UNIVERSITY OF NAIROBI

A ROBUST CHEMOMETRICS APPROACH TOWARDS TRACE BIOMETAL AND ASSOCIATED SPECIATION ANALYSIS IN SOFT TISSUE UTILIZING ENERGY DISPERSIVE X-RAY FLUORESCENCE AND SCATTERING SPECTROMETRY

By

SICHANGI, Edward Kokonya BSc (Hons), PGDE (Nairobi)

Thesis submitted for examination in partial fulfillment of the requirements for the award of the degree of Master of Science in Physics of the University of Nairobi.

September, 2017

DECLARATION

"I hereby declare that this thesis submitted for the degree of MSc at Department of Physics, University of Nairobi is my own work and has not previously been presented to any other institution of higher education."

SICHANGI, Edward Kokonya

(I56/81139/2012)

Department of Physics, University of Nairobi

Signature…… Date……

This thesis is submitted with approval of the undersigned supervisors for examination,

Dr. Angeyo H. Kalambuka,

Department of Physics, University of Nairobi.

Signature…… Date……

Dr. Alix D. Massop,

Department of Physics, University of Nairobi.

Signature…… Date……..

Mr. Michael J. Mangala

Institute of Nuclear Science & Technology, University of Nairobi.

Signature…… Date……

DEDICATION

To my late daughter and son both of whom passed away on delivery in 2010 and 2015 respectively.

ACKNOWLEDGEMENTS

This thesis, like any other, owes much to the contributions and participation of many people; Dr. Angeyo H. Kalambuka, Dr. Alix D. Massop and Mr. Michael Mangala for their invaluable inputs throughout the time of this research. I profoundly thank Dr. Angeyo H. Kalambuka for his guidance on problem formulation and proposal writing, Dr. Dehayem Massop for taking time off her busy schedule to accompany me to the Institute of Nuclear Science laboratory for sample preparation and guidance on proposal writing and Mr. Mangala for guiding me on proposal and report writing.

I thank the Director and laboratory staff of the Institute of Nuclear Science and Technology of the University of Nairobi for allowing me to access to their laboratory facilities and specifically the Chief Technologist Mr. S. K. Bartilol for his guidance and assistance in EDXRF spectrometry.

I also thank the staff of the department of Human Anatomy and Vetinary Medicine laboratories of the University of Nairobi for assisting me in sample preparation.

Much thanks also to my departmental colleagues Mr. Okonda Justus, Mr. Ian Kaniu and Ms. Emily Otieno for helping me in developing my proposal, sample preparations and use of chemometrics.

My gratitude goes to the chairman and administrative staff of the Department of Physics for facilitating much of the work and to all who gave me moral and material support. Finally, yet important, I'm deeply touched and indebted to my wife; Resah, my daughters Risper, Nimpha, Esther and my son Solomon for being patient and understanding during my absence.

ABSTRACT

The goal of this study was to utilize a robust multivariate chemometrics approach towards direct, rapid and accurate quantitative determination of Na, Mg, Zn, Fe, Cu, Mn and Co and speciation of Cu, Mn and Fe in complex light element matrix materials in this case soft body tissues. This study has relevance in disease diagnostics in body tissues utilizing trace biometals as the disease biomarkers. The technique in use is Energy Dispersive X-ray Fluorescence and Scattering (EDXRFS) spectrometry.

Direct, rapid and simultaneous determination of trace biometal concentrations and their speciation in human body tissues utilizing Energy Dispersive X-ray fluorescence and Scattering (EDXRFS) spectrometry is challenging. This is because the spectra are characterized by analyte peak overlaps, weak fluorescence peak signals and extreme matrix effects due to the predominance of low-Z elements (dark matrix), poor signal-to-noise ratio (SNR) of the analyte peaks, and imprecise sample geometry. Moreover, there is not yet available a direct method for speciation analysis in EDXRF spectroscopy.

The utility of chemometrics in EDXRF analysis is however still under development. High-noise and low-resolution EDXRF spectra are amenable to multivariate chemometric methods for qualitative and quantitative analysis. Simple univariate, classical linear regression and multivariate regression methods are restricted to linear relationships and are therefore inapplicable.

Samples of soft body tissues prepared as thin (5μm), intermediate thick (10μm) and thick (20 μm) were analyzed. Robust chemometrics methods namely combined use of wavelet transform (WT), principal component analysis (PCA), independent component analysis (ICA) for spectral processing and artificial neural network (ANN) and partial least squares (PLS) for multivariate calibration based on the use of paraffin wax ‘standards’ spiked with Fe, Cu, Mn, Zn, Na, Mg and Co were utilized.

WT was used in de-noising and resolution enhancement of the spectra to optimize them for the determination of the elemental concentration of Na, Mg, Mn, Fe, Cu, Co and speciation of Fe, Mn, Cu. When WT and PCA were combined, there was improved signal–to–noise ratio (SNR) of analyte as well as scatter peaks. ICA was used for both pattern recognition (classification) of tissues into those with lower and higher speciation of Fe, Mn, Cu and for spectral preprocessing and/or dimension reduction when combined with WT to optimize the spectra for

qualitative analysis. PCA was used for reduction of spectral data dimension and pattern recognition.

The preprocessed spectra were used as an input to artificial neural networks (ANN) and partial least squares (PLS) models for development of calibration strategy for direct quantitative analysis using fluorescence spectral signature regions and Compton scatter peaks. Both ANN and PLS calibrations gave results for trace element concentration better than when raw spectra was utilized i.e. (R^2 values for the elements Fe, Mg, Mn, Na, Co, Zn and Cu were; $R^2 \sim 0.889 – 0.951$ before and ~ 0.989 - 0.997 after preprocessing for ANNs; and $R^2 \sim 0.876$ - 0.931 before, and ~ 0.969 – 0.993 after preprocessing for PLS). The results also indicate that there was improvement in determination of low–Z elements (Na and Mg) when the preprocessed spectra of both fluorescence and scatter regions were utilized simultaneously i.e. ($R^2 = 0.976$ for featured fluorescence and $R^2 = 0.994$ for both featured and Compton scatter for Na; and $R^2 = 0.932$ for featured fluorescence and $R^2 = 0.995$ for both featured and Compton scatter for Mg utilizing ANN model). Normally Na and Mg cannot be analyzed by classical EDXRF spectroscopy. Quantitative analysis Oyster tissue by the analytical approach was in agreement with certified values of the analytes in the standard reference material (≤ 6 % or less for most elements). The results are independent of sample thickness.

Quantitative analysis of dog tissues indicate that mammary and prostate cancer tissues were dominated with high concentration of Zn, Fe, Mg and Cu as compared to healthy mammary and prostate tissues. The results of speciation analysis indicate that mammary and prostate cancer tissues were rich in high oxidation state (Cu^{2+}, Fe^{3+}, Mn^{7+}). The analytical approach reported here is novel for direct rapid analysis of concentration levels and speciation alterations of selected trace elements (Fe, Cu, and Mn) in the context of cancer characterization.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

ANN	Artificial neural networks
AXIL	Analysis of X-ray spectra by iterative least squares
BP-ANN	Back-propagation artificial neural network
CRM	Certified reference material
DWT	Discrete wavelet transforms
EDXRF	Energy dispersive X-ray fluorescence
EDXRFS	Energy dispersive X-ray fluorescence and scattering
FASTICA	FAST independent component analysis
FP	Fundamental parameters
HPGe	High –purity Germanium
ICA	Independent component analysis
KNN	K-nearest neighbor
LDA	Linear discriminant analysis
PCA	Principal component analysis
PCR	Principal component regression
PLS	Partial least square
QXAS	Quantitative X-ray analysis system
SNR	Signal-to-noise ratio
SR-XRF	Synchrotron radiation X-ray fluorescence
SVM	Support vector machines
TXRF	Total reflection X-ray fluorescence
μ-XRF	Micro X-ray fluorescence
WD-XRF	Wavelength dispersive X-ray fluorescence
WT	Wavelet transform
XANES	X-ray absorption near-edge structure
XRF	X-ray fluorescence

CHAPTER ONE: INTRODUCTION

1.1 XRF spectroscopy

X-ray fluorescence (XRF) spectrometry is an elemental analysis technique that is based on the principle that atoms/elements, when excited by an external ionizing radiation, emit X-ray radiation of characteristic energy (or wavelength) which when analyzed, the elements present in the excited sample can be quantified. XRF spectroscopy is capable of analyzing solid and liquid samples for major, minor and trace (i.e. $\leq$ 1000 µg/g) elements in a sample. XRF spectrometry is widely used for quantitative and qualitative determination of all but low (i.e. Z $\geq$13) atomic number elements with detection limits down to µg/g or ng/g levels depending on the sample form and the element of interest (Wobrauscheck and Christina, 2010). The inapplicability of XRF spectroscopy to low- Z elements is due to enhanced scattered radiation and low fluorescence yield of the low Z- elements (Beckhoff, 2006).

The advantages of XRF spectroscopy in material analysis compared to other competing methods are its large dynamic range (ng/g to %), high precision (~ 0.1 %) and minimal requirement for sample preparation. However, XRF spectrometry has some drawbacks namely; spectral interferences of element characteristic lines due to the poor resolution of the normally used solid-state detectors and matrix effects. Matrix effects make a direct conversion of fluorescence peak intensities into elemental concentrations difficult. This is because secondary X-rays from the sample to the detector are attenuated by absorption and X-ray fluorescence enhancement resulting into characteristic peak intensity that is a function of both the element's concentration and that of the other elements in the sample (Potts and Webb, 1992). Sample destructivity, weak excitation of low-Z elements and stringent sample geometry requirements for the XRF- FP method are other limitations associated with the method (Beckhoff, 2006). Due to these drawbacks, a new method, energy dispersive X-ray and fluorescence scattering

(EDXRFS) spectrometry was recently developed that exploits both fluorescence and scatter radiation for material analysis. In this technique fluorescence as well as scatter radiation are utilized simultaneously and complimentarily for direct transformation of EDXRFS spectra to analytes concentration (Kaniu *et al.,* 2011; Kaniu *et al.,* 2012; Kaniu *et al.,* 2014; Angeyo *et al.,* 2012).

1.1.1 Scatter radiation in XRF analysis

In XRF spectroscopy, scatter radiation occurs when photons incident to a sample interact with loosely bound electrons of the target element in the sample matrix thus resulting into Rayleigh (coherent) and Compton (incoherent) scattering (Anjali, 2014). Scattered radiation is considered a drawback in elemental quantification analysis, for it introduces elevated background in the spectrum (Nielson, 1977), which masks weak fluorescent peak of analytes, that are the basis of quantitative XRF analysis.

Samples especially those of organic nature such as soft body tissues are mostly composed of low-Z elements which enhance Compton scattering due to the low X-ray absorption cross-section of such elements (Bueno *et al.,* 2006) and enhanced background radiation. This in effect, masks, the especially subtle (trace and ultratrace) fluorescent peaks of analytes making it difficult to realize quantitative analysis based on direct correlation of the analyte peak fluorescence and the concentration of the analyte in the sample, as peak statistics and SNR of the analyte is very poor.

1.1.2 Quantitative XRF spectrometry of materials

In principle, elemental spectral peak intensities in an XRF energy spectrum are determined using, for example spectrum evaluation software such as AXIL. The determination of elemental concentrations is then done using QXAS software (Van Espen *et al.,* 1977). However, the

relationship between the intensity of an element's X-ray characteristic peak and its concentration is not a simple linear function in practice (Debertin and Helmer, 1988).

Generally when a sample contains a large proportion of light elements ($Z < 13$), the incident excitation radiation beam is mainly scattered and highly absorbed in a sample. The number of fluorescence photons emitted from the analyzed sample is affected by the so called "matrix'' effects categorized as: (i) X-ray absorption (ii) enhancement and (iii) sample macroscopic effects which constitute sample inhomogeneity (Brouwer, 2003).

The matrix effects may be corrected using methods such as Influence Coefficients (IC) and Fundamental Parameter (FP) (Lachance, 1999). Use of mathematical expressions is the basis of FP method in which fluorescence and emission are quantified in terms of fundamental physical parameters of X-ray emission process. Mass absorption coefficients, fluorescence yields, absorption jump ratios, line emission probabilities and characteristic X-ray energies are referred to as the fundamental parameters (Nielson, 1977).

The Influence Coefficients method enables quantification of matrix effects of each element individually i.e. the matrix effects of element *j* and *k* on analyte *i* etc. The concentration of an element *j* may be expressed as a function of its X-ray characteristic line intensity corrected for the concentration of other elements by influence coefficients (Gerald, 1999). The approach leads to accurate results but it is time consuming and expensive.

The goal of this study was to design and exploit robust chemometrics methods in (i) preprocessing of EDXRFS spectra to optimize for qualitative and quantitative analysis of soft body tissue (ii) realizing a direct rapid and accurate simultaneous determination of trace biometals and associated speciation in soft body tissue prepared as thin, intermediate thick and thick sections.

1.2 Statement of the problem

The analytical challenges encountered in the determination of the trace and major biometal and associated speciation in complex light element matrices such as soft human body tissue through non-invasive, rapid and direct energy dispersive X-ray fluorescence (EDXRF) spectrometry necessitate the development of robust multivariate chemometrics techniques to aid in the analysis. EDXRF spectra obtained from complex low-Z element matrices are characterized by spectral overlaps, weak fluorescence signals of the analytes of interest and background noise which masks the normally subtle fluorescence peaks of analytes of interest thus making it difficult to perform quantitative analysis directly. Also the method is not applicable to speciation analysis. In principle the utility of EDXRF is limited to high Z elements, but poor in performance for low-Z element. Consequently determination of trace biometal and associated speciation is an analytical challenge in the EDXRF analysis of complex low-Z dominated matrices.

1.3 Objectives

1.3.1 General objective

The goal of this study was to design and utilize a robust multivariate chemometrics approach in energy dispersive X-ray fluorescence and scattering spectrometry to realize direct, rapid and more accurate simultaneous determination of trace and major biometals (namely: Na, Mg, Zn, Fe, Cu, Mn and Co) and associated speciation in soft body tissues prepared as thin (5 μm), intermediate thick (10 μm) and thick (20 μm) sections.

1.3.2 Specific objectives

(i) To develop an analytical strategy for the hybrid use of WT, PCA and ICA in preprocessing of EDXRFS spectra to optimize for quantitative determination of Na, Mg, Zn, Fe, Cu, Mn and Co and speciation of Fe, Cu and Mn in soft body tissues.

(ii) To apply the developed analytical strategy towards comparative direct and rapid quantification of trace and major biometals (namely: Na, Mg, Zn, Fe, Cu, Mn and Co) in soft body tissue utilizing EDXRFS spectrometry in conjunction with chemometrics techniques (ANN and PLS).

(iii) To develop multivariate calibration for speciation analysis of Fe, Cu and Mn in soft body tissue utilizing EDXRFS spectrometry and chemometrics techniques.

(iv) To apply the developed analytical methods above on native soft body (cancerous and non-cancerous) tissue so as to characterize them in the context of cancer prognostics and diagnostics.

1.4 Justification and significance of the study

The use of EDXRF spectroscopy for elemental determination in complex light element matrix samples requires proper spectrum evaluation. However, the complex nature of the processes that results in the realization of an X-ray fluorescence spectrum make it hard to get an accurate estimate of the analyte line intensities that can be utilized in determination of the content of the sample especially at trace and ultra-trace levels in low-Z dominated samples. Another major drawback of EDXRF spectrometry in elemental analysis is its inability of distinguishing elements in different valence states. This is because of poor resolution of e.g. K_α and K_β fluorescence peaks of analytes. Multiple scattering and prominent Bremsstrahlung contribute

to the continuous portions of spectrum that overlap the subtle characteristics X-ray fluorescence lines of interest (Jahan *et al.,* 2013).

In this work a robust chemometrics approach for EDXRFS spectral preprocessing and determination of trace biometals namely Na, Mg, Co, Zn, Cu, Fe and Mn and associated speciation in soft body tissues utilizing fluorescence and scatter has been exploited. This approach is improvement to the common EDXRF practices. The approach involves hybrid use of chemometrics techniques (WT, PCA, ICA, ANN and PLS) in preprocessing of EDXRFS spectra and quantitative determination of trace biometals and associated speciation in soft body tissues. The ability to perform quantitative and speciation analysis of soft body tissue provides necessary diagnostic information about cancer.

1.5 Scope and limitations of the study

The study was limited to seven biometals namely; Fe, Cu, Mn, Zn, Mg, Co and Na. This is because these elements are essential trace materials in the body tissue and play a major role in body health. Mg and Na were included in the study to represent low-Z elements. The speciation analysis of trace biometals was limited to three elements namely; Cu, Fe and Mn since they are the most involved in generation of free radicals in Fenton reaction that initiates tissue pathogenesis.

CHAPTER TWO: LITERATURE REVIEW

2.1 Chapter overview

This chapter presents three subsections that review the basis and context of the research presented in this thesis. Section 2.2 highlights utility of energy dispersive X-ray fluorescence scattering (EDXRFS) spectrometry in elemental analysis of material. Section 2.3 reviews analytical chemometrics in spectroscopy while section 2.4 highlights robust chemometrics approaches in EDXRFS spectrometry for complex matrix analysis.

2.2 Utility of EDXRFS spectrometry in elemental analysis of material

EDXRF spectrometry is suitable method for analyzing biometals in a variety of samples (Wobrauscheck and Christina, 2010). The technique utilizes conjointly fluorescence and the X-ray scatter peaks acquired from target samples to build a quantitative analysis calibration approach for trace elements (Kaniu, 2011). The method is an extension of EDXRF to the scatter domain for material characterization in terms of elemental concentration and related physico-chemical properties.

Traditionally in the EDXRF technique, spectral deconvolution methods such as analytical functions or least squares fitting to reference spectra are used to obtain the net fluorescence peak intensities (Vekemans *et al.*, 1994), whereas methods such as fundamental parameters (FP) or empirical models are used in quantitative determination of elements (Rousseau, 2006). However, FP method and empirical coefficients relies on the assumptions of sample homogeneity, plain sample surface, negligible microstructural effects and complete definition of the sample matrix composition. These assumptions limit the utility of EDXRF in the full quantitative analysis of "complex matrix" samples such as body tissues. By ''complex matrix'' (with respect to XRF spectroscopy) mean those which are dominated by severe matrix effects, overlapping peaks, 'dark matrix' problems, poor signal-to-noise ratio (SNR) of the X-ray

fluorescence spectra and imprecise analysis geometry. Such samples have enhanced Compton scatter and elevated background as well as severe matrix effects (Verbi *et al.,* 2005). Consequently direct EDXRF analysis at trace and ultra-trace levels is a challenge.

In principle, there are two ways to determine the speciation of an element in X-ray spectrometry; the direct method which utilizes the X-ray absorption near edge structure (XANES) technique but which is slow, costly and not easily available (Bernaus *et al.,* 2006) and the second other method is based on XRF but it is done indirectly; i.e. it calls for the element species to be physically separated prior to irradiation of the sample (Grafe *et al.,* 2008) and therefore is time consuming and it is not applicable to direct and simultaneous speciation analysis.

Quantitative analysis of material in XRF spectroscopy has traditionally entailed two processes namely (i) spectrum analysis and (ii) matrix corrections. Analysis of spectrum is meant to reduce background noise and obtain the net peak intensity of fluorescence spectral peaks from elements existing in the sample. Next, the resulting fluorescence peak intensity from a collection of reference samples are used in an algorithm for matrix correction to resolve interfering peaks and to establish the relationship between each elements net peak intensity and concentrations linked to them.

Matrix effects in XRF analysis of samples organic in nature cause spectral complexity, which consequently makes spectrum evaluation and subsequent deconvolution of the net intensities into concentration and speciation difficult. Such samples are characterized by enhanced Compton scatter and high background (Verbi *et al.,* 2005). Although scatter peaks are always regarded as a disadvantage, their utility in the analysis of light element matrices has been demonstrated (Kessler *et al.,* 2002). Scatter peaks contain among others, information about low Z – elements in terms of their concentration and speciation.

The fundamental parameters method, based on the Sherman equation (Sherman, 1955), has been applied to correct matrix effects in XRF (Nielson, 1977). The equation allows for the calculation of net X-ray intensities emitted by each element from a specimen of known composition. A major problem is that the expressions involve multiple integrals and hence cannot be applied easily in practice. However, the application of FP method is hindered by lack of knowledge of the sample composition, elimination of the constraint of normalizing the concentrations to 100 % and matching theory with each spectrometer. Influence coefficients has also been applied to correct matrix effects. The method quantifies matrix effects individually. However, coefficients methods require more standards than FP.

The limitations of FP and influence coefficients demanded a different approach that can combine practically flexible influence coefficient concept and the theoretical exactness of the FP method to be proposed (Rousseau, 1984). Such methodology should allow for example the calculation of theoretical influence coefficients within another algorithm referred to as the fundamental algorithm. These coefficients were referred to as multi-element influence coefficients since they depend on the full composition of the matrix (Rousseau, 2009) .

The properties of the solid-state detector and the nature of the detection process result in characteristic X-ray signals that exhibit spectral overlap. For an accurate quantitative and qualitative analysis, it is required that precise and non-biased peak intensities of the fluorescence lines are derived. There are different methodologies that exist for obtaining the intensity or area of an isolated peak, i.e. spectrum stripping, integration of peak, deconvolution and least-squares fitting. However, stripping of spectrum is difficult to use since accumulation of error occurs as stripping proceeds and thus it is not accurate. Peak deconvolution and integration are fundamentally different applications of the same principle since both of them rely on the model of detector to find peak intensity by mathematical methods; interpolating the

background under the peak and adding over the peak width, the contents of background corrected channel. The detector response model determines the accuracy of the two methods. Restricted by its functionality, resolution of overlapping peaks cannot be achieved by these two methods hence only part of the spectral information is used (Bechoff, 2006).

In samples of organic origin, for example, the dark matrix is made of mostly carbon, hydrogen and oxygen (C-H-O) and low-Z elements ($Z < 10$) which cause enhanced Compton scattering due to low X-ray absorption coefficients (Vekemans *et al.,* 1994). The intensity of the scattered peaks/scattering cross-section is inversely proportional to Z. The low Z elements have high background radiation associated with the scattering which gets to the detector as well. Hence, the signal-noise ratio (SNR) and the analyte detection limits are drastically affected in the analytes. The enhanced background masks and makes the tiny/subtle fluorescent peak intensity from the trace analytes difficult to determine.

EDXRF spectra extracted from geological materials which were irradiated using a radioisotope source had a challenge of peak deconvolution. This was demonstrated when the problems related to utilization of pure element spectra for the evaluation of peak overlap factors were described and a detector peak model that might be used derived (Boyle, 1999). The model included correcting of dead time, sum peaks and escape peaks. The proposed model was effectively used to determine fifteen minor and trace elements in geological materials resulting in remarkably low detection limits. However, the result would have improved further if the approach was incorporated with chemometrics for spectral processing and quantitative analysis.

The use of scattered radiation for matrix effects correction is well known (Nielson, 1977). It was applied to samples of varying composition by utilizing for instance, recently, Monte Carlo simulation of XRF spectra in studying photon scattering at high X-ray energies (Vincze *et al.,*

1999). The simulation data was in agreement with data that was obtained from samples of copper, aluminium and polypropylene which were excited by monochromatic synchrotron radiation and an HPGe detector. Multiple Rayleigh and Compton scattering were witnessed in the light matrices of polypropylene and aluminium and Monte Carlo code was successfully used to model these effects. Matrix correction using scattered radiation involves the use of the intensity of the scattered radiation to correct for perturbations of the fluorescence intensities.

In other spectroscopic techniques, many approaches based on chemometrics have been studied and applied to non-linear multivariate calibrations, which could be applied to XRF analysis. Matrix effects corrections in X-ray fluorescence spectrometry to perform accurate quantitative and qualitative analysis is the same as solving the non-linear effects in multivariate calibration (Kaniu, 2011; Okonda, 2015).

2.3 Review of analytical chemometrics in spectroscopy

The development of modern analytical instrumentation has brought up scientific challenges especially in the efficient utilization of the data acquired. Chemometrics uses statistics and mathematical procedures to obtain maximum information by analyzing data obtained from spectroscopic measurements (Kurt and Peter, 2008).

Chemometrics methods can extract vital features (for instance underlying chemical properties) from complex spectral data and therefore can be developed for use in performing rapid and stable analyses. Chemometrics methods have an advantage of representing of multivariate data into few dimensions in a graphical interface. Incorporating chemometrics into analytical spectroscopy has other advantages which include the ability to reduce spectral noise, handle spectral interferences and outliers, and to achieve multivariate calibration (Kowalik and Einax, 2006).

In an attempt to overcome the limitations of the classical EDXRF techniques discussed in section 2.2, multivariate chemometrics techniques such as principal component analysis (PCA), partial least squares (PLS), artificial neural network (ANN), wavelet transform (WT), etc., have been applied to analyze XRF spectra for both multivariate calibration, pattern recognition and spectral preprocessing. The PLS and ANN techniques are characterized by calibration and prediction steps. During the calibration stage, a model is build, describing the relation between the EDXRF spectrum and the analyte concentration for a given set of samples. During the prediction step, the actual analysis, elemental concentrations in the unknown samples are determined. Compared to the classical methods like FP, these techniques enables spectrum evaluation, qualitative and quantitative analysis in one step (Kaniu, 2011).

Antonio and Bruno (2001) were able to fit EDXRF spectra with a genetic algorithm (GA). The GA technique used was based on a set of genetic operators. It worked well even when the knowledge of the peaks was incomplete. The result obtained was superior with respect to a standard implementation of a Marquardt-Levenberg algorithm. However the method had a drawback in that the speed of convergence was slower than the Marquardt-Levenberg technique.

Daniel and Thomas (2011) used handheld EDXRF spectrometry in the analysis of fossil bones from which important spectral information was obtained. Through statistical analysis of spectral data (i.e. normalization and mean centering) and multivariate processing (PCA) it was possible to interpret trends in the dataset. The two sample sites were segregated by PC1 score values and the distribution of score values was associated with geological control over bone chemistry.

Wavelets transforms (WT) have progressively depicted remarkable application in the area of analytical chemistry. WT is usually used instead of other algorithms that exist for enhancing

resolution, noise reduction and compression of data. This is because of its efficiency, availability of large number of functions and pace in data treatment. Discrete wavelet transform (DWT) was effectively used for reducing noise in short wave NIR reflectance spectra and the resulting analyzed spectra were used as the input to support vector machine (SVM) (Vu, 2014). Signal processing of spectra by WT is nonetheless under development.

A technique based on wavelet de-noising was applied to quantum-cascade laser spectrometer for *in situ* and real-time atmospheric trace-gas measurements (Li *et al.,* 2012). The wavelet digital-filter technique in signal processing demonstrated to achieve better precision measurement with better detection without decreasing the fast temporal response, unlike other normally used digital-filter techniques (i.e., Wiener filter, moving average and Kalman filter). It was found that continuous wavelets transform (CWT) has better resolution of space-time and is moderately easier to perform as compared to DWT for reduction of the varying background in NIR spectra (Ma and Shao, 2004). However wavelets transform has been hardly applied to XRF spectral processing.

X-ray fluorescence spectrometry combined with chemometrics has been applied to forensic analysis i.e., glass fragments identification at crime sites. Glass fragments retrieved from clothes of people who were suspected of committing offences, were rapidly analyzed by energy dispersive X-ray fluorescence spectrometry as a non-destructive method. Unlike routine refractive index which was not able to identify any of the samples, 112 pairs out of 129 samples were identified by Linear Discriminant Analysis (LDA) and neural networks (Grzegorz, 2007). The weakness of this approach is that the spectral data was not preprocessed and therefore the results were less accurate.

Long *et al.* (1998) utilized ANN for deconvolution of low resolution (equivalent to spectral overlap e.g. of two speciation of the same element) XRF spectra. Instead of analyzing peak and fitting the experimental results to a mathematical function as is common when conventional algorithms are used, the ANN used the entire spectrum, comparing its shape with the patterns learnt during the network training. The results exhibited the potential for ANN to be used in XRF spectrum deconvolution. The combinations of multi-variables gives better correlation than using single variables (Kurt and Peter, 2008).

Bos and Weber (1991) compared two ANN optimization procedures namely genetic algorithms and backward error propagation for quantitative XRF spectrometry. The authors used already published data from thin-film Fe-Ni-Cr samples for which the genetic algorithm approach performed poorly. They found that (as expected) the larger the training set, the better were the predictions. They furthermore, pointed out that samples outside the training set could not be predicted. Neural networks were nonetheless shown to be useful as empirical calibration models with good quantification abilities with sufficient accuracy because of their ability to correct nonlinearities. The method might have given better results if a preprocessing step was added to the analytical procedure before fitting data to ANNs.

Bueno *et al.* (2006) designed a quantitative method for elemental analyses in food combining XRF analysis with PLS. Six PLS models were created based on XRF spectra and the known elemental concentrations of metallic (Fe, Mg, Ca, K, Mn, Zn) analytes. The study demonstrated the ability of PLS in conjunction with XRF analysis to predict macro (low-Z) element concentrations of Ca, K and Mg.

Simultaneous determination of rare earth elements (Pr, Nd and Sm) simulated by their (similar) chemical properties and hence difficulty in quantitative analysis was conducted by Schimidt *et al.* (2003). Rare earth elements produce spectra which are composed of very complex

absorption and emission and thus separation of such spectra is challenging due to the absence of selective reagents. The oxides of these metals were diluted in a silica gel matrix and acquired spectral data were treated by ANNs and PLS. The back propagation-single component (BP-SC) ANN showed the best performance by giving the least root mean square error of prediction (REMSEP) of 17.5 % for Pr, 12.5 % for Nd and 12.6 % for Sm. BP-SC ANN made it possible the resolution of non-linearities in spectral calibration, thus effectively overcoming the challenge of spectral overlaps.

In another study, Facchin *et al.* (1999) showed the predominance of ANN in the modeling of complex EDXRF spectral data for the simultaneous determination of (and therefore resolution) of Pb and S in solid samples. The study was stimulated by the common nature of intense overlapping of spectral lines for these elements which complicates their direct and simultaneous determination. The applicability of ANNs to this problem was evaluated and compared to five other data treatment methods; polynomial PLS (POLYPLS), partial least squares (PLS), PLS neural networks (NNPLS), linear regression (LR) and corrected intensity (CI). The standard errors of prediction (SEP) for ANNs compared to the other (linear) methods was lowered by 35 % for Pb and 100 % for S. The study demonstrated the applicability of ANN in modeling non-linear (matrix effects) data in EDXRF spectroscopy.

Kessler *et al.* (2002) demonstrated the applicability of multivariate chemometrics calibration to extract latent relationships in EDXRF spectra, and the feasibility of forming a chemical classification based on characteristic fluorescence and scattered (coherent and incoherent) X-ray radiation. PCA was exploited to distinguish between spectra of different chemical compounds and also played an important role in the estimation of influences of the different physical (mean atomic number, μ/ρ, main element concentration) and technical parameters (grain size, sample position, packing wall thickness) on spectra of different chemical

compounds without packing. A multivariate relationship was built between the physical parameters of chemical compounds and the location of their spectra in PCA data space, thus building a spectra library data set which was used for fast and safe identification of chemical compounds in powdered chemicals packing plants.

Maria *et al.* (2005) investigated mild variations in organic matrices which results from alterations in X-ray Raman scattering. Principal component analysis (PCA) and hierarchical cluster analysis (HCA) were utilized to picture out the effects. Conventional energy-dispersive X-ray fluorescence spectrometer was used, whereby organic compounds produced intense scattering of the X-ray from the source. Chemometrics showed the ability to classify natural samples despite their complexity.

The greatest potential of application of chemometrics in X-ray spectrometry has been demonstrated in the fields of multivariate calibration, prediction and pattern recognition. This was shown in paint classification by use of XRF and chemometrics, resulting in a reliable, fast and easy-to-perform method (Verbi *et al.,* 2005). Moreover, analysis of low-Z elements in organic liquid samples by principal component regression (PCR) coupled with EDXRF spectrometry using backscattered radiation revealed the parts of the scatter peaks that are important in determining H, C and O (Molt and Schramn, 1999). Further, utilization of partial least squares regression (PLSR) for quantitative EDXRF analysis of liquids has demonstrated the potential of chemometrics in qualitative analysis of organic and inorganic matrices (samples) (Lemberg and Van Espen, 1999). The enhanced analytical ability resulting from integrating chemometrics in XRF spectroscopy analysis of materials has moreover been shown in soils classification (Custo *et al.,* 2002), monitoring of sucrose conversion (Karen *et al.,* 2007), plant classification, characterization of honey (Enrich *et al.,* 2007), classification of archaeological pottery (Bakraj, 2006) and Portland cement characterization (Karen *et al.,*

2006). Chemometrics in this case was mainly utilized to perform quantitative analysis of analytes signals with interfering peaks.

Utility scatter peaks of XRF spectroscopy in conjunction with use of chemometric techniques has been demonstrated in the determination of low-Z elements (Karen *et al.,* 2006). However, the potential of chemometrics XRF spectroscopy is still under development and has not yet largely been fully exploited. Robust multivariate chemometrics has analytical utility in EDXRF as it can enable the transformation of spectral responses directly to concentrations and other material properties contained in the spectral characteristics of the matrix (Verbi, 2005; Kaniu, 2011).

2.4 Robust chemometrics approach to spectroscopy for complex matrix analysis

It is clear from the above studies that chemometrics can solve a number of problems that pose a challenge when using traditional spectroscopy. However, the utility of chemometrics coupled with XRF spectroscopy in material analysis remains largely unexplored. XRF spectra with elevated background and poorly resolved peaks are amenable to multivariate chemometrics spectral quantitation. Simple univariate, classical linear regression and multivariate regression methods are limited to linear relationships and therefore cannot be easily applied. The major drawback of XRF is its inability of distinguishing elements in different valence states. This is because of poor resolution of K_α fluorescence peaks of the analytes.

Chemometrics-assisted EDXRFS spectrometry was utilized for direct rapid analysis of complex matrix liquids (Angeyo *et al.,* 2012). The technique exploits X-ray fluorescence and scatter profiles obtained non-invasively from complex matrices to correct for matrix effects observed in the deconvolution of intensity signals to concentration and to develop a multivariate calibration strategy for analyses of chemical property of the sample matrix (concentration and speciation) utilizing multivariate chemometrics techniques (Okonda, 2015).

However, robust chemometrics spectral preprocessing techniques have not been well exploited and/ or developed to analyze spectra for accurate determination of trace element and their speciation. In this study therefore, robust chemometrics-assisted energy dispersive X-ray fluorescence and scattering (EDXRFS) spectroscopy is proposed as a non-destructive method for direct rapid and simultaneous determination of trace elements concentration and speciation which has hardly been exploited in EDXRF. The method reported in this study makes hybrid use of WT, PCA and ICA for spectra preprocessing to optimize them for qualitative analysis by PCA and ICA and quantitative analysis by ANN and PLS. The method has enabled trace and major biometal and speciation analysis in soft body tissue. The technique affords rapid analyses of large number of complex matrices for trace elements (Na, Mg, Fe, Mn, Co, Zn, Cu) in a manner that is applicable to disease characterization and diagnostics in native soft body tissue.

CHAPTER THREE: THEORETICAL BACKGROUND

3.1 Chapter overview

In this chapter, we outline the basic principles of X-ray fluorescence and scattering spectroscopy for analysis of complex matrix samples for determination of elemental content. In addition, we outline the principles of multivariate chemometrics techniques i.e. PCA, ICA, WT, ANNs and PLS for their utility in modeling XRF spectra to develop a robust chemometrics approach to complex matrix materials analysis.

3.2 Principles of X-ray fluorescence (XRF) spectroscopy

3.2.1 Interaction of X-rays with soft body tissue

When X-rays propagate through soft body tissue, a reduction in the intensity of the incident X-ray beam occurs. This attenuation process leads to emission of secondary X-ray radiation by the soft body tissues. Two main processes are involved in the interaction. The X-rays may be photoelectrically absorbed. This process may cause (i) fluorescent emission of X-rays characteristic of the elements in the tissue, with energies that are independent of the primary radiation source ($E_{\text{fluorescence}} < E_{\text{primary radiation}}$) and (ii) emission of photoelectrons and Auger electrons. The X-rays may also scatter. This scattering can be (i) coherent if there is no loss of energy for the incident X-ray photon ($E_{\text{coh scattering}} = E_{\text{primary radiation}}$) or (ii) incoherent when there is a small loss of energy ($E_{\text{incoh scattering}} < E_{\text{primary radiation}}$) (Jenkins, 1986).

3.2.2 Photoelectric absorption of X-rays

The ability of X-rays to penetrate and even pass through matter is one of their best known properties. Part of this radiation will be absorbed by the material. The decrease in intensity (attenuation) of the incident X-ray beam is directly proportional to the thickness of the soft tissue. Considering an almost ‘infinitely’ thin layer of pure, single element material of thickness

dx and a monochromatic beam of X-ray radiation of intensity I_0, the incremental loss of intensity, dI will be:

$$dI = -\mu I dx \tag{3-1}$$

where μ is the linear absorption coefficient. If τ represents the photoelectric absorption and σ the total scatter cross section, then $\mu = \tau + \sigma$. In most cases, the scattering effect is small compared with the absorption effect and, for practical purposes, it is possible to say that $\mu \approx \tau$.

3.2.3 Scattering of X-rays

Scattering of the incident radiation can also occur when an X-ray photon interacts with an atom of the sample. The scattering cross-section (σ) is mathematically described as:

$$\sigma = Zf^2 + (1 - f^2) \tag{3-2}$$

where f is the atomic scattering factor (ratio of the amplitude of the wave scattered by an atom, to the wave scattered by a free electron), and Z is the atomic number.

Coherent scattering is always present in the X-ray fluorescence spectrum and is the main cause of background signals. Incoherent scattering or the Compton effect (second term of equation 3-2) occurs when the X-ray photon loses part of its energy in the collision. The electron involved in the scattering process will gain some of the energy of the photon, as the total momentum is to be maintained. The wavelength of the scattered photon will increase according to the expression:

$$\Delta\lambda = 0.0243(1 - cos\ \phi) \tag{3-3}$$

where ϕ is the angle of scatter and λ is the wavelength measured in Å. This type of scatter tends to displace the continuum background to the low energy regions. Broad peaks, corresponding to the characteristic primary wavelengths of the anode material of the X-ray tube, can also be found in that area of the spectrum.

Incoherent scatter becomes more manifest for those samples with a high proportion of low-Z elements (i.e. $Z < 13$). Since all elements present in the sample contribute to both coherent and incoherent scattering (scatter peaks), the measured scatter peaks may be used to obtain information about overall sample composition (Beckoff *et al.*, 2006).

3.2.4 Quantitative approaches in EDXRF analysis

In quantitative EDXRF spectroscopy, the net fluorescence spectral peak intensities are deconvolved and mathematically related to the corresponding elemental concentrations. The normally used procedure is to calibrate the spectrometer first by measuring certified reference materials. Through calibration procedure, the relationship between the concentrations of analytes and the corresponding intensity of the fluorescent lines of the elements is established. Once the relationship is known, the concentrations of unknown can be determined.

3.3 Matrix effects and matrix correction models

The transformation of X-ray fluorescence intensities into elemental concentrations is not direct. The number of photons emitted from the sample is also affected by physical properties of the sample (‘‘matrix effects’’). This takes place in three categories namely X-ray absorption, X-ray enhancement and sample macroscopic effects.

Absorption process attenuates the secondary X-rays from the sample. In order to calculate the absorption for a multi-element sample, the composition of the sample must be known. For analysis of an unknown sample, an iterative procedure is used. Sample macroscopic

effects comprise of effects of inhomogeneities of the sample, and non-uniform representation of the analytes at the sample surface.

Enhancement arises when the secondary X-rays emitted by a heavier element have sufficient energy to excite additional secondary emission from a lighter element (Potts and Webb, 1992). As a result, characteristic peak intensity of an element is a function of both the element's concentration of the other element in the sample. The use of a simple univariate linear function to correct matrix effects is often not viable and alternative approaches have to be used, ranging from the use influence (alpha) coefficients to fundamental parameter (FP) models based on a comprehensive knowledge of characteristics of the instrument and the sample (Rousseau, 2006).

3.3.1 Fundamental parameters (FP) model

The fundamental parameters method is based on mathematical expressions, quantifying fluorescence emissions in terms of fundamental physical parameters of X-ray emission process and instrumental as well as sample parameters. Fundamental parameters are the mass absorption coefficients, absorption jump ratios, fluorescence yields, line emission probabilities and characteristic X-ray energies (Nielson, 1977).

Since the composition of the sample is required to be known before creating a calibration model, the FP method is affected by uncertainties in mass absorption coefficients and fluorescence yields of the individual elements which results in a limited accuracy of the results of elemental composition. An additional problem is that of the so- called ‘‘dark matrix’’ (low-Z matrix). In this respect, low–Z matrix samples pose a problem since most of the matrix consists of organics and other light elements. Moreover, the FP approach requires homogeneous and flat samples (Markowicz, 2011). The proposed method in this study incorporates robust chemometrics approaches to solve this geometric limitations.

3.3.2 Influence Coefficient model (IC)

Influence coefficients quantify matrix effects individually, i.e. the matrix effect of element *j*, the matrix effect of element *k*,…on analyte *i*. The concentration of an element *j* can then be written as a function of its characteristic line intensity, corrected for the concentrations of other elements by influence (alpha) coefficients (Lachance, 1999).

The corrected concentration of element *i* (C_i) is then estimated from elements concentrations.

$$C_i = R_i \left(1 + \alpha_{ij} C_j + \alpha_{ik} C_k + ..\right) = R_i \left(1 + \sum_j C_j\right) \quad (3\text{-}4)$$

R_i is the fluorescence intensity of element *j* relative to the pure analyte *i*. In general the coefficients methods require more standards than the fundamental parameters method. If the coefficients cannot be calculated from theory, at least one standard per influence coefficient is needed (one less if elimination of the major element is applied). Quite often, these empirical coefficients are system-specific, i.e. not transferable to other systems (Lachance, 1999).

3.3.3 Compton scatter method

This is an alternative method that is used to correct matrix effects which does not require major elements composition data and is especially applied in analysis of trace elements. The method relies on the direct measurement of intensity of the Compton K_α scatter peak from the sample undergoing analysis rather than basing on a mathematical model of absorption or enhancement effects. This method was derived from the observation that the intensity of Compton scatter peak is inversely proportional to the mean mass absorption coefficient of the sample (Andermann and Kemp, 1958). The matrix effects can be corrected by normalizing of the fluorescence line intensity to that of the Compton scatter peak with measurements being made on the sample being analyzed. Compton scatter approach has a limitation in that it fails if the measurement of analyte line and Compton scatter peak intensities are made on different sides of the major element's X-ray absorption edge.

3.4 Chemometrics in XRF analysis

It has been noted that spectral lines from the X-ray tubes excitation source can be exploited to achieve classification and quantification of organic and low atomic number elements traditionally not possible in conventional EDXRF. This is achieved by analyzing data using chemometrics tools such as PCA, ANNs and SVM (Sussulini *et al.,* 2009).

With chemometrics, it is possible to transform the EDXRF spectral data directly into elemental concentrations using implicit modeling techniques. In the present study, the applicability of robust chemometrics techniques in the modeling of EDXRF and scatter spectral data to perform trace biometal and speciation analysis of selected elements in complex matrices materials namely soft body tissue has been demonstrated.

3.4.1 Principal component analysis (PCA)

PCA is a bilinear modeling which gives an interpretable overview of spectral data in a reduced multidimensional space. The method is used to perform dimension reduction of the original data, modeling of data, detection of outliers, selection of main variables, classification, validation and prediction of samples. PCA creates new coordinate axes called principal components (PCs) that are uncorrelated; measuring different dimension (Brereton, 2003). Thus chemometrics is very important when analyzing complex matrices as it helps to visualize and obtain information from the original spectra data by applying PCA. Centering of original data matrix **X** is achieved by subtracting the mean spectrum x from each spectrum, after which the resulting centered data matrix $\mathbf{X_c}$ is decomposed to a score matrix **T** and loading matrix **P** (Brereton, 2003).

$$\boldsymbol{X}_c = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} \tag{3-5}$$

The residuals are gathered in a matrix **E** and that the position of the samples in the new coordinate system described by **T**. The loading matrix (**P)** defines the new axis (Virendra *et al.*, 2011).

The importance of PC is expressed in terms of residual and explained variance. The residual variance explains the variation in the data that remains unexplained once the current PC has been taken into account whereas the explained variance, mostly measured as a percentage of the total variance in the data, is a measurement of the proportion of data variation accompanied for by the current PC. Variances are used to express how much of the information available in the data is described by the model.

Loadings reflect both the contribution of variable to PC and how well the PC takes into account the variation of the same variable over the whole points in the data. In geometric terms, loading is simply the cosine of the angle lying between the variable and the current PC. Therefore smaller angle has a larger loading. Hence loadings can only vary from -1 to +1.

Line plots of the loadings illustrate the importance of the original variables for each PC and may be used to deduce the quantitative differences causing the clusters or trends in the data. The loading vectors are considered to be hidden spectral profiles that cuts across sample spectra- i.e. the loadings contain qualitative information such as elemental concentrations giving rise to the spectra peaks.

2-D scatter plots of the score vectors (in different combinations) depict covariance between samples, providing a data overview. Patterns and clusters of samples as well as outliers are easily identified in the score plots which enable the exploration of the expected and unexpected trends in the data.

3.4.2 Wavelet transforms (WT)

Wavelet transform (WT) is one of the latest techniques for signal processing. WT is mathematical functions that divides up data into components of different frequency and studying the components with resolution matched to their scales (Graps, 1995). WT signal is reliable in signal-processing due to its unique properties (i.e., wavelets can be regular or irregular, sharp or smooth and symmetric or asymmetric). They have ability to extract transient features of signals and facilitate analysis of signals at multiple resolutions. X-ray fluorescence (XRF) spectra can been analyzed based on continuous wavelet transform filters to filter the signal and noise components of the spectrum (Sergey *et al.*, 2011).

Wavelet analysis is efficient in uncovering aspects of the data that alternative signal analysis techniques don't, for example aspects like data trends and self-similarity. Wavelet analysis is capable of reducing noise in a signal without considerable loss of information.

WT is categorized into discrete wavelet tools and continuous wavelet tools. Discrete wavelet tools are used for both signal analysis and signal processing such as peak detection, noise reduction and data compression. On the other hand, Continuous wavelet tools are used for signal analysis, such as time-frequency and self-similarity analysis (Goswami and Chana, 2011).

Considering a time varying signal, f (t), the Continuous Wavelet Transform (CWT) of such signal can be expressed as:

$$W_f(\tau, s) = \frac{1}{\sqrt{|S|}} \int_{-\infty}^{+\infty} f(t)\Psi(\frac{t-\tau}{S})dt \qquad (3\text{-}6)$$

where τ and *S* are refer to as translation (or time location in this case energy) factor and the scaling (or dilation) factor respectively.

To achieve multiresolution analysis, the wavelet must be efficient in both translation and scaling (i.e. stretching and shrinking). The factor /s/$^{-1/2}$ represnts the energy normalization across the various scales, whereas Ψτ,s (t) is obtained by translations and dilations of a single function Ψ (t) called ‘‘mother wavelet’’:

$$\Psi_{\tau,S}\ (t) = \frac{1}{\sqrt{|S|}}\Psi(\frac{t-\tau}{S}) \tag{3-7}$$

To measure the performance of the WT algorithm in spectral preprocessing, SNR was computed using the equation (3-8) (Kabir and Shahnaz, 2012). The definition of SNR is based on the height of the peaks and an estimation of the noise (Elise *et al.*, 2010).

$$SNR\ = 10log_{10}\sum_{i=1}^{N} X^2(i) \tag{3-8}$$

where $X(i)$ is the EDXRFS spectra and N is is the width of the EDXRFS signal.

Wavelets separate the signal peaks from the normally high frequency spectral noise and reduces the baseline without necesarily modeling the baseline. The separation of signal peaks from the noise presents an opportunity to build a model that can determine the total element from XRF spectra without being distracted (Sergey *et al.*, 2011). WT can be used to reduce the high background noise from the EDXF spectrum.

3.4.3 Artificial Neural Networks (ANNs)

Artificial neural networks are input-output–based empirical models that are used for modeling relationships for multi-input and multi-output. The information they contain and process is

distributed over a large number of model parameters which accounts for their great flexibility. ANNs mimic human cognitive process and as such inconsistent (fussy) data (Liu *et al.*, 1993). These characteristics cannot be expressed through a classical well-defined algorithm; rather they are based on experience. It is commonly argued that ANNs can implement solutions to problems that are difficult to solve in a classical algorithmic way (Jose *et al.*, 2009). The potential of ANNs as a modeling tool for multivariate calibration is well established; efforts are now directed on developing proper methodologies that ensure ideal conditions for calibration are met (Despagne and Massart, 1998). ANNs can be used to build empirical multivariate calibration models of the form.

$$\boldsymbol{Y} = F(\boldsymbol{X}) + \varepsilon \tag{3-9}$$

with F being the network function, **X** matrix of the analyte measurements performed on a series of n samples, **Y** the vector matrix containing sample responses e.g. concentration of target analytes and ε, the error of calibration.

ANNs are based on the principle that each neuron receives a series of inputs that are dynamically weighted. The ANNs compare the weighted sum of its inputs to a given threshold and apply a non-linear function to compute the output. The calibration of ANN comprises of iteratively optimizing the weights as a function of the difference between the predicted and the known values of concentration in calibration samples. Lastly the ability of the network to make prediction is evaluated from the validation data set (Beale and Jackson, 1991). Various ANNs techniques exist e.g. Back propagation networks BP-ANN, Kohen self-organizing maps (SOMs) and counterpropagation networks (CP-ANN). The commonly used technique is BP-ANN because of its ability to be applied in regression and supervised classification.

3.4.3.1 Back propagation networks (BP-ANN)

Error back-propagation artificial neural networks (BPNs) constitute without doubt the paradigm most widely applied to develop ANNs nowadays (Ramadan *et al.,* 2005). It has a capability to solve a wide range of problems. The name refers to feed-forward ANNs which are trained using an error back-propagation algorithm. BPNs were developed in the 1960s (Werbos, 1974). BPNs 'learn' how to relate a predefined set of spectra in the calibration set with their associated target values (e.g. concentrations) using a supervised two-phase cycle. The neural network consists of three layers of nodes: an input layer, a hidden layer and an output layer in the case of XRF analysis.

Back propagation (BP) training functions is utilized in training feed forward networks (non-linear regression). The training process consists of four stages namely; (i) collection of training data sets, (ii) creation of new network, (iii) training of the network and finally (iv) simulating the trained network response to new unknown inputs. To determine the number of neurons to be used in the hidden layer, the lowest value of minimum square error (MSE) and network performance for each trained network are utilized.

In this study, ANNs was used for multivariate calibration. The input variables are the EDXRFS spectra at selected *n* energies depending on the elements of interest. The input variables were weighted and transferred to one or several hidden layers which are internal to the network.

3.4.4 Independent component analysis (ICA)

ICA is a deep-rooted statistical signal processing technique which decomposes multivariate signals into statistically independent components with little loss of information. In EDXRFS spectrum, the different energy (keV) positions can be considered to be different independent variables and while intensities relates to the magnitudes of the variables. ICA model can be illustrated using the following mathematical equation:

$$x = A * s \tag{3-10}$$

where *x* refers to the mixed signal source vector, *A* the mixing matrix and *s* is the source signal.

The objective of ICA technique is to determine a demixing matrix A^{-1} that separates the source signal vector S into statistically independent set of sources. Therfore, this implies that A^{-1} is the inverse matrix that goes from the source vector S to the signal vector *X*. Computed S is based on the following expression (Guonqing *et al.*, 2008)

$$S = A^{-1} * X \tag{3-11}$$

The algorithm in this work is based on the *FASTICA*. This method was used in EDXRFS data dimension reduction, fluoreescence peak resolution and denoising without having a negative effect on the appropriate information obtained.

3.4.5 Partial least squares (PLS)

PLS is a multivariate regression method for creating a relationship between a set of independent variables *X*, which can be for example chemical measurements and a set of dependent variables *Y* illustrating the class in which the object belongs; where the object is *X* and the class is *Y*. Hence PLS may be utilized to model a relationship between measured EDXRFS spectral data and known concentration for trace bio-metals. This is different from conventional quantitative EDXRF methods that are based on the deconvolution of net characteristic peak intensities. The spectral data variables may be collected as the matrix *X* with a row number equal to the number of samples and a column number equal to channels in the spectra.

The *Y*-matrix is used in approximating the response variables to make sure that the first components are the ones which are vital for predicting *y*-variables. The PLS method can model one component at a time (i.e. PLS-1) or it can determine two or more analytes (i.e. PLS-2)

simultaneously. In the case of PLS-1, there is only one response variable *Y* (Kurt and Peter, 2008). In PLS-2, a matrix *Y* instead of vector *y* is used, multivariate *x* - and *y* -data are given by the matrix *X* of dimension $n\ X\ m$ and the matrix *Y* of the size $n\ X\ q$

The purpose of calibration in PLS is to find a linear relation for prediction i.e. *Y* from *X*:

$$Y = X.B + E \tag{3-12}$$

between the *X*- and *Y*- variables, using an *m x q* matrix B of regression coefficients, and an error matrix *E*. *Y* represents the concentrations and *X* is the calibration described by their scores.

PLS method is also characterized by a calibration and prediction step. A model is built in calibration step which describes the relation between the response (in this case, the EDXRFS spectrum) and the analyte concentration for a given set of samples. The resulting model is then evaluated by a test set, with known values for *X* and *Y*. During the prediction step, the actual analysis, in this case the trace biometal concentrations in unknown samples are determined from the calibration model. The quality of the regression model is tested by, for example, root mean square error of RMSE which indicates the average error of prediction for the standard samples contained by the calibration model.

3.4.5.1 Validating the PLS regression model

Validating the PLS model is concerned with the selection of the optimal number of PLS components. Quite often, more than 1 PLS component are used to model *Y* by *X*. The optimum number of PLS components is estimated by cross validation (CV) method which divides data into different calibration and test sets. A calibration set creates an optimized model which is applied to the corresponding test set (Kurt and Peter, 2008). The relation between the scores then becomes

$$\mathbf{U = TD + H} \tag{3-13}$$

with **D** being a diagonal matrix with elements $d_1, d_2, \ldots\ldots d_n$, **T** score matrix and **H** the residual matrix. Furthermore, validation provides a value for the prediction error enabling the evaluation of the ability of the model to make prediction. An independent test data set is normally recommended for model validation because in CV, no test set data are used and thus the prediction performance is usually too optimistic.

3.4.5.1.1 Root Mean Square Error (RMSE)

The determination of the optimum number of PLS components is done by calculation of the RMSE. This value is a measure of the goodness of fit between the known and the predicted property:

$$RMSE = \sqrt{\frac{\sum(Z - \hat{Z}_i)^2}{n}} \tag{3-14}$$

with Z_i being the reference value, $\hat{Z}_i$ predicted value and n number of validation samples.

The RMSE value is then plotted against number of scores. The corresponding score number is at its optimum when the RMSE value reaches a minimum or plateau. The RMSE value is a measure of predictive performance of the PLS model and it gives the expected average error when constituent properties of samples with unknown composition are to be predicted.

CHAPTER FOUR: MATERIALS AND METHODS

4.1 Chapter overview

In this chapter, we present the procedures used in the preparation of standard samples for calibration training of spectral data for subsequent sample spectral data analysis for trace biometal and speciation determination. The utility of robust chemometrics methods using X-ray fluorescence spectra for quantification and speciation of trace elements has been investigated using calibration models developed from prepared simulate samples. X-ray fluorescence spectral data were obtained after sample irradiation for 50 s and 100 s live time using Amptek X-123 Mini-X spectrometer. Quantification method was validated for accuracy of measurements by analysis of NIST Oyster tissue 1566b as a certified reference material (CRM). All chemicals used in the sample preparation in this study were of analytical grade.

4.2 Energy dispersive X-Ray fluorescence (EDXRF) instrumentation

In this study the Energy dispersive X-ray fluorescence spectrometer (Amptek X-123 spectrometer Mini-X) was used for measurements. The spectrometer consists of the following:

(a) The Mini- X which includes the X-ray tube, the power supply, the control electronics and the USB communication to the computer. It features 50 kV/ 80 µA power supply, a silver (Ag) transmission target and Be end window with collimator and filters to facilitate its use in XRF. To simplify further the use of Mini-X an AC adaptor is provided to supply the 12 VDC needed to power the system.

(b) The X-ray spectrometer detector and signal processor consists of silicon drift detector which is air cooled mounted behind the Be window on the front of the extender. Resolution of detector is 125 eV FWHM at 5.9 keV. The signal processor consists of a shaping amplifier, high performance low power digital pulse processor, a multichannel analyzer and power supply. The XRF spectrometer incorporates FP quantitative analysis software,

test stand with shielding and sample enclosure. The spectrometer MCA unit has a memory of 2048 channels.

All measurements were done in air and voltage were auto adjusted during acquisition of the spectrum. Sample exposure area was maintained at 2 mm diameter throughout the measurement in this study (Figure 4.2.1).

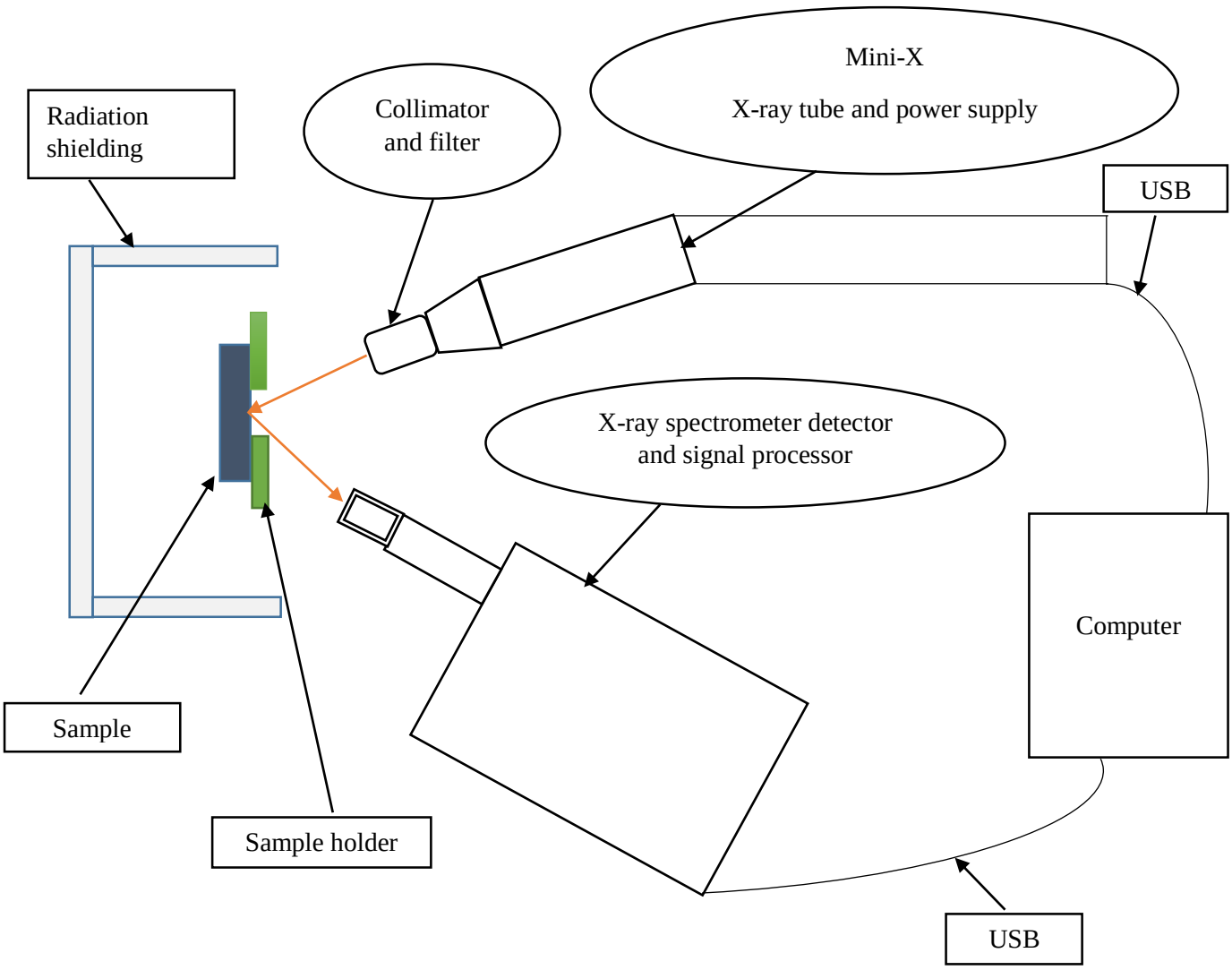


Figure 4.2.1: Schematic presentation of EDXRF instrumentation.

4.3 Sample preparation for calibration purposes

Tissue simulates were prepared in two sets separately one for lower speciation (Fe^{2+}, Cu^{+} and Mn^{2+}) and the other for higher Speciation (Fe^{3+}, Cu^{2+} and Mn^{7+}) using high purity paraffin wax as a "base matrix" which is spiked with multi-element stock solution of the elements (Fe, Mn, Co, Zn, Na, Mg and Cu) at various concentration as they occur in body tissues (John *et al.,* 2006; Mona *et al.,* 2013; Banas *et al.,* 2001; Leitao *et al.,* 2014). Table 4.3.1 shows the chemicals used in the preparation of tissue simulates.

Table 4.3.1: Chemicals used in preparation of tissue simulates

Source of cation	Chemical formula	Relative formula mass (g/mol)	Cation
Cuprous chloride	CuCl	99.00	Cu^{+}
Cupric nitrate	$Cu(NO_3)_2.3H_2O$	241.60	Cu^{2+}
Manganese (II) chloride	$MnCl_2.4H_2O$	197.70	Mn^{2+}
Potassium permanganate	$KMnO_4$	158.03	Mn^{7+}
Ammonium ferrous sulphate	$(NH_4)_2.SO_4FeSO_4.6H_2O$	392.14	Fe^{2+}
Ferric chloride	$FeCl_3.6H_2O$	270.32	Fe^{3+}
Cobalt (II) chloride	$CoCl_2.\ 6H_2O$	237.93	Co^{2+}
Zinc nitrate	$Zn\ (NO_3)_2.\ 6H_2O$	297.48	Zn^{2+}
Sodium acetate 3-hydrate	$CH_3COONa.3H_2O$	136.10	Na^{+}
Magnesium chloride	$MgCl_2.6H_2O$	203.30	Mg^{+}

Preparation of standards demands that (i) they should have a matrix that has same chemical and physical properties as the samples for analysis and (ii) that they are homogenous. In this case paraffin wax was a suitable base matrix chosen for preparation of tissue simulation because it mimics soft tissue properties due to its organic matrix.

4.3.1 Preparation of stock solutions and calibration standards

Various concentrations of analyte stock solutions were prepared by dissolving a mass *m* of the salt in 10 ml of ethanol and the contents thoroughly shaken to enhance homogeneity. Salts of ammonium ferrous sulphate and cuprous chloride which were not directly soluble in ethanol were dissolved in a mixture of HCl and ethanol mixed in the ratio 1:4 to attain solubility. This procedure was developed in the course of coming up with a method of dissolving insoluble salts in ethanol.

The mass *m* in grams (g), of the salt was weighted using the following formula;

$$m=\frac{C_S \, x \, V \, x \, rmm}{1\,000\,000 \, x \, ram \, x \, pp} \tag{4-1}$$

where C_s is the concentration of the stock solution, *ram* is the relative atomic mass of the trace element of interest, *V* is the volume of stock solution in ml, *rmm* is the relative molecular mass of the salt and *pp* is the percentage purity. Table 4.3.2 shows the analytes stock solutions used in this study.

Table 4.3.2: Elemental concentration of the stock solution

Analyte	Calculated mass *m* of the compound salt (mg)	10 ml Stock solution (µg/g)
Cu^{+}	12.600	800
Cu^{2+}	30.569	800
Mn^{2+}	22.260	600
Mn^{7+}	17.430	600
Fe^{2+}	142.570	2000
Fe^{3+}	97.788	2000
Co^{2+}	12.220	300
Zn^{2+}	91.420	2000
Na^{+}	2391.900	3600
Mg^{+}	170.700	2000

4.3.2 Spiking scheme of the trace biometals

After preparing the stock solution of the analytes, the following elements of interest; Mg, Na, Mn, Fe, Co, Cu and Zn were spiked in the concentration range in which they occur in human tissue (Banas *et al.,* 2001; Leitao *et al.,* 2014***)*** as shown in Tables 4.3.3 and 4.3.4 using the serial dilution formula below.

$$C_1 \times V_1 = C_2 \times V_2 \qquad (4\text{-}2)$$

where C_1 and C_2 are the concentrations of the stock solution and sample respectively, V_1 and V_2 are the volumes of the stock solution used and the volume of the sample made respectively. To attain homogeneity solutions were shaken thoroughly.

Trace elements were spiked at various concentrations and oxidation states using the following scheme in Tables 4.3.3 and 4.3.4. Calibration standards were labelled as LA to LO to represent low speciation and HA to HO to represent higher speciation (Tables 4.3.3 and 4.3.4). Co was spiked in only one oxidation state (i.e. Co^{2+}) while Na, Zn and Mg which have only one oxidation state appears in both low and high speciation sets for calibration purposes.

Table 4.3.3: Calibration design for simulate tissue for Na^{+}, Mg^{2+}, Cu^{+}, Mn^{2+}, Fe^{2+}, Co^{2+}, Zn^{2+}

Ion of interest	Sample concentration (µg/g)														
	LA	LB	LC	LD	LE	LF	LG	LH	LI	LJ	LK	LL	LM	LN	LO
Na^{+}	12	20	30	40	50	60	70	90	110	120	150	160	170	180	200
Mg^{2+}	10	20	30	40	50	60	70	80	90	100	120	130	140	150	160
Mn^{2+}	3	4	5	6	8	10	12	14	16	18	20	22	24	26	28
Fe^{2+}	12	20	30	40	50	60	70	90	120	110	150	160	170	180	200
Co^{2+}	1	3	4	6	9	10	12	14	16	18	21	22	24	27	30
Cu^{+}	4	10	15	20	25	30	35	40	45	50	55	60	65	70	80
Zn^{2+}	8	20	30	40	50	60	70	80	90	100	130	150	170	180	200

Table 4.3.4:Calibration design for simulate tissue for Na^{+}, Mg^{+}, Cu^{2+}, Mn^{7+}, Fe^{3+}, Co^{2+}, Zn^{2+}

Ion of interest	Sample concentration (µg/g)														
	HA	HB	HC	HD	HE	HF	HG	HH	HI	HJ	HK	HL	HM	HN	HO
Na^{+}	12	20	30	40	50	60	70	90	110	120	150	160	170	180	200
Mg^{2+}	10	20	30	40	50	60	70	80	90	100	120	130	140	150	160
Mn^{7+}	3	4	5	6	8	10	12	14	16	18	20	22	24	26	28
Fe^{3+}	12	20	30	40	50	60	70	90	120	110	150	160	170	180	200
Co^{2+}	1	3	4	6	9	10	12	14	16	18	21	22	24	27	30
Cu^{2+}	4	10	15	20	25	30	35	40	45	50	55	60	65	70	80
Zn^{2+}	8	20	30	40	50	60	70	80	90	100	130	150	170	180	200

4.3.3 Simulate sample preparation

The multi-elemental calibration standard samples solutions in vials were thoroughly stirred to enhance homogeneity and approximately 0.6 ml of the sample from each vial was pipetted onto steel mould and about 0.6 ml of acetone added to further dissolve the salts and also for dehydration. The sample content in the mould was placed in an oven for some time to raise their temperature to avoid instant solidification of molten paraffin wax on mixing, after which it was embedded in 0.6 ml of molten paraffin wax at 56 ºC for 10-15 minutes by placing the mould on an adjustable heater until all the acetone boiled off (evaporated) at this temperature leaving the elements of interest in paraffin wax. The mixture was stirred continuously using forceps to enhance homogeneity until cooled to room temperature.

The cooled mixture in the mould was further cooled in a freezer at -20 ºC for 10-15 minutes to enhance effective solidification. This resulted into a tissue simulate block which was removed from the mould and kept in a sealed labeled plastic bag.

The solidified simulate block samples were then sectioned into 5 µm, 10 µm and 20 µm specimens to represent thin, intermediate and thick samples approximately (Magdalena, 2012), for each element standard with a microtome and mounted on a 2 µm Mylar foil for analysis. The various thickness were based on the fact that absorption and enhancements effects vary from thin, intermediate to thick samples (Markowicz, 2011). The intensity of X-rays leaving light elements (low –Z) is low because the elements have low X-ray energies and mostly give rise to high Compton scatter and therefore these elements are affected most than heavy elements in thin samples (Kvetoslav, 1999). The decrease in intensity (attenuation) of the incident X-ray beam is directly proportional to the thickness of the absorbing matter. Standard reference biological material (NIST Oyster tissue 1566b) was prepared in three replications for use in method validation.

4.4 Sampling and preparation of native soft body tissue (domestic dog tissues)

In this study, approximately 50 g of healthy prostate tissues, healthy mammary tissues, prostate cancer tissues and mammary cancer tumors samples ten in number were surgically removed from prostate/mammary organs from domestic dogs at the Small Animal Clinic of Faculty of Veterinary Medicine, University of Nairobi, in June, 2013 following in patient admission and were later euthanized. The selected tissue samples were fixed in formalin for preservation of cellular details (Su *et al*., 2004) and later freeze dried at -80^0C for 48 hours (time optimized for removal of water and to stop cellular activities) in petri dishes.

The freeze dried samples weighing approximately 30 g were thawed at room temperature prior to sample preparation at Kenya Medical Research Institute (KEMRI) histology laboratory. The samples were then embedded in highly purified paraffin wax (Su *et al.,* 2004) to provide the supportive and protective aid during sectioning and sectioned to 5 µm, 10 µm and 20 µm thick tissue sample sections using a microtome and then mounted on a 2 µm Mylar foil for EDXRFS analysis with another thin section preserved for histopathological analysis.

4.5 Procedure of EDXRFS spectroscopy

The samples in this study were irradiated for 50 s and 100 s with same operating conditions. The spectral data were recorded in the energy range of 0-42 keV with an interval of 0.02 keV and data acquisition dead time of less than 1 %. Standard reference material (NIST Oyster tissue 1566b) were similarly prepared in triplicate and analyzed for method validation.

4.6 Robust multivariate chemometrics analysis of EDXRFS spectra

Analysis of EDXRFS spectral data obtained following sample irradiation time of 50 s and 100 s, 50 keV and 80 µA tube current was done using robust multivariate chemometrics methods. Based on the fact that full spectral data require a large memory and the analytes of interest do not cover the whole spectral region, spectral data compression was done by selecting the

appropriate spectral regions corresponding to each selected elements of interest for determination of their concentration and speciation as shown in Table 4.3.5;

Table 4.3.5: Selected spectral regions of interest of elements (Na, Mg, Mn, Fe, Co, Cu and Zn)

Element		Energy (keV)	Channel numbers
Na	K_α	1.000	48-52
	K_β	1.100	53-56
Mg	K_α	1.250	57-63
	K_β	1.300	64-68
Mn	K_α	5.895	272-288
	K_β	6.492	297-318
Fe	K_α	6.400	292-316
	K_β	7.059	323-345
Co	K_α	6.925	319-343
	K_β	7.649	352-373
Cu	K_α	8.041	372-392
	K_β	8.907	413-431
Zn	K_α	8.631	395-422
	K_β	9.572	442-466

For Compton scatter region, channel numbers (956-999) was selected. The inclusion of Compton scatter region is associated with the fact that most biological samples are composed of low-Z elements and other trace elements which are not observed in EDXRF spectral photo peak (Banas *et al.,* 2001). However, Compton scatter peak has potential utility in quantitative

determination of light elements when treated with chemometrics. The presence of enhanced scatter in the spectrum is an indication of dominance with dark matrix.

For determination of speciation of trace metals Cu, Fe and Mn, spectral data compression was done by selecting the appropriate spectral regions corresponding to each selected elements of interest shown in Table 4.3.5.

4.6.1 EDXRFS spectral preprocessing

4.6.1.1 Classical multivariate wavelets analysis

Wavelet analysis of spectral data was achieved using Matlab version R2014a software specifically for de-noising and resolution of $K_{\alpha 1}$ and $K_{\alpha 2}$. In this work, a data matrix consisting of columns corresponding to *n* samples and 2048 rows corresponding to elements energy channels was constructed. This formed multivariate signal ready for the analysis.

The critical issue is how to choose wavelets as different wavelets serve different purposes. For denoising and resolution enhancement, the appropriate wavelet function should resemble the shapes of true peaks, i.e. a cone shape in our EDXRFS peaks, thus symlet wavelet family with different parameters in MATLAB was chosen (Zhi *et al.*, 2012). The de-noising consists of the steps: (i) transforming the spectra to the wavelet domain using 2D wavelets and (ii) applying soft thresholding methods.

For the purpose of speciation analysis, the EDXRFS spectra was analyzed via classical wavelet transform analysis combined with ICA (WT-ICA) based on the fact that de-noising was carried out without smoothing the sharp structures (peaks). The result is a cleaned-up signal that still contains analytes information.

4.6.1.2 Combined use of wavelet transform (WT) and principal component analysis (PCA) for quantitative analysis

Signals from EDXRFS measurements are usually superposed with high frequency noise and / or trends (like offsets or baselines). Elevated background masks tiny photo peaks making it difficult to perform quantitative analysis as this distracts analytical model development.

In connection to the above challenge PCA was combined with WT to provide more enhanced noise reduction by combining with soft thresholding on detail coefficients as compared to wavelet analysis alone (classical wavelet analysis). This optimizes the spectra towards total elemental determination.

The SNR values of wavelet transform principal component assisted analysis of EDXRFS spectra K_α fluorescence peaks of the analytes were computed and compared with those from classical multivariate wavelet analysis and were used as a measure of algorithm performance.

4.6.1.3 Independent component analysis (ICA) of EDXRFS spectra

Independent component analysis in this work was done using *fastica* algorithm in Matlab version (R2014a software). Data dimension reduction, computing of the final independent components and resolution enhancement was done using the following stages (i) the spectral data was mean centered (ii) then whitened (iii) iteration until convergence (iv) computing independent components.

ICA has the ability to identify the components of a mixture and hence demixing is an effective denoising (Uzay *et al.*, 2003). SNR of K_α photo peaks (fluorescence) before and after ICA analysis were computed and used as a measure of the algorithm performance. ICA was conjointly used with WT for processing of spectra towards speciation analysis.

4.6.2 Multivariate calibration for quantitative analysis

4.6.2.1 ANNs calibration for elemental concentration determination

The ANNs analysis of spectral data in this work was realized using Matlab version R2014a software. The calibration dataset consisted of simulate tissue spectra which had been preprocessed via combined use of wavelet and principal component analysis. ANN network for the elements (Mn, Co, Fe, Zn, Na, Mg and Cu) was created. Back propagation neural network (BP-ANNs) with a two-layer (input and output with a hidden layer) was utilized. The training process comprised of the following four steps namely; (i) organizing of the training data (ii) building the network (iii) training the network and finally (iv) simulating the network response to new inputs (Kaniu *et al.*, 2012).

Training process of ANN was achieved after repeating the training of a number of times and the resultant trained network with the best performance (i.e. low MSE) was reserved for utilization. Simulating the output of the neural network with the measured outputs was used to test the results after network selection. A correlation coefficient, R^2, was used evaluate the performance of the network.

4.6.2.2 Multivariate calibration by PLS for elemental concentration determination

PLS in this study was performed using both selected fluorescence and selected fluorescence in conjunction with Compton scatter peaks in the energy range (0 keV – 42 keV) for the 30 soft body simulate spectra calibration samples within 2048 variables (energy channels). The samples and the variables were defined in terms of selected fluorescence region of interest (ROI). Compton scatter and total scatter regions were used to characterize the low Z-elements. The calibration dataset was 30 simulate tissue spectra which had been preprocessed via wavelet assisted principal component analysis.

The PLS model was developed from a training set of X-variables (spectral intensities) and Y-variables (known concentrations). These training data formed two matrices **X** and **Y**. The regression models were calibrated and validated by the CV method. The ability of the regression models to make prediction was then assessed using coefficient of multiple determination (R^2).

4.6.3 Multivariate calibration for elemental speciation analysis

4.6.3.1 EDXRF spectrum analysis by PCA

Prior to PCA analysis, the selected spectral region of interest (fluorescence signatures) were analyzed via combined use of WT and ICA for noise reduction and peak resolution. Principal component analysis (PCA) was used to identify possible patterns within the data for classification (speciation) and exploratory purposes. PCA was done using Matlab version R2014a software. PCA model was built for speciation analysis using simulate samples and the results validated by full cross validation.

4.6.3.2 Independent component analysis (ICA) of EDXRFS spectral data

For purposes of pattern recognition (classification), ICA was done using jade algorithm coded in Matlab version R2014a software. Prior to ICA analysis, the selected spectral region of interest intensities were analyzed via combined use of WT and ICA using *fastica* algorithm for noise reduction and resolution of $K_{\alpha 1}$ and $K_{\alpha 2}$ fluorescence peaks. ICA in this study was performed using the selected fluorescence (region of interest) in the energy range (0 keV – 42 keV) for the 30 simulate soft body simulate spectra calibration samples.

4.6.4 Summary of robust chemometrics-EDXRFS calibration approach

Figure 4.6.1 shows a summary of the robust chemometrics approach used in this study for trace biometal and speciation analysis of native soft body tissues.

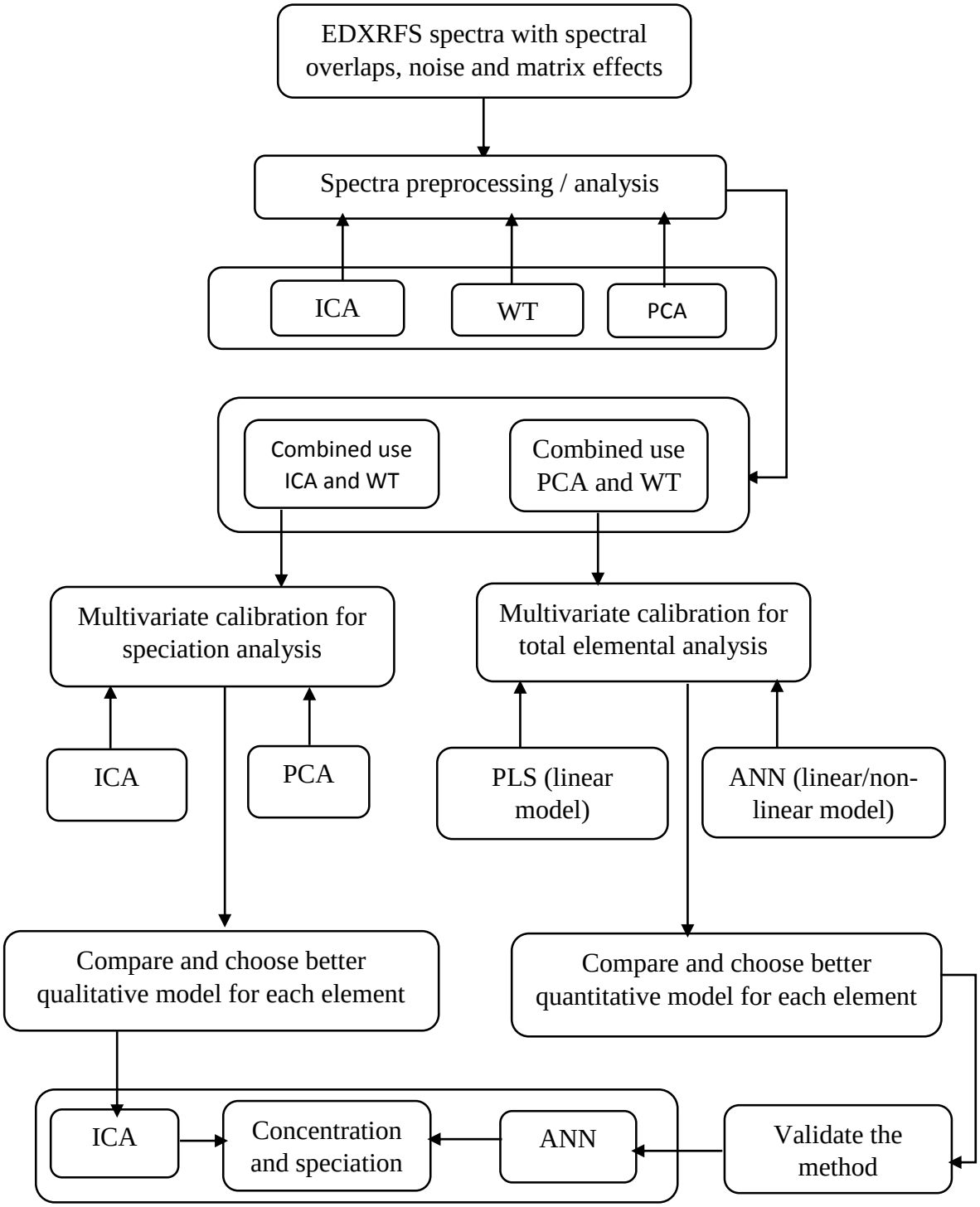


Figure 4.6.1: Schematic diagram in the development of a robust chemometrics analytical procedure.

CHAPTER FIVE: RESULTS AND DISCUSSION

5.1 Chapter overview

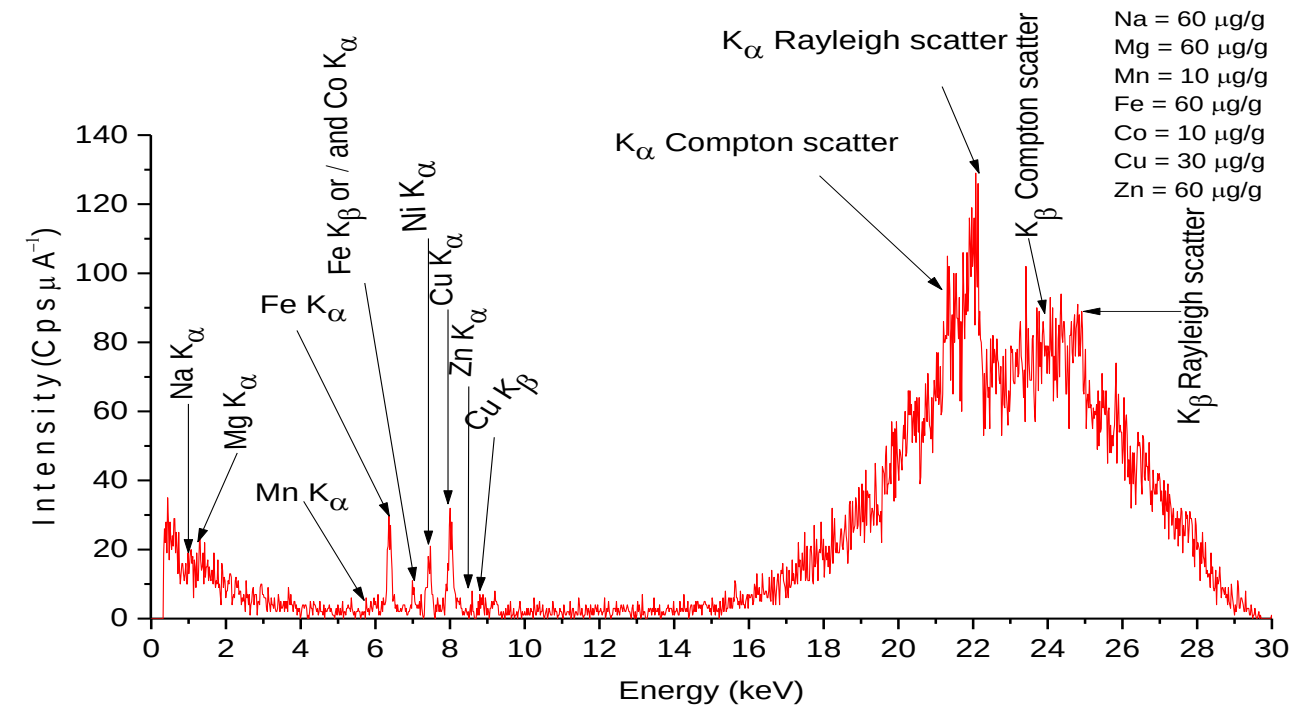
In this chapter, we present the results of the robust multivariate chemometrics approaches in EDXRFS analysis of trace biometals and their speciation in both model and native soft tissue. Included, also are results of EDXRFS spectral data of simulate tissue samples analyzed with (PCA, WT, and ICA) for spectral preprocessing to optimize for both qualitative and quantitative calibration. Results of quantitative multivariate calibration (PLS and ANN) of EDXRFS spectral data for elemental concentration and (ICA and PCA) for speciation analysis of simulate samples are presented. Finally, the application of the robust analytical chemometrics - EDXRFS method on both healthy and diseased (cancerous) native soft body tissue samples is reported.

5.2 Spectral preprocessing of EDXRFS data from simulate soft tissue

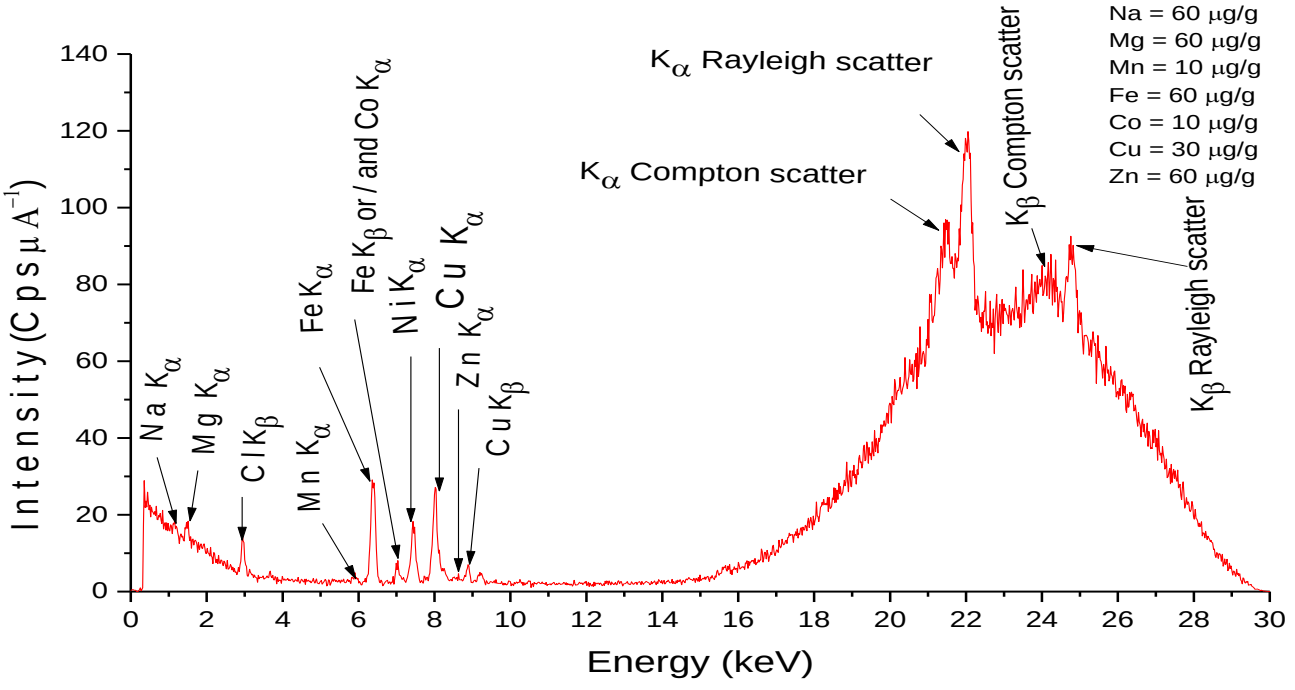
5.2.1 Classical wavelet (CW) analysis

Typical EDXRFS spectrum for trace (Mn, Cu, Fe, Zn and Co) and major (Na and Mg) elements before preprocessing, spiked in simulate tissue at concentrations shown in Table 4.3.3 and Table 4.3.4 respectively is shown in Figure 5.2.1(i). The fluorescence and scatter spectra were obtained by irradiating the simulate samples for 50 s live time.

Figure 5.2.1 (ii) shows classical wavelet analyzed spectrum for simulate tissue sample. To test the performance of the wavelet analysis, SNR of K_α peaks for each analyzed element was computed in Matlab before and after spectral preprocessing. The background noise after WT analysis is reduced with improved values of SNR (Table 5.2.1). Background is more pronounced in the typical sample spectrum as compared to the preprocessed one.



(i)



(ii)

Figure 5.2.1: (i) Raw spectrum and (ii) Classical wavelet (CW) analyzed spectrum of simulate sample 10 µm thick at a live time of 50 s.

The effectiveness of this preprocessing process depends on the selection of the mother wavelet and thresholding of the detail coefficients (Miguel and Lucia, 2011). SNR of K_α peaks for each element before and after CW analysis for 5 µm, 10 µm and 20 µm thickness were calculated using algorithm shown in Appendix 1 and are presented in Table 5.2.1.

Table 5.2.1: SNR for raw and classical wavelet (CW) analyzed EDXRFS spectra of simulate tissue samples at alive time of 50 s

Peak	Signal to noise ratio (SNR)					
	5 µm		10 µm		20 µm	
	Raw	CW	Raw	CW	Raw	CW
Zn K_α	0.13	0.83	0.34	0.98	0.72	2.00
Cu K_α	3.20	12.12	4.74	15.00	6.50	14.14
Co K_α	0.46	9.30	0.73	11.52	0.89	10.12
Fe K_α	4.91	20.32	6.40	23.49	9.86	22.31
Mn K_α	0.18	6.41	0.52	7.62	3.16	8.02
Mg K_α	0.29	2.30	0.81	2.89	1.05	3.03
Na K_α	0.16	2.07	0.86	2.41	1.61	3.12

Table 5.2.1 shows that sample thickness influences the value of SNR when raw spectra is utilized. The smaller the thickness the lower the SNR for unpreprocessed spectral data and more pronounced in low – Z elements due to low fluorescence yield (Kvetoslav, 1999) since Compton scattering is enhanced because of low X-ray absorption cross section of such elements hence high background radiation (Bueno, 2006).

Further, Table 5.2.1 shows an improvement in the SNR for the various thickness after preprocessing with classical wavelet (CW) for all analytes. There is much improved signal to

noise ratio for Cu, Mn, Co and Fe peaks. This can be attributed to the fact that these are heavy elements and therefore the intensities of X-rays fluorescence leaving them is high.

Table 5.2.2: SNR of classical wavelet analyzed spectra of 5 µm thick simulate tissue irradiated at a live time of 50 and 100 seconds respectively

Peak	SNR of classical wavelet analysis of spectra	
	5 µm 50 s	5 µm 100 s
Zn K_α	0.83	0.97
Cu K_α	12.12	13.86
Co K_α	9.30	10.50
Fe K_α	20.32	21.95
Mn K_α	6.41	7.49
Mg K_α	2.30	2.85
Na K_α	2.07	2.39

Table 5.2.2 shows that SNR results for 100 s are slightly better than for 50 s and therefore development of rapid robust chemometrics enabled EDXRFS analytical procedure was based on the spectra acquired at 50 s irradiation time since longer times (100 s) do not change the performance of the method significantly.

5.2.2 EDXRFS spectra analysis combining wavelets and principal component analysis

Since background is still pronounced in the CW analyzed sample spectrum with low SNR, EDXRFS spectrum of a simulate tissue sample, 10 µm thick, obtained by irradiating the sample at a live time of 50 s was preprocessed by combined use of wavelet transform and principal component analysis and the resulting spectrum is shown in Figure 5.2.2. PCA was used to remove components without vital information of the signal via dimension reduction. This

allows the expression of the data set in new reduced space that maximizes its variance. In this work PCA was performed on approximation coefficients to keep only the most important features of the spectra. Wavelet denoising reduces the noise in the signal while preserving the signal characteristics irrespective of its frequency (Jeana *et al.*, 2013).

To compare the spectrum of classical wavelet analysis, Figure 5.2.1 (ii) with that of combined use of wavelet transform and principal component analysis (WT-PCA), Figure 5.2.2, SNR were used. Table 5.2.3 shows results of SNR values for WT-PCA preprocessed spectrum are higher when compared with that obtained for CW alone.

Table 5.2.3: SNR results of raw spectra and combined WT and PCA of simulate tissue samples at various thickness at a live time of 50 s

Peak	Signal to noise ratio (SNR)					
	5 µm		10 µm		20 µm	
	CW	PCA + WT	CW	PCA + WT	CW	PCA + WT
Zn K_α	0.83	1.11	0.98	1.12	2.00	3.32
Cu K_α	12.12	30.36	15.00	34.12	14.14	34.56
Co K_α	9.30	27.76	11.52	34.16	10.12	34.28
Fe K_α	20.32	28.24	23.49	33.99	22.01	37.04
Mn K_α	6.40	9.60	7.62	12.23	8.02	14.00
Mg K_α	2.30	4.64	2.89	6.11	3.03	7.46
Na K_α	2.07	4.69	2.41	5.98	3.12	6.33

Table 5.2.3 shows improved SNR for all the three sample thickness when combined use of wavelets and principal component analysis is utilized as compared to wavelet transform alone

for all elements. Reducing noise from a signal aims at improving signal to noise ratio (SNR) and also better peak detection (Elise *et al.*, 2010).

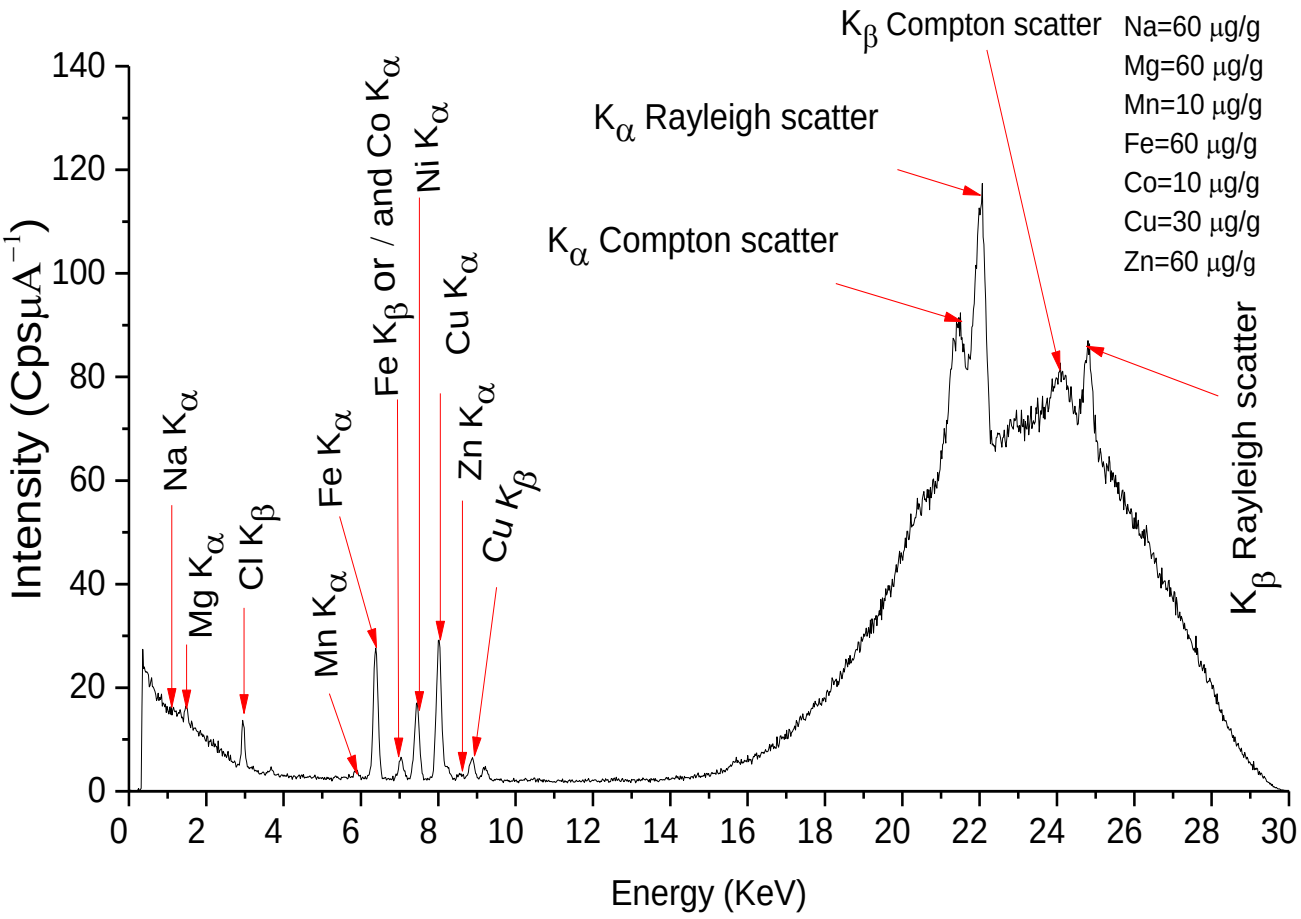


Figure 5.2.2: EDXRFS full spectrum of de-noised simulate tissue at live time of 50 s via wavelets combined with PCA.

PCA was executed on both the coarser approximation coefficients matrix in the wavelet domain and the final reconstructed matrix. By careful selection of the numbers of retained principal components, resulting simple signals are reconstructed (Hyvarinen and Oja, 2000; Aminghafari *et al.*, 2006).

In summary therefore, it can be deduced that WT assisted PCA (Table 5.2.3) produces better results as compares to classical wavelet (CW) analysis (Table 5.2.1) in spectral preprocessing towards total elemental determination since elevated background which masks subtle peaks making quantification in EDXRFS difficult is eliminated. In general, for unpreprocessed spectral data, sample thickness has a profound influence on the SNR which when after preprocessed there is no significant difference for the sample thickness variation between 5-20 μm. Therefore, 10 μm thickness which gave on average better SNR of K_α peaks of the elements analyzed was utilized and thus was adopted for analysis. Biological specimens such as soft body tissue obtained from human body are recommended to be extracted in small quantity to avoid body damage (Markowicz *et al.,* 1993).

5.2.3 ICA of simulate tissue samples

The use of ICA analysis was motivated by the fact that the EDXRFS spectral data consists of a set of independent signals additively combined $K_{\alpha 1}$ and $K_{\alpha 2}$ to form K_α line. ICA identifies the components of a mixture and so demixing is effectively a de facto denoising. The spectral data used in this study meets the conditions of ICA in separation of source namely: (i) the components should be statistically independent (ii) mixing at the detectors is linear. The ICA algorithm does not differentiate between noise and signal but simply separates components as compared to the wavelet denoising technique (Uzay *et al.,* 2003).

Therefore ICA was used to extract spectral line components contribution to $K_{\alpha 1}$ and $K_{\alpha 2}$ from atomic element transitions which consists of electronic random noise. To quantify the ability of ICA in spectral preprocessing (i.e. noise reduction), the performance measure i.e. signal to noise ratio (SNR) of K_α peaks of the elements of interest were computed and compared for unpreprocessed spectra and ICA analyzed spectra. Results of SNR before and after application of ICA are shown in Table 5.2.4.

Table 5.2.4: SNR for various peaks for 10 µm thick simulate tissue, 50 s live time before and after ICA

Peak	SNR unpreproceesed spectra	SNR after (ICA)
Zn K_α	0.34	1.27
Cu K_α	4.74	14.61
Co K_α	0.73	1.36
Fe K_α	6.40	16.35
Mn K_α	0.52	1.26
Mg K_α	0.81	1.07
Na K_α	0.86	1.03

The extracted signal obtained when ICA seperates all the sources that make up an observed mixture becomes the denoised signal. Results of Table 5.2.4 show that SNR values after ICA are more than before ICA analysis although less than combined use WT and PCA (Table 5.2.3). ICA simply seperates the components (resolving K_α and K_β energy peaks) which is not sufficient to denoise the data and this explains the low values of SNR (Uzay *et al.*, 2003). However, this preprocessing was aimed at resolving the K_α and K_β energy peaks for speciation purposes.

This sub-section presented a robust chemometric approach in EDXRFS spectra preprocessing towards quantification and speciation analysis of trace and major biometals. Combined use of wavelet analysis and principal component analysis of EDXRFS spectra was adopted as preprocessing technique towards determination of trace biometals since the technique reduces backgrounds which masks subtle peaks and also corrects matrix effects which makes quantification analysis difficult. Combined use of wavelet analysis and independent component analysis was utilized as a preprocessing technique toward determination of speciation of trace

biometals. This is because resolution of K_α peak as well noise reduction was enhnaced. For rapidness of analytical procedure towards trace biometal and speciation analysis, 50 s was adopted since results of 100 s for preprocessing did not change significantly. Furthermore 10 µm sample thickness was also utilized since there was little change in the SNR results for all the thickness after preprocessing.

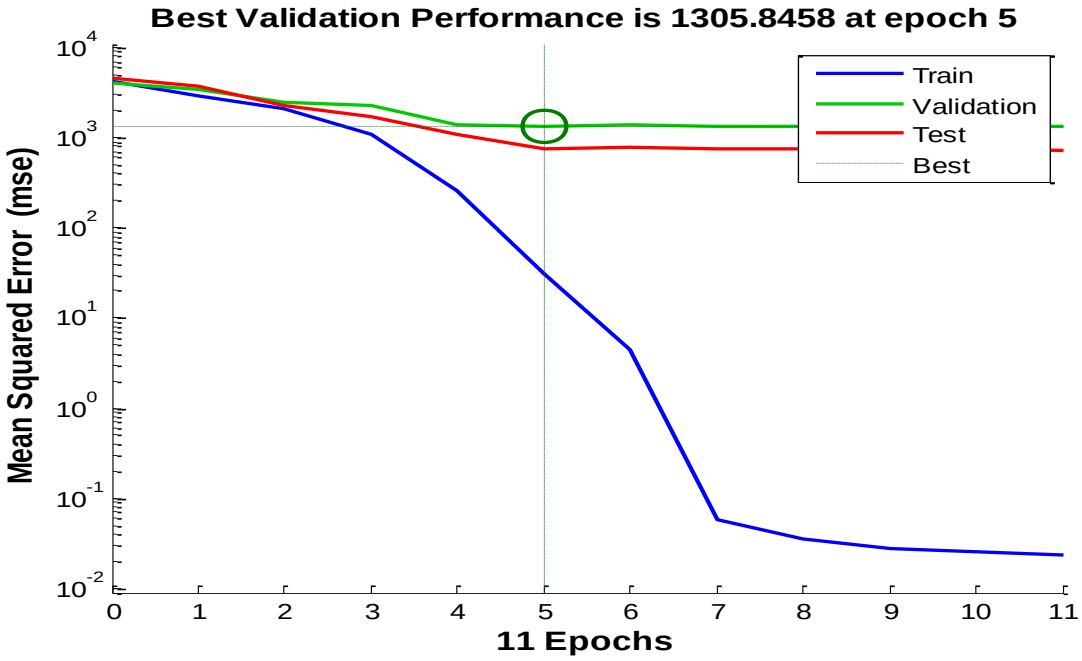
5.3 Multivariate calibration for quantitative analysis

5.3.1 ANNs for quantitative analysis of trace biometals in simulate tissues

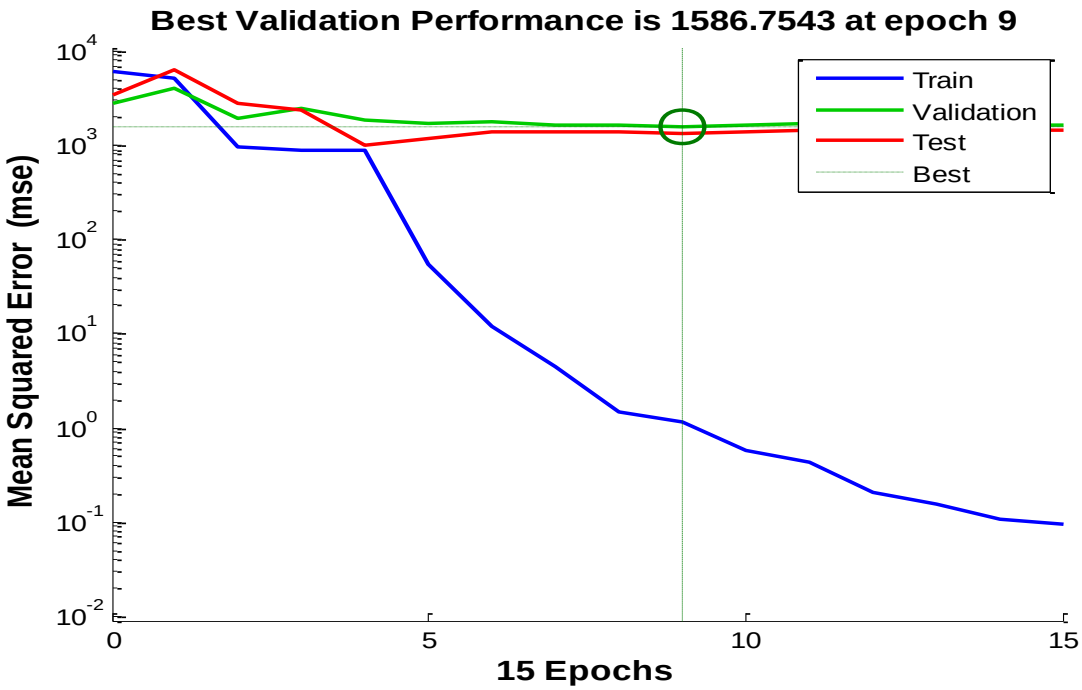
Owing to the fact that EDXRF spectra is prone to background noise and poor resolution, preprocessing was performed prior to ANN analysis. This was done via combined use of wavelet transform and principal component analysis which resulted in high SNR when compared to classical wavelets.

In principle ANNs technique needs a computation capacity that is high to enable handling of large data set. In this study smaller data set was used instead referred to as regions of interest (ROI) of fluorescence region corresponding to preselected peak regions of (Na, Mg, Fe, Cu, Zn, Mn and Co) as shown in Table 4.3.5 and Compton scatter regions to determine the least mean square error (MSE). Further, the ROI was utilized to determine whether it is useful for quantitative determination of trace biometal or not.

A sample training performance plots are shown in Figure 5.3.1.1 (i) and Figure 5.3.1.1(ii). The plots show the MSE of the network starting at a large value and decreasing to a smaller value, an indication that the network is learning. The plots have three lines which represents the randomly divided inputs and targets into three sets, i.e. 70 % being used to train the network, 15 % used to validate how well the network is generalized and 15 % to test how the network will perform on new set of data.



(i)



(ii)

Figure 5.3.1.1: Performance plot for ANNs training errors exploiting (i) selected fluorescence regions of Na, Mg, Fe, Cu, Zn, Mn and Co plus Compton scatter peaks and (ii) selected fluorescence regions of Na, Mg, Fe, Cu, Zn, Mn and Co.

Table 5.3.1 shows the results obtained following training the neural network at various neurons in the hidden layer. The efficiency in predicting is based on the R value which a measure of how variation in the output is well explained by the targets. A numerical value approximately equal to 1 shows a good correlation between targets and the outputs. The best model corresponded to low mean-squared error, high R value and fewer neurons in the hidden layer. Hidden layer with high number of neurons is likely to cause over-fitting whereas too few would not be capable in achieving an accurate classification. A model with 5 neurons was therefore selected.

Table 5.3.1: Analytical performance indices for different ANN models

Model	Number of neurons in hidden layer	Best validation performance mean squared error	Number of epochs/ iterations	Correlations coefficient			
				Training R-value	Validation R-value	Test R-value	Overall R-value
(i)	1	0.0101	11	0.9979	0.8881	0.9303	0.9503
(ii)	2	0.608	8	0.9752	0.9681	0.9353	0.9658
(iii)	3	0.0258	11	0.9989	0.9254	0.8821	0.9270
(iv)	4	0.485	10	0.9803	0.7582	0.8715	0.9320
(v)	**5**	**0.0245**	**12**	**0.9919**	**0.9515**	**0.9814**	**0.9815**
(vi)	6	0.323	16	0.9917	0.9786	0.9405	0.9489
(vii)	7	0.032	15	0.9983	0.9528	0.9206	0.9741

Figure 5.3.1.2 shows the linear regressions between the network outputs and corresponding trace element targets for four datasets (training, validation, test and total response) utilizing selected fluorescence region alone.

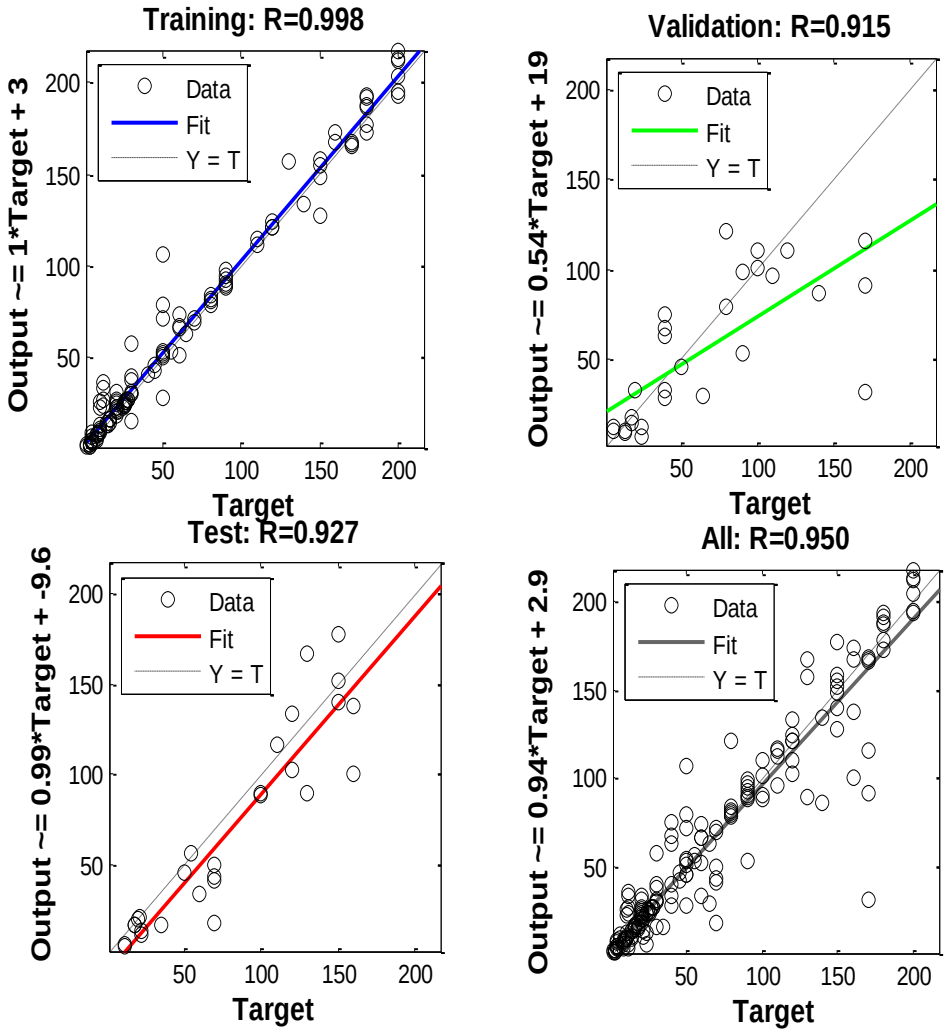


Figure 5.3.1.2: Linear regression of ANN outputs and corresponding trace elements targets for simulate tissues utilizing the selected fluorescence regions.

ANN was able to simultaneously train the model for the trace and major biometals namely; Na, Mg, Co, Fe, Mn, Cu and Zn with R - Value of 0.998. This was followed by a good validation and prediction with R - value of 0.915 and 0.927 respectively which resulted to an overall R

value of 0.95. Thus the output tracked the targets satisfactory for training and validation and the R – value was over 95 % for the total response.

It can be deduced from Figure 5.3.1.3 that selected fluorescence regions plus Compton gives a better model with R values so close to 1; 0.996, 0.944 and 0.963 for training, validation and prediction respectively. This resulted to a good overall R value of 0.984 which is over 98 % for the total response. Compton scatter has utility in quantitative analysis of especially low- Z elements.

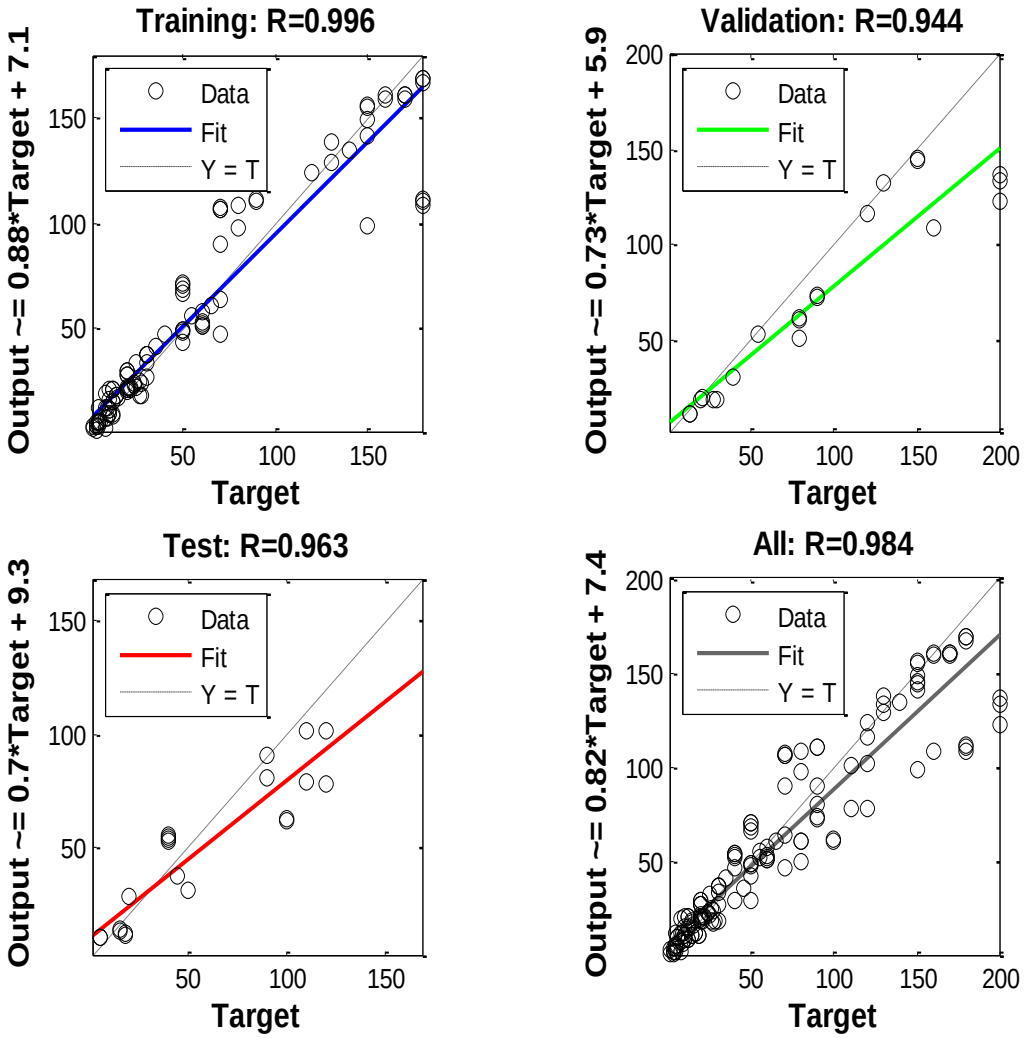


Figure 5.3.1.3: Linear regression of ANN outputs and corresponding trace elements targets for simulate tissues utilizing the selected fluorescence regions plus Compton scatter.

In general, the utility of Compton scatter peak in addition to the selected fluorescence region (K_α and K_β) for each analyzed trace and major biometal has a better utility in development of quantitative model for elemental determination especially for low –Z elements which is not possible with traditional XRF. This can be viewed from the fact that the output satisfactorily tracked the targets (Figure. 5.3.1.3) for an overall R value of over 98 % for the total response. This is an improvement to overall R value obtained by Kaniu (2011) and Okonda (2015) when both featured fluorescence and Compton scatter were utilized conjointly. The improved R value is as a result of utilization hybridized chemometrics spectral preprocessing approach. The preprocessing approach improves peak statistics by reducing elevated background which masks the normally subtle fluorescence peaks making quantitative analysis in EDXRFS difficult.

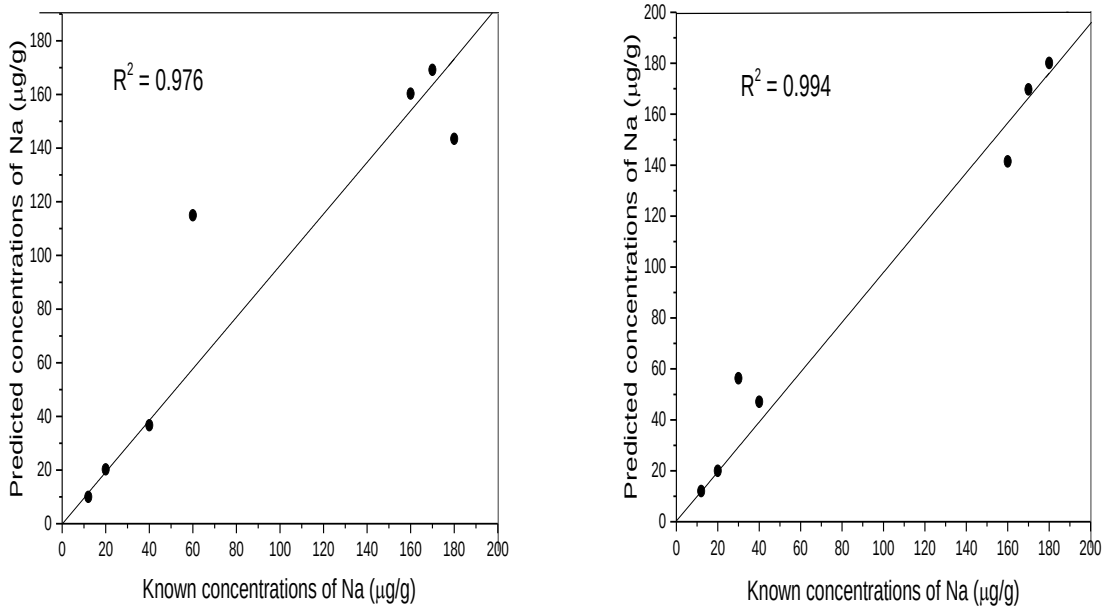
ANN quantitative analysis of EDXRFS spectra before spectral preprocessing was done and the results were compared with those obtained after WT-PCA analysis of EDXRFS spectra. The results clearly show that the preprocessed spectra data gives improved results. The values of R^2 were used as a measure of the performance (Table 5.3.2).

Table 5.3.2: Regression coefficients of ANN for feature selected fluorescence plus Compton scatter before and after WT-PCA of spectra for 10 μm thick simulate tissue at a live time of 50 s

Element / and flurescence signature	Regression coefficient (R^2)	
	Before spectral analysis	After spectral analysis
Na (K_α and K_β)	0.889	0.994
Mg (K_α and K_β)	0.890	0.995
Mn (K_α and K_β)	0.921	0.994
Fe (K_α and K_β)	0.916	0.989
Co (K_α and K_β)	0.951	0.993
Cu (K_α and K_β)	0.923	0.997
Zn (K_α and K_β)	0.903	0.996

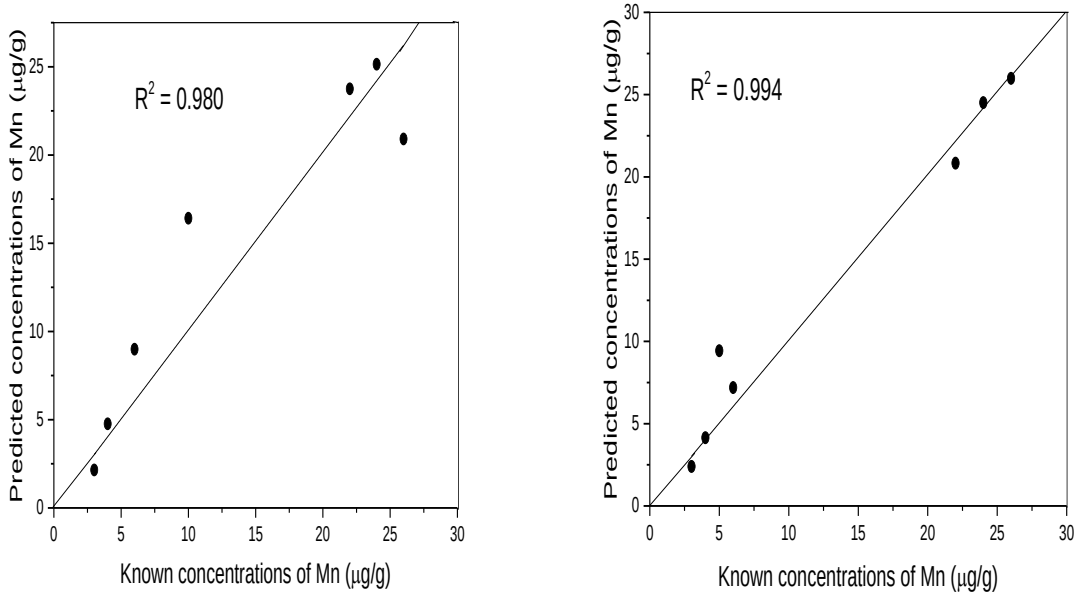
Table 5.3.2 shows improved regression coefficients for all the analyzed elements (Na, Mg, Mn, Fe, Co, Cu and Zn) after spectral preprocessing as compared with before spectral preprocessing utilizing selected fluorescence and scatter region conjointly. Thus the use of WT-PCA in preprocessing of EDXRFS spectral data in conjunction with ANN affords direct and rapid quantification of the trace biometals including the low – Z elements which is a difficult task for classical XRF.

The predicted verses known concentrations regression plots (Figure 5.3.1.4-10) show the ability of ANN model in predicting Na, Mg, Mn, Fe, Co, Cu and Zn in simulate tissue samples 10 μm thick and how well the model is likely to perform in analysis of tissue sample of similar matrix composition.



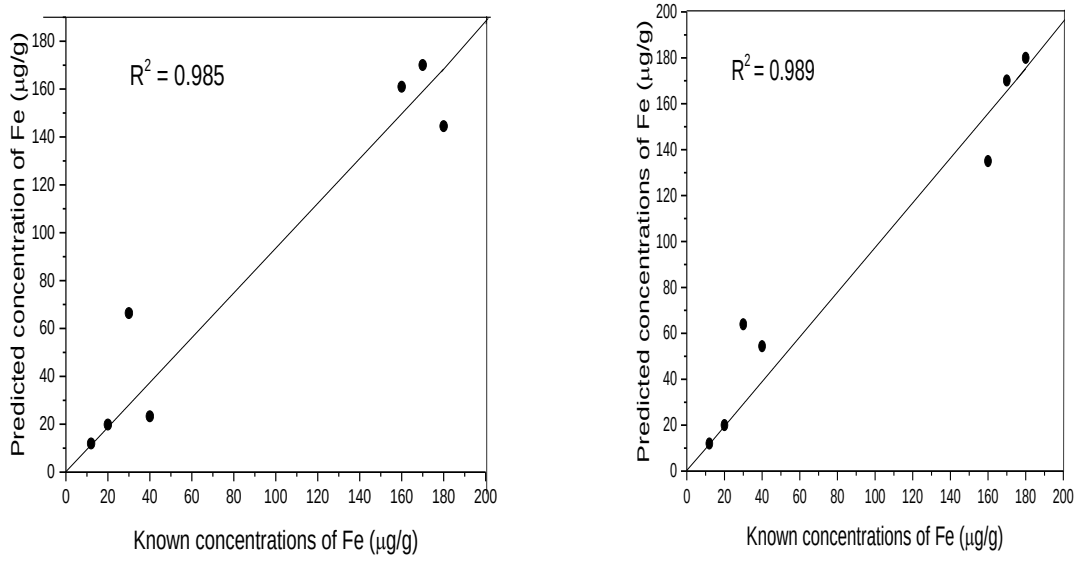
(i) (ii)

Figure 5.3.1.4: ANN regression plots for predicted verses known concentration of Na utilizing (i) selected fluorescence (ii) selected fluorescence plus Compton scatter at a live time of 50 s.



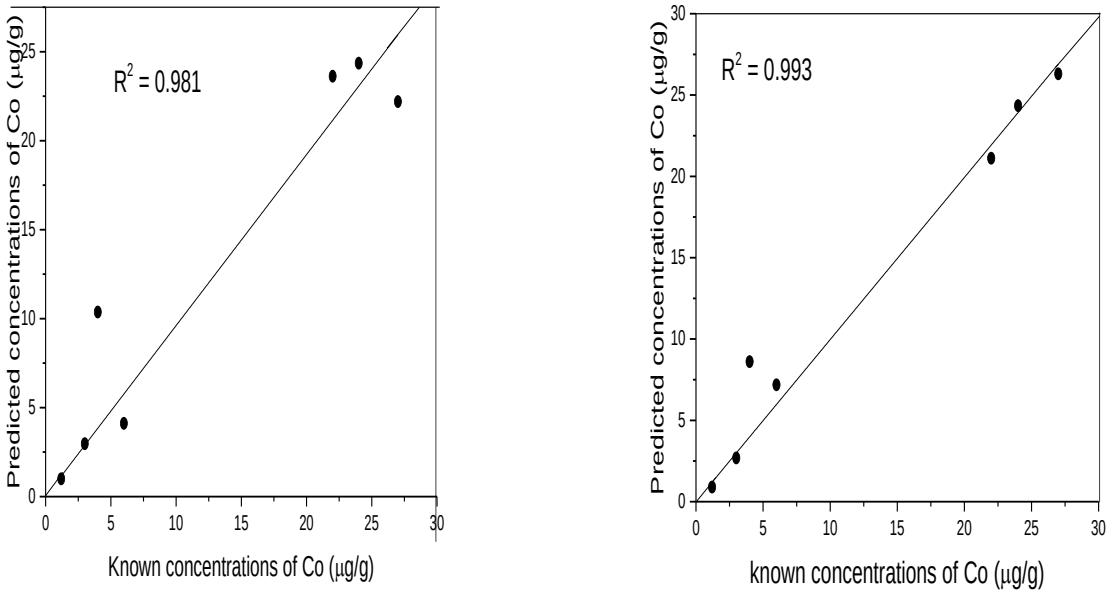
(i) (ii)

Figure 5.3.1.5: ANN regression plots for predicted verses known concentration of Mn utilizing (i) selected fluorescence (ii) selected fluorescence plus Compton scatter at a live time of 50 s.



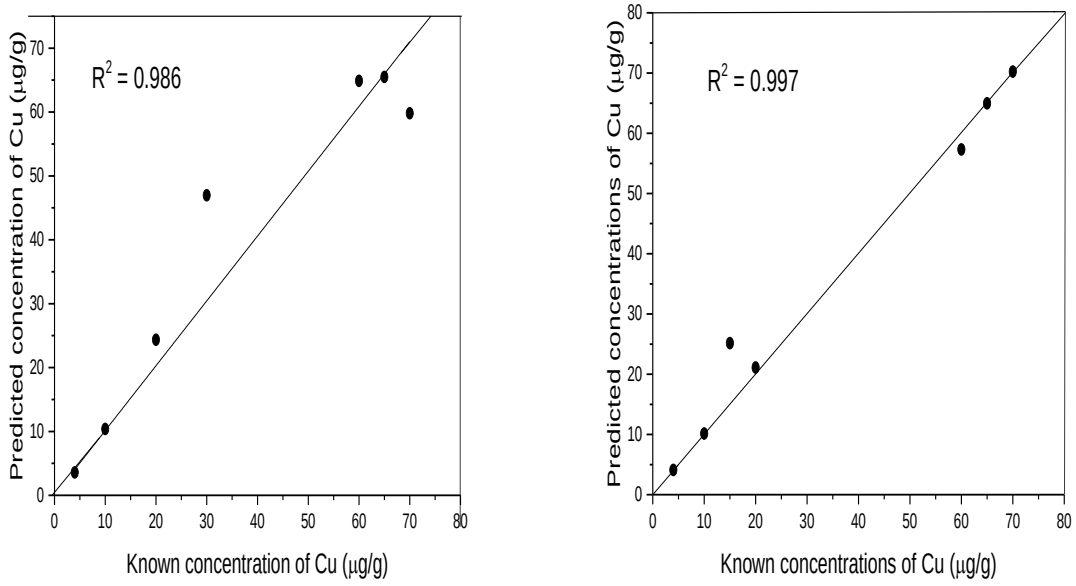
(i) (ii)

Figure 5.3.1.6: ANN regression plots for predicted verses known concentration of Fe utilizing (i) selected fluorescence (ii) selected fluorescence plus Compton scatter at a live time of 50 s.



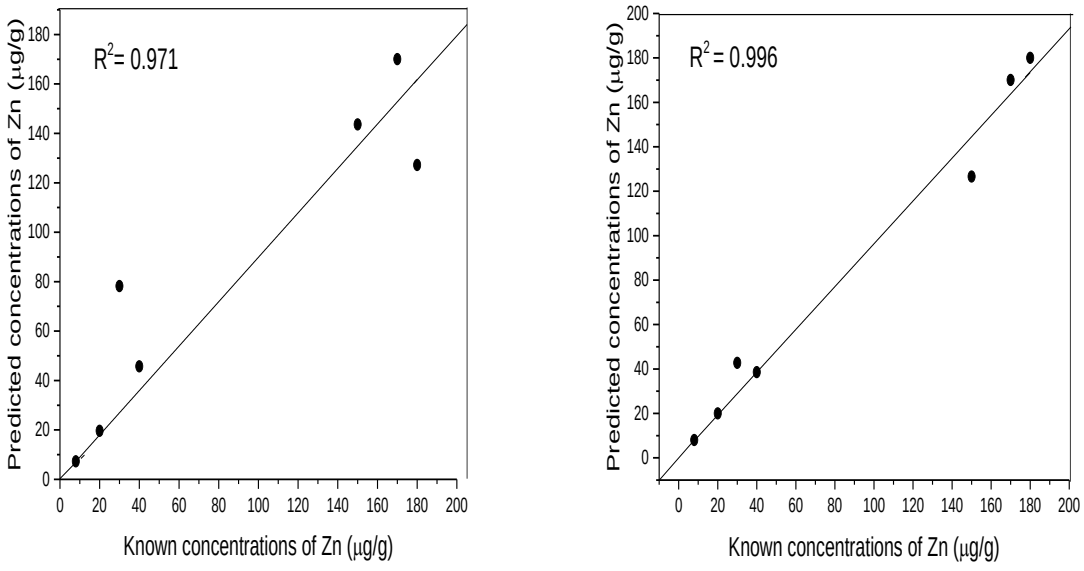
(i) (ii)

Figure 5.3.1.7: ANN regression plots for predicted verses known concentration of Co utilizing (i) selected fluorescence (ii) selected fluorescence plus Compton scatter at a live time of 50 s**.**



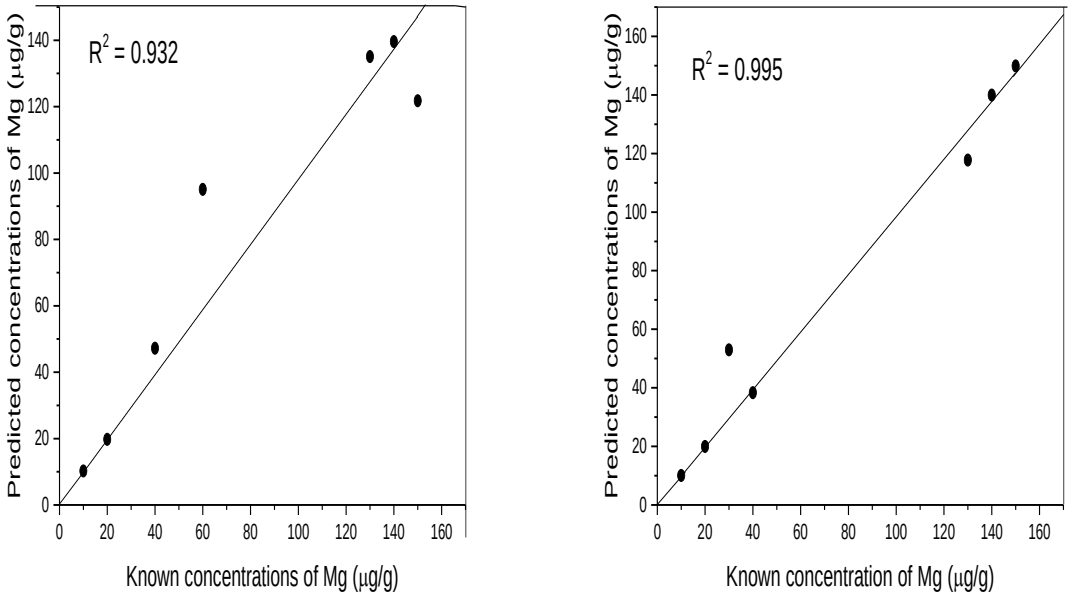
(i) (ii)

Figure 5.3.1.8: ANN regression plots for predicted verses known concentration of Cu utilizing (i) selected fluorescence (ii) selected fluorescence plus Compton scatter at a live time of 50 s.



(i) (ii)

Figure 5.3.1.9: ANN regression plots for predicted verses known concentration of Zn utilizing (i) selected fluorescence (ii) selected fluorescence plus Compton scatter at a live time of 50 s.



(i) (ii)

Figure 5.3.1.10: ANN regression plots for predicted verses known concentration of Zn utilizing (i) selected fluorescence (ii) selected fluorescence plus Compton scatter at a live time of 50 s.

Results from Figure 5.3.1.4 and Figure 5.3.1.10 show that utility of Compton scatter peak in combination with selected fluorescence region in the context of EDXRFS spectrometry enabled by robust chemometrics has potential in developing a quantitative model for trace biometal determination for low - Z elements (Na and Mg). The low values of R^2 for low – Z elements when selected fluorescence region alone is utilized is as a result of emission of low X-ray fluorescence as compared to heavy elements thus mostly giving rise to high Compton scatter (Kvetoslav, 1999). The R^2 values improve when selected fluorescence region plus Compton scatter peak are utilized.

Table 5.3.3 summarizes R^2 values for the biometals when selected fluorescence signature alone is utilized and when selected fluorescence and Compton scatter regions are utilized conjointly after spectral preprocessing. There is an improved R^2 value for Na and Mg from 0.976 and

0.932 respectively for selected fluorescence region alone to 0.994 and 0.995 respectively for selected fluorescence plus Compton scatter region. This shows the utility of scatter region in quantitative analysis of biometals especially the low –Z elements (Kaniu, 2011).

Table 5.3.3: Regression coefficients of ANN for processed EDXRFS spectra for 10 μm simulate tissue

Element	Regression coefficient (R^2)	
	Selected flourescence region	Selected flourescence plus Compton scatter peak
Na	0.976	0.994
Mg	0.932	0.995
Mn	0.980	0.994
Fe	0.985	0.989
Co	0.981	0.993
Cu	0.986	0.997
Zn	0.971	0.996

The accuracy and precision of the ANNs multivariate calibration technique depends on the quality of the input data, spectra and sample reproducibility and the ability to fit the known data. Errors originating from preparation of simulate tissue models were largely due to manual spiking during sample preparation procedure which was subject to error and thus inaccurate. Errors due to spectral noise were reduced by combined use wavelet analysis and PCA and the non-linear (training functions) handling capabilities of ANNs.

5.3.2 PLS for quantitative analysis of trace biometal in simulate tissues

Owing to the fact that spectral data is prone to elevated background, poor resolution of (K_α and K_β) energy peaks, a preprocessing step (section 4.6.1.2) was performed prior to PLS analysis to develop a calibration model. The processed EDXRFS spectra were used as an input to building a PLS-2 model for biometal analysis in soft body tissue.

PLS was performed using (i) selected fluorescence peaks (K_α and K_β) alone and (ii) selected fluorescence plus Compton scatter peak regions of biometals (Na, Mg, Mn, Fe, Co, Cu, Zn). The regressions of Y (concentrations) was based on selected PLS components of X (Figure 5.3.2.1 and 5.3.2.2) instead of original X-variables by cross-validation approach. Thus 8 PLS with the largest variances to explain as much of the total variation of the X-variables as possible for good correlation with the dependent Y-variables was utilized.

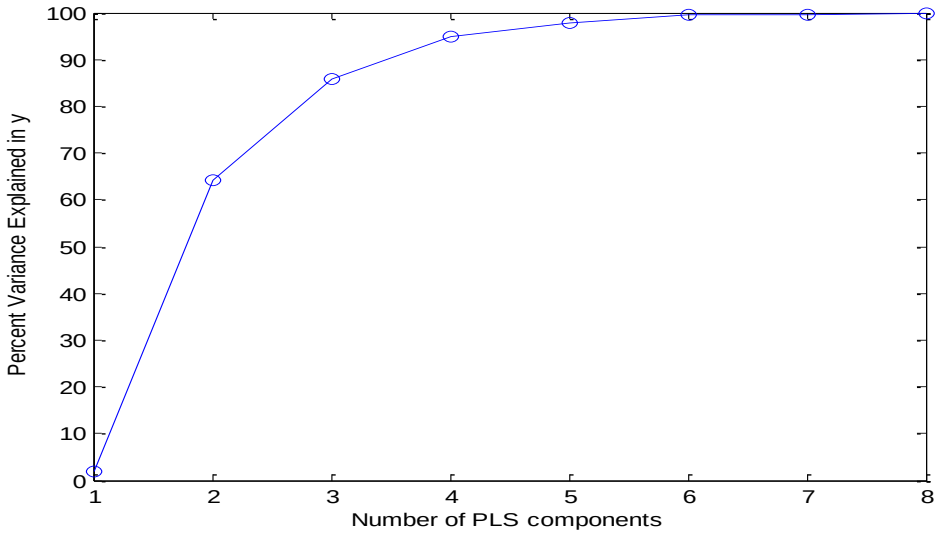
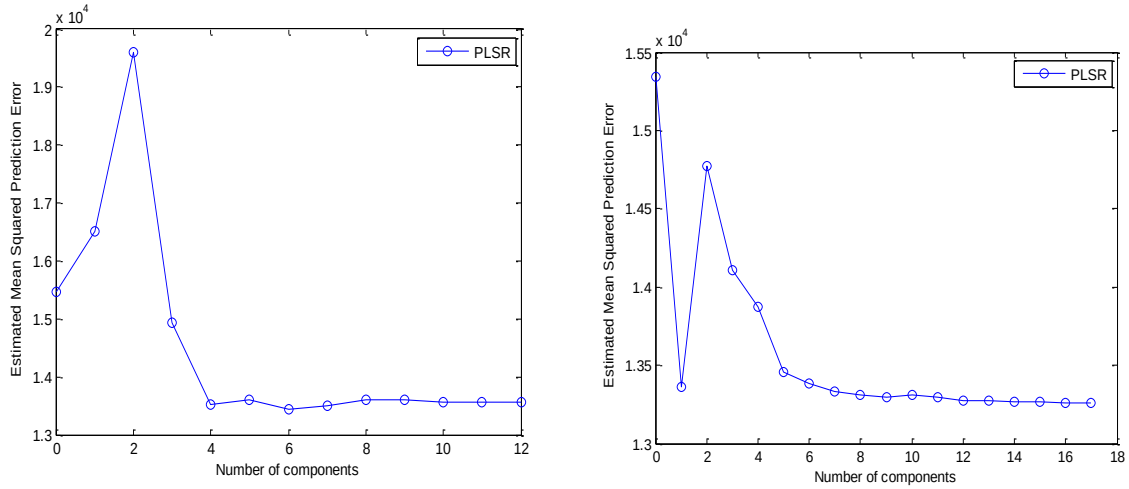


Figure 5.3.2.1: Percentage variance explained in y vs number of PLS components

Figure 5.3.2.1 suggests that PLS with four components explains most of the variance in the observed y. The mean squared prediction error (MSEP) curve (Figure 5.3.2.2) for PLS indicates that selected fluorescence region plus Compton scatter peak gives lower (MSEP) when

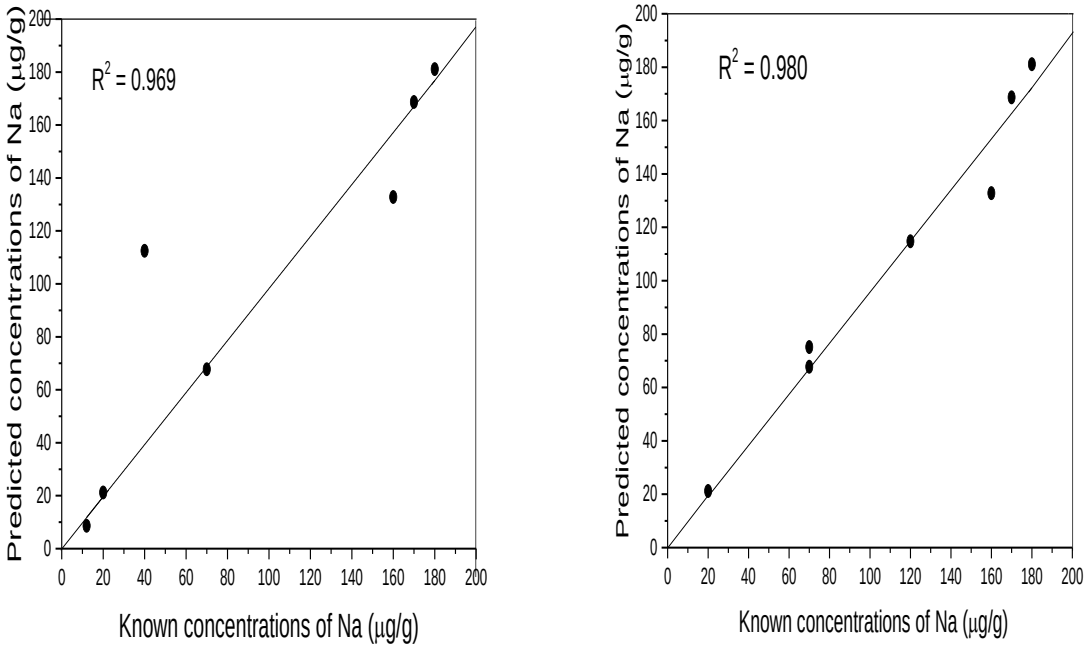
compared to selected florescence region alone. The regions were therefore utilized in developing PLS calibration. The 8 PLS components were then regressed against the given elemental concentrations to get the regression coefficients which were used on the test data to predict elemental concentrations.



(i) (ii)

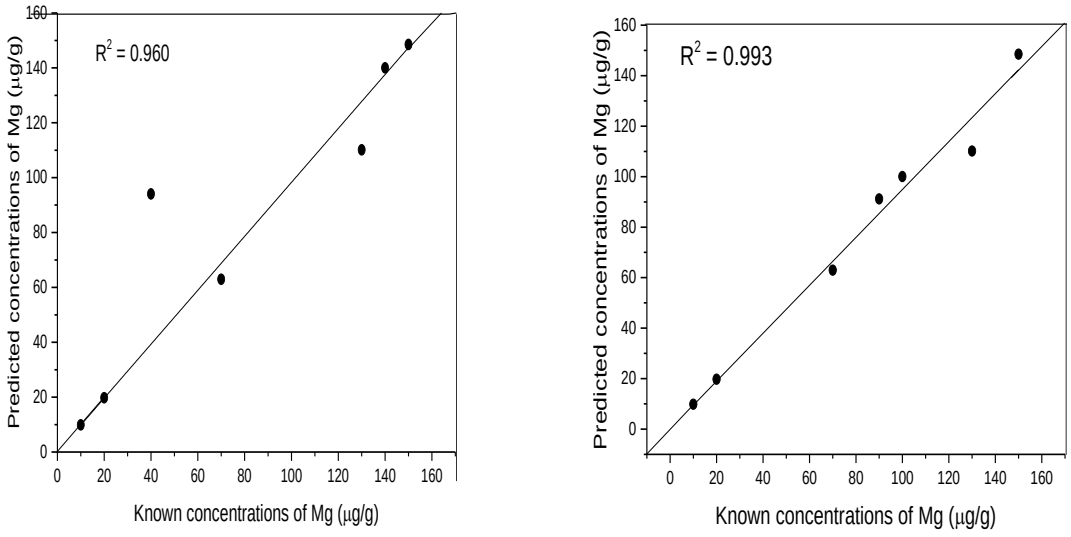
Figure 5.3.2.2: Mean squared prediction error (MSEP) curve showing number of PLS components in PLS analysis of simulate tissue samples utilizing (i) featured fluorescence and (ii) featured plus Compton scatter.

The predicted verses known concentrations regression plots (Figure 5.3.2.3-9) show the ability of PLS model in predicting Na, Mg, Mn, Fe, Co, Cu and Zn in simulate tissue samples 10 µm thick (i) utilizing preselected fluorescence region and (ii) preselected fluorescence region plus Compton scatter and how well the model is likely to perform in analysis of tissue sample of similar matrix composition.



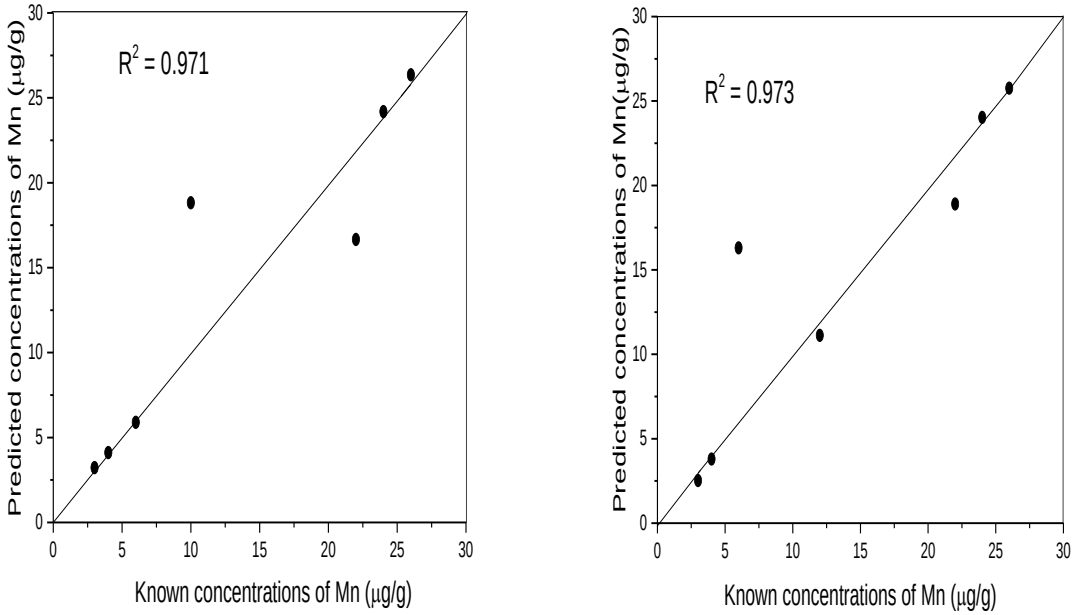
(i) (ii)

Figure 5.3.2.3: PLS regression plots for predicted verses known concentration of Na utilizing (i) selected fluorescence (ii) selected fluorescence plus Compton scatter at a live time of 50 s.



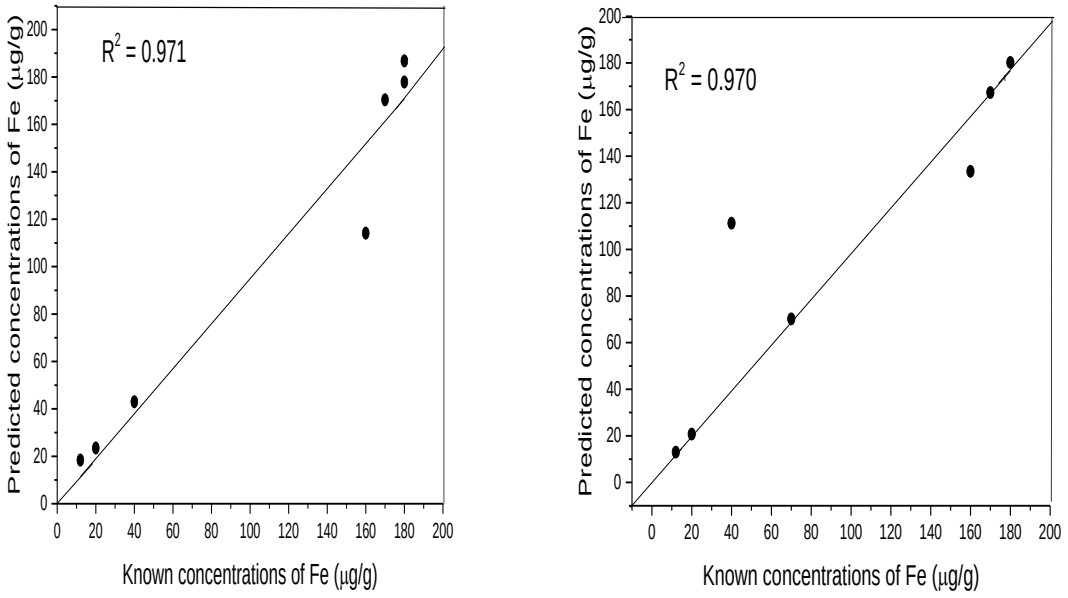
(i) (ii)

Figure 5.3.2.4: PLS regression plots for predicted verses known concentration of Mg utilizing (i) selected fluorescence (ii) selected fluorescence plus Compton scatter at a live time of 50 s.



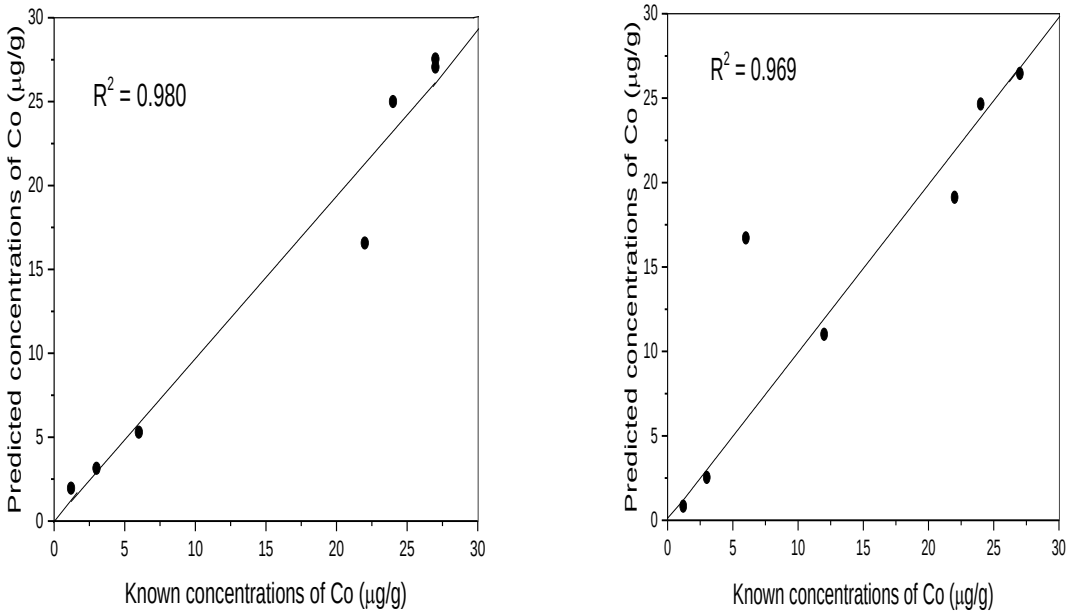
(i) (ii)

Figure 5.3.2.5: PLS regression plots for predicted verses known concentration of Mn utilizing (i) selected fluorescence (ii) selected fluorescence plus Compton scatter at a live time of 50 s.



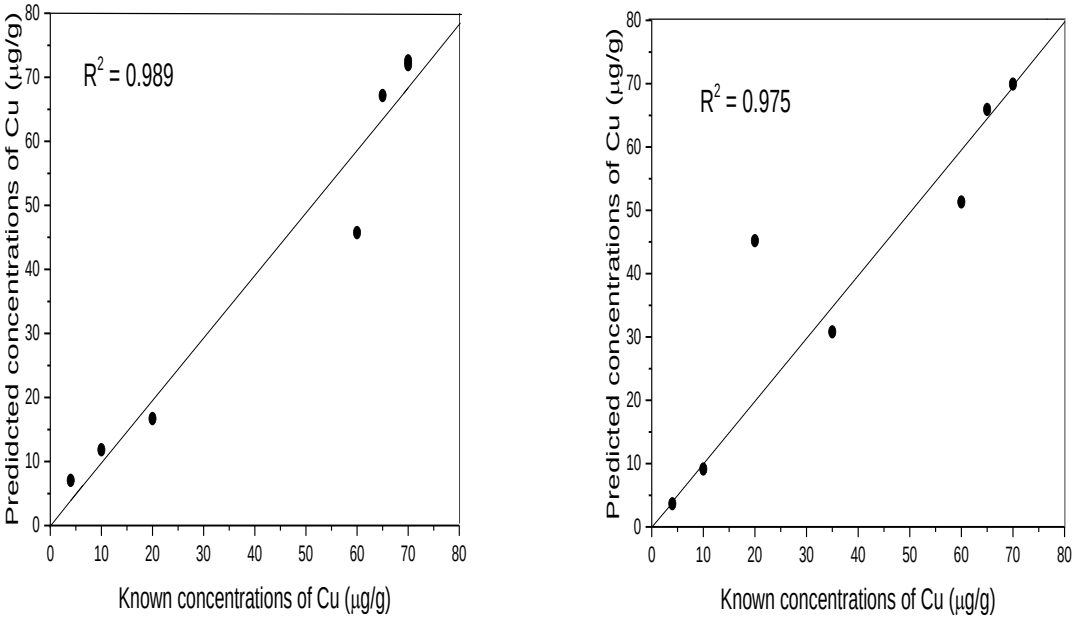
(i) (ii)

Figure 5.3.2.6: PLS regression plots for predicted verses known concentration of Fe utilizing (i) selected fluorescence (ii) selected fluorescence plus Compton scatter at a live time of 50 s.



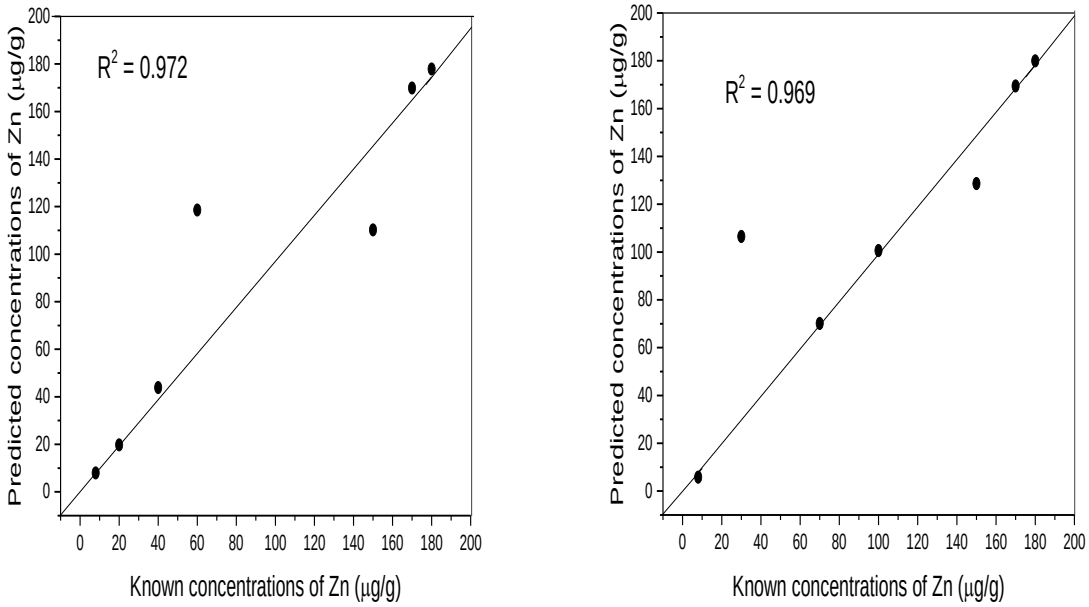
(i) (ii)

Figure 5.3.2.7: PLS regression plots for predicted verses known concentration of Co utilizing (i) selected fluorescence (ii) selected fluorescence plus Compton scatter at a live time of 50 s.



(i) (ii)

Figure 5.3.2.8: PLS regression plots for predicted verses known concentration of Cu utilizing (i) selected fluorescence (ii) selected fluorescence plus Compton scatter at a live time of 50 s.



(i) (ii)

Figure 5.3.2.9: PLS regression plots for predicted verses known concentration of Zn utilizing (i) selected fluorescence (ii) selected fluorescence plus Compton scatter at a live time of 50 s.

Results show that utility of Compton scatter peak in combination with selected fluorescence region in the context of EDXRFS spectrometry enabled by chemometrics has utility in developing a quantitative model for trace biometal determination for low - Z elements (Na and Mg), Figure 5.3.2.3 and 5.3.2.4. The R^2 values improve when selected fluorescence region plus Compton scatter peak are utilized.

Table 5.3.4: Regression coefficients of PLS calibration before and after WT-PCA preprocessing of selected fluorescence (K_α and K_β) and Compton scatter peak

Element	Regression coefficient (R^2)	
	Before WT-PCA preprocessing	After WT-PCA preprocessing
Na	0.876	0.980
Mg	0.890	0.993
Mn	0.931	0.973
Fe	0.884	0.970
Co	0.921	0.969
Cu	0.910	0.975
Zn	0.892	0.969

Table 5.3.4 shows better prediction of elemental concentrations when preprocessed spectral data is utilized as compared to before preprocessing. This is because the use of WT-PCA in spectral preprocessing optimizes the spectra towards quantification analysis. Background which masks subtle peaks making it difficult to perform quantitative analysis is reduced.

The results of ANN and PLS calibration models were compared to determine the most appropriate multivariate method for quantitative analysis of each of the elements of interest; Na, Mg, Fe, Cu, Mn, Zn and Co which would be applied in analysis of native soft body tissue samples. The results of comparison are shown in Table 5.3.5.

Table 5.3.5: Comparison of regression coefficients of ANN and PLS calibration for selected fluorescence plus Compton scatter peak

Element	Regression coefficient (R^2)	
	PLS	ANNs
Na	0.980	0.994
Mg	0.993	0.995
Mn	0.973	0.994
Fe	0.970	0.989
Co	0.969	0.993
Cu	0.975	0.997
Zn	0.969	0.996

The results show that ANN calibration model gives better prediction for elemental concentrations of Na, Mg, Mn, Fe, Cu, Zn, Co overall than PLS calibration model. The results are accurate for most elements of interest with ($R^2 > 0.99$) than the corresponding values obtained through PLS model ($R^2 < 0.98$) as shown in Table 5.3.5. This is because ANN model considers both linear and non-linear relations. The lower R^2 values of PLS model is due to non-linearity of the model as it could not deal with nonlinear spectral relations to analyze concentrations in the calibration samples. However, there is little difference in R^2 values of both models due to the utility of the combined use of WT and PCA in preprocessing of EDXRFS spectral data for quantification analysis.

5.4 Analysis of NIST Oyster tissue using WT-PCA and ANNs

The accuracy and reliability of the use of WT-PCA in spectral preprocessing and ANN calibration model in prediction of concentrations of the biometals namely Fe, Co, Cu, Mn, Zn, Na, Mg in soft tissue was evaluated by analysis of Oyster tissue standard reference material

(NIST 1566b) for comparison with fundamental parameter (FP) – EDXRF method. The concentrations for most biometals obtained using robust chemometrics approach are in good agreement (≤ 6% or less for most elements) with the certified values (Table 5.4.1). However, the concentrations of the biometals using classical FP-EDXRF method are not in agreement with the certified values in Oyster tissue (Table 5.4.1). Thus FP- EDXRF is inapplicable for direct rapid analysis of biometals especially the low-Z elements in soft body tissue. This is because the FP-EDXRF approach assumes X-ray fluorescence peak intensity as a linear function of concentrations (He and Van Espen, 1990) which is not easily achieved in direct biometal analysis of complex matrices (soft body tissues).

Table 5.4.1: Comparison of elemental concentrations in CRM Oyster tissue (NIST1566b) utilizing robust chemometrics enabled EDXRFS approach and classical EDXRF fundamental parameter method

Element	Certified value ± standard deviation (µg/g)	EDXRF-FP values ± standard deviation (µg/g)	EDXRFS measured value ± standard deviation (µg/g)	% deviation from certified values
Zn	1424± 46	1270 ± 16	1376 ± 25	3.3
Mg	1085 ± 23	-	1036 ± 21	4.6
Mn	18.5 ± 0.2	23.3 ± 6.5	18.2 ± 0.8	1.6
Fe	205.8 ± 6.8	231 ± 27	225 ± 25	3.5
Co	0.371 ± 0.009	-	0.354 ± 0.086	4.5
Cu	71.6 ± 1.6	77.9 ± 3.5	73 ± 3	2.0
Na	3292 ± 53	-	2950 ± 31	5.8

5.5 Multivariate calibration for speciation analysis

5.5.1 Principal component analysis (PCA) analysis of simulate samples

Principal component analysis (PCA) was used to reduce the spectral data dimensions for sample exploration in order to identify possible discriminating features between the 2 sets of simulate calibration standard samples (higher and lower speciation) at irradiation time of 50 s. The irradiation time of 50 s was chosen based on the earlier research which had shown that an increasing irradiation time doesn't have an effect on speciation analysis (Okonda, 2015). Before performing PCA, the spectra was preprocessed by combined use of WT and ICA for s resolution K_α and K_β peaks and noise reduction. The PCA classification results of simulate tissue obtained are based on the analyte preselected fluorescence peaks K_α and K_β for the elements (Mn, Fe, and Cu) as inputs to PCA.

For speciation analysis of Cu, the preselected fluorescence region of interest for Cu K_α and K_β peaks were utilized as input to PCA. As shown in Figure 5.5.1, two clusters are evident which are identified as Cu^+ and Cu^{2+} oxidation state respectively; explained by variance of 96 % (92 % and 4 % for PC1 and PC2) respectively. The outliers constituted to the remaining 4 % which can be attributed to noise. The two clusters are influenced by both K_α and K_β intensity signals of Cu as shown in on PCA loadings plot in Figure 5.5.2. Positive loadings of the Cu in PC1 (92 %), Figure 5.5.2, significantly contribute to the classification of Cu simulate samples into lower and higher oxidation states. Cu loading have negative influence on PC2 clustering.

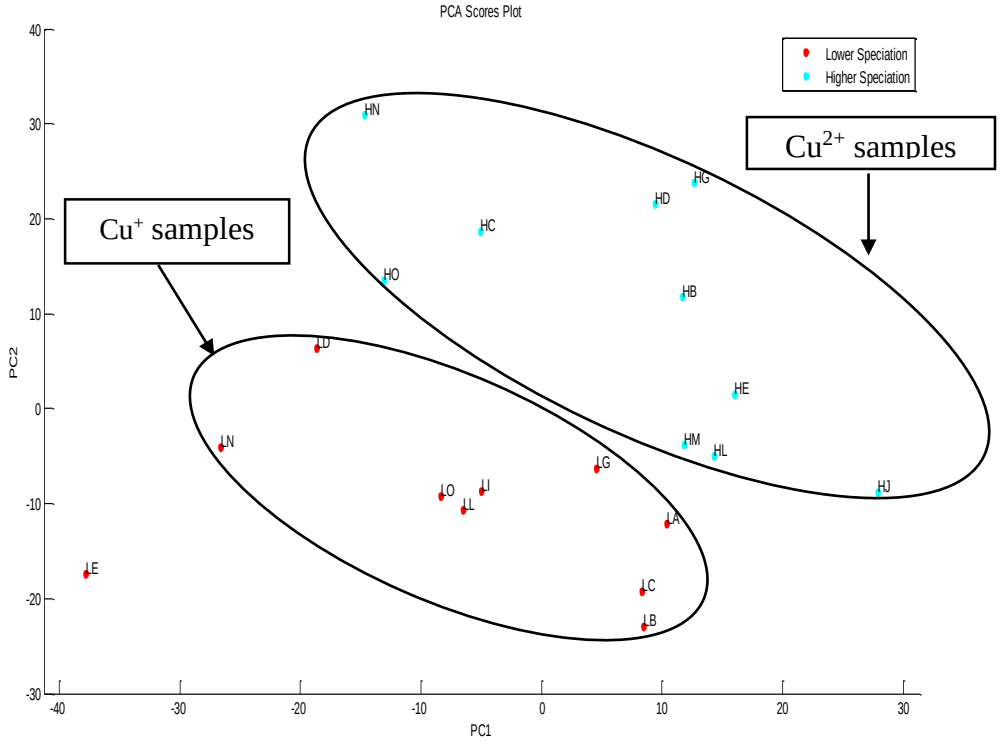


Figure 5.5.1: PC1 (92 %) × PC2 (4 %) score plot of speciation analysis of Cu simulate samples 10 µm thick at live time of 50 s.

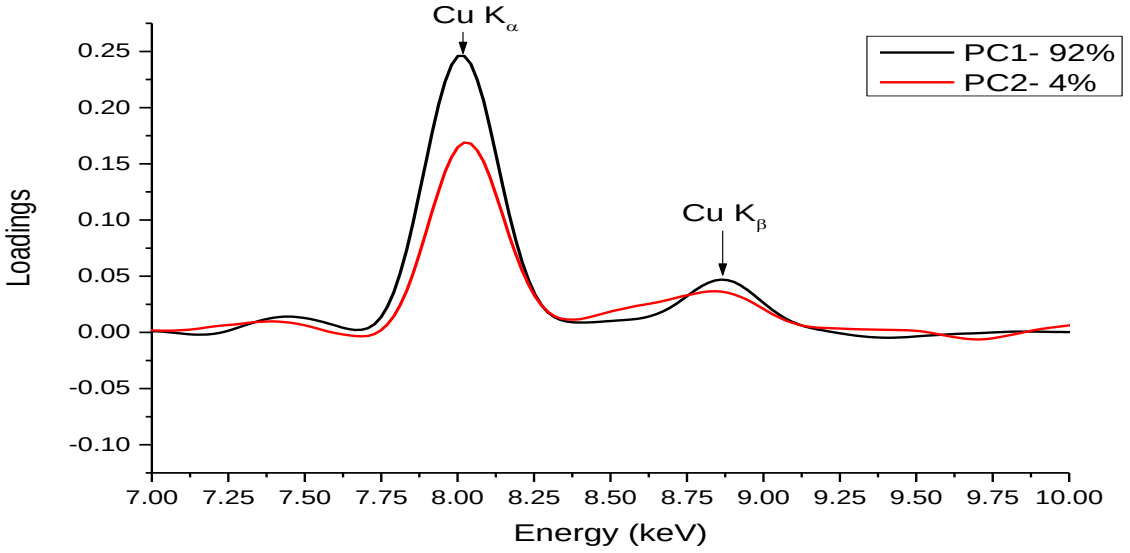


Figure 5.5.2: PCA loading plot showing the variables of simulate samples simulate samples 10 µm thick at live time of 50 s using preselected fluorescence signature of Cu.

To perform speciation analysis of Fe, similarly the preselected fluorescence region (K_α and K_β peaks) of Fe were utilized as input to PCA. Figure 5.5.3 shows two clusters of Fe^{2+} and Fe^{3+} respectively; with an explained variance of 97 % (93 % and 4 % for PC1 and PC2). The outliers constituted the remaining 3 % which can be attributed to noise. The two clusters in Figure 5.5.3 are influenced by both K_α and K_β intensity signals of Fe as shown in on PCA loadings plot in Figure 5.5.4. Positive loadings of the Fe in PC1 (93 %), Figure 5.5.4, significantly contribute to the clustering of Fe simulate samples into their respective oxidation states. Fe loading have positive influence on PC2 clustering.

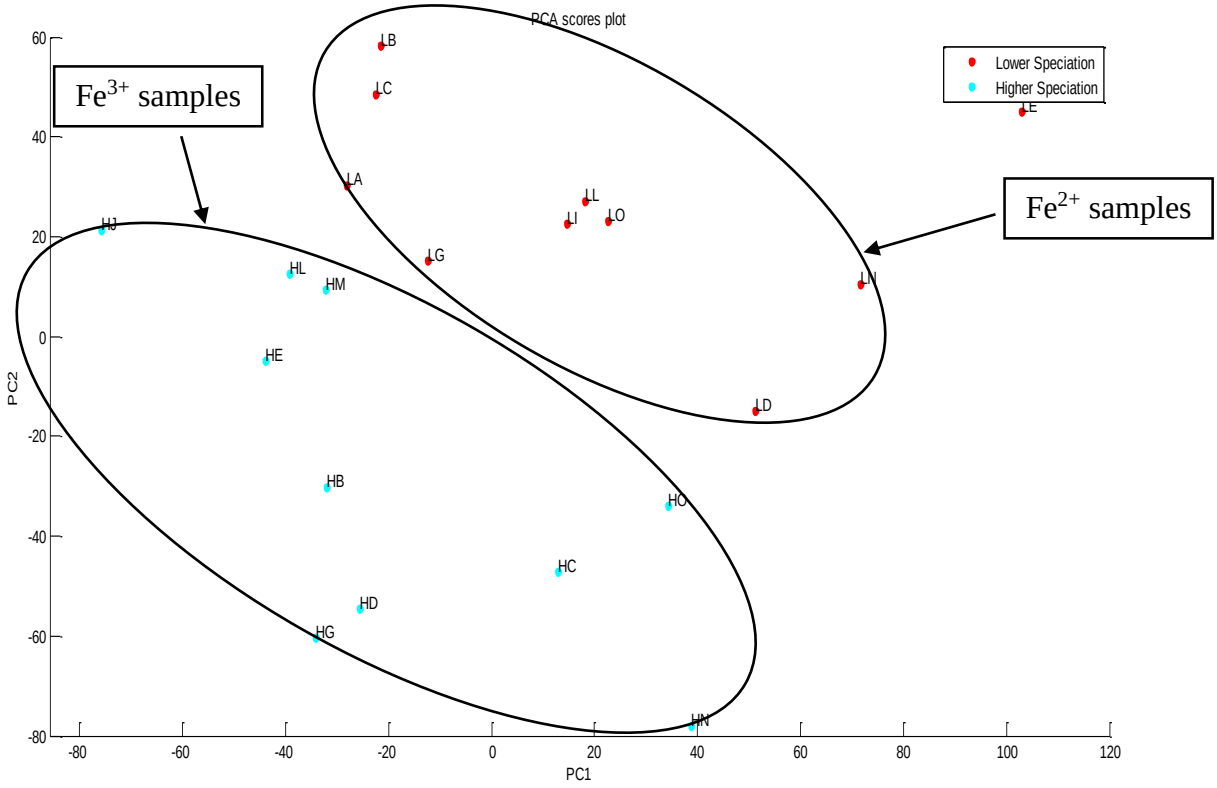


Figure 5.5.3: PC1 (93%) × PC2 (4%) score plot of speciation analysis of Fe simulate samples 10 µm thick at live time of 50 s.

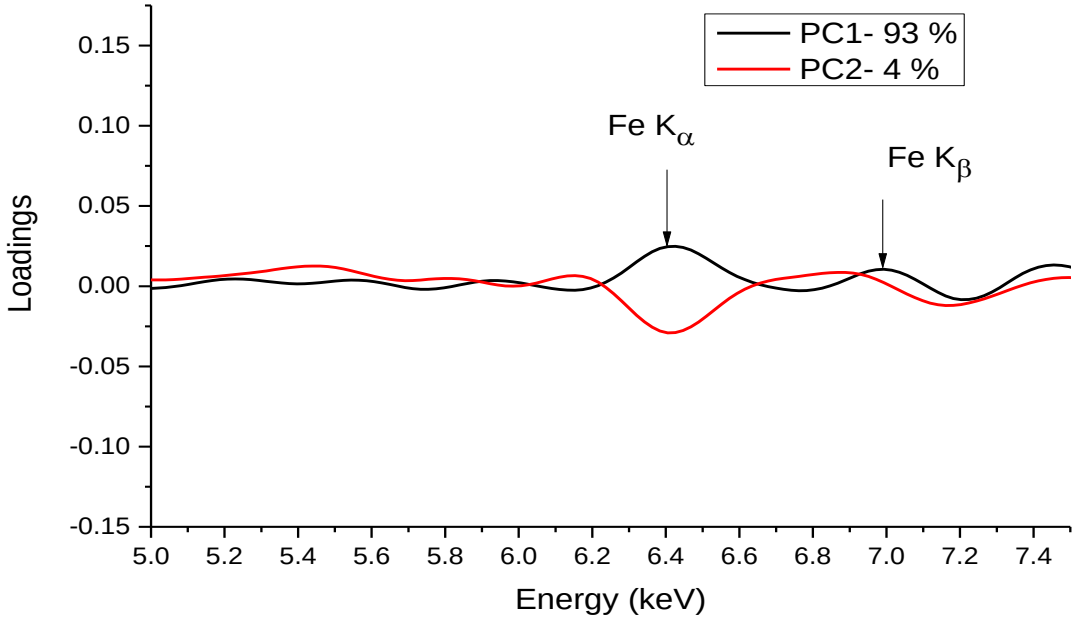


Figure 5.5.4: PCA loadings plot showing the variables of simulate samples 10 µm thick at a live time of 50 s using preselected fluorescence signature of Fe.

Likewise speciation analysis of Mn was carried out by preselecting fluorescence region of Mn K_α and K_β peaks as inputs to PCA. PCA score plot of simulate samples in Figure 5.5.5 shows classification of Mn into the two clusters of Mn^{2+} and Mn^{7+} respectively based on composition of Fe in the simulate samples. The explained variance is 95 % i.e. 87 % and 8 % for PC1 and PC2 respectively. The outliers constituted to the remaining 5 % which can be attributed to noise. The two clusters in Figure 5.5.5 are influenced by both $K_{\alpha 1}$ and $K_{\alpha 2}$ intensity signals of Fe as shown in on PCA loadings plot in Figure 5.5.6. Positive loadings of the Mn in PC1 (87 %), Figure 5.5.6, significantly contribute to the clustering of Mn simulate samples into their respective oxidation states.

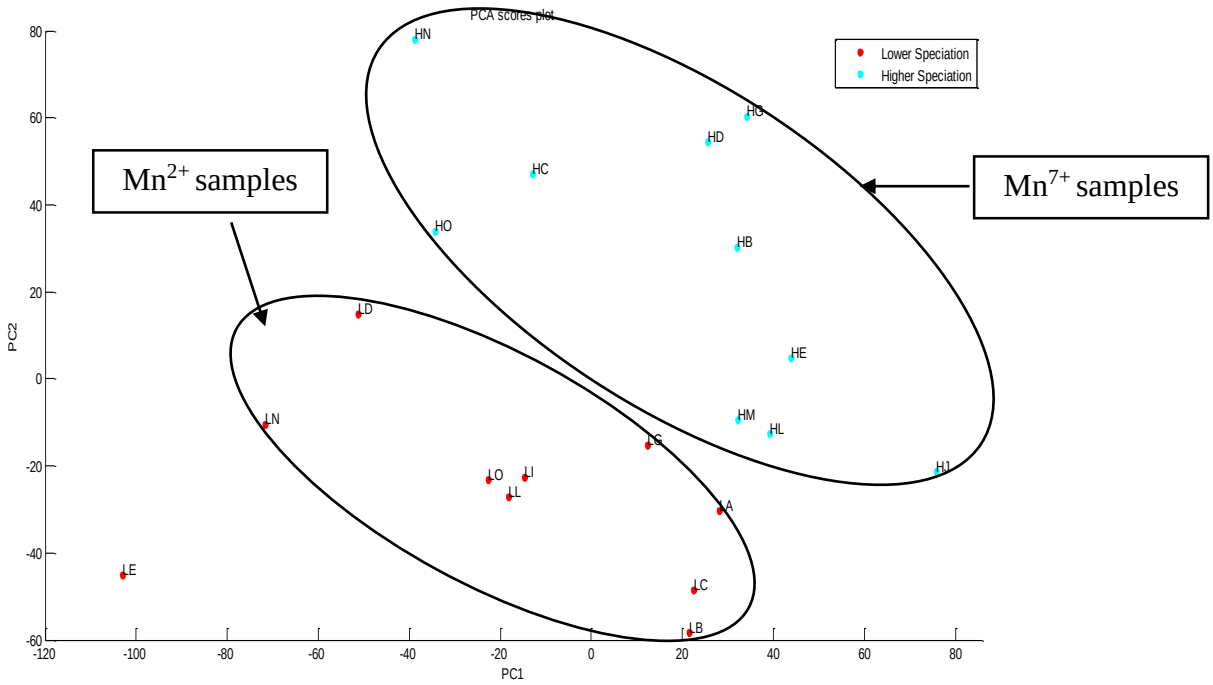


Figure 5.5.5: PC1 (87 %) × PC2 (8 %) score plot of speciation analysis of Mn simulate samples 10 µm thick at live time of 50 s.

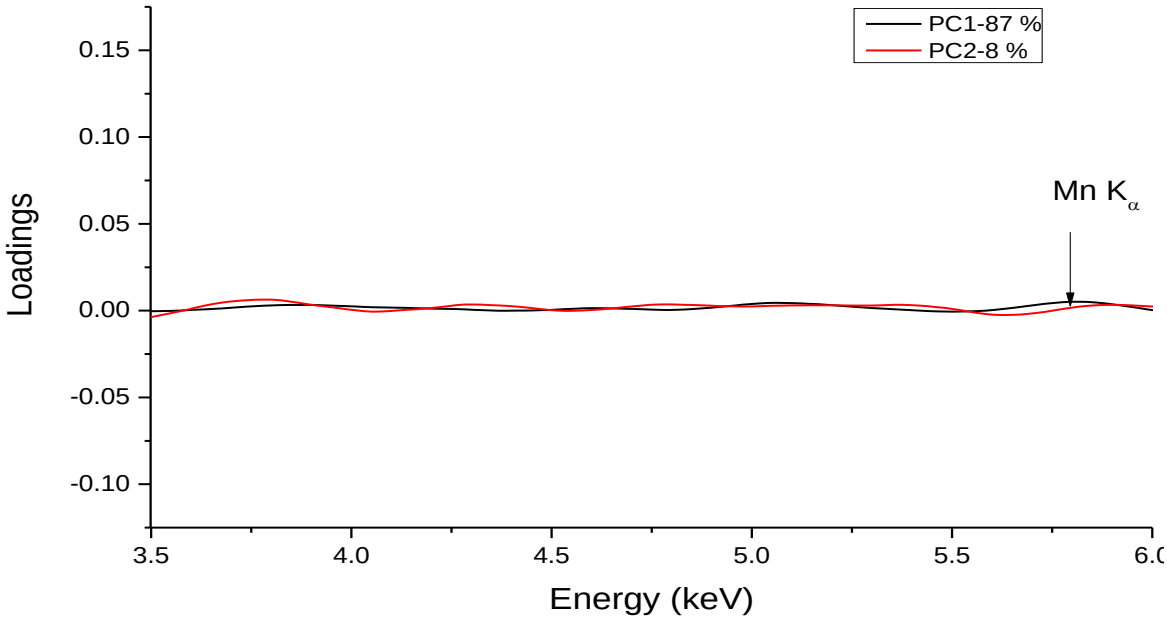


Figure 5.5.6: PCA loadings plot showing the variables of simulate samples simulate samples 10µm thick at live time of 50 s using preselected fluorescence signature of Mn.

In general, the loading plots in Figures 5.5.2, 5.5.4 and 5.5.6 indicate that PC1 and PC2, based on 2 major clusters were able to distinguish between simulate samples containing either the lower speciation (Fe^{2+}, Cu^{+} and Mn^{2+}) or higher speciation (Fe^{3+}, Cu^{2+} and Mn^{7+}). The 2 clusters were influenced by K_{α} intensities of Fe^{2+}, Cu^{+} and Mn^{2+} and Fe^{3+}, Cu^{2+} and Mn^{7+} together with total elemental content of spiked trace biometals based on the positive contribution of PC1 loadings. The classified oxidation states of the trace biometals in simulates were in agreement with the known spiked oxidation states thus this generally validates PCA as a tool for speciation analysis in soft tissue which has an application in disease prognostic and diagnosis such as cancer.

5.5.2 Speciation analysis of model soft tissue via Independent component analysis (ICA)

Before performing speciation analysis using ICA, the EDXRFS spectra were preprocessed via combined use of WT and ICA (WT-ICA) for noise reduction and peak resolution enhancement whereby noise in the signal was reduced while preserving signal characteristics. WT-ICA was preferred to wavelet transform assisted principle component analysis (WT-PCA) because the information useful for speciation analysis is retained unlike WT-PCA which performs well in denoising but removes information useful for speciation analysis i.e. the subtle peaks of $K_{\alpha 1}$ and $K_{\alpha 2}$.

Speciation analysis of Mn in model soft tissue samples was performed by preselecting the fluorescence region of K_{α} and K_{β} peaks of the EDXRFS spectra for Mn. The spectra was then preprocessed by the combined use of wavelet transform and ICA analysis for noise reduction and resolution of the peaks. Figure 5.5.7 shows the classification of Mn samples into low and high oxidation states via ICA. The samples were effectively discriminated into low speciation (containing Mn^{2+}) and high speciation (containing Mn^{7+}). The two clusters are influenced by K_{α} intensity of Mn.

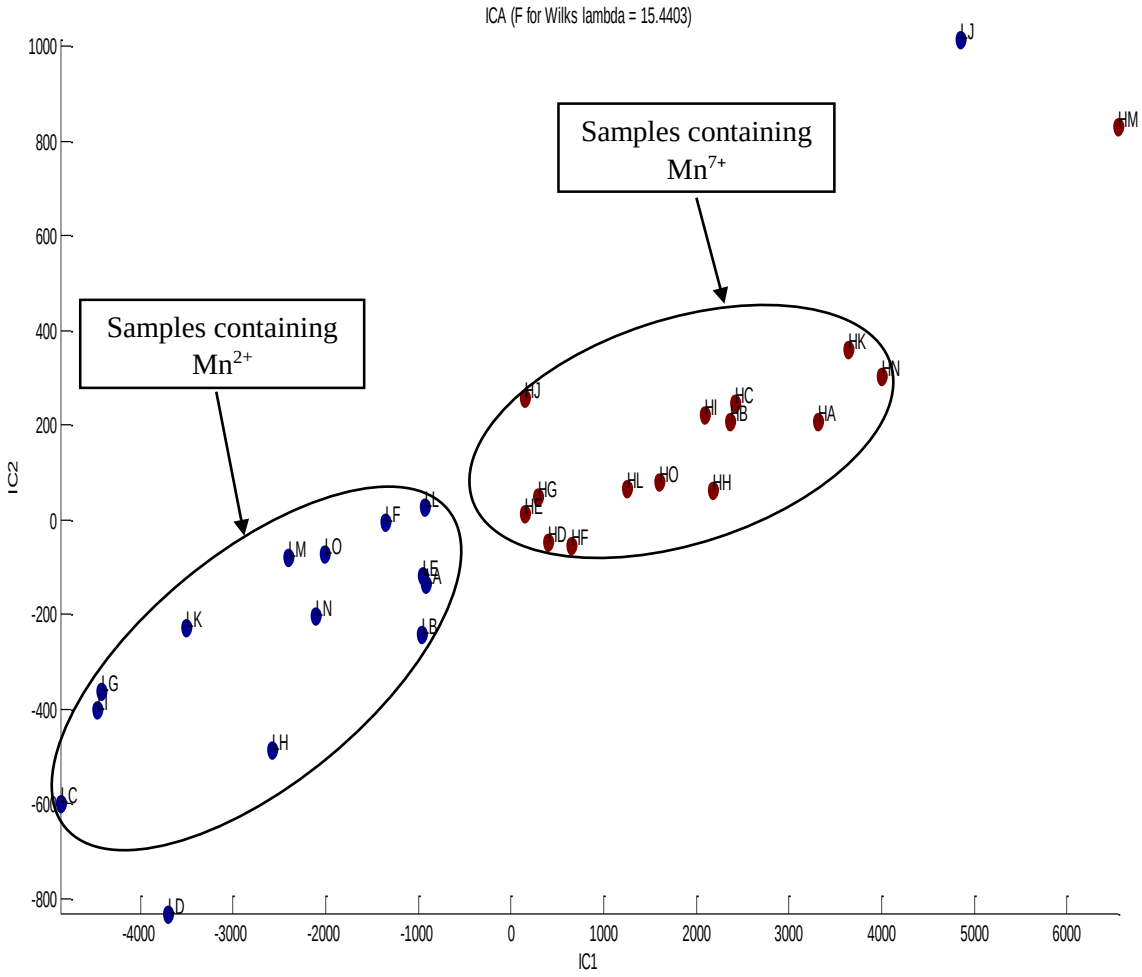


Figure 5.5.7: ICA scores for speciation analysis model of Mn simulate tissues 10 µm thick at 50 s utilizing selected fluorescence.

To perform speciation analysis of Cu in model soft tissue samples, preselected fluorescence region of K_α and K_β peaks of Cu from EDXRFS spectra were used for model input. ICA scores for the speciation model of Cu in Figure 5.5.8 shows the discrimination of Cu into low speciation (samples containing Cu^{+}) and high speciation (samples containing Cu^{2+}). The two clusters are influenced by K_α and K_β intensities of Cu.

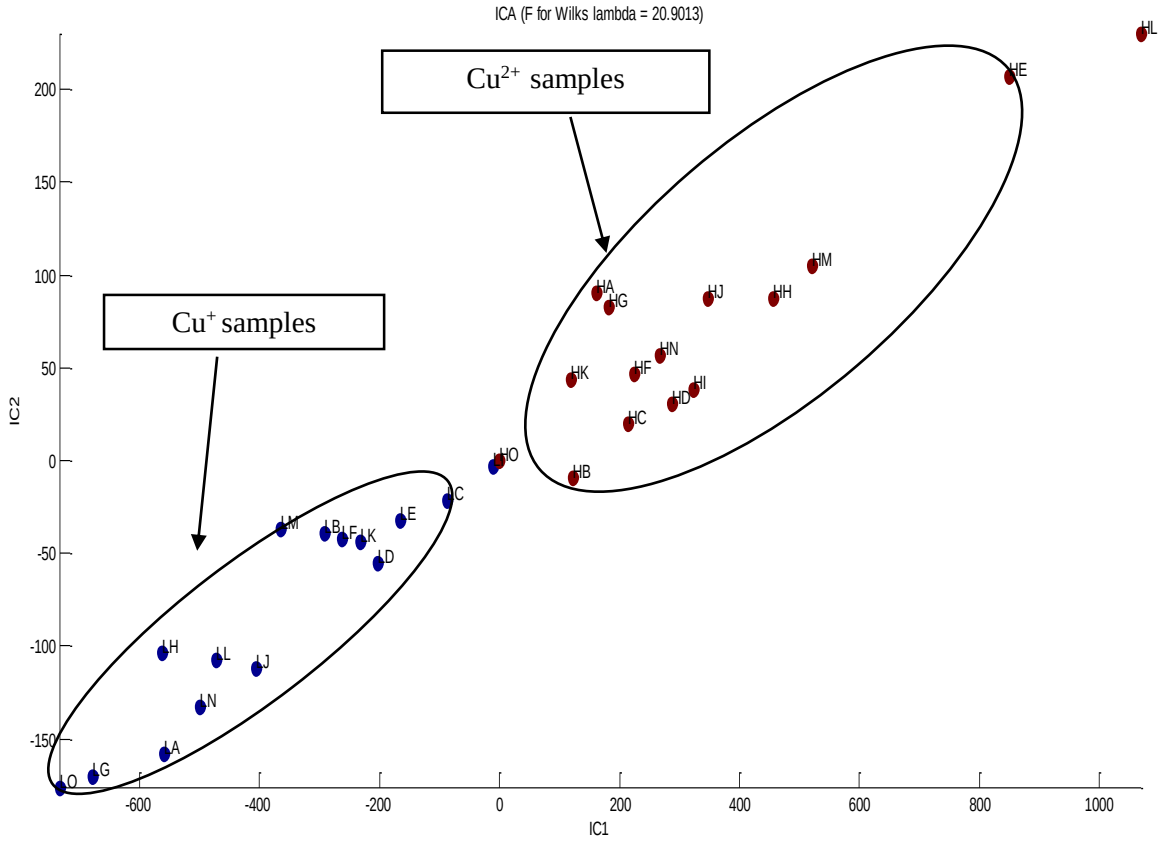


Figure 5.5.8: ICA scores for speciation analysis model of Cu simulate tissues, 10 µm at 50 s utilizing selected fluorescence.

For speciation analysis of Fe, selected fluorescence region of Fe peaks was utilized as input into the ICA model. Figure 5.5.9 shows the classification of Fe simulate samples into low and high oxidation states via ICA. The samples were effectively discriminated into low speciation (Fe^{2+}) and high speciation (Fe^{3+}) influenced by K_α and K_β intensity signals and total elemental content.

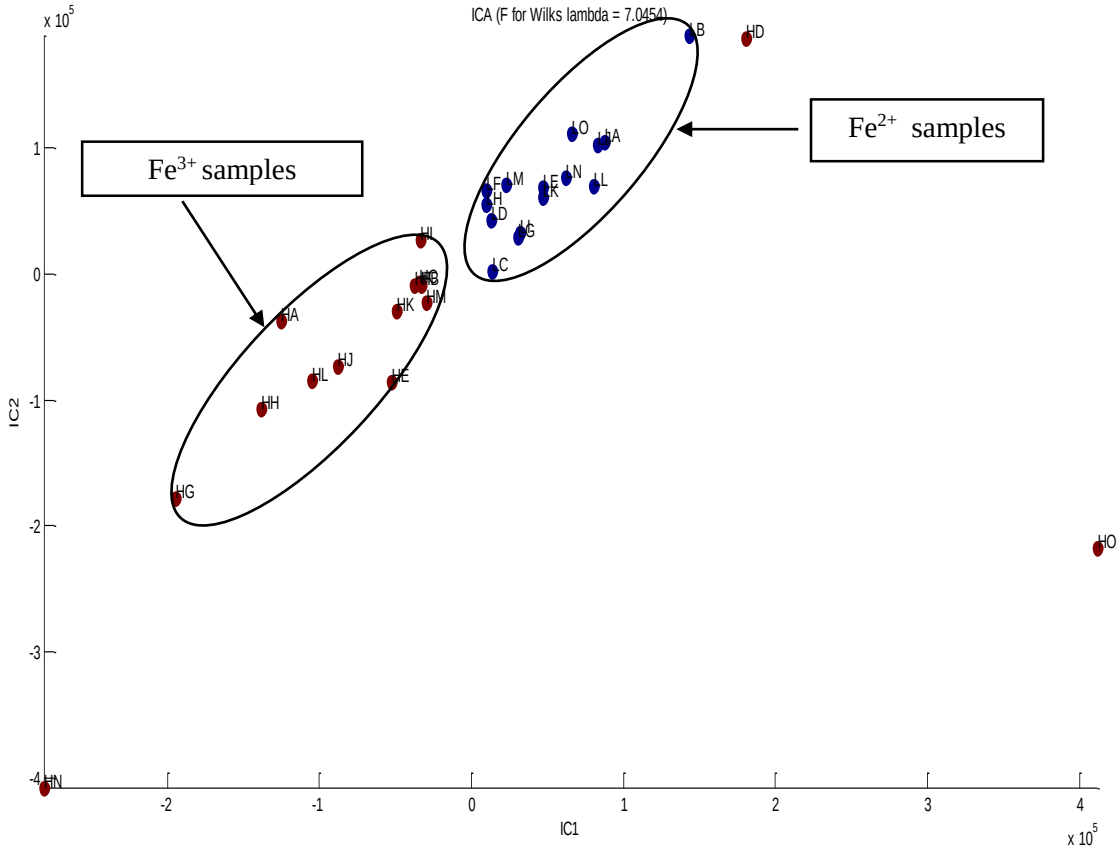


Figure 5.5.9: ICA scores for speciation analysis model of Fe simulate tissues 10 µm at 50 s utilizing selected fluorescence.

Figures 5.5.7-9 show the clustering of simulate samples for the biometals (Mn, Cu, Fe) into lower speciation (Fe^{2+}, Cu^{+} and Mn^{2+}) or higher speciation (Fe^{3+}, Cu^{2+} and Mn^{7+}). ICA ellipsoids corresponding to the two oxidation states are significantly smaller. This means that ICA models have greater stability and reliability. ICA enables the separation of samples that overlap in the space of PCA scores. If the values of ICA scores of new samples fall into an ellipsoid of smaller size, that indicates a higher probability of identification of sample group affiliation in comparison with PCA model.

In general, the above results are in agreement with the known speciation of Cu, Mn and Fe in the calibration samples. This therefore validates ICA as a tool to qualitatively determine the speciation of Cu, Mn and Fe in unknown tissue samples of the similar matrix. Therefore ICA was consequently applied to perform classification of native body tissues to identify ones that were having lower or higher oxidation states of the biometals. This analytical results has great significance in disease diagnostics utilizing tissue trace biometal and associated speciation as disease biomarkers.

Furthermore it was found that two ICA models makes the attributes of the samples to lie inside a subspace which is limited (Frank, 2005). It is concluded therefore that two components are inadequate in describing the data and all the results henceforth are from ICA model with three components.

5.6 Application of robust chemometrics approach in analysis of domestic dog tissue

In practice EDXRF spectra of soft body tissue is characterized by spectral overlaps, weak fluorescence signals and background noise making it difficult to perform quantitative analysis. The use of EDXRF method is not applicable to speciation analysis in principle. Therefore this calls for development of a robust multivariate chemometrics techniques towards elemental analysis for low-Z elements in biological samples. The domestic dog was chosen in this study as it has the same pathogenic base for cancer as humans.

5.6.1 Combined use of WT-PCA and ANN for quantitative analysis of trace biometals

Before analysis of the EDXRFS spectral data by ANN for determination of concentration of the biometals, the spectra were preprocessed by the combined use of WT and PCA for background noise reduction and resolution enhancement which makes quantitative analysis difficult. The elemental concentrations in domestic dog tissues were obtained by using the validated ANN which was the better model as compared to PLS. Table 5.6.1 shows the results

of trace elements: Fe, Cu, Na, Mg, Zn, Mn, Co concentrations in cancerous and healthy domestic dog tissue.

Based on the histopathological reports, the dog tissues were classified as cancerous or non-cancerous. Results of concentration levels are presented in Table 5.6.1 in which there is a significant differences in concentrations of the biometals namely; Fe, Zn, Cu and Mg between healthy and cancerous tissues. As shown in Table 5.6.1, concentrations of Fe, Zn, Cu and Mg in mammary cancer tissues were high as compared to normal mammary tissues in the ratio 3:1, 2:1, 3:1 and 2:1 respectively; showing that cancer is strongly affected by the level of concentrations present in the tissues.

Similarly, concentrations of Fe, Zn, Cu and Mg in prostate cancer tissues (Table 5.6.1) were high as compared to normal prostate tissues in the ratio 5:2, 2:1, 2:1, and 2:1 respectively; also signifying that the prostate cancer is strongly affected by the level of concentrations of these elements present in the tissues. Thus the biometals are the cancer biomarkers.

Results in Table 5.6.1 shows elevated concentrations of Fe, Zn, Cu and Mg in mammary cancer and prostate cancer tissues as compared to normal mammary and prostate tissues; suggesting that they may strongly be affected by the type of cancer.

Table 5.6.1: Concentrations of biometals for histopathologically classified dog tissue samples analyzed using selected fluorescence and scatter peaks

Health status	Sample name	Elemental concentrations ± standard deviation (μg/g)						
		Na	Fe	Mn	Co	Zn	Mg	Cu
Normal mamary	SD	34.8±0.5	45.4±3.0	29.2±3.5	64.3±3.6	165.4±8.0	80.6±5.0	39.3±4.0
	SI	34.8±1.5	46.5±3.0	29.9±2.5	65.2±4.2	171.1±10.1	70.1±5.0	40.4±2.7
Mammary cancer	SA	37.4±1.0	137.0±8.2	30.5±3.6	61.8±5.1	289.3±4.0	110.3±5.0	127.9±10.0
	SC	35.4±1.0	139.3±6.2	28.8±4.1	60.8±6.5	300.6±3.0	109.0±7.8	131.4±6.5
	SG	35.7±1.7	138.4±9.2	29.8±2.5	59.8±4.1	288.6±11.1	112.2±3.5	120.2±4.5
Normal prostate	SE	34.5±0.8	50.3±2.0	30.3±2.0	68.6±1.5	159.6±3.0	122.8±7.5	48.8±3.5
	SH	37.3±1.0	58.7±5.0	30.1±2.5	74.5±6.0	108.3±7.5	79.7±5.2	67.90±4.0
Prostate cancer	SB	35.4±1.0	152.5±9.1	29.9±2.5	71.3±3.6	301.4±4.0	125.4±10.0	148.4±7.5
	SJ	37.4±1.0	145.1±3.8	29.8±2.5	68.9±3.5	296.0±15.0	113.4±5.0	134.9±7.0
	SF	36.4±1.0	144.9±3.2	29.8±2.5	70.3±1.7	291.5±5.6	111.5±5.4	132.3±8.0

The results show concentration of Zn was relatively high in all the tissues (108.3±7.5 - 301.4±4.0) as compared to any other biometal (e.g. Fe at 45.4±3.0 -152.5±9.1) which is the closest in comparison particularly in prostate and mammary cancer tissues. This can be attributed to the fact that Zn is vital for growing cells and cancer is known to be uncontrolled growth of cells, thus accumulation of Zn in those tissues implies that Zn participates in tumorigenesis. Zn is found to accumulate in mammary tumors and support tumor growth (Lee *et al.,* 2003). The accumulation of Zn in the healthy prostate as well is because it plays an important role for male fertility.

There was increased level of Fe in both mammary cancer tissues (SA, SC and SG) and prostate cancer tissues (SB, SJ and SF) as compared to healthy tissues (SD, SE, SH and SI). Fe is

associated with tumorigenesis since it is the source of formation of mutagenic hydroxyl radical which interferes with the repairing of DNA and thus affecting the signal transduction in cancer cells by acting as a nutrient that proliferates tumors (Zhang and Zhang, 2015).

Increased concentrations of Cu in mammary and prostate cancer tissues as compared to healthy mammary and prostate tissue was evident. Cu has been found to increases in malignant states. Reducing agents such as ascorbate or superoxide radical reduces copper complexes to the cuprous state. The complexes react with hydrogen peroxidase forming hydroxyl radicals as product that damage DNA. OH radicals may perhaps cause double strand breaks in the cellular DNA initiating malignant process (Ehud *et al.,* 1983).

Figure 5.6.1 summarizes how the concentration of Na, Mg, Co, Zn, Fe, and Cu are varying with the status of the tissue (i.e. normal mamary, mammary cancer, normal prostate and prostate cancer).

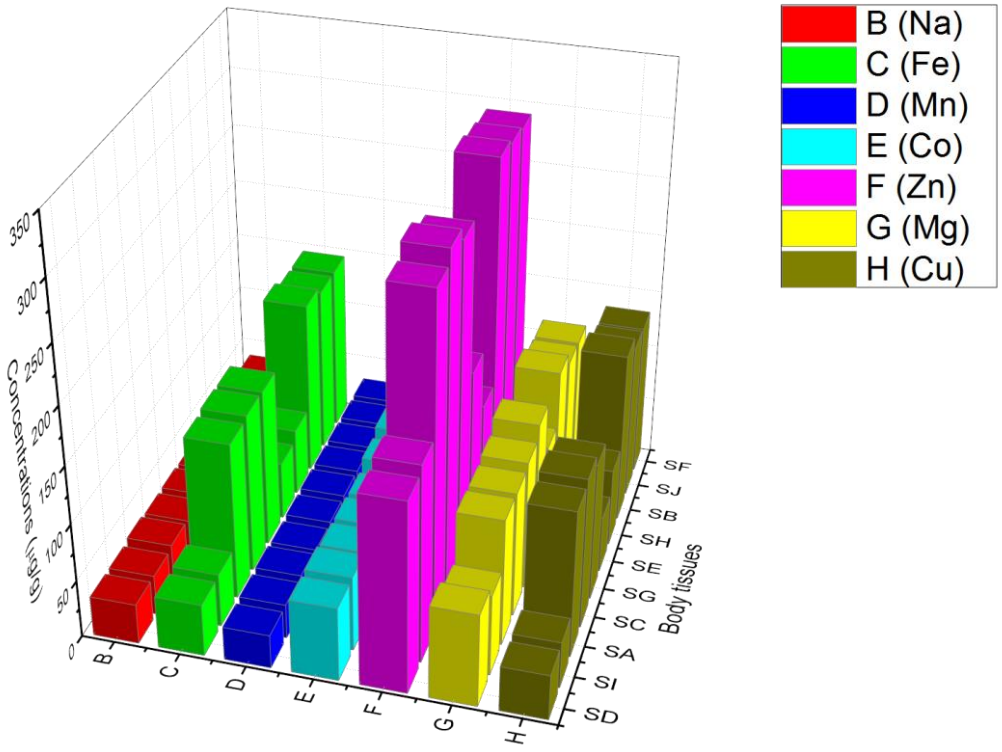


Figure 5.6.1: 3-D Grouped bar graph showing the variation in concentration of trace and major elements in cancerous and healthy tissues.

In general, most of the above biometals are important for various biological and enzymatic processes hence changes in concentrations of these elements (Mn, Fe, Cu, Zn, Mg) in cancerous tissue are realistic. This therefore strongly supports the fact that the trace biometals are biomarkers for disease diagnosis (cancer).

The relationship between biological processes and biometals concentrations in tumor in native soft body tissue (ten dog tissues) were determined using correlation tests between biometals (Mn, Fe, Cu, Zn, Co, Mg and Na) concentrations in malignant tissues. Pearson correlation coefficients between the biometal concentrations in cancerous tissues are presented in Tables 5.6.2 and 5.6.3.

Table 5.6.2: Pearson correlation coefficients between biometals concentrations in mammary cancer tissue

Element	Pearson correlation coefficients between biometal concentrations						
	Na	Fe	Mn	Co	Zn	Mg	Cu
Na	1						
Fe	-0.967	1					
Mn	0.884	-0.974	1				
Co	0.788	-0.604	0.410	1			
Zn	-0.574	0.765	-0.890	0.052	1		
Mg	0.032	-0.287	0.495	-0.590	-0.837	1	
Cu	0.073	0.184	-0.400	0.672	0.775	-0.994	1

Table 5.6.3: Pearson correlation coefficients between biometal concentrations in prostate cancer tissue

Element	Pearson correlation coefficients between biometal concentrations						
	Na	Fe	Mn	Co	Zn	Mg	Cu
Na	1						
Fe	-0.854	1					
Mn	-0.867	0.999	1				
Co	-0.995	0.801	0.814	1			
Zn	-0.545	0.901	0.891	0.462	1		
Mg	-0.796	0.995	0.992	0.734	0.941	1	
Cu	-0.781	0.992	0.989	0.718	0.949	0.999	1

In prostate cancer, the Pearson correlation coefficient between Mn and Fe, Zn and Fe, Cu and Mn and Cu and Zn was found to be positive (0.999, 0.901, 0.989 and 0.949) respectively. There

is strong positive correlation between Mn with Fe and Cu. This is a result of Mn being a co-factor of these elements in cancerous prostate tissues thus resulting to increased metabolic activities for carcinogenesis.

In mammary cancerous tissues, the Pearson correlation coefficient between Zn and Fe (0.765), Cu and Zn (0.775), Mn and Na (0.884) and Co and Na (0.788) was evident. This can be related to their co-factor roles to development of of mammary cancer.

However, the most significant differences in elemental associations between prostate and mammary cancer can possibly be based on the strong positive correlation between Cu and Mg (0.999) and, Mn and Fe (0.999) in prostate tumors with corresponding negative correlations in mammary tumors; Cu and Mg (-0.994), Mn and Fe (-0.974). Furthermore, there exist strong positive correlations between Cu and Zn, Fe and Zn in both prostate and mammary cancer tissues. This can be related to their co-factor roles to development of cancer.

In conclusion therefore, the strong positive correlation between the biometals Mn and Fe, Cu and Zn, Cu and Mn , Mn and Zn supports the fact that the coexistence of these trace biometals in tissues can be used as biomarker for diagnostic of prostate cancer. The positive correlations between Zn and Fe, Cu and Zn, Mn and Na , Co and Na can also be used as biomarkers for diagosis of mammary cancer.

5.6.2 Speciation analysis of dog tissues spectra by independent component analysis (ICA)

The determination of the speciation of Fe, Cu and Mn was aimed to give better insight in the mechanism of cancer development. The results of section 5.5.2 were in agreement with the known speciation of Cu, Mn and Fe in the calibration samples. This therefore validates ICA as a tool to qualitatively determine the speciation of Cu, Mn and Fe in unknown tissue samples of the similar matrix. Thus validated ICA multivariate technique was utilized for determination of

the speciation of Fe, Cu and Mn in native cancerous soft tissues to find their role in progression of cancer.

Prior to ICA speciation analysis, the EDXRF spectral was preprocessed via wavelet transform for noise reduction and ICA for $K_{\alpha 1}$ and $K_{\alpha 2}$ peaks resolution. The native domestic dog tissue were identified either as of low speciation (Cu^{+}, Fe^{2+} and Mn^{2+}) or high speciation (Cu^{2+}, Fe^{3+} and Mn^{7+}) depending on whether they were classified together with the known speciation of the calibration samples.

Figure 5.6.2 shows the classification of the dog tissue samples into high speciation and low speciation of Fe via ICA. The samples; SA, SB, SC, SD, SC, SE, SF, SG, SI, SJ are native soft dog body tissues. The spectra was preprocessed via WT-ICA before performing speciation analysis.

Figure 5.6.2 shows that two clusters were identified which represents the samples with high (Fe^{3+}) and low (Fe^{2+}) speciation respectively. The samples SH, SE, SI and SD were found to have (Fe^{2+}) and while samples SA, SB, SC, SF, SG and SJ were found to be dominated by (Fe^{3+}) as they were classified together with respective high and low oxidation states of the known simulate samples. According to histopathological results of analyses, SD, SI are healthy mammary tissues and SE, SH are healthy prostate tissues. The samples SA, SC, SG are mammary cancerous tissues while SB, SF and SJ are prostate cancer tissues.

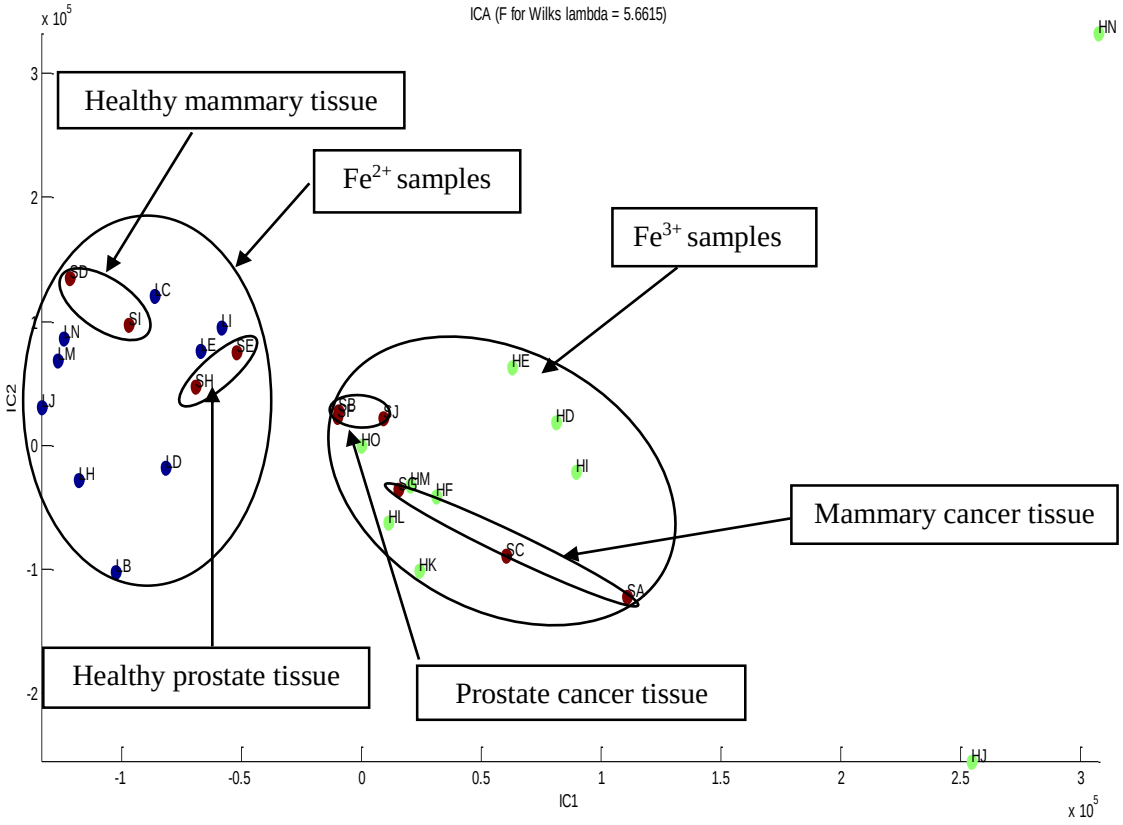


Figure 5.6.2: ICA score plot for speciation of Fe in domestic tissue utilizing selected fluorescence.

Figure 5.6.3 shows two clusters were identified which represents the samples with low (Mn^{2+}) and high (Mn^{7+}) oxidation states respectively. The samples SH, SE, SI and SD were found to have (Mn^{2+}) and while samples SA, SB, SC, SF, SG and SJ were found to be dominated by (Mn^{7+}) as they were classified together with respective high and low speciation of the known simulate samples.

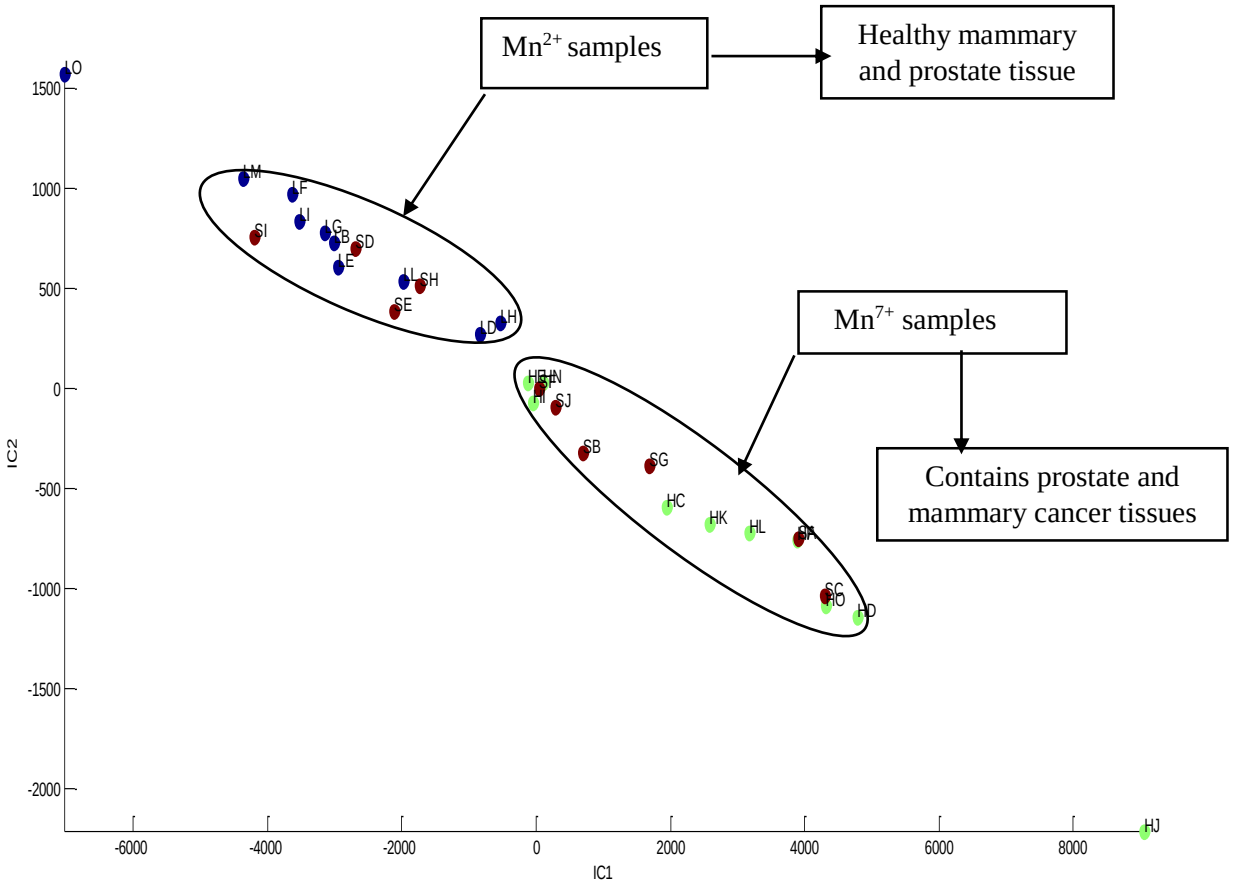


Figure 5.6.3: ICA score plot for speciation of Mn in domestic tissue utilizing selected fluorescence.

Figure 5.6.4 shows two clusters were identified which represented the samples with high (Cu^{2+}) and low (Cu^{+}) speciation respectively. The samples SH, SE, SI and SD were dominated by (Mn^{2+}) and while samples SA, SB, SC, SF, SG and SJ consisted mostly of (Mn^{7+}) as they were classified together with respective low and high speciation of the known simulate samples.

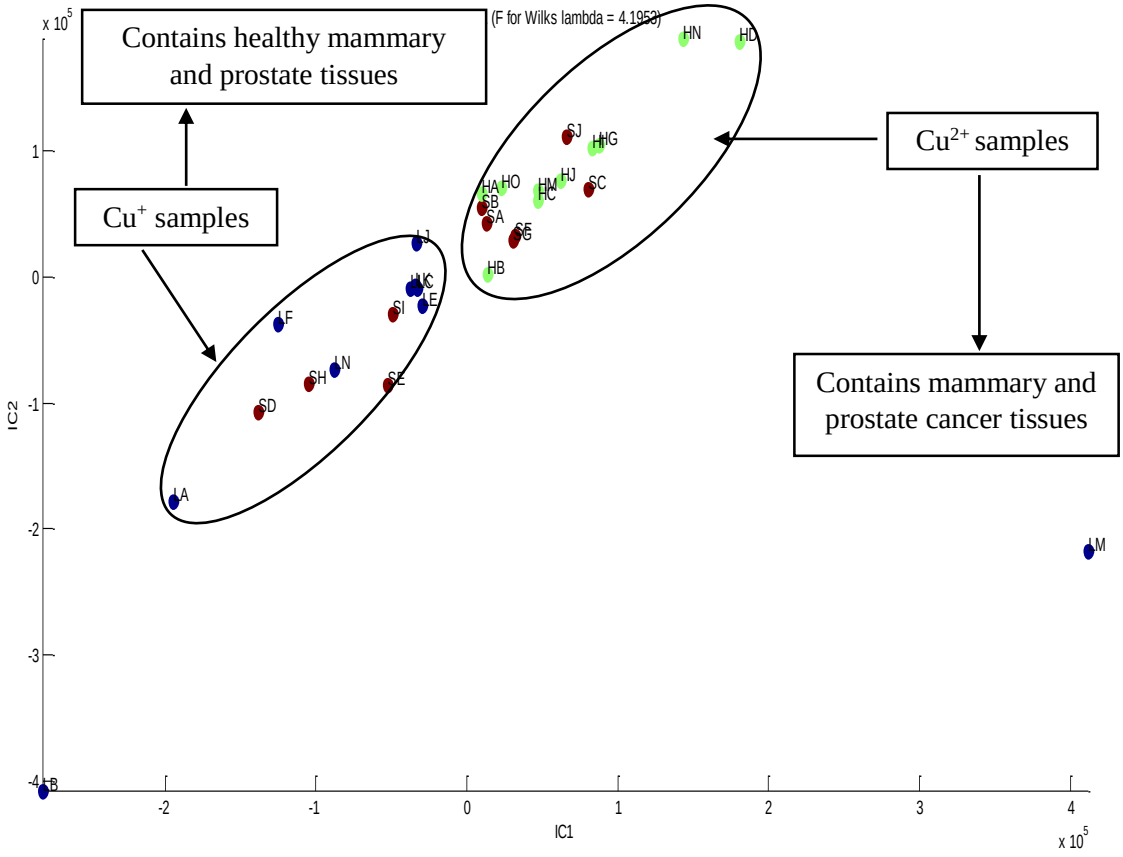


Figure 5.6.4: ICA score plot for speciation of Cu in domestic tissue utilizing selected fluorescence.

The speciation results above show that mammary and prostrate cancer tissues were rich in Cu, Mn and Fe occurring mostly in the Fe^{3+} , Mn^{7+} and Cu^{2+} oxidation states respectively which demonstrates the possible production of free radicals in cancerous tissues in Fenton reaction based on the oxidation states of Cu and Mn (Rojas *et al*., 1999). Due to the ability of Cu to change between its two oxidation states; Cu^{+} and Cu^{2+}, Cu has been discovered to cause generation of Reactive Oxygen Species (ROS) which produce hydroxyl radicals that modify proteins, lipids, and nucleic acids thus leading to DNA damage (Armendariz and Vulpe, 2003). The results are in agreement with an earlier study using XANES (Kwiatek *et al*., 2004), in which the cancerous tissue were found to be dominated by Fe^{3+}.

The ability of the analytical method to perform speciation analysis enables the development of an approach for spectral diagnosis of cancer utilizing tissue biometal speciation profiles as cancer biomarkers. From these results it is clear that speciation alterations, as well as correlations of trace biometals in body tissues are cancer biomarkers.

5.6.3 Principal component analysis (PCA) of dog tissue

Figure 5.6.5 shows classification of the native soft dog tissue based on the tissue pathological state, in which 3 classes have been identified in the PC1×PC2 score plot with explained variance of 99 % (91 % and 8 % for PC1 and PC2 respectively).

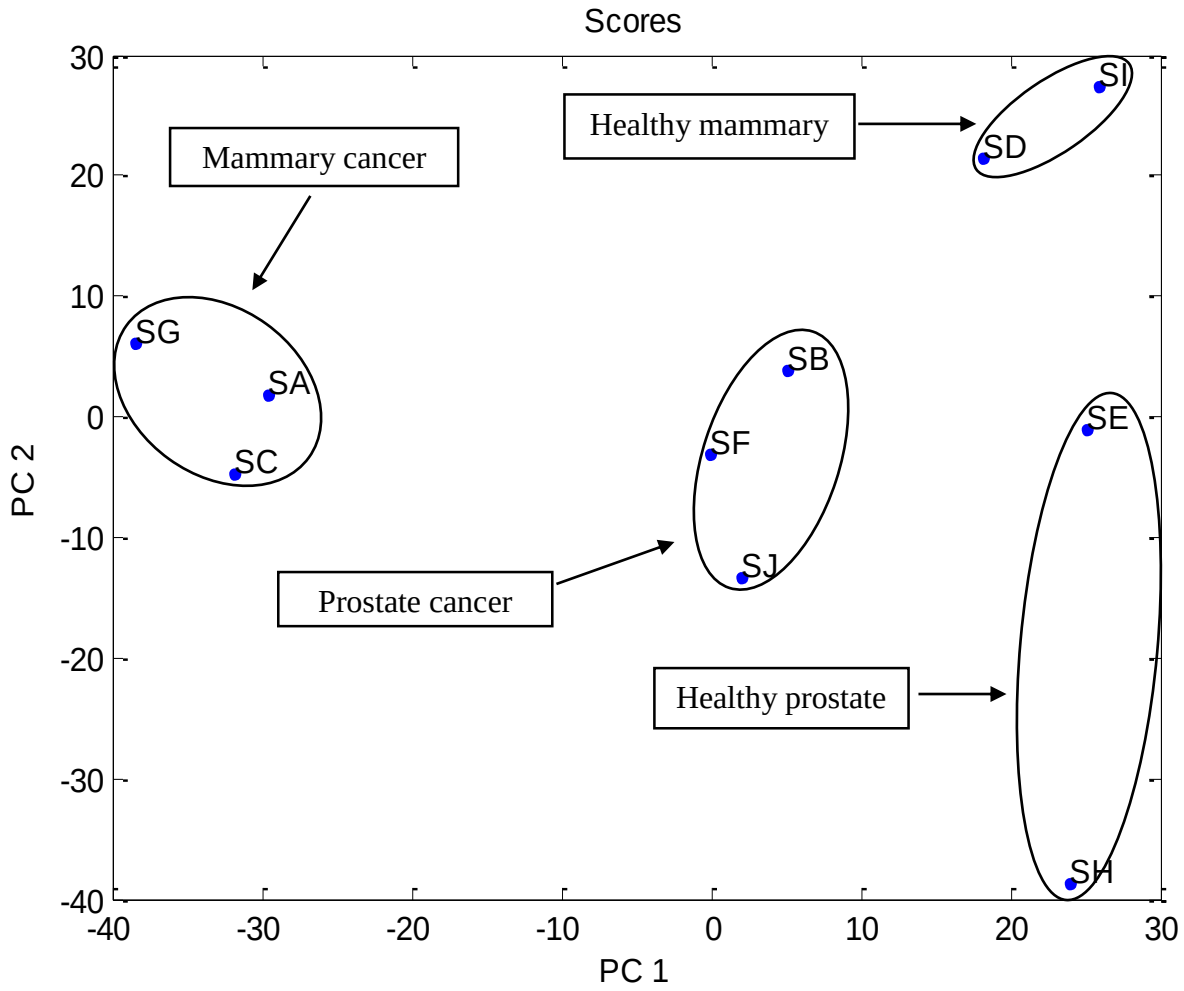


Figure 5.6.5: PC1 (91%) × PC2 (8%) score plot analysis for native soft dog tissue samples using selected fluorescence signature plus Compton scatter at a live time of 50 s 10 µm thick.

Prior to PCA of the spectral data, the spectra was first preprocessed by combined use of WT and ICA for noise reduction and resolution of peaks. The preselected fluorescence region of interest for the elements Na, Mg, Fe, Mn, Cu, Zn and Co was used as an input into PCA model. Fig. 5.6.6 shows the variables biometals responsible for the above classification of dog tissues. Based on the PC1 loadings plot, trace elements; Cu, Mn, Fe and Zn determine the health state of tissues depending on the concentration levels. Prostate and mammary cancer are influenced by concentrations of the biometals Cu, Zn and Fe. The results clearly show that biometals especially; Fe, Cu, Mn, and Zn are valuable biomarkers for cancer diagnosis

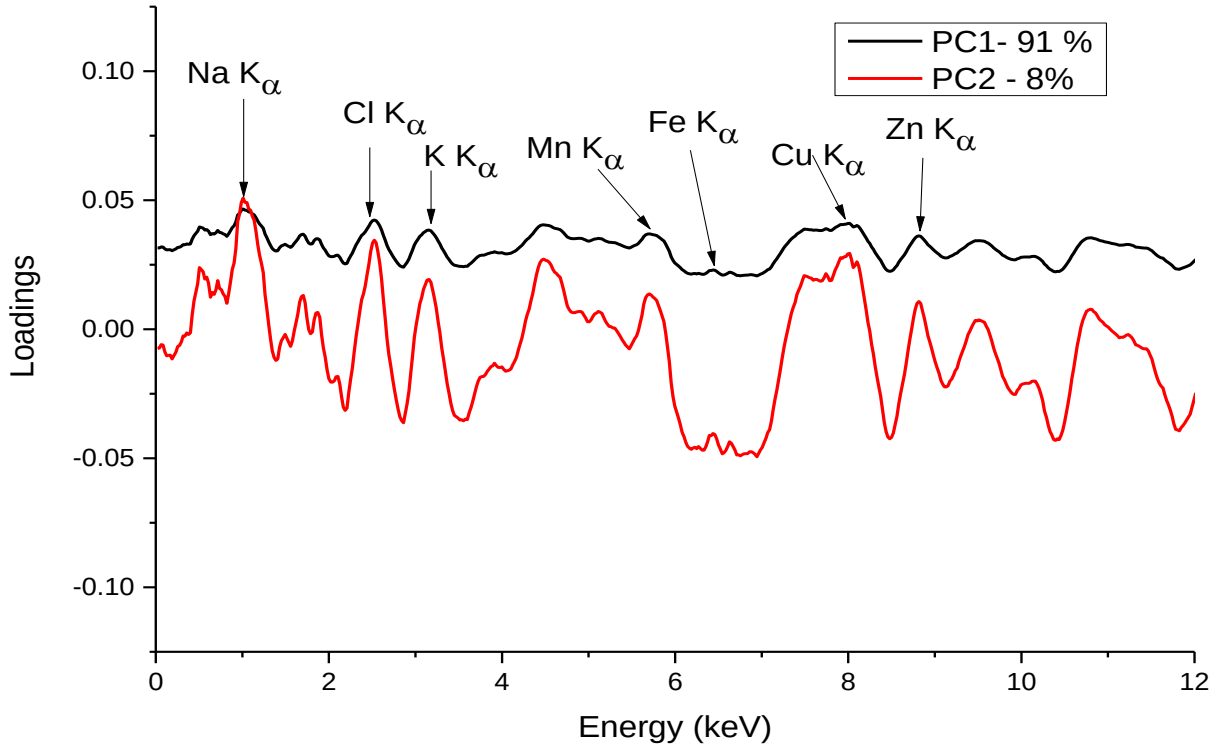


Figure 5.6.6: PCA Loadings spectrum showing the variables in domestic dog tissue samples 10 µm at a live time of 50 s using selected fluorescence peaks plus Compton scatter.

Figure 5.6.7 shows clustering of the native soft dog tissue based on the tissue pathological state, where 3 classes have been identified in the PC1×PC2 score plot with explained variance of 97 % (90 % and 7 % for PC1 and PC2 respectively). The result of classification shows that

exclusion of Compton scatter has a slight impact on classification with reduced explained variance from 99 % to 97 %. This can be attributed to the fact Compton scatter contains information about low- Z elements (Na and Mg) which were present in both healthy and cancerous tissues.

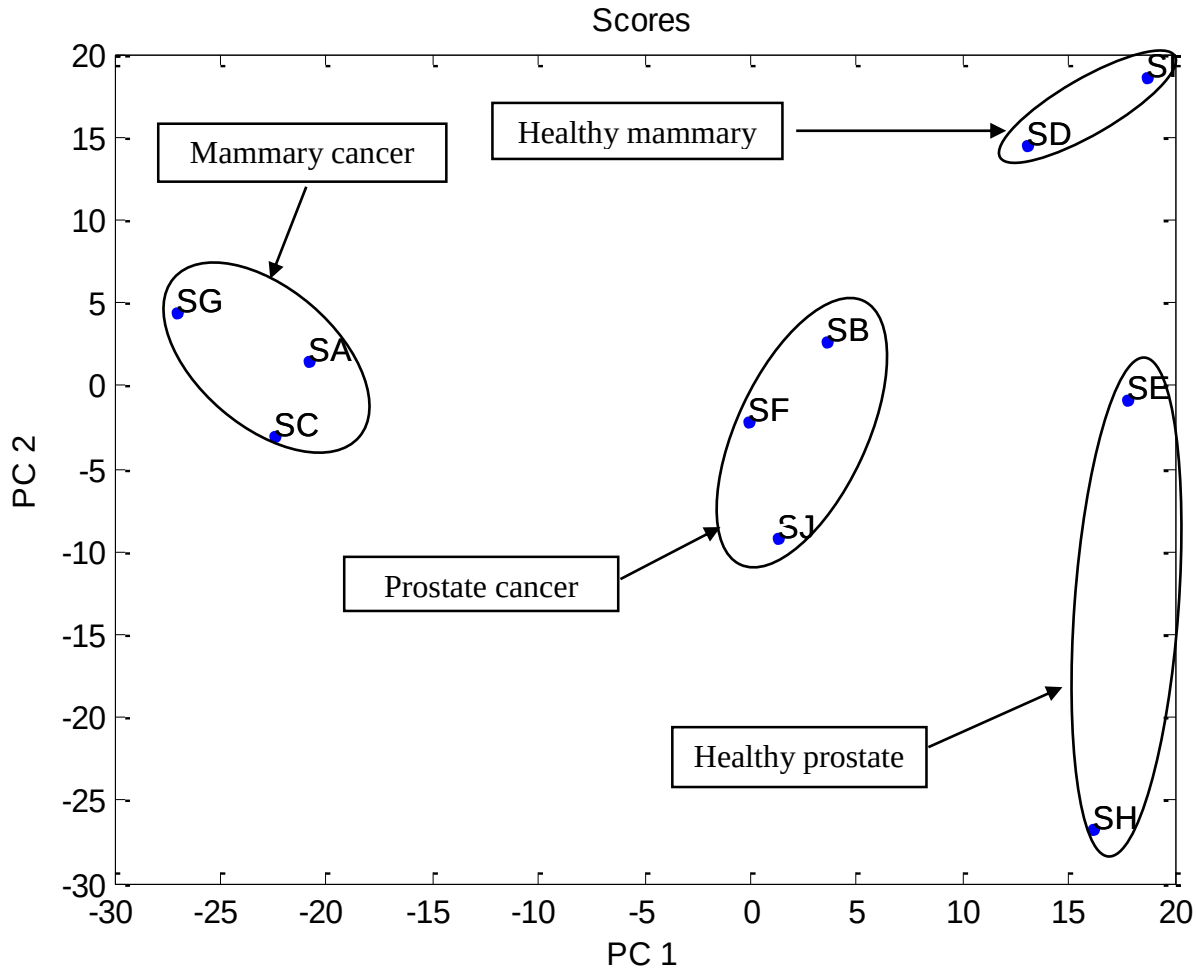


Figure 5.6.7: PC1 (90%) × PC2 (7%) score plot analysis for native soft dog tissue samples using selected fluorescence signature at a live time of 50 s 10 µm thick.

In general, these results show that PCA exploratory analysis can be used in discriminating body tissue samples based on their healthy status and thus can have an application in cancer diagnosis. Based on the results of exploratory, quantitative and speciation analysis of biometals

Mn, Zn, Na, Mg, Cu, Co and Fe in soft body tissue (simulate and dog) using the robust chemometrics approach, can be utilized in diagnosis of prostate and mammary cancer.

CHAPTER SIX: CONCLUSION AND RECOMMENDATIONS

6.1 Conclusions

This study was undertaken to develop a robust chemometrics approach in energy dispersive X-ray fluorescence and scattering (EDXRFS) spectrometry of complex matrices such as native soft body tissues towards direct, rapid and noninvasive determination of the concentration and speciation of selected trace elements (Zn, Fe, Na, Mg, Cu, Mn and Co) in various thickness. Methods for spectral preprocessing and multivariate calibration based on chemometrics techniques were developed. Chemometrics techniques namely principal component analysis (PCA), wavelet transforms (WT), independent component analysis (ICA), artificial neural networks (ANN) and partial least squares (PLS) were utilized.

Combined use of wavelet analysis and independent component analysis was utilized in spectral preprocessing towards speciation analysis which was reliable in resolution K_α and K_β peaks and noise reduction. However the signal to noise ratio (SNR) values were low and the algorithm retained significant amount of noise. ICA simply seperates the components which is not sufficient to denoise the data and this explained the low values of SNR. The combined use of multivariate wavelet and principal componet analysis (PCA) processing of EDXRFS spectra was able to reduce the noise present in the signal while preserving the signal characteristics irrespective of its frequency with much improved SNR and it optimized well the spectra for determination of elemental concentration by reducing background noise which masks subtle peaks making quantitative analysis difficult.

ICA speciation analysis results of Mn, Fe and Cu after spectral preprocessing were better than that of PCA. This could be applied to characterization of cancer based on concentration of trace elements (Fe, Cu, Zn, Mn, Co, Mg and Na) and speciation of Cu, Fe and Mn qualitatively, using the featured X-ray fluorescence and scatter peaks conjointly.

The X-ray scatter peaks had a strong relationship with low-*Z* elements. Preselected fluorescence (Fe, Cu, Zn, Mn and Co) signature performed correctly for ANNs for the high-*Z* elements, and fluorescence plus Compton scatter peak for the low-*Z* elements (Na and Mg) making it possible to quantify the low-Z elements which is difficult task with classical XRF. The non-linear quantitative relationships (matrix effects) between trace and major elements concentration and EDXRFS spectra were corrected significantly by spectral preprocessing procedures and ANNs. The ANNs models results were significantly improved as compared to those of PLS in predicting of major and trace biometals in tissues used.

The developed analytical strategy which involved the hybrid use of chemometrics techniques in spectral preprocessing towards quantitative and qualitative speciation analysis of biometals performed better compared to the traditional XRF for the two samples (simulate and native soft body tissues), with the former furnishing elemental concentration and speciation information rapidly i.e. shorter sample analysis times, 50 s compared to classical EDXRF, 2 000 s.

6.2 Recommendations

For future result reliability, further research is recommended on the development of proper simulate samples encompassing a wider range of the trace and major elements in the representative concentrations and speciation. Alternative chemometrics techniques needs to be utilized especially for pattern recognition since PCA and ICA cannot fully quantify how much of each sample contains in terms of particular ionic species of the trace elements for total elemental speciation analysis. The two are only able to classify as either low or high speciation. Most promising would be incorporating in the methodology the use of K- nearest neighbor (KNN) technique.

REFERENCES

Abbott, P. H., Adams M. J. (1997). AXIS: Automated XRF interpretation of spectra. *X-ray Spectrometry*: **26**: 125–131.

Aminghafari, M., Cheze, N., and Poggi, J. M. (2006). Multivariate denoising using wavelets and principal component analysis. *Computational Statistics & Data Analysis, **50***: 2381-2398.

Andermann, G. and Kemp J.W. (1958). Scattered X-rays as internal standards in X-ray emission spectrometry, Anal.chem **30**: 1306-1309.

Angeyo K. H., Gari S., Mangala J. M. and Mustapha A. O. (2012). Feasibility for direct rapid energy dispersive X-Ray fluorescence and scattering analysis of complex matrix liquids by partial least squares. Applied Radiation & Isotopes **70**: 2596-2601.

Anjali, S. (2014). Determination of minor and trace elements concentrations in kidney stones using elemental analysis techniques, Msc. Thesis, Missouri University of Science and Technology.

Antonio, B., and Bruno G. (2001). Fit of EDXRF spectra with a genetic algorithm, *X-Ray Spectrom, **30***, 32–36.

Armendariz, D. and Vulpe, A. D. (2003). 11th International Symposium on Trace Elements in Man and Animals Abstracts. *The Journal of Nutrition* **133(5)**: 203E–282E.

Banas, A., Kwiatek, W. M. and Zajac W. (2001). Trace element analysis of tissue section by means of synchrotron radiation: the use of GNUPLOT for SRIXE spectra analysis, *Journal of Alloys and Compounds, **328***: 135–138.

Barkraj, E. H., (2006). Application of multivariate statistical methods to classify archeological pottery from tel-Alramad site, based on X-ray fluorescence analysis, *X-Ray Spectrom,* **35**: 190-194.

Beale, R. and Jackson, T. (1991). *Neural Computing: Data Analysis for the Laboratory and Chemical Plant*, John Wiley and Sons, Bristol.

Bechoff, B., Kanngierfer, B., Weddel, R., Wolff H., and Langhoff, N. (2006). *Practical X-ray Fluorescent Analysis*, Springer, Berlin.

Bernaus, A., Gaona, X., Van R. D., and Valiente, M. (2006) Determination of mercury in polluted soils surrounding a chlor-alkali plant-direct speciation by X-ray absorption spectroscopy techniques and preliminary geochemical characterization of the area, *Anal chim Acta,* **565**: 73-80.

Bernhard, S., Burges, C. C. and Smola, A. J. (1999). *Pairwise classification and support vector machines*. The MIT Press, C. Massachusetts, London England, 255-268.

Bos, M., and Weber, H. T. (1991). Comparison of the training of neural networks for quantitative X-ray fluorescence spectrometry by a genetic algorithm and backward error propagation, *Anal.chim.Acta,* **247(1)**: 97-105.

Boyle, J. F. (1999). Isotope –source energy –dispersive XRF analysis of geological materials using gas-filled proportional counters: signal deconvolution using simulated peak shapes, *X-Ray Spectrom,* **28(3)**: 178-182.

Bueno, M. I. M. S., Fabiola, M. V. P., Edenir, R. P. F. (2006). Development of a methodology for calcium, iron, potassium, manganese and zinc quantification in teas using X-ray spectroscopy and multivariate calibration. J. *Agric. Food Chem*, **54**: 5723-5730.

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

Brouwer, P.N. (2003). Theory of XRF, Getting Acquainted with the Principles, PANalytical BV: Almelo, pp. 23 – 30.

Cortes, C., and Vapnik, V. (1995). Support vector network. *Machine learning*, **20**: 273-297.

Custo, G., Boeykens, S., Cicerone, D., Vazquez, C. (2002). Combining XRF analysis and chemometric tools for preliminary classification of Argentine soils, *X-Ray Spectrom,* **31**:132-135.

Criss, J. W., and Birks, L. S. (1968). Calculation methods for fluorescent X-ray spectrometry. Empirical coefficients versus fundamental parameters. *Anal. Chem*, **40**: 1080-1086.

Daniel, B., and Anusuya, C. (2011). Chemometrics analysis of EDXRF measurements from fossil bone. *X-ray spectrom,* **40**: 441-445.

Debertin K and Helmer R.G (1988). Gamma and X-ray Spectrometry with Semiconductor Detectors, Elsevier Science Publishers B.V., Amsterdam: 29-37.

Despagne, F., and Massart, D. L. (1998). Neural networks in multivariate calibration. *Analyst,* **123(11)**: 157-178.

Devos, O., Mouton, N., Sliwa, M., Ruckebusch, C. (2011). Baseline correction methods to deal with artifacts in femtosecond transient absorption spectroscopy, Anal. Chim. Acta **705**: 64–71.

Ehud, J., Margalioth M. D., Joseph G., Schenker M. D., and Mordechai C. (1983). Copper and zinc levels in normal and malignant tissues. *AmericanCancer society*, **52**: 868-872.

Elise, M., Caroline, T., Herve, C., and Patric, D. (2010). Multivariate denoising methods combining wavelet and Principal component analysis for mass spectrometry data. *Proteomics*, **10**: 2564-2572.

Enrich, C., Boeykens, S., Caracciolo, N., Custo, G., Vasquez, C. (2007). Honey characterization by total reflection X-ray fluorescence; evaluation of environmental quality and risk for the human health, *X-Ray Spectrom,* **36**: 215-220.

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, *X-Ray Spectrom*, **28**: 173–177.

Frank, W. (2005). Independent component analysis and regression applied on sensory data. *J.Chemometrics*, **19**: 171-179.

Gerald, R. (1999). Demystification of algorithms and influence coefficients in quantitative XRF analysis, *JCPDS-International centre for Diffraction Data*, 718-731.

Goswami, J. C., and Chan, A. K. (2011), *Fundamentals of Wavelets: Theory, Algorithms, and Applications*, Second Edition, John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/9780470926994.index.

Grafe, M., Tappero, R.V., Marcus, M. A., and Sparks, D. L. (2008). Arsenic speciation in multiple metal environments-II, Micro-spectroscopic investigation of a CCA contaminated soil, *J.Colloid interface Sci*, **321**: 1-20.

Graps, A. (1995). An introduction to wavelets, *IEEE Computational Science and Engineering*, **2(2):** 50–61.

Grzegorz, Z. (2007). Glass analysis for forensic purpose-a comparison of classification Methods, *J. chemometrics,* **21**: 174-186.

Guong, W, Qingzhu, D., and Zhenyu, H. (2008). Independent component analysis and its applications in signal processing for analytical chemistry, TrAC *Trends in Analytical Chemistry*, **27(4)**: 368-376.

He, F., Van Espen, P. (1990). An integrated system for quantitative EDXRF analysis based on fundamental parameters. Nucl. Instrum. Methods Phys. Res. Sect. A: Accel. Spectrom. Detect. Assoc. Equip. **299 (1‑3)**: 580‑583.

Hibbert, D. B., Minkkinen, P., Faber, N. M., and Wise, B. M. (2009). IUPAC project: A glossary of concepts and terms in chemometrics. *Anal. Chim. Acta,* **642**: 3-5.

Hyvarinen, A., and Oja., E. (2000). Independent Component Analysis: Algorithm and Applications. *Neural Networks*, **13**: 411-430.

Jahan, B. G., Mohammad K. R., and Nayereh, A. (2013). Using chemometric methods for overlap correction of Sodium-Zinc spectral lines generated by wavelength dispersive X-ray fluorescence in mineral samples. *X-ray spectrometry*, **43(3)**: 131-137.

Jeena, J., Salice, P., and Neetha., J. (2013). De-noising using soft thresholding, *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, **2(3)**: 1027-1032.

Jenkins, R. (1986). *An Introduction to X-ray Spectrometry*, John Wiley & Sons, New York, USA.

Jens, L., Brian, T. V., Niels, P., Kresten, K., and Hans, C. B. H. (2001). Rapid method for EDXRF analysis of clayey and sandy soil, *X-Ray spectrom,* **30**:186-189.

John, G., Jan, N., Vera, S., Anette, L., Eleonore, B., and Marita, E. (2006). Increased levels of transition metals in breast cancer tissue. *Neuroendocrinology Letters,* **27(1)**: 101–104.

Jose, M. A. G. and Neil, W. B. (2009). Basic Chemometric Techniques in Analytical Spectroscopy. *RSC Analytical Spectroscopy Monographs* **10**: 244-263.

Jordanov, J. U., Tsanov, T. S., Stefanov, R., and Jordanov, N. (1987). Problems of au-tomatic qualitative X-ray fluorescence analysis, *X-ray Spectrometry,* **16(6)**: 255–259.

Kabir, M. A and Shhnaz, C 2012. Comparison of ECG signal denoising algorithm in EMD and wavelet domains. *IJRRAS,* **11(3**): 499-516.

Kaniu, M. I. (2011). Development of a chemometric energy dispersive X-ray fluorescence and scattering spectroscopy method for rapid soil quality assessment, MSc. Thesis, University of Nairobi.

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 spectrometry,* **40**: 432-440.

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 PLS and ANN analytical modeling approaches, *Talanta,* **98**: 236-240.

Kaniu I. M. and Angeyo, K. H. Challenges in Rapid Soil Quality Assessment and Opportunities Presented by Chemometric Energy Dispersive X-Ray Fluorescence and Scattering Spectroscopy. Geoderma (2014) 241-242: 32-40.

Karen, G., Thais, L. A., and Maria, I. S. B. (2007). Sucrose inversion monitored by X-ray scattering and chemometrics, *X-Ray Spectrom,* **36**: 132-135.

Karen, G., Alline, S. L., Cintia, A. S., Mariana, G. S., Vivianne, P. S., Jose, C. V., and MariaIzabel, M. S. B. (2006). Characterization of Portland cement by X-Ray spectrometry allied to Chemometrics, *J. Chemometrics,* **20**: 455-463.

Kessler, T., Hoffman, 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,* **3**1: 383-390.

Klaunig, J., Kamendulis, M., and Hocevar, B. (2010. Oxidative Stress and Oxidative Damage in Carcinogenesis, *Toxicologic Pathology,* **38**: 96-109.

Kowalik, C., and Einax J. W. (2006). Modern chemometric data analysis-methods for the objective evaluation of load in river systems. *Acta hydrochim, hydrobiol,* **34**: 425-436.

Kump P., Ljublyano, Stroenia (1993). Quantitative analysis of environmental samples by X-ray spectrometry, *Applied Spectroscopy*, **143**: 3-34.

Kurt, V. and Peter, F. (2008). Introduction to multivariate statistical Analysis in chemometrics, CRC press, New York.

Kvestoslav, R.S. (1999). *Analytical chemistry of Aerosols*. CRC press LLC. Boca Raton, Florida, USA.

Kwiatek, W. M., Banas, A., Banas, K., Gajda, M., Gałka, M., Falkenberg, G. and Cichocki, T. (2005). Iron and other elements studies in cancerous and non-cancerous prostate tissues. *Journal of Alloys and Compounds* **40,** 178–183.

Liu, Y., Upadhaya, B. R., and Masoud, N. (1993). Chemometric data analysis using artificial neural networks, *J.App.spetroscopy*, **47**: 12-23.

Lachance, G. R (1999). Demystification of algorithms and influence coefficients in quantitative XRF analysis, *JCPDS-International Centre for Diffraction Data,* 718-731

Lee, R., Woo, W., Wu, W. B., Kummer, A., Duminy, H., and Xu, Z. (2003). Zinc accumulation in Nmethyl-N-nitrosourea-induced rat mammary tumors is accompanied by an altered expression of ZnT-1 and metallothionein, *Exp Biol Med,* **228**: 689-696.

Leitao, R. G., Palumbo A. J., Souza, P. A. V. R., Pereira , G. R C., Canellas , G. L., Anjos , M. J., Nasciutti, L.E., and Lopes, R. T. (2014). Elemental concentration analysis in prostate tissues using total reflection X-ray fluorescence, *Radiation Physics and Chemistry*, **95**: 62-64.

Lemberge, P. (2000). Quantitative X-Ray fluorescence analysis using partial least squares and Monte Carlo simulations, PhD Thesis, Universiteit Antwerpen.

Lemberg, P., and Van Espen, P. (1999). Quantitative EDXRF analysis of liquids using partial least squares, *X-Ray Spectrom,* **28**: 77-55.

Li, J., Parchatka, U., and Fischer, H. (2012). Applications of wavelet transform to quantum cascade laser spectrometer for atmospheric trace gas measurements, *Appl. Phys. B,* **108**: 951–963.

Long, X., Huang, N., and Peng, X. (1998). An artificial neural networks analysis of low-resolution X-ray fluorescence spectra. *Adv. X-ray anal,* **40**: 307-414.

Ma, C. and Shao, X. (2004). Continuous wavelet transform applied to removing the fluctuating background in near-infrared spectra, *J. Chem. Inf. Comput. Sci*, **44**: 907–911.

Magdalena, S. B. (2012). Sample thickness considerations for quantitative X-ray fluorescence analysis of the soft and skeletal tissues of the human body – theoretical evaluation and experimental validation. X-Ray Spectrom, **41**: 328–337.

Maria, I. M. S., Martha, T. P. O., Aline, M. S., Erica, B. S. and Alete P. (2005). X-ray scattering processes and chemometrics for differentiating complex samples using

conventional EDXRF equipment, *Chemometrics and Intelligent Laboratory Systems,* **78**: 96–102.

Markowicz, A. (2011). An overview of quantificationmethods in energy-dispersive X-ray fluorescence analysis. *PRAMANA-journal of Physics,* **76(2)**: 2321-329.

Marini, F., Bucci, R., Magri, A. L., and Magri, A. D. (2007). Artificial neural networks in chemometrics: History, examples and perspectives, *Microchemical,* **88(2)**:178-185.

Miguel, M., and Lucia, M. (2011). Wavelet multivariate-denoising scheme for detection of fractures in concrete. 5th Pan American Conference for NDT, Cancun, Mexico.

Molt, K., and Schramn, R. (1999). Determination of light elements in organic liquid matrices by PCR in EDXRF using backscattered radiation, *X-ray Spectrom*, **28**: 59-63.

Mona., A., Liwaa., H. And Ayad., M.J. (2013). Determination some minerals in cancer tissue, *Journal of Biology and Medical Sciences – JBMS,* **1**: 13-16

Nielson, K. K. (1977). Matrix corrections for energy dispersive X-ray fluorescence analysis of environmental samples with coherent/incoherent scattered X-rays. *Anal. Chem.* **49(4):** 641-648.

Okonda, J. (2015). Development of chemometrics aided energy dispersive X-ray fluorescence and scattering (EDXRFS) method for rapid diagnostics of cancer, MSc thesis, University of Nairobi.

Potts, P. J and Webb, P. C. (1992), X-ray fluorescence spectrometry, *J. of Geochemical Exploration,* **44**: 251-296.

Ramadan, Z., Hopke, P. K., Johnson, M. J. and Scow, K. M. (2005). Applications of PLS and back-propagation neural networks for the estimation of soil properties, *Chemom.Intell.Lab.syst,* **75(1)**: 23-30.

Rindby, A. (1989). Software for energy-dispersive X-ray fluorescence, *X-ray Spectrometry*, **18**:113–118.

Rousseau, R. M. (2006). Corrections for matrix effects in X-ray fluorescence analysis, A tutorial. *Spectrochim. Acta. Chem,* **30:** 1306-1309.

Rousseau, R. M. (1984). Fundamental algorithm between concentration and intensity in XRF analysis – theory and Practical application, *X-Ray Spectrom,* **13**: 115-125.

Rousseau, R. M. (2009), The Quest for a Fundamental Algorithm in X-Ray Fluorescence Analysis and Calibration, *The Open Spectroscopy Journal*, **3**: 31-42.

Rudolf, O. M. (1972). *Spectrochemical Analysis by X-ray Fluorescence,* Plenum Publishing, Corporation New York, New York.

Sergey, A., Xiang, L. and John, F. (2011). Rapid limit tests for metal impurities in pharmaceutical materials by X-ray fluorescence spectroscopy using wavelet transform filtering, *Anal. Chem,* **83**: 1061–1068.

Schimidt, F., Ponce, L. C., Bueno, M. I. M. S., Poppi, R. J. (2003). Determination of rare earth elements by energy dispersive X-ray fluorescence and artificial neural networks, *X-Ray Spectrom.* **32**: 423-427.

ShengXiang Bao, (1999). Combination of corrections for absorption, overlap and background in XRF spectrometry, *X-ray spectrom,* **28(3)**: 141-144.

Sherman, J. (1955). The theoretical derivation of fluorescent X-ray intensities from mixtures, *Spectrochin. Acta*, **7**:283-306.

Su, J. M. F., Perlaky, L., Li, X. N., Leung, H. C. E., Antalffy, B., Armstrong, D. and Lau, C. C. (2004). Comparison of ethanol versus formalin fixation on preservation of histology and RNA in laser capture micro dissected brain tissues, *Brain Pathol*, **14**: 175–18.

Sussulini, A., Lima, A. G., Figueiredo, E. C., Fenandes, H. L., Pinheiro, S. C. L., Bueno, M. I. M. and Pereira, F. M. V. (2009). X-ray scattering information for powdered fruit juice mixes, *X-ray spectrom***, 38**:254.

Torres, E. L., Fuentes, M. V., Greaves, E. D. (1998). SAX, software for the analysis of X-ray fluorescence spectra, *X-ray Spectrometry,* **27**: 161–165.

Uzay, E., Ceyhun B., Ata, A., Ayşın, E., Bülent, S. and Kerem, H. (2003). Wavelet denoising vs ICA denoising for functional optical imaging. Neural Engineering. Conference Proceedings. First International IEEE EMBS Conference

Van Espen, P., Nullens, H. and Adams, F. (1977). A computer analysis of X-ray fluorescence spectra. *Nucl.Instrum.Methods*, **142**: 243-250.

Vekemans, B., Janssens, K., Vincze, L., Adams, F., Van Espen, P. (1994). Analysis of X-ray spectra by iterative least squares (AXIL): new Developments, *X-Ray Spectrom,* **23**: 278-285.

Verbi, F. M., Pereira-Filho, E. R., Bueno, M.I.M. (2005). Use of X-ray scattering for studies with organic compounds: a case study using paints, *Microchim.Acta,* **150**: 131-136.

Vincze, L., Janssens, Vekemans, B., Adams, F. (1999). Monte Carlo simulation of X-ray fluorescence spectra: Photon scattering high energy X-ray energies. *Spectrochim. Acta Part* B, **54(12):** 1711-1722.

Virendra, S., Anrawal, H., Joshi, C., Sudersha, M. and Sinha, K. (2011). Elements profiles of agricultural soil by EDXRF technique and use of the principal component analysis 113 (PCA) method to interpret the complex data, *Applied Radiation and Isotopes,* **36**: 181-194.

Werbos, P., (1974). Beyond regression, new tools for prediction and analysis in the behavioral sciences, PhD thesis, Harvard University, Cambridge, MA.

Wobrauscheck, P. and Christina, S., (2010). Energy dispersive X-ray fluorescence analysis, *X-ray spectrometry*, John Wiley & Sons Ltd, 1-17.

Zhang, C., and Zhang, F. (2015). Iron homeostasis and tumorigenesis: Molecular mechanisms and therapeutic opportunities, *Protein cell*, **6(2)**: 88-100.

Zhao, B., Liu, Y., and Xia, S. W. (2000). Support vector machines and its application in handwritten numerical recognition, *In Proceedings of 15th Int. Conference on Pattern Recognition*, **2**: 720-723.

Zhi, L., Ahmed, A., Bing-Yi, J., and Xin, Ga. (2012). WaVPeak: picking NMR peaks through wavelet-based smoothing and volume-based filtering, **8(7)**: 914–920.

APPENDICES

Appendix 1: Wavelet transform Training Algorithm Used in this Work

```
% load a multivariate signal
>> load XC; % input spectral data
>> level=5;
>> wname='sym4';
>> tptr='rigrsure';
>> sorh='s';
%set PCA parameters by retaining all the principal component analysis
>> npc_app=4;
>> npc_fin=4;
>> npc_fin=4;
>> npc_fin=4;
%Finally perform multivariate denoising
>> XC_den=wmulden(XC,level,wname,npc_app,npc_fin,tptr,sorh);
%Improve the first result by retaining fewer principal components
[XC_den,npc,nestco]=wmulden(XC,level,wname,npc_app,npc_fin,tptr,sorh
);
%  Use kais rule
>> npc_app='kais';
>> npc_fin='kais';
>>
[XC_den,npc,nestco]=wmulden(XC,level,wname,npc_app,npc_fin,tptr,sorh
);
% computing signal to noise before wt analysis
>>r=snr(XC(:)); % db
>> rr=10.^(r/10); % from db to linear;
% computing signal to noise ratio after wt analysis
>> R=snr(XC_den(:)); % db;
>> RR=10.^(R/10); % from db to linear
```

Appendix 2: ICA spectral preprocessing Training Algorithm Used in this Work

```
>> m = mean(XC)';
>> Xc = XC'-repmat(m,1,size(XC,1)); % Center and transpose
>> R = cov(Xc'); % Whitening
>> [V,D] = eig(R);
>> WhiteT = V*diag(diag(D).^-0.5)*V'; % Whitening transform
>> Xw = WhiteT*Xc;
>> Ww = fastica(Xw); % Finally perform fastica
>> W = Ww'*WhiteT; % Reconstruct
```

Appendix 3: ANNs Training Algorithm Used in this Work

```
net=fitnet(7,'trainlm'); % Create a feedforward network with 7 hidden neurons, 3 output neurons and assign the Levenberg-Marquardt training function - TRAINLM
net.divideParam.trainRatio=.7;% divide samples into training
net.divideParam.valRatio=.15; ;% divide samples into validation
net.divideParam.testRatio=.15; ;% divide samples into test
[net,pr]=train(net,input,target); % Train the network.
net=init(net);% initialize the net if it is not satisfactory and
              train again
[net,pr]=train(net,input,target);
output=net(sample); % Simulate the trained network.
```

Appendix 4: PLS Algorithm used in this study

```
% load data containing spectral intensities
X = input;% input spectral data
y = target; % target concentration
Z=Test data; % test spectral data
% choosing number of components with cross-validation
>>[Xl,Yl,Xs,Ys,beta,pctVar,PLSmsep] = plsregress(X,y,17,'CV',10); % Transpose X and y to have same number of rows
% Plot MSEP curve for PLSR to determine the number of PLS compnents
>> plot(0:17,PLSmsep(2,:),'b-o');
```

```
xlabel('Number of components');

ylabel('Estimated Mean Squared Prediction Error');

legend({'PLSR'},'location','NE');

>> [XL,yl,XS,YS,beta,PCTVAR,MSE] = plsregress(X,y,8); % Transpose X and y to have same number of rows

% Plot the percent of variance explained in the response variable as a function of the number of components.

>>plot(1:8,cumsum(100*PCTVAR(2,:)),'-bo');

xlabel('Number of PLS components');

ylabel('Percent Variance Explained in y');

>> yfit = [ones(size(Z,1),1) Z]*beta;

% Plot fitted vs. observed response for the PLSR  fit.

>> plot(y,yfit,'o');xlabel('Observed Response');

ylabel('Fitted Response');

legend({'PLSR with 8 Components'},'location','NW');
```

Appendix 5: ICA classification Training Algorithm Used in this Work

```
% Load (meas) spectral data

>>Species=[zeros(15,1);ones(15,1)];

 [u,s,v] =  svd(meas,'econ');

     t=u*s;

>> ICs=3;

>>  Options.Method ='Normal';

     Options.Data= 'Data';

[Scores] =  jadeR_2005 (meas,ICs);  % ensure that the current path is JadeR_2005

>>figure;

     imagesc(Scores(:,1:2)),axis tight;

     xlabel('ICs');

      title('ICA');

>>figure;

     plot(Scores(:,1:2)),axis tight;

     xlabel('Samples');
```

```
     title('ICA');
>>[lambda_ICA,F_ICA,pr_ICA,df,Fdf]                                    =
manova_Strauss(Scores(:,1:2),Species,0); % Transpose scores first for
inner matrix dimension to match
>>figure;
scatter(Scores(:,1),Scores(:,2),200,Species,'Filled'),axis tight;
xlabel('IC1');
ylabel('IC2');
title(['ICA (F for Wilks lambda = ',num2str(F_ICA),')']);
% Load meas
>>Species=[zeros(10,1);ones(10,1); 2*ones(10,1)];
 [u,s,v] =  svd(meas,'econ');
     t=u*s;
>> ICs=3;
>>   Options.Method ='Normal';
     Options.Data= 'Data';
[Scores] =  jadeR_2005 (meas,ICs);  % ensure that the current path is
JadeR_2005
>>figure;
     imagesc(Scores(:,1:2)),axis tight;
     xlabel('ICs');
      title('ICA');
     figure;
     plot(Scores(:,1:2)),axis tight;
     xlabel('Samples');
     title('ICA');
>>[lambda_ICA,F_ICA,pr_ICA,df,Fdf]                                    =
manova_Strauss(Scores(:,1:2),Species,0); % Transpose scores first for
inner matrix dimension to match
>>figure;
scatter(Scores(:,1),Scores(:,2),80,Species,'Filled'),axis tight;
xlabel('IC1');
ylabel('IC2');
```

```
title(['ICA (F for Wilks lambda = ',num2str(F_ICA),')']);
```

Appendix 6: PCA classification Training Algorithm Used in this Work

```
% replacing the 1s and 0s in species with desired phrases
spec = cell(size(species));
keys = {'Lower Speciation','Higher Speciation'};
for n = 1:length(species)
    if species(n) == 0
        spec{n} = keys{1};
    else
        spec{n} = keys{2};
    end
end
species = spec;
X = meas;
X = bsxfun(@minus, X, mean(X));    % zero-centered data
[~,S,V] = svd(X,0);                % singular value decomposition

[S,ord] = sort(diag(S), 'descend');
pc = V(:,ord);                     % principle components
latent = S.^2 ./ (size(X,1)-1)     % variance explained
score = X*pc;                      % projected data
gscatter(score(:,1), score(:,2), species, [], [], [], 'on', 'PC1',
'PC2')
```