# UNIVERSITY OF NAIROBI

# Design of a Radioisotope-Excited EDXRF System for Rare Earth Elements Analysis in Geological Samples

**BY**

**TARUS BRIAN KIPLABAT**

**S56/78344/2015**

**B.Sc (Hons), Physics**

A Thesis Submitted in Partial Fulfillment of the Requirements for Award of the Degree of Master of Science in Nuclear Science of the University of Nairobi.

# DECLARATION

I declare that this thesis is my original work and has not been submitted elsewhere for examination, award of a degree or publication. Where other people's work or my own work has been used, this has properly been acknowledged and referenced in accordance with the University of Nairobi's requirements.

Signature: ……………………………… Date: 21/07/2022

**Tarus Brian Kiplabat**

**S56/78344/2015**

Department of Electrical and Information Engineering

Faculty of Engineering

University of Nairobi

This thesis is submitted for examination with our approval as research supervisors:

Signature Date

Dr. M. I. Kaniu
Department of Physics
University of Nairobi
P.O Box 30197-00100,
Nairobi, Kenya.
ikaniu@uonbi.ac.ke

……………………… 21st July 2022

Prof. M. J. Gatari
Department of Electrical &
Information Engineering
University of Nairobi
P.O Box 30197-00100,
Nairobi, Kenya.
mgatari@uonbi.ac.ke

……………….… *21 July 2022*

# DEDICATION

*“In memory of Mr. David Maina who initiated and immensely inspired this research”*

# ACKNOWLEDGEMENT

First and foremost, I’d like to offer special thanks to Mr. David Maina, who, although no longer with us, inspired me to undertake this research, guided its progress, and offered many of the resources this research depended upon.

Immense gratitude goes to Dr. Ian Kaniu for taking over the main supervisory role at a short notice, lending his vast knowledge in chemometric XRF spectral analysis, introducing me to his team, advice on machine learning methods, fast tracking, and following up every week for the past year. A wholehearted big thank you to Prof. Michael Gatari for supervision, providing resources, and charting the course to the successful completion of this research. I will forever be indebted to your mentorship and immense contribution in this research.

I am grateful to the University of Nairobi, this scholarly journey was possible courtesy of the MSc scholarship without which my studies at the Institute of Nuclear Science and Technology wouldn’t have been possible. Special thanks to Mr. Simion Bartilol and Mr. Michael Mangala for their suggestions on lead block pressing and insightful questions on machine learning. I was impressed by the teams at the mechanical engineering and carpentry workshop for the impeccable technical abilities. Many thanks to Messrs. Ndung’u Ndegwa (MSc.), Kennedy Odongo (MSc.), and Robert Ikedi (MSc.) for their codes, techniques, and comradely support.

Special thanks to my family for their immense support during the tough and uncertain times. My other family, the Ogenches’, a thank you won’t be enough, I am forever grateful. To Cynthia, you’ve made many sacrifices this far, thank you, *mabuch*, it is your turn. To my son Barak, so much can be achieved in your sleep ☺☺☺.

# ABSTRACT

Quantification of rare earth elements (REEs) using Energy Dispersive X-ray Fluorescence (EDXRF) is severely affected by presence of other elements, concentration, nature and energy of the excitation source, detection system, and analysis technique. Conventional REE analysis methods such as NAA, XRF, ICP-(OES and MS) are expensive, unavailable, and involve lengthy sample preparation. Robust elemental quantification using EDXRF and machine learning techniques have been demonstrated in many settings. Therefore, this study aimed at designing a radioisotope excited EDXRF instrument and using chemometrics and machine learning (ML) to quantify REEs in geological materials and starch. The instrument was built using an annular Americium-241 excitation source with an activity of 106 mCi and a peltier cooled SDD Detector. Analytical samples were prepared by schematically mixing REEs salts; Dy, Y, and Ce in geological and starch matrices. The EDXRF setup was used to acquire spectra and R software was used for data visualization, feature selection, scatter ratio correction, performance of PCA for dimension reduction, and to build ML models; SVR, ANN, and RF. Instrument shielding resulted in reduction of doses from 1.68 mSv/h with sample chamber door open to 250 nSv/h while closed. Results of scatter ratio correction established that regions for rock and starch matrices were different, 16.4~17.4 keV for rock and 18.7~20.6 keV for starch. RF model of Cerium in rock attained lowest root mean squared error of prediction (RMSEP) of 106 ppm at 57% accuracy using 9 PCs with limit of detection (LoD) of 9 ppm. RF model of Dy in rock attained the lowest RMSEP of 79 ppm at an accuracy of 41% using 3 PCs with LoD of 20 ppm. RF model of Y in rock attained lowest RMSEP of 140 ppm at an accuracy of 99.9% using 2 PCs with LoD of 64 ppm. RF model of Ce in starch attained lowest RMSEP of 30 ppm at 90% accuracy using 7 PCs with LoD of 6 ppm. NN model of Dy in starch attained lowest RMSEP of 25 ppm at 95% accuracy using 5 PCs with LoD of 7 ppm. NN model of Y in starch attained lowest RMSEP of 112 ppm at an accuracy of 99.99% using only 1 PC with LoD of 71 ppm. RF model of Ti in starch attained lowest RMSEP of 41 ppm at 78% accuracy using 5 PCs with LoD of 7 ppm. NN model of Nb in starch attained lowest RMSEP of 14 ppm at 98% accuracy using 9 PCs with LoD of 4 ppm. This study showed that with limited resources, an XRF instrument setup can be used with machine learning techniques to quantify REEs in geological and other matrices.

## TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| ANN | Artificial Neural Networks |
| EDXRF | Energy Dispersive X-Ray Fluorescence and Scattering |
| FWHM | Full Width at Half Maximum |
| HREE | Heavy Rare Earth Element |
| ICP-MS | Inductively Coupled Plasma Mass Spectrometry |
| ICP-OES | Inductively Coupled Plasma Optical Emission Spectrometry |
| INAA | Instrumental Neutron Activation Analysis |
| IUPAC | Union of Pure and Applied Chemistry |
| LREE | Light Rare Earth Element |
| MLR | Multilinear Regression |
| NORM | Naturally Occurring Radioactive Material |
| PC | Principal Component |
| PCA | Principal Component Analysis |
| PCR | Principal Component Regression |
| PLSR | Partial Least Squares Regression |
| RE | Rare Earth |
| REE | Rare Earth Element |
| RMSEP | Root Mean Squared Error of Prediction |
| RNAA | Radiochemical Neutron Activation Analysis |

| | |
|---|---|
| SDD | Silicon Drift Detector |
| SRM | Standard Reference Material |
| TVT | Tenth Value Thickness |
| UV-VIS | Ultraviolet–visible spectroscopy |
| WDXRF | Wavelength dispersive X-Ray fluorescence |
| XRF | X-Ray Fluorescence |

# CHAPTER 1
# INTRODUCTION

## 1.1 Background to the study

Energy Dispersive X-Ray Fluorescence (EDXRF) spectroscopy is an effective non-destructive method in the analysis of elements in materials. The quantitative analysis of rare earth elements (REEs) in geological samples in EDXRF however is challenging because of severe matrix effects, spectral overlap and similarities among the elements (Zawisza *et al.*, 2011). These challenges can be overcome using the relatively new field of chemometrics, a mathematical method applied to chemical problems used to uncover patterns and relationships in EDXRF analysis data (Andrade-Garda, 2009; Angeyo *et al.*, 2012; Kaniu *et al.*, 2012; Kaniu and Angeyo, 2015).s

EDXRF being a non-destructive technique that requires little sample preparation and with a high throughput makes it one of the cheapest spectroscopic methods available (Goldstein and Sivils, 2002). Traditional wet-chemical quantification of REEs e.g., inductively coupled plasma optical emission spectrometry (ICP-OES) and inductively coupled plasma mass spectrometry (ICP-MS) involve lengthy separation and purification treatments that bring great complexity and analytical difficulty, beside overlooking the high chance of contamination errors due to the chemicals used in sample pretreatment (Zawisza *et al.*, 2011). Naturally occurring radioactive materials (NORMs), uranium (U) and thorium (Th) found alongside REEs forms complexes with most reagents used in optical spectroscopy and thus presents obstacles in measurements (Castor and Hedrick, 2006). These methods including Neutron Activation Analysis (NAA) methods are ideal for REE analysis; they are however sophisticated, expensive and unavailable in many laboratories in developing low-income countries. It is thus necessary to develop rapid, robust, and direct techniques that EDXRF and machine learning are readily offering to boost REE analysis in geological samples and subsequently enhance their exploration locally and at a cost that is reachable by many interested persons.

Radioisotope excited EDXRF spectra consist of fluorescent peaks, spectral overlaps, and scatter peaks. Fluorescent peaks originate from the sample being analyzed and the excitation source. Spectral overlap is mainly due to similarities among elements while scatter peaks depend on the energy of excitation source. Ideally, Am-241 emits 59.54 keV gamma rays and is the best excitation source for REEs $K_\alpha$ and $K_\beta$ (except the $K_\beta$ of lutetium) emission lines. Frequency of spectral overlap is greatly reduced with higher excitation energies considering the resolution state of presently available detectors. This however is not the case for such encapsulated excitation sources. It mainly suffers from multiple peaks of ranging energies (van Grieken and Markowicz, 2002). These challenges can be overcome using scatter correction and multivariate techniques (Andrade-Garda, 2009; Angeyo *et al.*, 2012; Kaniu *et al.*, 2012; Kaniu and Angeyo, 2015).

Coherent to incoherent scatter ratio are two widely used corrections in XRF. Markowicz, (2011) noted that the ratio of coherent-to-incoherent scatter can be used for matrix correction. Similarly Sitko, (2006) proposed a model for correction of matrix effects using the ratio of coherent to incoherent scatter. Multivariate techniques have been used in EDXRF to correct for different types of matrices, geometrical setup, sample preparation and much more. Angeyo *et al.*, (2012), used principal component analysis (PCA) and showed that the scatter region (low-Z elements) possessed the most important spectral signatures. Traore *et al.*, (2014) used EDXRF, PCA, and hierarchical clustering analysis (HCA) to discriminate between various matrices. Kaniu and Angeyo, (2015), used EDXRF with PCA, HCA, soft independent modeling of class analogies (SIMCA), PLS regression, PLS discriminant analysis, and ANN in analysis of soils. Various other studies demonstrate the ability of multivariate techniques used in conjunction with EDXRF; Henrich *et al.*, (2000), Kessler *et al.*, (2002), Custo *et al.*, (2002), Goraieb *et al.*, (2007), Enrich *et al.*, (2007), van Es *et al.*, (2009), etc.

## 1.2 Statement of the Problem

In energy dispersive X-ray fluorescence analysis of rare earth elements in geological samples, their excitation energies remain a challenge. The commonly available 5-50 keV tube excitation systems are not sufficient in efficiently exciting the K emission lines of REEs ($K_{\alpha1}$ and $K_{\beta1}$ for Yttrium, Cerium, and Dysprosium ranging between 14.96 ~ 52.18 keV) and thus require a better excitation source (higher energy, >50 keV, tube excitation systems are available but are not cost effective). EDXRF systems are mainly supplied as prepackaged units and thus limits the spectroscopist to the design specifications of the manufacturer.  Spectral overlap is another challenge, the L lines of REEs ($L_{\alpha1}$ and $L_{\beta1}$ for Yttrium, Cerium, and Dysprosium ranging between 1.92 ~ 7.25 keV) overlap with low-atomic-number element lines at low energies. A high energy radioisotope excitation source like the Am-241 is an excellent rare earth element analysis tool, but inherently suffers from its multiple peaks associated with elements in its decay chain. The relatively new field of chemometrics and machine learning can be used to analyze the EDXRF spectrum that results from using such sources. Therefore, this study aimed at designing a radioisotope excited (based on Am-241) EDXRF spectroscopy system and further use chemometrics to determine the quantity of REEs in geological and starch samples.

## 1.3 Research Hypothesis

Given a radioisotope excitation source and an X-ray detector, it possible to design and assemble a simple, cost-effective, and a readily available system that can be combined with machine learning techniques for direct and rapid REE analysis.

## 1.4 Research Objectives

### 1.4.1 General Objective

The main objective of this study was to design a radioisotope excited energy dispersive X-ray fluorescence (EDXRF) system based on Am-241 coupled with machine learning to determine the elemental concentrations of rare earth elements in geological samples.

### 1.4.2 Specific Objectives

(i) To design (fabricate parts, assemble, and calibrate) a radioisotope excited-EDXRF unit based on an annular Am-241 source.

(ii) To develop predictive models of REEs (Ce, Dy, Y, Ti and Nb) based on selected ML techniques namely, support vector regression, artificial neural networks, and random forests.

(iii) To compare the predictive performance of the models in sp. objective (ii) above.

## 1.5 Justification and Significance

REEs are used to manufacture, catalysts, rare earth magnets, phosphors, hard drives, lasers, hybrid engines, optical fibres and many others. They are thus vital and indispensable in modern technologies and are widely utilized in the energy, military, and manufacturing industry (Szumigala and Werdon, 2010). China is currently the worlds' largest producer of REEs, accounting for over 90% of the worlds’ total production. In 2014, the Chinese government imposed restrictions in the export of REEs and the prices shot up over ten times owing to a demand driven market (Haque *et al.*, 2014). Kenya, especially Kwale County has been shown to contain some of the largest rare earth deposits in the world. Cortec Mining Company was in 2012 awarded a license to mine REEs in Mrima Hill in Kwale County (Jha, 2014). Exploration of REEs is especially difficult owing to the unavailability of affordable methods of analysis. EDXRF can be an excellent pre-surveyor technique for potential REEs as it promises a non-destructive, low-cost, accurate, and a high throughput elemental analysis. Combining this with chemometrics would further enhance the robustness of this technique (Zawisza *et al.*, 2011). Using XRF and chemometric techniques, this study furthers knowledge in development of a complete REE analysis equipment that is readily field deployable.

## 1.6 Scope of Study

This study is a report on the design of a conceptual lab-based radioisotope excited EDXRF system based on Am-241 excitation. Simulate REE samples of Cerium, Dysprosium, and Yttrium were developed for geological and starch matrices. Simulate Niobium and titanium in starch were also developed due to their energy similarity to REEs.

# CHAPTER 2
# LITERATURE REVIEW

## 2.1 Overview

This chapter reviews literature on occurrence, uses and exploration of rare earth elements, methods used in REE analysis, design setup of radioisotope excited XRF equipment, matrix correction techniques, and application of multivariate techniques in XRF spectral analysis.

## 2.2 Occurrence, uses and exploration of rare earth elements

Rare earth elements (REEs) are a group of 15 elements, the lanthanides; lanthanum (La), cerium (Ce), praseodymium (Pr), neodymium (Nd), promethium (Pm), samarium (Sm), europium (Eu), gadolinium (Gd), terbium (Tb), dysprosium (Dy), holmium (Ho), erbium (Er), thulium (Tm), ytterbium (Yb), and lutetium (Lu). Considered as REEs, scandium and yttrium are two other elements found in mineral assemblages as the lanthanides and possess similar physical and chemical properties (Szumigala and Werdon, 2010). The International Union of Pure and Applied Chemistry (IUPAC) classifies REEs into light rare earth elements (LREE) and heavy rare earth elements (HREE). LREEs are classified as Lanthanum to Gadolinium (atomic numbers 57-64) and the rest as HREEs including Yttrium (atomic numbers 65-71, and 39) excluding Scandium which is not classified into either category (Pollard and Mapleson, 2013).

REEs are moderately abundant in the earth’s crust. Cerium is the 25th most abundant element of the 78 most common elements at 60 ppm. Thulium and lutetium are the least abundant rare earth elements at about 0.5 ppm (Pollard and Mapleson, 2013). Rocks, minerals, ores, coal, and sand have been shown to contain REEs. Minerals containing HREEs include gadolinite, euxenite, xenotime, fergusonite, yttrotungstite, samarskite, yttrotantalite, and yttrialite. Those containing LREEs include bastnasite, cerianite, allamite, monazite, loparite, lanthanite, cerite, fluocerite, stillwellite, chevinite, and britholite. However, monazite, xenotime and bastnasite are currently the most extracted minerals in the world (Zawisza *et al.*, 2011). About 850 potential deposits have been identified around the world but very few are operational mines. The Bayan Obo (China), Mountain Pass (USA), and Mount Weld (recently opened in Australia) are the prominent operational mines in the world. China is currently the worlds' largest producer of REEs, accounting for over 90% of the worlds' total production (Haque *et al.*, 2014).

The uptake of rare earth elements in modern technologies is ever increasing and governments all over the world are looking to tap into this growing market by exploring their natural rare earth resources. REEs in conjunction with other elements are widely utilized in the manufacturing, defense, and energy industries. REEs are used in the manufacture of catalysts, rare earth magnets, phosphors, hard drives, lasers, hybrid engines, and optical fibers among many others (Castor and Hedrick, 2006; Szumigala and Werdon, 2010). These elements have shown no satisfactory element substitutes and it is predicted that they will be in technological demand over the next half century (Greenfield and Graedel, 2013). The advantages of exploration and subsequent mining of REEs have a direct impact on the environment, therefore an analysis technique is also required to mitigate harmful effects of mining (Liang *et al.*, 2014).

Kenya, especially Kwale County has been shown to contain some of the largest rare earth deposits in the world. The history of exploration and analysis for REEs and other elements in Kwale began in 1934 by Kenya Mines and Geological Department. In 1955 Binge and the Geological Survey of Kenya developed a base Rare Earth Oxide (REO) inventory of Mrima at 32 million tonnes at concentration of 3.1% (Binge and Joubert, 1966). Between 1968 and 1971 Pechiney estimated europium oxide deposits at 12,000 tonnes at 800ppm (Pechiney, 1971). In 1998, United States Geological Survey (USGS) estimated REO size of 50.8 metric tonnes at concentration of 0.59% (Singer, 1998). In 2012, Pacific Wildcat Resources Corporation reported an estimated 2.15 million tonnes of total rare earth oxide (TREO) at average concentration of 4.4% (Pollard and Mapleson, 2013). Cortec Mining Company was in 2012 awarded a license to mine REEs in Mrima Hill in Kwale County (Jha, 2014). Exploration and environmental analysis of REEs is thus vital to the sustainable economic growth of Kenya (Government of Kenya, 2015).

## 2.3 Methods of REE Analysis

Many techniques are dedicated to the analysis of REEs; Instrumental Neutron Activation Analysis (INAA), Radiochemical Neutron Activation Analysis (RNAA), X-ray Fluorescence (XRF), inductively coupled plasma optical emission spectrometry (ICP-

OES), inductively coupled plasma mass spectrometry (ICP-MS), and Ultraviolet–visible spectroscopy (UV-VIS) are the most frequently employed analytical techniques for determination of REEs in geological materials. Outlined below are the various methods in REE analysis.

INAA technique involves non-destructively subjecting REEs to a constant flux of neutrons. A standard REE with known concentration is similarly subjected to the neutron flux. The result of neutrons impinging on REEs results in a radioactive nuclide that emits neutrons, protons, or gamma rays that are characteristic of the nuclide. Quantification follows the comparative analysis of the characteristic gamma rays emitted by the sample and the standard (Gordon *et al.*, 1968). RNAA technique involves radiochemically separating nuclides after subjecting them to a neutron flux, like INAA. This is however a more expensive and destructive method unlike INAA but with a better limit of detection and accuracy (Leclercq and Meyers, 2006). ICP-OES (also known as ICP-AES) technique uses plasma at high temperatures at 6000~10000 Kelvin that excite REEs to produce characteristic electromagnetic radiation aiding in their quantification (Lieser, 2001). ICP-MS technique, similar to the plasma in ICP-OES, ionizes and atomizes elements that are subsequently detected on a mass spectrometer. The techniques is especially favored for its sensitivity and can be used to detect isotopes (Houk, 1986). UV-VIS technique uses the absorbance characteristic particular to all elements. A UV source is shone on a sample, the portion that is transmitted is measured. The measured spectra will not contain the absorbance regions of the elements present in the sample thereby aiding in its detection and quantification (Misra and Dubinskii, 2002). Table 2.1 presents a summary of REE analysis methods. Neutron Activation Analysis (NAA) and wet-chemical analysis techniques using ICP-OES and ICP-MS are the most ideal in the analysis of REEs but they involve lengthy separation and purification treatments that bring complexity and analytical difficulty. Moreover, there are high chances of contamination errors due to the chemicals used in sample pretreatment. These methods are very sophisticated and expensive, therefore unavailable in most laboratories in developing countries (Zawisza *et al.*, 2011). XRF technique involves the use of X-rays or gamma rays to excite elements to produce characteristic X-rays that readily identify elements. This techniques is the subject of the remainder of this work.

Table 2.1: Summary of methods in REE analysis

| Method | Principle | Result | Strength | Weakness |
|---|---|---|---|---|
| INAA | Neutron Impingement | Radioactive nuclide that emits neutrons, protons, or gamma rays | Easier handling of radioistopes than RNAA, and sophisticated | Expensive, lengthy separation |
| RNAA | Radiochemical separation after neutron impingement | Similar to INAA | Sophisticated, and lower limit of detection | Lengthy separation, expensive, and high chance of contamiantion |
| ICP-OES | Plasma at high temperature | Excitation of atom to produce characteristic electromagnetic radiation | Higher detection limit compared to ICP-MS and sophisticated | Expensive, lengthy separation, and high chance of contamination |
| ICP-MS | Plasma at high temperature with mass spectrometer | Mass spectrometry of excited atoms | Very low detection limits (including isotopes) and sophisticated | Expensive, lengthy separation, and high chance of contamination |
| UV-VIS | UV absorbance characteristic of elements | Measurement of un-absorbed UV | Sophisticated | Not ideal |

## 2.4 Design Setup of Radioisotope Excited XRF Units

Various components make up a radioisotope excited XRF unit. Components include; radioactive sources, detectors, electronic instrumentation, windows, X-ray filters for energy selection, cooling systems, and computation. This section describes studies that have employed various designs and components of radioisotope excited XRF.

Szökefalvi-Nagy *et al.*, (2004) used an annular Fe-55 radioisotope excitation source in combination with a Canberra dipstick Si(Li) detectors with a resolution of 175 eV to study originality of paints. The analysis area was Aluminium collimated while the signals were processed by Canberra electronics and the final spectra was collected in a Canberra 8100 multichannel analyzer (MCA). In the same study, Szökefalvi-Nagy *et al.*, (2004) assembled a more compact radioisotope excited device. They used a smaller RITVERC Fe-55 ring source and a 190 eV resolution thermoelectrically cooled AMPTEK XR-100 detector with a PXT/CR power supply unit. The spectra were collected using a Canberra 35+ MCA and a simple computer was used for analysis.

In their soil quality assessment study, Kaniu et al., (2012) used a 25 mCi Cd-109 radioisotope source with a 190 eV (Mn Ka X-ray) resolution EG&G ORTEC Si(Li) detector in a cryostat. The detector was furnished with a beryllium entrance window with a gold contact. Canberra devices used included an amplifier, a power supply, and an analog-to-digital converter (ADC) interfaced to an MCA card. Kaniu and Angeyo, (2015) and Kaniu *et al.*, (2011) used a similar setup to assess chemometrics viability in rapid soil quality assessment.

Durak and Şahin, (1997) in their K-shell fluorescence yields measurements for Ba, Ce, Nd, Gd, Dy, Er and Yb used a 100 mCi Co-57 radioisotope source and a 190 eV resolution (Fe X-ray at 5.9 keV) Ge( Li) detector coupled to a 4096 channel MCA. Aluminium shield was used to suppress low energies of the Co-57 source and lead was used to shield against stray radiation. Durak and Şahin, (1998) used the same setup to perform K-shell fluorescent yield measurements for Cs, Sm, Eu, Ho, Ta, W, Hg, and Pb.

YAP and TANG, (1985) and YAP, (1986) in their studies of Chinese porcelains used an annular 30 mCi Am-241 source. The study used a Si(Li) detector with a beryllium window and microprocessor based MCA coupled to a desktop computer. AXIL (Analysis of X-ray spectra by Iterative Least-squares fitting) computer program was used for spectra analysis. Commonly used photonic radioisotope excitation sources are summarized in Table 2.2 (IAEA, 1970).

Radioisotope excited EDXRF have been in use since the 1970s. The main focus has always been on radiation safety of people and the environment during use and storage of the radioisotope source. The IAEA recommends practices that minimize exposure to radiation, i.e., time, distance, and shielding for such sources as the Am-241 that are used in the lab for XRF (IAEA, 2007).

Radioisotopes are widely used in lab based and portable XRF analyzers. Some common uses include elemental analysis in; alloy sorting and identification, mining and minerals, pulp and paper, environment, fibers, films and coatings, chemicals and process control, plastics, agriculture, cosmetics, pharmaceuticals, and petroleum products (van Grieken and Markowicz, 2002).

Table 2.2: Nuclear properties of radioisotopes used in XRF.

| Radioisotope | Half Life (yr) | Decay Mode | Photon Energy (keV) |
|---|---|---|---|
| $^{55}Fe$ | 2.7 | Electron Capture (EC) | 6 Mn K X-rays |
| $^{238}Pu$ | 86.4 | $\alpha$ | 12-17 U L X-rays |
| $^{109}Cd$ | 1.27 | EC | 88 |
| | | | 22 Ag K X-rays |
| $^{125}I$ | 0.16 | EC | 35 |
| | | | 27 Te K X-rays |
| $^{210}Pb$ | 22 | $\beta$ | 47 |
| | | | 11-13 Bi L X-rays plus bremsstrahlung up to 1.17 MeV |
| $^{241}Am$ | 458 | $\alpha$ | 60 |
| | | | 14-21 Np L X-rays 662 |
| $^{153}Gd$ | 0.65 | EC | 103 |
| | | | 97 |
| | | | 70 |
| | | | 41 Eu K X-rays |
| $^{57}Co$ | 0.74 | EC | 700 |
| | | | 122 |
| | | | 14 |
| | | | 6.4 Fe K X-rays |

## 2.5 Matrix Correction Techniques

Practical XRF experiments especially of medium to thick environmental samples are subject to contain many other elements other than those of interest that ultimately interfere with detection and quantification of elements of interest. Minimizing or correcting these effects are crucial in XRF. Feature selection and coherent to incoherent scatter ratio are two widely used corrections in XRF. This section explores some select works that have used these techniques.

Mikhailov *et al.*, (2002) conducted studies using incoherent to coherent scatter ratios to determine masses of light elements in crystals. Similarly, Mikhailov *et al.*, (2018) proposed a technique to determine the mass of ash in fuels using the incoherent to coherent scattering ratios. The study further proposed a linear calibration algorithm of incoherent to coherent scattering ratios using standard samples. Mikhailov *et al.*, (2020) investigated the dependence of incoherent to coherent scattering ratios to the scatterer atomic number. They found that both single and double composition standards showed similar scattering ratios dependence on the atomic number.

Sitko, (2006) proposed an empirical model for correction of matrix effects on thin membranes using the ratio of coherent to incoherent scatter in XRF. The developed matrix correction method satisfactorily quantified elements of geological origin in the range $19 < Z < 92$ on thin membranes.

Markowicz, (1984) evaluated matrix correction using Compton scattered x-rays for thick samples. Markowicz, (2011) noted that the ratio of fluorescent-to-Compton scatter can be used for matrix correction. Notable demonstration of incoherent to coherent scatter ratios was performed by Burkhalter, (1971). In the study, silver ore concentration in silicaeous matrices with varying concentrations of other elements was to be determined. By developing standards that contain 0.1% of silver and 5% of the other elements, Burkhalter determined that the ratio of incoherent to coherent scatter did not change much ( ±6%) as shown in Table 2.3 (Burkhalter, 1971).

Table 2.3: Coherent to incoherent scatter ratios of silver standard**.**

| **Sample Matrix** | $I_{Ag\ K\alpha}$ **(counts/s)** | $I_{Compton}$ **(counts/s)** | $I_{Ag\ K\alpha}/I_{Compton}$ |
|---|---|---|---|
| $SiO_2$ | 200.0 | 1890 | 0.106 |
| $SiO_2$ + 5% Fe | 150.0 | 1420 | 0.106 |
| $SiO_2$ + 5% Zr | 90.1 | 454 | 0.106 |
| $SiO_2$ + 5% Ba | 139.0 | 1440 | 0.096 |
| $SiO_2$ + 5% Pb | 83.8 | 754 | 0.111 |

## 2.6 Multivariate Techniques in XRF Analysis

Matrix correction techniques have been used extensively with approximation approaches to detect and quantify REEs. REE approximation analysis in XRF considers various factors such as; geometry of source-sample-detector, activation source, configuration of measurement (WDXRF or EDXRF), sample preparation, and more to work. Moreover, approximation techniques are complex and require highly controlled calibration setups. Multivariate techniques works around this problem by introducing a simple learn-by-example approach. This section reviews literature on EDXRF with application of multivariate techniques.

Variable selection is a very important step in XRF analysis using multivariate techniques. Adams and Allen, (1998) showed the dependence of partial least square (PLS) models on efficiency of spectral variables choice in quantification. They obtained results comparable to those of the Lucas Tooth and Price models. Schimidt *et al.*, (2003) used EDXRF and artificial neural networks to determine the concentration of promethium, neodymium, and samarium in silica matrix achieving standard errors of less than 17.5% for all elements. Using PCA and PLS modelling, Goraieb *et al.*, (2007) in their study of sugar sweetness, found that the variables at the scattering regions of the spectrum highly influenced the discrimination of sugars. Angeyo *et al.*, (2012), demonstrated the use of principal component analysis (PCA) with EDXRF in lubricant oil (complex matrix). PCA in their study showed that the scatter region (low-Z elements) possessed the most important spectral signature.

XRF has potential applications in classification and clustering. Henrich *et al.*, (2000) used EDXRF, PCA, and regularized discriminant analysis (RDA) to identify chemical compounds in sealed bottles. They identified Compton and Rayleigh scattering as important spectral regions in classification of chemical content. In a follow up study, Kessler *et al.*, (2002), used EDXRF, PCA, and RDA to classify chemical compounds by their fluorescent peaks. They similarly affirmed the importance of Compton and Rayleigh scatter in classification. Custo *et al.*, (2002) using a WDXRF setup, chemometrics, and simple sample preparation, classified Argentine soils by considering minor and trace elements. Vázquez *et al.*, (2002) used a TXRF setup, PCA, and cluster analysis to demonstrate clustering of various polymers using their coherent and incoherent scatter. Verbi *et al.*, (2005) using EDXRF and PCA on their study of the effects of treatment on paints, reported a clear discrimination of the paints according to treatment using the Compton and Rayleigh scatter regions. Goraieb *et al.*, (2006) in their study of Portland cements using EDXRF and PLS, were able to differentiate cement according to the producer and calcium content. Goraieb *et al.*, (2007) using XRF, PCA, and PLS modelling, demonstrated the ability to classify sugar structures and degree of sweetness presenting an alternative to dependence on human sensory ability. Enrich *et al.*, (2007) used TXRF and PCA to discriminate heavy metal contaminants in honey according to their geographical origin in Argentina. van Es *et al.*, (2009) evaluated the discriminative power of LA-ICP-MS and XRF on document paper (types of matrices) and showed that XRF possessed a fairly good discriminant ability despite lower sensitivity. Traore *et al.*, (2014) using EDXRF, PCA, and hierarchical clustering analysis (HCA) discriminated Senegalese magmatic, metamorphic, and mineralized metamorphic rock matrices.

Study of various matrices such as soil present an avenue for application of multivariate techniques. Kaniu *et al.*, (2011) demonstrated the potential of chemometrics with energy dispersive x-ray fluorescence and spectroscopy (EDXRF) in soil quality assessment (SQA) by analyzing micronutrients (Fe, Cu, and Zn) and macronutrients ($NO_3^-$, $SO_4^{2-}$, and $H_2PO_4^-$) as soil quality indicators. Further, a rapid method for trace macronutrients (C, N, Na, Mg, P) analysis based on chemometrics and EDXRF was developed by Kaniu *et al.*, (2012). Kaniu and Angeyo, (2015), using EDXRF, PCA, HCA, soft independent modeling of class analogies (SIMCA), PLS regression, PLS discriminant analysis, and ANN found a plausible solution to the point of care for soils in a field portable XRF spectrometer.

Matrix interference is a major problem in XRF. Facchin *et al.*, (1999) in their complex matrix studies, compared the performance of ANN against other multivariate techniques in quantifying lead and sulfur in their overlapping region (sulfur K line at 2.31 keV, and lead M line at 2.35 keV). They reported better performance overall of ANN over PLS, POLYPLS (polynomial partial least squares), NNPLS (partial least square neural networks), LR (linear regression) and CI (corrected intensity) models. Nagata *et al.*, (2006) demonstrated the power of multivariate technique (PLS1) over univariate techniques using a synchrotron TXRF setup. They investigated the severe effect of bromine matrix in the quantification of lead and arsenic. They reported lower overall RMSEPs for the multivariate technique (PLS1) than the univariate technique. This demonstrated that multivariate techniques were capable of overcoming severe matrix effects unlike traditional methods.

China being the largest producer of REEs, WU *et al.* (2010), reviewed 20 years of the XRF use in China's REE industry. WU *et al.* (2010) explored XRF use in REE analysis of metals and alloys, ores, soils, concentrates, ore separations, etc. The analyses were grouped into matrix type, measured elements, measurement methods, and calibration methods. Measurements were performed on pressed powders, fused beads, thin films, and solids. Various REE occurrence matrices were considered; geological, biological, petrochemical, electrical, and ferrous and non-ferrous matrices. The long and difficult separation encountered in chemical separation techniques unlike pelletized samples was recognized in heavy-element matrices. Light element matrices showed superior performance in solution paper filter (WU *et al.*, 2010).

This study brings together two aspects of XRF. The first aims at assembling a radioisotope excited XRF system by designing and fabricating parts from readily available materials. This would result in an affordable setup achievable in many laboratories. The second aims at using the developed XRF setup to analyze REEs in complex matrices. The second part combines three distinct techniques; feature selection, coherent-to-incoherent scatter ratio technique, and a multivariate approach. It is the first time that such a combination of techniques are used in REE analysis.

# CHAPTER 3
# THEORETICAL BACKGROUND

## 3.1 Overview

This chapter introduces the concept of X-Ray fluorescence and further focuses on radioisotope excited XRF. Due to importance of radiation protection, a subsection is dedicated. Machine learning is introduced with focus on the techniques used in this study. Metrics in regression modelling are also discussed.

## 3.2 X-Ray Fluorescence (XRF)

X-rays rays are electromagnetic radiation typically at wavelengths between 0.005 to 10 nm with higher wavelengths bordering or overlapping ultraviolet radiation while lower wavelengths overlapping gamma radiation.

X-rays are primarily produced in x-ray tubes. Two types of x-rays are produced; continuous and characteristic x-rays. Continuous x-rays (also known as bremsstrahlung) are produced when electrons are decelerated in a target atom. The minimum possible bremsstrahlung wavelength (reflecting the highest energy) is defined as:

$$\lambda_{min} = \frac{hc}{eV_o} \tag{3.1}$$

Where $h$ is the Planck's constant, $c$ the speed of light, $e$ the electron charge, and $V_o$ the potential difference applied to the X-ray tube. Bremsstrahlung is independent of the composition of the target element. Unlike bremsstrahlung, characteristic x-rays are produced when accelerated electrons eject inner shell electrons in the target. The produced x-rays are measured as definite spikes characteristic of the element(s) used as target (van Grieken and Markowicz, 2002).

### 3.2.1 Interaction of X-ray (and Gamma) Radiation with Matter

X-ray (and gamma) radiation interact with matter through various modes; photoelectric effect, Compton scattering, pair production, Auger effect, and Rayleigh scattering. These modes of interactions are discussed below.

Photoelectric effect is the most favorable mode of interaction in XRF analysis (van Grieken and Markowicz, 2002). Photoelectric effect occurs when an atom absorbs an incident photon dislodging an electron in one of its inner orbitals creating a vacancy. The excited atom rearranges its electron structure by having an electron from a higher energy level transit and filling the created vacancy through which an x-ray photon is emitted. The photon emitted is a characteristic of that atom which enables its identification (Moseley, 1913). Figure 3.1 demonstrates the photoelectric effect (Source: van Grieken and Markowicz, 2002).

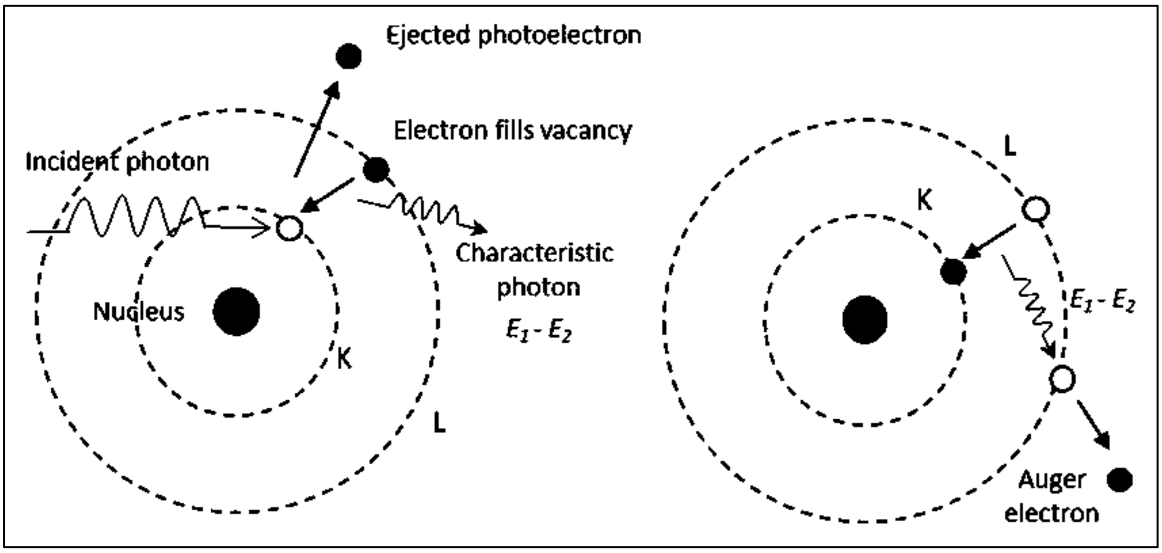


Figure 3.1: Photoelectric and Auger Effect demonstration.

The equation governing production of characteristic X-rays in elements was first established by Moseley in 1913. It states:

$$\frac{1}{\lambda} = K(Z - \delta)^2 \tag{3.2}$$

Where atomic number Z is the atomic number, λ the wavelength of the photon, K is a constant taking on different values for particular spectral series, and δ being the shielding constant. The production of photoelectrons (and subsequent photoelectric effect) can also be achieved using gamma rays, electrons, and other energetic particles.

Compton scattering occurs when photons (mainly gamma and X-rays) partially lose their energy to an electron. This leads to an increase in their wavelength. Figure 3.2 demonstrates the Compton scattering effect (Vandegrift, 2015).

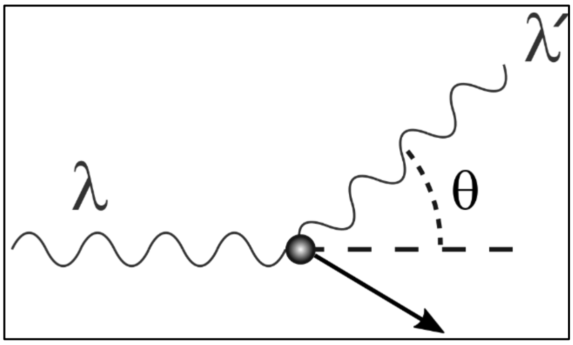


Figure 3.2: Depiction of Compton scattering.

Compton derived the equation to this effect by stating:

$$\lambda' - \lambda = \frac{h}{m_e c}(1 - \cos\theta) \tag{3.3}$$

Where $\lambda$ is the initial wavelength, $\lambda'$ is the wavelength after scattering, $h$ is the Planck constant, $m_e$ is the electron rest mass, $c$ is the speed of light, and $\theta$ is the scattering angle. The quantity $\frac{h}{m_e c}$ is a constant called the Compton wavelength which is equal to$2.43\ x\ 10^{-12}\ m$. at $\theta$ = 0 there is no Compton scattering thus no increase in wavelength while at $\theta$ = 180°, the Compton scatter increases the wavelength by twice the Compton wavelength (Compton, 1923).

Pair Production occurs when high energies photons can create subatomic particles. Primarily, electron-positron pairs are produced when high energy photons interact with nuclei. The condition for this to happen is that the photon has to possess energy higher than the rest mass energy of the particles it produces. The photoelectric effect occurs at relatively lower energies, pair production occurs at high energies, while Compton scatter occurs at energies in-between and depends mainly on the atomic number of the interacting nucleus (Choppin *et al.*, 2013).

Figure 3.3 shows the three competing effects of photon interaction with matter (Choppin *et al.*, 2013).

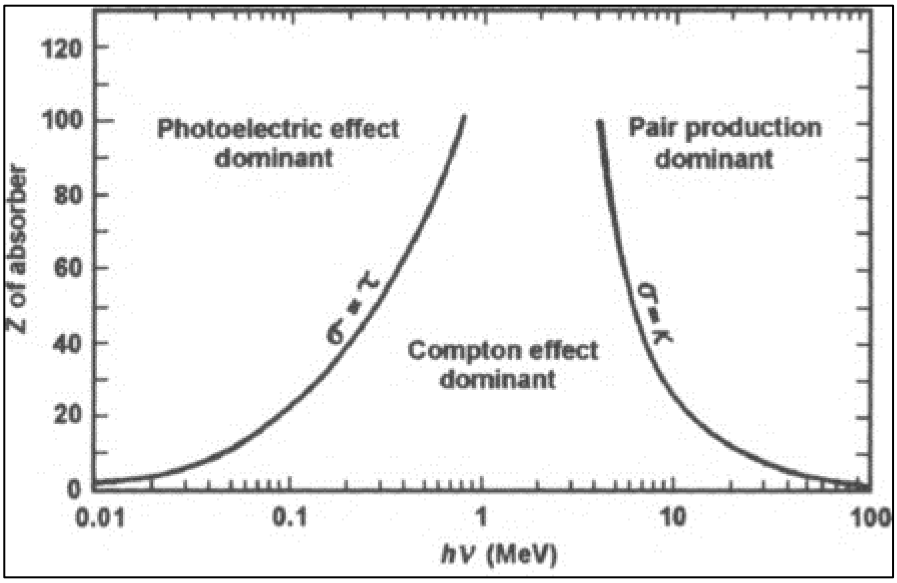


Figure 3.3: Competing photon interaction effects.

Auger effect can be thought of as “double photoelectric effect”. As shown in Figure 3.1, a photoelectron is first emitted by which a characteristic photon is also emitted from the inner-shell electron transition. However, instead of the characteristic photon being emitted entirely it ejects an electron from an outer shell of the atom which results in two photoelectrons (IUPAC, 1997).

Rayleigh scattering (elastic scattering) occurs when electromagnetic interaction undergoes elastic scattering when it interacts with particles whose size is smaller than its wavelength. This phenomenon dominates the weaker energy regions with longer wavelengths.

### 3.2.2 Detection of X-rays (and gamma rays)

The aim of XRF spectroscopy is to measure the characteristic radiation from samples. Measurement techniques are many, however, the goals of the analysis play a major role. The goals influence the detector type of choice. In XRF, the detector is mainly used to separate and measure the characteristic radiation from samples which aids in their identification and quantification.

XRF detectors work mainly by ionization, i.e., a characteristic photon with energy E proportionally produces ionizations in the detector;

$$N = E/e \tag{3.4}$$

Where $N$ is the number of ionizations due to photon of $E$ at a threshold of production of energy $e$.

The standard deviation of the number of ionizations determines the resolution of a detector;

$$\sigma = \sqrt{(FN)} \tag{3.5}$$

Where $F$ is the Fano factor and depends on the energy loss of photons not being purely statistical.

Three factors are critical in detector selection, i.e., resolution, efficiency, and dead time. The resolution is measured as the full width at half maximum (FWHM);

$$FWHM = 2.35\ x\ \sqrt{(eFE)} \tag{3.6}$$

Where the terms are as aforementioned.

The efficiency of a detector depends on the energy of the radiation and the density of the type of detector. The detector dead time is the pulse processing time of the arriving radiation (Lindon *et al.*, 2016).

Three types of detectors are mainly used in XRF spectrometry (gas-filled, scintillation, and semiconductor detectors).

Gas filled detectors take advantage of the effect of radiation in ionization and excitation of gases along the tracks of particles or photons. Three main types of gas-filled detectors are in use; ion chambers, proportional counters, and Geiger-Mueller counters. Ion chambers are the simplest of gas filled detectors and are normally based on the collection of all charges created through direct ionization by application of an electric field. Proportional counters use a stronger electric field than ion chambers to achieve gas multiplication (avalanche) allowing individual released electrons to further create an avalanche of charges independent of each other (Knoll, 2011). Geiger-Mueller (G-M) counters are similar to proportional counters but with a far stronger field and a saturated avalanche. G-M counters produces similar counts irrespective of the original ionization unlike proportional counters (Wilkinson, 1996). The main difference between gas-filled detectors are that ion-chambers are mainly operated in current mode while proportional counters and Geiger-Mueller counters almost exclusively operate in pulse mode.

Scintillation detectors take advantage of some materials producing light upon interaction with radiation. Such materials can be categorized broadly into either organic or inorganic scintillators. Organic scintillators produce light in transition of the energy levels of individual molecules and are independent of the physical state of the molecule. They include pure organic crystals (e.g., anthracene), liquid organic solutions, plastic scintillators, thin-film scintillators, and loaded organic scintillators. Inorganic scintillators on the other hand produce light based on the regular lattice structure of a crystal. These crystals are activated to introduce defects that allow preferred energy transitions to produce light. Examples include alkali hallides (NaI(Th), Cs(Th), etc.,), slow inorganics (BGO, $CdWO_4$, etc.,), unactivated fast organics ($BaF_2$, CsI, etc.,), cerium-activated fast inorganics (GSO, YAP, YAG, etc.,), and glass scintillators (Ce activated Li glass, Tb activated glass, etc.,). The light produced can then be detected by secondary systems using photomultiplier tubes and photodiodes (Knoll, 2011).

Semiconductor diode detectors. Some imitations of scintillation detectors are their bulky nature and poor resolving power that emanate from the inefficient number of events that take place between interaction and signal processing. The resolution of scintillation detectors is fundamentally limited by statistics. Semiconductor detectors overcome this problem by increasing the number of information carriers per radiation interaction. To achieve this, semiconductor-based detectors just like gas-filled detectors create electron-hole pairs within a semiconductor crystal which are accelerated to electrodes in an electric field. This is achievable by exploiting the band structure of solids. The band structure of elements such as Silicon can be changed to have a very low energy requirement for electron-hole pair production. This makes semiconductor detectors versatile in resolution. Examples include Silicon and Germanium based detectors (Knoll, 2011).

XRF can be broadly categorized into two, wavelength dispersive x-ray fluorescence (WDXRF) and energy dispersive x-ray fluorescence (EDXRF). WDXRF uses a crystal monochromator to diffract x-rays to be measured at specific angles while EDXRF uses a semiconductor-based detector to distinguish different x-ray energies emanating from a sample (van Grieken and Markowicz, 2002). Table 3.1 shows excitation energies of REEs that are of interest in XRF (Rover, 2016; Amptek, 2006).

Table 3.1. Excitation energy values of REEs.

| Element | Energy (KeV) | | | |
|---|---|---|---|---|
| | $K_{\alpha 1}$ | $K_{\beta 1}$ | $L_{\alpha 1}$ | $L_{\beta 1}$ |
| Scandium | 4.09 | 4.46 | 0.40 | |
| Yttrium | 14.96 | 16.74 | 1.92 | 2.00 |
| Lanthanum | 33.44 | 37.80 | 4.65 | 5.04 |
| Cerium | 34.72 | 39.26 | 4.84 | 5.26 |
| Praseodymium | 36.02 | 40.75 | 5.03 | 5.49 |
| Neodymium | 37.36 | 42.27 | 5.23 | 5.72 |
| Promethium | 38.65 | 43.96 | 5.43 | 5.96 |
| Samarium | 40.12 | 45.40 | 5.64 | 6.21 |
| Europium | 41.53 | 47.03 | 5.85 | 6.46 |
| Gadolinium | 42.98 | 48.72 | 6.06 | 6.71 |
| Terbium | 44.47 | 50.39 | 6.28 | 6.98 |
| Dysprosium | 45.99 | 52.18 | 6.50 | 7.25 |
| Holmium | 47.53 | 53.93 | 6.72 | 7.53 |
| Erbium | 49.10 | 55.69 | 6.95 | 7.81 |
| Thulium | 50.73 | 57.58 | 7.18 | 8.10 |
| Ytterbium | 52.36 | 59.35 | 7.41 | 8.40 |
| Lutetium | 54.06 | 61.28 | 7.65 | 8.71 |

### 3.2.3 Quantification of Elements in X-ray Fluorescence

Quantitative analysis of elements using x-rays is potentially a complex procedure especially in multi-element samples. This section presents a primer on the relationship between the intensity of the excitation source (based on photons) and the measured characteristic x-rays from the element(s) being analyzed.

Considering a polychromatic source used to excite an element *i* in a homogenous matrix of thickness $T$ (cm), the fluorescent radiation yield is given by:

$$I_i(E_i)d\Omega_1 d\Omega_2 = \frac{d\Omega_1 d\Omega_2}{4\pi}\frac{\varepsilon(E_i)}{\sin\psi_1}\ x \int_{E_{c,i}}^{E_{max}} a_i(E_o)\frac{1-\exp\left[-\rho T(\mu(E_o)\csc\psi_1+\mu(E_i)\csc\psi_2)\right]}{\mu(E_o)\csc\psi_1+\mu(E_i)\csc\psi_2} I_o(E_o)d\mathrm{E}_O \tag{3.7}$$

And that:

$$a_i(E_o) = W_i\tau_i'(E_o)\omega_i p_i\left(1-\frac{1}{j_i}\right) \tag{3.8}$$

Where; $d\Omega_1$ – differential solid angle for the primary (incident) radiation; $d\Omega_2$ – differential solid angle for the characteristic secondary (emerging) radiation; $\varepsilon(E_i)$ – intrinsic detector efficiency for photons at energy $E_i$; $E_{c,i}$ and $E_{max}$ – critical absorption energy of element $i$ and the maximum energy in the excitation spectrum; $\rho$ – density of the specimen (in g/cm$^3$); $\psi_1$ and $\psi_2$ – effective incidence and takeoff angles; $\mu(E_o)$ and $\mu(E_i)$ – total mass attenuation coefficients (in cm$^2$/g) for the whole specimen at energies $E_o$ and $E_i$ respectively; $I_O(E_O)d\mathrm{E}_O$ – number of incident photons per second per steradian in the energy interval $E_o$ and $E_o + d\mathrm{E}_O$ ; $W_i$ – weight fraction of the *i*th element; and $\tau_i'(E_o)$ – total photoelectric mass absorption coefficient for the *i*th element at the energy $E_o$ (in cm$^2$/g).

Equation (3.7) shows that the emerging characteristic radiation is modified by the total mass attenuation coefficients $\mu(E_o)$ and $\mu(E_i)$ which is a major source of the matrix effects in XRF technique.

Equation (3.7) can be further simplified for thin and thick samples. Thin samples eliminate the matrix effects problem while thick samples benefit from the fact that beyond certain thickness there is no practical increase in the intensity of the analyte (saturation mass) (Jenkins *et al.*, 1981; Tertian and Claisse, 1982).

## 3.3 Radioisotope Excited X-ray fluorescence

Radioisotope excited XRF is a vital tool in elemental analysis of materials. It is particularly preferred over x-ray tube excitation for its cost, simplicity, stability of x-rays, monoenergetic rays, size, weight, and ruggedness. Semiconductor based detectors have been widely used in conjunction with radioisotope excited XRF systems (Knoll, 2011).

### 3.3.1 Elemental analysis in radioisotope excited XRF

Moseley first established the relationship between atomic number (Z) and the wavelength (λ) of the emitted photon (Moseley, 1913) according to Equation (3.2). This property readily identifies the atoms present in a sample. However, this is not a simple procedure when considering samples like soils and rocks with many elements which interfere spectrally with each other. Spectral overlap, noise, low detector resolution, sample grain size, and geometry of detection are some effects that a spectroscopist has to deal with. Filtering, smoothing, and background removal techniques have been adopted to ease the identification of peaks associated with particular atoms in EDXRF spectroscopy (Gauglitz and Tuan, 2003).

In quantitative analysis, the fact that fluorescent intensities in XRF spectra are a simple reflection of an elements' concentration in a sample is a gross assumption when considering geological materials. It is a complicated process because measured intensities do not depend solely on the concentration of the analyte. Measured intensity depends on many factors; accompanying elements (the matrix), type of sample (solid, liquid, powder, etc.), sample preparation method, size and shape of sample, geometry of measurement setup, irradiation flux, irradiation size, spectral distribution of the exciting radiation, and resolution of the detection system (van Grieken and Markowicz, 2002). The matrix effect presents the greatest challenge in XRF analysis especially in many-element matrices.

The concentration of an analyte in a sample is determined by its fluorescent intensity in combination with the fluorescent x-rays of the matrix elements and compton-scattered x-rays. When the *K* shell x-rays of an analyte *i* is excited by monoenergetic beam of x-rays in an infinitely thick sample, and the incident and emitted rays are normal to the surface of the sample the detected intensity $I_i$ of the $K_\alpha$ x-rays of the analyte $i$ is approximately given by;

$$I_i = \frac{G\varepsilon(E_i)a_i(E_o)I_o(E_o)}{\mu(E_o) + \mu(E_i)} \tag{3.9}$$

where; $G$- Geometrical constant, $\varepsilon(E_i)$- intrinsic efficiency of the detector to the x-rays of the analyte $i$; $a_i(E_o)$ - $W_i\tau_i'(E_o)\omega_i\rho_i\left(1 - \frac{1}{j_i}\right)$; $I_o(E_o)$ -the source emission (photons/s); $\mu(E_o), \mu(E_i)$- mass absorption coefficients for the exciting radiation with energy $E_o$ and the characteristic radiation with energy $E_i$, respectively, in the sample (cm$^2$/g); $\tau_i'(E_o)$- total photoelectric mass absorption coefficient for the $i$th element at energy $E_o$ (cm$^2$/g); $\omega_i$- the $K$ shell fluorescent yield for the analyte $i$; $\rho_i$- relative transition probability for $K_\alpha$ lines of analyte $i$; and $j_i$- jump ratio.

Enhancement has been assumed to be negligible in Equation (3.9). For L and M shell fluorescence x-rays, their intensities can be calculated similarly. For radioisotopes having more than one energy, intensities can be calculated separately for each emitted energy and the total intensity is determined by summing the products of intensities and the probability of emission of that energy by the radioisotope.

X-rays are scattered from the sample and its surroundings to the detector by mechanisms of coherent and incoherent (compton) scattering. There is no energy loss in coherent scattering unlike incoherent scattering (compton scattering). The compton scattered energy is given by;

$$E = E_o/(1 + \gamma(1 - cos\theta)) \tag{3.10}$$

Where; $E_o$ is the incident x-ray energy, and $\theta$ the scattering angle. $\theta$ is measured from the direction of the incident x-ray photon and in most radioisotope XRF systems ranges from 90$^o$ – 150$^o$. The intensity of scattered radiation $I_s$ from an infinitely thick sample is given by;

$$I_s = \frac{GI_o(E_o)T_s\varepsilon_s \sum(\mu_{si}(\theta)W_i)}{\sum[(\mu_i + \mu_{si})W_i]} \tag{3.11}$$

where; $GI_o, (E_o), W_i$-same as in Equation (3.9); $T_s$-transmission of the scattered x-rays through the filter and the detector window;$\varepsilon_s$- efficiency of the detector for the scattered x-rays; $\mu_{si}(\theta)$-differential scattering cross section for the x-rays scattered by the ith element toward the detector; and $\mu_{si}$- mass absorption coefficient of the scattered x-rays for the ith element of the sample.

Compton scattering is an important interaction in XRF as discussed in sections above. The concentration of an analyte *W<sub>i</sub>* in a sample is determined by its fluorescent intensity *I<sub>i</sub>* in combination with the fluorescent x-rays of the matrix elements and Compton-scattered x-rays $I_{com}$:

$$W_i \cong k \frac{I_i}{I_{com}} \tag{3.12}$$

where $k$ is a constant. $I_s$ is replaced by $I_{com}$ because Equation (3.11) holds for both coherent and Compton scattered x-rays when the appropriate cross section is used. Solving Equation (3.12) directly for the concentration of the analyte is difficult because the intensity of its characteristic line is not only a function of the analyte but of the other elements present in the sample. By adopting simple approximation such as proposed by Lucas-Tooth and Price, this problem can be solved (van Grieken and Markowicz, 2002). The approximation states that, “because the intensities of x-rays of elements are functions of their respective concentrations, one can substitute their measured x-ray intensities for concentrations of matrix elements” (Lucas-Tooth and Price, 1961). With calibration samples, the L-TP model can be used to calibrate benchtop and portable x-ray analyzers because of its simplicity and ruggedness (van Grieken and Markowicz, 2002).

### 3.3.2 Matrix Correction using Scattered X-rays

According to Andermann and Kemp (1958), the equation for scattered intensity per atom is given by:

$$S_a = I_e F^2 \tag{3.13}$$

Where $S_a$ is the coherent scattering intensity per atom, $I_e$ the electronic scattering intensity, and F the atomic structure factor, and;

$$S_c = I_e R(1 - \sum f_n^2) \tag{3.14}$$

Where $S_c$, is the incoherent scattering intensity per atom, $R$ is the recoil factor, and $\sum f_n^2$ is the incoherent scattering function. However, considering an assemblage of atoms;

$$I_0 \ \alpha \ \frac{S_a}{2\mu_0} + \frac{S_c}{\mu_0 + \mu_c} \tag{3.15}$$

Where $I_0$ is the scattered intensity, $\mu_0$ the absorption coefficient at the observed wave length, and $\mu_c$ the absorption coefficient at the Compton wavelength. The fluorescent intensity is then given by;

$$I_f \ \alpha \ \frac{t}{\mu_i + \mu_o} \tag{3.16}$$

Where, $I_f$ is the intensity of fluorescence, $t$ the emission coefficient, $\mu_i$ the absorption coefficient for the incident radiation, and $\mu_0$ is the absorption coefficient for the observed radiation. The relationship of the S's and $\mu$'s to the atomic number (Z) is given by;

$$\begin{gathered} S_a + S_c \ \tilde{\alpha} \ Z^{(1 \text{ to } 2)} \\ \mu \ \tilde{\alpha} \ Z^4 \end{gathered} \tag{3.17}$$

While the coherent and incoherent scattering intensities are given by;

$$\begin{gathered} I_0 \ \tilde{\alpha} \ Z^{-(3 \text{ to } 2)} \\ I_f \ \tilde{\alpha} \ Z^{-4} \end{gathered} \tag{3.18}$$

The ratio of coherent to incoherent scatter is then given by;

$$\frac{I_f}{I_o} \ \tilde{\alpha} \ Z^{-(1 \ to \ 2)} \tag{3.19}$$

The ratio equation of coherent to incoherent scatter has evolved since the 1950's and has been demonstrated experimentally.

## 3.4 Radiation Shielding

Measurement of radiation is affected by many factors, one being external radiation from the environment of a detector collectively called background radiation. Background radiation originate from other radiation emitting devices, terrestrial sources, cosmic sources, natural activity of the detector materials, and the air around the detector. These forms of radiation have to be reduced or eliminated to effectively perform XRF spectroscopy. Common materials used for shielding include Lead (Pb), steel, Tungsten, and concrete (Knoll, 2011). These materials are categorized as passive shields. Shielding is also considered to minimize the chemical and biological effects of radiation.

Shielding is one component in the three simple concepts that should be followed in radiation protection. Two other concepts are; time by minimizing the time spent using radiation sources, and distance by maximizing the distance between operator and source of radiation. These two concepts are achievable through practice. Shielding is the ultimate defense against effects of radiation. In gamma and x-ray shielding, considerations are made on the cost and benefits of protection (IAEA, 2018).

The intensity of radiation that is attenuated by a material is given by:

$$I = BI_o e^{-\mu x} \tag{3.20}$$

Where $I$ is the intensity after attenuation, $B$ is the buildup factor ($B \geq 1$), $I_o$ is the original intensity, $\mu$ is the linear attenuation coefficient, and $x$ is the thickness of the shielding material. The buildup factor is determined through experimentation or calculation. Figure 3.4 shows the buildup factors for lead (James, 2007).

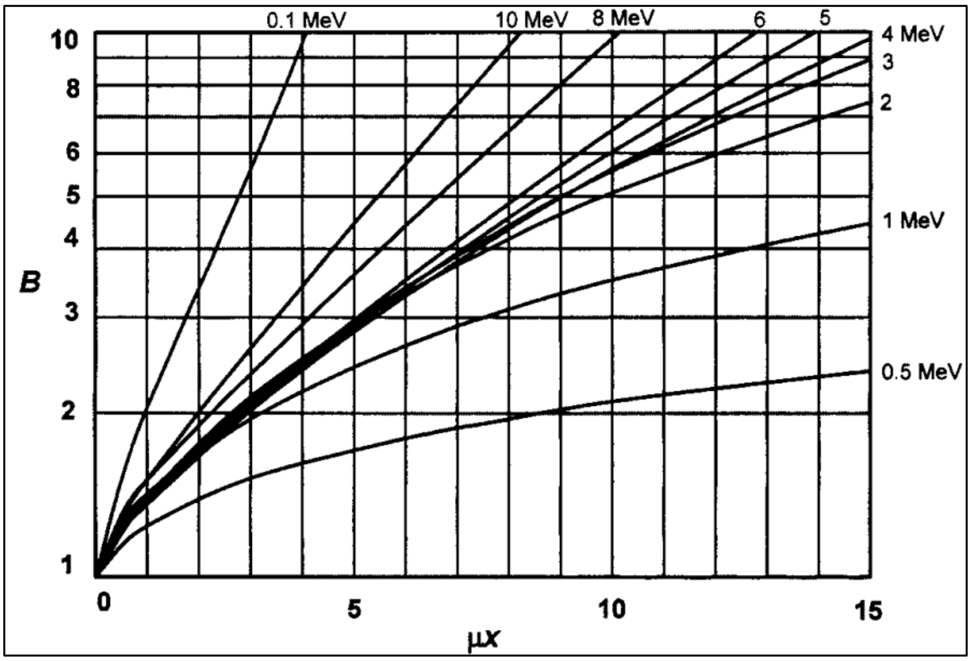


Figure 3.4: Buildup factors of lead using monoenergetic point sources.

Similar curves can be determined for other materials like steel, tungsten, and concrete.

Radiation effects are quantified by their energy, activity, and distance. For point sources of gamma rays and x-rays, the exposure rate is calculated using the formula:

$$\dot{X} = \frac{0.5CE}{r^2}\ R.h^{-1} \tag{3.21}$$

Where C is the activity of source in curie, E is the energy of the radiation, and r is the distance from the source. $R.h^{-1}$ is Roentgen per hour, the unit of measurement (James, 2007).

To protect against the effects of radiation, limits of exposure have been set. These limits are set in a way that over the lifetime of an individual, the risk of death due to exposure to radiation does not exceed the chance of death of any occupation. The International Commission on Radiological Protection (ICRP) and the National Council on Radiation Protection and Measurements (NCRP) have recommended set of limits of exposure over a duration of time. Table 3.2 shows the exposure limits according to NCRP (report 116) and ICRP (publication 60) (ICRP, 1991; NCRP, 1993).

Table 3.2: Exposure limits according to NCRP and ICRP.

| | | **NCRP-116** | **ICRP-60** |
|---|---|---|---|
| Occupational Exposure | | **Effective Dose** | |
| | Annual | 50 mSv | 50 mSv |
| | Cumulative | 10 mSv x age (y) | 100 mSv in 5y |
| | | **Equivalent Dose** | |
| | Annual | 150 mSv lens of eye;<br>500 mSv skin, hands, feet | 150 mSv lens of eye;<br>500 mSv skin, hands, feet |
| Public Exposure | | **Effective Dose** | |
| | Annual | 1 mSv if continuous<br>5 mSv if infrequent | 1 mSv; higher if needed provided 5-year annual average ≤ 1 mSv |
| | | **Equivalent Dose** | |
| | Annual | 15 mSv lens of eye;<br>50 mSv skin, hands, feet | 15 mSv lens of eye;<br>50 mSv skin, hands, feet |

## 3.5 Machine Learning Approaches

Recent advances in computation power and data collection have revolutionized statistics. This revolution has brought about the entirely new field of machine learning, a technique that applies computer algorithms on data to perform predictive analytics. Algorithms are shown data which enable their optimization in a process called training, they are then tested in various ways. The algorithms having been trained can predict new datasets. Machine learning is a broad field and it can be categorized into supervised, unsupervised, and reinforcement learning algorithms (Christopher, 2006; Ethem, 2010).

### 3.5.1 Supervised Learning

Supervised learning uses labeled data for classification or regression. Labeled data can be categorical or continuous variables. Categorically labelled datasets are used in classification while continuous label datasets are used in regression. (Stuart and Peter, 2010). Given a set of N datapoints represented as;

$$(x_i, y_i) \dots (x_N, y_N) \tag{3.22}$$

A supervised learning function can be represented as;

$$g: X \to Y \tag{3.23}$$

Where $X$ is the input variable space while $Y$ is the output space, $g$ is any function that can map the variable space to the output space. Many problems in XRF are either categorical or continuous.

Example of XRF classification is in matrix classification where algorithms are presented with various categories of matrices analyzed in XRF. The classification algorithms are then trained on the categorically labelled XRF spectra and can then be used to predict the matrix category of new XRF spectra. Example of XRF regression is in element quantification in which algorithms are presented with continuously varying element concentrations spectra analyzed in XRF. The regression algorithms are then trained on the continuously labeled XRF spectra and can be used to predict the concentration of the element in new XRF spectra.

The following algorithms are widely used in supervised learning; Support-vector machines, Linear regression, Logistic regression, Naive Bayes, Linear discriminant analysis, Decision trees, K-nearest neighbor algorithm, Neural networks, Random Forests, and many others.

### 3.5.2 Unsupervised Learning

Unsupervised learning, unlike supervised learning uses unlabeled data. The unsupervised learning algorithms used attempts to discover previously unknown patterns, groupings, and information in the data (Stuart and Peter, 2010).

Example of unsupervised learning in XRF is the analysis of ancient exchange relationships in materials such as ornaments and carvings. XRF spectra of such materials are collected and algorithms are used to discover clusters of related material that signify similar origins. Other famed use of unsupervised learning in XRF is in dimension reduction of the normally highly multivariate and correlated XRF data.

The following algorithms are widely used in unsupervised learning; Hierarchical clustering, k-means, Mixture models, Local Outlier Factor, Isolation Forest, Principal component analysis, Independent component analysis, Non-negative matrix factorization, Singular value decomposition, among others.

### 3.5.3 Semi-supervised Learning

Many real-world datasets are a combination of labeled and unlabeled data. Labeled data collection is resource intensive because it requires human input. Unlabeled data is vastly available. Combining the two is based on the limited amount of labeled data and the amount of useful information contained in the unlabeled data (Cabannes *et al.*, 2021). Given a set of N labeled datapoints represented as;

$$(x_i, y_i) \dots (x_N, y_N) \tag{3.24}$$

And a set of M unlabeled data points given as;

$$(x_{N+1}) \dots (x_{N+M}) \tag{3.25}$$

A semi-supervised learning function attempts to map the input variables to the output for the unlabeled data.

In XRF, semi-supervised learning can be used to improve the accuracy of training labeled data. Example is the availability of vast amounts of unlabeled XRF data and small amount of labeled data. The first step in semi-supervised learning involves training models on labeled data then using the model to predict pseudo-labels for the unlabeled data. The final steps involve merging the labeled and pseudo-labeled datasets and retraining the model to achieve low-error rates. Examples of algorithms used in semi-supervised learning are therefore similar to the above mentioned.

### 3.5.4 Reinforcement Learning

Reinforcement Learning (RL) pertains an agents' action given a set of rules that tries to maximize the chance of success. RL algorithms are tuned to explore the set-out rules (environment) and for every action it takes a reward mechanism is attached. The previous step and the reward act as input in deciding the best action to take in the agents' subsequent steps. RL is not often encountered in XRF, it is however one of the key pillars in machine learning.

A reinforcement learning process involves: a set of environment and agent states, $S$; a set of actions taken by the agent, $A$; a change of state from an initial, $S_t$, to a next state $S_{t+1}$, in an interval $t$ under action $A$; and a reward mechanism between transitions of states, $R$. Figure 3.5 shows the representation of a reinforcement learning problem (Shweta Bhatt, 2018);

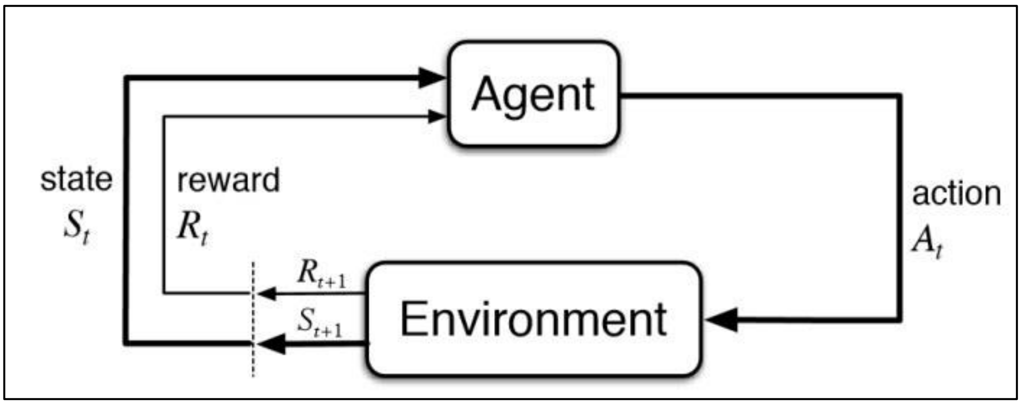


Figure 3.5: Formulation of a basic reinforcement learning problem.

## 3.6 Chemometrics

Chemometrics is the application of mathematical and statistical techniques in analytical science. Data generated by analysis instruments is large and diverse. Multivariate data such as EDXRF spectra contain several thousand variables; this brings complexity to traditional data analysis tools. Multivariate data are used for various tasks, for example in discrimination analysis of soil or classification of soils into groups with similar properties (Brereton, 2003; Kaniu and Angeyo, 2015). The advent of powerful personal computers has pushed data processing to new heights. These processing capabilities include simultaneous analysis of variables from several samples. The aim of chemometrics is to develop calibration models that can be used to predict properties of interest present in a chemical system. Principal component analysis (PCA), Support Vector Regression (SVR), Artificial Neural Networks (ANN), Random Forests (RF) among others are techniques employed in chemometrics for calibration and prediction (Miller and Miller, 2010).

### 3.6.1 Principal Component Analysis

Due to the amount of information in spectroscopic analyses, relationships and patterns are hard to observe in multivariate data. EDXRF spectral data present such datasets. Three problems arise with such data; it is not possible to graphically represent more than three variables, many statistical methods fail when there are high correlations between variables, and many of the variables carry little or information of no use (Okonda, 2015).

PCA is a data reduction technique when there are correlations in data (Filzmoser and Varmuza, 2016). PCA aims at finding principal components (PCs), $PC1$, $PC2$,…, $PCn$ which are linear combinations of the original variables, $X_1$, $X_2$,…, $X_n$, i.e.,

$$\begin{aligned} PC1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1n}X_n \\ PC1 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2n}X_n \\ &. \quad . \quad . \quad . \quad . \quad . \\ PCn &= a_{n1}X_1 + a_{n2}X_2 + \cdots + a_{nn}X_n \end{aligned} \qquad (3.26)$$

Coefficients, $a_{11}$, $a_{12}$,…, $a_{nn}$ are chosen such that the new variables are not correlated with each other unlike the original variables. Creating *n* new variables for the original *n* variables seems like a pointless task. It is however important to note that the first principal component $PC1$ accounts for most of the variance in the data, the second principal component $PC2$ accounts for the second greatest variation in the data and so on. Therefore, the number of useful principal components to use is determined when a significant correlation between them is established. This property of PCA enables the reduction of variables (than the original) for analysis (Myatt and Johnson, 2008).

It is interesting to note that the PCs are orthogonal to each other. PCs are determined from the covariance matrix which is a measure of joint variance between two variables. The eigenvectors of the covariance matrix are the principal components. For each eigenvector, an eigenvalue is associated with it; this is the amount of variance in the dataset explained by the principal component (Miller and Miller, 2010).

For a typical spectral dataset, it can be represented as a matrix;

$$X = C.S + E \tag{3.27}$$

where $\boldsymbol{X}$ is the original data matrix, $\boldsymbol{C}$ the profile of a variable, $\boldsymbol{S}$ is the matrix of each spectra and $\boldsymbol{E}$ is an error matrix. Graphically;

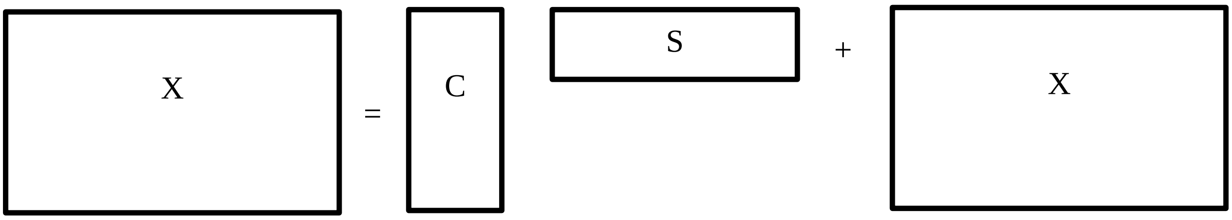


Figure 3.6: Multivariate Data such as occurs in EDXRF spectroscopy

The above equation can be re written as

$$\boldsymbol{X} \approx \widehat{\boldsymbol{C}}.\widehat{\boldsymbol{S}} = \boldsymbol{T}.\boldsymbol{P} \tag{3.28}$$

where $\widehat{\boldsymbol{C}}$ and $\widehat{\boldsymbol{S}}$ are preferred notations in chemometrics. $\boldsymbol{T}$ are the scores (having the same dimensions as $\boldsymbol{C}$) and $\boldsymbol{P}$ are the loadings (having the same dimensions as $\boldsymbol{S}$) of the matrix (Brereton, 2003).

### 3.6.2 Support Vector Machines (SVM)

Support vector machines (SVM) is a supervised learning technique broadly used in classification, outlier detection, and regression. SVM as a classification tool on a dataset, creates a separation hyperplane in which dissimilar data are separated by a plane that maximizes the distance (margin) between the closest points.

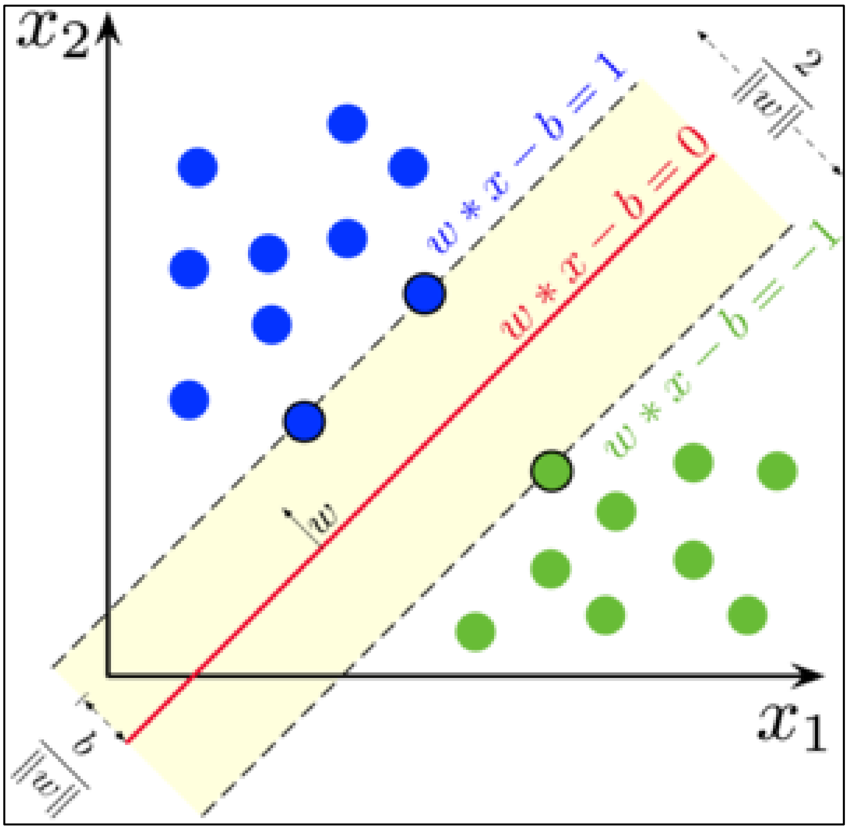


Figure 3.7: SVM illustration in 2 dimensions.

Points that lie by the margin are called the support vectors (SVs) and the central axis is the optimal separation plane (Savan, 2017). SVM as a regression tool (abbreviated SVR), unlike classification, creates a plane in which distance between datapoints are minimized (Meyer *et al.*, 2014).

Given a set of N datapoints represented as $(x_i, y_i) \dots (x_N, y_N)$, where X is the input variable space while Y is a categorical output space (say 1 and -1). SVM in classification attempts to find a hyperplane that maximizes the margin between the input variables.

The hyperplane is defined as;

$$\mathbf{w}^T\mathbf{x} - b = 0 \tag{3.29}$$

Where **w** is a normal vector to the hyperplane as shown in Figure 3.7. **x** is the input variable space, and $b$ is an offset term from origin. The classification can be generalized to regression problems with a few tweaks to the equation.

### 3.6.3 Artificial Neural Networks (ANN)

Natural neurons are the inspiration behind artificial neural networks (ANN). Natural neurons have dendrites that receive signals using synapses from its surroundings. If such signals surpass a certain threshold, the neuron activates and sends a signal through its axon to other synapses and possibly activating other neurons. Figure 3.8 shows a representation of a natural neuron (Source: Miller and Miller, 2010).

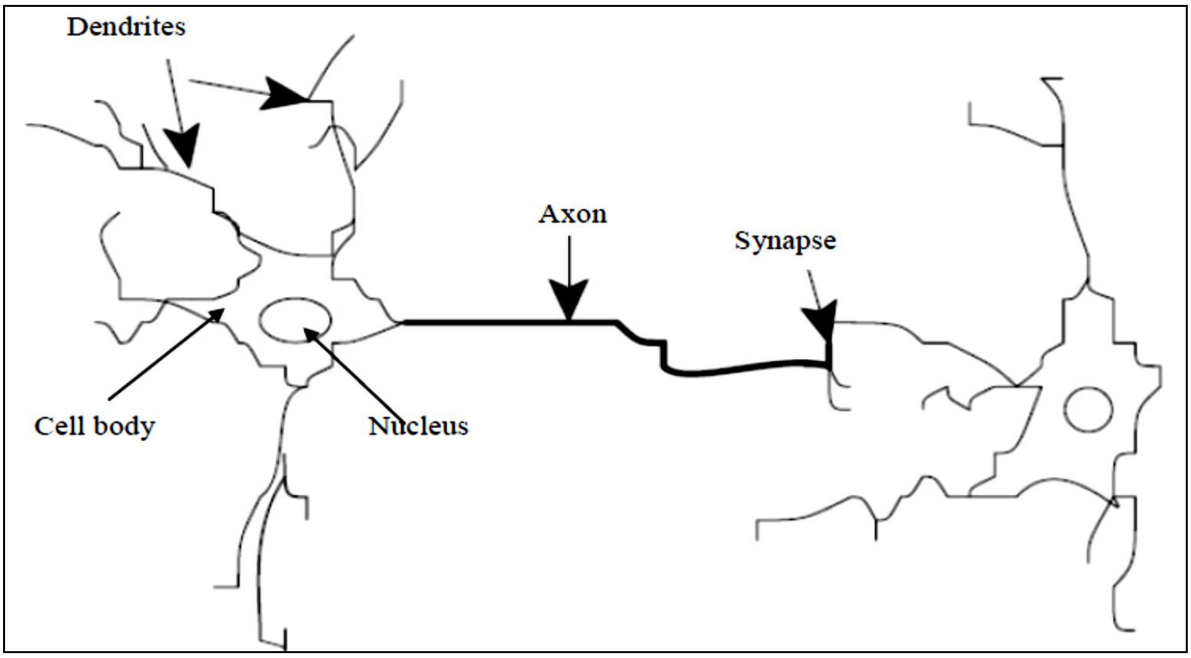


Figure 3.8: Artistic concept of natural neuron.

Modelling artificial neurons is the same idea, they have inputs (the synapses) which are then weighted (threshold of the signal) and a mathematical function is used to compute the activation of the neuron. The output (axon) is then determined by another function. The weights by which the neuron depends on for activation is determined by algorithms, which constantly adjust in a process called training/learning (Andrade-Garda, 2009).

A neural network consists of three layers, an input layer, a hidden layer, and an output layer. In the case of EDXRF analysis, the input is the spectra and the output is the concentrations of elements of interest as represented in Figure 3.9 (Shweta Bhatt, 2018). The backpropagation algorithm, is used in such layered ANNs. This means that the neurons are arranged in layers and send their signal forward and propagates errors backwards. The Backpropagation algorithm basically tries to minimize the error between the input and the expected output and it does this successively using the training data until the ANN has learnt (Miller and Miller, 2010). The artificial neural network algorithm is implemented in R by the neuralnet package (Stefan *et al.*, 2019).

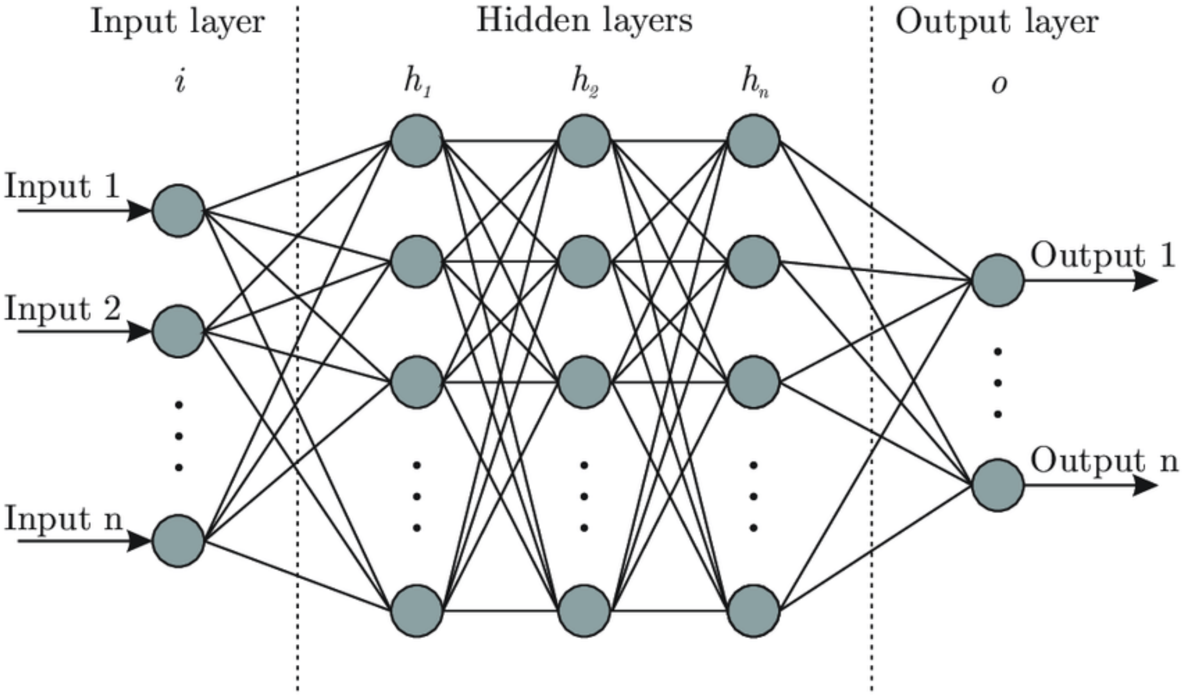


Figure 3.9: Illustration of Neural Network (Source: ResearchGate.net).

For an N input neural network represented as $X_1$, $X_2$,… $X_N$ and an output as $Y$, a neural network algorithm acts in a way to transform the input layer using weights and a biasing to activate a neural network node. This transformation occurs as;

$$\begin{gathered} X_1 \rightarrow X_1 * W_1 \\ X_2 \rightarrow X_2 * W_2 \\ \cdots \\ X_N \rightarrow X_N * W_N \end{gathered} \tag{3.30}$$

The output of the node is then a combination of all the weighted inputs with biasing.

$$Y = f(X_1 * W_1 + X_2 * W_2 + \cdots + X_N * W_N + B) \tag{3.31}$$

### 3.6.4 Random Forests (RF)

Random forests is a supervised learning technique that can be used for both classification and regression. RF has origins in decision trees. According to (Breiman, 2001), a random forest algorithm independently samples random vectors to determine a tree predictor in a combination of tree predictors in which all trees have a similar distribution in a forest. The algorithm converges to a limit as the number of trees in the forest grows large. Decision trees unlike random forests depends on a series of decision steps to reach a specified result.

RF is defined as a classifier that consists of many tree classifiers;

$$h(\mathbf{x}, \Theta_k), k = 1, \dots \tag{3.32}$$

Where $\Theta_k$ are randomly distributed vectors and in which every tree votes for the best class of input $\mathbf{x}$ . In regression mode, the random forest algorithm takes the tree predictor $h(\mathbf{x}, \Theta_k)$ as a numerical value than a class label. The mean over k of the trees then form the predictor of the RF. Figure 3.10 is a representation of an RF algorithm (Shweta Bhatt, 2018).

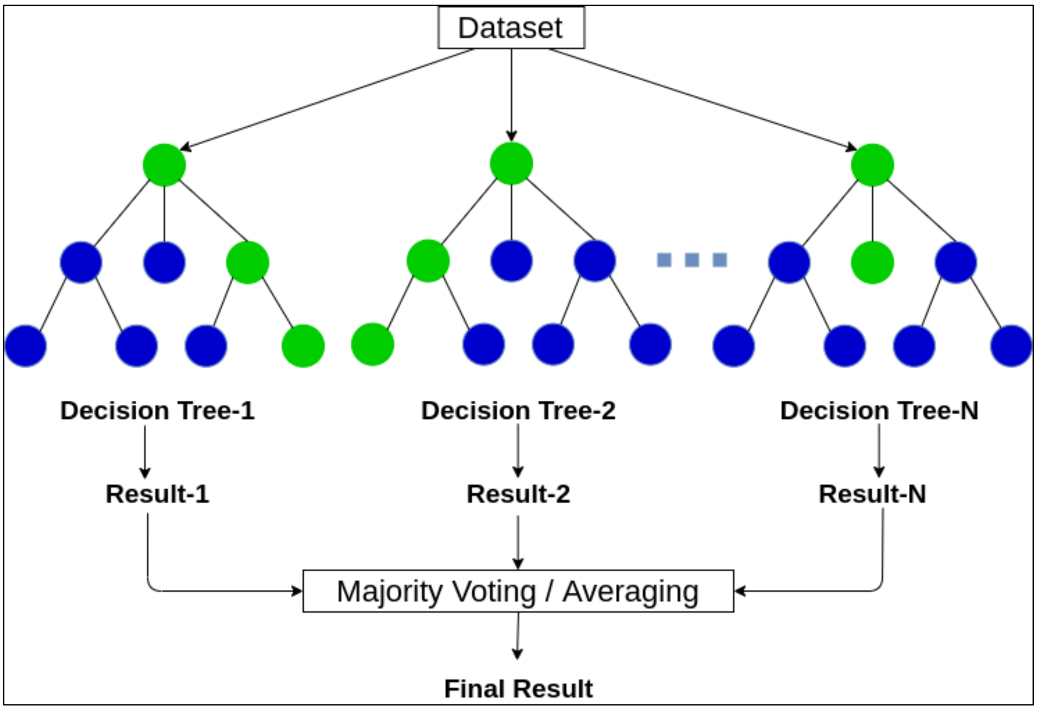


Figure 3.10: Random Forest as collection of Decision Trees.

## 3.7 Metrics in Regression Modelling

This study focused on regression modelling. To evaluate regression models, various techniques are used. The three main metrics are the Mean Square Error/Root Mean Square Error, Coefficient of Determination (R$^2$), and Mean Absolute Error.

### 3.7.1 Mean Square Error (MSE)/Root Mean Square Error (RMSE)

MSE also known as Mean Square Deviation (MSD) is a measure of a models' predictive ability that takes a models' result, compares it with the expected result, squares the error, sums the squared errors, and averages the sum.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{X}_i\right)^2 \tag{3.33}$$

Where $n$ is the number of predictions, $i$ is the observation, $X$is the observed number, and $\hat{X}$ is the predicted number.

RMSE also known as the Root Mean Square Deviation (RMSD) is the square root of the averaged mean square error.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \hat{X}_i\right)^2}{n}} \tag{3.34}$$

Where the symbols take the same meaning (Hyndman and Koehler, 2006).

### 3.7.2 Coefficient of Multiple Determination (R$^2$)

R squared value gives the amount of variance in the dependent variable that can be predicted from independent variables.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{3.35}$$

Where $SS_{res}$ is the residual sum of squares, $SS_{tot}$ is the total sum of squares, $y_i$ are the observed quantities, $f_i$ are the predicted values, and $\bar{y}$ is the mean of the observed quantity (Glantz and Slinker, 2001).

$R^2$ is cautiously applied as a metric in machine learning since it may give a good result which may not be the case (Kvålseth, 1985). The adjusted $R^2$ value is an extension of $R^2$ value that attempts to solve the problem of $R^2$ increasing as the number of variables increase in a model (Theil, 1961).

### 3.7.3 Mean Absolute Error (MAE)

MAE is the average error between observed and predicted values (Hyndman and Koehler, 2006). It is calculated as:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{3.36}$$

Where $n$ is the number of observations, $y_i$ is the observed value, and $x_i$ is the value.

# CHAPTER 4
# METHODOLOGY

## 4.1 Chapter Overview

This chapter is divided into three parts. The first covers instrument design, fabrication of associated parts, software setup, and calibration. The second covers simulate sample preparation and spectra acquisition. The third combines exploratory data analysis, data preprocessing, and machine learning. This chapter is arranged in such a way to achieve each specific objective.

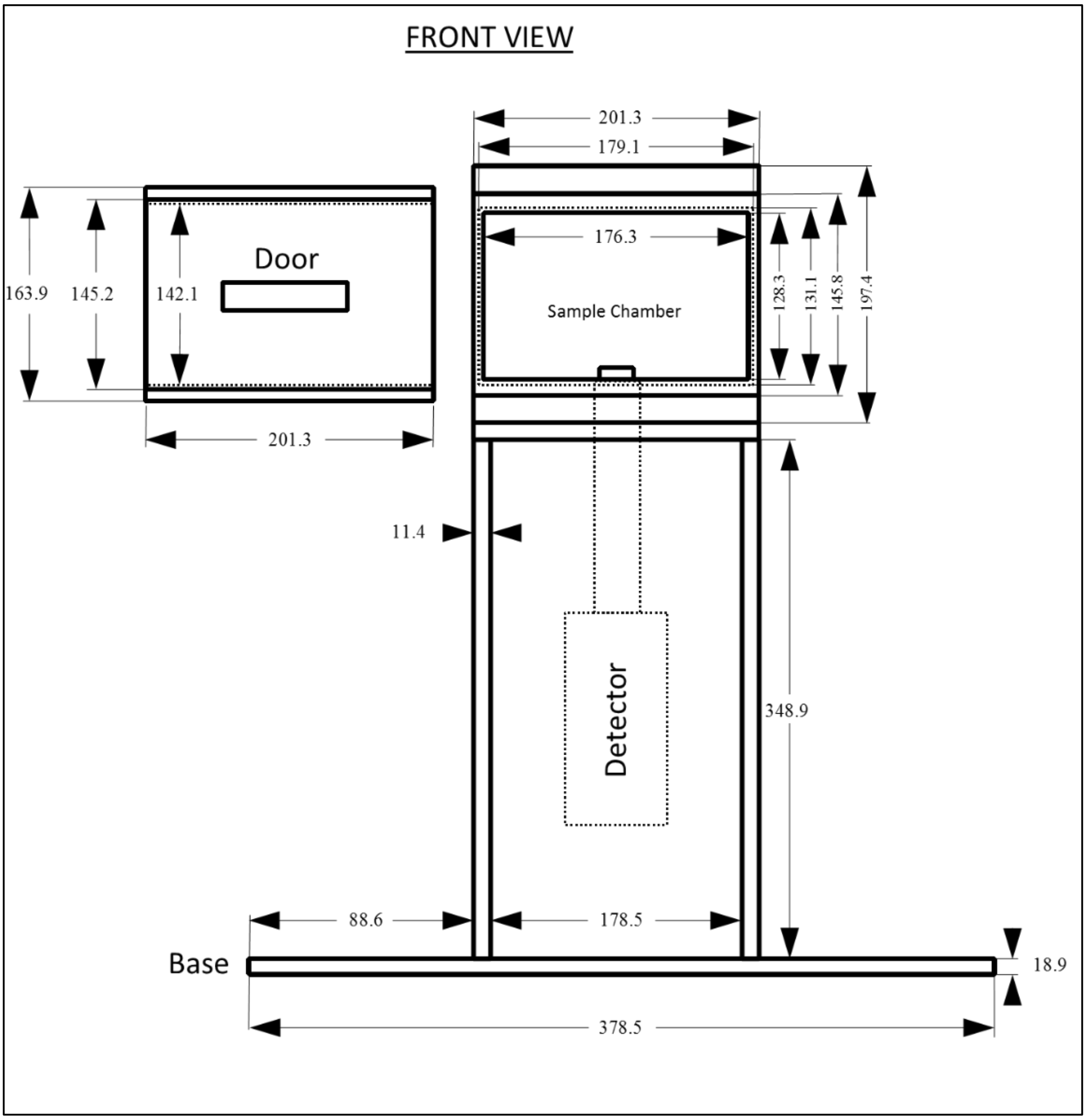
## 4.2 Instrumental Design and Setup

An annular Americium-241 excitation source with an activity of 106 millicurie (mCi) was available at the Institute of Nuclear Science and Technology, University of Nairobi (the activity was adjusted due to decay from the date of manufacture to date of use). A thermoelectrically cooled Silicon Drift Detector (Ketek's AXAS-M SDD, FWHM 130eV; Mn Kα 5.9 keV) was also available (KETEK, 2008). The two parts formed the bases of designing the instrument.

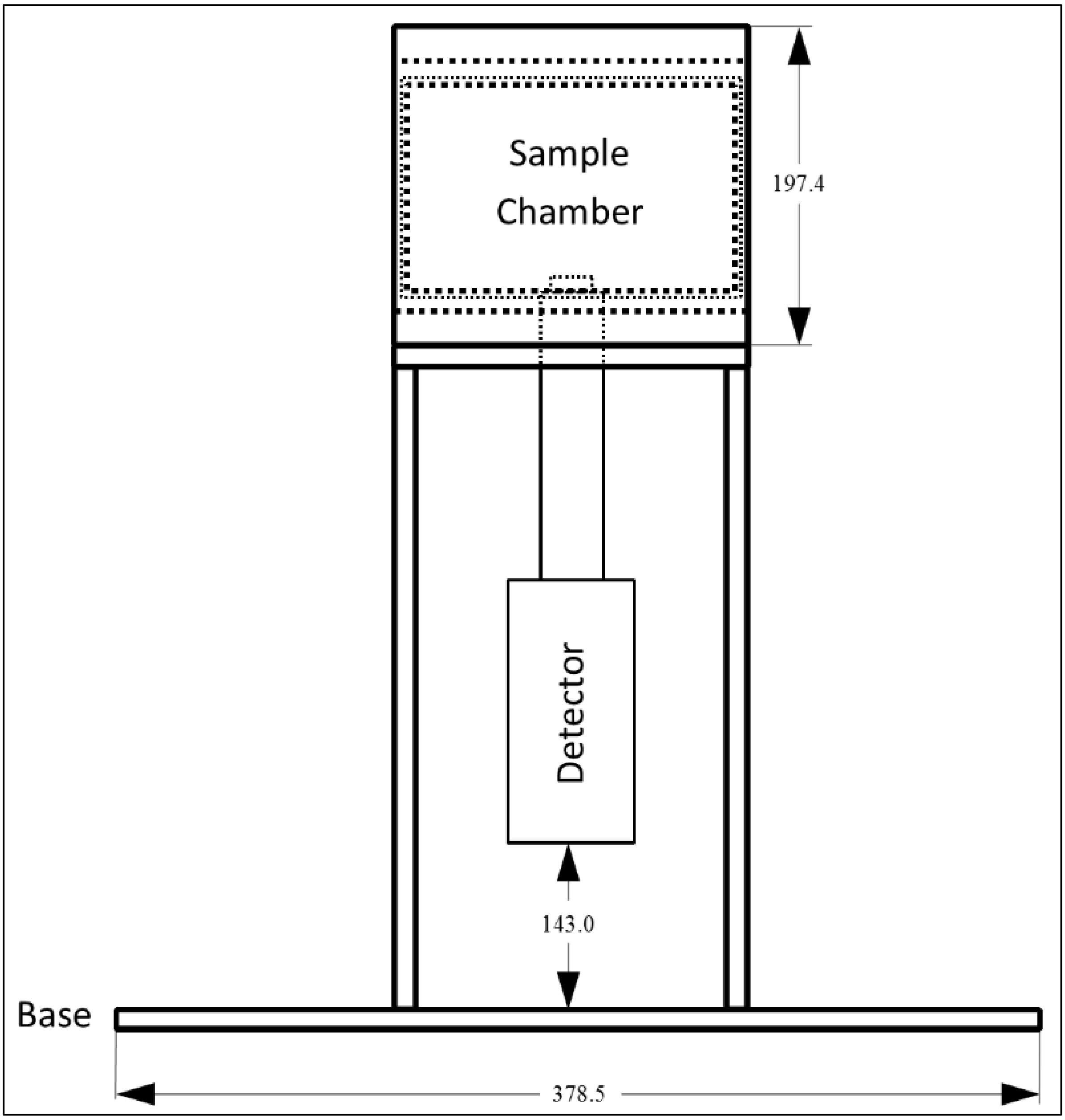
Four additional materials were necessary to couple the base parts. They included; wood, steel, lead, and aluminium. The materials were chosen based on their cost, availability, and workability. Wood provided the core support to the detector and shielding unit. Steel and lead were designed to provide shielding and to house the excitation source, aluminium parts, and sample. Aluminium parts were fabricated to protect the detector, take the weight of excitation source, and to hold the sample during analysis.

### 4.2.1 Design Concept

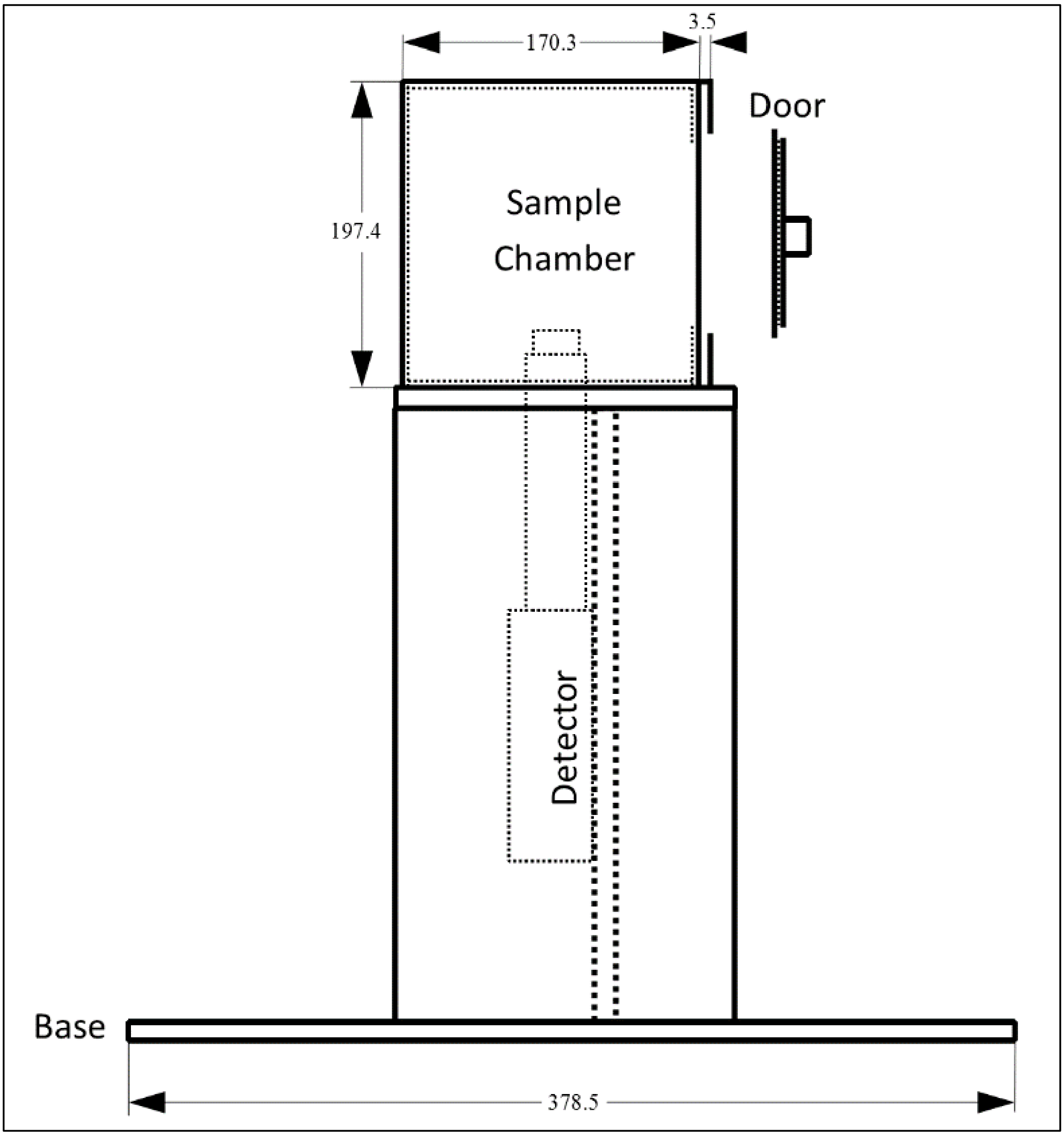
The instrument was designed according to the concept drawings and dimensions in Figure 4.1 (a). Other view angles are shown in Figure 4.1 (b), Figure 4.1 (c), and Figure 4.1 (d).


FRONT VIEW
201.3
179.1
176.3
Sample Chamber
128.3
131.1
145.8
197.4
Door
163.9
145.2
142.1
201.3
11.4
Detector
348.9
88.6
178.5
Base
18.9
378.5

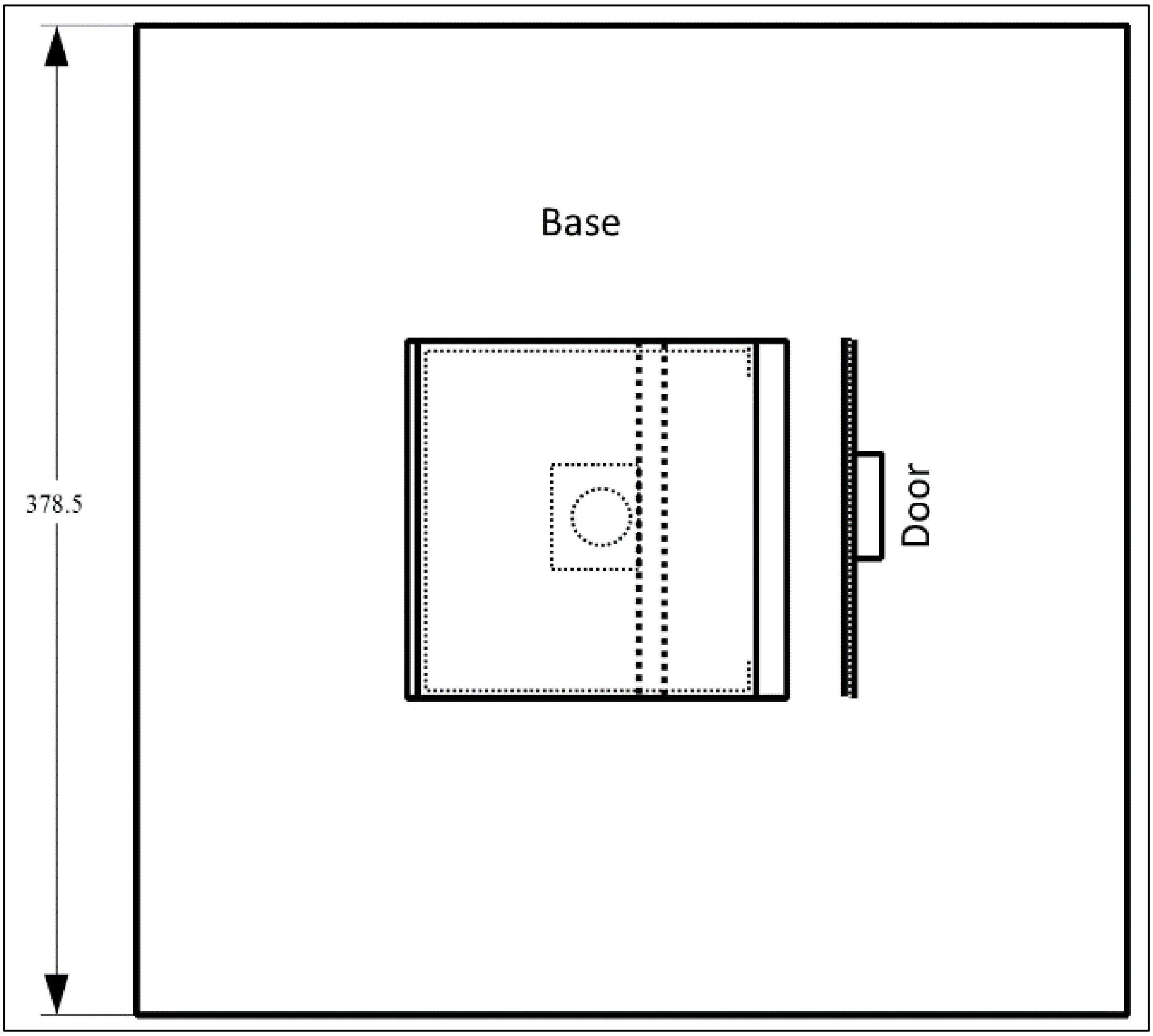
(a)



(b)



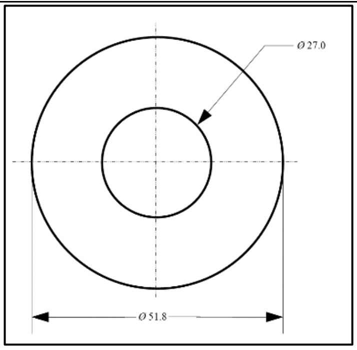
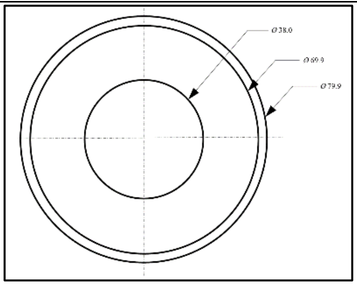
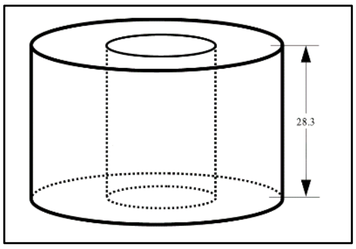
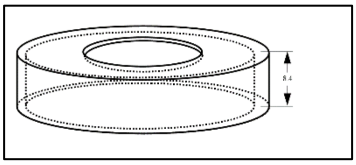
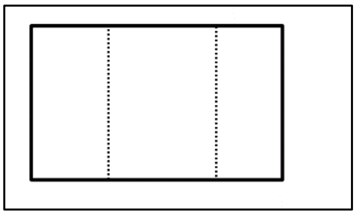
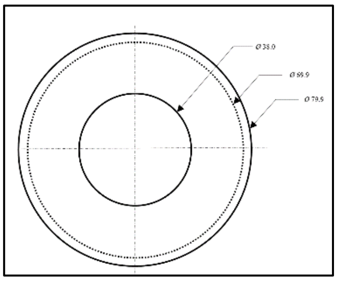
(c)



(d)

Figure 4.1: Instrument design (a) front, (b) rear, (c) side, and (d) top views (in mm).

Table 4.1. Aluminium additional parts (in mm)

| **Detector protector and Radioisotope holder** | **View** | **Sample holder** |
|---|---|---|
|  | Top |  |
|  | Angle |  |
|  | ←Side<br>Rear→ |  |

The shielding made up the radioisotope and sample chamber. Its outer surface was made from 1.5 mm thick steel and the inner surface was lined with 1.5 mm thick lead sheet. The inner upper surface was additionally lined with aluminium to reduce the intensity of backscattered radiation.

### 4.2.2 Fabrication of Parts and Instrument assembly

#### *4.2.2.1 Wood*

Cypress timber and plywood were preferred because they were soft and didn't change shape when dry. Professional carpenters were hired to cut out the required pieces. They followed the drawing design for all the dimensions and thickness. Several devices, nails, and wood glue were used to put the pieces together. The overall aim of the wooden structure was sturdiness and withstanding rough handling.

#### *4.2.2.2 Steel*

Steel sheet, 1.5mm thick was bought from *"Jua Kali"* artisans. The dimensions of the steel needed to change on the overlapping sections. Experienced lab technicians were employed to assess the dimensions and determine the overall cutting technique to use. The dimensions were slightly changed to cater for the overlaps and welding sections. Owing to the thickness of the steel, an electric guillotine shearing machine was used to cut the steel. Hydraulic sheet metal bending machine was used to bend the cut steel sheets. The sections were arc and spot welded together to form the sample housing. The lead lining was laid in during the welding.

#### *4.2.2.3 Lead*

Lead sheets ware obtained from lead blocks. Since lead is easily a malleable metal, a manual metal rolling mill machine was used. Experienced lab technicians set up the block for rolling while taking precautions to avoid lead contamination. Protective gloves were worn and no eating was allowed during lead handling. The block was rolled on successively decreasing thicknesses. The final roll was set at 1.5mm, which was beyond the TVT of the excitation source. The lead sheet was cut using a manual guillotine machine and lined inside the steel housing. Care was taken to avoid smoke from the arc welding process by working in a well-ventilated room. Hands were thoroughly washed after handling lead.

Through testing using a radiation dosimeter, radiation dose rates were measured around the unit at different locations. This testing serves as a quality control measure of the unit.

#### *4.2.2.4 Aluminium*

Detector protector/radioisotope holder and sample holder were made from aluminium. Starting from square blocks of aluminium, the dimensions were determined and marked by a lab technician. Metal lathe machine was used to carve out the shapes.

A hole for the detector finger was drilled. The wooden part and the steel/lead housing were screwed together. The sample chamber door was made of lead sandwiched between steel (giving better shield performance in direction of user). The aluminium parts did not need anchorage since it was anticipated that they'd be removed often.

### 4.2.3 Calibrations: Radioisotope, Sample, and Detector Height

Through testing, higher X-ray intensity was crucial because of the low-activity radioisotope. Two positions were adjustable; the sample holder height above the radioisotope and the detector height below the sample and radioisotope.

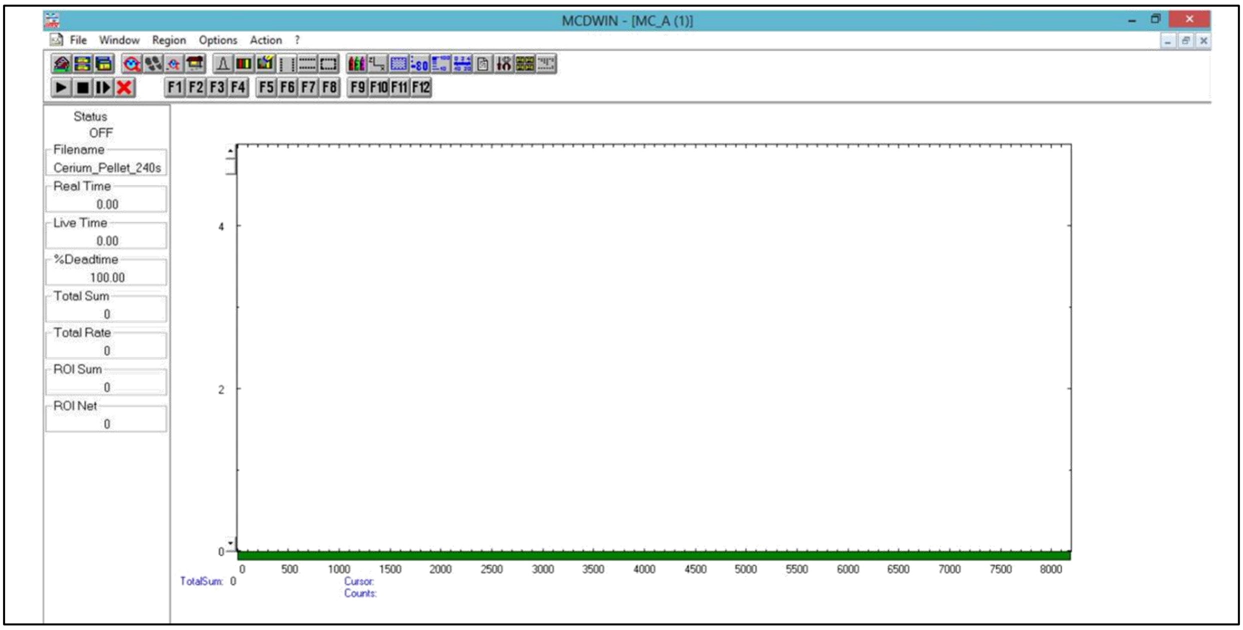
Sample holder height was reached upon by taking spectra at different sample heights above the excitation source at low, mid and high energies. Low energies were that of copper (8.05 keV), mid energies were of Zirconium (15.77 keV) and high energies were of Tin (25.27 keV). The classification of these energies was reached based on the instrument limit of energy detection at 26 keV. The heights were at 1, 2, 3, 4, 5 and 6 mm above the radioisotope source.

Detector height, similar to the task of machining the sample holder, the height at which to bolt the detector for maximum intensity was needed. The height was reached upon by taking spectra by decreasing the distance at three equidistant points (2.1, 1.1, and 0.1 cm) to the nearest position (0.1 cm) of the base of the radioisotope holder.

### 4.2.4 Software Setup and Energy Calibration

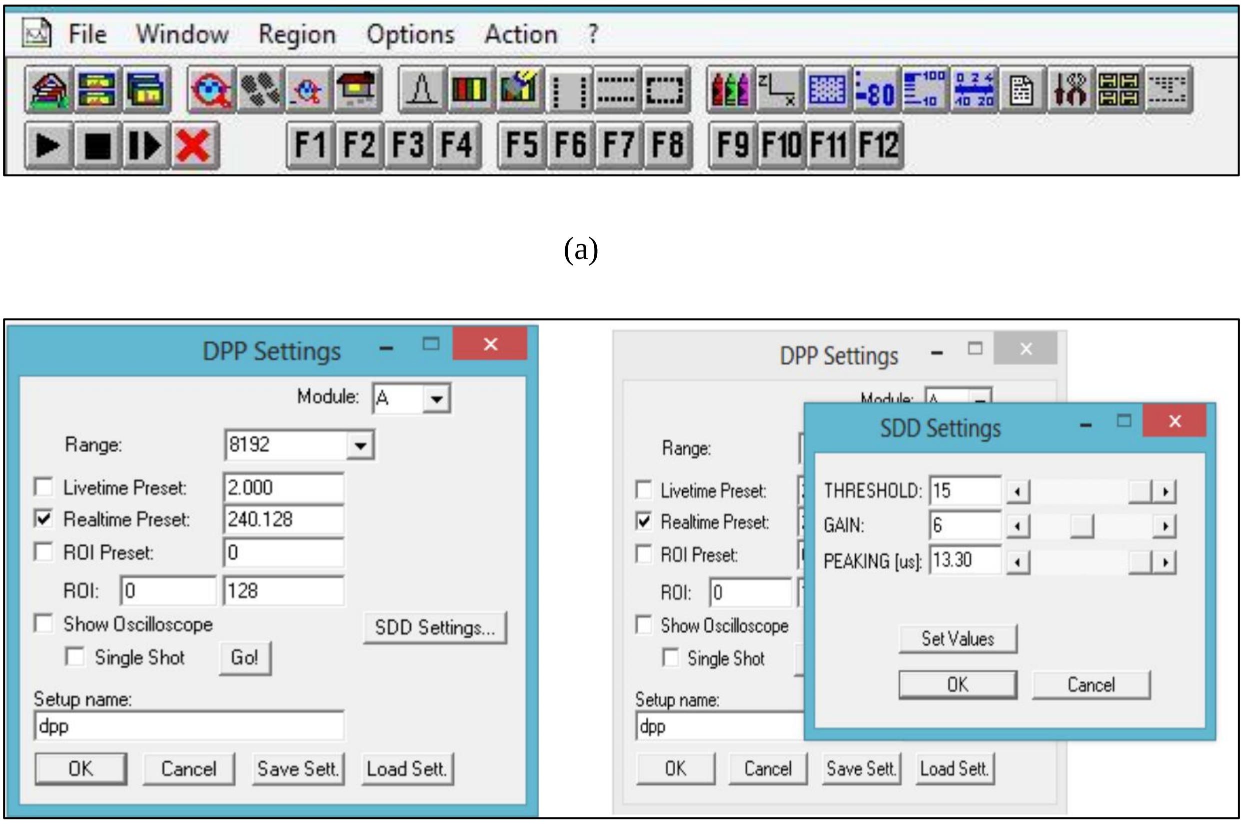
The detector was supplied with a power supply unit, multichannel analyzer and software (MCDWIN software) to run them. A desktop computer was used to install the software for control and spectrum acquisition. Several controls were provided to calibrate the software.

The software allowed for instrument setup of gain, threshold, peaking time, range, presets and more. The cursor can be readily used to understand the function of each button. Figure 4.2 and Figure 4.3 (a, b, and c) show various components of the software user interface.



Status: OFF
Filename: Cerium_Pellet_240s
Real Time: 0.00
Live Time: 0.00
%Deadtime: 100.00
Total Sum: 0
Total Rate: 0
ROI Sum: 0
ROI Net: 0

Figure 4.2: Instrument control interface.



(a)

DPP Settings

Module: A

Range: 8192

□ Livetime Preset: 2.000

☑ Realtime Preset: 240.128

□ ROI Preset: 0

ROI: 0 128

□ Show Oscilloscope

□ Single Shot

Setup name: dpp

SDD Settings

THRESHOLD: 15

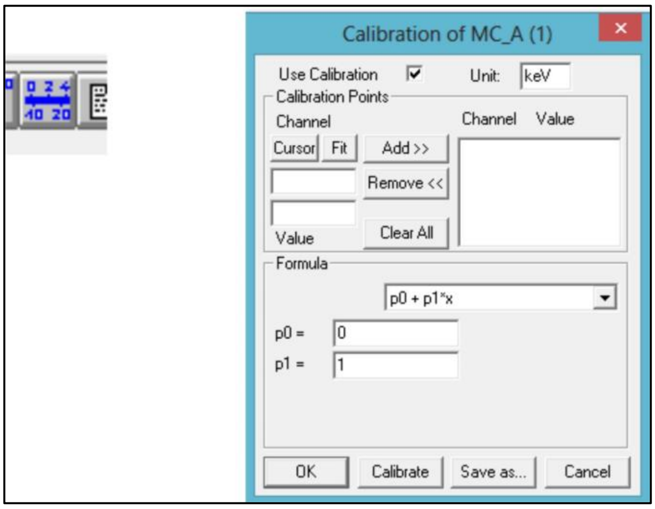
GAIN: 6

PEAKING [us]: 13.30

(b), (c)

Figure 4.3: (a) Instrument control bar, (b) Digital Pulse Processing, and (c) Silicon Drift Detector settings.

Element identification requires energy calibration. This can easily be achieved by clicking on the energy calibration icon and running a sample with a known energy. Two points are required to successfully calibrate the instrument for energy.


Calibration of MC_A (1)

Use Calibration ☑ Unit: keV

Calibration Points

Channel

Value

Channel Value

Formula

p0 + p1*x

p0 = 0

p1 = 1

Figure 4.4: Energy calibration settings.

The instrument was calibrated for energy using copper and tin with purity of 99.99%. Copper has a prominent Kα peak at 8.05 keV while Tin has one at 25.27 keV. This was sufficient to cater for the energy range of the detector.

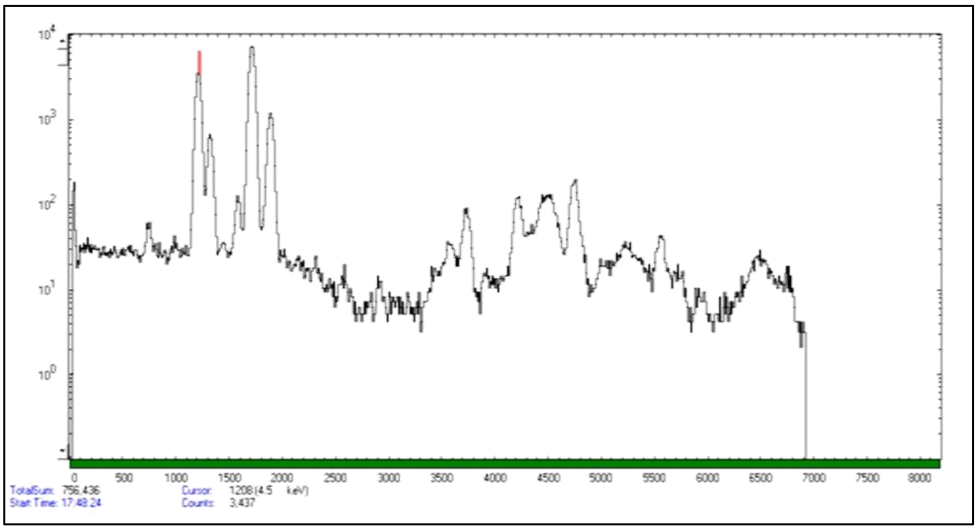


Figure 4.5: Logarithmic spectrum of a rock sample.

Figure 4.5 shows logarithmic spectrum of a rock sample acquired after energy calibration clearly showing elevated levels of Iron (6.4 and 7.0 keV) and Titanium (4.5 and 4.9 keV) among other elements.

The spectrum can be saved with its metadata readily by clicking on the *'Save As…'* button in the *'File'* menu. The spectrum can be retrieved similarly on the *'File'* menu by clicking *'Load…'* button. The spectrum is saved with a *'.asc'* file extension while the spectrum metadata is saved separately as *'.mcd'* file. The *'.mcd'* is used to read the *'.asc'* during loading of the spectrum to the software interface.

## 4.3 Predictive Modelling

### 4.3.1 Sample Preparation

Three REEs (Dysprosium, Yttrium, and Cerium) were used and two other elements Titanium (Kα1-4.51, Kβ1-4.93, and Lα1-0.45 keV) and Niobium (Kα1-16.62, Kβ1-18.62, Lα1-2.17, and Lβ1-2.26 keV) were included since their x-ray emission lines closely border those of REEs. Standard reference material (SRM) salts of Dysprosium, Yttrium, Cerium, Titanium, and Niobium were first acquired and powdered using pestle and mortar. Finely powdered starch and a rock SRM were available for use as matrices.

Powdered starch matrix and rock matrix (plus 40% starch binder by mass in rock matrix) were weighed (approximately 5g) into 30 glass mixing cylinders and labelled. The salts (cerium, dysprosium, yttrium, titanium, and niobium) were then weighed into glass cylinders to achieve different concentrations of the REEs. The concentration (in ppm) of an element in the matrix was determined using the formula:

$$C = \frac{(m \times \%_m)}{M \times 10^{-4}} \qquad (4.1)$$

Where; $m$ is the mass to be measured of SRM (salt containing rare earth element), $\%_m$ is the percentage of rare earth in the salt, $M$ is the mass of the matrix without rare earth element.

A summary of the concentration of REEs in the starch and rock matrices is shown Table 4.2 and Table 4.3. Ethanol of 99.99% purity was then added to the glass cylinders and a vibration-based mixer was used to thoroughly churn the mixture. The contents of the glass cylinder were then emptied to petri dishes and evaporated in an oven at 40℃ for 8 hours.

The dry samples were scrapped and powdered using pestle and mortar and transferred to glass cylinders. Each sample was then weighed and made into a pellet using an 8 ton/cm2 hydraulic press. Each sample was weighed so as to produce about 4 pellets at different masses. *'Blank matrices'* of starch and rock SRM at various masses were also pelletized. Each pellet was analyzed for 240 seconds except those that contained titanium and niobium that were acquired for 200 seconds since they were performed separately. The spectra with its metadata were saved to a local directory.

The spectra (with metadata), concentrations, and masses were then merged to form a table in the R Programming Language. Each row of the table contained the spectrum, metadata, name of sample, mass, and concentration (in ppm) of REEs in it.

Table 4.2. Rock matrix scheme of concentrations (ppm) and masses (mg)

| Sample ID | Ce | Dy | Y | Nb | Ti | Mass |
|---|---|---|---|---|---|---|
| 6_C | 16 | 9 | 18275 | 0 | 0 | 1226.9 |
| 3_C | 12 | 10 | 18254 | 0 | 0 | 1165.1 |
| 1_E | 365 | 7 | 11843 | 0 | 0 | 1184.7 |
| 11_C | 8 | 33 | 9483 | 0 | 0 | 1142.7 |
| 14_C | 56 | 14 | 9296 | 0 | 0 | 1162.6 |
| 15_C | 99 | 102 | 4858 | 0 | 0 | 1140.8 |
| 4_C | 19 | 54 | 3562 | 0 | 0 | 1195.2 |
| 2_B | 8 | 380 | 2397 | 0 | 0 | 1069.3 |
| 7_C | 60 | 19 | 1577 | 0 | 0 | 1158.9 |
| 10_B | 31 | 186 | 1151 | 0 | 0 | 1012.7 |
| 12_C | 184 | 12 | 761 | 0 | 0 | 1189.1 |
| 5_C | 235 | 361 | 598 | 0 | 0 | 1317.9 |
| 13_C | 31 | 191 | 405 | 0 | 0 | 1340.3 |
| -13 | 0 | 0 | 0 | 0 | 0 | 1140.0 |
| -12 | 0 | 0 | 0 | 0 | 0 | 1069.3 |
| -13 | 0 | 0 | 0 | 0 | 0 | 1140.0 |
| -14 | 0 | 0 | 0 | 0 | 0 | 1270.0 |
| -15 | 0 | 0 | 0 | 0 | 0 | 1370.5 |
| -16 | 0 | 0 | 0 | 0 | 0 | 1356.6 |

Table 4.3. Starch matrix scheme of concentrations (ppm) and masses (mg)

| **Sample ID** | **Ce** | **Dy** | **Y** | **Nb** | **Ti** | **Mass** |
|---|---|---|---|---|---|---|
| 3B | 8 | 7 | 18390 | 0 | 0 | 996.7 |
| 6B | 17 | 10 | 18309 | 0 | 0 | 1017.0 |
| 1C | 378 | 9 | 11939 | 0 | 0 | 912.5 |
| 14B | 45 | 9 | 9559 | 0 | 0 | 1006.4 |
| 8B | 23 | 186 | 9376 | 0 | 0 | 955.3 |
| 11B | 4 | 28 | 9349 | 0 | 0 | 1032.2 |
| 15B | 109 | 96 | 4970 | 0 | 0 | 1073.5 |
| 4C | 14 | 45 | 3688 | 0 | 0 | 1131.0 |
| 2B | 17 | 370 | 2324 | 0 | 0 | 1095.8 |
| 12B | 192 | 10 | 650 | 0 | 0 | 1101.0 |
| 5B | 241 | 396 | 463 | 0 | 0 | 1023.3 |
| 9B | 188 | 21 | 439 | 0 | 0 | 1074.1 |
| 13B | 33 | 214 | 434 | 0 | 0 | 1065.3 |
| -11C | 0 | 0 | 0 | 285 | 55 | 1109.0 |
| -12C | 0 | 0 | 0 | 348 | 111 | 1246.4 |
| -2C | 0 | 0 | 0 | 10 | 146 | 1064.7 |
| -4C | 0 | 0 | 0 | 35 | 349 | 1072.9 |
| -5C | 0 | 0 | 0 | 55 | 185 | 1090.4 |
| -8C | 0 | 0 | 0 | 145 | 236 | 977.4 |

### 4.3.2 Spectra data acquisition

To acquire raw spectral data of a sample (for sizes that fit inside the sample chamber), no much sample preparation is required. However, if it’s impractical to fit samples in the chamber, preparation becomes necessary. Acquiring spectra is useful if secondary analysis is to be performed on other platforms/software. Figure 4.6 shows a simplified procedure of spectrum acquisition.

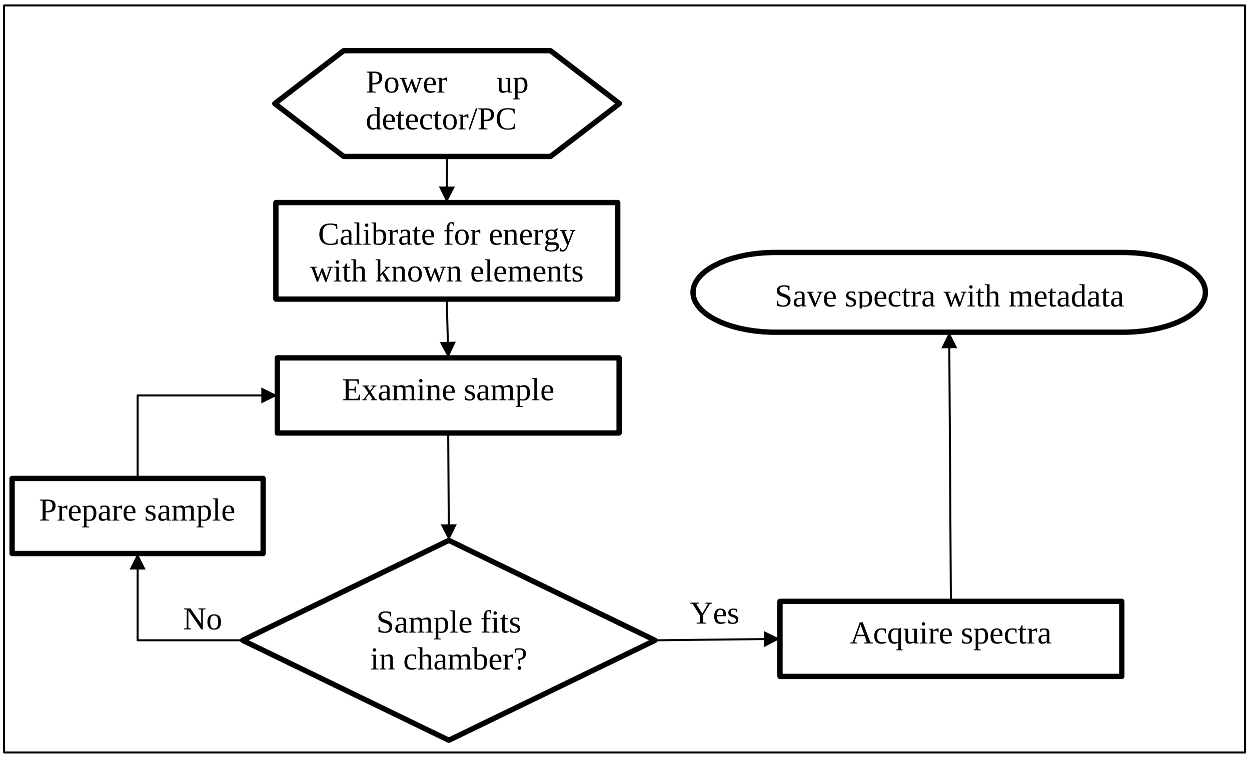


Figure 4.6: Flowchart for acquiring spectrum of a sample.

The spectrum produced required pre-processing i.e., data cleaning. This included extracting spectrum and metadata. Spectrum was collected over 8192 channels which presented a wide feature space. The spectrum was therefore averaged by serially summing 8 channels to form 1024 channels (this can be changed to any number of channels). The metadata contained useful information on the nature of spectrum collected. The *‘.mcd’* file contained the detector settings, names of file, sum of counts, maximum count, detector real time and live time, date and time, length of channels used, and other comments. The detector configuration was maintained by one file throughout the analysis. For reproducibility of the results in this study, this file would be necessary. The annotated code attached in APPENDIX B was used to extract the metadata and the spectrum by passing in the directory of the files. The masses and concentrations file were also processed by the code. The mass and concentration file were in the same folder as the spectrum and metadata files, named *‘scheme.csv’*.

### 4.3.3 Spectral Data Preprocessing

R was used for spectra preprocessing and predictive modelling. The open-source software R and integrated development environment (IDE) for R, RStudio (free and open source software) was downloaded from http://www.rstudio.com. Since R was used for the remaining part of the work, the spectrum and metadata are easily read into R as a dataframe objects using traditional table-like reading functions, e.g., *‘Spectrum <- read.table(“Spectrum.asc”).* The generation of graphs to determine positions of maximum intensity in both sample and detector were all done in R (R-Core-Team, 2021). Among other dependent libraries, many functions were borrowed from the libraries in Table 4.4;

Table 4.4: List of R libraries (other than base packages) used.

| Library | Description |
|---|---|
| e1071 | Functions for implementation of support vector machines |
| neuralnet | Functions for training of neural networks with flexible choice of parameters |
| plyr | Provides tools for splitting, applying and combining data |
| dplyr | Provides tools for easy manipulation of dataframes |
| reshape2 | “melt” and “Dcast” for flexible aggregation and restructuring of data |
| ggplot2 | Provides a system for graphic declaration using “The Grammar of Graphics” |
| caret | Convenient training function for many classification and regression models |
| ChemoSpec | Collection of functions for exploratory data analysis of spectral data |

Since spectral data present a wide feature space, so much computational time is spent on unnecessary or correlated data. Spectral data was first reduced by choosing specific spectral lines corresponding to X-ray emission lines of the elements. The scatter regions were also extracted. Each element spectral line region was then divided by the Compton scatter regions (ratio of signal to scatter). The ratio regions were then subjected to principal component analysis to determine the best scatter regions to use. The best ratio region was then used for modelling in conjunction with concentration data. The code used is attached in APPENDIX C.

## 4.4 Comparison of Predictive Models

Developing chemometric models is quite a laborious process. This process can be simplified by having a method that can work iteratively without much user input. The process used in this study is summarized in the flowchart in Figure 4.7. Models that pass through the process successfully are saved since re-training them is time consuming and computationally intensive. The R code is also readily optimized to cater for spectroscopists without programming background. This process can also be readily optimized to develop models for other elements.

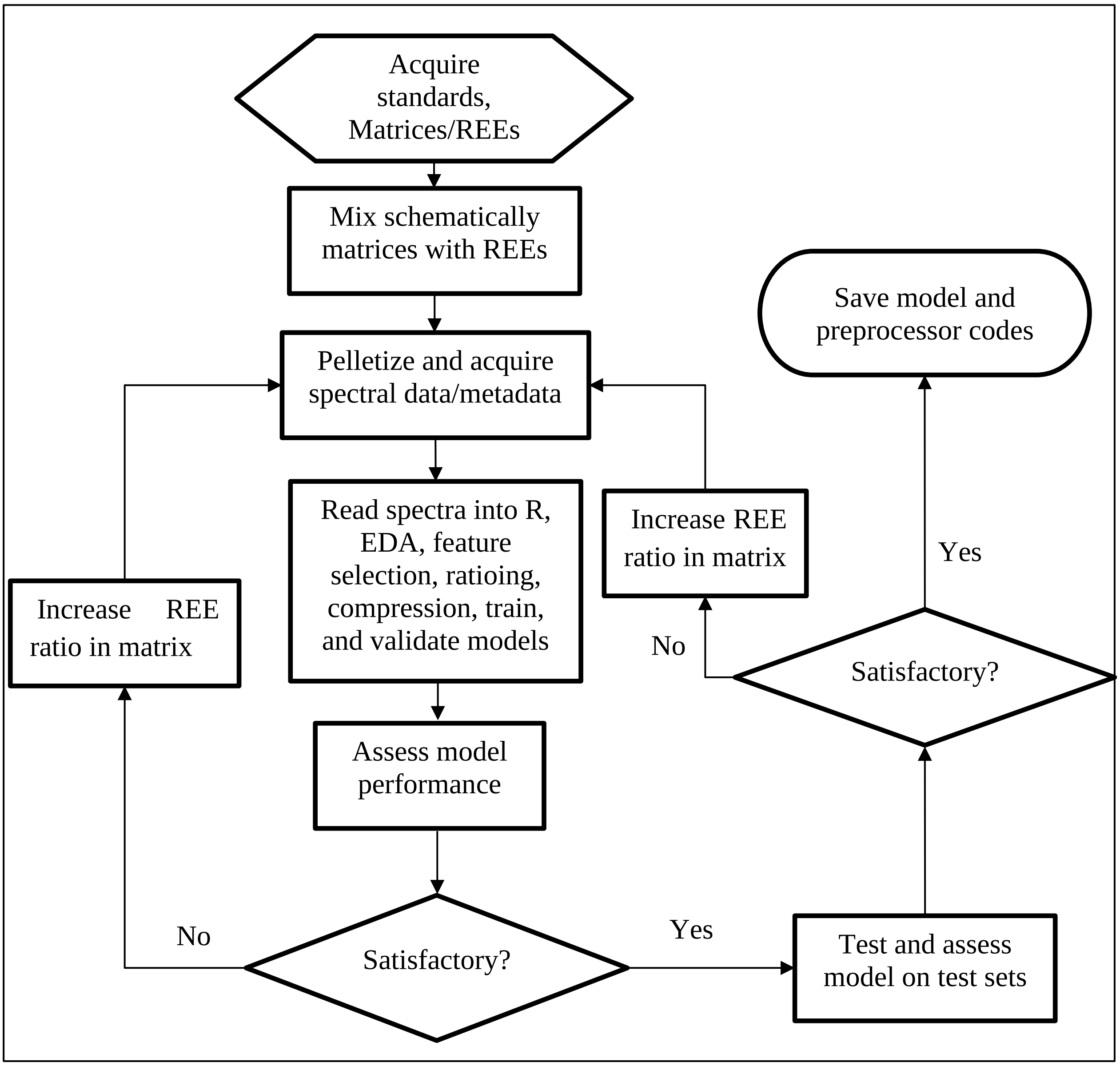

Acquire standards, Matrices/REEs
Mix schematically matrices with REEs
Save model and preprocessor codes
Pelletize and acquire spectral data/metadata
Read spectra into R, EDA, feature selection, ratioing, compression, train, and validate models
Increase REE ratio in matrix
Yes
Increase REE ratio in matrix
No
Satisfactory?
Assess model performance
No
Yes
Satisfactory?
Test and assess model on test sets

Figure 4.7: Flow chart diagram for developing chemometric models.

Prediction model codes for SVR, ANN, and RF were written for Dysprosium, Yttrium, Cerium, Titanium, and Niobium. Three R packages were important; *RandomForest* (for RF models), *e1071* (for SVM models), *neuralnet* (for ANN models), and *caret* (for model training optimization).

The code implemented predictive models and assessed their performance. All elements were similarly treated by sequentially changing the names and symbols in the prediction formula. Hyperparameter tuning was handled automatically by caret's train function.

Production chemometric models take new data and give estimates of concentrations. Saved models need no retraining unless new parameters are presented. The structure of the spectral data/metadata have to be the same as those used in model development. The same preprocessor codes are therefore used. The advantage of using R is the ability to perform the analysis by calling one function only. This is advantageous for spectroscopists without knowledge of R. Figure 4.8 shows the procedure for predicting concentration of elements in new samples.

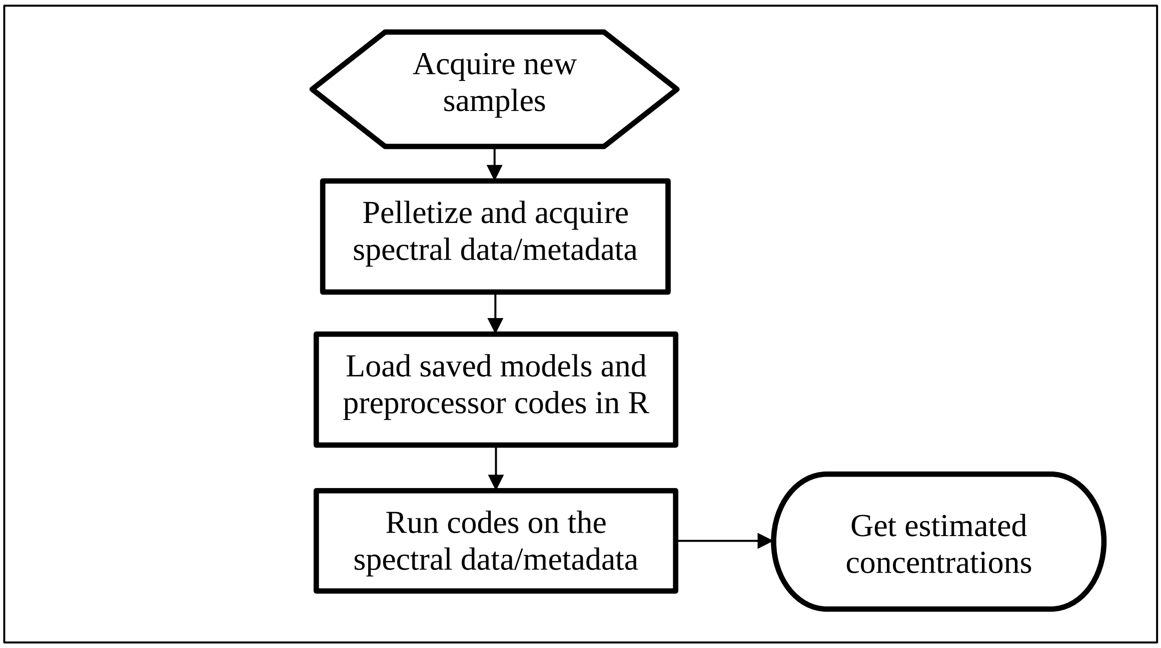

Acquire new samples
Pelletize and acquire spectral data/metadata
Load saved models and preprocessor codes in R
Run codes on the spectral data/metadata
Get estimated concentrations

Figure 4.8: Flow chart diagram of predicting concentrations of elements in unknown samples

# CHAPTER 5
# RESULTS AND DISCUSSION

## 5.1 Overview

This chapter presents the findings of the study in two parts. First part concerns the findings of the instrument design, associated parts, calibrations, and efficiency measurements. The instrument design formed the basis of collecting XRF spectra. Second part presents findings of data pre-processing, XRF spectral pre- and post-processing, and machine learning models applied in both stages.

## 5.2 XRF Instrument Setup

### 5.2.1 XRF Instrument Design and Setup

The unit to hold the detector, shielding, and the radioisotope was successfully built as shown in Figure 5.1. The broad base acts in such a way as to offer a mechanically stable configuration because of the weight of the shielding unit on top. There is also weight contribution from the detector, Americium source and samples to be analyzed, therefore the unit needed to have excellent stability. The height of the detector dictated the height of the unit and subsequently the base area of the unit. The unit without the samples to be analyzed can be safely tilted to an angle of $22^o$ to the horizontal before it tips on all sides.

The shielding made up the radioisotope and sample chamber. Its outer surface was made from 1.5 mm thick steel and the inner surface was lined with 1.5 mm thick Lead sheet. The inner upper surface was additionally lined with aluminium to reduce the intensity of backscattered radiation. The tenth value thickness (TVT) of Lead (Pb) required to reduce the intensity of radiation from the 60 keV gamma of Am-241 is less than or approximately 1 mm at a distance of 1 meter from the source. To shield by steel and achieve the same result requires about 3 mm of steel. Table 5.1 shows the overall performance of the shielding unit.

The shield performance measurements show that there are two main concerns, dose rate of 1.68 mSv/h (near the shield at 10 centimeters while door is open) and 3.7 µSv/h (seated at 1 meter while shield door is open). The doses are received mainly by the hands during sample change (500 mSv yearly limit). Doses are also received by the eyes (20 mSv over

5 years or not exceeding 50 mSv in any single year dose limits) when placing the source in the chamber (IAEA, 2018). Minimizing the time spent in transferring a sample for analysis is very critical (about 5-10 seconds per sample). The angle of approach should be aimed from underneath the source. Doses to the eyes can be minimized by wearing leaded goggles. Dose rates can also be minimized by leaving the room while sample is being analyzed.

Table 5.1. Dose rate measurements for shield performance

| Location | Dose rate (μSv/h) |
|---|---|
| Outside/ Background/ Outdoor | 0.20 |
| Indoors/With shield door closed – 3~4 meters | 0.20 |
| Closer-to-detector/With shield door closed – 10 cm | 0.25 |
| Shield door open/Dose on the shield door – 10 cm | 1680.00 |
| Seated while shield door is closed/head level – 1 meter | 0.20 |
| Seated while shield door is open/head level – 1 meter | 3.70 |
| Radioisotope outside with its cover – 5 cm | 0.60 |

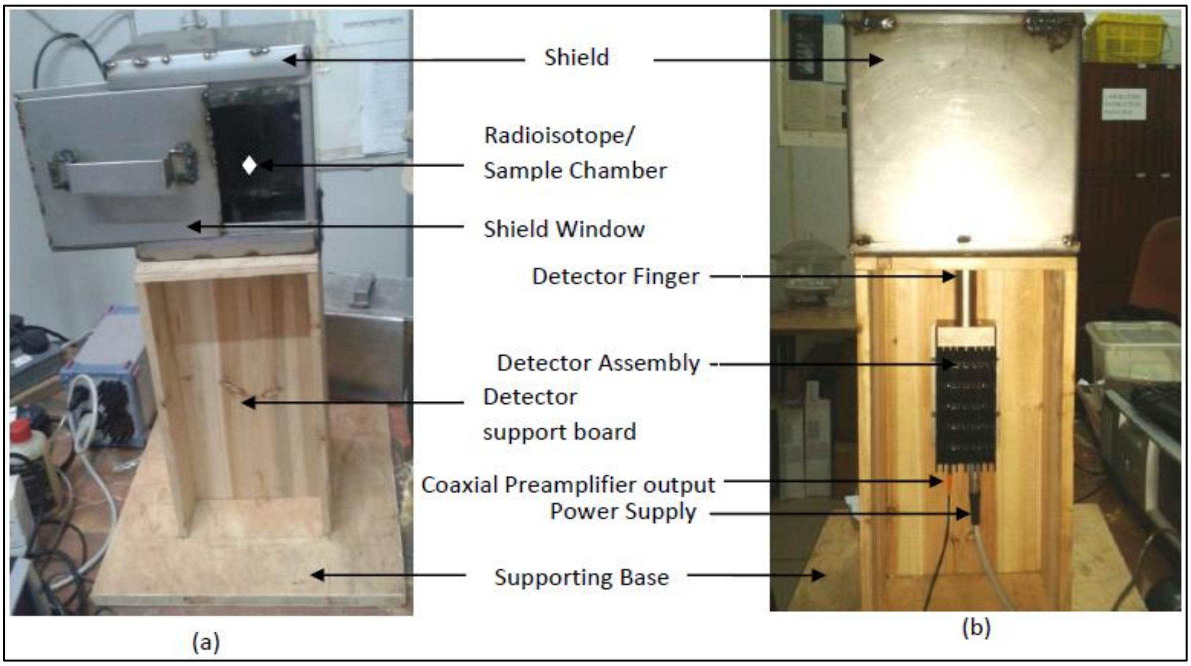


Figure 5.1: Instrument design. (a) front and (b) back view showing all the parts.

### 5.2.2 Fabricated Parts

Two aluminium parts were fabricated, detector protector and the sample holder. The aluminium detector protector also doubled as the radioisotope holder. It was machined to fit to both the detector finger on the inside and the cavity of the radioisotope source on the outside. The sample holder was designed to fit to a secondary sample tray and the radioisotope source.

Detector protector part is shown in Figure 5.2. Detector protector functions as to protect the detector finger from wobbling and to transmit the weight of the load (radioisotope and sample) to the instrument skeleton away from the fragile detector window. It was covered with mylar film to protect the detector from dust and sample fragments that occasionally break off.

Figure 5.2: Part offering protection for the detector finger

Sample holder, the part that holds a pelletized sample, is shown in Figure 5.3. The height of this part was reached upon by experimenting at different sample heights above the radioisotope source at low, mid and high energies. The best position was determined to be at 3mm.

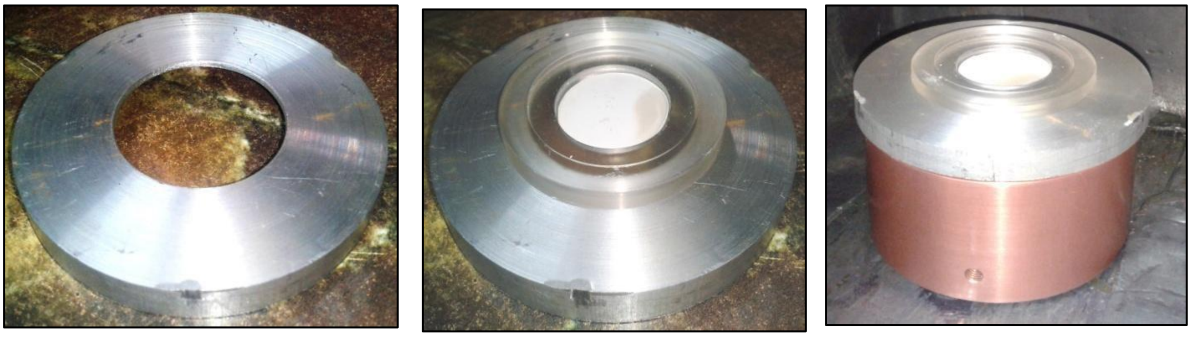


Figure 5.3: Part holding the sample above the radioisotope source and detector.

### 5.2.3 XRF Intensity and Energy Calibration

The height to machine the sample holder was determined based on the graphing on Figure 5.4. The height that maximized intensity for the detector was determined to be at the closest position near the base of the isotope. The sample height was determined to be at 3mm above the radioisotope source.

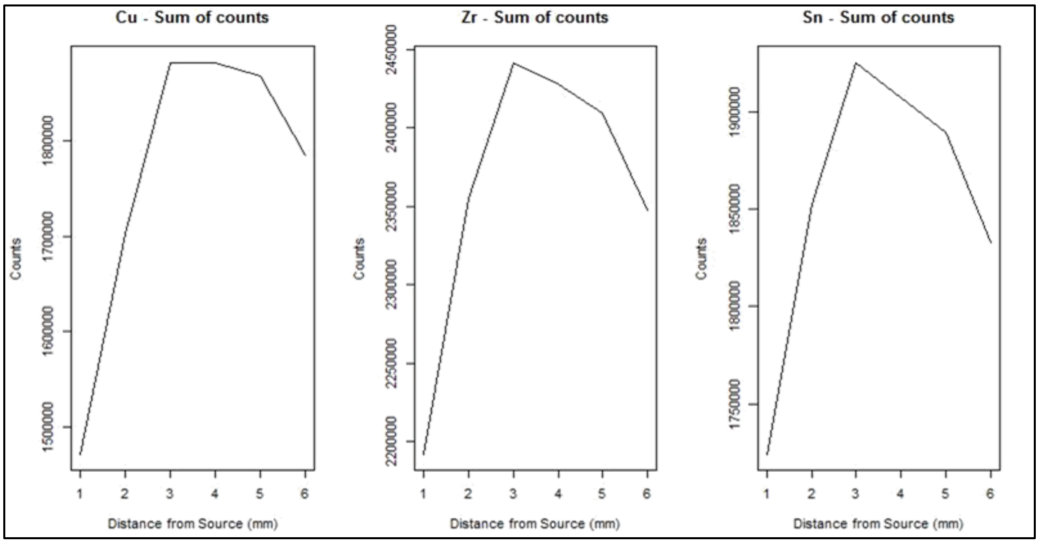


Figure 5.4: Variation in sum of counts at different distance from the source.

The height of the detector was also determined. The graphs below suggest that the highest intensity position was the nearest to the base of the radioisotope (1 mm just below the base of the radioisotope holder). Since the detector window is fragile, 1 mm room between the detector and the radioisotope was allowed. The maximum intensity is preferred because it minimizes analysis time. The background is a starch sample.

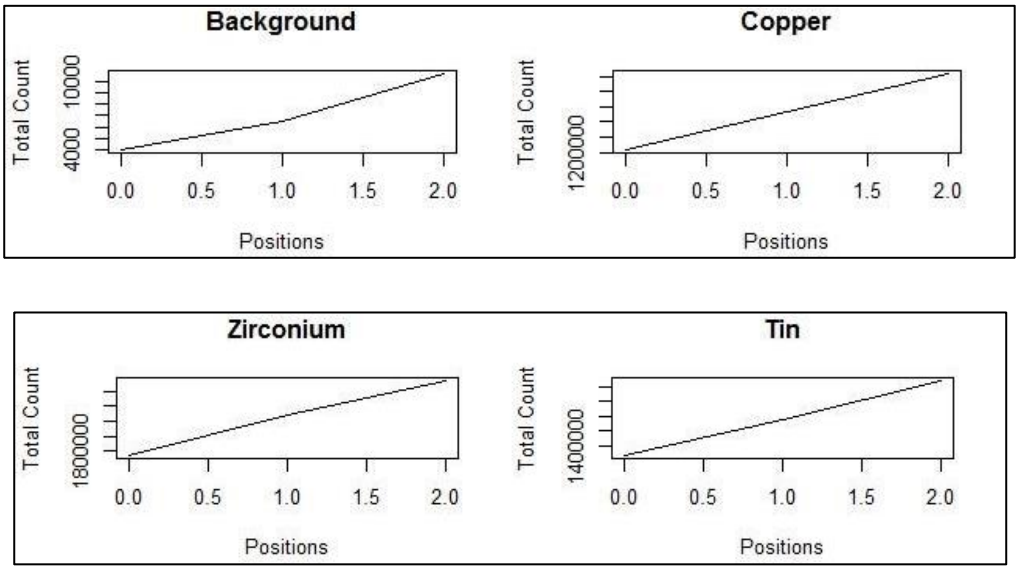


Figure 5.5: Detector positioning to determine the highest intensity.

## 5.3 Predictive Modelling

### 5.3.1 Simulate Sample Spectra Acquisition

Acquired spectrum and x-ray emissions table were used to identify the peaks associated with the elements of interest and the elements present in the matrices. Superimposed on the spectrum were Compton-scatter peaks and silicon escape peaks. Figure 5.6 shows the REE labeled XRF spectra typically generated by the instrument (the detector used in this study had a range of 0-26 keV). The figure also shows spectral peak and superimposed features directly identifiable through visual inspection. Table 5.2 shows the X-ray peak energy values and their identification.

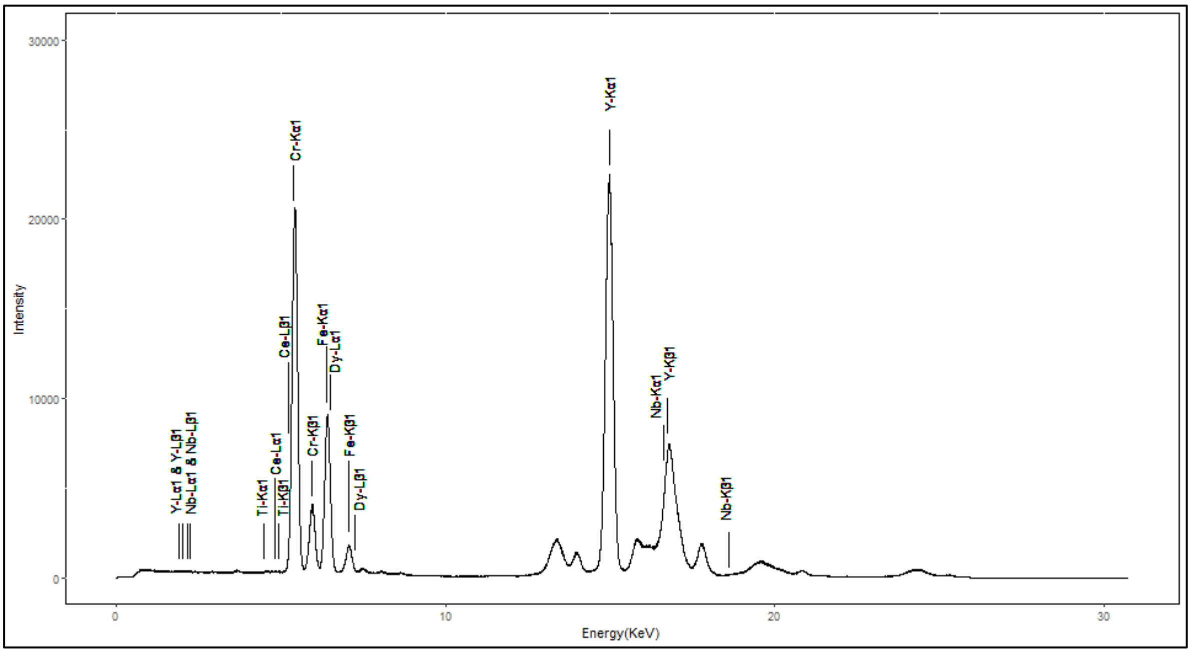


Figure 5.6: Sample spectra indicating regions of REEs of interest.

Table 5.2. Spectral artefacts identification of sample backscattered x-rays

| Identification | Peak energy (KeV) | Identification | Peak energy (KeV) |
|---|---|---|---|
| Ti-Lα1 | 0.45 | Dy-Lα1 | 6.51 |
| Y-Lα1 | 1.92 | Fe-Kβ1 | 7.11 |
| Y-Lβ1 | 2.01 | Dy-Lβ1 | 7.26 |
| Nb-Lα1 | 2.16 | Compton Np Lα | 13.45 |
| Nb-Lβ1 | 2.25 | Np Lα | 13.9 |
| Ti-Kα1 | 4.50 | Y-Kα1 | 14.91 |
| Ce-Lα1 | 4.83 | Si escape from 17.8 | 15.96 |
| Ti-Kβ1 | 4.92 | Nb-Kα1 | 16.62 |
| Ce-Lβ1 | 5.25 | Y-Kβ1 | 16.75 |
| Cr-Kα1 | 5.43 | Np Lβ1 | 17.80 |
| Cr-Kβ1 | 5.97 | Np-Ly1 | 20.92 |
| Fe-Kα1 | 6.42 | Compton 26.3 | 24.40 |

### 5.3.2 Region of Interest and Ratioing

Three regions of interest (ROIs) were identified in the rock and starch matrices; cerium-chromium (4.3-5.6 keV), dysprosium-iron (6.2-7.5 keV), and yttrium-niobium (14.6-18.8 keV). The ROIs were arrived at based on visual inspection, x-ray emission tables, and sections under curves. The ROIs were ratioed according to suspect Compton scatter regions; 12.7-13.7, 13.7-14.2, 16.4-17.4, 17.4-18.3, 18.7-20.6, 18.7-21.2, 20.6-21.2, 23.6-24.9, 23.6-25.5, and 24.9-25.5 keV. These regions are areas under particular peaks.

These regions were arrived at based on visual inspection of XRF spectral peaks that did not match any known peak from the source or sample nor any associated XRF effects.

Having ratioed for different regions in the spectra, Figure 5.7 shows the final ideal ratioing regions for the rock and starch matrices. This was a key result that was evaluated using PCA's ability to differentiate the various matrices according to element composition.

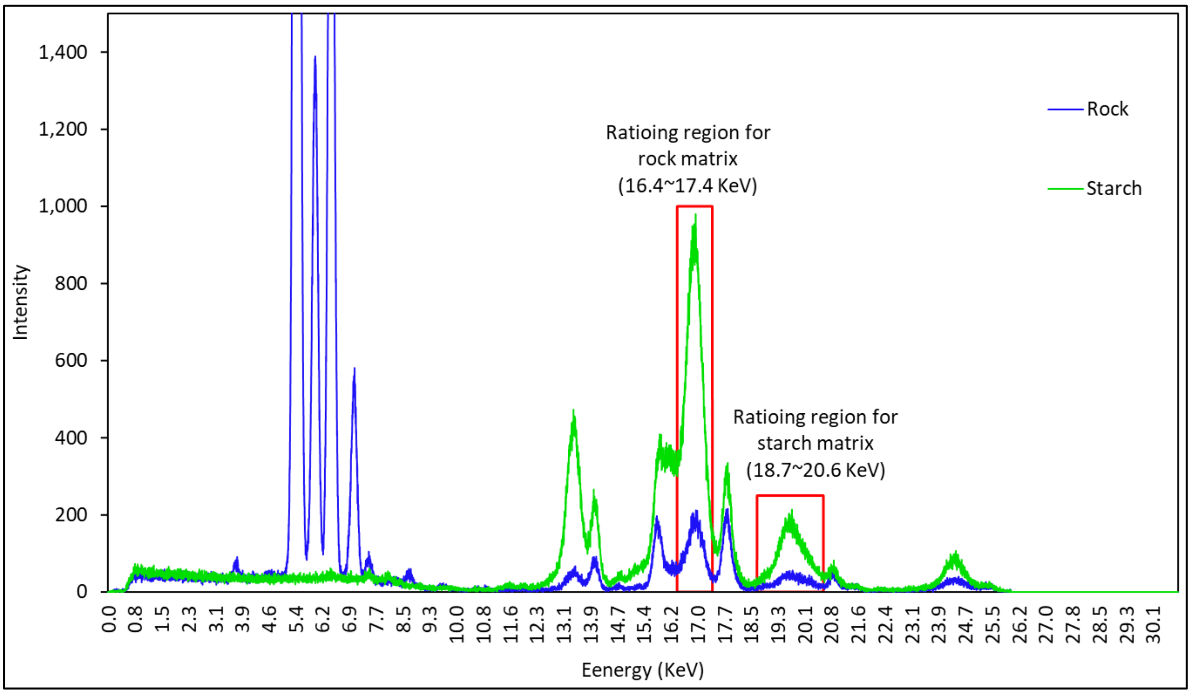


Figure 5.7: Starch and rock matrices profiles with identified ratioing regions

### 5.3.3 Principal Component Analysis

PCA was aimed at further reducing the number of variables by removing the effects of highly correlated variables. Furthermore, machine learning models perform better when the number of predictor variables are higher than the number of samples in training. The spectra were split into samples without elements and those with elements of interest so as to assess PCA performance in the ratioed regions. Rock and starch matrices PCA on their respective ratioed ROIs were performed. Two PCA scores were used to show the power of ratioing combined with PCA on the spectral data.

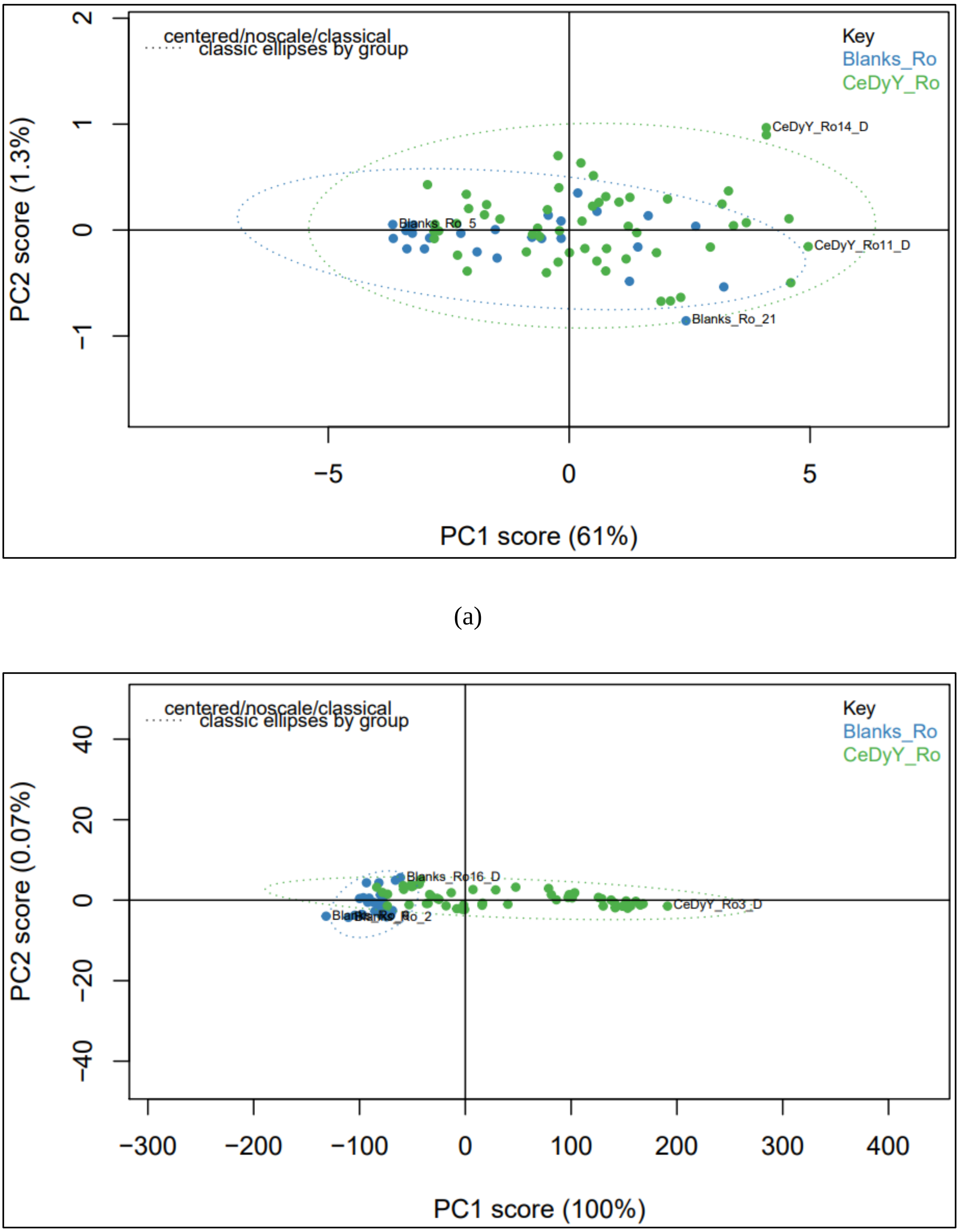
### 5.3.4 Rock Matrix PCA

PCA was first performed on rock samples without ratioing for Compton scatter. This step was combined with differentiating the samples into those that contained elements of interest and blanks. Secondly, the PCA results with ratioing applied on the spectra showed improved results overall indicating the dependence of XRF on Compton scatter. Three ROIs were subjected to PCA.

#### *5.3.4.1 Cerium-Chromium (4.3-5.6 keV) ROI*

Figure 5.8 shows the performance of the first two principal components (PCs). Figure 5.8 (a) shows the raw un-ratioed spectra PCA scores while (b) shows ratioed spectra PCA scores. It is evident that samples (blanks and element containing) could not be discerned without ratioing. In Figure 5.8 (b), PC1 explains all the variability in the data accounting for nearly 100%. PC1 clearly separates the blank sample from those containing elements however with an overlapping.

Figure 5.9 shows the PC loadings plot. The variability in spectral data between 5.2-5.6 keV represents the peak of cerium and chromium. Chromium being the dominant element in the matrix and it being at a slightly higher energy (5.43 keV, Cr-Kα1) than cerium (5.25 keV, Ce-Lβ1) produces significant enhancement effect, enough to differentiate the blank matrices from those containing elements of interest.



(a)

(b)

Figure 5.8: Rock Cerium-Chromium  PCA scores plot without (a) and with (b) ratioing.

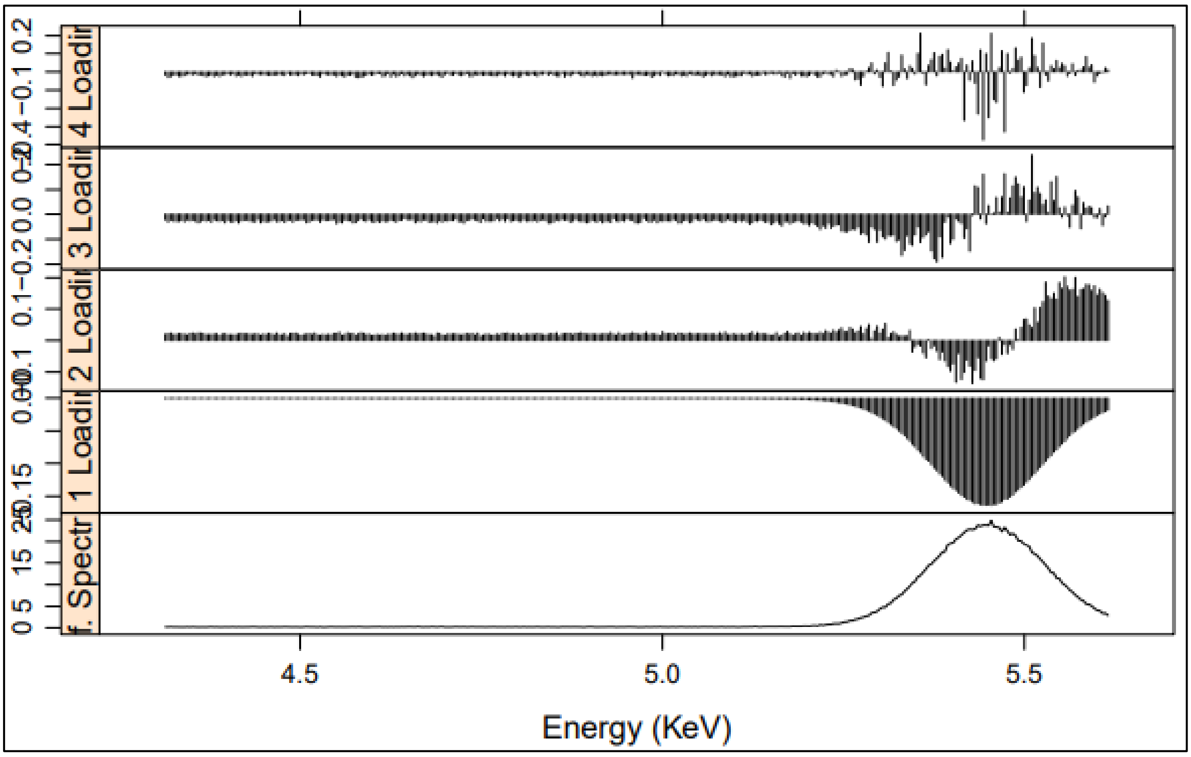
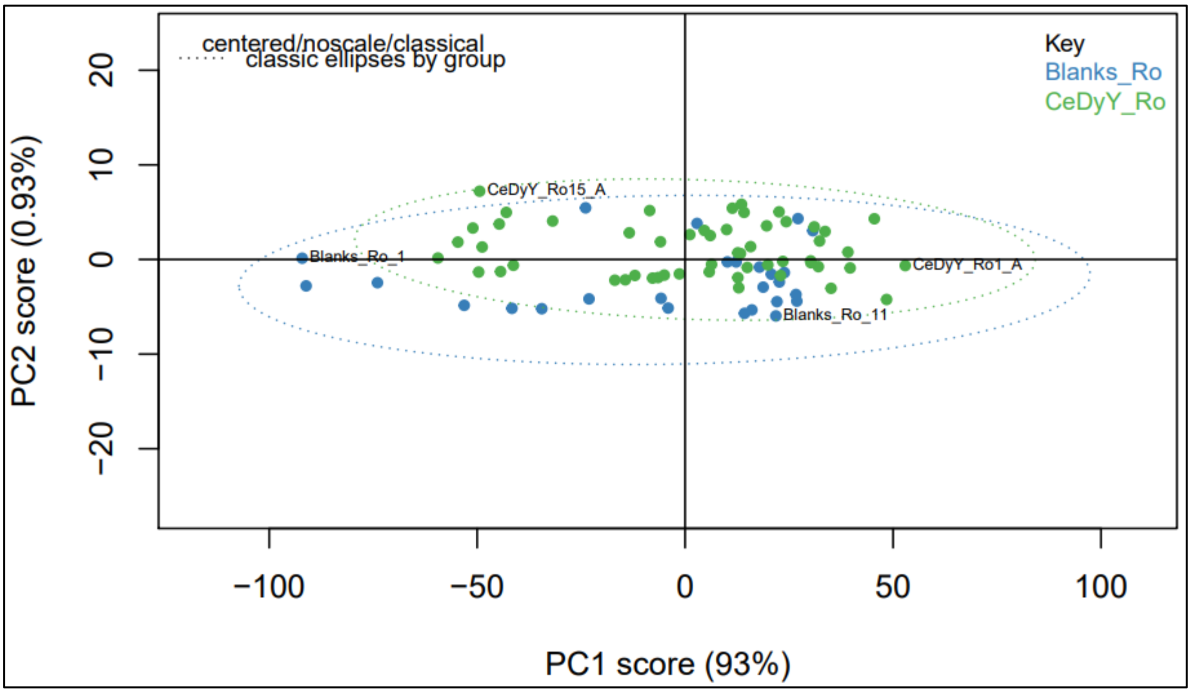


Figure 5.9: Rock Cerium-Chromium PCA loadings plot with ratioing.

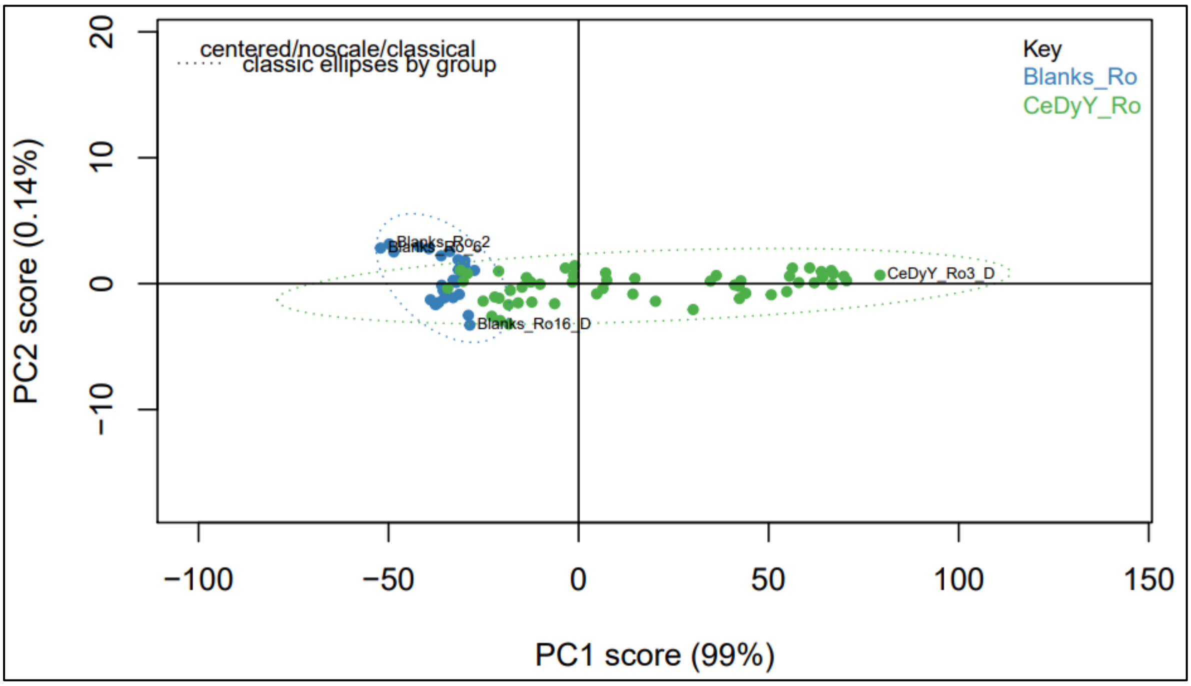
*5.3.4.2 Dysprosium-Iron (6.2-7.5 keV) ROI*

Figure 5.10 bears semblance to Figure 5.8. Ratioing combined with PCA presented similar results. In Figure 5.10 (b), similar to Figure 5.8 (b), PC1 accounts for almost all variability in the element concentration in the samples. PC1 similarly accounted for most of the variability at 99%. Two regions can be discerned however with overlapping.

Figure 5.11 shows the PCA loadings plot. The loadings plot of PC1 on the spectral region between 6.2-6.7 keV and 7.0-7.3 keV represent the energies of Dy-Lα1 (6.51 keV), and Dy-Lβ1 (7.26 keV). PC1 therefore explains the variability in the data owing to the presence of dysprosium. Higher order PCs explain the variability in concentration, mass, grain size, etc.



(a)



(b)

Figure 5.10: Rock Dysprosium-Iron PCA scores plot without (a) and with (b) ratioing.

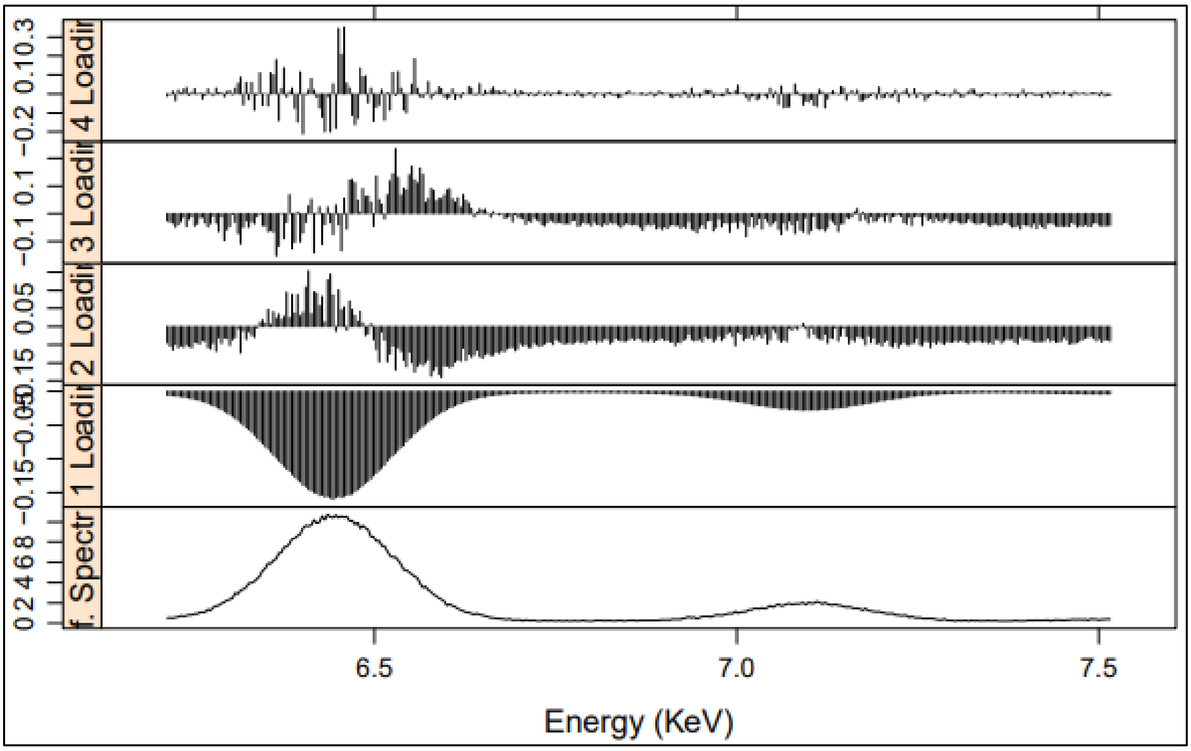
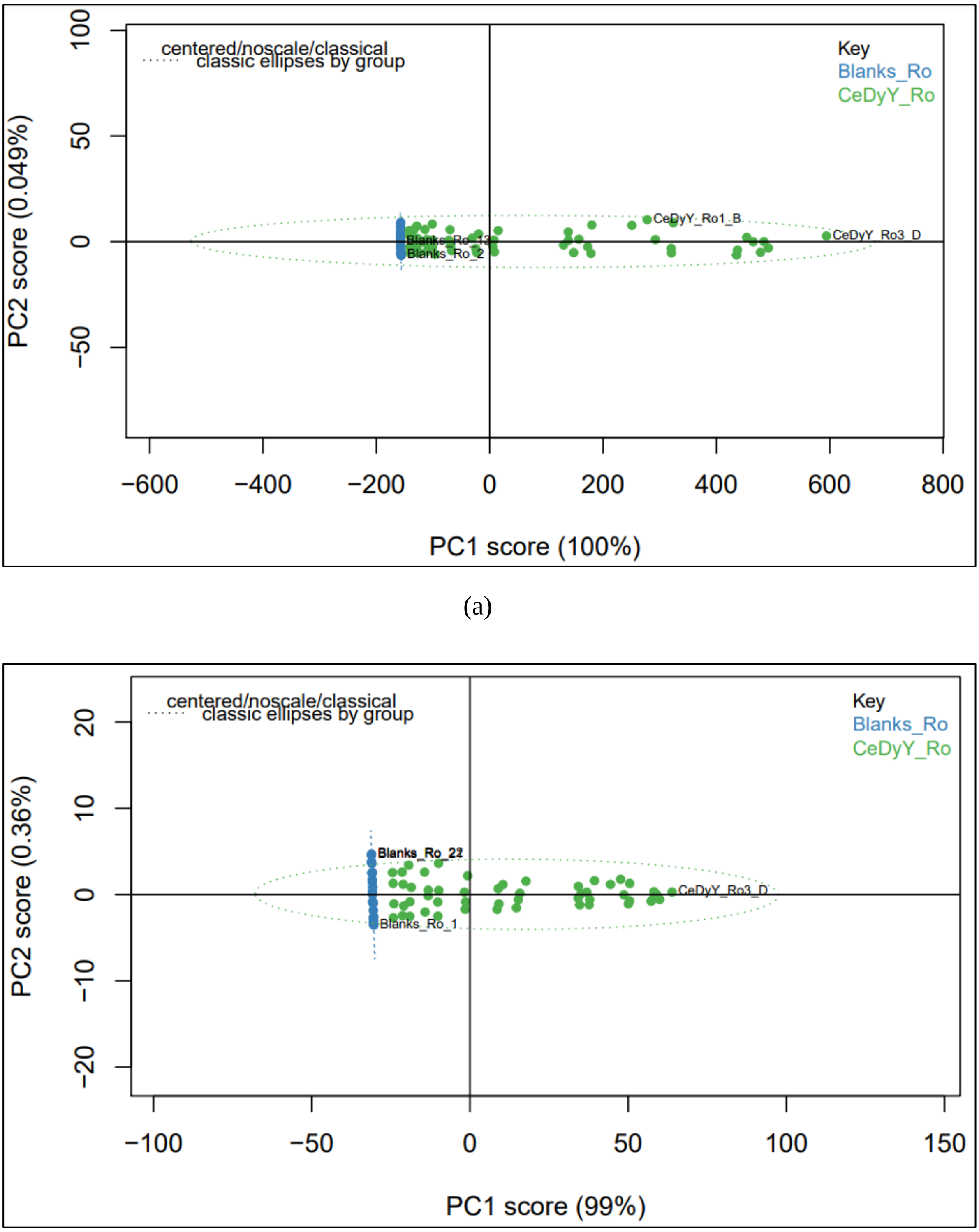


Figure 5.11: Rock Dysprosium-Iron PCA loadings plot with ratioing.

*5.3.4.3 Yttrium (14.6-18.8 keV) ROI*

The rock matrix contained Yttrium as the element of interest in the ROI, however, 15.9 keV peak identified as the silicon escape from 17.8 keV (due to the excitation source nuclei-Np Lβ1) existed. Yttrium has two prominent peaks at 15.00 keV (Y-Kα1) and 16.75 keV (Y-Kβ1). Figure 5.12 shows the score plots of the PCs. It can be seen clearly that by ratioing, PCA could clearly mark out two distinct groups of samples. This performance can be attributed to the overall higher concentration of Yttrium in the samples with PC1 explaining its variability.

Figure 5.13 shows the PCA loadings plot with PC1 weights at the two peaks of Yttrium at 15.00 keV (Y-Kα1) and 16.75 keV (Y-Kβ1).



(a)

(b)

Figure 5.12: Rock Yttrium PCA scores plot without (a) and with (b) ratioing.

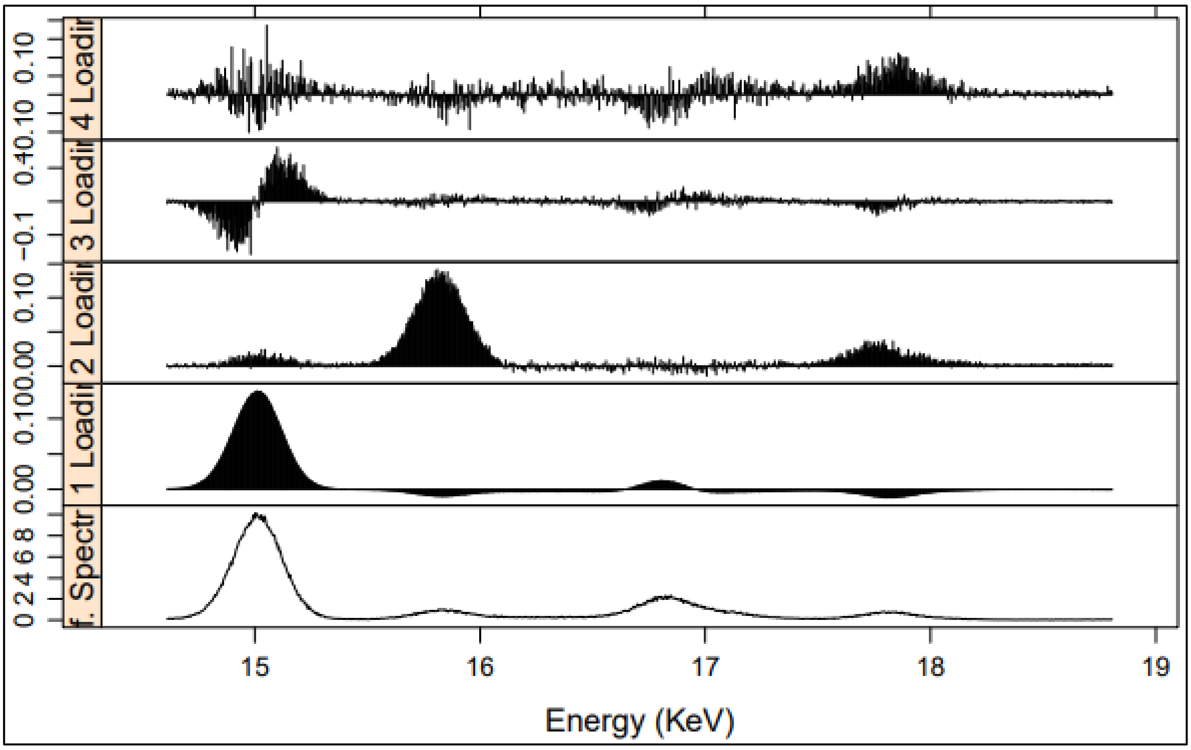


Figure 5.13: Rock Yttrium PCA loadings plot with ratioing.

### 5.3.5 Starch Matrix PCA

PCA of starch matrices involved the ROIs of Cerium-Titanium (4.3-5.6 keV), Dysprosium (6.2-7.5 keV), and Yttrium-Niobium (14.6-18.8 keV). Titanium and Niobium are two additional elements that were added to the matrix to simulate energy regions not covered by the available elements. Starch matrix did not contain Chromium and Iron. Three ROI were subjected to PCA.

#### *5.3.5.1 Cerium-Titanium (4.3-5.6 keV) ROI*

Figure 5.14 shows the starch cerium-titanium PCA scores plot of the first two PCs. PC1 accounts for 83% of the variability in the spectral data responsible for differentiating matrices that contain cerium, titanium, and blanks. Beyond PC2 explaining variance of 0.9%, the amount of variance explained does not increase significantly.

Figure 5.15 shows PC2 loadings in the region between 4.4-4.7 keV and 4.8-5.0 keV indicating Titanium lines centered at 4.5 keV (Ti-Kα1) and 4.9 keV (Ti-Kβ1) respectively. PC2 therefore explains much of the variability in titanium content. It is also worth noting that titanium and cerium were not mixed in one sample, rather in different samples of the same matrix. PC1 therefore mainly accounts for the concentration of cerium while PC2 accounts for the concentration of titanium.

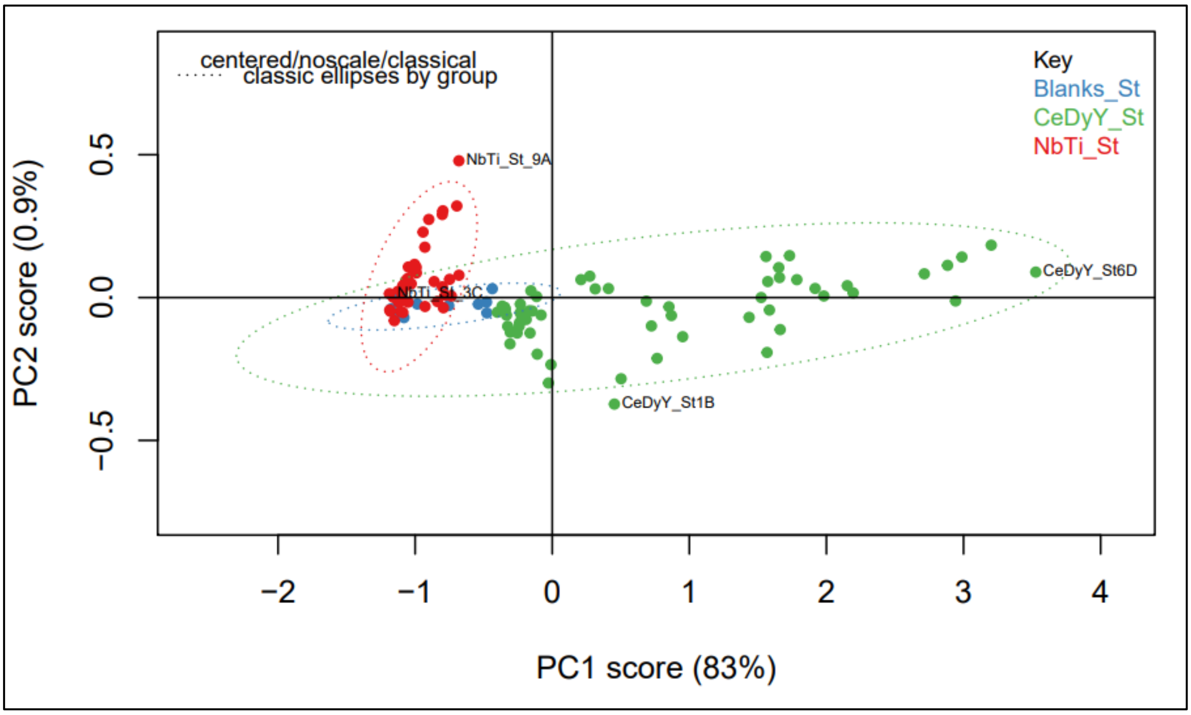


Figure 5.14: Starch Cerium-Titanium PCA scores plot with ratioing.

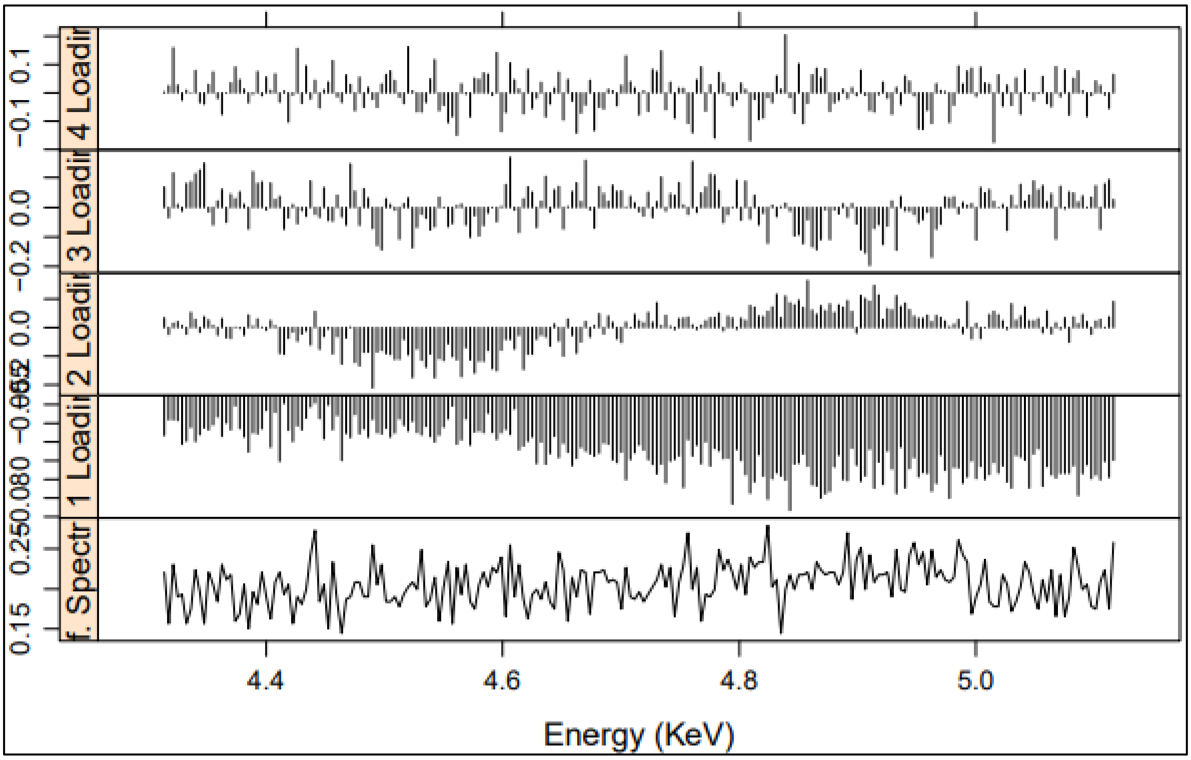


Figure 5.15: Starch Cerium-Titanium PCA loadings plot with ratioing.

### *5.3.5.2 Dysprosium (6.2-7.5 keV) ROI*

Figure 5.16 shows the first two PCs of the dysprosium ROI PCA. PC1 accounts for 80% of the variability in the spectral data. This is enough to differentiate samples containing dysprosium and blanks. The first two PCs account for 85% variability in the spectral data, rendering PCA a very important step in machine learning. Figure 5.17 shows the loadings plot of the first 4 PCs. The region 6.5 and 7.3 keV represent the Dy-Lα1 and Dy-Lβ1 lines of dysprosium. PC2 scores and loadings plot show that the spectra is weighted in both the positive and negative region of the dysprosium peaks. PC2 therefore accounts for the concentration of dysprosium in the samples.

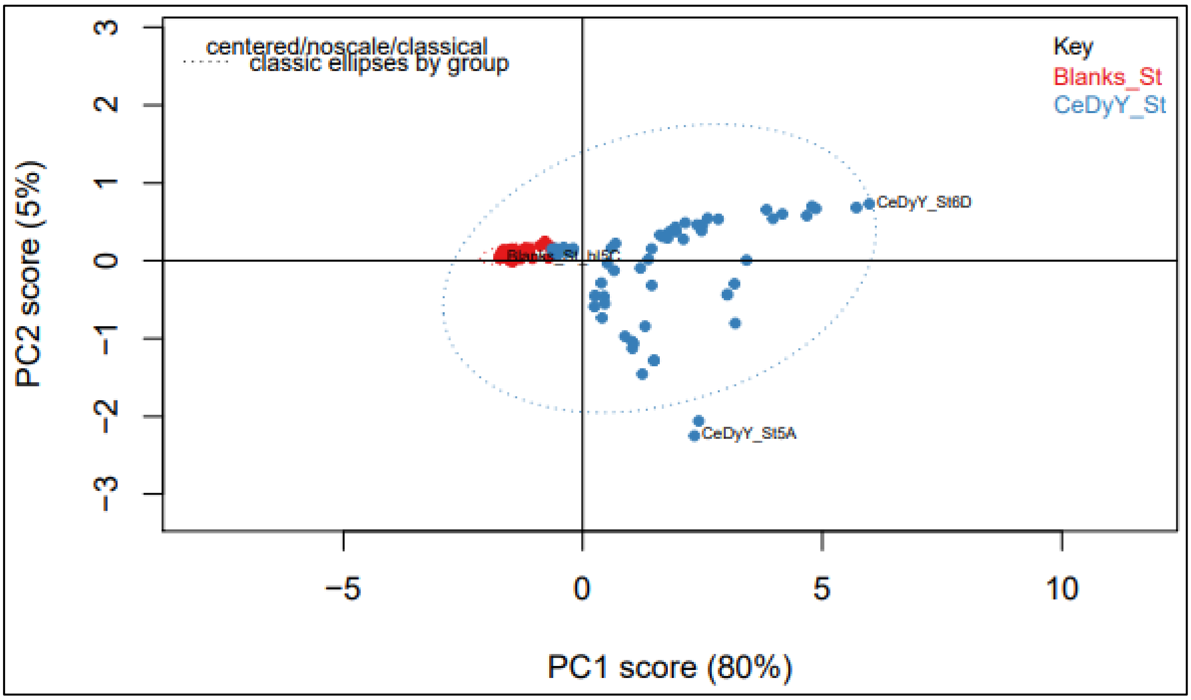


Figure 5.16: Starch Dysprosium scores plot with ratioing.

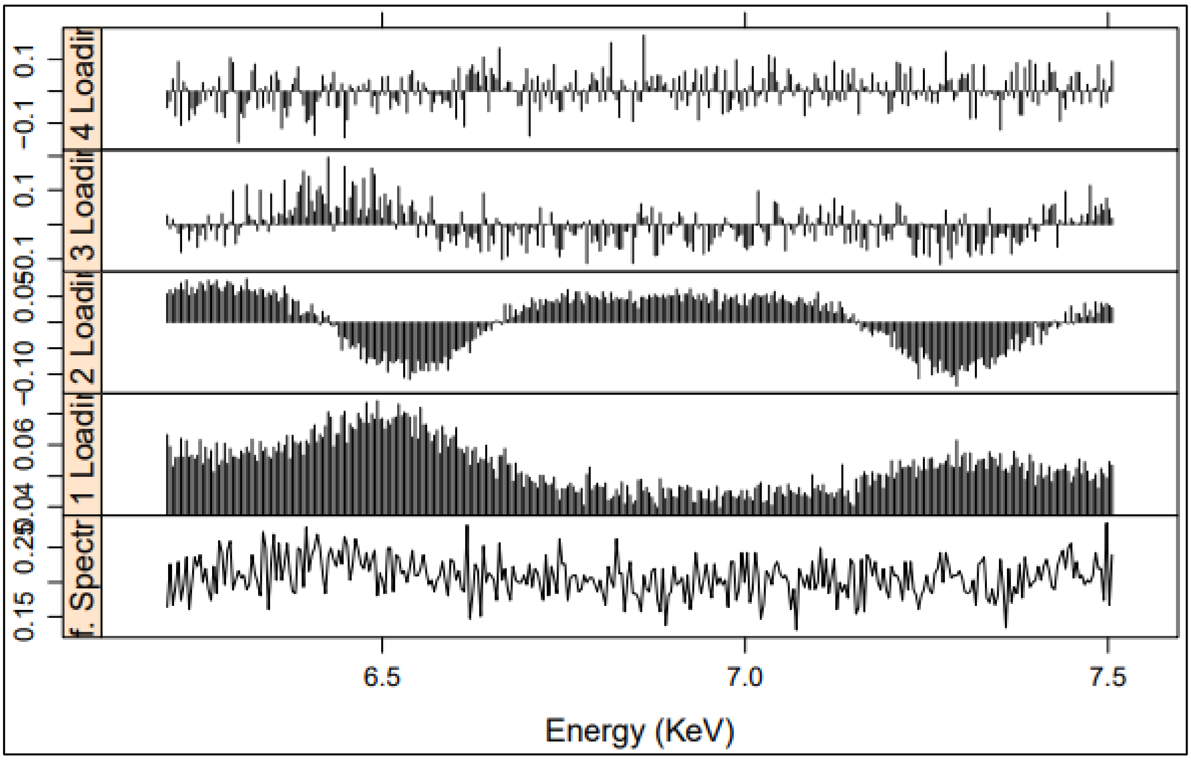


Figure 5.17: Starch Dysprosium PCA loadings plot with ratioing.

*5.3.5.3 Yttrium-Niobium (14.6-18.8 keV) ROI*

Figure 5.18 is the PCA scores plot of the first two PCs. Due to the high concentration of yttrium in the samples, PC1 clearly distinguishes the samples based on it. The niobium containing samples cannot be distinguished by either of the two PCs because of its low concentration. It can be noted that PC1 explains nearly 100% of the variability in the spectral data, and manages to cluster along its negative component the lower concentrations of yttrium. PC1 clustered niobium samples at its extreme negative. It should be noted that yttrium and niobium were analyzed in separate matrices.

Figure 5.19 shows the PCA loadings plot. Positive PC1 shows it is weighted at the two peaks of yttrium (15.00 keV Y-Kα1, and 16.74 keV Y-Kβ1). Niobium peak (16.63 keV Nb-Kα1) was not visibly explained by the plotted PCs.

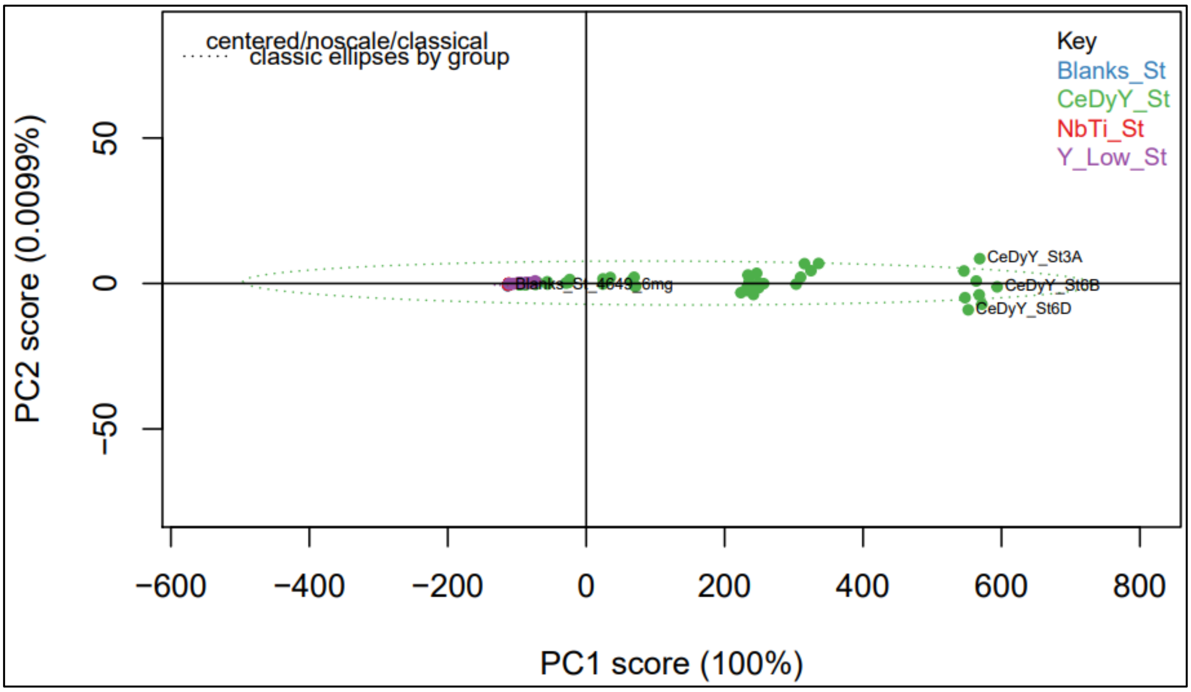


Figure 5.18: Starch Yttrium-Niobium scores plot with ratioing.

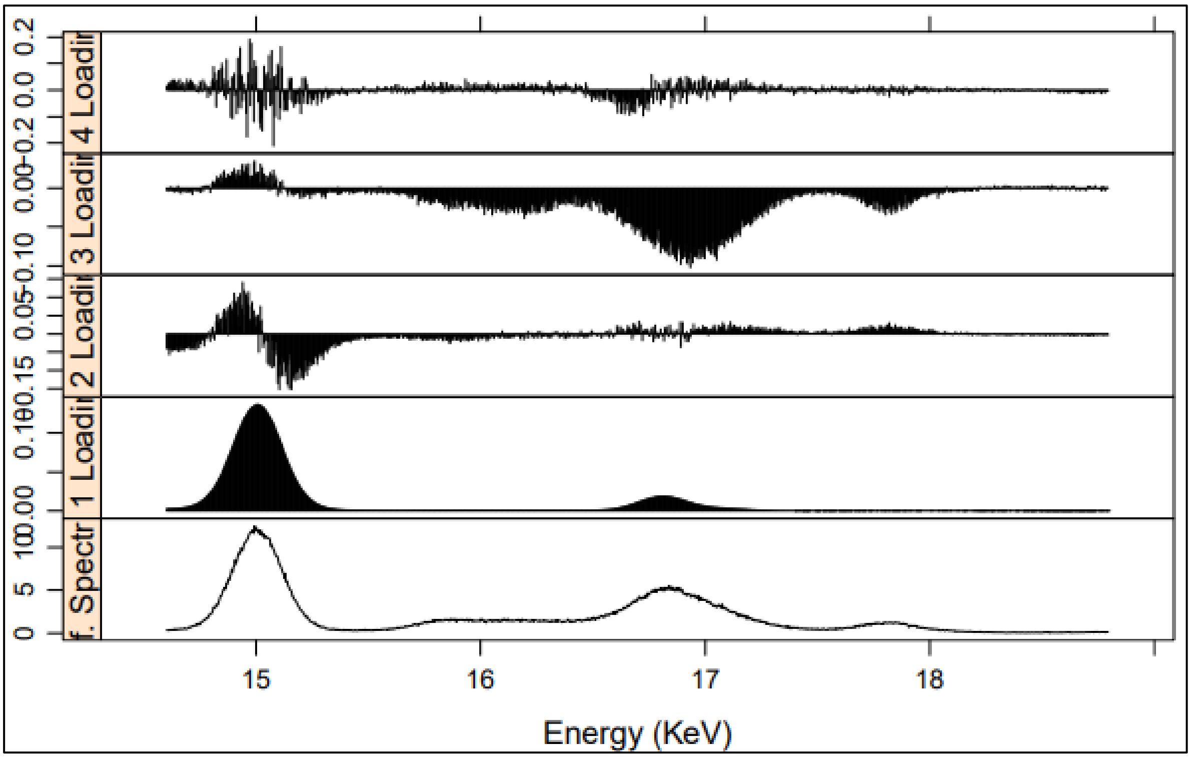


Figure 5.19: Starch Yttrium-Niobium PCA loadings plot with ratioing.

## 5.4 Comparison of Predictive Models

Predictive models based on the chemometric techniques; Artificial Neural Networks (ANN), Support Vector Regression (SVR), and Random Forests (RF) were implemented for Dysprosium, Yttrium, Cerium, Titanium, and Niobium. Their performance were divided into two types of matrices; rock and starch matrices. Each matrix results were further divided into respective element constituents. The rock matrix did not contain Titanium and Niobium, therefore, they were only reported in the starch matrix. The models were chosen based on iterative process of changing the number of principal components. Neural networks choice of number of layers and neurons per layer were automated by following the rule of thumb that; "number of layers should be about half of the input variables while the neurons per layer should be about two thirds of the preceding layer/input". SVR models were tuned by allowing the model to run on various values of cost and sigma. Random Forest models were tuned by allowing them to explore various values of tree splits (mtry). The models were trained with 10-fold cross validation. The models were evaluated on various metrics; Root Mean Square Error of Prediction (RMSEP), coefficient of determination ($R^2$), Limit of Detection (LOD), and Limit of Quantification (LOQ). RMSEP values were considered as the primary selection criteria of the models, followed by $R^2$ value.

### 5.4.1 Rock Matrix Predictive modelling

The rock matrix simulated the severe effects of the matrix on predicting concentrations in EDXRF. Three elements were considered; Cerium, Dysprosium, and Yttrium.

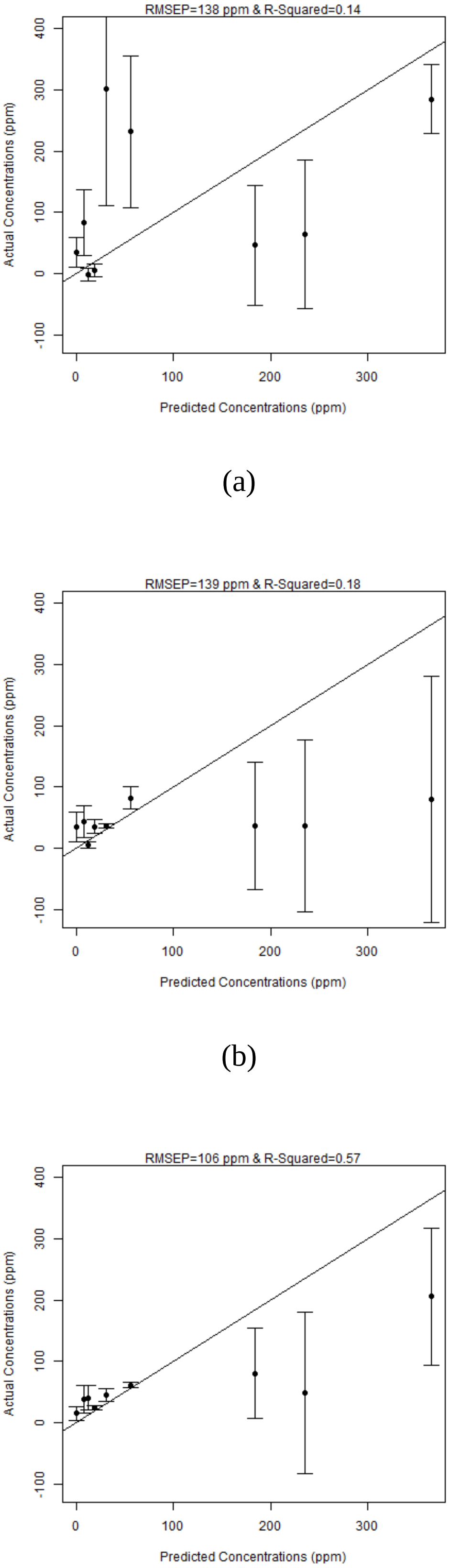
#### *5.4.1.1 Cerium*

Cerium in rock matrix ranged between 0 and 378 ppm. The number of PCs to use were varied from 1 to 11 principal components with the best performing model being random forests using 9 principal components. Table 5.3, Figure 5.20, and Table D.1 summarizes the performance of the models at 9 PCs. Running the models on the training dataset (control) indicates that the models do not overfit since the models were trained with validation. All Cerium testing models in the rock matrix achieved low RMSEP values of greater than 105 ppm. The models attained accuracies of below 20% except for the Random Forests Model at 57%. The models were considered unsuccessful.

Questions arise as to why the matrix of the blank and element-containing samples were separable with PCA while the quantifying prediction models could not. The main reason for this is the spectral overlap presented by chromium, a major constituent in the rock matrix. The K-lines of chromium interfered with the L-lines of cerium; therefore, a higher detection range of the detector could potentially solve the problem by measuring the K-lines of cerium. Increasing the number of PCs is a possible workaround but it would tremendously increase computational time and beat the logic of using PCA. Testing out the models up to 11 PCs was a computational time constraint, and can therefore be improved with better hardware and/or software. Other possibilities include increasing the concentration of REE in training with the consequence being lower detection limits. The effect of using higher concentration for training was clearly demonstrated by Yttrium in rock matrix (later in this section).

Table 5.3. Cerium models performance in rock matrix.

| | **NN** | **SVR** | **RF** |
|---|---|---|---|
| RMSEP_Test (ppm) | 138 | 139 | 106 |
| R_Sq_Test | 0.14 | 0.18 | 0.57 |
| LOD (ppm) | 7 | 17 | 9 |
| LOQ (ppm) | 24 | 57 | 30 |



(a)

(b)

(c)

Figure 5.20: Cerium (a) NN, (b) SVR, and (c) RF models on test dataset.

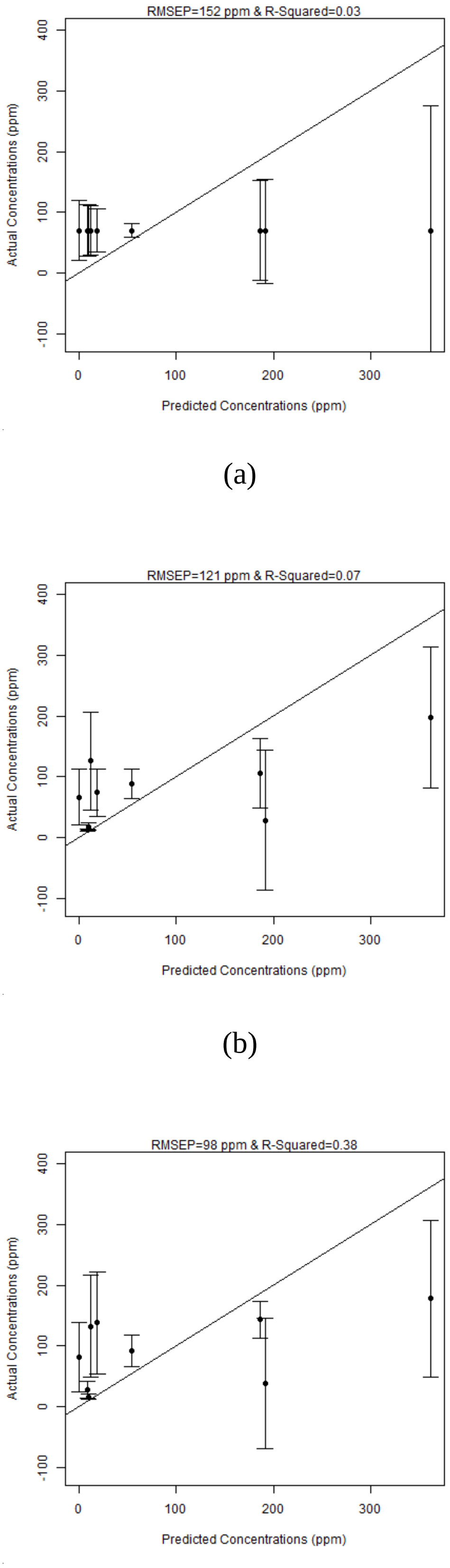
#### *5.4.1.2 Dysprosium*

Concentration of dysprosium in rock matrix ranged between 0 and 396 ppm. The number of PCs were varied from 1 to 11 with the best performing being 3 PCs with the random forest model. Table 5.4, Figure 5.21, and Table D.2 summarizes the performance of the models using 3 PCs. The RF model achieved the lowest RMSEP of 79 ppm at an accuracy of 41%. The model, having been cross validated, had a lower RMSEP of 52 ppm at 88% accuracy. All models achieved accuracies below 45% which were considered low and therefore unsuccessful.

Similar to Cerium, question arise of why the dysprosium containing samples were separable from the blanks but cannot be quantified by the models. Classically, the number of PCs can be increased, however, the more the PCs do not necessarily mean any improvement in predictive ability. In fact, the first PC accounts for 99% of the variability in the data. Beyond 3 PCs, the performance of the models got lower which means that the variability in the PCs beyond 3 explain noise. The major constituent in the rock matrix was iron which spectrally interferes with the emission lines of dysprosium. The effects of iron can be overcome using a detector with a higher range of detection that covers the K-lines of dysprosium. Other potential solutions involve having higher concentrations of analyte in the matrix and increasing the number of PCs. The number of PCs can be advantageous but comes with a computational time penalty and negates the logic of using PCA. Better hardware and software are advantageous in faster processing of the models.

Table 5.4. Dysprosium models performance in rock matrix

| | **NN** | **SVR** | **RF** |
|---|---|---|---|
| RMSEP_Test (ppm) | 111 | 96 | 79 |
| R_Sq_Test | 0.16 | 0.13 | 0.41 |
| LOD (ppm) | 38 | 24 | 20 |
| LOQ (ppm) | 127 | 79 | 66 |



(a)

(b)

(c)

Figure 5.21: Dysprosium (a) NN, (b) SVR, and (c) RF models on test dataset.

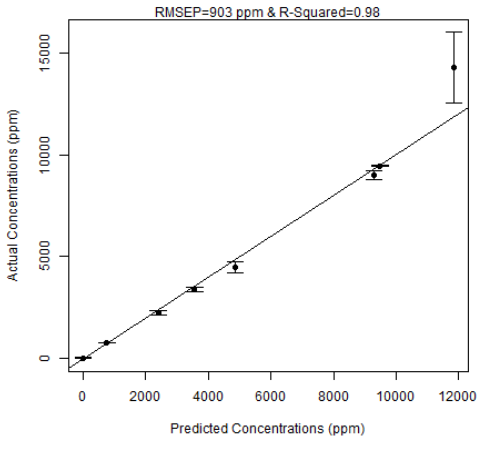
*5.4.1.3 Yttrium*

The concentration of yttrium in rock matrix ranged between 0 and 18276 ppm. This is a significantly high concentration range unlike cerium and dysprosium. Modelling was performed considering 1 to 6 PCs with the best results achieved with 2 PCs. The number of PCs was chosen using the model with the lowest RMSEP. Table 5.5, Figure 5.22, and Table D.3 summarize the results of using 2 PCs. The lowest RMSEP was achieved by random forest model at 140 ppm at an accuracy of 99.9%. Neural network performed fairly at 903 ppm and support vector regression at 1215 ppm with accuracies of 98% and 99% respectively. The performance of the models on the training dataset was inexact because the models were internally validated. The analysis was considered as a success.

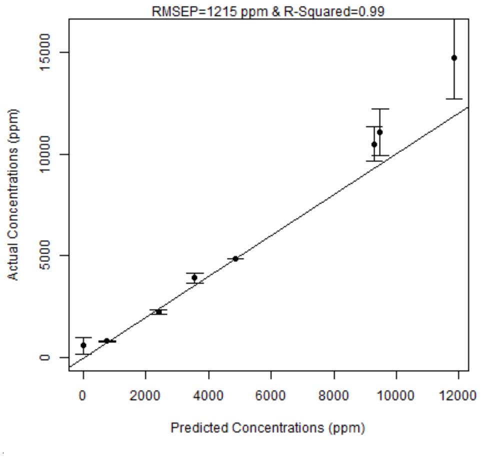
Evidently, the results of yttrium modelling, unlike cerium and dysprosium, benefited from a wider range of concentration and lower matrix interference. This was due to the abundant availability of yttrium for training. The PCA results of Figure 5.12 show that PC1 explains 99% of the variability in the data. However, as much as PC1 neatly separates the two groups of samples (with and without yttrium), 2 PCs were optimal for best results. In fact, using 1 PC achieved a slightly lower accuracy of 146 ppm at 99.9% accuracy.  It is important to note the low LOD (64 ppm) and LOQ (214 ppm) of RF model by considering the concentration range of yttrium (0~18275 ppm). The significantly high concentrations of Yttrium proved useful in understanding the lower limits of detections of models. The model however would respond very differently if it were exposed to data from different matrices.

Table 5.5. Yttrium models performance in rock matrix

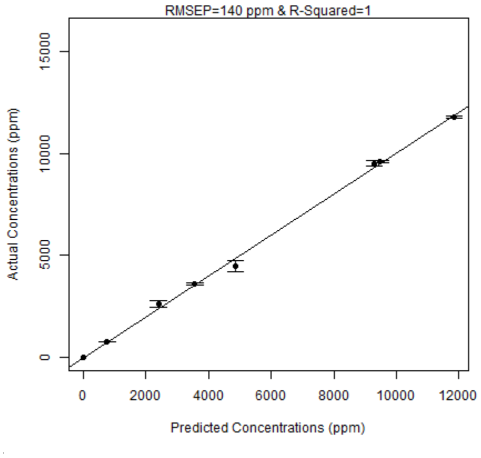
| | **NN** | **SVR** | **RF** |
|---|---|---|---|
| RMSEP_Test (ppm) | 903 | 1215 | 140 |
| R_Sq_Test | 0.98 | 0.99 | 1.00 |
| LOD (ppm) | 269 | 297 | 64 |
| LOQ (ppm) | 896 | 991 | 214 |



(a)



(b)



(c)

Figure 5.22: Yttrium (a) NN, (b) SVR, and (c) RF models on test dataset.

### 5.4.2 Starch Matrix Predictive modelling

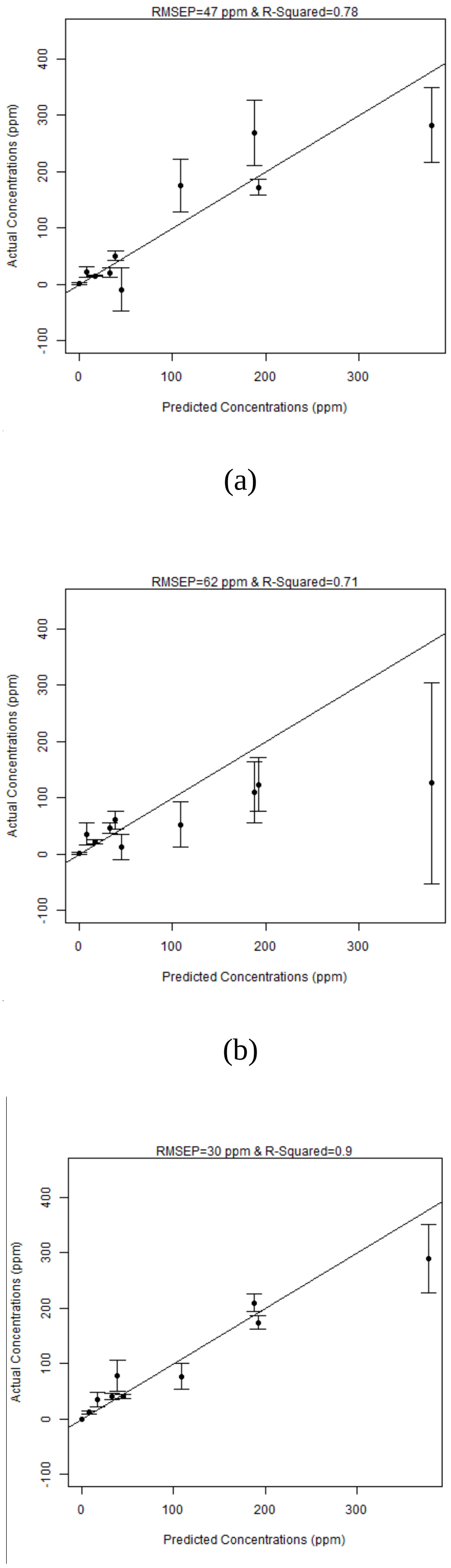
The starch matrix was considered a transparent matrix unlike the rock matrix. Analysis for elements was therefore deemed free of matrix interference. This dramatically improved the predictive ability of models. Accompanying elements however were present, which inevitably led to interference. The models were evaluated similar to the rock matrix elements.

#### *5.4.2.1 Cerium*

Cerium concentration in starch matrix ranged between 0 and 378 ppm. The number of PCs was varied between 1~12 PCs with the best performing model at 7 PCs. Random Forest model presented the best predictive ability with RMSEP value of 30 ppm at 90% accuracy. Table 5.6, Figure 5.23 and Table D.4 summarize the performance of the three models. The neural network with 7 PCs performed second best with RMSEP of 47 ppm at 78% accuracy while the support vector regression model performed least with 62 ppm at 71% accuracy. According to Figure 5.14, PC1 explains 83% of the variability in the spectral data while PC2 only explains 0.9%. Finding that 7 PCs were required to achieve good results indicate that higher order PCs are essential in many situations. In fact, using 1 PC only achieved RMSEP of 120 ppm at 25% accuracy although it could visually separate the samples (those with and without cerium). The performance of the models on the training set proves the importance of validation in training. The control results show the models did not overfit and the models were deemed successful. The emission lines of cerium were not clearly discernible from visual inspection. This presented difficulty in spectral peak identification. Additionally, Titanium which spectrally interfere with cerium was prepared independently but its spectrum was added to spectral data. Despite the interference, the samples were distinguishable by PCA and quantifiable by machine learning models.

Table 5.6. Cerium models performance in starch matrix

| | **NN** | **SVR** | **RF** |
|---|---|---|---|
| RMSEP_Test (ppm) | 47 | 62 | 30 |
| R_Sq_Test | 0.78 | 0.71 | 0.90 |
| LOD (ppm) | 7 | 14 | 6 |
| LOQ (ppm) | 23 | 46 | 21 |



(a)

(b)

(c)

Figure 5.23: Cerium (a) NN, (b) SVR, and (c) RF models on test dataset.

*5.4.2.2 Dysprosium*

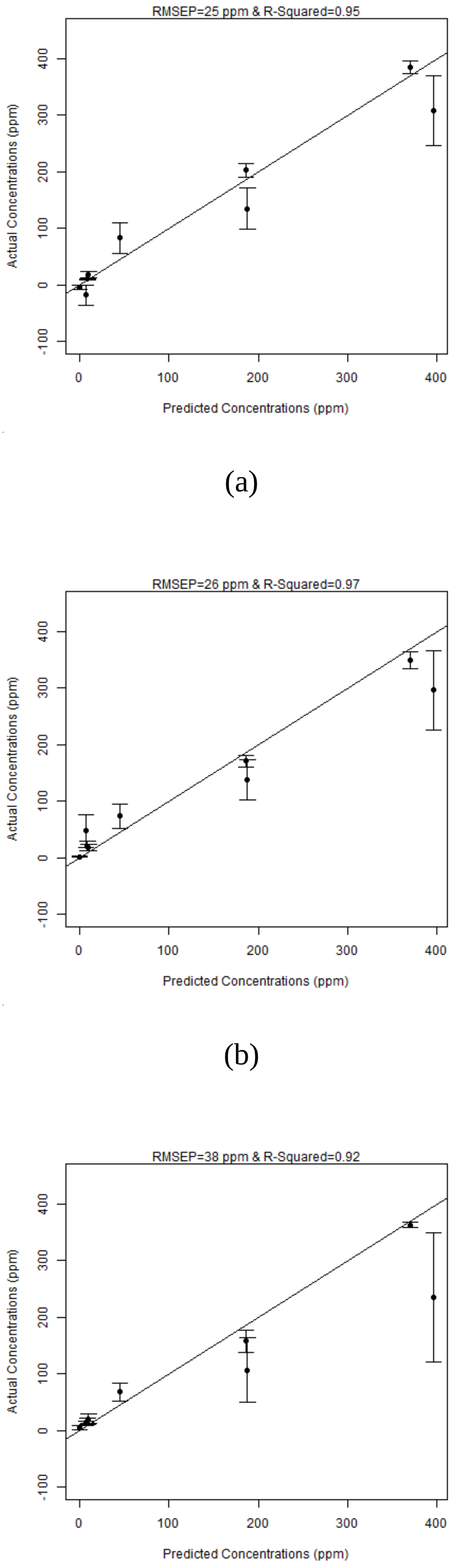
Concentrations of Dysprosium used to develop the models ranged from 0 – 396 ppm. The number of PCs was varied between 1 and 7 PCs with 5 PCs producing the best results. The neural network performed best with RMSEP of 25 ppm at 95% accuracy. SVR and RF models RMSEP values were 26 and 38 ppm at accuracies of 97% and 92% respectively. Table 5.7, Figure 5.24, and Table D.5 summarizes the results. According to Figure 5.16 PC1 and PC2 respectively account for 80% and 5% of the variability in the spectral data. The two PCs were sufficient to distinguish the presence of dysprosium in samples. However, using the two PCs for predictive modelling produced RMSEP of 31 ppm at 99% accuracy with the best model. The finer details were therefore tucked in PC3 to PC5. Beyond PC5, the models performed poorly. The models were deemed successful.

Visual inspection of spectral data did not readily identify dysprosium peaks. The only “visible” lines were the L-lines (due to detector limitations). However, combined with ratioing, PCA could resolve the low intensity peaks, and went the extra extent of successfully quantifying dysprosium. Earlier analyses that did not apply these techniques produced high RMSEP values with low accuracies.

Parallels were drawn between dysprosium in rock and starch matrices. The good results of dysprosium in starch and not in rock matrix shows how matrix effects can bring complexity to analyses.

Table 5.7. Dysprosium models performance in starch matrix

| | **NN** | **SVR** | **RF** |
|---|---|---|---|
| RMSEP_Test (ppm) | 25 | 26 | 38 |
| R_Sq_Test | 0.95 | 0.97 | 0.92 |
| LOD (ppm) | 7 | 8 | 6 |
| LOQ (ppm) | 24 | 28 | 19 |



(a)

(b)

(c)

Figure 5.24: Dysprosium (a) NN, (b) SVR, and (c) RF models on test dataset.

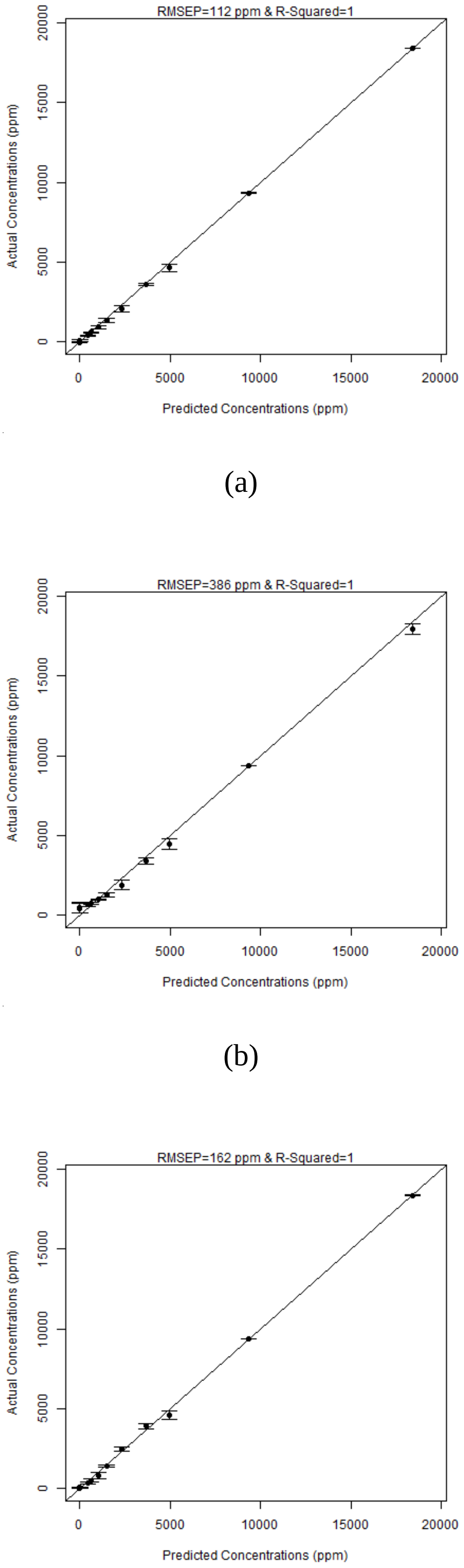
#### *5.4.2.3 Yttrium*

Concentrations for Yttrium modelling varied from 0 ~ 18390 ppm, the highest in this study. Input variables were varied between 1 and 5 PCs with the best model using only 1 PC. Neural network using only 1 PC achieved lowest RMSEP of 112 ppm at an accuracy of 99.99%. The RF and SVR achieved RMSEP of 162 and 386 ppm respectively both at accuracy of 99.99%. Table 5.8, Figure 5.25, and Table D.6 summarizes the model performance of using 1 PC. The results according to PCA results in Figure 5.18 show that PC1 explains all the variability in the spectral data (PC2 explains only 0.0099%). All the models were considered successful. However, neural network was the ideal model to use because it outperformed the RF and SVR models.

The excellent response of Yttrium to modelling is directly attributed to the higher concentration range coupled with a transparent matrix. Besides, Niobium which would spectrally interfere with it was prepared in different samples. This ideal situation is undesirable in geological matrices because other elements exist in the X-ray emission vicinity of Yttrium emission lines.

Table 5.8. Yttrium models performance in starch matrix.

| | **NN** | **SVR** | **RF** |
|---|---|---|---|
| RMSEP_Test (ppm) | 112 | 386 | 162 |
| R_Sq_Test | 1 | 1 | 1 |
| LOD (ppm) | 71 | 99 | 34 |
| LOQ (ppm) | 236 | 332 | 112 |



(a)

(b)

(c)

Figure 5.25: Yttrium (a) NN, (b) SVR, and (c) RF models on test dataset.

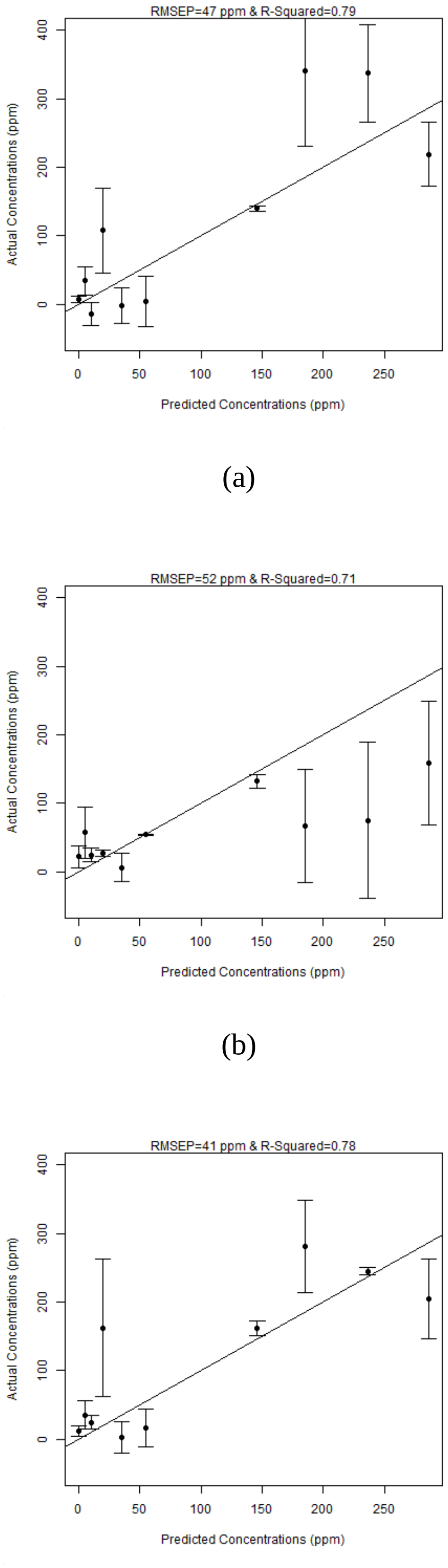
#### *5.4.2.4 Titanium*

The concentration of titanium ranged between 0 and 349 ppm. The number of PCs were varied between 1 and 10 PCs with 5 PCs producing the best results. Table 5.9, Figure 5.26, and Table D.7 summarize the modelling results using 5 PCs. Random forest performed best with RMSEP of 41 ppm at 78% accuracy. NN and SVR models achieved RMSEP of 47 and 52 ppm at accuracy of 79% and 71% respectively. Figure 5.14 shows the PCA score plot of cerium-titanium spectra with PC1 explaining 83% of spectral data variability and PC2 explaining 0.9%. Similar to cerium, higher order PCs were required to quantify titanium. At 2 PCs, the NN model achieved best results with RMSEP of 52 ppm at 73% accuracy. Testing the models on the training dataset shows that the validation procedure worked and does not overfit. The models were considered successful.

Titanium and cerium spectrally interfere. Titanium was prepared independent of cerium in all samples which would have made it easier for the training. However, the spectra were merged and PCA performed together. Successful prediction in this case means that the models were robust enough to quantify elements.

Table 5.9: Titanium models performance in starch matrix.

| | **NN** | **SVR** | **RF** |
|---|---|---|---|
| RMSEP_Test (ppm) | 47 | 52 | 41 |
| R_Sq_Test | 0.79 | 0.71 | 0.78 |
| LOD (ppm) | 11 | 15 | 7 |
| LOQ (ppm) | 38 | 49 | 24 |



(a)

(b)

(c)

Figure 5.26: Titanium (a) NN, (b) SVR, and (c) RF models on test dataset.

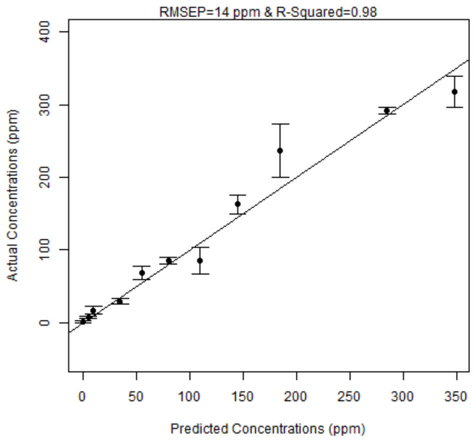
*5.4.2.5 Niobium*

Concentrations for Niobium models varied from 0 ~ 348 ppm. The number of PCs were varied from 1 to 9 PCs. The neural network performed best with an RMSEP of 14 ppm at 98% accuracy. RF and SVR models achieved RMSEP of 25 and 32 ppm at lower accuracies of 93% and 94% respectively. Table 5.10, Figure 5.27, and Table D.8 summarizes the modelling results using 9 PCs. According to the PCA score plot in Figure 5.18, PC1 explains 100% in the spectral data variability. PC1 successfully differentiates the various element containing (and blanks) in starch matrix. PC2 explains 0.0099% of spectral variability which raises the question of why 9 PCs ultimately showed good results while Yttrium only required 1 PC. Using 1 PC for niobium achieved RMSEP of 66 ppm at 24% accuracy while 2 PCs achieved RMSEP of 99 ppm at 29% accuracy. This affirms that even though PCA could separate the samples using 1 PC, it requires higher order PCs to successfully quantify elements.

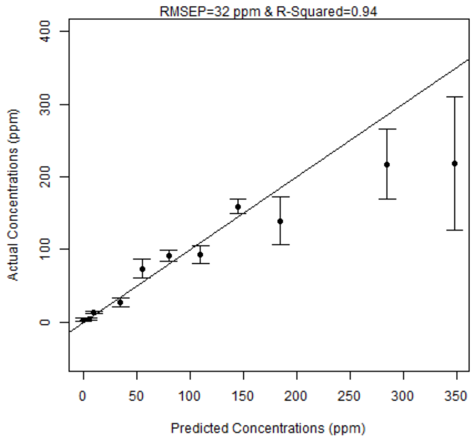
The models were considered successful at the concentrations involved. Niobium emission lines border those of Yttrium. There is no Yttrium in the samples of Niobium and vice versa, therefore spectral interference was not achieved from within the samples analyzed but from modelling which would potentially present challenges if mixed. Since Niobium and Yttrium occupy similar energy range, it implies that emission lines in this range would similarly preform using the instrument set-up.

Table 5.10. Niobium models performance in starch matrix.

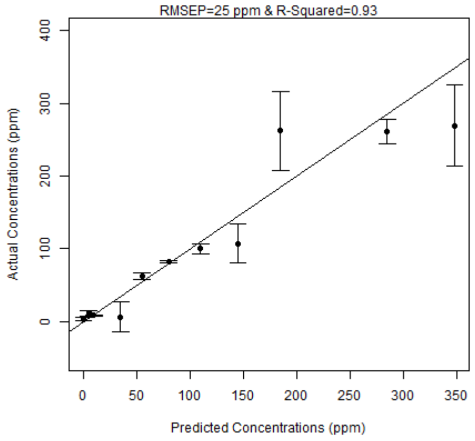
| | **NN** | **SVR** | **RF** |
|---|---|---|---|
| RMSEP_Test (ppm) | 14 | 32 | 25 |
| R_Sq_Test | 0.98 | 0.94 | 0.93 |
| LOD (ppm) | 4 | 8 | 4 |
| LOQ (ppm) | 13 | 26 | 13 |



(a)



(b)



(c)

Figure 5.27: Niobium (a) NN, (b) SVR, and (c) RF models on test dataset.

### 5.4.3 Summary Comparison with other Instruments

Techniques such as NAA and ICP-MS have very low detection limits. However, they involve lengthy sample preparation, have a high chance of contamination, and expensive. Table 5.11 presents a comparison of detection limits of REEs between NAA, other EDXRF studies, and the results of this study. The EDXRF system in this study performed very well in comparison.

Table 5.11: Comparison of results with other studies

| Element | EDXRF LoD (ppm) and context of study reference | NAA LoD (ppm) and context of study reference | EDXRF LoD (ppm) rock matrix in this study | EDXRF LoD (ppm) starch matrix in this study |
|---|---|---|---|---|
| Ce | 9 (Starch matrix using Lα lines (Schramm, 2016)) | 0.0014 (Bulska *et al.*, 2012) | 7 (Using Lα lines at low accuracy) | 6 (Using Lα lines) |
| Dy | 1 (Starch matrix using Lα lines (Schramm, 2016)) | 0.0014 (Bulska *et al.*, 2012) | 20 (Using Lα lines at low accuracy) | 6 (Using Lα lines) |
| Y | 46 (Thin sample analysis using K lines (Xiong *et al.*, 2020)) | 0.0036 (Dybczyński *et al.*, 2010) | 64 (Using K lines) | 34 (Using K lines) |

# CHAPTER 6
# CONCLUSIONS & RECOMMENDATIONS

## 6.1 Overview

This study aimed at designing a radioisotope excited x-ray fluorescence system based on Am-241 and further use chemometrics in energy dispersive x-ray fluorescence (EDXRF) spectroscopy to determine the quantity of rare earth elements (REEs) in geological samples. This chapter draws conclusions from the methods used and results. Recommendations for improving the equipment, methods, and results are also presented. Limitations to this study are also presented.

## 6.2 Conclusions

Building a unit to hold the detector, the shielding and the Am-241 radioisotope source was successful. Four main factors identifiable in instrument design were; sturdiness, shielding, additional parts, and intensity. The choice of timber used has proven to be very successful. The instrument has been used for over four years (2016-2020) and still holds its structure. There have been no weakened parts whatsoever. The steel housing also provides a rigid sample chamber that is firmly bolted to the wooden structure. The steel door has withstood many open and closing slide cycles and has held up relatively well. The door will need replacement once the sliding panels wear out. Alternatively, a hinged door design can be used in future iterations of the equipment.

Calibrating the radioisotope excited XRF system was performed successfully. Additional parts were required to operate the instrument safely and optimally. Aluminum used for the additional parts is cheap and readily available.  The sample holder was machined to precisely hold the sample at specific height while the detector-protector held the radioisotope in place away from the fragile detector window. Optimizing the instrument for intensity is crucial since the radioisotope source has a relatively low activity. The gamma ray paths to the sample and the x-ray path to the detector needs to be minimal. The annular source geometry has a radial gamma field that is intense at a particular height just above the source, and was determined by progressively increasing height based on different energies. The detector on the other hand is parallel to the sample. Since x-rays are emitted in all directions when gamma rays hit the sample, the closer the detector, the better the intensity of signal acquired.

Performing Principal Component Analysis and coming up with predictive models based on the chemometric techniques; Support Vector Regression (SVR), Artificial Neural Networks (ANN), and Random Forests that can quantify REEs in model matrices was successfully achieved. Spectral preprocessing, ratioing, and PCA were the key steps that improved the performance of models. Spectral preprocessing involved acquiring the spectra, compressing them, and selecting the region of interest. Ratioing involved dividing the ROI by Compton scatter regions. Ratioing in particular was responsible for much of the good results of chemometric modelling, underscoring the importance of information contained in Compton scatter peaks. Various analyses with and without ratioing clearly demonstrate this. PCA involved identifying PCs that could separate blanks from element-containing samples. 2 PCs in all matrices were sufficient in separating elements from blanks, however, more PCs were needed for element quantification. Varying the number of PCs uncovered the needs and inadequacies of randomness and reproducibility. Most models in two or more iterations of training do not necessarily yield the same results. This phenomenon is both desirable and undesirable. Desirable because in modelling with randomness, the models tend to yield better real-world results. Undesirable because reproducible models tend to overfit in the long run since training and testing data remain static.

Models were successfully created to predict REEs on test samples with known concentrations of REEs. NN and RF models presented better overall results than SVR. The models could satisfactorily predict all the elements in the matrices except cerium and dysprosium in rock matrix. Automatic functions helped in tuning the models thereby reducing time required to change variable that go into machine learning models. The “caret” package train function specifically reduced the time-consuming hyperparametric tuning. Faster modelling is desirable and can also be achieved, apart from PCA, by adopting general good computational techniques e.g., scaling of data.

This study was however conducted with limitations. The shielding was made of primarily lead (Pb) and steel. Lead blocks were pressed to get Lead sheets since they were unavailable and expensive to purchase. The results of dose rate measurements proved that the unit adheres to radiation protection requirements during operation. The unit can in-fact be used to store the radioisotope. The most challenging aspect in this study was the unavailability of REE salts. Cerium and Dysprosium, the REEs of interest in this study, were limited in quantity and were sparingly used. The results were therefore limited from a budgetary standpoint. Addition of titanium and niobium were an attempt at obtaining a comparable result. The detector used in this study operated in the energy range 0-26 keV. This energy range excludes the reliable K-lines of most REEs. The initial arrangement was to view the K-lines and therefore the setup was inherently limited in energy detection. Mixing the elements into one sample was a debatable arrangement especially for the lower concentration samples

## 6.3 Recommendations and Future Prospects

Instrument setup in this study was performed under uniform configuration conditions. The radioisotope source, detector, and sample height used were also constant. Spectrum acquisition time varied for primary elements of interest (cerium, dysprosium, and yttrium) at 240 seconds while additional elements (titanium and niobium) at 200 seconds. This presents an opportunity to vary the parameters and observe the effects.

Advantage of using the instrumental setup is the ability to swap the radioisotope source with other types of sources. The geometry of the excitation source should be annular, which many commercially available radioisotope shapes take. Similar to the excitation source, the detector can be swapped for other detector types. However, the height to bolt the detector below the sample does not have a dedicated rail which can allow moving the detector up and down. This can be a future iteration of the instrument setup. The excitation source can also have a dedicated winch to moves the radioisotope up and down. Such a combined rail-winch system can be used to observe the effects of geometry in detection efficiency and quantification.

Cerium and dysprosium in rock matrix particularly were separable (blanks and element containing) using the ratioing technique. However, their quantification was problematic. This can be attributed to chromium and iron that spectrally interfered. Additionally, computational time constraints of using higher order PCs hampered the analysis. Better hardware/software have the potential to overcome the computational time constraints. Higher concentration of analyte is another potential direction for further research.

Mass of sample (that determines thickness) and spectrum acquisition time were not included in the analysis. All models except cerium and dysprosium in rock performed satisfactorily, indicating that the models could harmonize different masses. This implies that the models have the potential to be used to determine thicknesses of samples. In fact, radioisotope excited XRF has been used for determination of coating thickness. The models could also harmonize the time of analysis, which implies that it would be possible to infer time of analysis in time-unlabeled spectra. A combination of mass and time using the experimental setup can be researched further.

This study mixed various elements in one sample but varying the concentrations in the samples. Cerium, dysprosium, and yttrium were mixed in similar samples while titanium and niobium were mixed into a second sample set. The correlation of sample concentrations was very low, <30%, between samples. This was an attempt at simulating matrix interference which in part succeeded. The combination of matrix spectra, rock and starch, can be an avenue for further study.

This study attempted to use machine learning to classify and cluster matrices. However not reported, these supervised and unsupervised learning techniques showed potential of using the setup in differentiating matrix types. This is an avenue for further research.

# REFERENCES

Adams, M.J. and Allen, J.R. (1998), Variable selection and multivariate calibration models for X-ray fluorescence spectrometry, *J. Anal. At. Spectrom.*, **Vol. 13 No. 2**, pp. 119–124.

Amptek. (2006), Periodic Table and X-Ray Emission Line Lookup Chart.

Andermann, G. and Kemp, J. (1958), Scattered X-Rays as Internal Standards in X-Ray Emission Spectroscopy, *Anal. Chem.*, **Vol. 30 No. 8**, pp. 1306–1309.

Andrade-Garda, J. (Ed.). (2009), *Basic Chemometric Techniques in Atomic Spectroscopy*, Royal Society of Chemistry, Cambridge, available at:https://doi.org/10.1039/9781847559661.

Angeyo, K.H., Gari, S., Mustapha, A.O. and Mangala, J.M. (2012), Feasibility for direct rapid energy dispersive X-ray fluorescence (EDXRF) and scattering analysis of complex matrix liquids by partial least squares, *Appl. Radiat. Isot.*, **Vol. 70 No. 11**, pp. 2596–2601.

Binge, F.W. and Joubert, P. (1966), The Mrima Hill Niobium Deposit, Coast Province, Kenya, *Minist. Nat. Resour. Mines, Geol. Dep.*, pp. 2–51.

Breiman, L. (2001), Random forests, *Mach. Learn.*, **Vol. 45 No. 1**, pp. 5–32.

Brereton, R.G. (2003), *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*, John Wiley & Sons, Chichester.

Bulska, E., Danko, B., Dybczyński, R.S., Krata, A., Kulisa, K., Samczyński, Z. and Wojciechowski, M. (2012), Inductively coupled plasma mass spectrometry in comparison with neutron activation and ion chromatography with UV/VIS detection for the determination of lanthanides in plant materials, *Talanta*, **Vol. 97**, pp. 303–311.

Burkhalter, P.G. (1971), Radioisotopic X-Ray Analysis of Silver Ores Using Compton Scatter for Matrix Compensation, *Anal. Chem.*, **Vol. 43 No. 1**, pp. 10–17.

Cabannes, V., Bach, F. and Rudi, A. (2021), Disambiguation of weak supervision with exponential convergence rates, available at: http://arxiv.org/abs/2102.02789.

Castor, S.B. and Hedrick, J.B. (2006), “Rare earth elements.” Industrial minerals volume, 7th edition: Society for mining, metallurgy, and exploration, Littleton, Colorado, *Ullmann’s Encycl. Ind. Chem.*, Vol. 7, Society for Mining, Metallurgy, and Exploration, Littleton, pp. 769–792.

Choppin, G., Liljenzin, J.O., Rydberg, J. and Ekberg, C. (2013), *Radiochemistry and Nuclear Chemistry: Fourth Edition, Radiochem. Nucl. Chem. Fourth Ed.*, Fourth., available at:https://doi.org/10.1016/C2011-0-07260-5.

Christopher, M.B. (2006), *Pattern Recognition and Machine Learning*, Springer.

Compton, A.H. (1923), A Quantum Theory of the Scattering of X-rays by Light Elements, *Phys. Rev.*, **Vol. 21 No. 5**, pp. 483–502.

Custo, G., Boeykens, S., Cicerone, D. and Vázquez, C. (2002), Combining XRF analysis and chemometric tools for a preliminary classification of argentine soils, *X-Ray Spectrom.*, **Vol. 31 No. 2**, pp. 132–135.

Durak, R. and Şahin, Y. (1997), Measurement of K-shell fluorescence yields for Ba, Ce, Nd, Gd, Dy, Er and Yb using radioisotope XRF, *Nucl. Instruments Methods Phys. Res. Sect. B Beam Interact. with Mater. Atoms*, **Vol. 124 No. 1**, pp. 1–4.

Durak, R. and Şahin, Y. (1998), Measurement of [Formula Presented]-shell fluorescence yields of selected elements from Cs to Pb using radioisotope x-ray fluorescence, *Phys. Rev. A - At. Mol. Opt. Phys.*, **Vol. 57 No. 4**, pp. 2578–2582.

Dybczyński, R.S., Czerska, E., Danko, B., Kulisa, K. and Samczyński, Z. (2010), Comparison of performance of INAA, RNAA and ion chromatography for the determination of individual lanthanides, *Appl. Radiat. Isot.*, **Vol. 68 No. 1**, pp. 23–27.

Enrich, C., Boeykens, S., Caracciolo, N., Custo, G. and Vázquez, C. (2007), Honey characterization by total reflection x-ray fluorescence: Evaluation of environmental quality and risk for the human health, *X-Ray Spectrom.*, **Vol. 36 No. 4**, pp. 215–220.

van Es, A., de Koeijer, J. and van der Peijl, G. (2009), Discrimination of document paper by XRF, LA-ICP-MS and IRMS using multivariate statistical techniques, *Sci. Justice*, **Vol. 49 No. 2**, pp. 120–126.

Ethem, A. (2010), *Introduction to Machine Learning, Second Edition*, Second., The MIT Press, London.

Facchin, I., Mello, C., Bueno, M.I.M.S. and Poppi, R.J. (1999), Simultaneous Determination of Lead and Sulfur by Energy-Dispersive X-Ray Spectrometry. Comparison between Artificial Neural Networks and Other Multivariate Calibration Methods, *X-Ray Spectrom.*, **Vol. 28 No. 3**, pp. 173–177.

Filzmoser, P. and Varmuza, K. (2016), Chemometrics: Multivariate Statistical Analysis in Chemometrics, available at: https://cran.r-project.org/package=chemometrics.

Gauglitz, G. and Tuan, V. (2003), *Handbook of Spectroscopy*, *Handb. Spectrosc.*, Wiley, available at:https://doi.org/10.1002/3527602305.

Glantz, S.A. and Slinker, B.K. (2001), *Primer of Applied Regression & Analysis of Variance, 2nd Edition*, McGraw-Hill.

Goldstein, S.J. and Sivils, L.D. (2002), A Non-Destructive X-Ray Fluorescence Method for Analysis of Metal Alloy Wire Samples, *Adv. X-Ray Anal.*, **Vol. 45**, pp. 457–462.

Goraieb, K., Alexandre, T.L. and Bueno, M.I.M.S. (2007), X-ray spectrometry and chemometrics in sugar classification, correlation with degree of sweetness and specific rotation of polarized light, *Anal. Chim. Acta*, **Vol. 595 No. 1-2 SPEC. ISS.**, pp. 170–175.

Goraieb, K., Lopes, A.S., Sato, C.A., Segatelli, M.G., Silva, V.P., Verzoto, J.C. and Bueno, M.I.M.S. (2006), Characterization of Portland cements by X-ray spectrometry allied to chemometrics, *J. Chemom.*, **Vol. 20 No. 11–12**, pp. 455–463.

Gordon, G.E., Randle, K., Goles, G.G., Corliss, J.B., Beeson, M.H. and Oxley, S.S.

(1968), Instrumental activation analysis of standard rocks with high-resolution γ-ray detectors, *Geochim. Cosmochim. Acta*, **Vol. 32 No. 4**, pp. 369–396.

Government of Kenya. (2015), *Kenya Mining: Investment Handbook 2015*, Ministry of Mining, available at: www.mining.go.ke.

Greenfield, A. and Graedel, T.E. (2013), The omnivorous diet of modern technology, *Resour. Conserv. Recycl.*, **Vol. 74**, pp. 1–7.

van Grieken, R.E. and Markowicz, A.A. (2002), *Handbook of X-Ray Spectrometry, Second Edition, Revised and Expanded*, Second Ed., Mercel Dekker Inc., New York and Basel.

Haque, N., Hughes, A., Lim, S. and Vernon, C. (2014), Rare earth elements: Overview of mining, mineralogy, uses, sustainability and environmental impact, *Resources*, **Vol. 3 No. 4**, pp. 614–635.

Henrich, A., Hoffmann, P., Ortner, H.M., Greve, T. and Itzel, H. (2000), Non-invasive identification of chemical compounds by energy dispersive X-ray fluorescence spectrometry, combined with chemometric methods of data evaluation, *Fresenius. J. Anal. Chem.*, **Vol. 368 No. 2–3**, pp. 130–138.

Houk, R.S. (1986), Mass Spectrometry of Inductively Coupled Plasmas, *Anal. Chem.*, **Vol. 58 No. 1**, pp. 97A-105A.

Hyndman, R.J. and Koehler, A.B. (2006), Another look at measures of forecast accuracy, *Int. J. Forecast.*, **Vol. 22 No. 4**, pp. 679–688.

IAEA. (1970), *TECHNICAL REPORTS SERIES No. 115: RADIOISOTOPE X-RAY FLUORESCENCE SPECTROMETRY*.

IAEA. (2007), *IAEA Nuclear Security Series No.5 - Identification of Radioactive Sources and Devices*, Vienna.

IAEA. (2018), *General Safety Guide No. GSG-7: Occupational Radiation Protection*, *IAEA Gen. Saf. Guid.*, Vienna.

ICRP. (1991), *1990 Recommendations of the International Commission on Radiological Protection,* edited by Smith, H., ICRP Publi., Pergamon Press, Oxford.

IUPAC. (1997), *Compendium of Chemical Terminology, 2nd Ed. (the "Gold Book"), XML on-Line Corrected Version (2006-), Blackwell Sci. Publ.*, available at: http://goldbook.iupac.org/PDF/goldbook.pdf.

James, E.T. (2007), *Atoms, Radiation, and Radiation Protection. Third, Completely Revised and Enlarged Edition.*, Third., Wiley.

Jenkins, R., Gould, R. and Gedcke, D. (1981), *Quantitative X-Ray Spectrometry*, Marcel Dekker, New York.

Jha, A.R. (2014), *Rare Earth Materials*, *Rare Earth Mater. Prop. Appl.*, CRC Press, available at:https://doi.org/10.1201/b17045.

Kaniu, M.I. and Angeyo, K.H. (2015), Challenges in rapid soil quality assessment and opportunities presented by multivariate chemometric energy dispersive X-ray fluorescence and scattering spectroscopy, *Geoderma*, **Vol. 241–242**, pp. 32–40.

Kaniu, M.I., Angeyo, K.H., Mangala, M.J., Mwala, A.K. and Bartilol, S.K. (2011), Feasibility for chemometric energy dispersive X-ray fluorescence and scattering (EDXRFS) spectroscopy method for rapid soil quality assessment, *X-Ray Spectrom.*, **Vol. 40 No. 6**, pp. 432–440.

Kaniu, M.I., Angeyo, K.H., Mwala, A.K. and Mangala, M.J. (2012), Direct rapid analysis of trace bioavailable soil macronutrients by chemometrics-assisted energy dispersive X-ray fluorescence and scattering spectrometry, *Anal. Chim. Acta*, **Vol. 729**, pp. 21–25.

Kaniu, M.I., Angeyo, K.H., Mwala, A.K. and Mwangi, F.K. (2012), Energy dispersive X-ray fluorescence and scattering assessment of soil quality via partial least squares and artificial neural networks analytical modeling approaches., *Talanta*, **Vol. 98**, pp. 236–40.

Kessler, T., Hoffmann, P., Greve, T. and Ortner, H.M. (2002), Optimization of the identification of chemical compounds by energy-dispersive x-ray fluorescence spectrometry and subsequent multivariate analysis, *X-Ray Spectrom.*, **Vol. 31 No. 5**, pp. 383–390.

KETEK. (2008), *DPP-Digital Pulse Processor Software Manual*, KETEK, GmBH, München.

Knoll, G.F. (2011), *Radiation Detection and Measurement, 4th Edition*, Fourth., Wiley.

Kvålseth, T.O. (1985), Cautionary Note about R 2, *Am. Stat.*, **Vol. 39 No. 4**, pp. 279–285.

Leclercq, L. and Meyers, R.A. (Eds.). (2006), *Encyclopedia of Analytical Chemistry*, Wiley, available at:https://doi.org/10.1002/9780470027318.

Liang, T., Li, K. and Wang, L. (2014), State of rare earth elements in different environmental components in mining areas of China, *Environ. Monit. Assess.*, **Vol. 186 No. 3**, pp. 1499–1513.

Lindon, J.C., Tranter, G.E. and Koppenaal, D.W. (2016), *Encyclopedia of Spectroscopy and Spectrometry*, edited by Lindon, J.*Encycl. Spectrosc. Spectrom.*, Second Edi., available at:https://doi.org/10.5860/choice.48-5433.

Lucas-Tooth, H.J. and Price, B.J. (1961), A mathematical method for the investigation of interelement effects in X-Ray fluorescence analysis, *Metallurgia*, **Vol. 64 No. 383**, pp. 149–161.

Markowicz, A. (1984), Theoretical evalution of the efficiency of compton scattered radiation method in EDXRF analysis, *X-Ray Spectrom.*, **Vol. 13 No. 4**, pp. 166–169.

Markowicz, A. (2011), An overview of quantificationmethods in energy-dispersive X-ray fluorescence analysis, *Pramana - J. Phys.*, **Vol. 76 No. 2**, pp. 321–329.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. and Leisch, F. (2014), e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-3. http://CRAN.R-project.org/package=e1071, *URL Http//CRAN.R-Project.Org/Package=e1071*, available at: https://cran.r-project.org/package=e1071.

Mikhailov, I.F., Baturin, A.A., Mikhailov, A.I., Borisova, S.S. and Fomina, L.P. (2018), Determination of coal ash content by the combined x-ray fluorescence and scattering spectrum, *Rev. Sci. Instrum.*, **Vol. 89 No. 2**, available at:https://doi.org/10.1063/1.4993101.

Mikhailov, I.F., Baturin, A.A., Mikhailov, A.I., Borisova, S.S. and Surovitskiy, S. V. (2020), Dependence of the Compton to Rayleigh intensity ratio on the scatterer atomic number in the range of 4(Be) to 31(Ga), *X-Ray Spectrom.*, **Vol. 49 No. 2**, pp. 284–290.

Mikhailov, I.F., Sobol, O. V., Varganov, V. V. and Fomina, L.P. (2002), Determination of mass fraction of light elements in crystalline materials by Compton-to Rayleigh scattering intensity ratio, *Funct. Mater.*, **Vol. 9 No. 4**, p. 651.

Miller, J.N. and Miller, J.C. (2010), *Statistics and Chemometrics for Analytical Chemistry*, Sixth., Pearson Education Limited, London.

Misra, P. and Dubinskii, M. (2002), *Ultraviolet Spectroscopy And Uv Lasers*, edited by Misra, P. and Dubinskii, M.A.*Ultrav. Spectrosc. Uv Lasers*, CRC Press, New York, available at:https://doi.org/10.1201/9780203908327.

Moseley, H.G.J. (1913), XCIII. The high-frequency spectra of the elements, *London, Edinburgh, Dublin Philos. Mag. J. Sci.*, **Vol. 26 No. 156**, pp. 1024–1034.

Myatt, G.J. and Johnson, W.P. (2008), *Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications*, *Mak. Sense Data II A Pract. Guid. to Data Vis. Adv. Data Min. Methods, Appl.*, John Wiley, New York, available at:https://doi.org/10.1002/9780470417409.

Nagata, N., Peralta-Zamora, P.G., Poppi, R.J., Perez, C.A. and Bueno, M.I.M.S. (2006), Multivariate calibrations for the SR-TXRF determination of trace concentrations of lead and arsenic in the presence of bromine, *X-Ray Spectrom.*, **Vol. 35 No. 1**, pp. 79–84.

NCRP. (1993), *Limitation of Exposure to Ionizing Radiation*, edited by Meinhold, C.B., Abrahamson, S., Adelstein, S.J., Bair, W.J., Fry, R.J.M., Boice, J.D., Hall, E.J., et al.*NCRP Rep.*, Bethseda, available at:https://doi.org/10.2307/3578722.

Okonda, J.J. (2015), *Development of Chemometrics Aided Energy Dispersive X-Ray Fluorescence and Scattering (EDXRFS) Method for Rapid Diagnostics of Cancer*, available at: http://erepository.uonbi.ac.ke.

Pechiney. (1971), The Beneficiation of the Mrima Hill Niobium and Rare Earths Deposits, Pechiney Saint Gobain.

Pollard, B. and Mapleson, D. (2013), Pacific Wildcat Resources Corp . NI 43 - 101 Technical Report for the Mrima Hill Niobium and Rare Earth Project, **No. 1**.

R-Core-Team. (2021), R: A language and environment for statistical computing. R Foundation for Statistical Computing, R Foundation, Vienna, available at: https://www.r-project.org/.

Rover, X.X. (2016), Amptek K and L Emission Line Lookup Chart X-Ray and Gamma Ray Detectors.

Savan, P. (2017), Chapter 2: SVM (Support Vector Machine) - Theory, in Machine Learning 101., available at: https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72 (accessed 3 May 2017).

Schimidt, F., Cornejo-Ponce, L., Bueno, M.I.M.S. and Poppi, R.J. (2003), Determination of some rare earth elements by EDXRF and artificial neural networks, *X-Ray Spectrom.*, **Vol. 32 No. 6**, pp. 423–427.

Schramm, R. (2016), Use of X-ray fluorescence analysis for the determination of rare earth elements, *Phys. Sci. Rev.*, **Vol. 1 No. 9**, available at:https://doi.org/10.1515/psr-2016-0061.

Shweta Bhatt. (2018), 5 Things You Need to Know about Reinforcement Learning, *KDnuggets*, available at: https://www.kdnuggets.com/2018/03/5-things-reinforcement-learning.html.

Singer, D.A. (1998), *Revised Grade and Tonnage Model of Carbonatite Deposits*, *United States Geol. Surv.*, United States Geological Survey.

Sitko, R. (2006), Correction of matrix effects via scattered radiation in X-ray fluorescence analysis of samples collected on membrane filters, *J. Anal. At. Spectrom.*, **Vol. 21 No. 10**, pp. 1062–1067.

Stefan, F., Frauke, G. and Marvin, N.W. (2019), neuralnet: Training of Neural Networks.

Stuart, J.R. and Peter, N. (2010), *Artificial Intelligence: A Modern Approach*, Third Edit., Prentice Hall.

Szökefalvi-Nagy, Z., Demeter, I., Kocsonya, A. and Kovács, I. (2004), Non-destructive XRF analysis of paintings, *Nucl. Instruments Methods Phys. Res. Sect. B Beam Interact. with Mater. Atoms*, **Vol. 226 No. 1–2**, pp. 53–59.

Szumigala, D.J. and Werdon, M.B. (2010), *A Brief Overview Including Uses , Worldwide Resources , and Known Occurrences in Alaska*.

Tertian, R. and Claisse, F. (1982), *Principles of Quantitative X-Ray Fluorescence Analysis.*, *Princ. Quant. X-Ray Fluoresc. Anal.*, Wiley-Heyden, London, available at:https://doi.org/10.1016/0016-7037(83)90078-9.

Theil, H. (1961), *Economic Forecasts and Policy*, Second Edi., Amsterdam, North-Holland Pub. Co.

Traore, A., Ndiaye, P.M., Mbaye, M., Diatta, F. and Wague, A. (2014), X-ray fluorescence combined with chemometrics for the characterization of geological samples: A case study in southeastern senegal, *Instrum. Sci. Technol.*, **Vol. 42 No. 5**, pp. 593–604.

Vandegrift, G. (2015), Compton Scattering SVG, Wikimedia, available at: https://commons.wikimedia.org/wiki/File:Compton-scattering.svg (accessed 9 June 2021).

Vázquez, C., Boeykens, S. and Bonadeo, H. (2002), Total reflection X-ray fluorescence polymer spectra: Classification by taxonomy statistic tools, *Talanta*, **Vol. 57 No. 6**, pp. 1113–1117.

Verbi, F.M., Pereira-Filho, E.R. and Bueno, M.I.M.S. (2005), Use of X-ray scattering for studies with organic compounds: A case study using paints, *Microchim. Acta*, **Vol. 150 No. 2**, pp. 131–136.

Wilkinson, D.H. (1996), The Geiger discharge revisited part III. Convergence, *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*, **Vol. 383 No. 2–3**, pp. 523–527.

WU, W., XU, T., HAO, Q., WANG, Q., ZHANG, S. and ZHAO, C. (2010), Applications of X-ray fluorescence analysis of rare earths in China, *J. Rare Earths*, **Vol. 28**, pp. 30–36.

Xiong, G., Jia, W., Shan, Q., Zhang, X., Tang, X. and Li, J. (2020), Equipment design and performance characterization of full field x-ray fluorescence (FF-XRF) element distribution imaging system with combined collimating lens (CCL), *Rev. Sci. Instrum.*, **Vol. 91 No. 12**, available at:https://doi.org/10.1063/5.0024461.

YAP, C.T. (1986), Xrf Analysis of Nonya Wares Using an Annular Americium Source, *Archaeometry*, **Vol. 28 No. 2**, pp. 197–201.

YAP, C.T. and TANG, S.M. (1985), ENERGY-DISPERSIVE X-RAY FLUORESCENCE ANALYSIS OF CHINESE PORCELAINS USING Am-241, *Archaeometry*, **Vol. 27 No. 1**, pp. 61–63.

Zawisza, B., Pytlakowska, K., Feist, B., Polowniak, M., Kita, A. and Sitko, R. (2011), Determination of rare earth elements by spectroscopic techniques: A review, *J. Anal. At. Spectrom.*, **Vol. 26 No. 12**, pp. 2373–2390.

----------------------------------------------------------------------------------------------------

# APPENDIX A : ELEMENT CONCENTRATIONS

Table A.1. Rock matrix scheme of concentrations (ppm) and masses (mg)

| Sample | Ce | Dy | Y | Nb | Ti | Mass |
|---|---|---|---|---|---|---|
| sample6_A.asc | 16 | 9 | 18275 | 0 | 0 | 524.3 |
| sample6_B.asc | 16 | 9 | 18275 | 0 | 0 | 788.9 |
| sample6_C.asc | 16 | 9 | 18275 | 0 | 0 | 1226.9 |
| sample6_D.asc | 16 | 9 | 18275 | 0 | 0 | 1409.3 |
| sample3_A.asc | 12 | 10 | 18254 | 0 | 0 | 587.6 |
| sample3_B.asc | 12 | 10 | 18254 | 0 | 0 | 866.2 |
| sample3_C.asc | 12 | 10 | 18254 | 0 | 0 | 1165.1 |
| sample3_D.asc | 12 | 10 | 18254 | 0 | 0 | 1591.0 |
| sample1_A.asc | 365 | 7 | 11843 | 0 | 0 | 2238.2 |
| sample1_B.asc | 365 | 7 | 11843 | 0 | 0 | 2176.0 |
| sample1_C.asc | 365 | 7 | 11843 | 0 | 0 | 483.3 |
| sample1_D.asc | 365 | 7 | 11843 | 0 | 0 | 800.4 |
| sample1_E.asc | 365 | 7 | 11843 | 0 | 0 | 1184.7 |
| sample1_F.asc | 365 | 7 | 11843 | 0 | 0 | 1817.5 |
| sample11_A.asc | 8 | 33 | 9483 | 0 | 0 | 502.9 |
| sample11_B.asc | 8 | 33 | 9483 | 0 | 0 | 791.9 |
| sample11_C.asc | 8 | 33 | 9483 | 0 | 0 | 1142.7 |
| sample11_D.asc | 8 | 33 | 9483 | 0 | 0 | 1966.7 |

| **Sample** | **Ce** | **Dy** | **Y** | **Nb** | **Ti** | **Mass** |
|---|---|---|---|---|---|---|
| sample14_A.asc | 56 | 14 | 9296 | 0 | 0 | 570.6 |
| sample14_B.asc | 56 | 14 | 9296 | 0 | 0 | 888.2 |
| sample14_C.asc | 56 | 14 | 9296 | 0 | 0 | 1162.6 |
| sample14_D.asc | 56 | 14 | 9296 | 0 | 0 | 1555.0 |
| sample15_A.asc | 99 | 102 | 4858 | 0 | 0 | 533.5 |
| sample15_B.asc | 99 | 102 | 4858 | 0 | 0 | 833.9 |
| sample15_C.asc | 99 | 102 | 4858 | 0 | 0 | 1140.8 |
| sample15_D.asc | 99 | 102 | 4858 | 0 | 0 | 1763.3 |
| sample4_A.asc | 19 | 54 | 3562 | 0 | 0 | 532.8 |
| sample4_B.asc | 19 | 54 | 3562 | 0 | 0 | 762.0 |
| sample4_C.asc | 19 | 54 | 3562 | 0 | 0 | 1195.2 |
| sample4_D.asc | 19 | 54 | 3562 | 0 | 0 | 1496.2 |
| sample2_A.asc | 8 | 380 | 2397 | 0 | 0 | 643.5 |
| sample2_B.asc | 8 | 380 | 2397 | 0 | 0 | 1069.3 |
| sample2_C.asc | 8 | 380 | 2397 | 0 | 0 | 831.8 |
| sample2_D.asc | 8 | 380 | 2397 | 0 | 0 | 1698.5 |
| sample7_A.asc | 60 | 19 | 1577 | 0 | 0 | 482.7 |
| sample7_B.asc | 60 | 19 | 1577 | 0 | 0 | 832.0 |
| sample7_C.asc | 60 | 19 | 1577 | 0 | 0 | 1158.9 |
| sample7_D.asc | 60 | 19 | 1577 | 0 | 0 | 1843.4 |

| Sample | Ce | Dy | Y | Nb | Ti | Mass |
|---|---|---|---|---|---|---|
| sample10_A.asc | 31 | 186 | 1151 | 0 | 0 | 521.6 |
| sample10_B.asc | 31 | 186 | 1151 | 0 | 0 | 1012.7 |
| sample10_C.asc | 31 | 186 | 1151 | 0 | 0 | 1265.8 |
| sample10_D.asc | 31 | 186 | 1151 | 0 | 0 | 1559.5 |
| sample12_A.asc | 184 | 12 | 761 | 0 | 0 | 518.7 |
| sample12_B.asc | 184 | 12 | 761 | 0 | 0 | 860.0 |
| sample12_C.asc | 184 | 12 | 761 | 0 | 0 | 1189.1 |
| sample12_D.asc | 184 | 12 | 761 | 0 | 0 | 1870.7 |
| sample5_A.asc | 235 | 361 | 598 | 0 | 0 | 547.4 |
| sample5_B.asc | 235 | 361 | 598 | 0 | 0 | 833.5 |
| sample5_C.asc | 235 | 361 | 598 | 0 | 0 | 1317.9 |
| sample5_D.asc | 235 | 361 | 598 | 0 | 0 | 1684.5 |
| sample13_A.asc | 31 | 191 | 405 | 0 | 0 | 522.9 |
| sample13_B.asc | 31 | 191 | 405 | 0 | 0 | 809.3 |
| sample13_C.asc | 31 | 191 | 405 | 0 | 0 | 1340.3 |
| sample13_D.asc | 31 | 191 | 405 | 0 | 0 | 1643.2 |
| Sample-1.asc | 0 | 0 | 0 | 0 | 0 | 354.8 |
| Sample-10.asc | 0 | 0 | 0 | 0 | 0 | 880.7 |
| Sample-11.asc | 0 | 0 | 0 | 0 | 0 | 977.9 |
| Sample-12.asc | 0 | 0 | 0 | 0 | 0 | 1069.3 |

| **Sample** | **Ce** | **Dy** | **Y** | **Nb** | **Ti** | **Mass** |
|---|---|---|---|---|---|---|
| Sample-13.asc | 0 | 0 | 0 | 0 | 0 | 1140.0 |
| Sample-14.asc | 0 | 0 | 0 | 0 | 0 | 1270.0 |
| Sample-15.asc | 0 | 0 | 0 | 0 | 0 | 1370.5 |
| Sample-16.asc | 0 | 0 | 0 | 0 | 0 | 1356.6 |
| Sample-17.asc | 0 | 0 | 0 | 0 | 0 | 1472.0 |
| Sample-18.asc | 0 | 0 | 0 | 0 | 0 | 1581.5 |
| Sample-19.asc | 0 | 0 | 0 | 0 | 0 | 1723.1 |
| Sample-2.asc | 0 | 0 | 0 | 0 | 0 | 375.8 |
| Sample-20.asc | 0 | 0 | 0 | 0 | 0 | 1799.3 |
| Sample-21.asc | 0 | 0 | 0 | 0 | 0 | 1935.8 |
| Sample-22.asc | 0 | 0 | 0 | 0 | 0 | 2004.9 |
| Sample-3.asc | 0 | 0 | 0 | 0 | 0 | 484.8 |
| Sample-4.asc | 0 | 0 | 0 | 0 | 0 | 500.9 |
| Sample-5.asc | 0 | 0 | 0 | 0 | 0 | 570.7 |
| Sample-6.asc | 0 | 0 | 0 | 0 | 0 | 402.4 |
| Sample-7.asc | 0 | 0 | 0 | 0 | 0 | 472.0 |
| Sample-8.asc | 0 | 0 | 0 | 0 | 0 | 677.6 |
| Sample-9.asc | 0 | 0 | 0 | 0 | 0 | 495.7 |
| sample16_A.asc | 0 | 0 | 0 | 0 | 0 | 564.4 |
| sample16_B.asc | 0 | 0 | 0 | 0 | 0 | 806.4 |

| Sample | Ce | Dy | Y | Nb | Ti | Mass |
|---|---|---|---|---|---|---|
| sample16_C.asc | 0 | 0 | 0 | 0 | 0 | 1252.7 |
| sample16_D.asc | 0 | 0 | 0 | 0 | 0 | 1644.2 |

Table A.2. Starch matrix scheme of concentrations (ppm) and masses (mg)

| Sample | Ce | Dy | Y | Nb | Ti | Mass |
|---|---|---|---|---|---|---|
| sample3A.asc | 8 | 7 | 18390 | 0 | 0 | 495.2 |
| sample3B.asc | 8 | 7 | 18390 | 0 | 0 | 996.7 |
| sample3C.asc | 8 | 7 | 18390 | 0 | 0 | 1559.0 |
| sample3D.asc | 8 | 7 | 18390 | 0 | 0 | 2245.6 |
| sample6A.asc | 17 | 10 | 18309 | 0 | 0 | 523.4 |
| sample6B.asc | 17 | 10 | 18309 | 0 | 0 | 1017.0 |
| sample6C.asc | 17 | 10 | 18309 | 0 | 0 | 1456.3 |
| sample6D.asc | 17 | 10 | 18309 | 0 | 0 | 2257.2 |
| sample1A.asc | 378 | 9 | 11939 | 0 | 0 | 528.2 |
| sample1B.asc | 378 | 9 | 11939 | 0 | 0 | 507.0 |
| sample1C.asc | 378 | 9 | 11939 | 0 | 0 | 912.5 |
| sample1D.asc | 378 | 9 | 11939 | 0 | 0 | 1274.0 |
| sample1E.asc | 378 | 9 | 11939 | 0 | 0 | 1920.9 |
| sample14A.asc | 45 | 9 | 9559 | 0 | 0 | 454.2 |
| sample14B.asc | 45 | 9 | 9559 | 0 | 0 | 1006.4 |

| Sample | Ce | Dy | Y | Nb | Ti | Mass |
|---|---|---|---|---|---|---|
| sample14C.asc | 45 | 9 | 9559 | 0 | 0 | 1472.6 |
| sample14D.asc | 45 | 9 | 9559 | 0 | 0 | 2048.2 |
| sample8A.asc | 23 | 186 | 9376 | 0 | 0 | 526.1 |
| sample8B.asc | 23 | 186 | 9376 | 0 | 0 | 955.3 |
| sample8C.asc | 23 | 186 | 9376 | 0 | 0 | 1463.4 |
| sample8D.asc | 23 | 186 | 9376 | 0 | 0 | 2077.7 |
| sample11A.asc | 4 | 28 | 9349 | 0 | 0 | 534.4 |
| sample11B.asc | 4 | 28 | 9349 | 0 | 0 | 1032.2 |
| sample11C.asc | 4 | 28 | 9349 | 0 | 0 | 1588.0 |
| sample11D.asc | 4 | 28 | 9349 | 0 | 0 | 1897.1 |
| sample15A.asc | 109 | 96 | 4970 | 0 | 0 | 466.7 |
| sample15B.asc | 109 | 96 | 4970 | 0 | 0 | 1073.5 |
| sample15C.asc | 109 | 96 | 4970 | 0 | 0 | 1550.0 |
| sample15D.asc | 109 | 96 | 4970 | 0 | 0 | 1846.4 |
| sample4A.asc | 14 | 45 | 3688 | 0 | 0 | 514.3 |
| sample4B.asc | 14 | 45 | 3688 | 0 | 0 | 805.0 |
| sample4C.asc | 14 | 45 | 3688 | 0 | 0 | 1131.0 |
| sample4D.asc | 14 | 45 | 3688 | 0 | 0 | 1711.2 |
| sample2A.asc | 17 | 370 | 2324 | 0 | 0 | 595.2 |
| sample2B.asc | 17 | 370 | 2324 | 0 | 0 | 1095.8 |

| **Sample** | **Ce** | **Dy** | **Y** | **Nb** | **Ti** | **Mass** |
|---|---|---|---|---|---|---|
| sample2C.asc | 17 | 370 | 2324 | 0 | 0 | 1479.1 |
| sample2D.asc | 17 | 370 | 2324 | 0 | 0 | 1775.8 |
| sample7A.asc | 74 | 21 | 1504 | 0 | 0 | 568.7 |
| sample7B.asc | 74 | 21 | 1504 | 0 | 0 | 1032.0 |
| sample7C.asc | 74 | 21 | 1504 | 0 | 0 | 1509.6 |
| sample7D.asc | 74 | 21 | 1504 | 0 | 0 | 1684.9 |
| sample10A.asc | 39 | 188 | 1070 | 0 | 0 | 559.9 |
| sample10B.asc | 39 | 188 | 1070 | 0 | 0 | 1026.6 |
| sample10C.asc | 39 | 188 | 1070 | 0 | 0 | 1346.1 |
| sample10D.asc | 39 | 188 | 1070 | 0 | 0 | 1938.4 |
| sample12A.asc | 192 | 10 | 650 | 0 | 0 | 465.3 |
| sample12B.asc | 192 | 10 | 650 | 0 | 0 | 1101.0 |
| sample12C.asc | 192 | 10 | 650 | 0 | 0 | 1418.2 |
| sample12D.asc | 192 | 10 | 650 | 0 | 0 | 1856.3 |
| sample5A.asc | 241 | 396 | 463 | 0 | 0 | 531.9 |
| sample5B.asc | 241 | 396 | 463 | 0 | 0 | 1023.3 |
| sample5C.asc | 241 | 396 | 463 | 0 | 0 | 1509.1 |
| sample5D.asc | 241 | 396 | 463 | 0 | 0 | 1928.6 |
| sample9A.asc | 188 | 21 | 439 | 0 | 0 | 536.7 |
| sample9B.asc | 188 | 21 | 439 | 0 | 0 | 1074.1 |

| **Sample** | **Ce** | **Dy** | **Y** | **Nb** | **Ti** | **Mass** |
|---|---|---|---|---|---|---|
| sample9C.asc | 188 | 21 | 439 | 0 | 0 | 1470.2 |
| sample9D.asc | 188 | 21 | 439 | 0 | 0 | 1766.7 |
| sample13A.asc | 33 | 214 | 434 | 0 | 0 | 482.6 |
| sample13B.asc | 33 | 214 | 434 | 0 | 0 | 1065.3 |
| sample13C.asc | 33 | 214 | 434 | 0 | 0 | 1602.5 |
| sample13D.asc | 33 | 214 | 434 | 0 | 0 | 1720.0 |
| sample_10A.asc | 0 | 0 | 372 | 0 | 0 | 434.2 |
| sample_10B.asc | 0 | 0 | 372 | 0 | 0 | 960.8 |
| sample_9A.asc | 0 | 0 | 303 | 0 | 0 | 439.0 |
| sample_9B.asc | 0 | 0 | 303 | 0 | 0 | 980.6 |
| sample_8A.asc | 0 | 0 | 243 | 0 | 0 | 457.9 |
| sample_8B.asc | 0 | 0 | 243 | 0 | 0 | 941.2 |
| sample_7A.asc | 0 | 0 | 170 | 0 | 0 | 540.2 |
| sample_7B.asc | 0 | 0 | 170 | 0 | 0 | 852.6 |
| sample_6A.asc | 0 | 0 | 127 | 0 | 0 | 476.5 |
| sample_6B.asc | 0 | 0 | 127 | 0 | 0 | 865.3 |
| sample_5A.asc | 0 | 0 | 79 | 0 | 0 | 455.1 |
| sample_5B.asc | 0 | 0 | 79 | 0 | 0 | 921.3 |
| sample_4A.asc | 0 | 0 | 54 | 0 | 0 | 467.9 |
| sample_4B.asc | 0 | 0 | 54 | 0 | 0 | 849.2 |

| **Sample** | **Ce** | **Dy** | **Y** | **Nb** | **Ti** | **Mass** |
|---|---|---|---|---|---|---|
| sample_3A.asc | 0 | 0 | 29 | 0 | 0 | 817.3 |
| sample_2A.asc | 0 | 0 | 17 | 0 | 0 | 473.9 |
| sample_2B.asc | 0 | 0 | 17 | 0 | 0 | 876.3 |
| sample_1A.asc | 0 | 0 | 5 | 0 | 0 | 437.7 |
| sample_1B.asc | 0 | 0 | 5 | 0 | 0 | 910.4 |
| sample-10A.asc | 0 | 0 | 0 | 235 | 5 | 593.0 |
| sample-11A.asc | 0 | 0 | 0 | 285 | 55 | 506.3 |
| sample-11B.asc | 0 | 0 | 0 | 285 | 55 | 543.4 |
| sample-11C.asc | 0 | 0 | 0 | 285 | 55 | 1109.0 |
| sample-12A.asc | 0 | 0 | 0 | 348 | 111 | 254.5 |
| sample-12C.asc | 0 | 0 | 0 | 348 | 111 | 1246.4 |
| sample-12D.asc | 0 | 0 | 0 | 348 | 111 | 2819.9 |
| sample-1A.asc | 0 | 0 | 0 | 5 | 83 | 470.3 |
| sample-1B.asc | 0 | 0 | 0 | 5 | 83 | 819.9 |
| sample-1C.asc | 0 | 0 | 0 | 5 | 83 | 1302.0 |
| sample-1D.asc | 0 | 0 | 0 | 5 | 83 | 2152.5 |
| sample-2A.asc | 0 | 0 | 0 | 10 | 146 | 617.9 |
| sample-2B.asc | 0 | 0 | 0 | 10 | 146 | 640.9 |
| sample-2C.asc | 0 | 0 | 0 | 10 | 146 | 1064.7 |
| sample-2D.asc | 0 | 0 | 0 | 10 | 146 | 2539.8 |

| **Sample** | **Ce** | **Dy** | **Y** | **Nb** | **Ti** | **Mass** |
|---|---|---|---|---|---|---|
| sample-3B.asc | 0 | 0 | 0 | 20 | 35 | 526.3 |
| sample-3C.asc | 0 | 0 | 0 | 20 | 35 | 955.7 |
| sample-3D.asc | 0 | 0 | 0 | 20 | 35 | 2890.7 |
| sample-4A.asc | 0 | 0 | 0 | 35 | 349 | 498.0 |
| sample-4B.asc | 0 | 0 | 0 | 35 | 349 | 649.3 |
| sample-4C.asc | 0 | 0 | 0 | 35 | 349 | 1072.9 |
| sample-4D.asc | 0 | 0 | 0 | 35 | 349 | 2585.3 |
| sample-5A.asc | 0 | 0 | 0 | 55 | 185 | 483.7 |
| sample-5B.asc | 0 | 0 | 0 | 55 | 185 | 660.6 |
| sample-5C.asc | 0 | 0 | 0 | 55 | 185 | 1090.4 |
| sample-5D.asc | 0 | 0 | 0 | 55 | 185 | 2587.2 |
| sample-6A.asc | 0 | 0 | 0 | 80 | 20 | 242.3 |
| sample-6B.asc | 0 | 0 | 0 | 80 | 20 | 688.4 |
| sample-6C.asc | 0 | 0 | 0 | 80 | 20 | 1123.5 |
| sample-7A.asc | 0 | 0 | 0 | 110 | 10 | 424.9 |
| sample-7B.asc | 0 | 0 | 0 | 110 | 10 | 582.5 |
| sample-7C.asc | 0 | 0 | 0 | 110 | 10 | 990.9 |
| sample-7D.asc | 0 | 0 | 0 | 110 | 10 | 2850.6 |
| sample-8B.asc | 0 | 0 | 0 | 145 | 236 | 591.0 |
| sample-8C.asc | 0 | 0 | 0 | 145 | 236 | 977.4 |

| **Sample** | **Ce** | **Dy** | **Y** | **Nb** | **Ti** | **Mass** |
|---|---|---|---|---|---|---|
| sample-9A.asc | 0 | 0 | 0 | 185 | 286 | 332.5 |
| sample-9B.asc | 0 | 0 | 0 | 185 | 286 | 514.0 |
| sample-9C.asc | 0 | 0 | 0 | 185 | 286 | 940.8 |
| sample-9D.asc | 0 | 0 | 0 | 185 | 286 | 3038.9 |
| sample_1446_4mg.asc | 0 | 0 | 0 | 0 | 0 | 1446.0 |
| sample_2173_5mg.asc | 0 | 0 | 0 | 0 | 0 | 2173.0 |
| sample_3055_0mg.asc | 0 | 0 | 0 | 0 | 0 | 3055.0 |
| sample_4649_6mg.asc | 0 | 0 | 0 | 0 | 0 | 4649.0 |
| sample_612_9mg.asc | 0 | 0 | 0 | 0 | 0 | 612.0 |
| sample_985_9mg.asc | 0 | 0 | 0 | 0 | 0 | 985.0 |
| sample16A.asc | 0 | 0 | 0 | 0 | 0 | 498.8 |
| sample16B.asc | 0 | 0 | 0 | 0 | 0 | 911.1 |
| sample16C.asc | 0 | 0 | 0 | 0 | 0 | 1663.5 |

# APPENDIX B : SPECTRAL PREPROCESSING CODES

```
#This function is executed by passing in a directory with'.asc', '.mcd', and 'scheme.csv' files
#In R, ‘/’ is used instead of ‘\’ to specify directory path
#Alternatively, this document can be knitted to 'Word', '.pdf' or '.html' from the directory it is located in.
#This Markdown is meant specifically for the rock matrix. For the starch matrix, its' Markdown file is located in that directory. It can thus be opened and executed-‘knitted’ from that directory.

#This step is only necessary if this code is executed as a script, it is therefore commented out. Care has to be taken with directory path as indicated in above comment.
#setwd("~Thesis/Data_Files_Processors/Rock_Matrix")

#-----------------------'.mcd' (metadata) processor function-------------
  #This function extracts data from the metadata file
  mcd <- function(){
    #List files in directory with '.mcd' extension
    files <- as.data.frame(list.files(pattern="\\.mcd$"))
    #Read the first file, it has unequal columns
    y <- as.data.frame(read.table(as.character(files[1,1]), sep = "=", fill = T))

    #Initiate a new dataframe to receive extraction metadata
    #At the same time, rename the file. This is useful for merging with spectral data.
    x <- as.data.frame(as.character(gsub(".mcd", ".asc", as.character(files[1,1]))))
    #Rename the column name of the receiving dataframe (x)
    colnames(x) <- "Name"
    #Create REALTIME column in (x) and save REALTIME data from the metadata file
    x$REALTIME <- as.numeric(gsub("REALTIME: ","",y[2,1]))
```

```r
    #Repeat the process for LIVETIME, TOTALSUM, ROISUM, NETTOSUM, and MAXVAL
    x$LIVETIME <- as.numeric(gsub("LIVETIME: ","",y[3,1]))
    x$TOTALSUM <- as.numeric(gsub("TOTALSUM: ","",y[4,1]))
    x$ROISUM <- as.numeric(gsub("ROISUM: ","",y[5,1]))
    x$NETTOSUM <- as.numeric(gsub("","",y[7,1]))
    x$MAXVAL <- as.numeric(gsub("MAXVAL: ","",y[8,1]))

    #Initiate a  'while loop' to extract metadata from all other files while updating the dataframe (x)
    n <- 2

    #Loop through all '.mcd' files in the directory
    while (n<=as.numeric(length(files[,1]))) {
      #Read the 'Nth' file in the loop, it has unequal columns
      y <- as.data.frame(read.table(as.character(files[n,1]), sep = "=", fill = T))

      #This process is repeated as before
      #Initiate a new dataframe (m) to receive extraction metadata
      #At the same time, rename the file. This is useful for merging with spectral data.
      m <- as.data.frame(as.character(gsub(".mcd", ".asc", as.character(files[n,1]))))
      colnames(m) <- "Name"
      m$REALTIME <- as.numeric(gsub("REALTIME: ","",y[2,1]))
      m$LIVETIME <- as.numeric(gsub("LIVETIME: ","",y[3,1]))
      m$TOTALSUM <- as.numeric(gsub("TOTALSUM: ","",y[4,1]))
      m$ROISUM <- as.numeric(gsub("ROISUM: ","",y[5,1]))
      m$NETTOSUM <- as.numeric(gsub("","",y[7,1]))
      m$MAXVAL <- as.numeric(gsub("MAXVAL: ","",y[8,1]))

      #Update the initial dataframe (x) with the 'Nth' file metadata
      x <- rbind(x,m)
      #Increment loop
      n <- n + 1
    }
    #Return the metadata dataframe
    return(x)
  }
#--------------------------end of metadata function-----------------------

#-------------------------asc (spectrum data) function-----------------------
  #This function extracts data from the spectrum file
  asc <- function(){
    #List files in directory with '.asc' extension
    files <- as.data.frame(list.files(pattern = "\\.asc$"))
    #Read the first file
    y <- read.table(as.character(files[1,1]))
    #Extract entire spectrum, this is the entire range of counting of the detector
    y <- as.data.frame(y[312:6911,])

    #Sum the first 8 entries and save to a dataframe
    x <- as.data.frame(sum(y[1:8,]))
    #Rename the column
    colnames(x) <- "V1"

    #Inititate a while loop starting at the 8+1 entry
    n <- 9
    #Loop through the entire dataframe length
    while (n <= nrow(y)) {
      #Write to a variable the ending value (the dataframe respects data entry)
      g <- n+7
      #Sum the values between 'Nth' entry and the 'last entry'-(value of variable ,g) and store in a dataframe
      z <- as.data.frame(sum(y[n:g,]))
      #Rename the dataframe to match the first (x) dataframe
      colnames(z) <- "V1"
```

```
   #Bind below the initial dataframe entry with the new dataframe entry
   x <- rbind(x,z)
   #Increment loop
   n <- n+8
  }

  #Overwrite y with the first averaged (by summing) file (x)
  y <- x
  #Rename the column with the name of the spectrum
  colnames(y) <- as.character(files[1,1])

  #Initiate while loop for the second spectral file
  num <- 2

  #Loop through the number of spectral files in the directory
  while(num <= as.numeric(length(files[,1])))
  {
   #Read the 'Nth' spectrum file to a dataframe (m)
   m <- read.table(as.character(files[num,1]));
   #Extract entire spectrum, this is the entire range of counting of the detector
   m <- as.data.frame(m[312:6911,])

   #PERFORM SIMILAR OPERATION TO AS IN THE FIRST SPECTRAL FILE
   x <- as.data.frame(sum(m[1:8,]))
   colnames(x) <- "V1"

   n <- 9
   while (n <= nrow(m)) {
    g <- n+7
    z <- as.data.frame(sum(m[n:g,]))
    colnames(z) <- "V1"
    x <- rbind(x,z)
    n <- n+8
   }
```

```
      #Overwrite (m) with contents of (x)
      m <- x
      #Rename column to the 'Nth' specral file
      colnames(m) <- as.character(files[num,1]);
      #Bind the 'Nth' dataframe to the side of the initiated dataframe
(y)
      y <- cbind(y,m);
      #Increment loop
      num <- num + 1;
    }
    #Transpose dataframe
    y <- as.data.frame(t(y))
    #Add a new column to hold the names of the dataframe
    y$Name <- rownames(y)
    #Return merged spectra dataframe
    return(y)
  }
#----------------------------end of asc (spectral) function------------
----

  #Call the functions
  y <- asc()
  x <- mcd()
  #Merge metadata with spectral data by their respective file names
  y <- merge(x, y, by="Name")

#----------File generators-----CAREFUL HERE, use folder (Starch or Rock
) specific names--------

  #Read in 'scheme.csv' file containing the Name of file, concentration
 of elements, and pellet mass
  dat <- read.csv("scheme.csv")
  #Extract data for Yttrium (Name, concentration, and Mass)
  Y <- as.data.frame(cbind(as.data.frame(dat$Name),as.data.frame(dat$Y)
, as.data.frame(dat$Mass)))
  #Rename columns
  colnames(Y) <- c("Name", "Y","Mass")
  #Merge 'scheme' data for Yttrium with spectra and metadata by "Name"
variable
  Y <- merge(Y, y, by="Name")
  #Write out a '.csv' file to the folder
  write.csv(Y, file = "Y-Ro.csv", row.names = FALSE, eol = "\r")

  #Repeat the same steps for Cerium, Dysprosium, Titanium, and Niobium
  Ce <- as.data.frame(cbind(as.data.frame(dat$Name),as.data.frame(dat$C
e), as.data.frame(dat$Mass)))
  colnames(Ce) <- c("Name", "Ce","Mass")
  Ce <- merge(Ce, y, by="Name")
  write.csv(Ce, file = "Ce-Ro.csv", row.names = FALSE, eol = "\r")

  Dy <- as.data.frame(cbind(as.data.frame(dat$Name),as.data.frame(dat$D
y), as.data.frame(dat$Mass)))
  colnames(Dy) <- c("Name", "Dy","Mass")
  Dy <- merge(Dy, y, by="Name")
  write.csv(Dy, file = "Dy-Ro.csv", row.names = FALSE, eol = "\r")

  Nb <- as.data.frame(cbind(as.data.frame(dat$Name),as.data.frame(dat$N
b), as.data.frame(dat$Mass)))
  colnames(Nb) <- c("Name", "Nb","Mass")
  Nb <- merge(Nb, y, by="Name")
  write.csv(Nb, file = "Nb-Ro.csv", row.names = FALSE, eol = "\r")

  Ti <- as.data.frame(cbind(as.data.frame(dat$Name),as.data.frame(dat$T
i), as.data.frame(dat$Mass)))
  colnames(Ti) <- c("Name", "Ti","Mass")
  Ti <- merge(Ti, y, by="Name")
  write.csv(Ti, file = "Ti-Ro.csv", row.names = FALSE, eol = "\r")

  #Merge all files with respective element concentrations and masses
  Ydycenbti <- merge(dat, y, by = "Name")
  #Write out master '.csv' file
  write.csv(Ydycenbti, file="Cedyynbti-Ro.csv" , row.names = FALSE, eol
 = "\r")
#---------------end of preprocessing and file generation------Ta da!---
-----
```

# APPENDIX C : PCA AND PREDICTIVE MODELING CODES

This code is annotated since some results are printed within it. This was obtained from a markdown file. This code was used for Yttrium in rock matrix. All other elements can similarly be treated (with care).

```
library(pls)

library(e1071)

library(neuralnet)

library(plyr)

library(reshape2)

library(ggplot2)

library(doParallel)

library(caret)

library(dplyr)

library(DT)

PCA - ChemoSpec Approach

Chemometrics, as defined by Varmuza and Filzmoser; “. . . the
extraction of relevant information from chemical data by mathematical
and statistical tools.”

ChemoSpec was developed for the chemometric analysis of spectroscopic
data, such as UV-Vis, NMR or IR data.

Data Preprocessing

Two csv files (rock and rock matrices) are used. They are obtained from
preprocessing raw spectral files. The spectral section of the files are
used. ChemoSpec ideally works on spectroscopic data.
```

*Several steps are required to process the data for ChemoSpec functions. The rock and rock matrix are preprocessed and merged with appropriate naming.*

*rock Matrix*

*File subsets are identifiable:*

i. *Samples prepared with Cerium, Dysprosium, and Yttrium (at high concentration).*

**ii.** ***Samples with no elements.***

iii. ***Samples containing Niobium and Titanium.***

**iv.** ***Samples containing only Yttrium (at low concentration).***

**IMPORTANT TO NOTE IS THE SAMPLE NOMECLATURE. 'Knitr' CONVERTS '-' TO '.' WHICH CAUSES ISSUES WITH CHEMOSPEC SPECTRA READING FUNCTIONS. PAY ATTENTION TO ALL SAMPLE NAMES SO AS TO REPLACE THEM APPROPRIATELY.*

*Ce, Dy, and Y in rock Matrix*

*This code extracts rock matrices containing Cerium, Dysprosium, and Yttrium (at high concentration) in one sample.*

```
df <- read.csv("Cedyynbti-Ro_Y-Nb ROI (14.6 - 18.8 keV)_matrix_cor_(16.4 - 17.4 keV region).csv", header = T, check.names=FALSE)

df <- as.data.frame(t(df))
df <- df[-1,]
df$Name <- row.names(df)



df2 <- read.csv("Mass_conc_ro.csv")

df3 <- merge(df2, df, by = "Name")
write.csv(df3, "Cedyynbti-Ro.csv", row.names = F)
```

```
require(ChemoSpec)

require(ggplot2)
#Read in csv file
df <- read.csv("Cedyynbti-Ro.csv")
CeDyY.Ro <- subset(df, Ce>0&Dy>0&Y>0) #Subset data
CeDyY.Ro <- data.frame(t(CeDyY.Ro[,-c(2:13)]))#Remove non-spectral columns and transpose
CeDyY.Ro[1,] <- gsub("sample", "CeDyY_Ro", CeDyY.Ro[1,])#Chane naming for ChemoSpec
CeDyY.Ro[1,] <- gsub("Sample", "CeDyY_Ro", CeDyY.Ro[1,])#Chane naming for ChemoSpec
CeDyY.Ro[1,] <- gsub("-", "_", CeDyY.Ro[1,])#Remove ".asc" tag
colnames(CeDyY.Ro) <- CeDyY.Ro[1,] #Set the column names to sample names
CeDyY.Ro <- CeDyY.Ro[-1,] #Remove first row containing names
CeDyY.Ro$Energy <- as.numeric(c((1:nrow(CeDyY.Ro))*0.0300111028647766/8))+14.61 #Add Energy column to replace channels
n.col <- ncol(CeDyY.Ro) #Get number of columns
CeDyY.Ro <- cbind(CeDyY.Ro[,n.col], CeDyY.Ro[,c(1:(n.col-1))]) #Make energy column first and sample spectra follows
```

Blanks rock matrix

This code extracts rock matrices containing no element.

```
Blanks.Ro <- subset(df, Ce==0&Dy==0&Y==0&Ti==0&Nb==0) #Subset data
Blanks.Ro <- data.frame(t(Blanks.Ro[,-c(2:13)]))#Remove non-spectral columns and transpose
Blanks.Ro[1,] <- gsub("sample", "Blanks_Ro", Blanks.Ro[1,])#Chane naming for ChemoSpec
Blanks.Ro[1,] <- gsub("Sample", "Blanks_Ro", Blanks.Ro[1,])
Blanks.Ro[1,] <- gsub("-", "_", Blanks.Ro[1,])
colnames(Blanks.Ro) <- Blanks.Ro[1,] #Set the column names to sample names
Blanks.Ro <- Blanks.Ro[-1,] #Remove first row containing names
```

## Nb-Ti in rock matrix

This code extracts rock matrices containing only Niobium and Titanium.

```r
NbTi.Ro <- subset(df, Ti>0&Nb>0) #Subset data
NbTi.Ro <- data.frame(t(NbTi.Ro[,-c(2:13)]))#Remove non-spectral columns and transpose
NbTi.Ro[1,] <- gsub("Sample", "NbTi_Ro", NbTi.Ro[1,])#Chane naming for ChemoSpec
NbTi.Ro[1,] <- gsub("-", "_", NbTi.Ro[1,])
colnames(NbTi.Ro) <- NbTi.Ro[1,] #Set the column names to sample names
NbTi.Ro <- NbTi.Ro[-1,] #Remove first row containing names
```

##No starch Matrix

All the datasets are merged and written to csv by the code below. This should ideally work for one of the ChemoSpec spectral reading functions. THIS FAILED INVOKING AN EXTRA STEP OF CREATING CSV SAMPLE FILES.

```r
bind.ed <- cbind(CeDyY.Ro, Blanks.Ro, NbTi.Ro)
write.table(bind.ed, "merged_matrix.csv", sep = ",", row.names = FALSE)
```

## Generation of csv files

This code chunk generates csv files from the merged dataset. This is due to the requirements in ChemoSpec spectra reading functions. The files are stored separately in a directory since the function reads all csv files in a path.

```r
merged <- read.csv("merged_matrix.csv")
set.wd <- setwd("./asc")
set.wd
```

```
## [1] "E:/Google Drive/Thesis/Data_Files_Processors/Combined_All_Starch and Rock-No blanks-sim/Ratio_Final/Cedyynbti-Ro_Y-Nb ROI (14.6 - 18.8 keV)_matrix_cor_(16.4 - 17.4 keV region)"
```

```r
#Create a loop
m <- 2
```

```
while (m <= ncol(merged)) {

  spec.asc <- merged[,c(1,m)]
  file.name <- colnames(spec.asc)[2]
  file.name <- gsub(".asc", ".csv",file.name)
  write.table(spec.asc, file = file.name, row.names = FALSE, col.names
= FALSE, sep = ",")
  m <- m+1
}
```

```
setwd("./asc")
file_s <- as.data.frame(list.files(pattern = "\\.csv$"))
file_s$Name <- 0
colnames(file_s) <- c("oldCol", "Name")
file_s$Name <- file_s$oldCol
file_s[,2] <- gsub(".csv", ".asc", file_s[,2])
file_s[,2] <- gsub("CeDyY_Ro", "sample", file_s[,2])
file_s[,2] <- gsub("Blanks_Ro_", "Sample-", file_s[,2])
file_s[,2] <- gsub("Blanks_Ro", "sample", file_s[,2])
file_s[,2] <- gsub("NbTi_Ro_", "sample-", file_s[,2])
file_s[,2] <- gsub("Y_Low_Ro", "sample", file_s[,2])
```

```
{r getNames} # write.csv(file_s, "filenames.csv") # file_s <-
read.csv("filenames.csv") #
```

## Getting Data into ChemoSpec

The csv files are ready for ChemoSpec. This is achieved by making sure the path points to the directory of interest.

```
require(ChemoSpec)
require(R.utils)
```

```
#Point to new directories for rock, rock, and combined matrices
setwd("./asc")
files <- as.data.frame(list.files(pattern = "\\.csv$"))
spect.ra <- files2SpectraObject(gr.crit =
c("NbTi_Ro","Blanks_Ro","CeDyY_Ro","Y_Low_Ro"), sep = ",", header =
FALSE, freq.unit = "Energy (KeV)", int.unit = "Intensity", debug = TRUE)
```

```
#spect.ra <- normSpectra(spect.ra, method = "TotInt")
sumSpectra(spect.ra)
```

Plot of Spectra

All spectral data plots.

```
png(filename = "spectra.png", height = 400, width = 580)
    plotSpectra(spect.ra,
                offset = 0,
                which = c(6,35,40),
                yrange = c(0,10000),
                lab.pos = 2200)
dev.off()
```

Figure: Spectral plot

PCA

Two options for PCA are available in the ChemoSpec package i. Classical methods. ii. Robust methods.

Classical methods use all the data to compute the scores and loadings while Robust methods focus on the core of the data, which means that some samples are downweighted. Robust methods tend to downweigh outlier samples (such samples contribute most variance in the dataset), thereby decreasing their influence on the PCs.

i. PCA Classical method

The dataset contains trailing zeroes due to the instrument detection range. This implies that autoscaling cannot be performed.

Classical method 2D scores plot

```
pca.spec <- c_pcaSpectra(spect.ra, choice = "noscale", cent = TRUE)
plotScores(spect.ra,
           pca.spec,
           pcs = c(1,2),
           ellipse = "cls",
           tol = 0.001)
abline(h=0, v=0)
```

Classical method 3D scores plot

```
plotScores3D(spect.ra, pca.spec, ellipse = FALSE)
```

Classical method Scree plot

```
plotScree(pca.spec)
```

Figure: Scree plot of classical method

Loadings Plot

```
plotLoadings(spect.ra,
             pca.spec,
             loads = c(1:4),
             ref = 40)
```

Extract Principal Components and merge with concentration data

```
pcs <- as.data.frame(pca.spec$x[,c(1,2,3,4)])
pcs <- as.data.frame(cbind(file_s$Name, pcs))
colnames(pcs) <- c("Name", "PC1", "PC2", "PC3","PC4")
conY <- read.csv("Mass_conc_ro.csv")
pcs <- merge(conY, pcs, by="Name")
```

Extract Variables to use: The concentration and principal components Extract Variables to use: The concentration and principal components

```
data <- pcs[,c("Y", "Mass", "PC1","PC2","PC3","PC4")]
#create a random scheme to sample
index <- sample(1:nrow(data),round(0.8*nrow(data)))
#Split data into testing and training set according to the initial
randomization
train <- data[index,]
test <- data[-index,]
```

```r
#---Build formula for prediction by listing them to a variable (n)
n <- names(data)
#Create formula used in most R predictive formulations
11

f <- as.formula(paste("Y ~", paste(n[!n %in% "Y"], collapse = " + ")))
#Scale the data for faster modelling. In Neural Networks, using the original datapoints will take too much computational time. The algorithm will not converge in time.
maxs <- apply(data, 2, max)#Get maximum for each column
mins <- apply(data, 2, min)#Get minimum for each column
#Scale data according to the scheme
scaled <- as.data.frame(scale(data, center = mins, scale = maxs - mins))
#Split data into testing and training set according to the initial randomization
train_ <- scaled[index,]
test_ <- scaled[-index,]
```

Models optimization

```r
ctrl <- trainControl(method="cv", number=10)
# train the NN
gridNN <- expand.grid(layer1 = c(1,2,3,4), layer2 = c(1,2,3,4), layer3 = c(1,2,3,4))
#gridNN <- expand.grid(layer1 = c(1), layer2 = c(1), layer3 = c(1,2))
NN <- train(f,
            data = train_,
            trControl = ctrl,
            method = "neuralnet",
            tuneGrid = gridNN)
# summarize results
print(NN)

#Predict using best model
predNN <- as.data.frame(predict(NN, newdata = test_))*((max(data$Y)-min(data$Y))+min(data$Y))
predNN
```

```
ctrl <- trainControl(method="cv", number=10)


SVR <- train(f,
             data = train_,
             method = "svmPoly",
             trControl=ctrl)

#Predict using best model
predSVR <- as.data.frame(predict(SVR, newdata = test_))*((max(data$Y)-
min(data$Y))+min(data$Y))
predSVR

ctrl <- trainControl(method="cv", number=10)


RMFR <- train(f,
              data = train_,
              method = 'rf',
              trControl = ctrl)
RMFR

#Predict using best model
predRMFR <- as.data.frame(predict(RMFR, newdata =
test_))*((max(data$Y)-min(data$Y))+min(data$Y))
predRMFR

#Write predicted values to dataframes
Measured <- as.data.frame(test$Y)


#Bind predictions
pred <- cbind(Measured, predNN, predSVR, predRMFR)


#Rename columns
colnames(pred) <- c("Measured", "NN", "SVR","RF")


#Write predictions to a file in the directory. Ideal for MS Excel users.
write.csv(pred, file = "predY.csv")
```

```r
#Plot prediction graphs of the test dataset
#The following libraries are required

#Melt PREDICTED data to three variables based on the MEASURED VARIABLE,
for easy ploting
df_melt <- melt(pred, id="Measured")
#Change column names of melted data
colnames(df_melt) <- c("Measured", "Model", "Predicted")

#Print to '.png' graphs of prediction using PCR, PLSR, SVM, and ANN
png(filename = "PredictionPlotY.png", height = 300, width = 400)
ggplot(data = df_melt, aes(x=Measured, y=Predicted, colour = Model))+
  geom_point()+
  #Add straight line through central ideal prediction
  geom_abline(lty=2)+
  #Add title
  ggtitle("Predicted vs Measured Yttrium")+
  #Format background and title position
  theme(panel.background = element_rect(fill = "white", colour="black",
linetype = 1), plot.title = element_text(hjust = 0.5))
dev.off()#Shut graphics device
```

# APPENDIX D : ADDITIONAL RESULTS

Table D.1. Predictive model performance (in ppm) for Cerium in rock matrix

| Sample ID | Measured | NN | SVR | RF |
|---|---|---|---|---|
| 5 | 0 | 34 ±24 | 34 ±24 | 15 ±11 |
| 10 | 8 | 83 ±53 | 44 ±26 | 38 ±22 |
| 16 | 12 | -2 ±10 | 5 ±5 | 40 ±20 |
| 26 | 19 | 5 ±10 | 35 ±11 | 24 ±4 |
| 27 | 31 | 301 ±191 | 36 ±4 | 45 ±10 |
| 31 | 56 | 232 ±124 | 82 ±18 | 61 ±4 |
| 35 | 184 | 46 ±98 | 37 ±104 | 80 ±74 |
| 40 | 235 | 64 ±121 | 36 ±141 | 49 ±132 |
| 46 | 365 | 285 ±57 | 80 ±201 | 206 ±112 |

Table D.2. Predictive model performance (in ppm) for Dysprosium in rock matrix

| Sample ID | Measured | NN | SVR | RF |
|---|---|---|---|---|
| 11 | 0 | 34 ±24 | 34 ±24 | 15 ±11 |
| 31 | 8 | 83 ±53 | 44 ±26 | 38 ±22 |
| 34 | 12 | -2 ±10 | 5 ±5 | 40 ±20 |
| 36 | 19 | 5 ±10 | 35 ±11 | 24 ±4 |
| 42 | 31 | 301 ±191 | 36 ±4 | 45 ±10 |
| 45 | 56 | 232 ±124 | 82 ±18 | 61 ±4 |
| 49 | 184 | 46 ±98 | 37 ±104 | 80 ±74 |
| 62 | 235 | 64 ±121 | 36 ±141 | 49 ±132 |
| 79 | 365 | 285 ±57 | 80 ±201 | 206 ±112 |

Table D.3. Predictive model performance (in ppm) for Yttrium in rock matrix

| Sample ID | Measured | NN | SVR | RF |
|---|---|---|---|---|
| 4 | 0 | 34 ±24 | 588 ±416 | 0 ±0 |
| 10 | 761 | 777 ±11 | 812 ±36 | 766 ±4 |
| 22 | 2397 | 2254 ±101 | 2254 ±101 | 2616 ±155 |
| 25 | 3562 | 3386 ±124 | 3918 ±252 | 3623 ±43 |
| 27 | 4858 | 4464 ±279 | 4867 ±6 | 4484 ±264 |
| 47 | 9296 | 9001 ±209 | 10487 ±842 | 9507 ±149 |
| 57 | 9483 | 9453 ±21 | 11071 ±1123 | 9596 ±80 |
| 68 | 11843 | 14283 ±1725 | 14734 ±2044 | 11759 ±59 |

Table D.4. Predictive model performance (in ppm) for Cerium in starch matrix

| Sample ID | Measured | NN | SVR | RF |
|---|---|---|---|---|
| 35 | 0 | 2 ±2 | 2 ±2 | 0 ±0 |
| 73 | 8 | 22 ±10 | 36 ±19 | 12 ±3 |
| 79 | 17 | 15 ±1 | 22 ±4 | 35 ±13 |
| 85 | 33 | 21 ±8 | 46 ±9 | 41 ±6 |
| 94 | 39 | 51 ±8 | 61 ±16 | 78 ±28 |
| 98 | 45 | -9 ±38 | 13 ±23 | 41 ±3 |
| 102 | 109 | 175 ±47 | 53 ±40 | 77 ±23 |
| 104 | 188 | 270 ±58 | 110 ±55 | 210 ±16 |
| 125 | 192 | 172 ±14 | 124 ±48 | 174 ±12 |
| 126 | 378 | 283 ±67 | 126 ±178 | 290 ±62 |

Table D.5. Predictive model performance (in ppm) for Dysprosium in starch matrix

| Sample ID | Measured | NN | SVR | RF |
|---|---|---|---|---|
| 36 | 0 | -4 ±3 | 2 ±1 | 5 ±4 |
| 74 | 7 | -18 ±18 | 48 ±29 | 13 ±4 |
| 83 | 9 | 10 ±1 | 21 ±9 | 17 ±6 |
| 84 | 10 | 18 ±6 | 18 ±6 | 21 ±8 |
| 90 | 45 | 83 ±27 | 74 ±21 | 68 ±16 |
| 95 | 186 | 203 ±12 | 171 ±11 | 158 ±20 |
| 98 | 188 | 135 ±37 | 138 ±35 | 107 ±57 |
| 101 | 370 | 385 ±11 | 349 ±15 | 363 ±5 |
| 106 | 396 | 309 ±62 | 297 ±70 | 235 ±114 |
| 123 | 0 | -4 ±3 | 2 ±1 | 5 ±4 |

Table D.6. Predictive model performance (in ppm) for Yttrium in starch matrix

| **Sample ID** | **Measured** | **NN** | **SVR** | **RF** |
|---|---|---|---|---|
| 46 | 0 | -19 ±14 | 442 ±312 | 0 ±0 |
| 49 | 5 | -9 ±10 | 446 ±312 | 0 ±4 |
| 66 | 17 | 19 ±1 | 459 ±313 | 11 ±4 |
| 68 | 29 | 84 ±39 | 489 ±325 | 63 ±24 |
| 72 | 463 | 399 ±45 | 653 ±134 | 357 ±75 |
| 74 | 650 | 585 ±46 | 761 ±78 | 506 ±102 |
| 87 | 1070 | 940 ±92 | 990 ±57 | 799 ±192 |
| 97 | 1504 | 1339 ±117 | 1280 ±158 | 1413 ±64 |
| 100 | 2324 | 2080 ±173 | 1894 ±304 | 2481 ±111 |
| 101 | 3688 | 3617 ±50 | 3390 ±211 | 3910 ±157 |
| 103 | 4970 | 4651 ±226 | 4478 ±348 | 4608 ±256 |
| 107 | 9349 | 9330 ±13 | 9353 ±3 | 9375 ±18 |
| 117 | 18390 | 18373 ±12 | 17936 ±321 | 18356 ±24 |

Table D.7. Predictive model performance (in ppm) for Titanium in starch matrix

| **Sample ID** | **Measured** | **NN** | **SVR** | **RF** |
|---|---|---|---|---|
| 1 | 0 | 7 ±5 | 22 ±16 | 12 ±8 |
| 4 | 5 | 34 ±21 | 57 ±37 | 35 ±21 |
| 13 | 10 | -14 ±17 | 24 ±10 | 24 ±10 |
| 16 | 20 | 108 ±62 | 27 ±5 | 162 ±100 |
| 23 | 35 | -2 ±26 | 6 ±21 | 2 ±23 |
| 27 | 55 | 4 ±36 | 54 ±1 | 16 ±28 |
| 32 | 146 | 140 ±4 | 132 ±10 | 162 ±11 |
| 34 | 185 | 340 ±110 | 67 ±83 | 281 ±68 |
| 38 | 236 | 337 ±71 | 75 ±114 | 245 ±6 |
| 45 | 286 | 219 ±47 | 159 ±90 | 204 ±58 |

Table D.8. Predictive model performance (in ppm) for Niobium in starch matrix

| **Sample ID** | **Measured** | **NN** | **SVR** | **RF** |
|---|---|---|---|---|
| 4 | 0 | 1 ±1 | 3 ±2 | 3 ±2 |
| 6 | 5 | 7 ±1 | 4 ±1 | 11 ±4 |
| 11 | 10 | 17 ±5 | 13 ±2 | 8 ±1 |
| 14 | 35 | 29 ±4 | 27 ±6 | 6 ±21 |
| 19 | 55 | 68 ±9 | 73 ±13 | 62 ±5 |
| 24 | 80 | 85 ±4 | 91 ±8 | 82 ±1 |
| 25 | 110 | 85 ±18 | 93 ±12 | 100 ±7 |
| 29 | 145 | 163 ±13 | 159 ±10 | 107 ±27 |
| 30 | 185 | 237 ±37 | 139 ±33 | 262 ±54 |
| 35 | 285 | 292 ±5 | 217 ±48 | 261 ±17 |
| 39 | 348 | 318 ±21 | 218 ±92 | 269 ±56 |