, including machine learning-assisted studies. Finally, I finish this section with some conclusions relating to the impact of the remote sensing data types on predicting lithology as well as pointing out ideas for future research projects.

4.1.1. Environmental and mineral prospection

The geophysical data were used for the prospecting of minerals. In (Bournas et al., 2019) study, they used high-resolution airborne geophysical data with a spatial resolution of 100 m. The data included reduced-to-pole (RTP) data and their derivatives, gamma ray spectrometric grid channels and their ratios, and Landsat ETM+ images. These data were implemented to an enhanced maximum likelihood classifier (EML) which was trained using training zones with known mineralized potential and known geophysical signatures. After that, the classifier generated prospectivity maps for the uranium as well as iron ore and strategic metals (i.e., titanium and vanadium). The results of this study showed that the produced prospectivity maps were able to delineate new potential zones for gold, uranium and other strategic metals. However, the authors insisted that their results should be combined with fieldwork to assess the accuracy of the predicted models. In a similar study concerned with minerals prospecting, Sun et al (2019) adopted a GIS-based methodology coupled with machine learning methods for mapping copper-bearing ores in Tongling district, east China. This work also included a comparison between several machine learning algorithms (MLAs), namely: support vector machine (SVM), artificial neural networks (ANN) and random forest (RF). These MLA were trained using a set of 12 input grid data, among which, the RTP was included. The comparison study showed that RF achieved not only the best overall accuracy, but also the best sensitivity, negative predictive value, and Kappa index. Moreover, the RF was the most efficient model in capturing copper deposits, as it was able to locate most of them within the smallest testing zones. Therefore, it was used to generate a prospectivity map in a follow-up exploration, in which, the RF prospectivity map was able to predict two newly discovered deposits within the area. As a conclusion for their work, the authors suggested that prospectivity MLA-based models can provide substantial aid for delineating and exploring copper deposits.

In an environmental study, using multiple remotely sensed data, Sahin (2020) carried out a study in which he focused on generating a landslide susceptibility map for Ayancik, Turkey. His study concentrated on using tree-based ensemble methods, such as random forest, gradient boosting machines (GBM) and extreme gradient boosting (XGBoost) methods. To train these MLAs, 105 locations in which landslides occurred were used. The researcher attributed information about the lithology, geomorphology (extracted from digital elevation model DEM), hydrogeology and land cover (extracted from satellite images) to these locations. The number of input data consisted of fifteen factors. These are landslide causative factors. In addition to that, and after training the MLAs using all factors, the researcher tested a feature selection method called *"Symmetrical uncertainty"* (SU). The SU calculated the feature importance of the input factors. After that, a logistic regression (LR) based method was used to choose the best subset of features that were proposed by the SU. As a result of the first experiment, and although a small difference between the MLAs performance was observed, the researcher declared that the XGBoost slightly outperformed the other ensemble methods. For the second experiment, by choosing only the subset of features proposed by the LR method, the accuracy of all MLAs was increased, and the training time was decreased. Therefore, Sahin concluded that employing the feature selection methods is advantageous for the performance of the MLAs because it reduces the dimensionality of the data which most remote sensing-based data are suffering from. He also concluded that it also increases the performance of the trained MLAs.

In a more recent study, Youssef and Pourghasemi (2021) used the same input data as Sahin (2020). In their study, seven MLAs, including support vector machine (SVM), random forest (RF), multivariate adaptive regression spline (MARS), artificial neural network (ANN), quadratic discriminant analysis (QDA), linear discriminant analysis (LDA), and naive Bayes (NB). The MLAs were trained to produce a landslide susceptibility map in Asir region, Saudi Arabia. The results of this study showed that random forest achieved the best score among all other MLAs. Interestingly, and despite using different methods to rank feature importance, the geomorphology information (extracted from DEM) such as the slope was chosen in both studies as the feature that most contributes to predicting the occurrence of a landslide whereas features like land cover/use and plan curvature were ranked as the least contributors. In contrast, the altitude feature in Sahin (2020) was chosen as having the second most contribution contrarily to Youssef and Pourghasemi (2021) where the same feature had less contribution.

Lee et al. (2019) and Joharestani et al (2019) also conducted two environmental studies using remote sensing data, and both studies employed random forest (RF), extreme gradient boosting (XGBoost) and deep neural networks (DNN). In the first study, the MLAs were used to predict total precipitable water (TPW) and water vapor content in the atmosphere. The second study used the algorithms to predict the concentration of fine particulate matter with a diameter less than 2.5 μm ($PM_{2.5}$) in air. Lee et al. (2019) employed images of Himawari 8, GeoKompsat-2A satellites to extract TPW. They used two different reference data to validate the predictions of the MLAs. Regardless of the reference data, the DNN achieved a correlation score of 0.96, by which, it outperformed the performance of both tree-based ensemble methods. The random forest achieved the worst performance. In addition to that, the researchers observed that the trained models overestimated the

TPW in regions with high altitudes. Therefore, they suggested that considering altitude data as a predictor for the TPW would have a positive effect on the predictions of the MLAs. In (Joharestani et al., 2019), the MLAs were trained with aqua satellite data, ground-measured $PM_{2.5}$ and metrological data such as, but not limited to, temperature, relative humidity and daily rainfall. Unlike the previous study, the XGBoost proved to be the best MLA for predicting $PM_{2.5}$, as it achieved the best correlation score. What is worth noting in this study; however, is the usage of the three different feature ranking methods with each giving a different feature ranking. Therefore, these ranking methods should be approached skeptically as not to eliminate features that might be deemed as having less contribution to the prediction process while, in fact, using them might be beneficial for MLAs.

(Kang et al., 2020) conducted another an environmental study using remote sensing data. They assessed the performance of six MLAs including: lasso, support vector regressor, random forest, XGBoost, long-short term memory (LSTM), and convolutional neural network (CNN). The purpose of his study is to predict maize yield using the MLAs. These MLAs were trained using environmental variables extracted from different remote sensing data. Unlike the previously overviewed studies, this study did not use a feature ranking method but instead depended on the study of the correlation score of the input data with output data (i.e., maize yield). The study resulted in building three space features: the first space contained environmental variable used to predict maize yield from previous studies; the second space contained the variables of the first space and the variables that showed a substantial correlation with the maize yield; and the last space contained all the input variables. The results of this study showed that not only the XGBoost performed better than the other MLAs but it was also computationally faster. Moreover, the XGBoost achieved the best score when trained using only the variables of the second space. This indicates the advantage of performing the correlation study. In contrast, although the MLAs based models (i.e., CNN and LSTM) achieved comparable scores as the SVM and the RF, their performance showed an increase the more we include features when training them. This shows that these MLAs have better stability when facing large features space.

4.1.2. Geological mapping without the aid of machine learning

In the following section, a review a collection of published papers that aimed to generate geological interpretations based on remote sensing data, mainly geophysical data, is included. In these studies, data-driven methods such as machine learning methods are not applied. The data was, instead, directly used to infer the geological interpretations.

In an attempt to map regolith (i.e., surficial deposits that cover bedrock) characteristics in Yilgran craton, west Australia, Dauth (1997) integrated airborne magnetic and spectrometric data and the satellite data of SPOT ("*satellite pour l'observation de la Terre"*) and Landsat. The researcher previously applied processing techniques to the data so that it would be more suitable for mapping.

For the magnetic data, he applied a line-based filter to the airborne magnetic data called *"REGMAG."* The filter was used to accentuate the response of surficial formations and to attenuate the response of bedrocks. For the radiometric data, he used a normalization algorithm to adjust the value of the radiometric ratio K/Th. For Landsat images, he utilized the representations of ternary images of the 5/7, 4/7 and 4/2 bands. The processing techniques and the data representation used in this study proved to be useful to produce remote sensing-based maps that can be used to map near-surface formations. However, a point of consideration before applying the REGMAG filter is that the survey data should have high spatial resolution and low system noise. In another study, (Asfirane & Galdeano, 1995) utilized airborne magnetic data to interpret the geology of the north of Algeria. They were able to classify the magnetic anomalies into two groups: short wavelength anomalies that correspond to volcanic rocks of the Triassic age and long wavelength anomalies related to the basement in the region. Despite not being able to invoke certain interpretations about the basement, the researchers were able to reach the conclusion that the basement in northern Algeria is highly magnetized, and its magnetic signature is similar to the basement observed in Europe.

(Slavinski et al., 2010) conducted a study in which they integrated Landsat 7 data, aerial photography, digital elevation model (DEM), and both the airborne magnetic and spectrometric data to produce a predictive geological map of the Baie Verte Peninsula. To achieve intended results, they used a methodology that consisted of confronting the geological map of the region. Although they couldn't outline new geological units, they could redefine the contacts between the known units. This came after inspecting the airborne magnetic and DEM data. Both data types proved useful for outlining the fault system as well as defining the limits of some lithological units.

Youssef and Elkhodary (2013) conducted a well-detailed study in which they used airborne gamma-ray spectrometric (AGRS) data to update the geological map of the southeastern desert in Egypt, and also to execute an environmental study. As a part of a qualitative analysis, the grids of the total count (Tc) and the three radioelements as well as their ratios and ternary images were inspected by extraction grid values corresponding to each lithological unit in the area. Taking the Tc grid as a reference, the contacts between lithological units were roughly mapped while ternary images were used to further refine these contacts. The results of this were that some lithological units, which were identified as a whole unit in the reference geological map, were subdivided into two distinct units, each having different radiometric signature. Moreover, to assess the distribution of the radioelements in these units, a quantitative analysis was carried out on the numerical data by calculating descriptive variables, such as the minimum, the maximum, the range, the standard deviation and the variation coefficient. Among all channels, only the thorium channel was characterized by a relatively homogenous distribution across all units. Following the analysis that was designed to redefine the lithological contacts, and to locate uriniferous zones, the researchers calculated a significant factor defined as the mean plus two times the standard deviation, for eU, eU/eTh and eU/K channels. Areas with values above the significant factor were deemed as having a potential for uranium mineralization. Based on this analysis, the researchers were able to locate nine potential zones. They recommended a follow-up field work to further inspect the located zoned. The final part of this study was the environmental study which was assessed by calculating the dose rates in the area. The researchers concluded that the majority of the lithological units exhibit a dose rate that does not exceed the safe limits. The only exception are some Precambrian rocks which are located in the northwestern and southern parts of the area.

In another well-detailed study aimed at geological mapping, Thomas (2020) presented a rigorous overview of the usefulness of satellite images to map geology. In their study, advanced spaceborne thermal emission and reflection radiometer (ASTER) and Landsat 8 products as well as a digital elevation model (DEM) were employed to map the geology of the Murchison Greenstone Belt region, in south Africa. During the processing, of satellite images, procedures like band rationing, false color composites (FCC), minimum noise fraction (MNF) and PAN sharpening are usually used. The last two procedures are only available for Landsat eight products. Using these procedures, the researchers produced sixteen sub-images from the ASTER data and twenty-six sub-images from the Landsat data. Using the DEM data, they produced a slope map with different azimuth directions. All of these images were presented as FCC in RGB. The result was maps that can distinguish geologies and structural features and protruded outcrops including dykes. In addition, the slope maps were found to be useful for identifying ridges, outcrops, and structural lineaments. The sharpened bands proved to show more distinct limits of the lithological units.

4.1.3. Geological mapping with the aid of machine learning

Unlike the previous section, this section is concerned with studies that used remote sensing data for geological mapping using data-driven algorithms. We start reviewing works that used mainly satellite images for the mapping task, and after that, we review studies that used airborne geophysical data in combination with other remote sensing data.

Due to the popularity of the maximum likelihood classifier (MLC), Oommen et al. (2008) conducted a comparative study between MLC and support vector machine (SVM). The performance of these classifiers was tested in classifying lithologies. To do that, both hyperspectral images (Hyperion EO1 satellite) and multispectral images (Landsat 7 satellite) were separately used to train them. The researchers illustrated numerous training methodologies and illustrated how the classification accuracy of the SVM was better than MLC when only multispectral images were used to train them. Moreover, when hyperspectral images were used, they showed how MLC was not always able to generate predictions. This was due to the large number of bands in Hyperion images, while SVM was not affected by the band numbers. As matter of fact, the accuracy of the SVM increased as the number of bands used to train. After identifying support vector points, which are observations with the most contribution to the classification, the researchers finished their study by showing that training SVM with these observations only generated the best accuracy despite the small number of observations. Therefore, they concluded that by following a certain approach for the identification of support vector observations (i.e., useful observations), the SVM would achieve a high classification score with a small size of training data.

In a similar study, Kovačevič et al. (2009) also tested the performance of the SVM to classify lithology in the South part of Saharan Atlas. In this study, the input data consisted of using the spatial coordinates (X and Y) as well as five Landsat 7 bands. Gaussian and linear kernels were used for the SVM. The MLA was trained three times with different input (only XY coordinates, Landsat 7 bands only and the combination of the two) and each time with different training data sizes. The results of these experiments showed that the SVM with the Gaussian kernel performed better than that with the linear kernel. This indicated the complexity and the non-linearity problem of predicting lithology. In addition, the SVM was able to predict lithology using only spatial information (i.e., coordinates), and the addition of the Landsat 7 had a small to no effect on the performance of MLA. This study was concluded by testing the SVM with the Gaussian kernel on the overall geology of the region. By doing so, the researchers illustrated how, at this time, incorporating the Landsat 7 data increased the performance of the MLA. Another usage of SVM to classify lithology was proposed by Ourhzif et al (2019). This time, it was used to compare the lithology predictions based on Landsat 8 and ASTER data. The bands of both satellite data were used to generate maps of ratios, false composite color (FCC) maps and minimum noise fraction (MNF) maps. The principal components analysis was used to process these maps. After that, these maps were all fed to the SVM to produce predictive lithology maps based on satellite data. The results of this comparison showed that both satellite data were able to capture surficial sedimentary rocks with some sedimentary units appearing better in the predictive map of the ASTER data than the Landsat 8 predictive map. Additionally, the use of the ASTER data allowed capturing basaltic units, while the Landsat 8 data proved to be more effective in capturing rhyolitic and schist units in the region. Bachri et al. (2019) also assessed the performance of the SVM for lithology mapping. To achieve that, they used Landsat 8 data, and with it, they incorporated geomorphological attributes (e.g., slope, curvature, and surface roughness) from the digital elevation model (DEM) obtained from ALOS/PALSAR satellite. Besides showing the advantageous use of the SVM for the lithological map, the researchers revealed the limitations of using satellite data for the task of geological mapping. These limitations are vegetation and weathering cover, atmospheric effect, heterogeneity of lithology at the pixel level of the satellite data, the similar mineralogical, and chemical composition of the lithological units which lead to a similar spectral response. Othman and Gloaguen (2017) conducted a similar study concerned with the effect of satellite data with geomorphological data on predicting lithology. They used ASTER satellite data and a geomorphological attribute called topographic position index TPI, which can be extracted from DEM data. The geological mapping was carried out using these datasets in conjunction with SVM and RF MLAs. The results showed that RF outperformed SVM, and also, including the TPI increased the prediction accuracy of the MLAs.

Yang, et al (1998) carried out an interesting study in which they tried to examine the effect of including different date types to map lithology. In their study, they used Landsat data and geophysical data, including gravity, magnetic and gamma-ray spectrometry. Using these data, they trained 4 artificial neural networks (ANN) models, and for each model, they used different input data. Their results showed that the visible (VIS) and near-infrared (NIR) bands of the Landsat data had small contribution to the classification, because excluding them was not determinantal for the performance of the classifier. As a matter of fact, including these bands for one of the trained models decreased its performance which indicated that, for some data types, excluding them might be more beneficial to the classifier than using them to expand the input space. On other hand, geophysical data, especially the radiation ratios, proved to be essential for generating accurate classification. Harris et al (2009) also used airborne magnetic data, gamma-ray spectrometry (ground and airborne) data and Landsat to produce a predictive geological map for Sekwi region, in Canada. In this study, the maximum likelihood (MLC) classifier was used. Some of the valuable insights of this study are: among all remote sensing data used in this study, gamma-ray spectrometry is a generally better choice for separating lithology. This is especially true in young terranes with well-outcropped sedimentary rocks intruded by granitic formations. The surficial and the vegetation cover do not affect the sensitivity of the airborne gamma ray (AGRS) data to bedrocks formations, as in this study, very high correlation score was obtained between the airborne data and ground data. By using a supervised classifier, MLC, to produce a predictive geological map, the AGRS data produced the most comparable predictive map to the reference map; however, incorporating all data types would further increase the accuracies of the predicted map. This indicates that using various data types, ones with geological affinity, would deliver the most accurate predictions.

The same researcher, (Harris et al., 2014), conducted another geological mapping based on remote sensing data in concert with machine learning. In their study, they focused on comparing data from different satellites, including Landsat 7, Landsat 8, ASTER data and SPOT-5. They also integrated airborne magnetic data. To carry out this comparison, they used random forest (RF), a robust classification method using a maximum likelihood classifier (RCM-MLC). This study demonstrated how the acquisition conditions (e.g., sun angle, time of acquisition) of the satellite data could affect the predictions issued from them, which was exhibited by the low prediction accuracy of ASTER data that were obtained at low sun angle. Also, this study showed the importance of spectral resolution over spatial resolution. This was indicated by the low prediction accuracy of SPOT data compared to Landsat data despite having higher spatial resolution. Moreover, incorporating band ratios with the raw data of the satellite proved to have a positive effect on the prediction accuracy. This indicates their usefulness for separating lithologies. For the MLA comparison, this study showed that both algorithms achieved high classification accuracy. The RF might; however, be preferred because it is more flexible, less prone to overfitting and does not suffer from the Hughes effect (i.e., does not suffer from increasing the dimension of the input space). Because of these advantages, the researcher, Harris, conducted another experiment of geological mapping in (Harris & Grunsky, 2015) using the RF. This study did not incorporate satellite data but instead geochemical data which was integrated with geophysical data. In addition, to train the RF, they followed two methodologies for choosing the training sample. The first one consisted of using regionally distributed samples. The second method consisted of using training samples from field observation from older geological expeditions. The results of this study showed that RF accuracy was higher when field observations were used. Geochemical data also proved to produce more accurate predictions than geophysical data in both training methods. Therefore, including the geochemical data, if present, would additionally contribute to the MLA-based lithology models.

Cracknell and Reading (2013) conducted an interesting study in which they revealed the relationship between the uncertainty of MLA-based lithology models and lithological contacts. In this study, they used airborne geophysical data (Magnetic and gamma-ray spectrometry), Landsat 7 and DEM data to train a random forest (RF) and a support vector machine (SVM) lithology model. The authors found that the RF uncertain predictions were caused by lithological contacts. They also discovered that the degree of the uncertainty increased the closer the observation to be predicted to the contact the more they decreased the number of observations incorporated in training the RF model. Uncertain predictions were also related to the heterogeneous nature of the lithological units as well as using data with different characteristics (i.e., geometry and resolution). In contrast, uncertain predictions of the SVM were not in function of the distance of the observation from the lithological contacts and, therefore, could not be used to identify them. As a result, this study showed the superiority of the RF in utilizing patterns within remote sensing data and using it to identify geological areas related to either drastic changes in lithology or contact between lithological units.

In another study, Cracknell and Reading (2014) carried out a thorough MLA comparative study in the context of geological mapping. They employed Naive Bayes, k-Nearest Neighbors, Random Forests, Support Vector Machines, and Artificial Neural Networks, and they trained them with the same data that were used in their previous study. The main focus of this study was to evaluate the effect of the spatial distribution of the training samples (i.e., clustered or random distribution across the study area) as well as the effect of including spatial coordinates in training MLAs. This study pointed out that to generate high classification scores and increase the generalization capabilities of the MLAs, the training samples should be distributed across the whole area. Moreover, the inclusion of spatial coordinates as input for MLAs did increase the performance of the MLAs. However, the authors recommended that incorporating them as input for MLA training should be approached with great care in order not to hinder the ability of the MLAs models to predict lithologies in regions which are not located in the proximity of the training region. From a comparative point of view, in this study, the RF outperformed all other MLAs, as it was able to generate at least comparable or better classification scores compared to other MLAs. It was also the least sensitive to model parameters and the most computationally efficient.

4.1.4. Discussions and conclusions

The studies we have covered so far used different remote sensing data types and different machine learning algorithms for geological mapping purposes. These studies revealed that:

- Magnetic data can be used to predict bed formations, predict and update the faults system as well as contacts between lithological units; (Slavinski et al., 2010)
- Gamma ray spectrometry data is the most used geophysical data for geological mapping and also, when used, is the data type with the most contribution to predicting lithologies; (Harris et al., 2009; Harris & Grunsky, 2015; Yang et al., 1998; M. A. S. Youssef & Elkhodary, 2013)
- Satellite data and the ability to produce sub-images using a multitude of methods can contribute to the distinction between lithologies, especially surficial sedimentary formations and other well-outcropped rocks; (Harris et al., 2014; Kovačevič et al., 2009; Ourhzif et al., 2019; Thomas, 2020)
- Digital elevation models (DEM) and its derivative, such as slopes, topography curvature, and topographic position index can be used for: predicting fault systems, outlining lithological units limits and increasing the performance of MLAs in the context of predicting lithology; (Othman & Gloaguen, 2017; Slavinski et al., 2010; Thomas, 2020)
- From a comparative point of view, different MLAs were compared, Random Forest (RF) and support vector machine (SVM) proved to achieve better performance than traditional classifiers like maximum likelihood (MLC), and also better than the well-known supervised classifier artificial neural networks (ANN). (Bachri et al., 2019; Cracknell & Reading, 2013, 2014; Harris et al., 2014; Oommen et al., 2008)

Geophysical data is frequently used with other remote sensing data in concert with machine learning algorithms for geological mapping. However, the studies that include geophysical data have not been sufficiently dissected, leaving room for future insights. As future perspectives, we can consider an intercomparison of the geophysical data (including: magnetic, gamma-ray spectrometry and gravimetric data) to evaluate the contribution of each type to classify lithology; inspect the sensitivity of remote sensing data, especially geophysical data, to each rock class (i.e., sedimentary, metamorphic and igneous); integrate into predictive modelling maps derived from magnetic/ gravimetric data such as apparent susceptibility and apparent density. Despite being numerical filters, they may possess additional information about the physical properties of rock units. Most of the MLAs-based geological mapping studies use artificial neural networks (ANN), support vector machines (SVM), random forest (RF) or traditional classifiers like maximum likelihood (MLC). Yet, more advanced MLAs, such as gradient boosting (GBM), extreme gradient boosting (XGBoost) and deep neural networks (DNN) do not appear to have been studied. In other research fields, these methods proved to outperform RF and SVM. In geological mapping applications, they have good potential to produce more accurate predictions especially in terms of boosting methods-based MLAs. The geological phenomena can be defined as an imbalanced problem — that is, the distribution of the geological units is heterogenous which causes the occurrence of certain lithologies more than others. MLAs like RF and SVM usually outlook low-occurrence lithologies during the training phase, because of the insufficient number of training samples corresponding to this lithology. However, the boosting methods are trained in a way that makes them pay more attention to minority class (i.e., low-occurrence lithologies) without missing the others, and therefore these methods have the potential to capture the geological phenomena more accurately.

Chapter 5 : Machine learning applications in the geological mapping

Part I : Overview of Machine learning theory and algorithms

5.1. Introduction

Artificial intelligence (AI) and machine learning (ML) are the part of computer science which we use to create intelligent systems. Although the terms are sometimes used interchangeably, there is a difference between AI and ML. AI is a broader concept that aims to create intelligent machines that can simulate human thinking and behavior, whereas ML is an application or subset of AI that allows machines to learn from data without being explicitly programmed (Figure 5-1).

Machine learning, artificial intelligence and data science have been in the top trends for the last few years because of the availability and the ease in which we can gain access to information (every aspect of our lives has been computerized). Computerized systems allow for greater efficiency in performing specific tasks both more accurately and more rapidly than doing the same task using manual methods. Besides that, it makes it possible to store large collection of data in a relatively small amount of space (e.g., hard drives and USB drives). With the continuous convergence towards the computerization, the desire to use those easily accessible large data has grown more important than ever; therefore, the concept of machine learning has flourished.

The term Machine learning dates back to 1952 when an IBM expert, Arthur Samuel, wrote a program for playing checkers. For a long time, it was just a mathematical concpets. However, when ML started to become more feasible thanks to advances in computers, numerous industries started looking for ways to apply this empowering technology for their business purposes.

Machine learning is a system of automated data processing algorithms that provide tools for deducing unknown patterns and potentially useful information, which are relevant to our problem, from stored databases. The process of developing these kinds of tools has evolved throughout several fields such as chemistry, computer science, physics, and statistics and has been called "*Machine learning," "Artificial intelligence," "Pattern recognition," "Data mining," "Predictive analytics"* and *"Knowledge discovery."* While each field approaches the problem using different perspectives and toolsets, the ultimate objective is to make accurate prediction. The main idea behind ML is to build a computer program, that automatically analyses data to identify patterns that occur frequently. This is the learning process, where the computer program tries to learn how to identify these specific patterns. Consequently, make more accurate predictions about the future and similar data.

MLAs guide us toward more accurate and faster predictions, but like every new technology, they have their limitations, mainly because of the nature of data being fed to the computer program. Some of the limitations include inadequate data preprocessing, and inadequate validation model, or most importantly, overfitting of the MLA model during the training phase. Furthermore, when searching for predictive relationships, modelers often fail to recognize all the existing models. This is usually due to insufficient knowledge about the field, which we derived the data from, having expertise in only a few models or the restrictions caused by the employed software (M. Kuhn & Johnson, 2013).

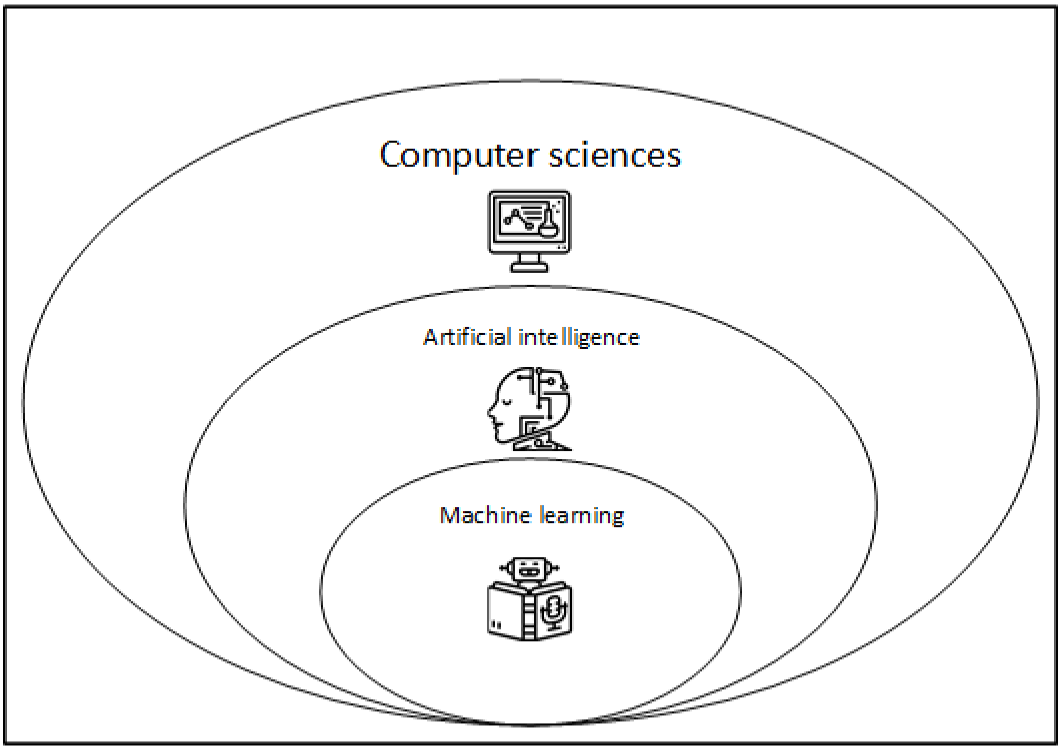


Figure 5-1 Representation of the computer science, the artificial intelligence and the machine learning, and how they are related.

5.2. Machine learning types

As mentioned before, after feeding the computer program, the latter tries to learn the patterns that reside within the data. There are many methods to achieve this, yet the literature recognizes three major categories: supervised learning, unsupervised learning and reinforcement learning.

5.2.1. Supervised learning

Supervised learning is the most popular paradigm for machine learning, because it is the easiest to understand, and to implement. It consists of using labelled data to train an MLA model (Figure 5-2). These data contain input observations as well as their corresponding output (i.e., label), allowing the computer program to produce a model that best estimates the relationship between the inputs and their outputs. The data used to train the model is referred to as *"training data."*

After training the MLA model, it is used to predict the output for unseen data with a similar structure as the training data called *"testing data."* Unlike the training data, the labels of these data are not provided to the trained model. By employing such a strategy, the classification accuracy of the MLA model can be assessed by comparing the outputs produced by the model and the real labels of the testing data. (Witten et al., 2011)

From a mathematical point of view, we describe the learning problem in supervised MLAs by:

- Input vector X of n instances X= ($x_{1,}$ x_2... x_n);
- An output vector Y of n instances corresponding to labels of the X vector Y= (y_1, y_2…. y_n);
- A learning machine (LM) capable of estimating a set of functions F (x, α), where α is a vector of parameters.

The problem of learning is to choose from a given set of functions F (x, α) the one that best approximates the real response Y. In other words, the learning process is an operation of minimizing the difference between Y and F(x, α), which is usually done by calculating the loss function defined as:

$$L(x,\alpha) \ = [Y - \ F(x,\alpha)] \tag{5-1}$$

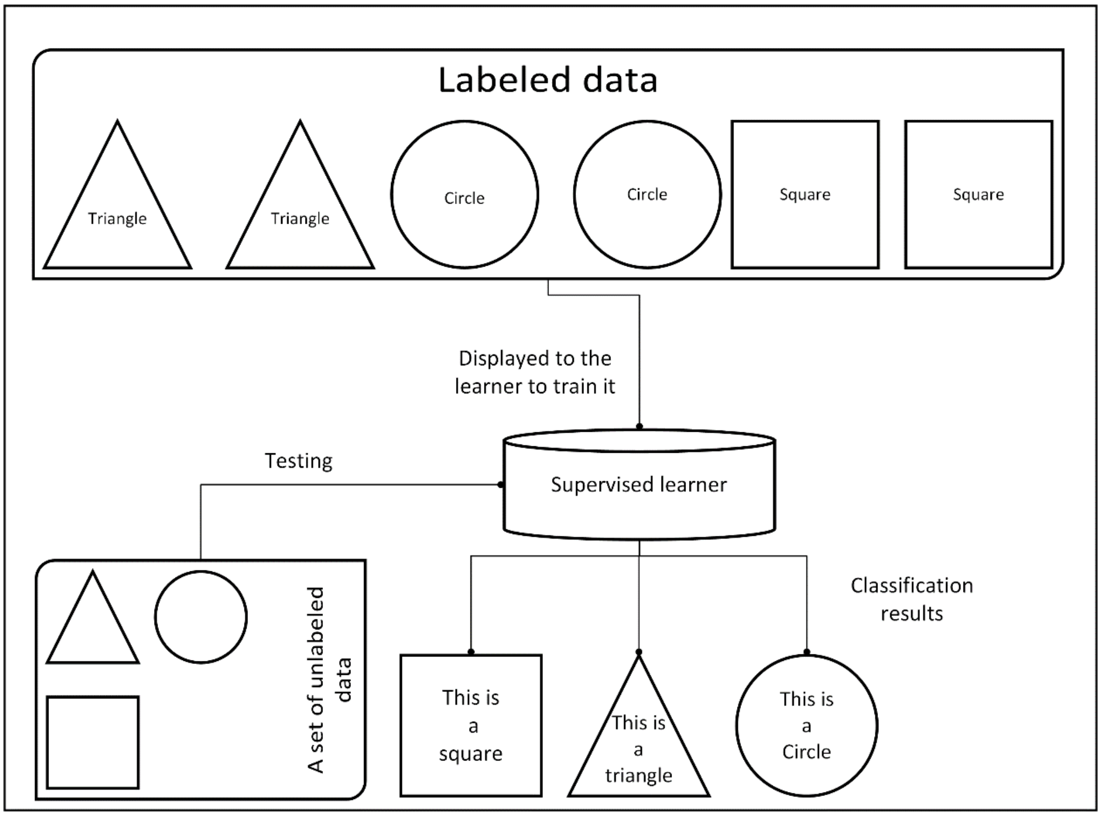


Figure 5-2 An example of a simple supervised classifier trained to recognize simple shapes.

In the MLA literature, when we try to predict a quantitative value, the learning process is called *"Regression* ", while learning a qualitative value is called *"classification* ". As a part of the research project, we are working on, we are concerned with predicting the type of lithologies/rocks which are provided as a string character (i.e., qualitative value); therefore, the learning problem we are facing is a classification problem.

5.2.2. Unsupervised learning

Unsupervised learners, unlike supervised learners, are not based on learning the relationships between the input-output pairs in the data, but rather on grouping observations with similar features in a way that great contrast between groups can be achieved. The groups are referred to as "clusters," and the function the governs the model assesses the contrast between them (Figure 5-3) (Dougherty et al., 1995). It should be noted that in the case of unsupervised learning, the procedure of splitting the data on training and testing sets, which is the common procedure to train supervised learners, is not applied, and the totality of data is used to train the model.

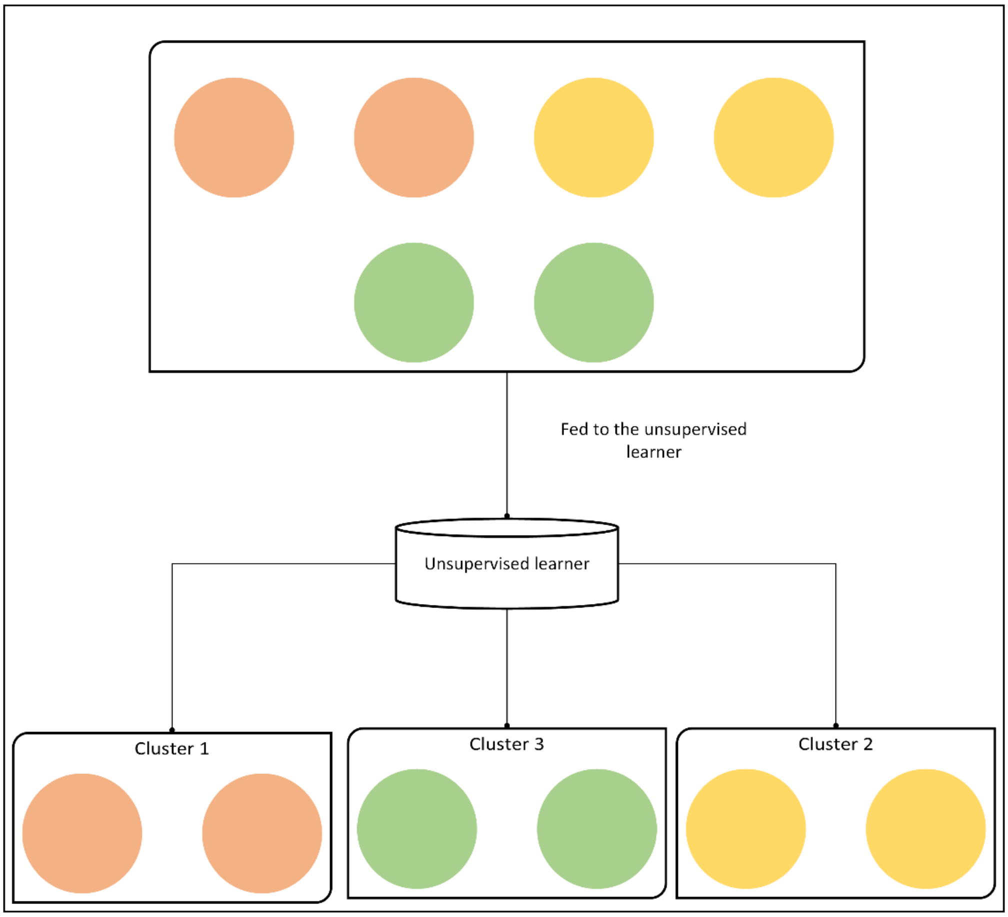


Figure 5-3 Illustration of how an unsupervised learner create clusters of observations with the same color.

MLAs based on unsupervised learning are usually utilized for a variety of reasons. For example, they are used to reduce the dimensionality of the features' space for visualizations purposes like the applications of the principal components analysis (PCA). Likewise, they are used to determine the distribution of the features' space like the application of K means clustering (Berry et al. 2020), and finally, they can also be used to render unlabeled data more suitable for the supervised learning algorithms. (Hofmann, 2001)

5.2.3. Reinforcement learning

Reinforcement learning fills the gap between supervised learning, where the algorithm is trained using correct answers provided in the target data, and unsupervised learning, where the algorithm clusters data based on similarities. The middle ground is where information is provided about whether or not the answer is correct, but not how to improve it. The reinforcement learner has to try out different strategies and see which works best. (Marsland, 2014)

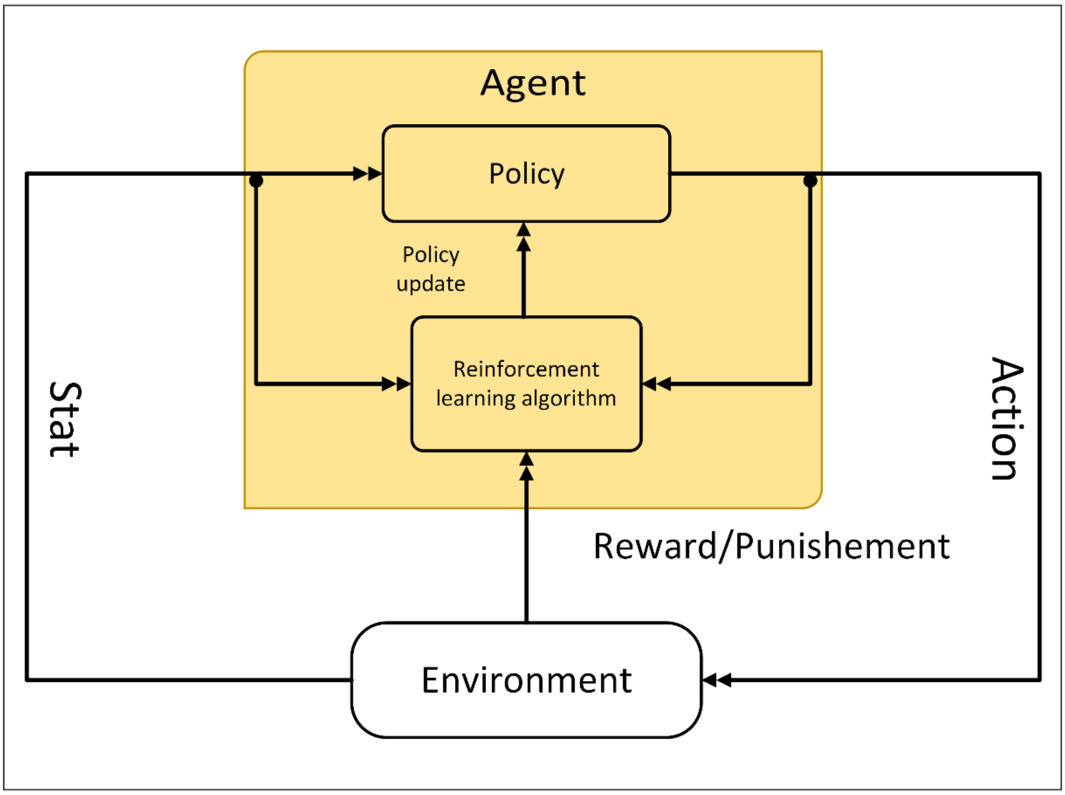


Figure 5-4 Illustration of a reinforcement learner working.

Reinforcement learning is a computational approach to learning from interactions with an environment. It includes four steps: a policy, a reward signal, a value function and an environment model (Figure 5-4). The policy defines the reactions that would be taken by the agent when facing specific states of the environment; The reward defines the goal (or punishement) in the reinforcement learning problem. Each time the agent behaves in a certain way within the environment, it would be rewarded (or punished) by the environment through a single number, and the main objective of the reinforcement learners is to maximize the total reward; the value function determines what is good in the long run by calculating the total amount of rewards an agent can expect to accumulate over time; the environment model controls how the environment will behave. This model is used for planning, by considering future situations before experiencing them and also predicting the next (state, reward). (Sutton & Barto, 2015)

5.3. Supervised learning strategies

These MLAs are classified into: statistical learning algorithms, instance-based learners, logic-based learners, Support Vector Machines and Perceptron (Kotsiantis et al., 2007). In this research project, we only utilized logic-based and Perceptron algorithms, so in the following section, we overview the theory behind these two strategies. The other strategies have not been considered within the scope of the objectives of this thesis.

5.3.1. Logic-based algorithms

This strategy is a way to represent rules dwelling in data which attaches a set of output classes (label) with its appropriate instance (input features) with a hierarchical sequential structure that recursively split data using a series of (*"if", "else if", "then"*) statements into more homogenous subdivisions. Decision Trees (DT), introduced by (Breiman et al., 1984), is one of the most used logic-based algorithms, and it is the foundation for many advanced algorithms.

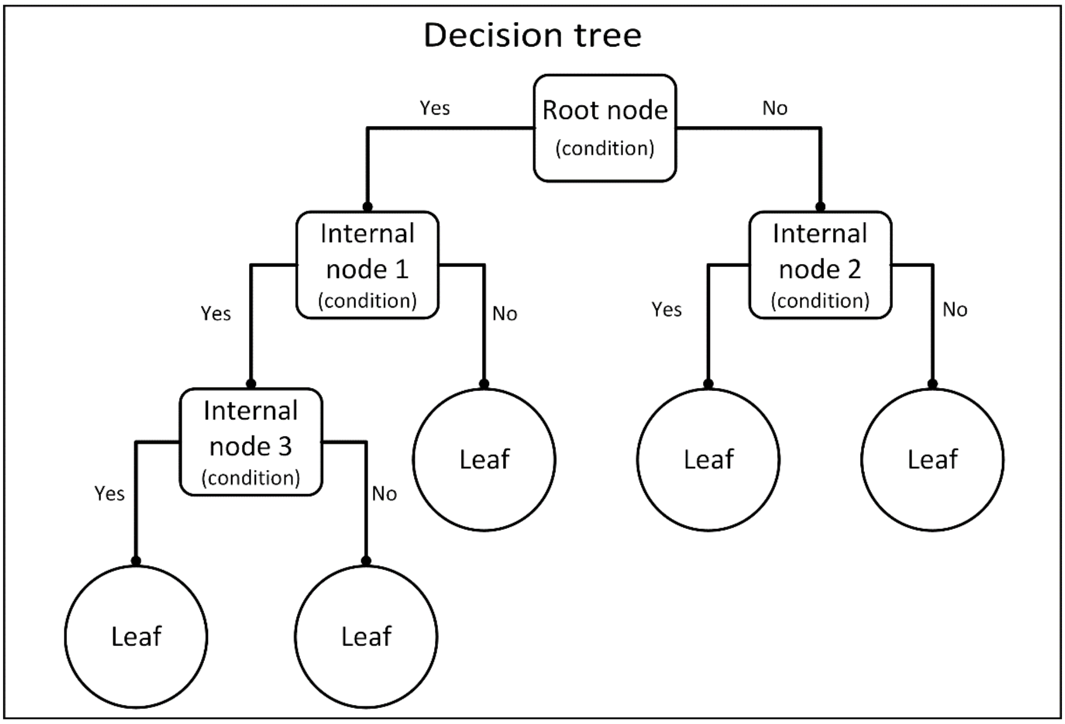


Figure 5-5 Illustration of a decision tree with five possible outcomes.

In a DT algorithm, the task of constructing a tree from the training set has been called tree induction, tree building and tree growing. Most decision trees algoirhtms proceed in a greedy top-down manner, also known as divide and conquer. Greedy algorithms work by making the decision that seems most promising at any moment without considering whether or not their decision would lead to the overall optimal solution. The top-down approach consists of splitting complex data into smaller fragments, called *"modules."* The same procedure is repeated on each module until it can no longer be split. The building process (Figure 5-5) starts by sorting the observations using their features, and then it searches for the feature that best split the training data. This feature would be the *"Root node."* By using the same procedure, the internal nodes (also known as non-terminals or test nodes) are created. Each branch represents a value that the node can assume. When the training data can no longer be discriminated, the resultant node is called the leaf node (also known as the terminal or decision node).

An observation x_i is classified by testing each of its features starting at the root node down to a leaf. The label at the leaf node at where the observation ends up is its output class. Each leaf can hold either a class output or a probability vector of multiple classes. In the case of univariate trees, choosing which feature to employ when we create the internal nodes corresponds to finding a decision rule based on feature value that best split the training set into more homogenous parts, while in the case of multivariate trees, finding a split corresponds to a combination of existing features. Most of the multivariate splitting criteria are based on a linear combination.

5.3.1.1. Splitting criteria

Splitting criteria are also known as goodness measures, feature evaluation criteria, feature selection criteria, impurity measures or splitting rules. There are multiple methods used to achieve the maximum node purity (Breiman et al., 1984), and the most common methods are:

- **Gini index:** it is the probability of incorrectly classifying a randomly chosen element in the dataset if it were randomly classified according to the class distribution in the dataset.

$$\text{Gini index} = 1 - \sum_{i=1}^{C} p_i{}^2 \qquad (5\text{-}2)$$

Where C is the number of classes, and p_i is the probability of randomly picking an element of class i.

When training a decision tree, the best split is chosen by maximizing the *"Gini Gain,"* which is calculated by subtracting the weighted impurities of the branches from the parent node. The weighting depends on the number of elements in each child node.

$$\text{Gini gain} = \text{Gini index(parent node)} - \text{Weighted Gini index(child nodes)} \qquad (5\text{-}3)$$

- **Information gain:** it is an impurity-based criterion that uses the entropy measure, which is the degree of uncertainty, impurity or disorder. It aims to reduce the level of entropy starting from the root node to the leaf nodes. It is used to determine which feature gives the maximum information about a class. For a given node, the entropy is calculated as below:

$$E(S) = -\sum_{i=1}^{C} p(i) * \log_2 p(i) \qquad (5\text{-}4)$$

Where C is the number of classes, and p_i is the probability of randomly picking an element of class i.

Next, we need to know how much entropy we removed. This is where Information Gain comes in. Mathematically it can be written as:

$$\text{Information gain} = \text{Entropy(parent node)} - \text{Weighted entropies(child node)} \qquad (5\text{-}5)$$

Higher information gain ties in with removing more entropy, which increases the purity of the child nodes. We can refer to the reduction of entropy as the amount of the earned information from the child node about the parent node.

5.3.1.2. Offsetting the over-fitting problem

The decision tree representation is often prone to over-fitting (Hastie et al., 2009), so the two following strategies have been used to circumvent this issue:

- Define a stopping criterion to prevent it from reaching a point where it perfectly fits the training data. The following stopping criteria are commonly used:
  - All instances in the training are classified into a single class;
  - The maximum tree depth has been reached;
  - The number of cases in the leaf node is less than the minimum number of cases in the parent node;
  - The number of cases in one or more child nodes is less than the defined threshold for the minimum number of cases for child nodes;
  - The splitting criteria value has not exceeded a certain threshold.
- Pruning the Decision Tree. It consists of defining a loose stopping criterion and letting the decision tree over-fit the training data. Then the over-fitted tree is cropped into a smaller tree by eliminating subdivisions (subtrees) that have a high misclassification rate.

The conducted studies indicate that in most cases, changing the splitting criteria will not make much difference in the overall performance. Because it is hard to come up with a balanced one, since employing tightly stopping criteria tends to create small and under-fitted decision trees, while using loosely stopping criteria tends to generate large decision trees that are over-fitted to the training set. However, using a pruning method proved to improve the performance of the decision tree.

5.3.2. Perceptron-based algorithms

The human brain is a parallel-distributed computational system composed of an astronomical number of interconnected neurons (approximately 86 billion). By assigning each of these neurons to solve specific problems, even complex processing problems such as image recognition can be carried out in a way that might seem immediate without exposing the system to a strenuous effort. Studying how this unique structure works inspired the development of the artificial neural network, thus the name *"Artificial neurons."*

The very first mathematical model of an artificial neuron was the "*Threshold Logic Unit,"* and it was proposed in 1943 by the American neurophysiologist Warren S. McCulloch (1898–1969) and the American logician Walter H. Pitts Jr (1923–1969). In 1958, F. Rosenblatt suggested the first model of a learning machine, called the *"Perceptron."* He described the model as a program for computers and demonstrated with simple experiments that this model can be generalized.

Typically, a simple artificial neural network (ANN) is composed of three layers, input, output and a hidden layer placed in the middle. Each layer is composed of a set of very simple processing elements called nodes. When there is more than one hidden layer, it is called a deep neural network (DNN). The nodes are interconnected with what is known as *"synaptic weights"*. They represent the parameter that the neural network seeks to optimize to increase the predictability power.

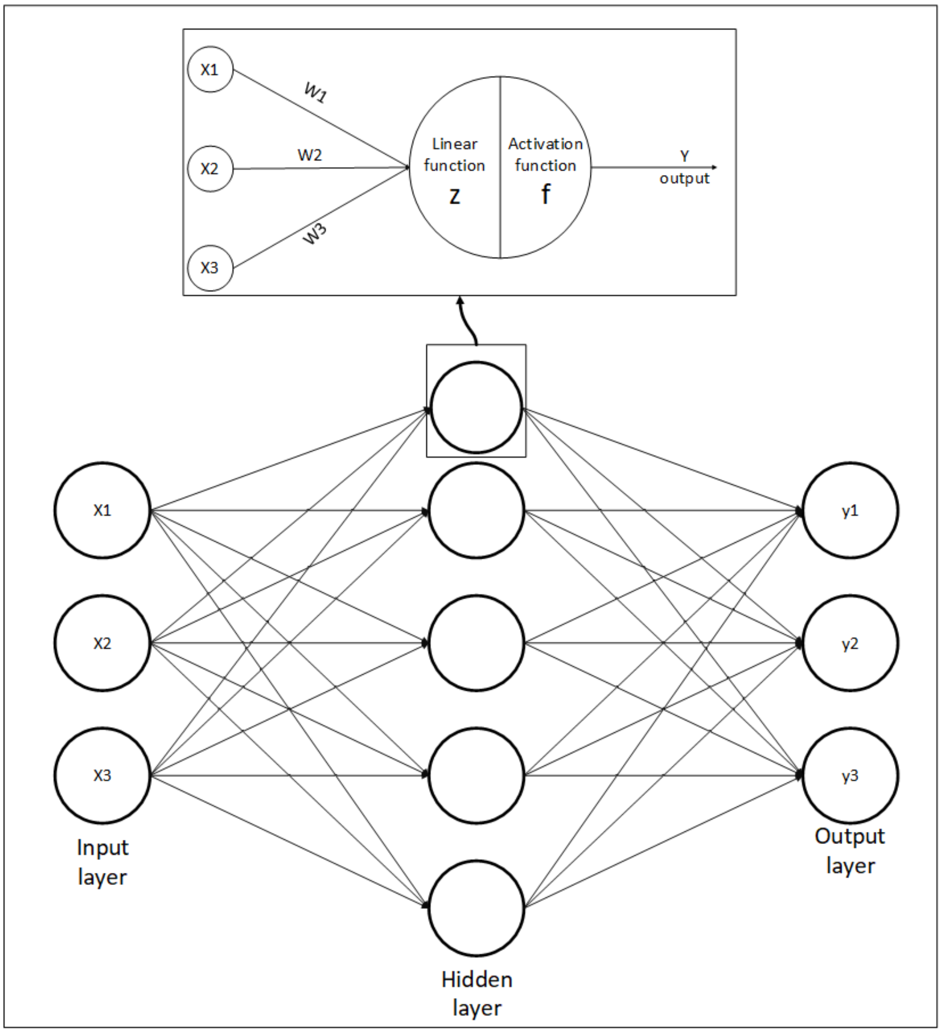


Figure 5-6 A simple ANN architecture with a single hidden layer, 3 input features and 3 possible outcomes. (The bias nodes are not shown in the figure)

In the neural network literature, the output of a single neuron is called *"Activity."* For instance, the activity of the input layer is represented by the actual value of the observations of the dataset introduced to the net, and each node represents a feature. This activity is passed to the adjacent layer (hidden layer) through the synaptic weights. Upon receiving the activity, the hidden nodes sum up the received weighted activities and perform a nonlinear transformation on the summation using an *"activation function."* This function usually serves to normalize the activation of the hidden nodes. All these procedures work for calculating the activity of the nodes in the hidden layer. Again, the hidden nodes are also linked to the output layer and the same calculations are repeated to calculate the final output, which is in turn compared to a pre-specified threshold to convert the final scoring into a class label (Figure 5-6). Generally, a single bias node is added to the input layer and every hidden layer in the net to increase its flexibility to better fit the data. It is a constant value that will always have the value 1.

There is a variety of activation functions, and the most commonly used ones are addressed in the table below:

Table 5-1 Examples of the most used activation functions in ANN-based algorithms.

Step function	$\mathrm{f(x)} = \begin{cases} 1, & \mathrm{x} \geq \theta \\ 0, & \mathrm{x} < \theta \end{cases}$
Sigmoid function	$f(x) = \dfrac{1}{1+e^{-x}}$
ReLu function (Rectified Linear Unit)	$f(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}$

5.3.2.1. The backpropagation algorithm

As mentioned before, the net seeks to enhance the accuracy of which it can predict the introduced data by tuning the weights. Many learning algorithms can be used, but the *"Backpropagation"* algorithms tend to work quite well in most cases.

The backpropagation is a gradient descent algorithm that uses derivatives to update the weights. By calculating the error, which is the difference between the predicted and real values, and using the *"Chain rule* "of derivatives, the weights can be updated. This process begins with the weights that connect the hidden nodes to the output node, and then it propagates to the nodes that link the inputs with the hidden nodes. This backward implementation is known as backpropagation.

Let: $e_i = \frac{1}{2}(y_i - f(x))^2$

Where: e_i is the squared loss function to be minimized;

y_i: the real value;

f(x)*:* the predicted value;

X_i: input value.

To update the weights, we need to differentiate the loss function according to two parameters w_{ij} and w_k which are respectively: the weights that link the hidden nodes with the input node, and the weight that connect the output nodes to the hidden nodes.

$$\frac{\partial e_i}{\partial w_{ij}} = \frac{d\left[\frac{1}{2}(y_i - f(x))^2\right]}{dw_{ij}} \quad (5\text{-}6)$$

$$\frac{\partial e_i}{\partial w_k} = \frac{d\left[\frac{1}{2}(y_i - f(x))^2\right]}{dw_k} \quad (5\text{-}7)$$

Where: $x = \sum f(z)\, w_k$, and: $z = \sum w_{ij} x_i$.

- The partial derivative of equation (5-6) gives:

$$\begin{aligned}\frac{\partial e_i}{\partial w_{ij}} &= -(y_i - f(x))\frac{\partial f(x)}{\partial w_{ij}} \\ &= -(y_i - f(x))f\grave{}(x)w_k f\grave{}(z)x_i\end{aligned}$$

With $\frac{\partial f(x)}{\partial x_{ij}} = \frac{df(x)}{dx}\frac{dx}{dw_{ij}}$, and:

$$\begin{aligned}\frac{dx}{dw_{ij}} &= w_k \frac{df(z)}{dw_{ij}} \\ &= w_k f\grave{}(z)\frac{dz}{dw_{ij}} \\ &= w_k f\grave{}(z)x_i\end{aligned}$$

- The partial derivative of equation (5-7) gives:

$$\begin{aligned}\frac{\partial e_i}{\partial w_k} &= -(y_i - f(x))\frac{\partial f(x)}{\partial w_k} \\ &= -(y_i - f(x))f\grave{}(x)f(z)\end{aligned}$$

With $\frac{\partial f(x)}{\partial w_k} = f\grave{}(x)f(z)$.

Consequently, we can obtain the new weights by multiplying the partial derivatives with the learning rate "η" and then subtracting it from the current weights

$$\Delta w_{ij} = \eta(y_i - f(x))f\grave{}(x)w_k f\grave{}(z)x_i$$

$$\Delta w_k = \eta(y_i - f(x))f\grave{}(x)f(z)$$

The choice of an optimal learning rate is crucial because a small value tends to make the search process very slow and more computationally intensive. Furthermore, the possibility of falling into a local minimum is more likely to happen. On other hand, a big value might completely overstep the optimal solution.

There are two approaches to updating the weights. The first one is called the *"on-line "or "Stochastic gradient descent"* method. The synaptic weights are updated on an observation-by-observation basis. The second approach is called *"batch learning."* In this approach, the weights are updated after presenting all the observations in the training data to the net, which implies that we need to store the data during the training of the network.

The on-line training is where data is provided sequentially but not in full, meaning that some of the data is left out. This training approach is most common when the dataset is too large to be handled or computed all at once, so we break up computations to handle the size of the data. The term online comes from some earlier (and still very relevant applications) where we did not "divide" the data due to technical constraints (storage and computational complexity), but because the data in its nature was obtained only sequentially and incompletely (data from the stock market is a very common example). On other hand, batch and mini-batch training are how we describe computations where it is possible to process the dataset altogether. In both mini-batching and batching, you are providing some portion of the dataset during each iteration (where generally, loss/gradients are computed, and back-propagation occurs). When we use the term "batching" , we are generally stating that this portion is 100% of the dataset, whereas in mini-batching, we process smaller portion, such as 10% of the dataset during each iteration.

5.3.3. Prediction error

In supervised machine learning problems, the prediction errors of the predictive models are due to three main components. These are: *"irreducible error," "bias"* and *"variance"* (equation 5-8). Irreducible error is the component of the random error brought by the noise of the data. This error cannot be estimated by any ML model; bias is brought by the limitation of the learning method utilized to estimate the relationship between inputs and outputs. In other words, this error component defines how much the learning method suits the learning problem; variance defines how the predictions of the learning method are affected by small changes of the training data (i.e., prediction stability). The last two components are inversely related, meaning that models with high variance will have low bias, and models with high bias will have low variance (M. Kuhn & Johnson, 2013; Witten et al., 2011).

$$Expected\ error = \sigma^2 + bias^2 + variance \tag{5-8}$$

Usually, complex MLAs with high variance (e.g., decision trees) have enough flexibility to learn complex data, so they tend to perfectly model the training data; however, they badly model the testing data, implying a weak generalization capability. These models are described as being *"over-fitted"*. On other hand, simple MLAs with high bias (e.g., linear models) do not have enough flexibility to model the training data. These models are described as being *"under-fitted"*. The final objective of modeling with MLAs is to find a good balance between prediction errors: namely bias and variance and to train a model that is neither over-fitted nor under-fitted.

5.3.4. Ensemble methods

The ensemble method is a machine learning technique that aims to minimize prediction error and produce a more stable and accurate ML model. To achieve that, the predictions of several ML models are combined, and depending on whether we are dealing with regression or classification problems, the outcome is inferred by averaging or by counting the votes of the ML models (Witten et al., 2011). The three main ensemble methods are: *"bagging," "boosting"* and *"staking."*

5.3.4.1. Bagging

Bootstrapping is a resampling method that consists of drawing a random observation with replacement from data, and the objective is to create another data with the same observation number as the original one. Owing to using random drawing with replacement, the new data, called *"bootstrap data, "*do not necessarily contain the same observation distribution as the original data, and a particular observation in the original one may occur multiple times or have no occurrence at all in the new data (the latter are called *"out-of-bag* "observations). Therefore, applying this resampling method to the same data multiple times can create each time different dataset with different observation distributions. One of the many usages of this method is to produce several data and use them to train ML-based models to build an ensemble model such as bagging. Bagging is an ensemble method that was proposed by (Breiman, 1996). The word bagging stands for *"bootstrap aggregating,* "and as the name suggests, this method utilizes bootstrapping in conjunction with a particular ML method, called the *"base model"* of the ensemble, to build an ensemble of predictive models. After that, the predictions issued from these models are aggregated to produce the final prediction of the bagged model. One of the most popular bagged models is bagged trees which has decision trees as a base model Figure 5-7 (M. Kuhn & Johnson, 2013).

The bagging method brings the advantage of greatly benefitting the performance of the ML-based models characterized by a high variance, such as decision trees, and allows training a stable ensemble model that is not affected by changes in the training data, meaning that when adding new observations, the predictions of the ensemble is rarely changed. On other hand, the bagging method is less beneficial for stable ML-based models (i.e., low variance). Another advantage of bagging is that it offers a built-in prediction accuracy evaluation. This advantage is due to the bootstrapping that gives each trained model of the ensemble the ability to evaluate their prediction accuracy on the out-of-bag observations that were not used to train them. The prediction accuracy of the ensemble is the averaged accuracy of the models that constitute the ensemble. This feature, however, comes with an intensive calculation and a big load on the storage capacity of the machine.

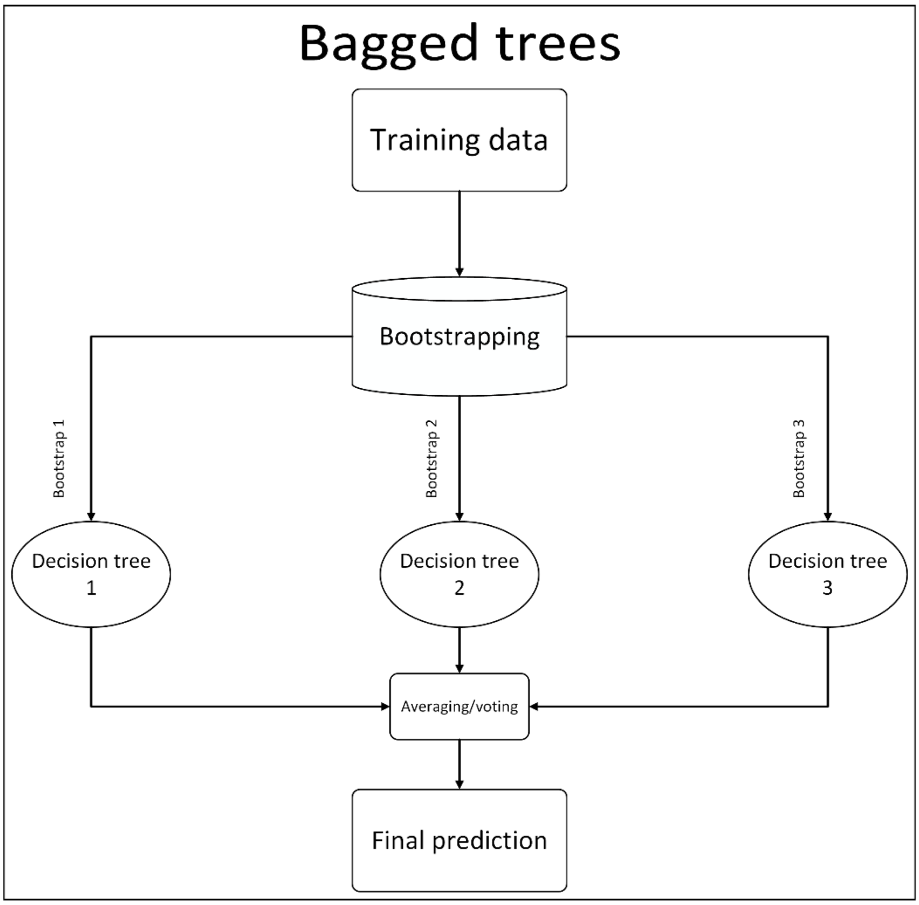


Figure 5-7 Illustration of a bagged trees (i.e., the decision tree is the base model of the ensemble). The training data is resampled using bootstrapping, and each decision tree is trained on different bootstrap data. The final prediction of the ensemble is the average/ majority votes of the three trees.

The purpose of introducing bootstrapping in the bagging method is to induce some randomization in the process of training the trees of the ensemble. The reason for that is to train an ensemble of independent trees with each tree, depending on its structure, having the ability to generate different predictions for the same observation. However, in big data in which the trees of the ensemble can, to a certain degree, estimate the relationship between inputs and output, and also the fact that all the features of the data are considered at every split of every tree, lead to producing trees that have a similar structure (i.e., same features used to split data at the root node as well as the internal nodes). This may generate similar predictions, and ultimately limit the bagging method of optimally reducing the variance of predictions (M. Kuhn & Johnson, 2013). This is referred to as *"tree correlation,"* and to ensure that bagging would not be affected by this, another aspect of randomization was implemented in the process of training the trees. Instead of considering all features at every split, only a random subset of features is used. This tweaking was proposed by (Breiman, 2001), and this new algorithm is known by the name *"Random forests."* According to the original author, random forests would not be affected by tree correlation, would suffer less from over-fitting when increasing the number of trees of the ensemble contrary to the original implementation, would not be prone to the noise of the data, would be more beneficial for low variance ML-based models than the original bagging algorithm, and because only a subset of features is considered at every split, it would be more computationally efficient than its predecessor algorithm.

5.3.4.2. Boosting

Boosting as an idea goes back to the 1990s. It consists of using a weak learner, defined as any ML method with a prediction accuracy slightly better than 50%, and constructing an ensemble of weak, yet complementary models in a way that each model tries to correct the prediction errors of the previous models. This idea was effectively implemented with the *"AdaBoost"* algorithm by (Schapire, 1999). Initially, AdaBoost was only designed for classification problems, but after that, (J. Friedman et al., 2000; J. H. Friedman, 2001) developed what is known as *"Gradient boosting machines (GBM),"* a modified version of AdaBoost, which is not only an expansion of boosting idea into regression problems but also it is the basis for more recent boosting-based algorithms.

Gradient boosting machines require a differentiable loss function (e.g., mean squared error) and a weak learner. Any ML method can be transformed into a weak learner, and subsequently, boosted into a stronger learner, but the tree paradigm is a popular choice for boosting because it can be easily transformed into a weak learner by just restricting their depth (M. Kuhn & Johnson, 2013).

The idea of building the ensemble in the GBM is based on finding a solution (i.e., predicted value) that minimizes the loss function across all the models of the ensemble. The algorithm builds the first model of the ensemble by minimizing the loss function according to the predicted value. This first model usually corresponds to a constant value, and it is taken as the initial prediction for all observations of the data. After that, the second model, and every subsequent model, are built following the next three steps:

- Calculate the prediction residuals of the previous model (the first model in this case) by taking the gradient of the loss function. If the mean squared error is chosen as the loss function, this step corresponds to calculating the difference between the real value and the prediction of the first model;
- By employing the residuals, the algorithm, after that, fits a model while taking these residuals as the outcome. If a decision tree is taken as a base model, the model is fit in the same way as shown in Figure 5-5;
- After fitting the model on the residuals, and similarly how the algorithm produced the initial prediction, the output of the model is calculated by minimizing the loss function according to the predicted value. This time, however, to take into account the contribution of the previous models, the previous predictions are also included in the minimization procedure.

These steps are repeated to train all the subsequent models, and the final prediction of the ensemble is obtained by summing the initial prediction and the output of the models that were calculated in the third step. To prevent the ensemble model from over-fitting the training data, the output of the models is scaled with what is known as the *"learning rate λ"* (equation 5-9) (M. Kuhn & Johnson, 2013).

$$Ensemble\ prediction = intial\ prediction + \lambda \sum output\ of\ the\ models \qquad (5\text{-}9)$$

Like bagging, boosting ensemble has a particular base model, and the final prediction is produced by aggregating predictions of the models of the ensemble. Unlike bagging, the models in boosting are trained sequentially and the performance of every model is dependent on the other models, whereas models in bagging are trained independently and the performance of every model is not related to the other models.

5.3.4.3. Stacking

The stacking idea, short for stacked generalization, was proposed by (Wolpert, 1992). Unlike bagging and boosting, it combines different ML methods into an ensemble model. In this way, the stacked model allows for minimizing the prediction bias by exploiting the compatibility of each ML method towards particular learning problems. Although the idea of stacking was proposed in the 1990s, it was not widely used like the other ensemble methods. The reason for that was the theoretical foundations behind these algorithms were not sufficiently developed; therefore, they were difficult to analyze. However, this changed when (Van Der Laan et al., 2007) introduced the *"super learner."* This work re-introduced the concept of stacking and set the foundation for newer stacked models.

The structure of the ensemble in a stacked model is built differently from the previous ensemble methods in how the prediction of the different models is aggregated. In stacking, each model is trained separately, and the predictions of these models, instead of simple aggregation, are fed to another algorithm to train a final model that would produce the final prediction of the ensemble. The models that were used to produce the first predictions are called *"level 0 models,"* and the model that was trained using level 0 predictions is referred to as the*" level 1 model"* or *"meta learner."* The objective of the latter is to learn how to combine the predictions of the level 0 models, and then leverage the predictions of the model with the most reliable predictions for the learning problem (Witten et al., 2011).

5.3.5. Model evaluation and experiments

As mentioned in section 5.2.1, the training of a machine learning model starts by confronting an algorithm with a set of labelled data from which it can learn the relationship between the inputs and their labels. After that, to evaluate the performance of the trained model, it is tested on other data with no labels. In this section, we present the most used parameters, called *"performance metrics,"* to evaluate the prediction performance of ML-based models. Since the learning problem in this dissertation is a classification problem, the metrics that are used to assess the performance of the regression problems are not included in this section. Moreover, throughout this section, we use *"iris"* data to demonstrate how performance metrics are used to assess the predictions of the MLAs.

Iris data was originally introduced by (Fisher, 1936). It includes the length and the width of the sepal and the petal of different species of iris, namely: *"setosa," versicolor"* and *"virginica"*. We use this data to train a decision tree (DT) model to predict species type, and after that, we evaluate the prediction performance of the model using different performance metrics. This data was chosen for two reasons: it has a small dimension (150 rows×5 columns), leading to decreasing both the computation time and the computation burden of training the MLA model; the output (i.e., species) is a categorical variable which implies that modelling this data is a classification problem; therefore, that would better serve the demonstration purposes of this section.

5.3.5.1. Training the model

Usually, before starting any modelling procedure, the data should be preprocessed by inspecting, and consequently, cleaning them from any anomalous values, including: dummy values, outliers and missing observations. This step, however, is skipped because the data at hand are considered clean and ready for modelling.

The original data is randomly split on two parts, namely: training and testing data. The proportions of these parts are 75% for the training data and 25% for the testing data. Also, the split is carried out in a way that the observation distribution according to the output column (i.e., species) in the training and the testing data is the same as the original data (Table 5-2). This splitting methodology is called *"stratified sampling.* "According to the species column (i.e., prediction target), there are three possible classes; therefore, the classification problem for the iris data is called *"multiclassification."* This term describes any data where the output column has at least 3 possible classes. Moreover, the observation count for these classes is equal. In the machine learning literature, this data structure is referred to as *"balanced data."*

While training MLA models, the main objective is to train a model that can capture the relationship between inputs and output without being over-fitted. This corresponds to searching for a vector of optimal parameters α that can minimize equation 5-1, and at the same time, not induce too much complexity to the model. This procedure is known by the name *"hyperparameter tuning."*

Hyperparameter tuning is carried out by setting a grid of random values. In this grid, the columns correspond to the parameters of the model, and the rows represent the parameters combination used to train the model. In our case, the grid values have three columns, indicating that a decision tree has three tunable parameters, and for each one, we chose 3 random values; therefore, the resultant grid has a dimension of (3×27). This procedure is usually very taxing on the machine and the majority of the training time is spent while tuning hyperparameters.

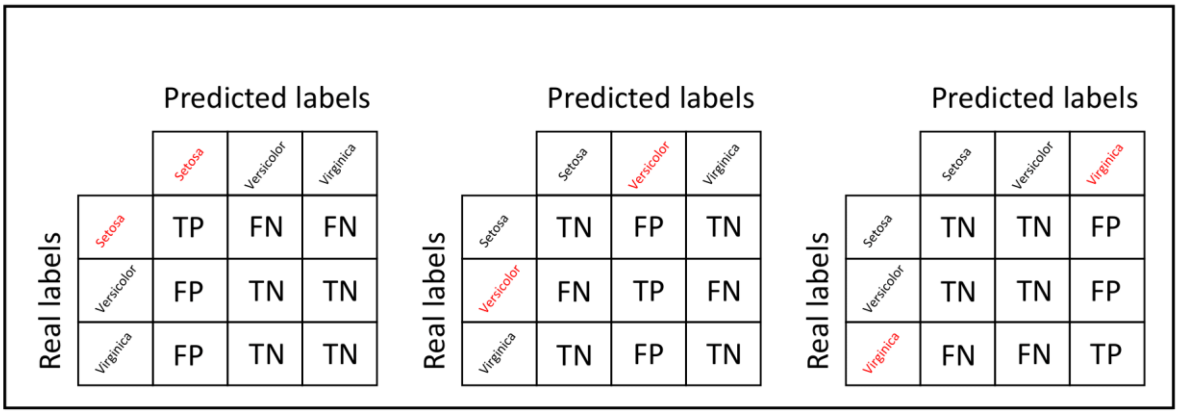
Table 5-2 The count of observations and their proportions according to the output column in the original, training and testing data.

Species	*Original data*		*Training data*		*Testing data*	
	Observation count	Proportions	Observation count	Proportions	Observation count	Proportions
Setosa	50	33.33%	37	33.33%	13	33.33%
Virginica	50	33.33%	37	33.33%	13	33.33%
Versicolor	50	33.33%	37	33.33%	13	33.33%

5.3.5.2. Evaluating the model

The evaluation procedure is executed on the testing data. As discussed before, the labels (i.e., species) of the testing data are not shown to the trained tree. Like this, we can use the model to predict species for the testing data, and afterwards, compare these predictions with the real labels. Usually, to compare the predicted labels and the real labels, they are presented in a tabular form called a *"confusion matrix."* From this matrix, we can define the following terms: *"true positive (TP)," "true negative (TN)," "false positive (FP),"* and *"false negative (FN)."*

During the evaluation, we usually assign one of the classes as being the positive case (i.e., the class that model should focus on while predicting), and all other classes would be set as the negative case; therefore, TP is the observations of the positive case that were correctly predicted; TN is the observations of the negative cases that were correctly predicted; FP is the observations that do not



	Setosa	Versicolor	Virginica
Setosa	TP	FN	FN
Versicolor	FP	TN	TN
Virginica	FP	TN	TN

	Setosa	Versicolor	Virginica
Setosa	TN	FP	TN
Versicolor	FN	TP	FN
Virginica	TN	FP	TN

	Setosa	Versicolor	Virginica
Setosa	TN	TN	FP
Versicolor	TN	TN	FP
Virginica	FN	FN	TP

Figure 5-8 the label of each cell in the same confusion matrix depending on which class is chosen as the positive case. The positive case is highlighted in red.

pertain to the positive case, but they were predicted as being positive cases; FN is the observations that do pertain to the positive case, but they were predicted as being negative cases.

For our tree model, the confusion matrix is shown in Table 5-3, and in Figure 5-8, we show for different choices of the positive case, the labels of each cell in the confusion matrix to facilitate the extraction of the TP, TN, FP and FN values.

Table 5-3 Confusion matrix of the trained tree. The rows correspond to the real labels of the testing data, and the columns correspond to predicted labels. The numbers are the observation count.

Real labels	*Predicted labels*		
	Setosa	Versicolor	Virginica
Setosa	13	0	0
Versicolor	0	12	1
Virginica	0	2	11

5.3.5.3. Performance metrics

Accuracy

The accuracy is the ratio of the correct predictions, for both the positive and the negatives cases, over the total number of observations in the data. It can be written as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{5-10}$$

$$Accuracy = \frac{13+12+11}{13+12+11+2+1} = 0.92$$

This metric gives an overall sense of the performance of the trained model, and it is the easiest metric to interpret. The main drawback of the accuracy is that it can be a misleading metric. For example, let's say we want to evaluate the predictions of data with two possible outcomes A and B, and the distribution of the observations according to these classes is highly unbalanced (e.g., 95% class B; 5% class A). Using the accuracy for such data would be misleading because if the model was able to correctly predict all B observations, and at the same time, fail to predict all the A observations, the testing performance would still indicate an almost perfect accuracy score. In the case when class A has the highest priority to be correctly predicted, using this model would not be ideal despite the near-perfect accuracy score.

Kappa statistic

The Kappa statistic, or Cohen’s kappa, is a metric that is used to measure the agreement between the predicted labels and the real labels. The main advantages of the kappa statistic come from the fact that it takes into account the class distribution of the output column and also subtracts the probability that the classification model produces a correct prediction by chance. It can be calculated as follows:

$$K = \frac{P_0 - P_e}{1 - P_e} \tag{5-11}$$

Where: P_0= the observed accuracy. It can be calculated using the formula in (5-10);

P_e= the expected accuracy that would be produced by chance.

Returning to our example, the probability that the classification model agrees with the real labels can be defined for every class by the probability of an observation being a Setosa and the probability of predicting a random observation as a Setosa; the probability of an observation being Versicolor and the probability of predicting a random observation as a Versicolor, or lastly, the probability of an observation being a Virginica and the probability of predicting a random observation as a Virginica. Using the confusion matrix in Table 5-3, we can calculate these probabilities as follows:

For the Setosa: the probability of an observation being a Setosa is: $\frac{13}{39} = 0.33$;

the probability of predicting a random observation as a Setosa is: $\frac{13}{39} = 0.33$.

For the Versicolor: the probability of an observation being a Versicolor is: $\frac{13}{39} = 0.33$;

the probability of predicting a random observation as a Setosa is: $\frac{14}{39} = 0.36$.

For the Virginica: the probability of an observation being a Virginica is: $\frac{13}{39} = 0.33$;

the probability of predicting a random observation as a Virginica is: $\frac{12}{39} = 0.30$.

The expected accuracy, therefore, can be calculated as follows:

$$P_e = (0.33 \times 0.33) + (0.33 \times 0.36) + (0.33 \times 0.30)$$
$$P_e = 0.33$$

Finally, the Kappa statistics is:

$$K = \frac{0.92 - 0.33}{1 - 0.33} = 0.88$$

The interpretation of the Kappa statistics is not as simple as the accuracy, but given that this metric takes values between -1 and 1, a value of 1 corresponds to a perfect agreement between the predictions and the real labels; a value of 0 implies that there is no agreement; values less than 0 are interpreted as the agreement between the predicted and real labels are worse than the expected agreement that would be produced by chance. As a last note, this metric is to be used when the class distribution is unbalanced because in the opposite case, the kappa statistic shows high values.

Specificity

The specificity is a measure of the model's ability to correctly capture all negative cases; therefore, it is also called the *"true negative rate."* For a given negative class, the specificity can be calculated as:

$$Specificity = \frac{TN}{TN+FP} \tag{5-12}$$

Sensitivity

The sensitivity, also called recall, is a measure of the model's ability to correctly capture all positive cases; therefore, it is also called the *"true positive rate."* For a given positive class, the sensitivity can be calculated as:

$$Sensitivity = \frac{TP}{TP+FN} \tag{5-13}$$

The sensitivity and the specificity are inversely related, meaning that a model with high sensitivity would have a low false negative count; however, it would have a high false positive rate and vice versa. Usually, to discuss the tradeoff between these two metrics, they are plotted on a special plot called *"receiver operating characteristic (ROC)."*

The ROC curve (Figure 5-9) plots the sensitivity against the *"false positive rate"* (defined as 1-specificity) for the same model but with different prediction thresholds.

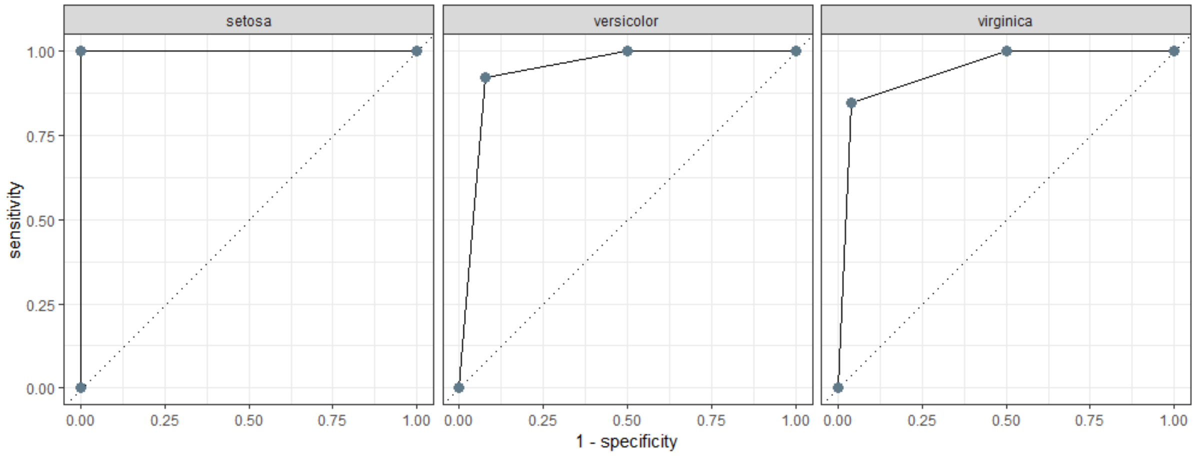


Figure 5-9 ROC curve of the decision tree model for the three classes. The pointed line (i.e., diagonal line) represent the roc curve for a random classifier, and every point on the plot represents a decision tree model with different threshold prediction.

An ideal model is located at the top left corner (i.e., a model that can correctly predict all the positive cases as well as all the negative cases); therefore, while tuning the model, we should aim to train the model as closely as possible to the left top corner. Moreover, the ROC curve is more commonly used to compare different MLA models for binary classification problems, but for multiclassification problems, especially when the number of the class is higher than 8, the ROC should be avoided, because it becomes hard to control and tune the model to capture a class at the expense of the others. (Landgrebe & Paclik, 2010)

Precision

For a given class, the precision shows the ratio of the correct predictions of all predictions made by the model. It is written as follows:

$$Precision = \frac{TP}{TP+FP} \tag{5-14}$$

Both precision and recall are commonly used to evaluate ML models; however, depending on the type of the learning problem, we usually want to improve one at the cost of the other. For example, when training a model that can predict whether a patient is sick or not, improving the recall of the model would be ideal, because the cost of the false negative (i.e., a sick patient predicted as healthy) is higher than the cost of the false positive (i.e., a healthy patient predicted as sick). Conversely, when we train a model to predict, and afterwards, show an advertisement according to the preference of the users. In this case, minimizing the false positive (i.e., an undesirable advertisement predicted as desirable) would be more advantageous than minimizing the false negative (i.e., a desirable advertisement predicted as undesirable).

Part II : Evaluating the contribution of the geophysical data and Landsat imagery in conjunction with machine learning for geological mapping. A case study: Silet, Central Hoggar.

5.1. Introduction

As discussed in previous chapters, most of the published works did not carry out a detailed methodology to study the contribution of the geophysical data to geological mapping. In this section, we propose multiple experiments to unveil their contributions to the geological mapping of the region of Silet. Wealso to assess the potential of the machine learning algorithm (MLA) *"Extreme Gradient Boosting (XGBoost)"* in predicting lithology. In other fields, this boosting-based algorithm proved to outperform the original implementation of the gradient boosting machines as well as other non-boosting algorithms. These experiments encompass the following methodology:

- By employing the XGBoost, we train multiple predictive models. These models are fed with different data types, including geophysical data, satellite data and digital elevation data. Also, they are trained to predict the lithology distribution in the region of Silet;
- The second experiment consists of comparing the XGBoost with another advanced MLA, namely Deep Neural Networks (DNN). The DNN proved to score excellent prediction scores in multiple fields, and based on our bibliographical research, this variation of the Artificial Neural Networks (ANN) has not been utilized for mapping geology;
- As assumed in the theory of the boosting ensemble methods, the performance of these MLAs should theoretically be better than bagging algorithms in capturing minority classes. Therefore, the last experiment aims to evaluate the prediction performance of the XGBoost, especially for low occurrence lithologies, against the random forest (RF), as it is considered, according to previous studies, the best MLA for the task of geological mapping.

5.2. Data and materials

We demonstrated in previous chapters all the processing schemes that we executed to improve the quality of the airborne geophysical data and to prepare it for the experiments shown in this section. In addition to the geophysical data, the images of the satellite Landsat 8 as well as the digital elevation model data, a product of the space Shuttle Radar Topography Mission (SRTM), were also included in the experiments. Concerning the Landsat 8 images, a *"level 2 collection 2"* version was downloaded, because these images are the target of processing and calibration methods that aims to improve their quality (Sayler, 2020; U.S. Geological Survey, 2016). Both datasets are freely available on (*https://earthexplorer.usgs.gov/*) with a spatial resolution of 30 m.

The airborne geophysical data includes both magnetic and spectrometric data. The first, after applying the reduced-to-pole transformation, were used to produce the apparent susceptibility and the apparent density grids, and the latter, after calculating the ground concentration of the radioelements, were used to calculate the radiation ratios Th/K, U/K and Th/U.

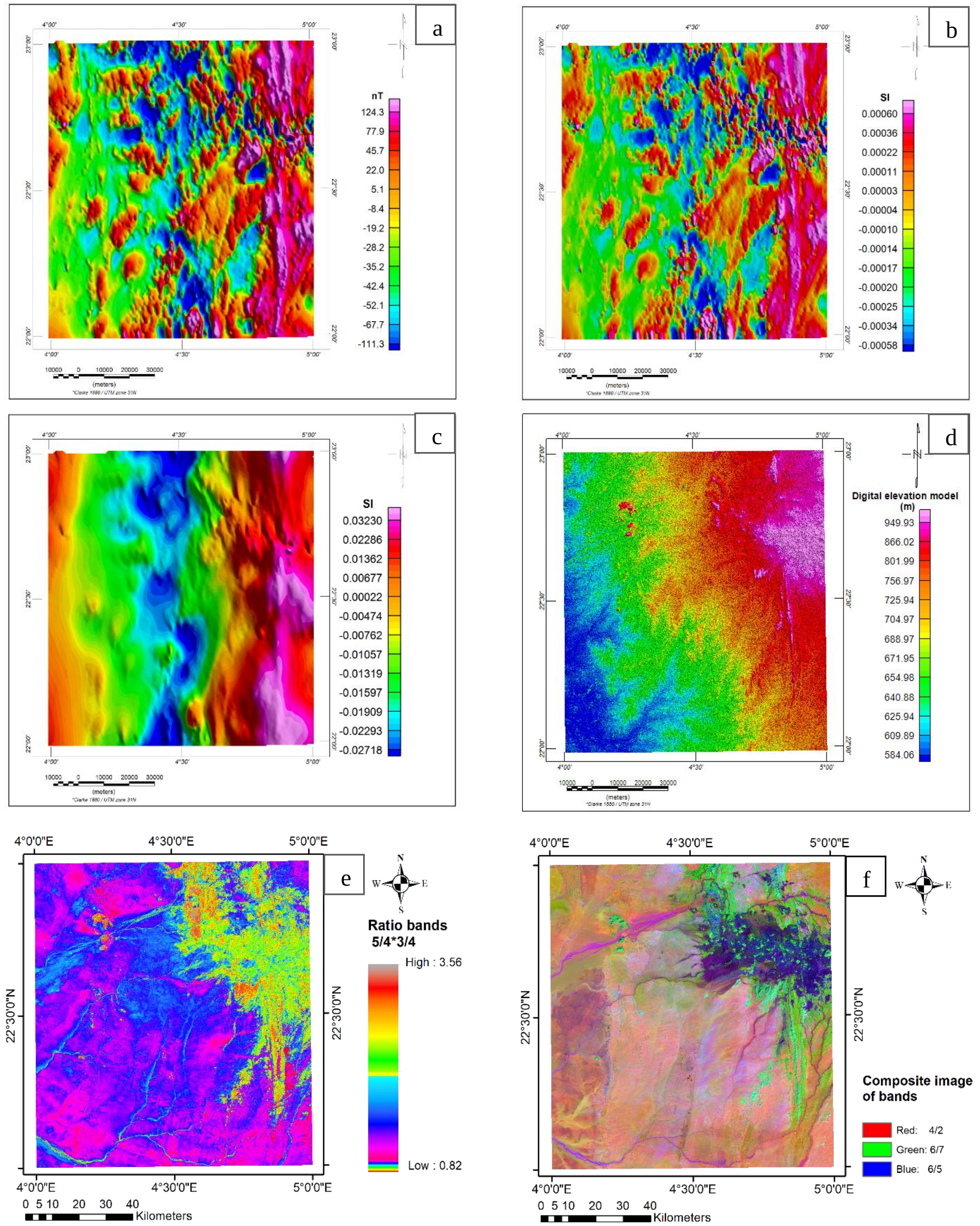


Figure 5-10 shaded grids of the MLAs input data. a) reduced to pole grid; b) apparent susceptibility grid; c) apparent density grid; d) digital elevation model (DEM); e) Landsat 8 band ratios 5/4*3/4; f) composite image of the three ratios 4/2, 6/7 and 6/5.

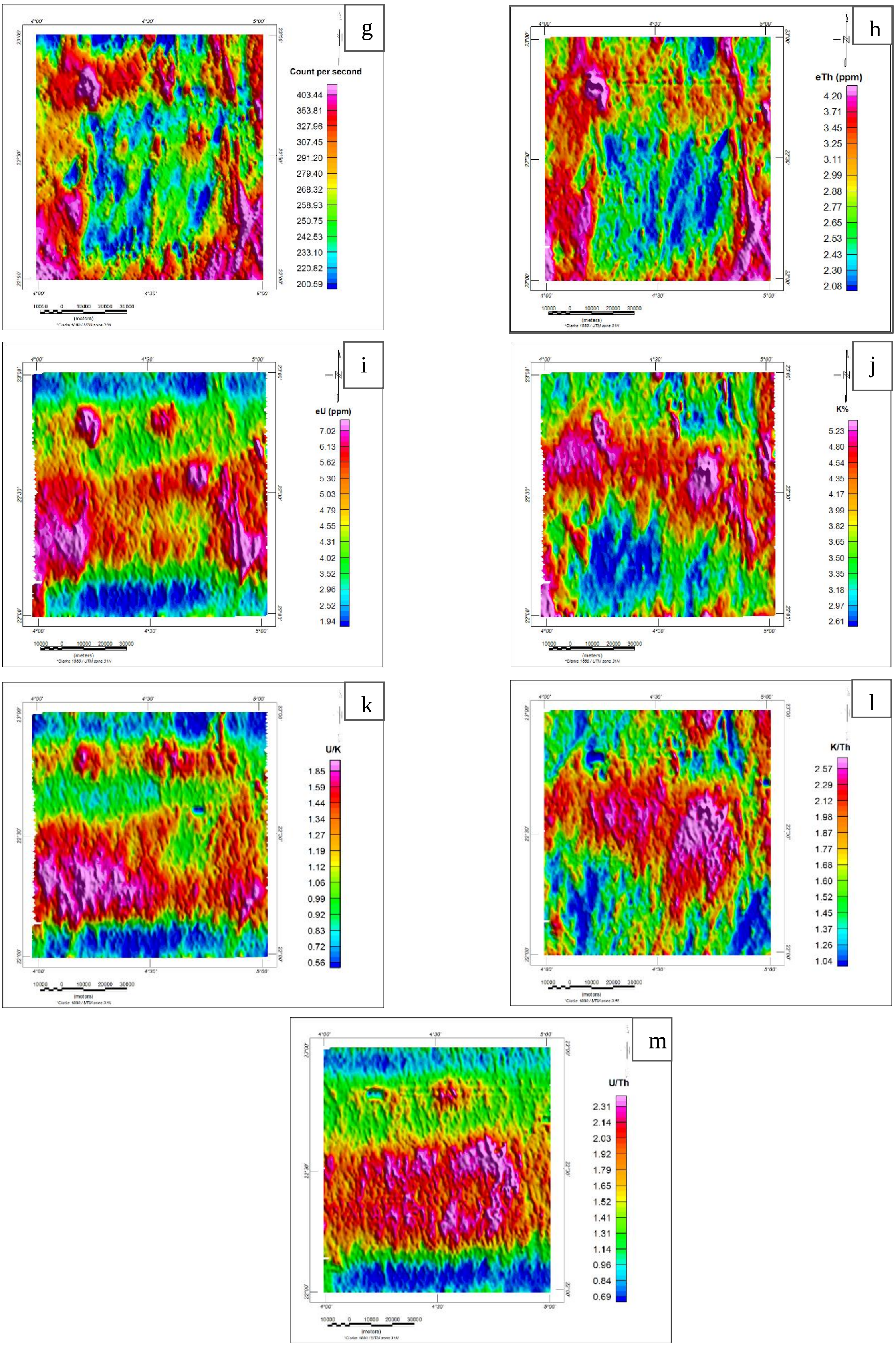


Figure 5-11 Shaded grids of the MLAs input data (continue). g) total count grid; h) equivalent in thorium grid; i) equivalent in uranium grid; j) equivalent in potassium grid; k) radiation ratio (U/K) grid; l) radiation ratio (K/Th) grid; m) radiation ratio (U/Th) grid.

The respective utilities of these ratios are: distinguishing the fresh and weathered mafic bedrocks (Dauth, 1997), identifying the fractionation zoning in felsic igneous rocks and mapping contacts between Precambrian rocks, sedimentary rocks and quaternary wadi sediments(M. A. S. Youssef & Elkhodary, 2013).

Table 5-4 Additional information about the input data concerning the spatial resolution and measurement units.

Data	*Spatial Resolution*	*Units*	*Data type*	
Reduced to pole map	200 m	nT	RTP	Geophysical data
Apparent density	200 m	SI	Apparent maps	
Apparent susceptibility	200 m	SI		
eU	200 m	Ppm	GRS data	
eTh	200 m	Ppm		
eK	200 m	Percentage		
Tc	200 m	Ppm		
Th/K	200 m	Ppm		
U/K	200 m	Ppm		
Th/U	200 m	Ppm		
Landsat 5/4*3/4°	~30 m	m	Landsat	Satellite data
Landsat 4/2-6/7-6/5*	~30 m	m		
SRTM	~30 m	m	DEM	
Geological map†	1:200,000	Character	Geology data	

Notes:

all data presented in the table are projected to a common projected coordinate system UTM zone 31°N and datum north Sahara 1959. Landsat 8 bands and SRTM products are resampled to 200 m cell size.

° multiplication of the two ratios (5/4) and (3/4)

* Composite image of the (4/2), (6/7), and (6/5)

† Rasterized to 200 m, the center of the cell is taken as the label for supervised learning.

Furthermore, the bands of the Landsat 8 were used to calculate the (5/4×3/4) ratio and the false-color composite image of the three ratios (4/2, 6/7, 6/5 in RGB), since they can be respectively used to distinguish between volcanic and metamorphic rocks from sedimentary rocks (Kusky & Ramadan, 2002), and to map iron oxides, clay minerals and ferrous minerals (Ourhzif et al., 2019). As a result of the applied transformations, the number of input data counted to 13 grids. They are presented in Figure 5-10 and Figure 5-11, and in Table 5-4, we include extra information concerning these data.

After obtaining the grids of the input data, they were compiled using an open-source environment for statistical computing and graphics (RStudio, 2019). All statistical analysis as well as machine learning modeling were carried out using this software. Moreover, since the Landsat 8 images and the digital elevation model have different spatial resolutions and also different projection methods from the geophysical data, they were the target of two transformations. The first consisted of re-projecting them to the same coordinate system as the geophysical data (i.e., UTM zone 31°N and datum NORTH SAHARA 1959), and the second transformation consisted of applying a resampling method to render their spatial resolution the same as the geophysical data. Lastly, the grids were organized in a tabular form where each column represents input data and each row represents an observation. The resulting database consisted of 284055 rows and 16 columns. 15 of the latter are input columns, and the last one is the lithology column (i.e., the outcome to-be-predicted).

Note: Throughout the experiments, and due to the big number of possible classes in the lithology column (i.e., 27 lithologies), the rock types are grouped into extrusive igneous, intrusive igneous, metamorphic, sedimentary and volcanic sedimentary sequences according to their respective rock class.

5.3. The XGBoost algorithm

The XGBoost is a boosting method that was introduced by (Chen & Guestrin, 2016) for sparse data and weighted quantile sketch for approximate tree learning. This algorithm is a new implementation of the Gradient Boosting Method (GBM) and compared to the classic GBM, the XGBoost is more advantageous for two main reasons. First, its loss function contains a regularization term that prevents it from overfitting, and the second is that its implementation is faster, more efficient and highly customizable. As demonstrated by (Chen & Guestrin, 2016), the XGBoost tries to minimize a loss function that is defined as:

$$L^t = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \Omega(f_t) \tag{5-15}$$

Where: l = convex loss function;

y_i= target value;

$\hat{y}_i$= predicted value. It can be calculated as follows: $\hat{y}_i = \hat{y}_i^{(t-1)} + f_t(x_i)$;

f_t = tree of structure q and leaf weight of w;

$\Omega\ (f_t) = \gamma T + \frac{1}{2}\lambda\|w\|^2$. It is a regularization term that facilitates the pruning process.

A second-order Tylor's approximation is used to simplify the L^t as follows:

$$L^t = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t) \tag{5-16}$$

Where g_i and h_i are respectively the first and the second-order partial derivatives of the loss function with respect to $\hat{y}_i$. Lastly, the optimal weight (output value) that minimizes the loss function can be calculated as:

$$w_i = -\frac{\sum_{i=1}^{j} g_i}{\lambda+\sum_{i=1}^{j} h_i} \tag{5-17}$$

The XGBoost calculates the observation log odds by summing the weights of its terminal leaves, and after that, the probability of belonging to a certain class *"i"* is estimated by calculating the logit transformation as follows:

$$p_i = \frac{e^{log(odds_i)}}{\left(1+e^{log(odds_i)}\right)} \tag{5-18}$$

5.4. First experiment

5.4.1. Methodology

This experiment has two main objectives. To evaluate the contribution of the geophysical data against/with the satellite data and also to evaluate the potential of the XGBoost-based models in geological mapping. The proposed methodology for doing that is to train nine XGBoost models, and to better distinguish the contribution of each data type, every one of these models is trained using different combinations of data types. We finish this experiment by employing the confusion matrix of the model that achieved the best overall performance to discuss and then propose some potential modifications to the original geological map. The models and also the data combination used to train them are presented in Table 5-5.

Table 5-5 Data combinations used to train every XGBoost model, and the labels that are used to refer to them in subsequent sections.

Data combination	*Label of the trained models*
GRS data	GRS model
RTP data	RTP model
GRS + RTP	GRS + RTP model
Landsat images	Landsat model
RTP + Apparent maps	RTP + Phy model
GRS+ RTP+ Apparent maps	Geophysical model
Landsat images + Apparent maps	Landsat + Phy model
GRS+ RTP + Landsat images	Geophysical + Landsat model
All data	All data model

5.4.2. Training and testing the XGBoost models

As in a typical supervised learning problem, the training phase starts by splitting the data into two parts, namely training and testing data. For our experiment, 20% of the data were used to train the XGBoost models, and the remaining 80% were kept for the evaluation phase. Also, the split was executed while preserving the distribution of the observations according to the lithology column (i.e., stratified sampling). The next step of the training phase is hyperparameter tuning. Knowing that the XGBoost has 7 tunable parameters, each one of the 9 models was tuned using a grid of random values with a dimension of (20×7). After finishing the training phase, multiple performance metrics were used to evaluate the performance of the trained models on the testing data. These are accuracy, kappa's statistics, sensitivity and precision.

5.4.3. Results

The first aspect of comparing the trained models is to use the accuracy and kappa's statistics and inspect which model, and consequently, which data type contributes the most to predicting geology. The measurements of the accuracy and kappa statistics on the testing data are presented in Figure 5-12, and in Table 5-6, we compare the computational time of the training phase to figure out which model achieved the best performance with the least training time.

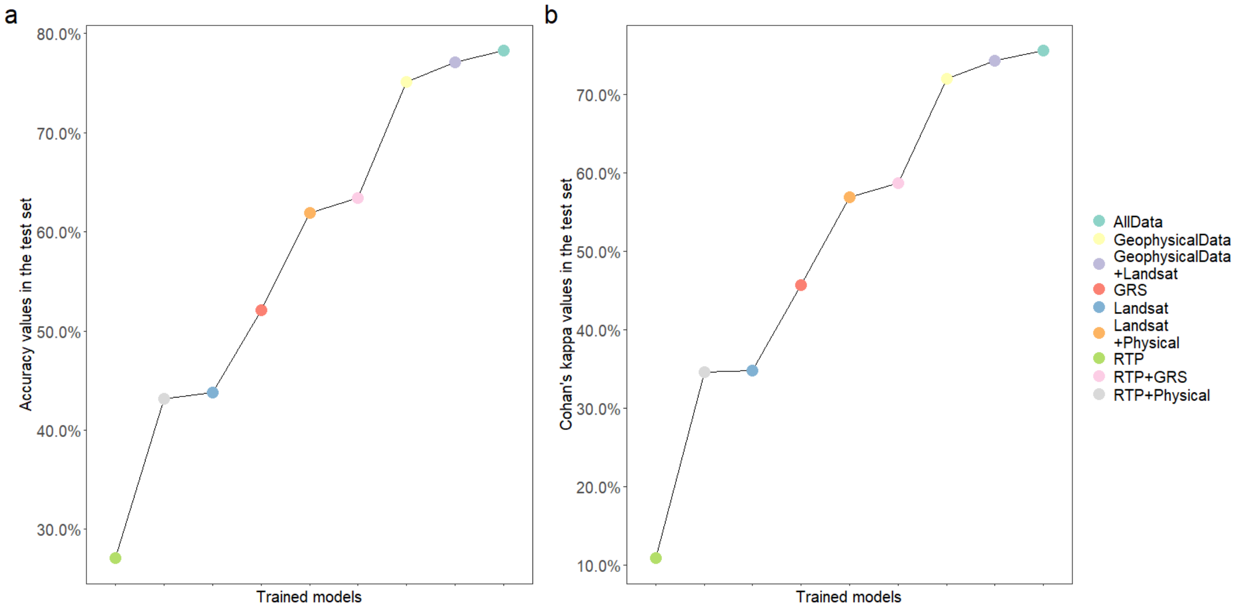


Figure 5-12 The accuracy and Cohen's kappa values of the 9 trained models in the testing set. a) Accuracy values. b) Cohen's kappa values. The points were enlarged for clarity purposes. Refer to Table 5-5 for more details about the trained models.

From Figure 5-12, we can see the advantageous effect of including more data to train the predictive models, as both the accuracy and the kappa's statistics increase with more data inclusion. Moreover, using a single data type (i.e., GRS, RTP or Landsat) demonstrates the limitation of each one to produce a good prediction score, because the trained models using one of these data scored a modest score that can barely exceed that mark of 50% for both metrics. However, by testing different combinations of these data, we can see a dramatic jump in both metrics. Taking the performance of the GRS model as a comparison reference, for example, we can see that by iteratively incorporating the RTP and then the apparent maps (i.e., apparent density and apparent susceptibility) with this model, the measurement for both metrics greatly increased, indicating that the RTP data and its derivatives have a complementary effect on the GRS data ability to capture lithology. Likewise, the model Landsat + Phy achieved an accuracy score above 60%, whereas the sole usage of the Landsat data alone produced a predictive model that did not even hit the 50% threshold for both metrics. Therefore, and similar to their effect on the GRS data, including RTP's derivatives could greatly benefit the ability of Landsat data to correctly capture lithology.

Table 5-6 Percentage increase in accuracy, Kappa's statistics and the Training Time.

	Compared to GRS model				*Compared to Landsat model*
	GRS + RTP	Geophysical data	Geophysical data + Landsat	All data	Landsat + Phy
Percentage increase of the accuracy %	21.15	44.23	48.08	50	40.91
Percentage increase of the training time %	6.11	15.74	37.07	37.25	10.36
Percentage increase of the kappa %	28.26	56.52	60.87	65.22	62.86

It is true that including more data generally means better prediction capabilities, but according to Table 5-6, different data types have more contribution to this increase than others. For instance, the table shows that by adding the apparent maps to the combinations (GRS + RTP) and also to the Landsat data, the prediction performance (accuracy and kappa's statistics wise) significantly increased. At the same time, including them did not induce a big training time increase, meaning that it was cost-effective to include them, and also that the apparent maps may offer a strong tool for the geological mapping. On the other hand, although increased both metrics, combining the Landsat data and the DEM data with geophysical data came at the expense of greatly increasing the training time, so in this case, we cannot say that including them was as beneficial as the apparent maps. In addition, we can also observe that the geophysical model, despite not achieving the best testing accuracy, still achieved the best percentage increase in accuracy while keeping the smallest increase in the training time.

The previous comparisons revealed the contribution of each data type to the overall capabilities of the predicted models on predicting lithology, but they did not reveal how each data type contributes to predicting each rock class. Therefore, the next aspect of comparing the trained models consists of using the performance metrics precision and sensitivity to unveil the utility of the input data in distinguishing between rock classes. Also, as mentioned before, and according to the geological map of Silet, there are 27 rock types. To facilitate the comparison procedures, they are grouped into extrusive igneous, intrusive igneous, metamorphic, sedimentary and volcano-sedimentary sequences according to their respective rock class. The mean precision and the mean sensitivity of the grouped classes are shown in Figure 5-13.

Figure 5-13 shows that the trained models tended to predict extrusive igneous rocks with higher precision values, meaning that when a particular observation is predicted as extrusive igneous rock, we can be very sure that the prediction is correct (i.e., prediction certainty). Moreover, higher sensitivity values were mainly related to metamorphic rocks, which indicates that if the trained models are confronted with metamorphic rock, there is a high probability that they would not misclassify it as being another rock class (i.e., capture rate).

The figure also demonstrates how the RTP data and their derived maps (i.e., apparent density and apparent susceptibility) seemed to increase the prediction certainty for the volcano-sedimentary and the metamorphic rocks, and to a lesser extent, the sedimentary rocks. On the other side of the coin, the RTP data and their derived maps seemed to greatly benefit the capturing rate for the metamorphic rocks, as the sensitivity values for this class were increased by a factor of 38 %. This was observed when comparing the precision and sensitivity values of the GRS model with the RTP+GRS model and the geophysical model relating to these rock classes. Furthermore, the derivatives of the RTP data substantially increased both the prediction certainty and the capturing rate of the Landsat model for the intrusive igneous rocks as well as the metamorphic rocks, since after including them, the precision values for these rock classes were respectively increased by 28 % and 20%, and the sensitivity values were increased by 30% for both classes.

What is worth noticing is that the precision values of the Landsat + Phy model for the extrusive igneous rocks is lower than the precision values of the Landsat model, which indicates that the apparent maps and Landsat might have a conflict towards this rock class, because including them seemed to increase the false positive count for this class (i.e., non-extrusive igneous rocks predicted as being one).

The effect of combining the Landsat data and the DEM with the geophysical data had a less important impact on the precision values as well as the sensitivity values, which was demonstrated by the small improvement of both metrics that range from 1 % to 5% for all rock classes. Therefore, we can say that, similarly to the improvement of the accuracy and kappa's statistics, incorporating them was also not cost-effective for improving the precision and the sensitivity.

The last aspect of comparing the predictions of the trained models is to visually compare them. This was executed by plotting the predictive maps produced by their predictions on the spatial domain. Each one of these maps was accompanied by an error map, a representation in two colours of the correct and incorrect predictions. These maps are presented in Figure 5-14.

Observing Figure 5-14 confirms that using each data type alone (i.e., RTP, GRS or Landsat) produced a mediocre predictive map at best. The predictive map produced by the RTP model (Figure 5-14.a) was the worst among the three, as clearly shown in the figure, since this model was unable to produce a comparable map to the original, reflecting the low accuracy score. Also, this model was unable to capture most of the rock units in Silet, as only three rock units, namely: the Arech-choum gneissic sequence in the western area (labelled as gi and presented in purple colour), Pelitic sandstone in the eastern area (labelled as P2GP and presented in grey colour) Tin Tikadiouit and Taket Granitoids complexes (labelled as GG1 and presented in orange colour) were present in the map. This indicates that the model predicted these rock units with a good sensitivity score, whereas for the uncaptured rock units, the sensitivity score was practically null (i.e., zero capturing rate). Compared to the RTP model, the predictive map of the Landsat model (Figure 5-14.b) did show some similarities with the original map, but it was a highly noisy map, and the limits of the major rock units were poorly delineated. In addition, some of the intrusive igneous rock units, especially dioritic units (labels start with the letter D), were nearly missed on the map, signifying that these rocks were misclassified as being another rock unit. The last predictive map produced by a single data type is the one predicted by the GRS model (Figure 5-14.c). This map was the most tolerable, the one that held the most similarities with the original map, the one with the best capturing rate and also the best delineation for the rock units compared to the other two, implying better accuracy and sensitivity scores. However, this model still had a high confusion between some rock units, especially: the calc-alkaline granite and granodiorite (labelled as G21 and presented in orange colour) with the granite and granodiorite of Tin Tikadiouit and Taket (labelled as G1and presented in heather purple colour), and the volcano-sedimentary sequence of Timeslarsine (labelled as P1VS and presented in yellow-green colour) with diorite associated with quartz diorite (labelled as D1 and presented in green colour). This confusion was slightly minimized in the predictive map of the GRS+RTP model (Figure 5-14.f), which further confirms that, despite producing the worst predictive map, the RTP data had a complementary effect with the GRS data. Their combination could, to a certain degree, produce an accurate predictive map.

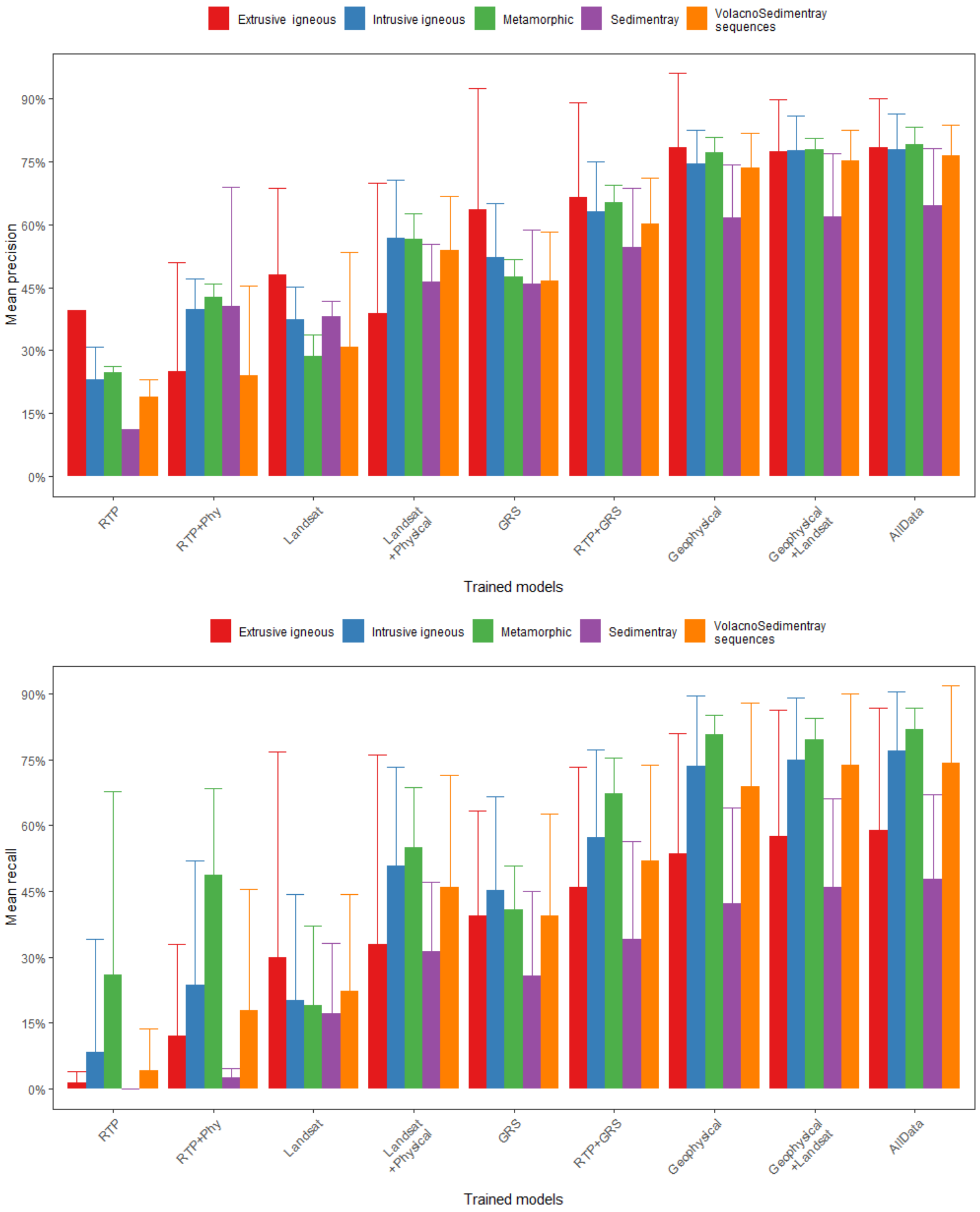


Figure 5-13 Mean precision values (top panel) and mean sensitivity values (bottom panel) of the 9 trained models related to the different rock classes. All rock types are grouped according to their rock class and the error bars represent the standard deviation of each rock class. Please refer to Table 5-5 for more details about trained models.

Note: The models trained with RTP, RTP+ Phy, or Landsat were unable to capture some of the lithologies present in the study area. Therefore, the mean and standard deviation values for these models do not represent the entire rock class.

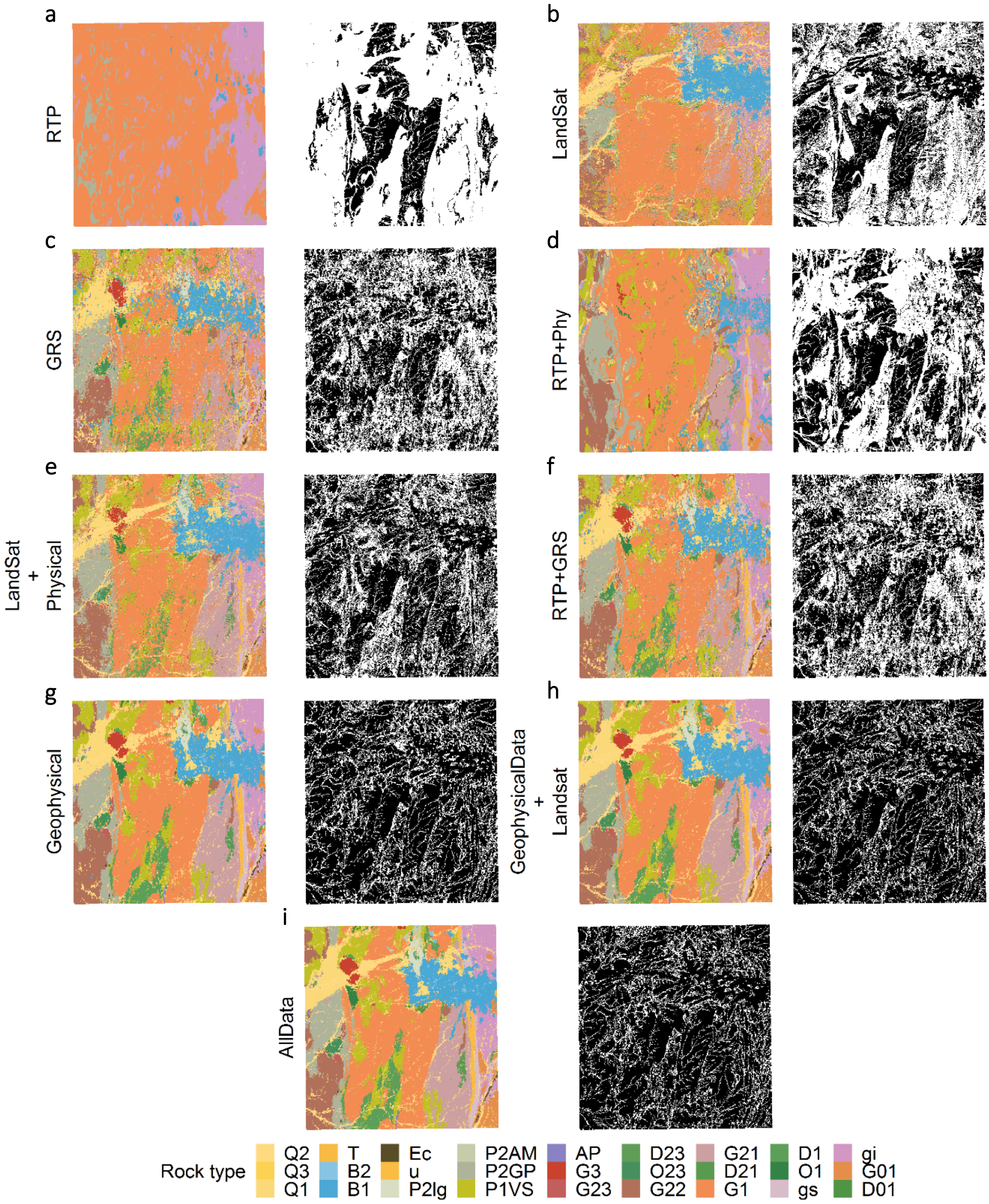


Figure 5-14 The final predicted maps of the 9 trained models plotted in the spatial domain The maps in black and white are a representation in two colors of the correctly and incorrectly predicted observations, where the black pixels represent the correctly predicted pixels. a) predicted map using the RTP data only; b) predicted map using the Landsat imagery only; c) predicted map using the GRS data only; d) predicted map using the combination of RTP data, apparent density and the apparent susceptibility; e) predicted map using the combination of the Landsat imagery, apparent density and the apparent susceptibility; f) predicted map using the combination of RTP and GRS data; g) predicted map using the combination of RTP, GRS data, the apparent density and the apparent susceptibility; h) predicted map using the combination of RTP, GRS data, the apparent density, the apparent susceptibility and the Landsat imagery; i) predicted map using all data.

The great impact of the RTP's derivatives on improving the predictions of the trained models was observed in (Figure 5-14-d), (Figure 5-14.e) and (Figure 5-14.g), the predictive maps for RTP+Phy, Landsat+Phy and Geophysical models. By the visual comparison of the error maps (i.e., black and white maps) for these models when compared to their peer models that did not include the RTP's derivatives, we can see that the white surface (i.e., incorrect predictions) was substantially decreased. Again, this confirms the strong potential for the RTP's derivative in aiding geological mapping.

By observing the predicted map of the geophysical data (Figure 5-14.g), we can see that this model successfully predicted the majority of the rock units and their spatial distribution, but the sedimentary rocks (presented as wadis in the lower left corner) were nearly overlooked by this model, which proves the limits of the geophysical data of providing a distinct signature that can be used to distinguish this rock class from the other. This limitation was roughly overcome by using the combination of the Landsat data as well as the geophysical data to train a model. Form the predicted map of this model (Figure 5-14-h), we can see that this combination provided a tool that can successfully delineate the limits of the major lithological units and also accentuate the sedimentary wadis in the area. Moreover, the predicted map of this combination, when compared to the predicted map produced by all data (Figure 5-14-i), showed that the DEM data had little effect on improving the predictive capabilities of the trained model, which was demonstrated by the none-observable difference between these two predictive maps.

According to Figure 5-12, the all-data model achieved the best performance; therefore, we conclude this experiment by discussing the confusion matrix of this model and then proposing some future modifications for the geological map of Silet. The confusion matrix is presented in Table 5-7.

The inspection of the table showed that the all-data model had an acceptable misclassification rate for all rock types; however, for the Trachyte and Phonolite (labelled as T), recent cone-shaped Basaltic associated with pyroclastic (labelled as B2), associated diorite facies (labelled as D23) and Æolian deposits (labelled as Q3), the model exhibited high misclassification rate (i.e., low sensitivity value). For the rock types T and B2, 34% and 49% of their respective observations were predicted as being old basalts sequence (labelled as B1). This high misclassification rate could be due to an erroneous prediction. However, as we can see in Figure 5-13, B1 predictions were characterized by a high certainty (i.e., precision values exceed 80%) Therefore, we can say that the predictions of the model were not necessarily wrong, and the geological limits of the T, B2 might have been wrongly mapped in the reference map. For the rock type D23, 23.4% of its observations were misclassified as being volcanogenic series of Ighellochem (labelled as P2lg), and 44.8% of the Q3 observations were misclassified as being the volcano-sedimentary series of Timeslarsine (labelled as P1VS). The model achieved a high precision score (around 80%) for both P2lg and P1VS. Therefore, the high misclassification rate of the D23 and the Q3 could also be due to errors in the reference map.

Table 5-7 Confusion matrix of the XGBoost trained with all data. The diagonal values represent the recall value related to each rock type, and the off-diagonal represent the ratio of the observations that were predicted as another rock type.

	Sedimentary				*Extrusive igneous*			*Metamorphic rocks*			*Volcano-sedimentary*				*Intrusive Igneous*												
	Q2	Q3	Q1	Ec	T	B2	B1	u	gs	gi	P2lg	P2AM	P2GP	P1VS	AP	G3	G23	D23	O23	G22	G21	D21	G1	D1	O1	G01	D01
Q2	35.9	12.5	10.6	17.9	1	4.1	1.5	3.6	6.3	7	3.1	7.1	5.5	5.9	15.8	2.9	12.1	9.2	11.1	4.3	4.6	4.8	4	6	9.6	9.3	25
Q3	0	34.5	0	0	0	0.1	0	0	0	0	0	0	0	0.3	0	0	0	0	0	0	0	0	0	0	0	0	0
Q1	4.8	0.4	75.7	0.2	0	0.7	1.9	0.2	0.1	0	0.4	0.1	0.8	1	15.8	4.1	0	0	0	0.2	0.2	0	0.2	0.7	2.7	0.1	4.4
Ec	0.1	0	0	45.5	0	0	0	0	0	0.2	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0.3	0
T	0	0	0	0	50.5	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B2	0	0	0.1	0	13.4	36.3	1.5	0	0	0	0.3	0.6	0	0	0	0	0	0.3	0	0	0	0	0.1	0.2	0	0	0
B1	2	0	5.4	0	34	49.4	90	1.2	4.3	2.9	3.9	1.3	0	0.1	0	0	0	3.5	0	0.1	0.7	0	0.1	0.2	0	3.1	5.3
u	0.4	0	0	0.2	0	0	0.1	81.6	5	0.3	0	0	0	0.1	0	0	0	0	0	0	0.6	0	0	0	0	0.2	0
gs	1.3	0	0.1	0	0	0.1	0.6	3.3	77	1	0.1	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0	2.4	0
gi	6.1	5.2	0	10.9	1	1.6	2	3.7	3.5	86.7	0	0	0	0	0	0	0	0	0	0	0.2	0	0	0.4	0	4.2	0
P2lg	0.3	0	0.1	0	0	0.7	0.4	0	0	0	87.1	3.8	0	0	0	0.3	0	23.4	0	0	0	0	0	0.1	2	0	0
P2AM	0.3	0	0	0	0	0.4	0	0	0	0	1.1	76.4	0	0.1	0	0	0	0.9	0	0	0	0	0.1	0	2.3	0	0
P2GP	4.8	0	0.9	10	0	0	0	0	0	0	0	0	86	0.3	0.2	0	8	0	3.6	1.9	0.1	0	0.5	0.6	0.1	0	0
P1VS	6.8	44.8	1.9	2.8	0	0.2	0	0.4	0	0	0.7	6.4	0.8	77.3	3.7	1.1	9.3	4.7	0.4	1.7	1.7	0	2.1	4.8	3.9	0	0.2
AP	0.1	0	0.2	0	0	0	0	0.1	0	0	0	0	0	0.1	44.3	0	0	0	0	0	0	0	0	0.2	0	0	0
G3	0.2	0	1	0	0	0	0	0	0	0	0	0	0	0.1	0	91	0	0	0	0	0	0	0	0.2	0	0	0
G23	0.2	0	0	0	0	0	0	0	0	0	0	0	0.2	0	0	0	59.9	0	1	0.1	0	0	0	0	0	0	0
D23	0.1	0	0	0	0	0.7	0	0	0	0	1.7	0	0	0	0	0	0	53.2	0	0	0	0	0	0	0.5	0	0
O23	0.4	0	0	0	0	0	0	0	0	0	0	0	0.2	0	0	0	3.9	0	82.2	0.1	0	0	0	0	0	0	0
G22	4.1	0	0.3	2	0	0	0.1	0	0.1	0	0	0	2.8	0.5	1.4	0	3.7	0	1.7	90.4	0.4	0	0	0.4	0	0	0
G21	8.1	0	0.3	0.2	0	0	0.7	3.9	0.6	0.2	0	0	0.2	2.3	0.2	0	0	0	0	0.9	90.1	16.8	0.1	0	0	0.5	0
D21	0.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.7	78.4	0	0	0	0	0
G1	17.9	2.6	2.5	1.4	0	4.1	0.5	0.8	1.4	0	0.5	3.6	2.8	9.3	11.3	0.2	3.2	3.5	0	0.1	0.5	0	91.8	9.2	5.7	0.5	5.5
D1	2.9	0	0.4	1.1	0	1.6	0	0.2	0	0.1	0.4	0.1	0.5	2.3	7.4	0.3	0	0.9	0	0.1	0	0	0.8	76.9	0.9	0.5	0
O1	0.4	0	0.2	0	0	0	0	0	0	0	0.5	0.5	0	0.2	0	0	0	0.3	0	0	0	0	0.1	0	72.4	0	0
G01	1.8	0	0	8.1	0	0.1	0.5	0.9	1.8	1.6	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0.1	0	78.4	0.7
D01	0.5	0	0.1	0	0	0	0	0	0	0	0.2	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0.4	58.9

5.5. Second experiment

In this experiment, we are concerned with comparing the performance of the XGBoost to the deep neural networks (DNN) in predicting geology. To do that, the same methodology (i.e., comparing the precision and the recall values as well as plotting the predictions in the spatial domain) that was used in the first experiment is also used here. In addition to that, the *"entropy"* metric is used to investigate the uncertainty of the predictions relating to both models.

5.5.1. Entropy

The entropy, or the information theory, computes the uncertainty related to the possible outcomes of certain variables. It was introduced by (Shannon, 1948), and it is calculated as follows:

$$\mathrm{H} = -k \sum_{i=1}^{n} p_i \, log \, p_i \tag{5-19}$$

Where p_i is the probability related to every possible outcome; n is the number of outcomes; k is an arbitrary number. Both k and the logarithm base are user-chosen parameters.

Besides predicting a class label, the XGBoost and the DNN models can also produce a vector of n elements corresponding to the probability of being one of the possible classes in the output column. The first is called *"hard prediction,"* and the latter is called *"soft prediction."* Adjusting both algorithms to produce the probability vector, these predictions can be injected in equation (5-19) to calculate the H value. In such case, n is taken as the number of rock types (i.e., 27), and p_i is the probability of a particular observation being a rock type i. The value H reaches a maximum value when all outcomes have an equal probability of happening and a minimum value if there is only one possible outcome. Moreover, to facilitate the interpretation of the entropy values, it was normalized to have values between 0 and 1. This was achieved by dividing the entropy values by $log_2(n)$. This normalization method was inspired by (S. Kuhn et al., 2018).

5.5.2. Training the MLA models

In this experiment, and unlike the first one, all input data were used to train the XGBoost and the DNN models. For the XGBoost model, we already showed in the first experiment how the hyperparameters were tuned, and for the DNN model, multiple architectures were tested using a random grid of values of dimension (5×20) to ensure better classification accuracy. All architectures were tested by setting the RELU function as an activation function and the ADAM optimizer, which was used to update the weights, as the optimization algorithm.

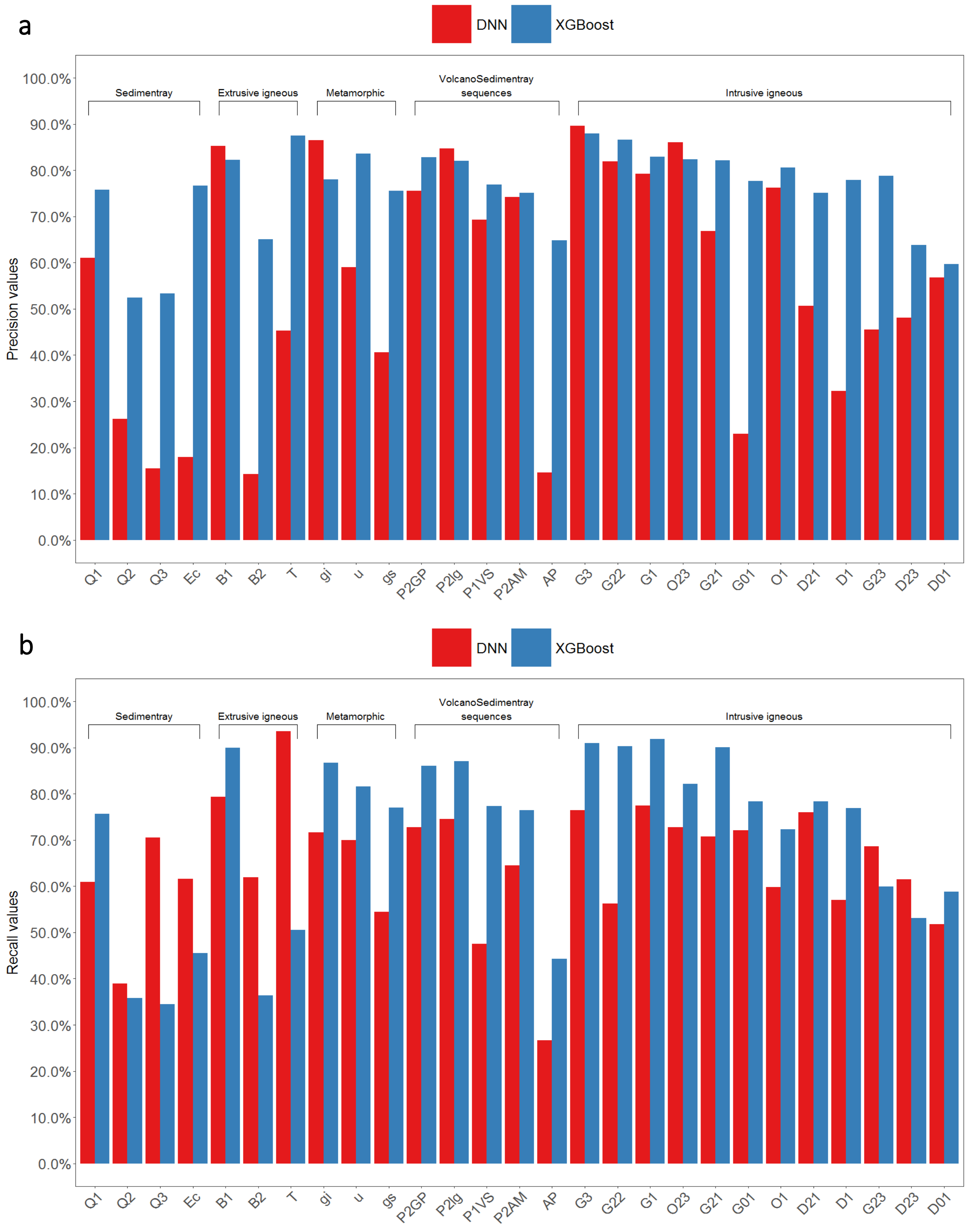


Figure 5-15 The precision and recall values relative to each rock for the XGBoost and DNN algorithms. a) precision values; b) recall values.

5.5.3. Results

Except for the old basalts sequence (labelled as B1), Arechchoum gneissic series (labelled as gi), Ighellochem volcanogenic series (labelled as P2lg), Taourirt granites (labelled as G3), gabbro (labelled as O23) where the DNN model slightly outperformed the XGBoost model, the Figure 5-15.a clearly shows that the XGBoost achieved better precision values. The difference between the two models, however, differs from one rock class to another. For example, for the volcano-sedimentary sequences, both models achieved comparable precision values, but for the other rock class, the difference of precision values between the models differed from one rock type to another with the XGBoost constantly achieving better precision scores.

Figure 5-15.b shows a comparison of the recall values between the two models. Like the precision values, the XGBoost was able to predict the majority of the rock types with better recall scores. Yet, the difference in the recall values between the two models was not as important as the precision values. This figure also demonstrates that the DNN model seemed to predict the sedimentary and extrusive igneous rocks with higher recall scores.

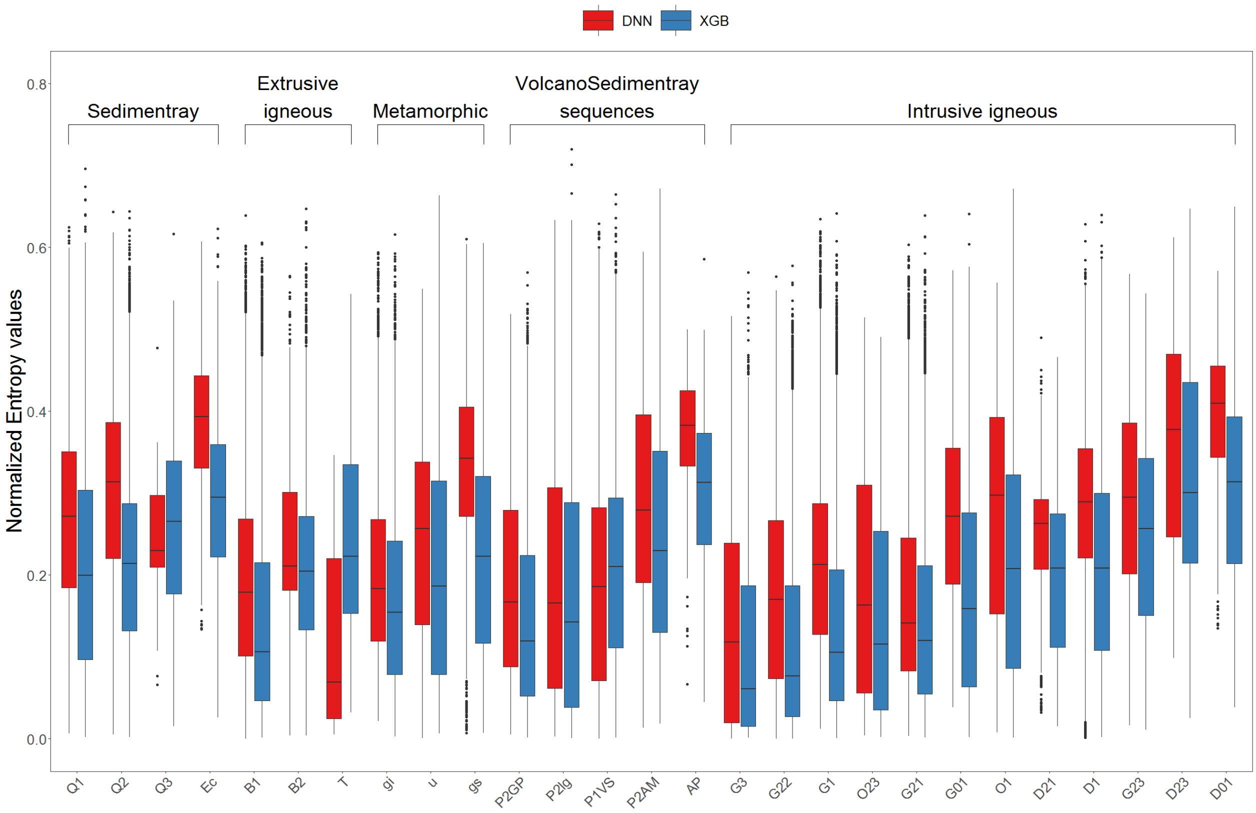


Figure 5-16 The normalized entropy values produced from the predictions of the XGBoost and the DNN models.

Figure 5-16 represents the distribution of the entropy values for the XGBoost and the DNN models. Overall, the median values differed from one rock type to another. However, both algorithms seemed to achieve lower entropy values (i.e., high certainty) for the intrusive igneous rocks and higher entropy values (i.e., low certainty) mainly for the sedimentary rocks. Moreover, the entropy values for the other rock classes appeared to vary roughly in the same range. The figure also demonstrates that the boxes, which are limited by the third quartile (i.e., 75% of the observations), corresponding to the XGBoost prediction were below those corresponding to the DNN predictions for all the rock's types except for the trachyte and phonolite (labelled as T). This indicates that the majority of the predictions of the XGBoost were characterized by higher certainty compared to the predictions of the DNN algorithm.

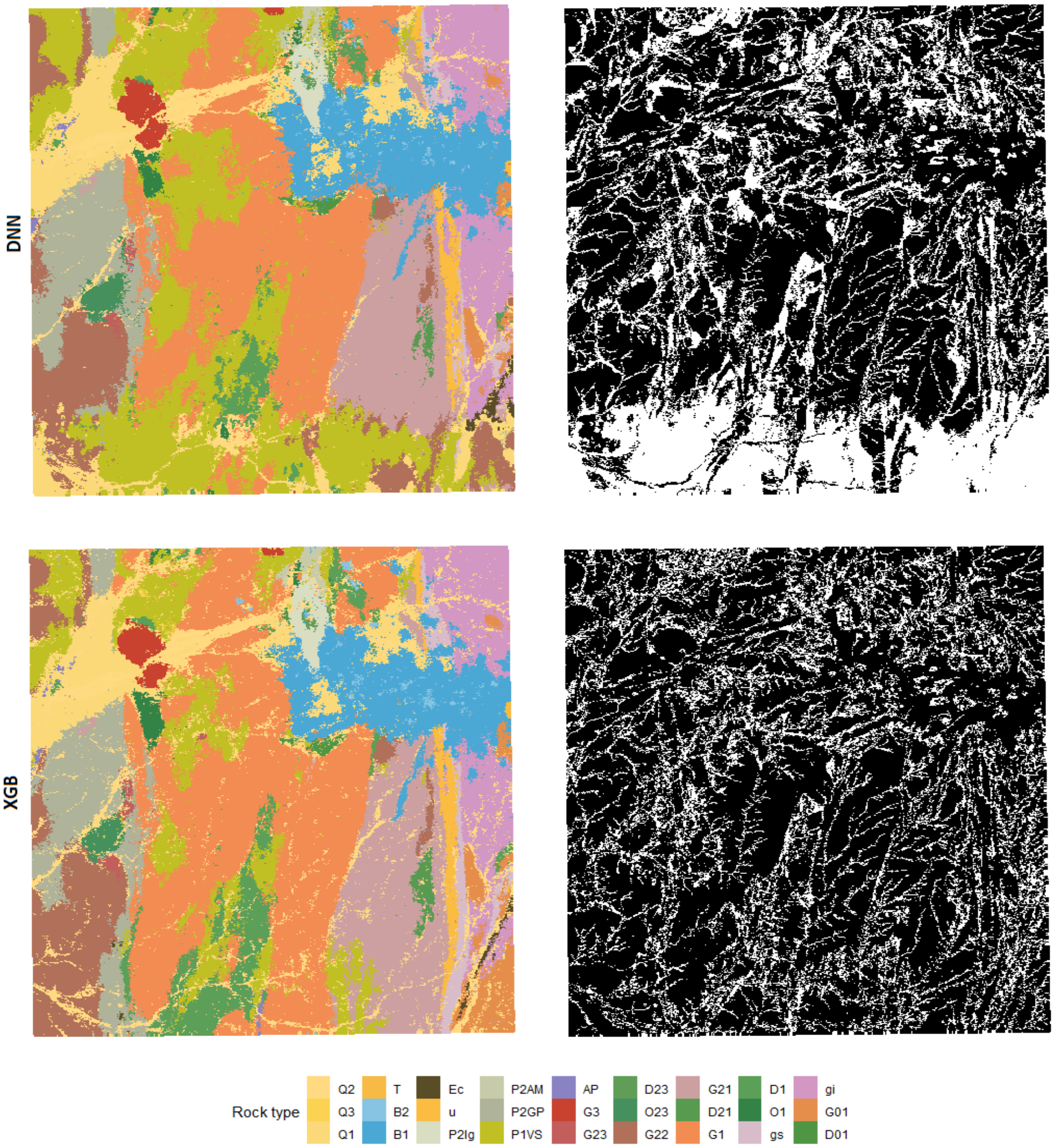


Figure 5-17 Predicted maps of the XGBoost and the DNN models plotted in the spatial domain. The maps to the right are a representation in two colors of the correctly and incorrectly predicted observations, where the black pixels represent the correctly predicted pixels.

Figure 5-17 shows the predicted maps of the XGBoost and the DNN algorithms as well as the error maps that indicate the failed predictions. The figure reveals that the predicted map of the XGBoost, which was generated from the predicted observations of the model, bore a better resemblance to the reference map than the predicted map of the DNN. The figure also reveals that failed predictions of both models were mainly due to the observations of the sedimentary rocks, especially the alluvium (labelled as Q2). In addition to that, the failed predictions of the DNN model were also induced by the misclassification of the observations of the calco-alkaline granitoids complex of the Isseimane river (labelled as G21), the granitoids complexes of Tin Tikadiouit and Taket (labelled as G1), diorite and quartz-diorite (labelled as D1) with the volcano-sedimentary series of Timeslarsine (labelled as P1VS). This was indicated by the high error rate (white surface in the error map) in the lower area of the study area.

5.6. Third experiment

The last experiment in this section aims to compare the XGBoost with the random forests (RF) to check whether or not the performance of the XGBoost, a boosting-based method, in capturing the low occurrence lithologies is better than the RF, a bagging-based method. Also, according to the previous studies, the latter is proved to be the best first choice for prediction lithology, so if the XGBoost did outperform the RF, the comparison discussed in this experiment could update the literature of applying machine learning in the context of geological mapping.

5.6.1. Methodology

Both the XGBoost and RF were trained using all input data presented in Table 5-4. To evaluate the performance of the trained models in capturing low occurrence lithologies, the receiver operating characteristic (ROC) curve was utilized to compare the capturing true positive rate (TPR) (i.e., sensitivity) and the false positive rate (FPR) (i.e., 1-specificity) for both models. Moreover, we summarized the ROC curves into a single numeric metric called *"area under curve (AUC)"*. The performance of the models was only compared for low occurrence lithologies which have an observation number of less the 2%. The number of observations relating to every rock type is shown in Table 5-8.

Table 5-8 Observation number of the rock types in Silet region and their proportions.

Rock types	*Observation number*	*Percentage of occurrence (%)*
G1	52179	22.96
Q2	31674	13.94
P1VS	21528	9.47
G21	19318	8.5
B1	16652	7.33
G22	14531	6.39
P2GP	13996	6.16
gi	11994	5.28
Q1	10965	4.83
D1	9797	4.31
G01	4640	2.04
gs	3697	1.63
u	3312	1.46
G3	1880	0.83
P2lg	1854	0.82
B2	1802	0.79
O1	1368	0.6
O23	1147	0.5
D21	973	0.43
P2AM	815	0.36
G23	778	0.34
Ec	653	0.29
D01	547	0.24
AP	488	0.21
D23	316	0.14
Q3	232	0.1
T	97	0.04

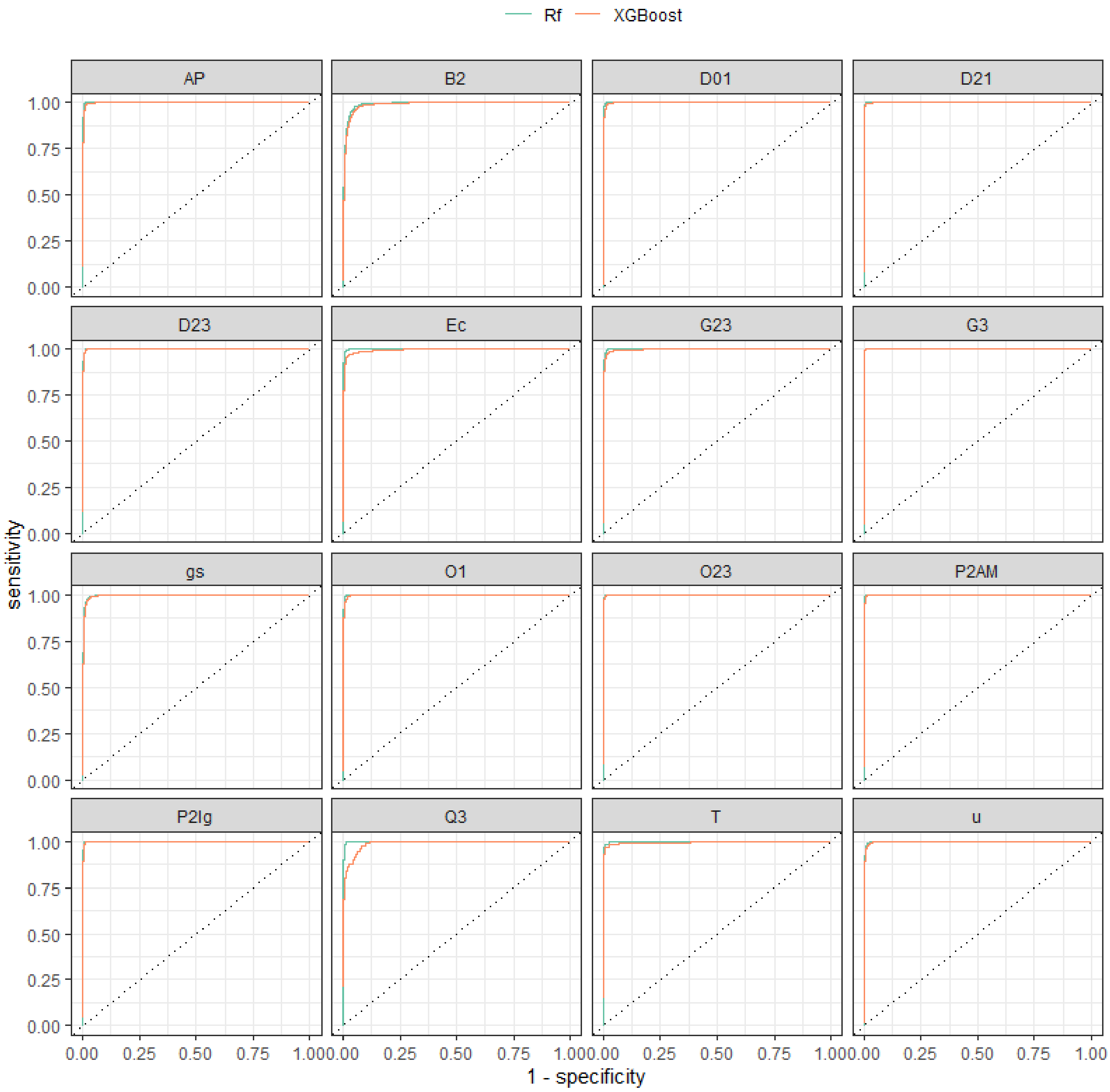


Figure 5-18 ROC curve of the XGBoost and the RF models relating to low occurrence lithologies (i.e., occurrence percentage < 2%)

5.6.2. Results

In Figure 5-18, we show the ROC curves of the trained model for low-occurrence lithologies. The figure shows that both models achieved an excellent classification score, as the curves for both models looked like the ROC curve of an ideal model (i.e., close to the top left corner). In addition, except for the rock type “Q3” and “Ec”, the figure also shows that the performance of the trained models was practically identical, indicating that the TPR and the FPR values were nearly the same. The identical performance was proved by AUC values, where the XGBoost and the RF achieved 0.986 and 0.99 respectively (an ideal model would achieve an AUC of 1).

5.7. Discussions and conclusions

This study described an application framework of the XGBoost algorithm for geological mapping. It achieved a good correlation of up to 78% with the existing geological map. This good correlation score is because our study area is dominated by igneous acidic rocks (i.e., granite). The igneous acidic rocks are often composed of the feldspar's minerals (K-spar family), micas minerals, and ferromagnesian silicate minerals (i.e., amphiboles and biotite). Thus, this rock class exhibits a strong magnetic and radioactivity signature, leading to making them be predicted with high certainty (i.e., low entropy values).

The first experiment focused on revealing the contributions and the limitations of the different data types to predict each rock class. Since Landsat is based on measuring the reflected fraction of the solar radiation, it has a low penetration depth which makes the Landsat images more suited for predicting the surficial lithologies, hence especially helpful for the sedimentary rocks among other surficial lithologies. On the other hand, the great penetration depth of the geophysical data (GRS and RTP) and the ability to derive the apparent maps, which can supply additional information about the physical properties of the rocks, make the geophysical data a valuable tool for predicting the lithologies located below the ground surface. Another advantage of geophysical data is its ability to generate good accuracy scores and noiseless predictions while keeping a short training time. Therefore, they maintain a good balance between achieving a good accuracy score and maintaining low computational expenses compared to other data combinations.

The contribution of the geophysical data to predicting lithology was demonstrated, when some of the granitic units in the region, namely the calc-alkaline granite and granodiorite of the Isseimane wadi (labelled as G21) and the granite and granodiorite of Tin Tikadiouit and Taket (labelled as G1) were set apart from other granitic units using the RTP. This might be due to the presence of the amphibole minerals, which are not present in the other types. The iron-rich minerals of this group usually display strong magnetic properties than the minerals with less iron (Rosenblum & Brownfield, 2000). Moreover, the RTP's derivatives (i.e., apparent density and apparent susceptibility) showed great potential in improving the accuracy, sensitivity and precision of the trained models in distinguishing between different rock classes. This was clearly shown in the observed visual improvement of the predicted maps when using the apparent maps and without using them in Figure 5-14. Although a positive effect is always observed when using the apparent maps, the proportion of this improvement differed from one rock class to another. For example, the igneous intrusive, metamorphic, and volcano-sedimentary rocks seemed to benefit more from including the apparent maps, which is to be expected, since igneous and metamorphic rocks tend to be more magnetic and denser than the other rocks classes; thus, they are easily differentiable. Conversely, the sedimentary rocks and the extrusive igneous rocks were the least beneficiaries from including the apparent maps. For the sedimentary rocks, this is caused by the heterogenous nature of the rock's parent material from which these sedimentary rocks were formed from, which causes a non-unified geophysical signature, and for the extrusive igneous rocks, this is caused by the similar mineralogical composition that they are sharing as well as the overlapped spatial distribution, leading to a high confusion rate among these rocks.

The first experiment also demonstrated the limitation of using a single data type to predict lithologies. Despite the high resolution that the Landsat 8 data usually come with, which allows fulfilling the task of mapping surficial lithologies in well-outcropped regions with good accuracy, the predictions produced using the Landsat data were noisy compared to geophysical data due to the limited outcropped rocks and the sedimentary cover in our study area. Likewise, the GRS data, despite generating a good predictive map compared to the other models that were trained with a single data type, a considerable amount of prediction noise was present in the predicted map. This may be due to its quality. Our data were recorded as a part of a regional scale survey, and thus it was covered by a wide flight line separation. As a result, the GRS data could be of medium quality. Another reason for the noisy predictions of the GRS data is the persistence of a low noise level after processing the numeric database of the GRS data. The Uranium channel (eU) and the Thorium channel (eTh) are the most affected by this noise.

The second experiment focused on comparing the XGBoost algorithm and the DNN algorithm. It showed that the XGBoost predictions, relating to each rock type, were characterized by a higher certainty which was indicated by the high precision values as well as the low entropy values. For the recall values, the DNN algorithm was able to achieve comparable values as the XGBoost and, for some rock types, slightly outperformed the XGBoost. Therefore, this study finds that the XGBoost is a better choice for geological mapping than the DNN algorithm.

In the last experiment, we conducted a comparison between the random forest and the extreme gradient boosting. It demonstrated that XGBoost and the Rf were able to achieve a comparable prediction performance since the overall accuracy and the area under curve for both models were practically identical (accuracy scores of 78% and 81%, and an AUC of 0.98 and 0.99 were respectively achieved for the XGBoost and the Rf). These results were as pertinent as those obtained by other published studies using different techniques of machine learning (Cracknell & Reading, 2014). In a conclusion, albeit with a small margin, our study further confirms that the random forest is the best first choice to train a lithology classifier.

As a last note, two key remarks should be taken into consideration:

- The gridding cell size is an important feature when using raster data. Bigger cell sizes tend to generate coarse grids that assign the same value over a wide surface. Therefore, in the case of the apparent density or susceptibility, a bigger cell size means a wide homogeneous surface, which might not be the case in the field, while smaller cell sizes tend to generate a more detailed grid. This requires more computational expenses when facing a spacious study area. Depending on the nature of the study, whether it is a regional or a narrow more detailed study, the tradeoff of the cell size should be considered with care.

- The two apparent maps were generated using a simple mathematical model that is based on the assumption that the responses are collections of vertical, square-ended prisms of infinite depth extent. Additionally, the magnetic response is due only to the local magnetization caused by the earth’s magnetic field (Yawsangratt, 2002). These assumptions are not always met; therefore, the obtained maps can only be considered as an approximation that does not fully reflect the real density values and susceptibility observed on the field.

Conclusions

Conclusions

In the dissertation, we executed multiple experiments to reveal the contribution of the geophysical data as well as the Landsat imagery to the geological mapping of Silet region. Our main findings can be summarized into the following:

- Due to the immense availability of different remote sensing data, and also the ready-to-use geophysical airborne data, employing machine learning can be regarded as an optimal tool to exploit these data and extract patterns that cannot be observed using traditional methods;
- Of all data types used to generate predictive maps in this study, the geophysical data proved to be superior in both overall accuracy and prediction certainty. Also, they can provide a strong means to produce very acceptable geological maps, depending on their quality;
- The methodology of combining different data types to predict lithology can almost always increase the performance of the machine learning-based model. However, the effect of including the geophysical data showed to be a balanced choice between increasing the performance and maintaining low computational expenses;
- From a machine learning algorithm prospective, this study described an application framework of the Xgboost algorithm for geological mapping. This algorithm achieved a good correlation of up to 78% with the existing geological map;
- The comparison of the Xgboost and the DNN algorithms showed that both algorithms can achieve comparable precision values, with the DNN having the slight advantage over the Xgboost; for the sensitivity values, the DNN seems to have the advantage of capturing sedimentary and extrusive igneous rocks more accurately, whereas for other classes, the Xgboost clearly have the edge.

Our results indeed came following previously published works; however, our proposed workflow in the dissertation emphasized the potential of different geophysical data types (i.e., airborne magnetic and gamma-ray spectrometry data) in predicting lithology against other data types. Previous works did not include any inter-comparison between geophysical data, and they have focused solely on employing geophysical data as a whole without including a detailed workflow as the one described in this dissertation. Therefore, our methodology should provide additional contributions to the already rich literature. In addition, besides the main question discussed throughout the dissertation, our work compares between bagging and boosting algorithms for mapping geology. Based on our current bibliographical review, the application of a boosting algorithm provides greater novelty to our study.

Employing remote sensing data (including geophysical data) in conjunction with machine learning algorithms has become the most practiced methodology in recent years. However, this methodology is not intended to replace traditional methods, but rather to complement them by producing objective and rapid predictive maps that can help to orient the fieldwork, especially in hard-to-explore terrains or regions with limited outcrops.

Perspectives and future works

As is the case with any empirical study, some limitations should be addressed in follow-up studies to further complement our findings. These are:

- The airborne data that were used in our study is issued from an old survey, so according to modern standards, their quality is of average quality. This can harm the prediction accuracy of the machine learning models and decrease their reliability. Therefore, employing newly acquired data which are issued from more sophisticated surveys should be the main goal for future works like the present one;
- According to the machine learning literature, our lithology models were inferential. Given the patterns in the training data, the purpose of such models is to affirm or denounce certain questions or hypotheses regarding the input-output relationship; however, because these models are tweaked in a way that only makes them priorities the extraction of the inherent relationships in the present data, their performance tends to degrade when facing new data (i.e., low generalizability). In our case, our models were able to perform well in classifying lithologies in Silet, but when applied in adjacent regions, the performance was considerably degraded. Future works should aim to train lithology models that are tweaked to have better generalizability. These are referred to as *"Predictive models,"* and the main objective of these models is to generate the most accurate predictions for both the present and future data;
- Nowadays, with the help of geographic system (GIS), integrating various remote sensing data types becomes easier than ever. Therefore, geological mapping practitioners should aim to experiment with other data types, such as geochemical data, gravity data, derivative of digital elevation models and other magnetic data derived maps to figure out which data combination that holds the most relevance to geology.

Finally, there is a need for a unified database of the major rock units encountered in Algeria which can link these units and their signature, if there are any, in the remote sensing data is necessary and should be the main focus before applying any modelling procedure. The reason for that is to assure that lithology models, especially machine learning-based ones, would produce reliable and objective predictions.

References

Allek, K. (2005). *Traitement et interprétation des données aéromagnétiques acquises au-dessus des régions de Tindouf et de l'eglab (sud-ouest de l'Algérie) – impact sur l'exploration du diamant* [Memoire de Magister]. Université des sciences et de la technologie Houari Boumediene (USTHB).

Ansari, A., & Alamdar, K. (2009). Reduction to the pole of magnetic anomalies using analytic signal. *World Applied Sciences Journal*, *7*(4), 405–409. http://www.idosi.org/wasj/wasj7(4)/2.pdf

Asfirane, F., & Galdeano, A. (1995). The aeromagnetic map of northern Algeria: Processing and interpretation. *Earth and Planetary Science Letters*, *136*(1–2), 61–78. https://doi.org/10.1016/0012-821X(95)00043-4

ASGA. (2019). *Catalogue des publications* (2019th ed., pp. 1-116 p). Agence des services géologiques de l'Algérie. https://doi.org/10.1016/0022-1694(69)90041-9

Bachri, I., Hakdaoui, M., Raji, M., Teodoro, A. C., & Benbouziane, A. (2019). Machine learning algorithms for automatic lithological mapping using remote sensing data: A case study from Souk Arbaa Sahel, Sidi Ifni Inlier, Western Anti-Atlas, Morocco. *ISPRS International Journal of Geo-Information*, *8*(6), 1–20. https://doi.org/10.3390/ijgi8060248

Baranov, V. (1957). a New Method for Interpretation of Aeromagnetic Maps: Pseudo-Gravimetric Anomalies. *Geophysics*, *22*(2), 359–382. https://doi.org/10.1190/1.1438369

Barraclough, D. R. (1987). International geomagnetic reference field: the fourth generation. *Physics of the Earth and Planetary Interiors*, *48*(3–4), 279–292. https://doi.org/10.1016/0031-9201(87)90150-6

Bechiri-benmerzoug, F. (2009). Pétrologie, géochimie isotopique et géochronologie des granitoïdes Pan-africains de type TTG de Silet : contribution à la connaissance de la structuration du bloc d'Iskel (Silet, Hoggar occidental) Algérie. *Université Des Sciences et de La Technologie HOUARI BOUMEDIENNE (USTHB), Mémoire de Doctorat*, 1-387 p.

Bertrand, J. M. (1974). Evolution polycyclique des gneiss précambriens de l'Aleksod (Hoggar central, Sahara algérien): aspects structuraux, pétrologiques, géochimiques et géochronologiques. *Université Montpellier II - Sciences et Techniques Du Languedoc*, *Mémoire de Doctorat*, 1-393 p.

Bertrand, J. M. L., Boissonnas, J., Caby, R., Gravelle, M., & Lelubre, M. (1966). Existence d’une discordance dans l’antécambrien du “fossé” pharusien de l’Ahaggar occidental (Sahara central). *C. R. Acad. Sc. Paris*, 2197–2200.

Bertrand, J. M. L., & Caby, R. (1978). Geodynamic evolution of the Pan-African orogenic belt: A new interpretation of the Hoggar shield (Algerian Sahara). *Geologische Rundschau 1978 67:2*, *67*(2), 357–388. https://doi.org/10.1007/BF01802795

Bertrand, J.-M., Michard, A., Boullier, A.-M., & Dautel, D. (1986). Structure and U/Pb geochronology of Central Hoggar (Algeria): A reappraisal of its Pan-African evolution. *Tectonics*, *5*(7), 955–972.

Black, R., Latouche, L., Liegeois, J. P., Caby, R., & Bertrand, J. M. (1994). Pan-African displaced terranes in the Tuareg Shield (central Sahara). *Geology*, *22*(7), 641–644. https://doi.org/10.1130/0091-7613(1994)022<0641:PADTIT>2.3.CO;2

Blakely, R. J. (1996). Potential Theory in Gravity and Magnetic. *Cambridge University Press*, 1-464 p.

Boissonnas, J. (1973). Les granites à structures concentriques et autres granites tardifs de la chaine pharusienne en Ahhagar (Sahara central, Algérie). *Companyéd. Bureau de Recherches Géol. et Minières, Centre Nat. de La Recherche Scientif*, 1-662 p. https://books.google.dz/books?id=w-gIxwEACAAJ

Bouhkalfa, L. (2002). *Les formations volcano-sédimentaires néoprotérozoïques de la branche orientale de la chaîne pharusienne (Hoggar occidental, Algérie) : lithologie et géochimie.* 9-31 p.

Bournas, N. (2001). Interpretation des donnees aerogeophysiques acquises au-dessus du hoggar oriental. *Université Des Sciences et de La Technologie Houari Boumediene (USTHB), Mémoire de Doctorat*, 1-250 p.

Bournas, N., Touré, A., Balboné, M., Zagré, P. S., Ouédraogo, A., Khaled, K., Prikhodko, A., & Legault, J. (2019). Use of machine learning techniques on airborne geophysical data for mineral resources exploration in Burkina Faso. *ASEG Extended Abstracts*, *2019*(1), 1–4. https://doi.org/10.1080/22020586.2019.12072949

Boyd, D. (1967). The contribution of airborne magnetic surveys to geological mapping. *Mining and Groundwater Geophysics*, 213–227.

Boyd, D. M., & Isles, D. J. (2007). "Proceedings of Exploration 07: Fifth Decennial International Conference on Mineral Exploration. *Geological Interpretation of Airborne Magnetic Surveys - 40 Years On*, 491–505.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/bf00058655

Breiman, L. (2001). Random Forests. *Machine Learning 2001 45:1*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees Regression Trees*. Wadsworth/Thomson Learning.

Briggs, I. C. (1974). Machine contouring using minimum curvature. *Geophysics*, *39*(1), 39–48. https://doi.org/10.1190/1.1440410

Caby, R. (2003). Terrane assembly and geodynamic evolution of central-western Hoggar: A synthesis. *Journal of African Earth Sciences*, *37*(3–4), 133–159. https://doi.org/10.1016/j.jafrearsci.2003.05.003

Caby, R., Andreopoulos-Renaud, U., & Gravelle, M. (1982). *Cadre géologique et géochronologique U/Pb sur zircon des batholites précoces dans le segment pan-africain du Hoggar central (Algérie)*. https://doi.org/https://doi.org/10.2113/gssgfbull.S7-XXIV.4.677

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chikhaoui, M. (1981). *Les roches volcaniques du Protérozoïque supérieur de la chaîne panafricaine du Hoggar et Anti-atlas.* Université. Montpellier.

Chuck, C., & Laura, C. (n.d.). *Magnetics Introduction to Filtering using the 2D Fourier Transform Potential Fields Objectives for this week*.

Cracknell, M. J., & Reading, A. M. (2013). The upside of uncertainty: Identification of lithology contact zones from airborne geophysics and satellite data using random forests and support vector machines. *Geophysics*, *78*(3), WB113--WB126.

Cracknell, M. J., & Reading, A. M. (2014). Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers and Geosciences*, *63*, 22–33. https://doi.org/10.1016/j.cageo.2013.10.008

Dallmeyer, R. D. (1990). The West African orogens and Circum-Atlantic correlatives. *Avalonian and Cadomian Geology of the North Atlantic*, 134–165. https://doi.org/10.1007/978-94-009-0401-9_8

Darnly, A. G. (1972). Airbone gamma-ray survey techniques. *Uranium Prospectiong Handbook*, 174–211.

Dauth, C. (1997). Airborne magnetic, radiometric and satellite imagery for regolith mapping in the yilgarn craton of western australia. *Exploration Geophysics*, *28*(2), 199–203. https://doi.org/10.1071/EG997199

Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and Unsupervised Discretization of Continuous Features BT - Machine Learning Proceedings 1995. *Machine Learning Proceedings 1995*, 194–202. https://doi.org/https://doi.org/10.1016/B978-1-55860-377-6.50032-3

Dupont, P. L. (1987). *Pétrologie et géochimie des ensembles magmatiques Pharusien I et Pharusien II dans le rameau oriental de la chaîne pharusienne (Hoggar, Algérie), Implications géodynamiques pour l'évolution d'une chaîne mobile au Protérozoïque supérieur.* Université Nancy.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, *28*(2), 337–407.

G. Nelson, D. R. (1991). *"Gamma-Ray Interactions with Matter", Passive Nondestructive Analysis of Nuclear Materials*. *I*.

Gabriel, D. (2007). *Mapping Petrological Patterns in the Refouge Granite By Integrating Geochemical, Vnir-Swir and Gamma- Ray Spectrometry Data Mapping Petrological Patterns in the Regoufe Granite by Integrating Geochemical, Vnir-Swir and Gamma-Ray Spectrometry Data*. *February*, 1–82.

Geosoft Inc. (2014). *Oasis montaj How-to guide* (p. 21). Geosoft Inc.

Gilmore, G. R. (2008). Practical Gamma-Ray Spectrometry: Second Edition. In *Practical Gamma-Ray Spectrometry: Second Edition*. https://doi.org/10.1002/9780470861981

Grasty, R. L., & Minty, B. R. S. (1995). *A Guide to the Technical Specifications for Airborne Gamma-Ray Surveys*. 89.

Gravelle, M. (1969). *Recherche sur la géologie du socle précambrien de l'Ahaggar Centro-Occidental dans la région de Silet-Tibehaouine.* Faculté des sciences de l'Université de Paris, Paris.

Greenly, E., & Williams, H. (1930). Methods in Geological Surveying with one coloured map and numerous illustrations in the text. *Geological Magazine*, *67*(9), 428–430. https://doi.org/10.1017/s0016756800100524

Groune, D. (2019). Application des Filtres Numériques aux Données Aérogéophysiques pour la Délimitation des Indices Uranifères dans le Hoggar Occidental. *UNIVERSITE M'HAMED BOUGARA-BOUMERDES, Mémoire de Doctorat*, 1-190 p. http://dlibrary.univ-boumerdes.dz:8080/bitstream/123456789/5665/1/Doctorat-Groune.pdf

Gunnarsdóttir, E. L. (2012). The Earth's Magnetic Field. *Háskóli Íslands Physics Department, Baccalaureate Thesis,* 1-54 p.

Harris, J. R., Ford, K. L., & Charbonneau, B. W. (2009). Application of gamma-ray spectrometer data for lithological mapping in a cordilleran environment, Sekwi Region, NWT. *Canadian Journal of Remote Sensing*, *35*(August 2009), S12–S30. https://doi.org/10.5589/m09-022

Harris, J. R., & Grunsky, E. C. (2015). Predictive lithological mapping of Canada's North using Random Forest classification applied to geophysical and geochemical data. *Computers \& Geosciences*, *80*, 9–25.

Harris, J. R., Juan, H. X., Rainbird, R., & Behnia, P. (2014). Remote predictive mapping 6. A comparison of different remotely sensed data for classifying bedrock types in canada's arctic: Application of the robust classification method and random forests. *Geoscience Canada*, *41*(4), 557–584. https://doi.org/10.12789/geocanj.2014.41.062

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2 nd, Vol. 2). Springer New York, NY. https://doi.org/https://doi.org/10.1007/978-0-387-84858-7

Hinze, W. J., Von Frese, R. R. B., & Saad, A. H. (2010). Gravity and magnetic exploration: Principles, practices, and applications. In *Gravity and Magnetic Exploration: Principles, Practices, and Applications*. https://doi.org/10.1017/CBO9780511843129

Hofmann, T. (2001). Unsupervised learning by probabilistic Latent Semantic Analysis. *Machine Learning*, *42*(1–2), 177–196. https://doi.org/10.1023/A:1007617005950

IAEA, 2003. (2003). *Guidelines for Radioelement Mapping Using Gamma Ray Spectrometry Data* (Issue 1363). International atomic energy agency. https://www.iaea.org/publications/6746/guidelines-for-radioelement-mapping-using-gamma-ray-spectrometry-data

Isles, D. J., & Rankin, L. R. (2013). *Geological interpretation of aeromagnetic data*. Society of Exploration Geophysicists and Australian Society of Exploration~….

Joharestani, M. Z., Cao, C., Ni, X., Bashir, B., & Talebiesfandarani, S. (2019). PM2.5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. *Atmosphere*, *10*(7), 373. https://doi.org/10.3390/atmos10070373

Kang, Y., Ozdogan, M., Zhu, X., Ye, Z., Hain, C., & Anderson, M. (2020). Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest. *Environmental Research Letters*, *15*(6). https://doi.org/10.1088/1748-9326/ab7df9

Kotsiantis, S., Zaharakis, I., … P. P. applications in, & 2007, undefined. (2007). Supervised machine learning: A review of classification techniques. *Books.Google.Com*, *31*, 249–268. https://books.google.com/books?hl=en&lr=&id=vLiTXDHr_sYC&oi=fnd&pg=PA3&dq=Supervised+Machine+Learning:+A+Review+of+Classification+Techniques&ots=CZqsAs_Ghi&sig=iGW00fIx3njgGUgQG3dbYX5TpKE

Kovačevič, M., Bajat, B., Trivič, B., & Pavlovič, R. (2009). Geological units classification of multispectral images by using support vector machines. *International Conference on Intelligent Networking and Collaborative Systems, INCoS 2009*, *May 2014*, 267–272. https://doi.org/10.1109/INCOS.2009.44

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. In *Applied Predictive Modeling*. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Kuhn, S., Cracknell, M. J., & Reading, A. M. (2018). Lithologic mapping using Random Forests applied to geophysical and remote-sensing data: A demonstration study from the Eastern Goldfields of Australia. *Geophysics*, *83*(4), B183–B193. https://doi.org/10.1190/geo2017-0590.1

Kusky, T. M., & Ramadan, T. M. (2002). Structural controls on Neoproterozoic mineralization in the South Eastern Desert, Egypt: An integrated field, Landsat TM, and SIR-C/X SAR approach. *Journal of African Earth Sciences*, *35*(1), 107–121. https://doi.org/10.1016/S0899-5362(02)00029-5

Landgrebe, T. C. W., & Paclik, P. (2010). The ROC skeleton for multiclass ROC estimation. *Pattern Recognition Letters*, *31*(9), 949–958. https://doi.org/10.1016/j.patrec.2009.12.037

Laurent Marescot. (2017). *Magnetic Surveying*. https://docplayer.net/20749830-Magnetic-surveying-dr-laurent-marescot-laurent-tomoquest-com.html

Lee, Y., Han, D., Ahn, M. H., Im, J., & Lee, S. J. (2019). Retrieval of total precipitable water from Himawari-8 AHI data: A comparison of random forest, extreme gradient boosting, and deep neural network. *Remote Sensing*, *11*(15), 1741. https://doi.org/10.3390/rs11151741

Lelubre, M. (1952). *Recherche sur la géologie de l'Ahaggar central et occidental*. Bulletin du service de la carte géologique de l'Algérie. https://books.google.dz/books?id=1fZ1zQEACAAJ

Liégeois, J. P. (2019). A new synthetic geological map of the tuareg shield: an overview of its global structure and geological evolution. In *Springer Geology* (pp. 83–107). https://doi.org/10.1007/978-3-319-96794-3_2

Liégeois, J. P., Latouche, L., Boughrara, M., Navez, J., & Guiraud, M. (2003). The LATEA metacraton (Central Hoggar, Tuareg shield, Algeria): behaviour of an old passive margin during the Pan-African orogeny. *Journal of African Earth Sciences*, *37*(3–4), 161–190. https://doi.org/10.1016/J.JAFREARSCI.2003.05.004

Lisle, R. J., Brabham, P., & Barnes, J. (2013). Basic Geologic Mapping, Fifth Edition. In *Environmental & Engineering Geoscience* (Vol. 19, Issue 2). https://doi.org/10.2113/gseegeosci.19.2.196

Love, J. J. (2008). Magnetic monitoring of earth and space. *Physics Today*, *61*(2), 31–37. https://doi.org/10.1063/1.2883907

Macleod, I. N., Jones, K. A., & Dai, T. (1993). 3-D analytic signal in the interpretation of total magnetic field data at low magnetic latitudes. *Exploration Geophysics*, *24*, 679–688.

Macmillan, S., & Finlay, C. (2011). The International Geomagnetic Reference Field. *Geomagnetic Observations and Models*, 265–276. https://doi.org/10.1007/978-90-481-9858-0_10

Marsland, S. (2014). *Machine Learning: An Algorithmic Perspective, Second Edition* (2nd ed.). Chapman & Hall/CRC.

Oommen, T., Misra, D., Twarakavi, N. K. C., Prakash, A., Sahoo, B., & Bandopadhyay, S. (2008). An objective analysis of support vector machine-based classification for remote sensing. *Mathematical Geosciences*, *40*(4), 409–424. https://doi.org/10.1007/s11004-008-9156-6

Othman, A. A., & Gloaguen, R. (2017). Integration of spectral, spatial and morphometric data into lithological mapping: A comparison of different Machine Learning Algorithms in the Kurdistan Region, NE Iraq. *Journal of Asian Earth Sciences*, *146*, 90–102. https://doi.org/10.1016/j.jseaes.2017.05.005

Ourhzif, Z., Algouti, A., Algouti, A., & Hadach, F. (2019). Lithological mapping using landsat 8 oli and aster multispectral data in imini-ounilla district south high atlas of marrakech. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *42*(2/W13), 1255–1262. https://doi.org/10.5194/isprs-archives-XLII-2-W13-1255-2019

Paquette, J. L., Caby, R., Djouadi, M. T., & Bouchez, J. L. (1998). U–Pb dating of the end of the Pan-African orogeny in the Tuareg shield: the post-collisional syn-shear Tioueine pluton (Western Hoggar, Algeria). *Lithos*, *45*(1–4), 245–253. https://doi.org/10.1016/S0024-4937(98)00034-6

Paterson, P., Grant, G., & Watson, W. (2003). *Paterson, Grant & Watson Limited. The GXperts Microlevelling for OASIS montaj ^{TM}.*

Pieter, P. (2021). *Butterworth Filters*. https://tttapa.github.io/Pages/Mathematics/Systems-and-Control-Theory/Analog-Filters/Butterworth-Filters.html

Podder, P., Mehedi Hasan, Md., Rafiqul Islam, Md., & Sayeed, M. (2014). Design and Implementation of Butterworth, Chebyshev-I and Elliptic Filter for Speech Signal Analysis. *International Journal of Computer Applications*, *98*(7), 12–18. https://doi.org/10.5120/17195-7390

Reeves, C. (2006). Aeromagnetic Surveys. *Earthworks*, 155.

Reford, M. S., & Sumner, J. S. (1964). Aeromagnetics. *Geophysics*, *29*(4), 482–516.

R.J., H., Schetselaar, E., & Behni, P. (2012). Remote Predictive Mapping: An Approach for the Geological Mapping of Canada's Arctic. *Earth Sciences*, *February*. https://doi.org/10.5772/25475

Rosenblum, S., & Brownfield, I. K. (2000). *Magnetic susceptibilities of minerals*. Citeseer. https://doi.org/https://doi.org/10.3133/ofr99529

RStudio. (2019). RStudio: Integrated development environment for R (Version 0.97.311). In *The Journal of Wildlife Management* (Vol. 75, Issue 8, pp. 1753–1766). http://doi.wiley.com/10.1002/jwmg.232

Sahin, E. K. (2020). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Applied Sciences*, *2*(7), 1–17. https://doi.org/10.1007/s42452-020-3060-1

Sayler, K. (2020). *Landsat 8 Collection 2 (C2) Level 2 Science Product (L2SP) Guide*. *2*(September).

Schapire, R. E. (1999). A brief introduction to boosting. *IJCAI International Joint Conference on Artificial Intelligence*, *2*, 1401–1406.

Seligman, H. (1992). Airborne gamma ray spectrometer surveying, technical reports series no. 323. *International Journal of Radiation Applications and Instrumentation. Part A. Applied Radiation and Isotopes*, *43*(3), 469. https://doi.org/10.1016/0883-2889(92)90124-W

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Silva, J. B. C. (1986). Reduction to the pole as an inverse problem and its application to low-latitude anomalies. *Geophysics*, *51*(2), 369–382.

Slavinski, H., Morris, B., Ugalde, H., Spicer, B., Skulski, T., & Rogers, N. (2010). Integration of lithological, geophysical, and remote sensing information: A basis for remote predictive geological mapping of the Baie Verte Peninsula, Newfoundland. *Canadian Journal of Remote Sensing*, *36*(2), 99–118. https://doi.org/10.5589/m10-031

Sun, T., Chen, F., Zhong, L., Liu, W., & Wang, Y. (2019). GIS-based mineral prospectivity mapping using machine learning methods: A case study from Tongling ore district, eastern China. *Ore Geology Reviews*, *109*, 26–49. https://doi.org/10.1016/j.oregeorev.2019.04.003

Sutton, R., & Barto, A. (2015). The Reinforcement Learning Problem. *Robotica*, *17*(January 1999), 229–235. https://doi.org/https://doi.org/10.1017/S0263574799271172

Swain, C. J. (1976). A FORTRAN IV program for interpolating irregularly spaced data using the difference equations for minimum curvature. *Computers and Geosciences*, *1*(4), 231–240. https://doi.org/10.1016/0098-3004(76)90071-6

Thomas, A. (2020). Processing and analysis of aster and landsat 8 scenes to aid in geological mapping: A case study of murchison greenstone belt area, South Africa. *Geomatics and Environmental Engineering*, *14*(3), 107–123. https://doi.org/10.7494/geom.2020.14.3.107

U.S. Geological Survey. (2016). Landsat 8 Data Users Handbook. *Nasa*, *8*(June), 97. https://landsat.usgs.gov/documents/Landsat8DataUsersHandbook.pdf

Van Der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, *6*(1). https://doi.org/10.2202/1544-6115.1309

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Morgan Kaufmann Publishers Inc. https://doi.org/https://doi.org/10.1016/C2009-0-19715-5

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Yang, G., Collins, M. J., & Gong, P. (1998). Multisource data selection for lithologic classification with artificial neural networks. *International Journal of Remote Sensing*, *19*(18), 3675–3680. https://doi.org/10.1080/014311698213885

Yawsangratt, S. (2002). *A gravity study of northern Botswana: a new perspective and its implications for regional geology*. http://www.itc.nl/library/papers/msc_2002/ereg/yawsangratt.pdf

Youssef, A. M., & Pourghasemi, H. R. (2021). Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Asir Region, Saudi Arabia. *Geoscience Frontiers*, *12*(2), 639–655.

Youssef, M. A. S., & Elkhodary, S. T. (2013). Utilization of airborne gamma ray spectrometric data for geological mapping, radioactive mineral exploration and environmental monitoring of southeastern Aswan city, South Eastern Desert, Egypt. *Geophysical Journal International*, *195*(3), 1689–1700. https://doi.org/10.1093/gji/ggt375

Zeghouane, H., & Hamis, A. (2009). *Geological map of Sahara; Silet area*.