UNIVERSITE D'ANTANANARIVO

FACULTE DES SCIENCES

FORMATION DOCTORALE EN PHYSIQUE

DEPARTEMENT DE PHYSIQUE

Laboratoire de Physique Nucléaire et de Physique de l'Environnement

THESE

pour l'obtention du diplôme de :

DOCTORAT DE PHYSIQUE

Option : ***Physique Nucléaire, Physique Théorique et Physique Appliquée***

sur l'

UTILISATION DE LA REGRESSION PLS EN SPECTROSCOPIE XRF : DEVELOPPEMENT D'UN LOGICIEL D'ANALYSE QUANTITATIVE ET APPLICATIONS A LA SPECTROMETRIE XRF A REFLEXION TOTALE

présentée par

RAKOTONDRAJOA Andrianiaina

devant la commission d'examen composée de :

Président :	**M. RAJAOBELISON Joël**	Professeur
Rapporteurs :	**M. RAVELOMANANTSOA Solofonirina Dieudonné**	Professeur
	M. RASTEFANO Elisée	Professeur
Examinateur :	**M. ANDRIANARY Philippe**	Professeur
Directeurs de thèse :	**M. RABOANARY Roland**	Professeur
	M. RAOELINA ANDRIAMBOLOLONA	Professeur titulaire de classe exceptionnelle

le 31 Mai 2008

instn Madagascar

UNIVERSITE D'ANTANANARIVO

FACULTE DES SCIENCES

FORMATION DOCTORALE EN PHYSIQUE

DEPARTEMENT DE PHYSIQUE

Laboratoire de Physique Nucléaire et de Physique de l'Environnement

THESE

pour l'obtention du diplôme de :

DOCTORAT DE PHYSIQUE

Option : ***Physique Nucléaire, Physique Théorique et Physique Appliquée***

sur l'

UTILISATION DE LA REGRESSION PLS EN SPECTROSCOPIE XRF : DEVELOPPEMENT D'UN LOGICIEL D'ANALYSE QUANTITATIVE ET APPLICATIONS A LA SPECTROMETRIE XRF A REFLEXION TOTALE.

présentée par

RAKOTONDRAJOA Andrianiaina

devant la commission d'examen composée de :

Président :	**RAJAOBELISON Joël**	Professeur
Rapporteurs :	**M. RAVELOMANANTSOA Solofonirina Dieudonné**	Professeur
	M. RASTEFANO Elisée	Professeur
Examinateur :	**M. ANDRIANARY Philippe**	Professeur
Directeurs de thèse :	**M. RABOANARY Roland**	Professeur
	M. RAOELINA ANDRIAMBOLOLONA	Professeur titulaire de classe exceptionnelle

le 31 Mai 2008

Ho an'i Dada sy Mama :

Hery lehibe ho ahy ny fahatokisanareo tsy niala tamiko !

Remerciements

Tout d'abord, j'aimerais remercier Notre Seigneur Dieu qui m'a donné la force et le courage pour aller au bout de ce que j'ai commencé.

Mes remerciements les plus chaleureux vont à l'endroit de Monsieur RAOELINA ANDRIAMBOLOLONA, Professeur Titulaire de classe exceptionnelle à la Faculté des Sciences de l'Université d'Antananarivo qui est le co-directeur de cette thèse. Je lui suis très reconnaissant de m'avoir donné ses précieux conseils pour le contenu de ce travail.

Ma profonde gratitude s'adresse aussi à Monsieur RABOANARY Roland, Professeur et Responsable de l'Option Physique Nucléaire, Physique Appliquée et Physique Théorique à la Faculté des Sciences de l'Université d'Antananarivo, qui m'a aussi encadré durant la réalisation de cette thèse. Je lui remercie de m'avoir accordé beaucoup de son temps pour m'aider à mener à bien ce travail malgré ses occupations multiples.

Je remercie M. RAJAOBELISON Joël, Professeur à la Faculté des Sciences de l'Université d'Antananarivo, qui a bien voulu accepter de présider le Jury de cette thèse et d'avoir sacrifié son temps pour examiner ce travail malgré ses différentes tâches.

Je présente aussi mes sincères remerciements à Monsieur RAVELOMANANTSOA Solofonirina Dieudonné, Professeur à la Faculté des Sciences de l'Université d'Antananarivo d'avoir bien voulu être mon rapporteur. Ses critiques constructifs m'ont été d'une grande aide pour l'amélioration de ce travail.

Que Monsieur RASTEFANO Elisée, Professeur à l'Ecole Supérieur Polytechnique de l'Université d'Antananarivo, trouve ici l'expression de mes reconnaissances les plus sincères, pour avoir accepté d'être le rapporteur de cette thèse et de m'avoir consacré beaucoup de temps malgré ses obligations.

J'adresse aussi ma sincère gratitude à Monsieur ANDRIANARY Philippe, Professeur et Chef du Département « Génie Chimique » à l'Ecole Supérieur Polytechnique de l'Université d'Antananarivo. Je lui suis reconnaissant d'avoir accepté d'examiner ce travail et d'avoir apporté sa contribution pour l'amélioration de cette thèse.

Le présent travail a été effectué en totalité au sein de l'Institut National des Sciences et Techniques Nucléaires (Madagascar-INSTN), je tiens à présenter mes sincères remerciements à son Fondateur et Directeur Général, le Pr. RAOELINA ANDRIAMBOLOLONA ainsi qu'à tout le personnel technique et administratif qui ont facilité l'accomplissement de mon travail. Mes reconnaissances vont plus particulièrement à tous les membres du département « Techniques de la Fluorescence X et Environnement » : Mme ANDRIAMANIVO Lucienne et son équipe, qui m'ont beaucoup aidé lors des différentes mesures et qui m'ont toujours donné leurs précieux conseils en matière de spectroscopie XRF ainsi qu'à toute l'équipe du département de « Maintenance et Instrumentation ».

Je ne saurais oublier d'adresser mes chaleureux remerciements à toute ma famille et à mes amis qui m'ont toujours supporté moralement et matériellement.

Enfin, à tous ceux qui, de près ou de loin, m'ont aidé à la réalisation de ce travail.

Merci !

Table des matières

Liste des tableaux

Liste des figures

Liste des annexes

Liste des abréviations

AIEA	: Agence Internationale de l'Energie Atomique
DSP	: Digital Signal Processing (Traitement numérique du signal)
ED-XRF	: Energy Dispersive X-Ray Fluorescence (Fluorescence X à dispersion d'Energie)
FWHM	: Full Width at Half Maximum
HPGe	: High Purity Germanium
LOO-CV	: Leave One Out Cross Validation (Validation Croisée Leave One Out)
MCA	: MultiChannel Analyzer (analyseur multicanal)
MCI	: Moindres Carrées Inverses
MCO	: Moindres Carrées Ordinaires
NILES	: Nonlinear estimation by Iterative Least Squares
NIPALS	: Nonlinear estimation by Iterative Partial Least Squares
PCR	: Principal Components Regression (Régression par Composantes Principales)
PF	: (méthode des) Paramètres Fondamentaux
PLS	: Partial Least Squares
PRESS	: PRediction Error Sum of Squares
RMSEC	: Root Mean Square Error of Calibration
RMSEP	: Root Mean Square Error of Prediction
R-X	: Rayons X
SDD	: Silicon Drift Detector
SIMPLS	: Straightforward Implementation of a statistically inspired modification of the PLS method
TXRF	: Total Reflection X Ray Fluorescence
XRF	: X Ray Fluorescence

Introduction

L'Institut National des Sciences et Techniques Nucléaires (Madagascar-INSTN) s'est doté du laboratoire des Techniques de la Fluorescence X en 1985. Depuis ce temps, les techniques utilisées sont restées les mêmes pour les analyses qualitative et quantitative. Ces techniques sont principalement le dépouillement des spectres à l'aide du logiciel AXIL[1] et l'application d'un modèle quantitatif pour la conversion des aires nettes en concentrations des éléments. La maîtrise de ces techniques demande beaucoup d'expériences après l'apprentissage basé sur les connaissances théoriques. Ceci entraîne que ces analyses, bien que faisant partie des travaux de routine, nécessitent un utilisateur expérimenté et qualifié. En outre, ces techniques demandent beaucoup d'interventions de la part de l'utilisateur, donc beaucoup de temps, pour passer du spectre aux concentrations des éléments à étudier. La véracité des résultats obtenus lors de ces analyses dépend ainsi en grande partie du savoir-faire de l'utilisateur.

Actuellement, la technique de la fluorescence X à dispersion d'énergie (ED-XRF) n'a pas cessé d'évoluer. Un développement considérable dans le domaine de l'électronique, des améliorations importantes pour les détecteurs ainsi que les tubes à rayons X ont abouti à des spectromètres très performants mais aussi très pratiques du point de vue utilisation. On peut citer comme exemples les spectromètres portatifs pour les mesures sur terrain mais aussi des spectromètres de laboratoire utilisant la technologie DSP (traitement numérique des signaux).

Le domaine d'application de la technique de la fluorescence X a aussi connu un élargissement important. En effet, outre les applications habituelles en environnement (eau, sol, air, plantes,..) et dans le domaine alimentaire, cette technique est aussi utilisée avec beaucoup de succès en archéologie, en industrie du ciment et des peintures,…

La plupart de ces applications et de ces nouvelles générations d'appareils nécessitent que les données obtenues lors des différentes mesures soient traitées le plus rapidement possible. La solution idéale serait alors de pouvoir convertir les spectres obtenus directement en concentrations des éléments sans intervention de la part de l'utilisateur.

[1] *Analysis of X-ray spectra by Iterative Least squares*

Cette solution aurait en effet comme avantages : le gain de temps lors des analyses mais aussi la facilité d'utilisation de la technique.

Le but du présent travail, qui a été entièrement réalisé au sein de Madagascar-INSTN, est d'apporter une méthode alternative permettant de passer directement du spectre aux concentrations sans intervention de l'utilisateur et de concevoir un logiciel pour la mise en œuvre de cette méthode pour l'analyse quantitative en ED-XRF.

Cette méthode qui s'appelle PLS (Partial Least Squares ou moindres carrés partiels) fût inventé par Herman Wold en 1975. Elle était alors utilisée en économétrie. Elle a été par la suite (dans les années 90) utilisée avec succès en spectroscopie proche infrarouge [3] et ultraviolet. Des résultats satisfaisants ont aussi été rapportés dans l'application de cette méthode en spectrophotométrie [24]. Ses applications en ED-XRF ne sont qu'à leurs débuts actuellement.

Dans le premier chapitre, quelques notions théoriques sur la technique de l'ED-XRF ainsi que les techniques classiques utilisées pour l'analyse quantitative sont exposées. Nous parlons aussi dans ce chapitre des développements récents en matière d'instrumentation en spectrométrie XRF.

Le deuxième chapitre est quant à lui consacré entièrement à la régression PLS. Les autres méthodes classiques de régression seront tout d'abord étudiées. C'est après ces études que le PLS sortira comme la solution la plus pertinente pour l'application en spectroscopie XRF. Les bases théoriques de la méthode PLS, les différents algorithmes pour son application pratique ainsi que les méthodes de détermination des incertitudes sur les prédictions sont traités en détail au cours de ce chapitre.

Nous donnons ensuite dans le chapitre 3 un guide pratique pour l'application de la méthode PLS en spectrométrie XRF.

Le chapitre 4 parle du logiciel X-PLS v1.0 que nous avons développé pour l'analyse quantitative en XRF. Les différents modules qui constituent ce logiciel sont détaillés au cours de ce chapitre. Les codes source des modules majeurs de ce logiciel sont présentés en Annexes.

Dans le 5è chapitre, la méthode de régression PLS est appliquée à l'analyse quantitative de quelques échantillons liquides, en utilisant le logiciel X-PLS v1.0. La technique XRF à réflexion totale y est utilisée à ce propos.

Enfin, outre les conclusions générales, quelques perspectives pour la continuation du présent travail ainsi que des propositions d'applications dans d'autres domaines sont émises dans le dernier chapitre.

Chapitre 1

La spectroscopie par fluorescence X

1.1. GENERALITES SUR LA SPECTROSCOPIE XRF :

Quand un atome est mis sous l'action d'une radiation incidente d'énergie suffisante, un négaton de sa couche interne est éjecté. L'atome se trouve alors dans un état instable dit excité. Pour revenir à l'état stable, un négaton d'une couche plus externe vient combler le trou laissé par le négaton éjecté. Ce phénomène s'accompagne de l'émission d'un rayonnement X qui est caractéristique de l'élément en question. Ces raies sont appelées raies de fluorescence X (XRF). L'analyse par fluorescence X ou la spectroscopie XRF est l'ensemble des mesures effectuées sur ces raies XRF.

La radiation incidente peut provenir de différents types de sources comme un radioisotope, un tube générateur de rayons X par exemple. L'analyse par fluorescence X peut être qualitative ou quantitative. Dans le premier cas, le but est d'identifier les éléments présents dans l'échantillon à analyser. L'analyse quantitative vise quant à elle à déterminer la quantité d'un élément quelconque dans l'échantillon en question.

Il existe deux méthodes majeures pour la spectroscopie XRF. La première est la spectroscopie XRF à dispersion de longueur d'onde (WD-XRF : Wavelength Dispersive X-Ray Fluorescence Analysis) qui consiste à déterminer la longueur d'onde des raies X émises. La deuxième méthode appelée spectroscopie XRF à dispersion d'énergie (ED-XRF : Energy Dispersive X-Ray Fluorescence Analysis) est quant à elle basée sur la mesure directe de l'énergie de ces raies. Nous n'allons nous intéresser qu'à cette dernière méthode tout au long de ce travail. C'est en effet la méthode qui est utilisée au sein de Madagascar-INSTN au sein duquel ce travail a été réalisé.

1.1.1. Instrumentation :

L'appareillage permettant d'effectuer l'analyse XRF, ou le spectromètre XRF, est basé sur le détecteur. C'est en effet cet élément qui interagit directement avec les raies X et permet de mesurer leurs énergies. La figure suivante montre la configuration typique d'un spectromètre ED-XRF.

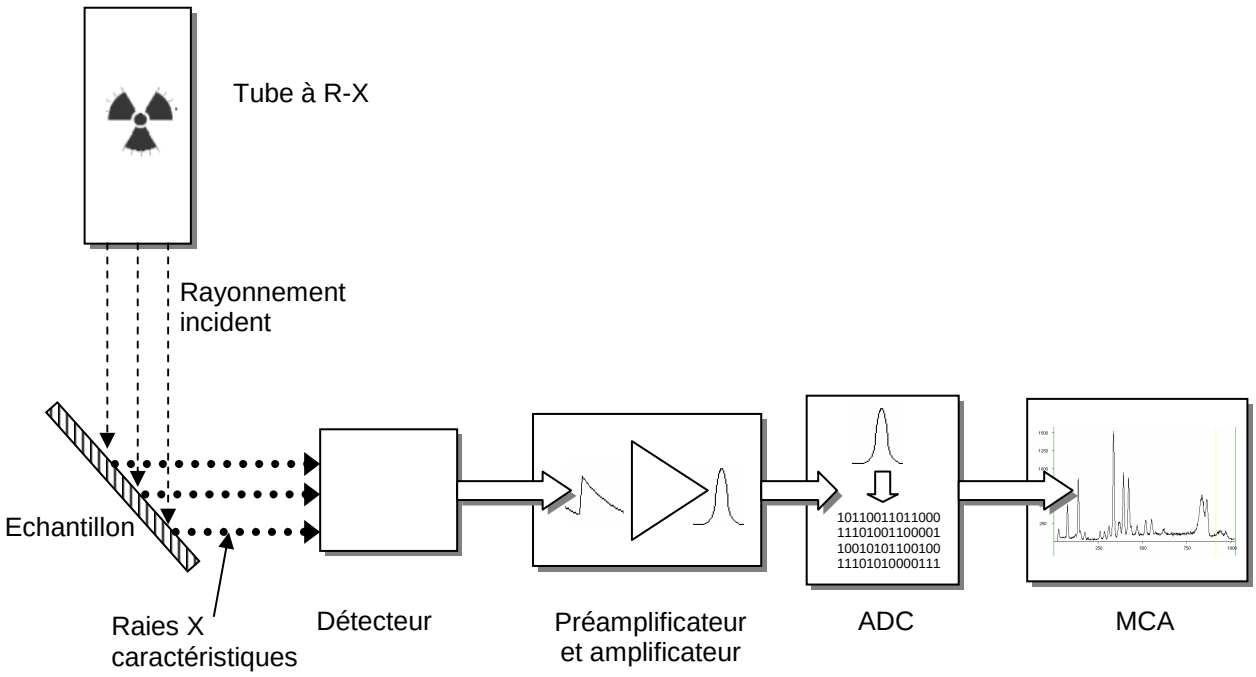


Figure 1.1 : Chaîne de spectrométrie ED-XRF typique

Une raie X produit des ionisations dans le détecteur qui convertit ces dernières en signaux électriques dont l'amplitude est proportionnelle à l'énergie de la raie X. Le préamplificateur et l'amplificateur amplifient et mettent en forme ces signaux pour qu'ils puissent être traités par le convertisseur analogique – digital (ADC). L'ADC convertit à son tours ces signaux analogiques au format numérique. C'est ce signal numérique qui est stocké dans l'analyseur multicanal (MCA). Ce dernier a en fait le rôle de trier tous les signaux et de les stocker dans différentes mémoires ou canaux. Un canal correspond alors à une énergie de raie X bien déterminée. La figure 1.2 représente un spectre obtenu par un spectromètre ED-XRF utilisant un générateur à rayons X (R-X). L'abscisse représente le numéro de canal. L'ordonnée donne le nombre de photons X de même énergie détectés par le détecteur. Pour un élément donné (représenté par un ou plusieurs pics sur le spectre), le

nombre de coups ou l'intensité (hauteur du pic) est relatif à la proportion de cet élément dans l'échantillon. L'analyse quantitative en spectroscopie XRF consiste à déterminer la concentration d'un élément donné par la mesure de l'intensité d'un pic ou de son aire nette tout en prenant en compte de plusieurs autres facteurs comme le bruit de fond, les chevauchements de pics,… En d'autres termes, l'analyse quantitative en spectroscopie XRF est la recherche d'une relation entre le spectre et la concentration d'un élément d'intérêt.

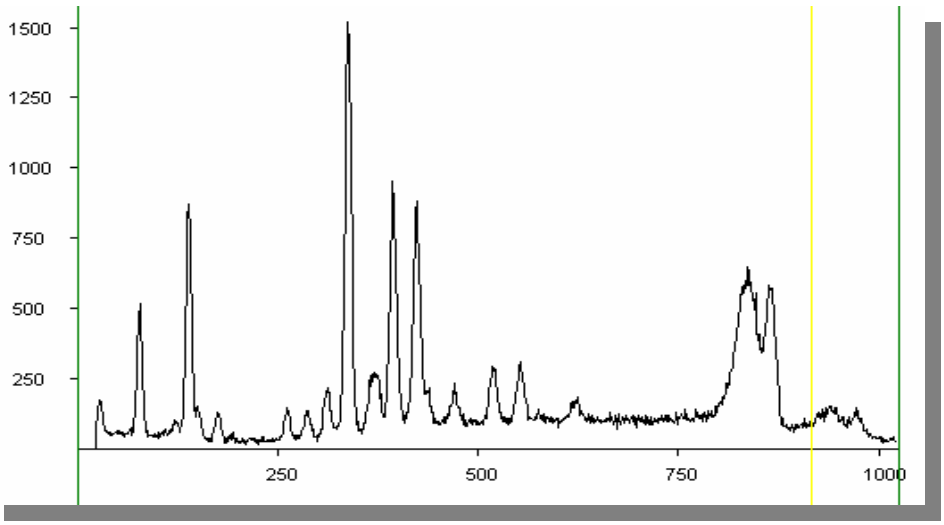


Figure 1.2 : Exemple de spectre XRF

1.1.2. Principes de la Fluorescence X à réflexion totale (TXRF) :

La TXRF est une variante de l'ED-XRF. Elle se distingue de la méthode conventionnelle par la manière dont le rayonnement primaire incident interagit avec l'échantillon. Elle est en d'autres termes une configuration géométrique particulière en ED-XRF. Cette configuration est caractérisée par un rayonnement primaire arrivant sous un angle d'incidence rasante sur l'échantillon qui est sous forme de couche fine déposée sur un porte-échantillon fait avec un matériau à haute réflectivité. Cet angle est plus petit ou proche de l'angle critique pour lequel on a une réflexion totale du rayonnement primaire. Cette configuration a comme avantage d'éviter au maximum la pénétration du rayonnement incident dans l'échantillon. Elle permet alors d'avoir un bruit de fond très réduit, améliorant ainsi la limite de détection. La méthode de la fluorescence X à réflexion totale est de ce fait très appropriée à l'analyse d'éléments en trace.

1.2. TECHNIQUES CLASSIQUES DE QUANTIFICATION EN ED-XRF :

La méthode classique de quantification en ED-XRF se divise en deux grandes étapes :

- le dépouillement du spectre
- la conversion de l'intensité en concentration

1.2.1. Le dépouillement du spectre :

Cette étape a pour but de déterminer l'aire nette des pics caractéristiques de l'élément d'intérêt. Plusieurs méthodes peuvent être utilisées pour cela.

La méthode la plus simple consiste à effectuer une interpolation du bruit de fond sous le pic en question. Il suffit ensuite d'additionner les contenus des canaux sous ce pic et de soustraire le bruit de fond pour avoir l'aire nette. Il est évident que cette méthode ne peut fonctionner que pour un pic isolé et sans la présence de chevauchement d'autres pics, qui est très souvent le cas en ED-XRF.

Une autre méthode plus performante est la méthode des moindres carrés utilisant des spectres de référence. Cette méthode suppose que le spectre d'un échantillon inconnu peut être modélisé comme une combinaison linéaire de spectres des éléments purs constituant cet échantillon. Cette méthode est dans la plupart des cas valable pour les lignes caractéristiques (pics) mais ne l'est pas pour le bruit de fond. L'application de cette méthode nécessite donc que le bruit de fond soit d'abord enlevé du spectre de l'échantillon inconnu. Le modèle du spectre est alors donné par l'équation suivante pour m éléments constitutifs.

$$\mathbf{y(i)} = \sum_{j=1}^{m} a_j x_{ji} \qquad (1.1)$$

où y(i) est le contenu du canal i

x_{ji} désigne le contenu du canal i pour le j-ème spectre de référence

a_j est le coefficient quantifiant la contribution du j-ème spectre de référence au spectre de l'échantillon inconnu

Les coefficients a_j sont obtenus par moindres carrés.

La méthode la plus robuste et la plus souple, et de ce fait la plus utilisée, pour la détermination des aires nettes est la méthode des moindres carrés utilisant des fonctions analytiques. Dans cette méthode, on modélise les pics caractéristiques et le bruit de fond en même temps. Pour le bruit de fond, on peut utiliser le plus souvent soit un polynôme, soit une fonction exponentielle. Les pics sont quant à eux modélisés par une fonction gaussienne ou une gaussienne modifiée. Comme une telle fonction n'est pas linéaire par rapport à ses coefficients, on ne peut pas utiliser les moindres carrés linéaires pour obtenir les solutions. On utilise plutôt une technique des moindres carrés non linéaires qui est une technique itérative. C'est cette dernière méthode de détermination de l'aire nette qui est implémentée dans le logiciel AXIL et est utilisée au sein de Madagascar-INSTN pour les analyses effectuées au département Techniques de la Fluorescence X et Environnement (TFXE). Bien que pouvant déterminer des petits pics au voisinage des pics importants, l'utilisation de cette méthode nécessite beaucoup d'interventions de l'utilisateur. Son automatisation est ainsi très difficile, voire impossible.

1.2.2. Conversion de l'intensité en concentration [23]:

Après avoir obtenu les aires nettes des pics à l'aide d'un logiciel de dépouillement de spectres, l'étape suivante est la conversion de ces quantités en concentration des éléments d'intérêt. En général, la concentration d'un élément donné dans un échantillon n'est pas une fonction linéaire de la mesure des aires nettes des pics correspondant à cet élément. Les méthodes utilisant la régression linéaire ne sont en effet valables que dans un intervalle limité de concentrations où l'effet de matrice est constant.

Les phénomènes suivants se produisent simultanément quand un échantillon est exposé aux rayons émis par un spectromètre XRF : les rayons primaires émis sont absorbés par l'échantillon. Ces rayons excitent un élément donné qui émet ses raies caractéristiques. Ce phénomène est la fluorescence primaire.

D'autre part, une partie des rayons émis par cet élément peut provenir de l'excitation par les radiations émises par d'autres éléments de l'échantillon. Ce phénomène est appelé fluorescence secondaire.

Il existe encore un autre phénomène plus complexe où la fluorescence primaire d'un élément produit une fluorescence secondaire pour un autre élément qui à son tour excite un

autre élément émettant par la suite ses raies caractéristiques. Ce phénomène est connu sous le nom de fluorescence tertiaire.

L'effet de ces phénomènes d'absorption et d'excitation sur le spectre émis par un échantillon est communément appelé « effet de matrice ».

Ainsi, l'aire nette des pics caractéristiques d'un élément donné est fonction de la concentration de cet élément mais aussi de la concentration des autres éléments de l'échantillon.

Il existe deux types de méthodes pour la conversion de l'aire nette en concentration qui tiennent compte de l'effet de matrice. Le premier type est constitué des méthodes analytiques et le deuxième comprend les méthodes mathématiques ou numériques.

1.2.2.1. Méthodes analytiques :

Dans cette catégorie de méthodes, on essaie d'éliminer ou d'évaluer l'effet de matrice par des techniques analytiques. Les plus utilisées de ces méthodes sont le standard interne et l'addition de standard dont les principes sont les suivants.

a. Standard interne :

Dans cette méthode, on ajoute la même proportion d'un élément donné dans les échantillons standard et les échantillons inconnus. Les échantillons standard servent à définir la relation entre l'intensité (ou l'aire nette) et la concentration : c'est l'étalonnage. On utilise ensuite cette relation pour les échantillons inconnus. L'élément ajouté, appelé standard interne, doit avoir les mêmes propriétés que l'élément à analyser en terme d'absorption. On peut écrire pour un élément d'intérêt i :

$$A_i = M_i C_i \quad (1.2)$$

et pour le standard interne : $A_s = M_s C_s$ (1.3)

où A_i et A_s sont les aires nettes, C_i et C_s les concentrations. Les coefficients M_i et M_s ne sont pas constants mais varient en fonction de l'effet de matrice. En divisant (1.2) par (1.3), on a

$$\frac{A_i}{A_s} = KC_i \qquad (1.4)$$

avec $K = \frac{M_i}{M_s C_s}$

Comme l'élément i et le standard interne ont à peu près les mêmes propriétés, le quotient $\frac{M_i}{M_s}$ ne varie que très peu en fonction du changement de l'effet de matrice et peut être considéré comme une constante. La constante est alors déterminée par régression linéaire.

b. Addition de standard :

Cette méthode consiste à ajouter une quantité connue de l'élément à analyser dans l'échantillon. Elle est utilisée surtout dans le cas où l'élément d'intérêt est à une très basse concentration et qu'on ne dispose pas d'échantillons standard adéquats. Le principe est le suivant : on obtient une augmentation ΔA_i de l'aire nette Ai en ajoutant une quantité ΔC_i de l'élément i. Si on a pour l'élément i :

$A_i = M_i C_i$

après l'addition, on a :

$$A_i + \Delta A_i = M_i(C_i + \Delta C_i) \qquad (1.5)$$

On suppose donc qu'on peut appliquer ce modèle linéaire dans l'intervalle contenant l'addition. On peut tirer de ces équations la concentration C_i. On répète cette procédure avec différentes quantités de l'élément i pour vérifier la linéarité de l'étalonnage.

1.2.2.2. Méthodes mathématiques :

Ces méthodes utilisent des formulations mathématiques pour calculer l'effet de matrice au lieu de l'éliminer ou de l'évaluer comme les méthodes précédentes. On distingue deux méthodes majeures dans ce domaine. Il s'agit de la méthode des paramètres fondamentaux et l'approche basée sur les coefficients d'influence.

a. Méthode des paramètres fondamentaux (PF):

La méthode des paramètres fondamentaux consiste à quantifier les rayonnements émis en fonction des paramètres physiques et instrumentaux intervenant lors de la mesure. Ces paramètres sont entre autres : les coefficients d'absorption massique, les rapports de saut d'absorption, les probabilités d'émission, les énergies des raies caractéristiques,… Ce sont Sherman et Shiraiwa & Fujino [23] qui ont mis au point la première formulation mathématique de ces intensités émises. L'application pratique de cette formulation fût pourtant très difficile car elle comportait des intégrations multiples et nécessitait la connaissance de la distribution spectrale du rayonnement incident.

C'est pour pallier à ce problème que Crin et Birks [23] ont proposé une autre formulation supposant que le rayonnement incident peut être divisé en un nombre fini d'intervalles d'énergies ΔE et remplacer les intégrales par des simples sommations. Cette relation est la suivante :

$$P_i + S_i = G_i C_i \sum_E \frac{\mu_{iE} I_E \Delta E}{\mu_S^{'} + \mu_S^{''}} \left\{ 1 + \sum_j \frac{1}{2} C_j p_{Ej} \mu_{iEj} \frac{\mu_{jE}}{\mu_{iE}} \left[\frac{1}{\mu_S^{'}} \ln\left(1 + \frac{\mu_S^{'}}{\mu_{sEj}} \right) + \frac{1}{\mu_S^{''}} \ln\left(1 + \frac{\mu_S^{''}}{\mu_{sEj}} \right) \right] \right\} \quad (1.6)$$

où

$$\mu_S' = \sum_i C_i \mu_i' \text{ avec } \mu_i' = \mu_{iE} \csc\psi'$$

$$\mu_S'' = \sum_i C_i \mu_i'' \text{ avec } \mu_S'' = \mu_{iE} \csc\psi''$$

P_i est l'intensité du pic pour l'élément i résultant de l'émission primaire, alors que S_i résulte des émissions secondaires provenant des autres éléments j dans l'échantillon.

G_i est une constante de proportionnalité

C_i la concentration de l'élément d'intérêt i

μ_{iE} est le coefficient d'absorption massique de i pour une énergie E

μ'_S désigne le coefficient d'absorption massique effectif de l'échantillon pour l'énergie incidente E

μ''_S est le coefficient d'absorption massique effectif de d'échantillon pour la raie caractéristique d'énergie E_i.

ψ' est l'angle d'incidence et ψ'' l'angle d'émergence.

L'application pratique de la méthode des PF nécessite deux étapes fondamentales: l'étalonnage et le calcul ou l'analyse.

Etalonnage :

Le but de l'étalonnage est l'établissement d'une relation entre les intensités mesurées et les intensités calculées à partir de la formulation mathématique ci-dessus pour des conditions de mesure et d'échantillonnage bien déterminées.

Pour ce faire, on utilise des échantillons standard, c'est-à-dire, des échantillons dont la composition est bien connue. On mesure à l'aide d'un spectromètre et d'un logiciel de dépouillement de spectre les intensités des éléments d'intérêts.

Analyse :

Cette étape consiste à déterminer les concentrations des différents constituants d'un échantillon inconnu. La méthode utilisée est un procédé itératif décrit par les points suivants :

- une première estimation de la constitution de l'échantillon est d'abord établie. Plusieurs méthodes peuvent être utilisées. On peut par exemple donner la même proportion à tous les constituants comme valeurs initiales des concentrations. Mais la méthode la plus utilisée consiste à mesurer les intensités et à en déduire les concentrations à l'aide de la courbe d'étalonnage et de la formulation mathématique des PF.
- Cette première estimation est ensuite utilisée pour calculer théoriquement les intensités qui sont par la suite converties en valeurs d'intensités mesurées à l'aide de la courbe d'étalonnage
- La différence entre ces deux valeurs (mesurées et calculées) permet de donner une nouvelle estimation de la composition de l'échantillon
- Ces étapes sont répétées jusqu'à la convergence.

b. Les coefficients d'influence :

La méthode des coefficients d'influence quantifie l'effet de matrice de chaque élément constituant l'échantillon individuellement. En d'autres termes, si l'élément d'intérêt est i et l'échantillon comprend les éléments j, k, l, …, les effets de matrices de j sur i, de k sur i, de l sur i,… sont calculés individuellement. La concentration de i peut ainsi s'écrire en fonction de l'intensité (ou l'aire nette) de sa ligne caractéristique et des concentrations des autres éléments avec leurs coefficients d'influence respectifs.

$$C_i = R_i[1+\sum_{j\neq i}\alpha_{ij}C_j] \qquad (1.7)$$

où C_i est la concentration de i,

R_i : l'intensité relative de i : intensité mesurée divisée par l'intensité de l'élément pur correspondant mesurée dans les mêmes conditions,

C_j : concentrations des éléments de l'échantillon

α_{ij} : coefficients d'influence indiquant l'effet de matrice de j sur i.

La sommation couvre donc tous les éléments de l'échantillon sauf i.

Cette équation est une formulation générale de la méthode des coefficients d'influence. Il existe en effet plusieurs algorithmes pour calculer ces coefficients. Certains de ces algorithmes comme ceux de Lachance-Trail et De Jongh [23] utilisent les paramètres fondamentaux pour le calcul des coefficients. C'est pour cela qu'ils sont parfois appelés coefficients d'influence théoriques ou fondamentaux. Il existe cependant des algorithmes qui proposent des formulations empiriques de ces coefficients. On peut citer par exemple les algorithmes de Lucas-Tooth et Price ainsi que ceux de Sherman [23].

1.3. DEVELOPPEMENTS RECENTS EN SPECTROSCOPIE XRF :

En ED-XRF, la qualité d'un spectromètre est définie principalement par la résolution en énergie et le taux de comptage. Les développements qu'ont connus les spectromètres les plus récents sont donc les fruits des recherches sur les améliorations de ces caractéristiques. Presque tous les modules constituant le spectromètre ED-XRF ont ainsi connu d'importantes innovations. Avant de détailler ces innovations, nous allons tout d'abord définir ces caractéristiques d'un spectromètre.

a. Résolution en énergie :

Théoriquement, on doit avoir à la sortie d'un spectromètre, pour un élément donné, un pic représenté par une ligne verticale. En effet, l'énergie de fluorescence d'un élément est connue avec précision et ne dépend d'aucun facteur externe. Ex : Mn K_α : 5,932 keV. On observe pourtant au lieu de cette ligne un pic plus ou moins étalé.

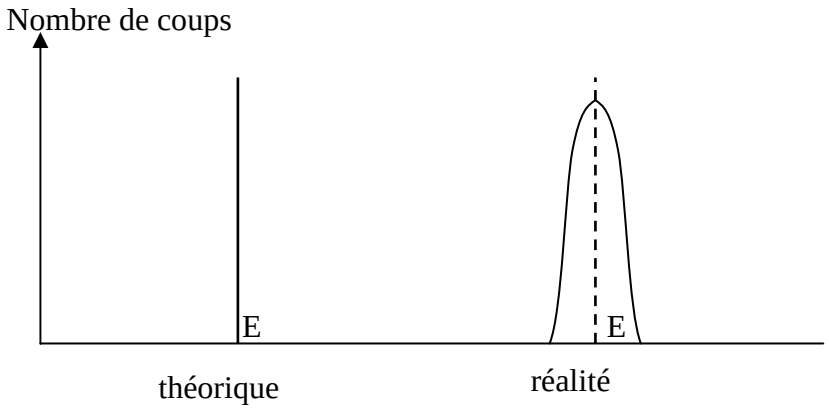


Figure 1.3 : Forme générale de la distribution en énergie de fluorescence X d'un élément

La résolution en énergie d'un spectromètre est définie comme sa capacité à distinguer ou à résoudre les différentes raies X caractéristiques issues d'échantillons contenant plusieurs éléments. Elle est alors spécifiée par la largeur à mi-hauteur (FWHM : Full Width at Half Maximum) d'un pic mesuré à une énergie bien déterminée qui est prise comme énergie de référence pour évaluer la qualité d'un détecteur. La référence la plus usuelle est donnée par la raie Kα du Mn qui est de 5,932 keV. Le FWHM total d'un spectromètre est la

convolution de la contribution du détecteur $FWHM_{det}$ avec celle du module électronique de traitement du signal $FWHM_{elec}$.

$$FWHM_{tot} = \sqrt{FWHM_{det}^2 + FWHM_{elec}^2} \quad (1.8)$$

Pour le détecteur, la résolution en énergie est reliée à la nature statistique de la production de paires négaton-trou par un photon incident. Elle est donnée par l'équation suivante :

$$FWHM_{det} = 2{,}35\sqrt{E\varepsilon F} \quad (1.9)$$

où E est l'énergie du photon incident
ε est la valeur moyenne de l'énergie nécessaire à la production de négaton-trou
et F le facteur de Fano

La contribution de la partie électronique est quant à elle due principalement aux fluctuations causées par effet thermique au sein des différentes composantes électroniques.

b. Débit de comptage :

On entend par débit de comptage (throughput en anglais), la capacité d'un spectromètre à traiter tous les événements se produisant au sein du détecteur. Un spectromètre a en effet des limitations pour le comptage de ces événements. Ces limitations sont dues au temps que met le spectromètre pour le traitement d'un événement (impulsion). Quand l'intervalle de temps entre deux impulsions est trop petit, on observe le phénomène de chevauchement qui donne un résultat erroné après le traitement par l'analyseur multicanal (MCA) du spectromètre. Il faut aussi remarquer que la performance en débit de comptage est reliée directement à la limite de détection. En effet, si toutes les impulsions émises par un élément donné sont comptées, sa limite de détection s'en trouve améliorée.

1.3.1. Détecteurs à semi-conducteur :

Les détecteurs à semi-conducteur ont fait leur apparition au début des années 70 avec l'arrivée du détecteur au Silicium dopé au Lithium (Si(Li)). Actuellement la plupart des spectromètres ED-XRF sont équipés de détecteurs à semi-conducteur qui sont typiquement le Si(Li) ou le Germanium pur HPGe. Pour pouvoir fonctionner, ces détecteurs ont besoin du refroidissement par azote liquide. Ce refroidissement représente pourtant leur plus

grand inconvénient. En effet, le récipient contenant l'azote liquide est assez encombrant et lourd car sa contenance est de 30 à 50 litres. Ceci rend la maniabilité de ces détecteurs très restreinte. Il existe déjà des récipients plus petits (aux environs de 5 à 10 litres) mais ces détecteurs sont encore difficiles à manier. Il faut rappeler que le rôle du refroidissement des détecteurs à semi-conducteur est de diminuer au maximum les courants de fuite dus aux effets thermiques qui détérioreraient la résolution en énergie.

C'est pour pallier à ces problèmes d'exploitation (encombrement, maniement, coût) que d'autres types de détecteurs à semi-conducteur ont été expérimentés. Ces détecteurs n'utilisent plus l'azote liquide comme refroidisseur mais sont dotés de système de refroidissement thermoélectrique par effet Peltier. Ils peuvent ainsi travailler à la température ambiante sans besoin du gros récipient pour azote liquide. Plusieurs matériaux composés ont été étudiés, comme le GaAs, CdTe, HgI_2. Bien que des détecteurs HgI_2 existent sur le marché, l'utilisation de ces types de détecteurs n'a pas connu le succès dans le domaine de l'ED-XRF.

Arrivèrent par la suite les détecteurs aux photodiodes Si-PIN (silicon positive intrinsic negative) qui sont aussi adaptés pour la détection des basses énergies (2-30keV). Ce type de détecteur a été rendu célèbre en 1997 durant la mission Pathfinder au cours de laquelle un détecteur Si-PIN a été utilisé pour l'analyse du sol et des roches de la surface de la planète Mars. La production du Silicium est une science déjà bien maîtrisée et très développée dans la majorité de l'industrie mondiale. En outres, le Si étant un élément simple et non un composé, il a de ce fait un degré élevé de stabilité. Pour ces raisons, les détecteurs au Si ont obtenu beaucoup de succès dans le domaine de l'ED-XRF. Le seul facteur qui peut affecter un détecteur Si-PIN est l'humidité. C'est pour cela que ces détecteurs sont produits avec un couvercle hermétique. Un exemple de ces détecteurs est le XR-100CR fabriqué par Amptek. Ce détecteur atteint la résolution de 185 eV à 5,9 keV avec son refroidissement thermoélectrique.

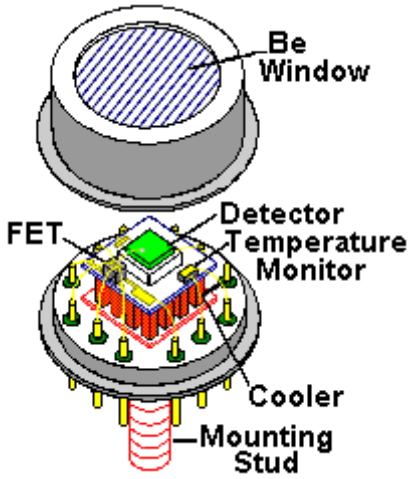


Figure 1.4 : Schéma d'un détecteur Si-PIN

Un autre type de détecteurs similaires au Si-PIN est ce qu'on appelle détecteur à semi-conducteur dopé (SDD : semiconductor drift detector). Les SDD ont la spécificité d'avoir une capacitance très basse pour diminuer les bruits de fond. Contrairement au Si-PIN qui a une surface active très petite (6-7mm²) pour diminuer la capacitance, les détecteurs SDD peuvent avoir une surface active pouvant atteindre 1cm² [4] augmentant ainsi leur performance en taux de comptage. On reporte une résolution en énergie de 145 eV à 5,9 keV pour ces types de détecteurs.

La plus récente et la plus impressionnante amélioration dans le domaine des détecteurs pour l'ED-XRF est le développement de détecteurs utilisant les superconducteurs. On peut citer par exemple le détecteur Nb/Al/AlOx/Nb [11] avec lequel une résolution en énergie aussi bonne que 29 eV à 5,9 keV a été obtenue avec un refroidissement à 6.10^{-2}°K. Le problème avec ces détecteurs est leur faible surface active ainsi que la nécessité de les maintenir autour de la température de 0°K.

1.3.2. Traitement numérique du signal :

Un autre développement très important dans le domaine de la spectrométrie ED-XRF est l'introduction du traitement numérique du signal (DSP : Digital Signal Processing). Dans les spectromètres analogiques, le signal (impulsions provenant du détecteur) est filtré, mis en forme, amplifié avant d'être numérisé. Dans le cas des systèmes numériques (DSP), le signal est aussitôt numérisé à la sortie du préamplificateur. C'est ce signal numérique

qu'on filtre et met en forme à l'aide d'algorithmes spéciaux pour être transféré au MCA à la fin du traitement. La majorité des tâches lors du traitement numérique du signal est actuellement effectuée par une seule puce appelée processeur numérique de signal. La taille des systèmes à DSP est ainsi très réduite et ils ne consomment que très peu d'énergie. Un autre avantage des spectromètres à DSP est leur stabilité par rapport à la température et le temps contrairement aux systèmes analogiques dont les comportements changent avec l'usure des composantes et la température. L'introduction du DSP dans les spectromètres ED-XRF a aussi permis une nette amélioration de la performance en taux de comptage. En effet, la fréquence d'échantillonnage lors de la numérisation du signal à la sortie du préamplificateur est relativement élevée (20MHz pour le DP4 de Amptek Inc. [27]). Ces spectromètres peuvent ainsi acquérir un signal dépassant la fréquence de 1MHz qui est très rarement atteinte dans les applications classiques de l'ED-XRF. Un exemple de ces types de spectromètres est le DP4 de Amptek Inc. qui a une dimension d'à peu près 9×6 cm² et une consommation de 400 mW.

Figure 1.5 : Le processeur numérique de signal DP4 de Amptek

La figure suivante montre le schéma d'un spectromètre complet autour du DP4.

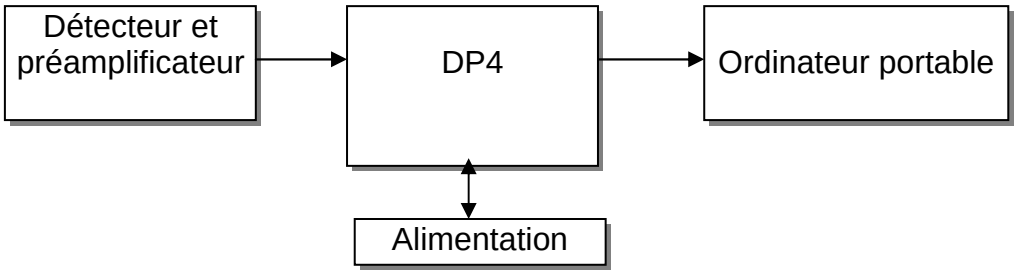


Figure 1.6 : Diagramme d'un spectromètre complet utilisant le DP4 pour le traitement du signal

1.3.3. Les tubes à rayons X :

Des recherches ont été aussi menées pour améliorer quelques caractéristiques des tubes à rayons X. Parmi ces caractéristiques, les plus importants sont la portabilité et la consommation en énergie. On utilise déjà au lieu des tubes à rayons X des sources radioactives scellées qui sont portatives comme l'^{241}Am, le ^{109}Cd, le ^{244}Cm, le ^{55}Fe. Le problème avec ces sources est le fait que les éléments qu'elles peuvent exciter sont très limités. Ceci n'est pas le cas des tubes à rayons X qui permet de détecter un très large éventail d'éléments. On peut trouver aujourd'hui sur le marché des tubes à rayons X miniaturisés. Un exemple est l'Eclypse III de Amptek Inc. [25]. Parmi ses caractéristiques techniques, on distingue sa consommation de 3W maximale avec un poids de 300g et une longueur de 15 cm environ.

Figure 1.7 : Le tube à rayons X Eclypse III

1.3.4. Spectromètres portatifs :

Avec tous ces modules miniaturisés et l'avènement des ordinateurs portables, on peut trouver actuellement des spectromètres complètement portatifs. Il existe même des systèmes comprenant à la fois le détecteur avec le préamplificateur et le module de traitement numérique du signal. Les spectromètres portatifs ont permis d'amener le laboratoire vers l'échantillon et non l'inverse. C'est en effet une exigence actuellement dans les applications comme les mesures sur terrain et les analyses non destructives surtout.

Un spectromètre portatif est composé par les éléments suivants :

- un détecteur avec le préamplificateur, un module de traitement de signal (numérique) qui comprend en gros un convertisseur analogique-digital, un circuit de mise en forme numérique du signal et un MCA. Dans la plupart des cas, ces différents modules se trouvent dans une seule unité.
- un ordinateur portable dont les rôles principaux sont la visualisation et le stockage des données

Un des domaines dans lesquels les spectromètres ED-XRF portatifs sont les plus utilisés actuellement est l'archéologie. Dans ce domaine, les échantillons sont dans la majorité des cas des objets dont l'aspect et les formes doivent rester inchangés ou qu'on ne peut même pas déplacer de leurs places d'origine (statuts, tableaux, murs,…). Plusieurs dispositifs ont été alors développés pour satisfaire aux exigences de ces types de mesures communément appelées tests non destructifs (NDT : Non-Destructive Testing). Ces exigences sont relatives au positionnement du détecteur et du système d'excitation (source radioactive ou tube à rayons X). Le dispositif sur la figure suivante montre un exemple de ce positionnement pour la mesure d'échantillons en archéologie.

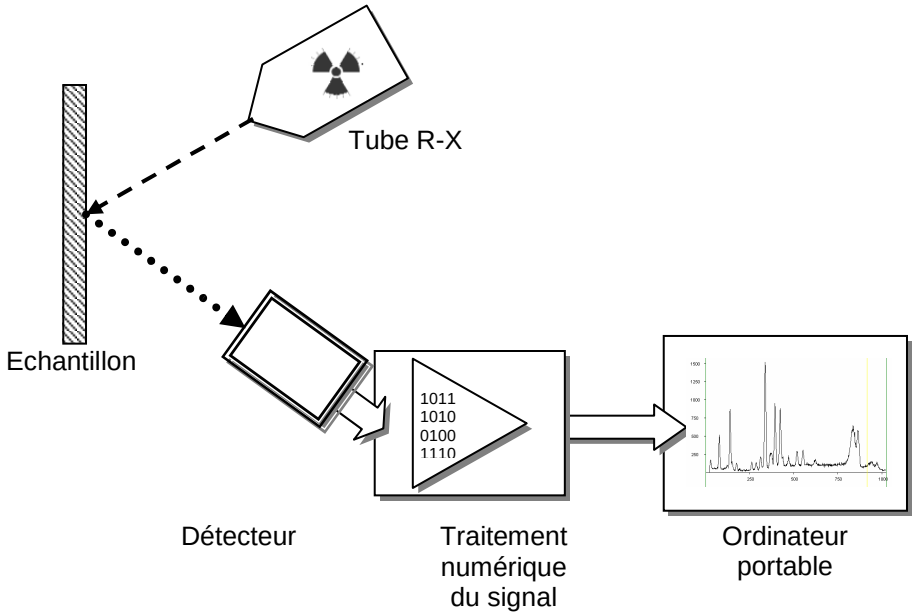


Figure 1.8 : Exemple de spectromètre portatif pour les mesures en NDT

L'échantillon doit être à une distance précise du dispositif de mesure pour que les rayons émis soient correctement captés par le détecteur. Le spectromètre utilise alors un système de pointage par rayon laser. L'échantillon est à la distance voulue quand les deux rayons laser parallèles au tube R-X et au détecteur coïncident.

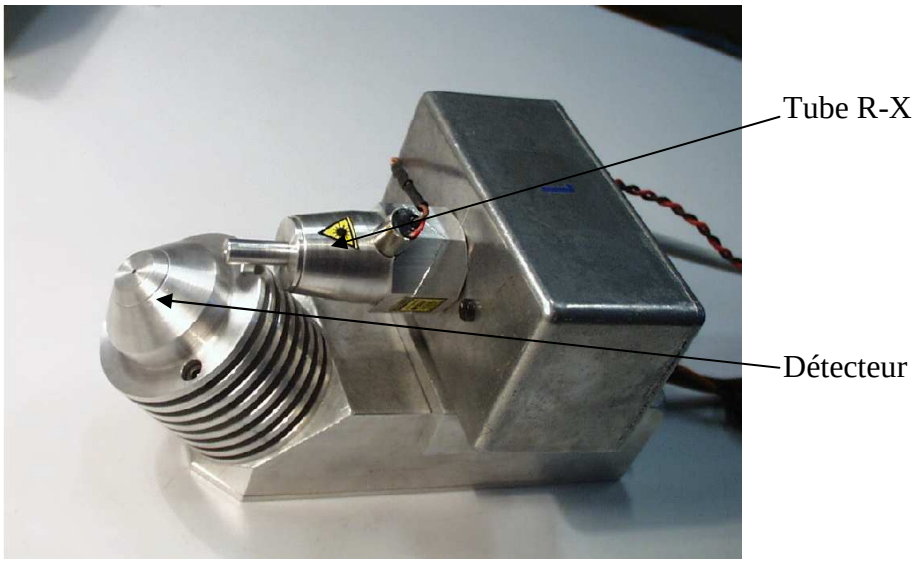


Figure 1.9 : Dispositif comprenant le tube R-X et le détecteur

L'exemple suivant montre la miniaturisation qu'ont pu atteindre les fabricants pour les spectromètres portables tout en gardant ou même améliorant les performances par rapport aux matériels de laboratoire. La figure suivante montre le modèle X-123 de Amptek Inc. [27]. Ce module est un spectromètre complet comprenant le détecteur avec son système de refroidissement, le préamplificateur, le traitement numérique du signal (comprenant le MCA).

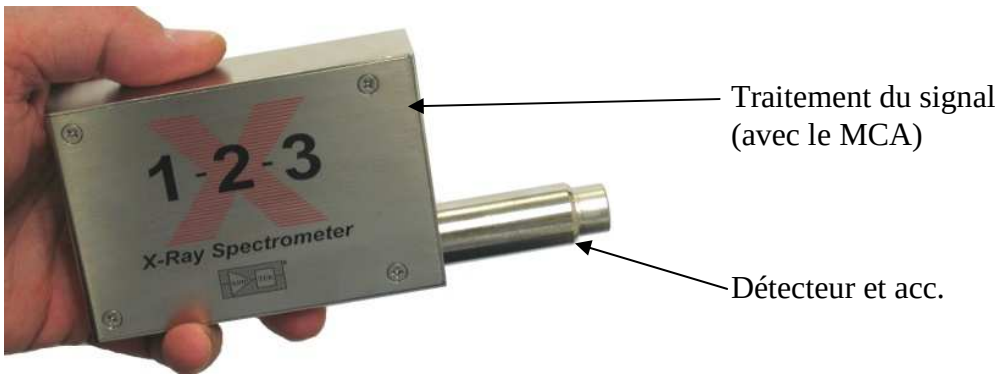


Figure 1.10 : Le spectromètre X-123

Chapitre 2

La régression PLS (Partial Least Squares)

2.1. GENERALITES :

2.1.1. Position du problème :

En science et en technologie ainsi qu'en d'autres domaines comme l'économie ou la démographie par exemple, les problèmes les plus souvent rencontrés sont des systèmes à entrées et sorties. Les entrées sont des phénomènes mesurables et souvent contrôlables et les sorties dépendent de ces entrées. Les méthodes de régression cherchent à évaluer les relations entre des variables indépendantes (entrées) et des variables dépendantes (sorties), en l'absence de modèles théoriques.

Le but de ces méthodes est ainsi de prédire les sorties d'un système donné chaque fois que les variables d'entrée sont mesurées, après avoir établi cette relation de régression. C'est pour ça que les variables d'entrée sont appelées « prédicteurs ». Dans le cas de la spectroscopie XRF, les prédicteurs sont les spectres qui seront représentés par la matrice [X] tandis que les sorties sont les concentrations des éléments chimiques étudiés et elles seront représentées par la matrice [Y].

Au cours de ce chapitre, nous allons considérer le cas général où [X] sera formée par des prédicteurs quelconques et [Y] les sorties correspondantes. Le problème se résume donc à la prédiction de [Y] à partir de [X] après avoir établi la relation de régression de la forme $[Y]=[X][\beta]+[E]$

où $[\beta]$ est le coefficient de régression et [E] le résidu de la régression de [Y] sur [X].

2.1.2. Notations et conventions :

Dans tout ce qui va suivre, nous allons adopter les conventions de notations suivantes.

Un vecteur sera noté comme suit :

$\overline{A}, \overline{B}, \overline{C}, \ldots$

Nous allons toujours utiliser les lettres majuscules pour les vecteurs.

Le i-ème élément d'un vecteur $\overline{A}$ sera ainsi représenté par la même lettre avec l'indice approprié : A^i

Les matrices seront représentées par des lettres majuscules entre crochets :

[A], [B], [C],…

Leurs éléments seront notés par les mêmes lettres avec des indices pour les lignes et les colonnes. Par exemple, les éléments de la matrice [X] sont X^i_j : i-ème ligne et j-ème colonne. Parfois, il est aussi pratique de définir une matrice comme étant une série de vecteurs. Une matrice [X] (n,p) peut ainsi être notée comme suit :

$$[X] = [\overline{X}_1, \overline{X}_2, ..., \overline{X}_p]$$

Le transposé d'une matrice [X] sera notée $[X]^t$

Une estimation ou prédiction de [X] sera notée : $[\hat{X}]$

[I] représente la matrice unité

2.2. LOI DE BEER-LAMBERT :

La loi de Beer-Lambert s'énonce comme suit :
Quand un échantillon est placé dans les rayons émis par un spectromètre, l'énergie qu'un élément de cet échantillon absorbe est proportionnelle à sa concentration à une longueur d'onde déterminée :

$$A_\lambda = \varepsilon_\lambda bC \qquad (2.1)$$

où A_λ est l'absorbance de l'échantillon pour la longueur d'onde λ
ε_λ : le coefficient d'absorption pour λ
b : la longueur du trajet à travers l'échantillon
et C la concentration du constituant considéré
ε_λ est différent pour chaque élément mais est constante pour une composition donnée et à la même longueur d'onde

Cette loi nous permet de supposer qu'une régression linéaire peut être établie entre la concentration d'un élément et les pics correspondant à cet élément dans un spectre obtenu à l'aide d'un spectromètre. Nous avons vu dans le paragraphe §1.2.2 (p.8) que ceci n'est valable que dans un intervalle limité de concentrations pour un élément donné. Quoi qu'il en soit, tout changement, qu'il soit dû à la concentration de l'élément d'intérêt ou aux autres éléments constituant la matrice, se répercute sur le spectre obtenu. En d'autres termes, le spectre contient toute l'information concernant l'échantillon à analyser. L'idée de l'utilisation des méthodes de régression est donc d'établir une relation de régression entre le spectre et la concentration de l'élément d'intérêt. Cette relation est de la forme

$$[Y] = [X][\beta] + [E] \qquad (2.2)$$

où [Y] représente la matrice formant les concentrations (variables réponses en général)

[X] est la matrice des spectres (variables prédictrices dans le cas général)

[β] est le coefficient de régression

[E] : une estimation de l'erreur commise lors de la régression

Différentes méthodes de régression vont ainsi être analysées pour l'établissement de cette équation de régression entre les concentrations et le spectre (ou absorbances) en spectroscopie XRF.

2.3. MOINDRES CARRES ORDINAIRES (MCO) :

2.3.1. Moindres carrés classiques (MCC) :

La relation (2.2) peut être écrite sous une autre forme:

$$[X]=[Y][\beta']+[E'] \quad (2.3)$$

où $[\beta']$ est une matrice (p,q) contenant les coefficients de régression

$[E_1]$ est une matrice (n,q) représentant les résidus de la régression

En MCO, $[\beta']$ est donné par la méthode des moindres carrés:

$$[\beta']= ([Y]^t[Y])^{-1}[Y]^t[X] \quad (2.4)$$

On peut ainsi prédire les sorties d'une série de prédicteurs $\overline{X}_i$ par

$$\overline{Y}_i = \left([\beta'][\beta']^t\right)^{-1}[\beta']\overline{X}_i \quad (2.5)$$

Cette formulation tend à impliquer que la méthode des MCC peut être utilisée en spectroscopie XRF car tout le spectre ([X]) est utilisé pour la modélisation des variables de sortie : les concentrations.

Le problème posé par cette méthode est que la matrice des coefficients de régression est obtenue à partir des variables réponses [Y]. En spectroscopie XRF, ceci équivaut à dire que les concentrations de tous les constituants présents dans les échantillons doivent être connues et que les échantillons doivent avoir exactement les mêmes constituants. Cette écriture de la relation de régression revient aussi à dire qu'un spectre peut être décrit comme un produit de spectres d'éléments purs (isolés) et de leurs concentrations respectives. Or, on sait qu'on ne peut pratiquement pas avoir de tels cas à cause des interactions entre les constituants (effet de matrice).

2.3.2. Moindres carrés inverses (MCI) :

La méthode des MCI est obtenue à partir de la relation (2.2): [Y]=[X][β]+[E]

L'avantage de cette méthode réside dans le fait qu'on peut traiter chaque colonne de [Y] indépendamment des autres colonnes :

$$\overline{Y} = [X]\overline{\beta} + \overline{E} \tag{2.6}$$

En spectroscopie XRF, cela veut dire qu'on peut modéliser les concentrations de chaque constituant même si on n'a aucune information ni sur les autres constituants ni sur la constitution même de l'échantillon.

Le coefficient de régression est donné par :

$$\overline{\beta} = \left([X]^t[X]\right)^{-1}[X]^t\overline{Y} \tag{2.7}$$

C'est le calcul de ce coefficient qui rend cette méthode pratiquement inutilisable en spectroscopie XRF. En effet, ce calcul nécessite l'inversion de la matrice $[X]^t[X]$ qui est une matrice carrée de dimension égale au nombre des longueurs d'ondes ou d'énergies utilisées pour avoir le spectre. Plusieurs colonnes de cette matrice sont colinéaires (raies K,L,… d'un même élément) la rendant ainsi singulière donc l'inversion n'est pas possible.

2.4. LA REGRESSION PAR COMPOSANTES PRINCIPALES :

La technique de la régression par composantes principales fait partie de ce qu'on appelle régression bilinéaire par opposition aux techniques des MCO qui sont des régressions linéaires. En effet, en plus du modèle linéaire entre les variables d'entrée et de sortie (modèle inter-groupe), chacune des variables est aussi décomposée en d'autres variables qui sont combinaisons linéaires des variables d'origine (modèle intra-groupe).

2.4.1. Analyse en composantes principales (ACP) :

Comme on a vu en MCO, on toujours des variables [X] dont les colonnes sont très corrélées. En ACP, on cherche des nouvelles variables de dimension très réduite par rapport à celle des variables d'origine, qui peuvent représenter au mieux les variations ou

informations contenues dans ces variables d'origine. Ces nouvelles variables, appelées composantes principales (CP) sont combinaisons linéaires des variables d'origine et sont choisies pour qu'elles soient orthogonales entre elles.

Soit $\overline{T} = [X]\overline{W}$ une combinaison linéaire quelconque de [X]

où $\overline{W}$ est un vecteur (p,1) appelé vecteur de poids (ou poids)

La variance de $\overline{T}$ est :

$$\text{var}(\overline{T}) = \text{var}([X]\overline{W}) = \overline{W}^t [X]^t [X]\overline{W} \quad (2.8)$$

Pour que $\overline{T}$ puisse représenter au mieux les variations contenues dans [X], il faut que cette variance soit maximale. On impose à $\overline{W}$ la condition $\overline{W}^t \overline{W} = 1$ pour éviter le cas $\overline{W}$ infini qui donne aussi un maximum. On montre [20] que cette maximisation conduit à une équation de la forme :

$$[X]^t [X]\overline{W} = \lambda \overline{W} \quad (2.9)$$

avec λ un scalaire quelconque.

$\overline{W}$ est donc vecteur propre de $[X]^t[X]$ correspondant à la valeur propre λ

(2.9) nous donne : $\text{var}(\overline{T}) = \lambda$ (2.10)

Donc, le premier vecteur propre $\overline{W}_1$ correspond à la plus grande valeur propre λ_1 de $[X]^t[X]$.

Les vecteurs de poids $\overline{W}_i$ (donc les composantes principales $\overline{T}_i$) sont ainsi extraits sous la contrainte d'orthogonalité et le i-ème vecteur correspond à la i-ème plus grande valeur propre de $[X]^t[X]$.

Si m vecteurs sont extraits, on a en écriture matricielle :

$[T_m]=[X][W_m]$ (2.11)

où $[T_m]$ est une matrice (n,m) des composantes principales

et $[W_m]$ une matrice (p,m) des vecteurs de poids

avec $m<<p$

La matrice des prédicteurs [X] peut donc se décomposer suivant :

$[X]=[T_m][W_m]^t$ (2.12)

Et comme le nombre maximal de composantes principales qu'on peut extraire est égal au rang de [X] : p, (2.12) est donc vrai pour m=p, mais pour m < p, on a :

$$[\hat{X}]_m = [T_m][W_m]^t \quad (2.13)$$

qui est la meilleure approximation de [X] dans le sens des moindres carrés.

L'analyse en composantes principales peut donc se résumer en un changement de variables ou plus précisément une projection des données [X] sur une nouvelle base formée par les $\overline{T}_i$.

2.4.2. Régression en composantes principales (PCR) :

La régression en composantes principales consiste à remplacer la variable [X] par ses composantes principales dans l'équation de régression entre [X] et [Y]

$$\overline{Y} = [X][\beta] + \overline{E} \quad (2.14)$$

sachant que $[X]=[T_m][W_m]^t$, on peut exprimer $\overline{Y}$ en fonction des composantes principales $[T_m]$:

$$\overline{Y} = [T_m]\overline{A}_m + \overline{E} \quad (2.15)$$

où $\overline{A}_m$ est un vecteur (m,1) des coefficients de régression de $\overline{Y}$ sur $[T_m]$

En utilisant les moindres carrés, la meilleure approximation de $\overline{A}_m$ est donnée par :

$$\overline{A}_{m,PCR} = ([T_m]^t[T_m])^{-1}[T_m]^t\overline{Y} \quad (2.16)$$

Comme les composantes principales sont orthogonales, on peut écrire :

$$[T_m]^t[T_m]=[L_m] \quad (2.17)$$

où $[L_m]$ est une matrice diagonale (m,m) avec les m premières valeurs propres de $[X]^t[X]$ comme éléments. On a donc :

$$\overline{A}_{m,PCR} = [L_m]^{-1}[T_m]^t\overline{Y} \quad (2.18)$$

Substituant $\overline{A}_m$ dans (2.15), on peut écrire la meilleure approximation de $\overline{Y}$ en PCR :

$$\overline{Y}_{m,PCR} = [T_m][L_m]^{-1}[T_m]^t\overline{Y} + \overline{E} \quad (2.19)$$

Et comme $[T_m]=[X][W_m]$, (2.19) devient :

$$Y_{m,PCR} = [X][W_m][L_m]^{-1}[W_m]^t[X]^t\overline{Y} + \overline{E} \quad (2.20)$$

Et si nous comparons cette équation avec (2.14), on déduit :

$$\overline{\beta} = [W_m][L_m]^{-1}[W_m]^t[X]^t\overline{Y} \quad (2.21)$$

Cette expression du coefficient de régression nous permet de souligner les caractéristiques suivantes de la régression PCR :

- la variable d'entrée [X] est utilisée en entier et ne nécessite pas d'inversion. La colinéarité des colonnes $\overline{X}_i$ de [X] ne pose donc aucun problème pour le calcul du coefficient de régression
- pour prédire une variable de sortie $\overline{Y}_i$ de [Y], la connaissance des autres colonnes de [Y] n'est pas nécessaire

La régression PCR combine donc les avantages des MCO et des MCI.

Un problème majeur survient pourtant quant à l'utilisation de cette méthode en spectroscopie XRF. Ce problème provient de la manière dont les composantes principales sont extraites. En effet, on a vu que les composantes principales sont choisies suivant l'importance de leurs variances, donc de leurs valeurs propres, sans aucune information sur leur capacité à prédire une variable de sortie donnée. Les composantes principales à faible variance sont ainsi éliminées.

Dans le cas de la spectroscopie XRF, prenons l'exemple de la modélisation de la concentration du Pb dans des échantillons de sols. Cet élément est responsable d'une toute petite variation seulement au sein du spectre. Les autres éléments comme le Si et le Fe par exemple produisent quant à eux la part la plus importante de la variation dans le spectre. Les premières plus grandes CP seront donc des « représentations » de ces éléments alors que ces CP serviront à modéliser la concentration du Pb.

C'est pour ça qu'une autre approche qui consiste à extraire les CP de telle sorte qu'elles soient en corrélation avec les variables de sortie a été développée. Ceci est le principe de base de la régression PLS.

2.5. LA REGRESSION PLS :

Comme en PCR, on a des variables très corrélées. On cherche donc à extraire des données originales des variables non corrélées appelées variables latentes qui ne sont autres que des composantes principales des variables d'origine. La régression PLS diffère de la régression PCR sur la construction de ces nouvelles variables. En effet, ces variables sont choisies pour représenter au mieux le groupe des prédicteurs [X] tout en modélisant la relation entre [X] et [Y]. En d'autres termes, la régression PLS réduit la dimension des variables (prédicteurs et réponses) en faisant une projection sur la direction suivant laquelle la covariance entre [X] et [Y] est maximum.

On cherche donc des composantes colinéaires à [X] : $\overline{T} = [X]\overline{W}$ et colinéaires à [Y] : $\overline{U} = [Y]\overline{C}$ telles que $\text{cov}(\overline{T}, \overline{U})$ soit maximum.

$$\left\{\overline{W}, \overline{C}\right\} = \arg\max \text{cov}\left([X]\overline{W}, [Y]\overline{C}\right) = \arg\max \overline{W}^t [S_{xy}]\overline{C} \tag{2.22}$$

sous les contraintes $\left\|\overline{W}\right\|^2 = \left\|\overline{C}\right\|^2 = 1$ (2.23)

avec $[S_{xy}]=[X]^t[Y]$

et $[S_{yx}]=[Y]^t[X]$

on a donc $[S_{xy}]^t=[S_{yx}]$

(2.23) peut s'écrire :

$$\overline{W}^t \overline{W} - 1 = 0$$

$$\overline{C}^t \overline{C} - 1 = 0$$

et en utilisant la méthode des multiplicateurs de Lagrange, on a :

$$L(\overline{W}, \overline{C}) = \overline{W}^t [S_{xy}]\overline{C} - \frac{\lambda_1}{2}(\overline{W}^t \overline{W} - 1) - \frac{\lambda_2}{2}(\overline{C}^t \overline{C} - 1) \tag{2.24}$$

Où λ_1 et λ_2 sont les multiplicateurs de Lagrange. Le maximum pour cette fonction est obtenu pour :

$$\begin{cases} \dfrac{\partial L}{\partial \overline{W}} = [S_{xy}]\overline{C} - \lambda_1 \overline{W} = 0 \\ \dfrac{\partial L}{\partial \overline{C}} = [S_{yx}]\overline{W} - \lambda_2 \overline{C} = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \overline{W}^t[S_{xy}]\overline{C} = \lambda_1 \\ \overline{C}^t[S_{yx}]\overline{W} = \lambda_2 \end{cases}$$

et comme $[S_{xy}]^t=[S_{yx}]$, on remarque que :

$$\lambda_1 = \lambda_1^{\ t}$$

$$\lambda_1 = \left(\overline{W}^t[S_{xy}]\overline{C}\right)^t$$

$$\lambda_1 = \overline{C}^t[S_{yx}]\overline{W}$$

$$\lambda_1 = \lambda_2$$

On peut donc écrire

$$\begin{cases} \overline{W}^t[S_{xy}]\overline{C} = \lambda \\ \overline{C}^t[S_{yx}]\overline{W} = \lambda \end{cases}$$

$$\begin{cases} [S_{xy}][S_{yx}]\overline{W} = \lambda^2\overline{W} \\ [S_{yx}][S_{xy}]\overline{C} = \lambda^2\overline{C} \end{cases} \qquad (2.25)$$

$\overline{W}$ et $\overline{C}$ sont donc vecteurs propres de $[S_{xy}][S_{yx}]=[X]^t[Y][Y]^t[X]$ et de $[S_{yx}][S_{xy}]=[Y]^t[X][X]^t[Y]$ respectivement, avec la même valeur propre λ^2.

Le premier vecteur propre $\overline{W}_1$ est $\overline{W}_1 = \dfrac{[X]^t[Y]}{\left\|[X]^t[Y]\right\|}$

La première composante PLS s'écrit donc : $\overline{T}_1 = [X]\dfrac{[X]^t[Y]}{\left\|[X]^t[Y]\right\|}$

Supposons obtenues les composantes $\overline{T}_1,...,\overline{T}_{h-1}$. On cherche la h-ème composante $\overline{T}_h = [X]\overline{W}_h$ orthogonale à $\overline{T}_1,...,\overline{T}_{h-1}$.

$\overline{W}_h$ doit donc vérifier : $\overline{W}_h{}^t[X]^t(\overline{T}_1,...,\overline{T}_{h-1}) = 0$

d'où $\overline{W}_h = ([I]-[Q_{h-1}])\overline{W}_h$

où $[Q_{h-1}]$ est l'opérateur de projection orthogonale sur l'espace engendré par $[X]^t\overline{T}_1,...,[X]^t\overline{T}_{h-1}$

Il en découle pour la covariance :

$$\mathrm{cov}\left([X]\overline{W}_h,[Y]\overline{C}_h\right)=\overline{W}_h{}^t[X]^t[Y]\overline{C}_h$$
$$\mathrm{cov}\left([X]\overline{W}_h,[Y]\overline{C}_h\right)=\overline{W}_h{}^t([I]-[Q_{h-1}])[X]^t[Y]\overline{C}_h$$
$$\mathrm{cov}\left([X]\overline{W}_h,[Y]\overline{C}_h\right)=\mathrm{cov}\left([X_{h-1}]\overline{W}_h,[Y]\overline{C}_h\right)$$

avec $[X_{h-1}]=[X]([I]-[Q_{h-1}])$ représentant le résidu de la régression de $[X]^t$ sur $[X]^t\overline{T}_1,...,[X]^t\overline{T}_{h-1}$

$\overline{W}_h$ est donc vecteur propre de $[X_{h-1}]^t[Y][Y]^t[X_{h-1}]$

d'où l'expression générale de $\overline{W}_h$ avant normalisation :

$$\overline{W}_h=[X]^t([I]-[Q_{h-1}])[Y]$$

$$\overline{W}_h=[X]^t\left([I]-\sum_{i=1}^{h-1}\frac{\overline{T}_i\overline{T}_i{}^t}{\overline{T}_i{}^t\overline{T}_i}\right)[Y] \qquad (2.26)$$

Le calcul des composantes s'accompagne donc d'une déflation de la matrice [X] à chaque obtention d'une nouvelle composante :

$$[X_{h+1}]=[X_h]-\overline{T}_h\overline{P}_h{}^t \qquad (2.27)$$

Avec $\overline{P}_h=\dfrac{[X]^t\overline{T}_h}{\overline{T}_h{}^t\overline{T}_h}$ (2.28)

On définit de la même manière les composantes $\overline{U}_h=[Y]\overline{C}_h$ en montrant que $\overline{C}_h$ est vecteur propre de $[Y_{h-1}]^t[X][X]^t[Y_{h-1}]$.

2.5.1. Les algorithmes PLS :

Il existe actuellement plusieurs types d'algorithmes de régression PLS : NIPALS, SIMPLS[2], kernel PLS,… Chaque type présente différentes variantes selon leur application. Nous allons nous intéresser à l'algorithme NIPALS (Nonlinear estimation by Iterative Partial Least Squares) qui est à l'origine de la régression PLS. Cet algorithme est aussi le plus utilisé grâce à sa facilité d'implémentation et sa stabilité.

[2] *Straightforward Implementation of a statistically inspired modification of the PLS method*

L'algorithme NIPALS a été présenté pour la première fois par l'inventeur de la régression PLS : Herman Wold en 1975 sous le nom de NILES (Nonlinear estimation by Iterative Least Squares). Il a été alors utilisé pour l'analyse en composantes principales. C'est en 1983 que Svante Wold a développé la version de NIPALS qui est actuellement le standard utilisé en régression PLS. Cet algorithme se présente sous deux formes :

- PLS1 : la variante utilisée pour le cas où on a une seule variable de sortie (un seul élément d'intérêt dans le cas de la spectrométrie XRF)
- PLS2 : la variante générale (modélisation de plusieurs éléments en même temps)

2.5.1.1. Le modèle PLS :

Comme on a vu au début de ce paragraphe, le modèle PLS consiste à chercher des nouvelles variables appelées variables latentes qui modélisent les données d'origine [X], [Y].

On part des composantes principales de [X] : $\overline{T}_i$. Ces composantes sont colinéaires à [X] et s'écrivent sous forme matricielle :

$$[T]=[X][W^*] \tag{2.29}$$

$[W^*]$ étant les vecteurs de poids (ou poids)

Ces composantes $\overline{T}_i$ ont les propriétés suivantes :

a) Comme l'a indiqué la relation (2.27), à chaque obtention d'une nouvelle composante $\overline{T}_i$ (ou $\overline{W}_i$), la matrice [X] est déflatée d'une quantité $\overline{T}_i \overline{P}_i^t$. On en déduit qu'après avoir obtenu toutes les composantes, [X] peut s'exprimer en fonction de [T] comme suit :

$$[X]=[T][P]^t+[E] \tag{2.30}$$

où [E] est la partie de [X] non expliquée par le modèle (résidu)

[E] doit donc tendre vers 0 pour que le modèle puisse être fiable

b) Ces composantes principales doivent aussi pouvoir modéliser la variable de sortie [Y] :

$$[Y]=[T][C]^t+[F] \tag{2.31}$$

où [F] est le résidu de la décomposition de [Y] sur [C]

Comme [Y] peut aussi s'écrire en fonction de ses composantes principales $\overline{U}_i$:

$[Y]=[U][C]^t+[G]$ (2.32)

En combinant (2.29) et (2.31), on a :

$[Y]=[X][W^*][C]^t+[F]$ (2.33)

$[Y]=[X][\beta]+[F]$

où $[\beta]$ est le coefficient de régression de [Y] sur [X]

2.5.1.2. L'algorithme NIPALS en régression PLS :

1. Initialiser le vecteur $\overline{U}$: on prend en général une colonne de [Y]
2. Calculer le poids de [X] : $\overline{W} = \dfrac{[X]^t \overline{U}}{\left\|\overline{U}\right\|^2}$
3. Calculer la composante principale $\overline{T} = [X]\overline{W}$
4. Les vecteurs de poids de [Y] : $\overline{C} = \dfrac{[Y]^t \overline{T}}{\left\|\overline{T}\right\|^2}$
5. Calculer la nouvelle valeur de $\overline{U}$: $\overline{U} = \dfrac{[Y]\overline{C}}{\left\|\overline{C}\right\|^2}$
6. Tester la convergence de $\overline{W}$: $\left\|\overline{W}_{old} - \overline{W}_{new}\right\| / \left\|\overline{W}_{new}\right\| < \varepsilon$

   Si cette condition est satisfaite, on continue à l'étape 7. Sinon, on retourne à l'étape 2
7. Enlever la contribution des composants obtenues de [X] et [Y] :

   $\overline{P} = \dfrac{[X]^t \overline{T}}{\left\|\overline{T}\right\|^2}$

   $[X] = [X] - \overline{T}\overline{P}^t$

   $[Y] = [Y] - \overline{T}\overline{C}^t$
8. Revenir à l'étape 1 pour la série de composantes suivante jusqu'au nombre de composantes maximales a qui sera déterminé par validation croisée.

2.5.2. Prédictions :

D'après ce qu'on a vu plus haut (cf. relation (2.26)), $\overline{W}_h$ est vecteur propre de $[X_{h-1}]^t[Y][Y]^t[X_{h-1}]$ associé à la plus grande valeur propre. On a ainsi : $\overline{T}_h = [X_{h-1}]\overline{W}_h$ où $[X_{h-1}]$ est le résidu de la régression après avoir extrait la (h-1)ème composante.

Mais la composante principale $\overline{T}_h$ peut aussi s'écrire en fonction de [X] (données d'origine) :

$\overline{T}_h = [X]\overline{W}_h^*$ ou en représentation matricielle :

$[T]=[X][W^*]$

Les $\overline{W}_h^*$ sont définies comme suit :

- on a vu qu'après obtention d'une composante, la matrice [X] est déflatée de $\overline{T}_h \overline{P}_h{}^t$. On a donc : $[X_h] = [X]\prod_{j=1}^{h}\left([I] - \overline{W}_j \overline{P}_j{}^t\right)$
- Il en résulte que : $W_h^* = \prod_{j=1}^{h-1}\left([I] - \overline{W}_j \overline{P}_j{}^t\right)\overline{W}_h$ pour h>1 et $\overline{W}_1^* = \overline{W}_1$ (2.34)

On montre [20] que les vecteurs de poids $\overline{W}_h^*$ peuvent s'écrire sous la forme matricielle suivante :

$[W_h^*] = [W_h]([P_h]^t[W_h])^{-1}$ (2.35)

(2.33) peut donc s'écrire :

$[Y]=[X][W]([P]^t[W])^{-1}[C]^t+[E]$

D'où le coefficient de régression:

$[\beta]=[W]([P]^t[W])^{-1}[C]^t$ (2.36)

L'équation de régression pour la prédiction d'une variable de sortie $\overline{Y}$ à partir d'une variable d'entrée $\overline{X}$ est ainsi : $\overline{Y} = \overline{X}\left([W]\left([P]^t[W]\right)^{-1}[C]^t\right)$ (2.37)

2.5.3. Validation du modèle – validation croisée :

Le problème qui se pose après avoir construit un modèle de prédiction est de connaître son aptitude à prédire les réponses de nouvelles variables d'entrée [X]. Pour le cas du PLS, ceci consiste surtout à évaluer le nombre optimal de composantes à inclure dans le modèle.

En effet, si trop peu de composantes sont incluses, il se peut que des phénomènes importants se produisant au sein des données étudiées ne soient pas modélisés. Si on prend par contre trop de composantes, on a le phénomène de surévaluation (overfitting) : on modélise les phénomènes sans importance comme les bruits et ceci peut fausser les prédictions.

La méthode la plus couramment utilisée pour évaluer le nombre optimal de composantes est la validation croisée ou cross-validation.

2.5.3.1. Principe de la validation croisée :

La validation croisée consiste à séparer les données {[X],[Y]} en deux groupes :

- 1 groupe d'étalonnage qui va servir à élaborer le modèle
- 1 groupe–test sur lequel le modèle sera appliqué pour évaluer sa capacité de prédiction

2.5.3.2. La validation croisée « Leave One Out » (LOO-CV) :

Le LOO-CV est une variante de la validation croisée dans laquelle le groupe d'étalonnage est composé des n-1 échantillons et le groupe–test est donc formé par l'échantillon restant.

Le modèle sera tout d'abord ajusté sur les n-1 échantillons. Les prévisions des variables [Y] seront calculées sur le groupe–test. La somme des carrés des erreurs (différence entre les [Y] calculés ou prédits et les valeurs connues des [Y]) sera ensuite évaluée. Cette procédure est répétée n fois jusqu'à ce que chacune des échantillons serve une fois de test. La somme de tous les carrés des erreurs obtenues forme ce qu'on appelle le PRESS : PRediction Error Sum of Squares.

Comme on a en général plusieurs constituants (plusieurs colonnes dans [Y]), l'expression du PRESS pour le k-ème constituant est :

$$PRESS_{kh} = \sum_{i=1}^{n} (\hat{Y}_{ih}^{k} - Y^{k})^2 \qquad (2.38)$$

où $\hat{Y}_{ih}^{k}$ est la prévision de Y^k en utilisant le i-ème échantillon comme groupe-test et avec un modèle PLS à h composantes.

Le PRESS est très utile pour la validation du modèle car c'est à partir de cette grandeur qu'on peut considérer qu'une composante PLS est significative ou non. En effet, cette quantité décroît en fonction du nombre de composantes pour atteindre une valeur minimale et se stabiliser par la suite. C'est ce minimum qui détermine le nombre de composantes à retenir pour le modèle.

Une méthode très pratique pour déterminer ce minimum est l'introduction de la quantité Q^2 [20]:

$$Q_h^2 = 1 - \frac{PRESS_h}{PRESS_{h-1}} \qquad (2.39)$$

où $PRESS_{h-1}$ est la somme des carrés des erreurs avec le modèle à h-1 composantes.

Cas du PLS1 (une seule variable Y) :

Une h-ème composante est jugée significative si Q^2_h est supérieure à une certaine limite (arbitraire) qui est fixée à $(1-0.95)^2=0.0975$ selon S.Wold et al. pour le logiciel SIMCA-P [20]

Ceci équivaut à dire que la h-ème composante est significative si :

$$\sqrt{PRESS_h} \leq 0.95\sqrt{PRESS_{h-1}}$$

Cas du PLS2 :

En PLS2, si on a q constituants : [Y] est de dimension (n,q), on a pour chaque variable $\overline{Y}_k$:

$$Q_{kh}^2 = 1 - \frac{PRESS_{kh}}{PRESS_{k(h-1)}}$$

et pour tout l'ensemble [Y] :

$$Q_h^2 = 1 - \frac{\sum_{k=1}^{q} PRESS_{kh}}{\sum_{k=1}^{q} PRESS_{k(h-1)}}$$

On décide ainsi que l'apport de la h-ème composante est significatif si $Q^2_h >= 0.0975$

ou si au moins un $Q^2_{kh} >= 0.0975$

Similairement au PRESS, on définit aussi le RMSEC (Root Mean Square of Error for Calibration). Cette quantité n'est autre que la racine carrée du PRESS et est utilisée au même titre que ce dernier.

$$RMSEC_{kh} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_{ih}^{k} - Y^{k})^2} \qquad (2.40)$$

2.5.4. Estimation des intervalles de prédiction :

En spectroscopie ED-XRF, comme dans tous les domaines de la science, l'expression numérique d'une quantité donnée n'a de signification que si l'on indique une estimation de l'incertitude. Ceci est valable que ce soit lors des mesures à l'aide d'un appareil quelconque ou lors de calculs numériques. En régression PLS, bien qu'on puisse obtenir sans problèmes majeurs le coefficient de régression [β], l'évaluation de l'incertitude commise lors d'une prédiction n'est pas du tout évidente. Contrairement à la régression linéaire pour laquelle les intervalles de prédiction sont donnés par des expressions bien connues en statistique, la seule quantité qui peut évaluer l'incertitude en PLS est le PRESS ou les autres quantités similaires comme le RMSEC et le RMSEP. Comme on a vu au paragraphe §2.5.3.2, le PRESS est une mesure globale de la fiabilité du modèle PLS qu'on a construit. Il ne peut pas ainsi être utilisé comme incertitude lors de prédiction d'échantillon inconnu. On aurait en effet à ce moment là une valeur constante de l'incertitude pour tout échantillon inconnu.

Plusieurs propositions ont été alors faites pour le calcul de l'intervalle de prédiction pour un échantillon donné.

2.5.4.1. Incertitude dérivée du PRESS [8]:

Pour cette méthode, on utilise le RMSEP (Root Mean Square Error of Prediction) qui n'est autre que la racine carrée du PRESS pour évaluer l'incertitude de prédiction pour un ensemble d'échantillons inconnus.

$$\mathrm{RMSEP} = \left[\frac{1}{n}\sum\left(\hat{Y}^i - Y^i\right)^2\right]^{1/2} \tag{2.41}$$

Pour tenir compte des incertitudes sur les valeurs des variables utilisées pour les standards (concentrations dans notre cas), une version corrigée du RMSEP est utilisée :

$$\mathrm{RMSEP}_{cor} = (\mathrm{MSEP} - \Delta Y)^{1/2} \tag{2.42}$$

où ΔY représente l'erreur globale sur les mesures des Y^i pour les standards.

On peut aussi déterminer l'incertitude commise au niveau d'un échantillon i. Cette incertitude liée à la prédiction de la quantité Y^i est donnée par :

$$\mathrm{SEP}_i = [(1 + h_i)\mathrm{MSEC}]^{1/2} \tag{2.43}$$

où MSEC est la somme des carrés des erreurs pour le modèle construit avec tous les échantillons standard

h_i est un facteur relatif à l'échantillon i. Ce facteur est considéré comme une pondération. Il est en effet proportionnel à l'écart par rapport à la moyenne des échantillons standard.

Comme pour l'application globale de l'incertitude pour un modèle donné, on introduit aussi le SEP corrigé :

$$\mathrm{SEP}_{cor,i} = [(1 + h_i)\mathrm{MSEC} - \Delta Y]^{1/2} \tag{2.44}$$

2.5.4.2. Développement linéaire du coefficient de régression [β] ([10][18]):

Une expression évidente de l'incertitude sur [β] et par conséquent sur la prédiction $\hat{Y}^u$ aurait été possible si [β] était une fonction linéaire de $\overline{Y}$. On a montré qu'une

linéarisation locale de la fonction $[\beta(\overline{Y})]$ est possible en développant $[\beta]$ en séries de Taylor d'ordre 1 en un vecteur $\overline{Y}_0$

$$[\beta(\overline{Y})] \approx [\beta(\overline{Y}_0)] + \left(\frac{\partial[\beta]}{\partial\overline{Y}}\right)_{\overline{Y}_0} (\overline{Y} - \overline{Y}_0) \tag{2.45}$$

D'après cette approximation linéaire, une évaluation de l'intervalle de prédiction a été obtenue. Elle est donnée pour un niveau de confiance α par:

$$\Delta\hat{Y}_u = t_{\alpha/2,df}\sigma\left[\frac{n+1}{n} + \overline{Y}_u{}^t\left(\frac{\partial[\beta]}{\partial\overline{Y}}\right)_{\overline{Y}_0}\left(\frac{\partial[\beta]}{\partial\overline{Y}}\right)^t_{\overline{Y}_0}\overline{Y}_u\right]^{1/2} \tag{2.46}$$

où $t_{\alpha/2,df}$ est la quantile pour une distribution de Student à df degré de liberté.

Le problème majeur pour cette expression de l'intervalle de prédiction réside dans le calcul de la matrice jacobienne $\frac{\partial[\beta]}{\partial\overline{Y}}$. Il se trouve en effet que ce calcul différentiel entre matrices n'est faisable que de façon itérative. Ceci implique que $[\beta]$ peut être obtenu itérativement. C'est pourtant uniquement l'algorithme SIMPLS qui permet cette façon de calculer $[\beta]$. Cette méthode d'évaluation de l'intervalle de prédiction n'est donc pas applicable à notre travail qui utilise l'algorithme NIPALS.

2.5.5. Utilisation de la méthode bootstrap :

2.5.5.1. Technique de rééchantillonnage :

Le rééchantillonnage est une technique statistique qui a pour but de quantifier l'incertitude commise lors de la prédiction d'un paramètre quelconque (moyenne, variance,…) ou de valider un modèle donné. En général, le rééchantillonnage consiste à la réutilisation des échantillons existants pour déterminer la fonction de distribution du paramètre en question.

Les différents types de rééchantillonnage sont utilisés en l'absence de conditions suffisantes pour effectuer des tests paramétriques. C'est le cas par exemple d'échantillon avec un petit nombre d'individus ou provenant de distribution non normale.

Les techniques de rééchantillonnage les plus utilisées sont le bootstrap, le jacknife et le test de permutation mais nous n'allons détailler que le bootstrap que nous allons utiliser pour la détermination d'intervalle de prédiction.

2.5.5.2. Le bootstrap :

Le bootstrap est une technique de rééchantillonnage qui consiste à créer plusieurs échantillons de même taille que l'échantillon initial par tirage aléatoire avec remise à partir de l'échantillon initial. Ces nouveaux échantillons sont appelés « échantillons bootstrap » et peuvent être décrits comme des versions aléatoires de l'échantillon initial. Soit par exemple $\mathbf{x}$ l'échantillon initial, avec $\mathbf{x}=(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$

Un échantillon bootstrap est noté
$\mathbf{x}^{*m}=(\mathbf{x}^*_1, \mathbf{x}^*_2, \ldots, \mathbf{x}^*_n)^m$
où m=1,..,b est le numéro de l'échantillon

$\mathbf{x}^*_i$ est tiré de $\mathbf{x}$ avec remise. Il se peut donc que des éléments $\mathbf{x}_i$ de $\mathbf{x}$ apparaissent plusieurs fois dans $\mathbf{x}^{*m}$ et d'autres pas du tout.

2.5.5.3. Réplication bootstrap :

Soit $s(\mathbf{x})$ un paramètre quelconque de $\mathbf{x}$: la moyenne ou la variance par exemple. $s(\mathbf{x}^{*m})$ est appelé réplication bootstrap de ce paramètre. Si on prend par exemple le cas de la moyenne : $s(\mathbf{x}) = \frac{1}{n}\sum_i \mathbf{x}_i$

La réplication bootstrap de la moyenne est alors la moyenne des $\mathbf{x}^*_i$:

$$s(\mathbf{x}^{*m}) = \frac{1}{n}\sum_i \mathbf{x}_i^{*m}$$

2.5.5.4. Algorithme général :

L'algorithme général pour l'estimation d'un paramètre statistique quelconque par la méthode bootstrap est le suivant :

- générer un certain nombre b d'échantillons bootstrap à partir de l'échantillon initial

- pour chacun de ces échantillons, déterminer les réplications bootstrap du paramètre en question
- faire une estimation bootstrap du paramètre en question. Cette étape consiste à définir la fonction de distribution du paramètre

2.5.5.5. Application du bootstrap à la régression PLS [3][10]:

Le but de l'application du bootstrap en régression PLS est de définir un intervalle de confiance des prédictions. On génère ainsi des échantillons bootstrap basés sur les résidus. Pour simplifier la compréhension du principe, nous allons considérer dans ce paragraphe le cas où on a une seule variable (PLS1). La régression PLS entre $\overline{Y}$ et les variables prédictrices [X] étant :

$$\overline{Y} = [X]\overline{\beta} + \overline{E}$$

Un échantillon bootstrap m est de la forme

$$\overline{Y}^{*m} = [X]\overline{\beta} + \overline{E}^{*}$$

avec $\overline{E}^{*} = \begin{pmatrix} E^{*1} \\ ... \\ E^{*n} \end{pmatrix}$

où E^{*i} est un terme aléatoire issu des résidus E^{i} de la régression initiale.

On a ainsi b échantillons $\left\{[X], \overline{Y}^{*m}\right\}$. On calcule ensuite pour chacun de ces échantillons, une réplication bootstrap du coefficient de régression $\overline{\beta}^{*m}$ tel que $\overline{Y}^{*m} = [X]\overline{\beta}^{*m}$

2.5.5.6. Intervalle de confiance du coefficient de régression :

Les réplications bootstrap $\overline{\beta}^{*m}$ obtenues plus haut permettent d'établir un intervalle de confiance du coefficient de régression. La méthode la plus simple est la méthode percentile. L'intervalle de confiance percentile est en effet délimité par la (1-α)r-ème et la αr-ème valeurs de la liste ordonnée des $\overline{\beta}^{*m}$ pour un niveau de confiance (1-2α).

L'autre approche appelée bootstrap percentile-t ou bootstrap-t consiste à établir l'intervalle de confiance à partir d'une nouvelle variable appelée souvent « pivot approximatif » :

$\bar{z}^{*m} = \dfrac{\bar{\beta}^{*m} - \hat{\beta}}{s(\bar{\beta}^{*m})}$ où $s(\bar{\beta}^{*m})$ est l'estimation bootstrap de l'écart type du coefficient de régression.

On estime par la suite le α percentile de la distribution de $\bar{z}^{*m}$ par rapport à $\hat{t}(\alpha)$ tel que le nombre des valeurs observées de $\bar{z}^{*m}$ soit égal à α % de r : $\#\left\{\bar{z}^{*m} \leq \hat{t}(\alpha)\right\} = \alpha r$

L'intervalle de confiance percentile-t pour le coefficient de régression est alors :

$$\left[\bar{\beta} - s(\bar{\beta})\hat{t}(1-\alpha), \bar{\beta} - s(\bar{\beta})\hat{t}(\alpha)\right] \qquad (2.47)$$

2.5.5.7. Intervalle de prédiction :

Deux approches peuvent aussi être utilisées pour la détermination de l'intervalle de confiance des prédictions en utilisant la méthode bootstrap. La première méthode est la méthode percentile qui consiste à estimer la distribution de l'erreur de prédiction à partir des erreurs bootstrapées pour construire un intervalle de prédiction. Comme pour les coefficients de régression, l'approche percentile-t peut aussi être utilisée pour l'intervalle de confiance de prédiction.

a. Intervalle de prédiction percentile :

La construction de l'intervalle de prédiction percentile part toujours des réplications bootstrap du coefficient de régression. La prédiction de $\bar{Y}$ pour des nouvelles observations ainsi que les nouvelles variables explicatives sont utilisées pour générer b échantillons bootstrap $\left\{[\Xi], \bar{\Psi}^*\right\}$ avec $\bar{\Psi}^* = \bar{\Psi} + \bar{E}^*$

si [Ξ] et $\bar{\Psi}$ représentent respectivement les variables prédictrices et réponses pour les nouvelles observations.

Les b réplications bootstrap sont ensuite utilisées pour calculer b estimations bootstrap de $\bar{\Psi}^*$: $\hat{\Psi}^{*m} = [\Xi]\bar{\beta}^{*m}$

Les b estimations bootstrap de l'erreur de prédiction $\overline{E}^{*} = \hat{\Psi}^{*m} - \overline{\Psi}^{*m}$ donnent enfin une estimation de la fonction de distribution de l'erreur et sont utilisées pour construire l'intervalle de prédiction. Soit G* cette fonction de distribution. Pour un niveau de confiance (1-2α), les quantiles $G^{*-1}(\alpha)$ et $G^{*-1}(1-\alpha)$ de cette fonction sont utilisées pour construire un intervalle de prédiction percentile qui est de la forme :

$$\left[\hat{\Psi} + G^{*-1}(\alpha), \hat{\Psi} + G^{*-1}(1-\alpha)\right] \tag{2.48}$$

Ces quantiles s'obtiennent facilement en faisant une liste ordonnée des estimations de l'erreur de prédiction. Il suffit alors de prendre le αr-ème et le (1-α)r-ème valeurs de cette liste.

b. Intervalle de prédiction percentile-t :

L'approche percentile donne des résultats satisfaisants pour les échantillons à distribution normale ou dont la taille n est élevée. Une autre approche qui est la méthode percentile-t est recommandée dans les cas ne satisfaisant pas ces conditions. Comme pour les coefficients de régression, on construit, pour chacune des b réplications de l'erreur de prédiction, la variable $\overline{z}^{*m} = \dfrac{\overline{E}^{*m}}{s(\overline{E}^{*m})}$ où $\overline{E}^{*m} = \hat{\Psi}^{*m} - \overline{\Psi}^{*m}$

$s(\overline{E}^{*m})$ étant l'écart type estimé des $\overline{E}^{*m}$.

$$s^2(\overline{E}^{*m}) = \sigma^{2*m}\left[[I] - [T] - ([T]^t[T])^{-1}[T]^t\right]$$

$$\sigma^{2*m} = \frac{1}{n-h}\left\|\overline{\Psi}^{*m} - \hat{\Psi}^{*m}\right\|^2$$

L'intervalle de prédiction percentile-t s'écrit alors :

$$\left[\hat{\Psi} - s(\overline{E}).\overline{z}^{-*\lceil(1-\alpha)r\rceil}, \hat{\Psi} - s(\overline{E}).\overline{z}^{-*\lfloor\alpha r\rfloor}\right] \tag{2.49}$$

Chapitre 3

Utilisation de la méthode PLS en analyse quantitative en spectroscopie par fluorescence X

3.1. INTRODUCTION :

En spectroscopie XRF, la méthode de quantification utilisant les paramètres fondamentaux que nous avons présentée dans le paragraphe §1.1.2.2. reste la plus fiable et la plus flexible en ce qui concerne les méthodes mathématiques. Il y a pourtant un inconvénient important de cette méthode et aussi des autres méthodes classiques (standard interne, addition de standard,…) : elles nécessitent beaucoup d'interventions de l'opérateur. Ces méthodes ne peuvent être utilisées que par un opérateur expérimenté en analyses qualitative et quantitative. Ces méthodes demandent aussi beaucoup de temps pour passer du spectre aux concentrations des éléments étudiés. En effet, il faut tout d'abord commencer par un logiciel d'évaluation de spectre pour avoir les intensités (ou aires nettes) des pics des éléments d'intérêt et c'est cette étape qui est la plus longue. Ensuite, ces intensités sont converties en concentrations par un modèle quantitatif donné.

Or actuellement, la technique de la spectroscopie XRF n'est plus limitée aux mesures en laboratoire. Les spectromètres portables sont de plus en plus utilisés pour des mesures sur terrain. Pendant ces mesures, les résultats doivent être obtenus le plus vite possible et avec le minimum d'intervention de la part de l'utilisateur.

La méthode de quantification idéale pour ces applications mais aussi pour les mesures en laboratoire serait donc une méthode qui peut transformer directement le spectre obtenu du spectromètre en concentrations des éléments d'intérêt sans intervention de l'opérateur.

Nous avons démontré dans le chapitre précédent que la méthode de régression PLS peut être la solution à ce problème. En effet, cette technique est basée sur l'existence d'une

relation linéaire entre le spectre et les concentrations des éléments étudiés. Il suffit ainsi de calculer le coefficient de la relation de régression et d'en tirer les concentrations à partir du spectre.

La figure 3.1 illustre l'avantage de la méthode de quantification par régression PLS par rapport aux méthodes classiques en spectrométrie XRF. Cette figure schématise en effet les étapes à suivre pour extraire les concentrations des éléments étudiés à partir d'un spectre obtenu par un spectromètre XRF pour les deux méthodes.

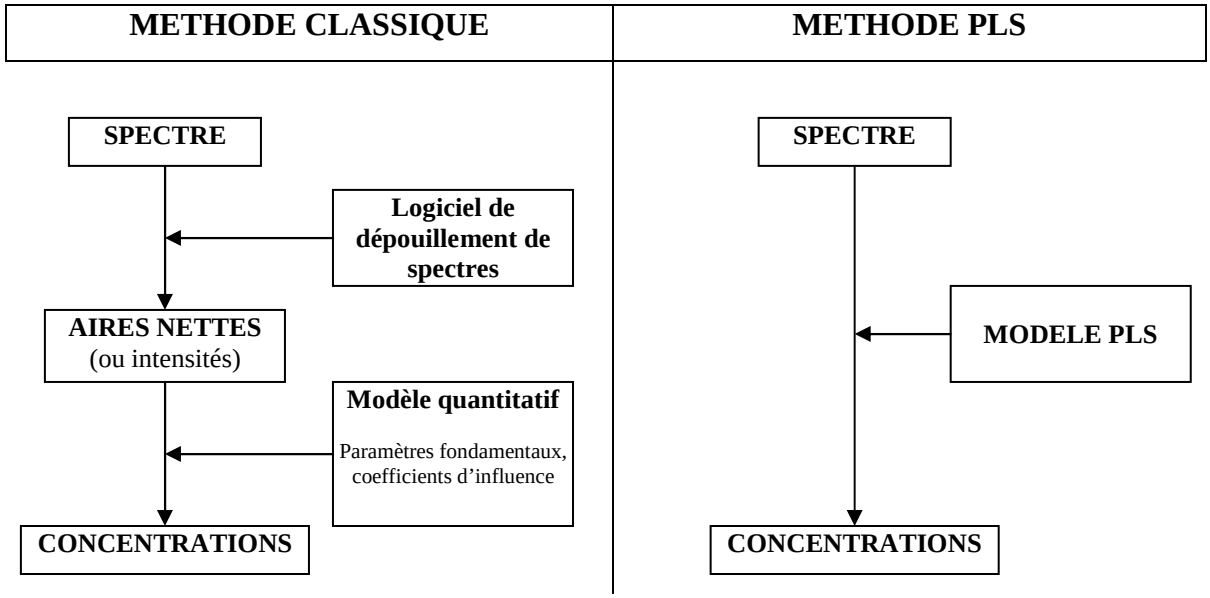


Figure 3.1 : Comparaison de la méthode classique et de la méthode PLS pour la quantification en spectrométrie XRF

La méthode PLS est une méthode directe et ne nécessite pratiquement pas d'intervention de la part de l'utilisateur. Un modèle PLS doit néanmoins être mis au point à l'avance pour des conditions de mesure et d'échantillonnage bien déterminées. Ce modèle est aussi construit pour faire uniquement l'analyse d'un ou plusieurs éléments bien définis à l'avance.

3.2. MISE AU POINT DU MODELE :

3.2.1. Principe :

Pour les mêmes conditions d'échantillonnage et de mesures, un changement de concentration d'un élément se reflète par un changement des intensités correspondantes dans le spectre. Le principe de la méthode PLS est de détecter ces variations au sein du spectre en faisant varier les concentrations des éléments. On pourra ainsi déterminer un modèle représentant la relation entre le spectre d'un échantillon et les concentrations de ses éléments constitutifs. Dès que ce modèle sera mis au point, le spectre suffira pour déterminer les concentrations des éléments d'un échantillon inconnu pour les mêmes conditions de mesure et d'échantillonnage. Comme on l'a décrit dans le chapitre 2, ce modèle est de la forme $[Y]=[X][\beta]+[E]$, où [X] représente le spectre, [Y] les concentrations, $[\beta]$ le coefficient de régression et [E] le résidu.

Les étapes suivantes sont nécessaires pour la mise au point de ce modèle.

3.2.2. Choix des échantillons standard :

Pour calculer le coefficient de régression $[\beta]$, on utilise des échantillons standard, c'est-à-dire des échantillons dont les concentrations des éléments sont bien connues à l'avance. Ces concentrations vont former la matrice [Y] alors que les spectres respectifs des échantillons donnent [X].

Le choix des concentrations des constituants pour les échantillons standard est crucial pour le modèle PLS. En effet, la capacité de prédiction du modèle dépend essentiellement de son aptitude à détecter les changements dans les spectres en corrélation avec les changements des concentrations de chaque élément des échantillons. La condition suivante doit ainsi être remplie pour les échantillons standard :

il faut éviter que les concentrations de deux ou plusieurs éléments changent en même temps dans le même sens. En d'autres termes, les éléments ne doivent pas rester en même proportion d'un échantillon à un autre. En effet, si de tels cas se produisent, le modèle détectera une seule variation au lieu de détecter les variations individuelles des concentrations. Une bonne pratique à appliquer est alors :

- soit d'attribuer les concentrations au hasard
- soit de faire changer la concentration d'un élément en gardant les autres constantes et de procéder ainsi de suite pour chaque élément.

Cette dernière méthode est la meilleure bien que nécessitant un nombre d'échantillons croissant en fonction du nombre d'éléments à étudier.

3.2.3. Prétraitement des données :

Le traitement des données avant la construction du modèle est très important en régression PLS. Il permet en effet d'éliminer les variations statistiques qui ne sont d'aucune utilité pour le modèle. Ces variations sont par exemple les incertitudes des comptages pour chaque canal ou les erreurs dues à la résolution en énergie pour chaque canal. Le but du prétraitement est ainsi de focaliser le modèle sur les variables qui sont vraiment pertinentes. Il existe en général deux grands types de prétraitement des données qu'on utilise en régression PLS : le centrage à la moyenne et la division par la variance. Le centrage à la moyenne permet d'avoir une distribution symétrique pour chaque colonne. Le deuxième traitement sert quant à lui à avoir les variables à variance uniforme. Cette variance est égale à l'unité pour la division par la variance. En spectroscopie XRF, comme le contenu de chaque canal (chaque colonne de [X]) suit une loi de Poisson, la variance est égale à la racine carrée de ce contenu. La racine carrée de chaque canal donne donc une variance de ½. La prise de la racine carrée permet donc également d'avoir une uniformisation de la variance.

3.2.4. Lissage des données :

Le lissage des données est une technique largement utilisée en traitement des données spectroscopiques. Cette méthode permet en effet d'éliminer les fluctuations statistiques entre des canaux voisins. Ces petites fluctuations peuvent nuire à la qualité du modèle surtout dans la détection des éléments à basses concentrations. Leur élimination permet ainsi de mettre en évidence les variations qui présentent vraiment la présence d'un élément au sein de l'échantillon. La méthode de lissage que nous utilisons dans ce travail est la méthode de filtrage de Savitsky et Golay [23]. Cette méthode est basée sur le fait qu'on

peut modéliser chaque intervalle d'un ensemble de données par un polynôme à condition que l'intervalle soit assez petit.

3.2.5. Etalonnage :

La construction proprement dite du modèle PLS s'appelle aussi étalonnage. Le résultat attendu à la fin de cette étape est le coefficient de régression [β]. Les échantillons standard sont mesurés par un spectromètre XRF pour avoir leurs spectres respectifs.

Supposons qu'on a n échantillons. Le spectre d'un échantillon i (i=1..n) sera représenté par le vecteur $\overline{X}_i = \begin{pmatrix} X_i^1 \\ ... \\ X_i^p \end{pmatrix}$ où p est le nombre de canaux du spectromètre.

La matrice représentant les spectres pour tous les échantillons est donc de la forme [X] (n,p)

$$[X] = \left(\overline{X}_1,...,\overline{X}_n\right) = \begin{pmatrix} X_1^1 & ... & X_n^1 \\ ... & ... & ... \\ X_1^p & ... & X_n^p \end{pmatrix}$$

Si on a q éléments à étudier, les concentrations pour un échantillon i seront représentées par le vecteur $\overline{Y}_i = \begin{pmatrix} Y_i^1 \\ ... \\ Y_i^q \end{pmatrix}$

La matrice des concentrations est donc [Y] (n,q)

$$[Y] = \left(\overline{Y}_1,...,\overline{Y}_n\right) = \begin{pmatrix} Y_1^1 & ... & Y_n^1 \\ ... & ... & ... \\ Y_1^q & ... & Y_n^q \end{pmatrix}$$

A partir de ces deux matrices [X] et [Y], on extrait les composantes PLS $\overline{W}, \overline{U}, \overline{C}, \overline{T}, \overline{P}$ comme décrites dans le paragraphe §2.5.1.2. Le nombre des composantes est aussi

déterminé selon les critères exposés au paragraphe §2.5.3.2. ou manuellement par l'observation de la variation des quantités comme le RMSEC, RMSEP en fonction de ce nombre de composantes.

Le coefficient est à la fin obtenu par l'équation (2.35)

3.2.6. Prédiction :

Après avoir mis au point le modèle, on pourra déterminer maintenant les concentrations des éléments d'intérêt pour un échantillon inconnu.

Soit $\overline{\Xi}$ le spectre de cet échantillon et $\hat{\psi}$ la prédiction des concentrations. On a : $\hat{\psi} = \overline{\Xi}\overline{\beta}$

La détermination des concentrations à partir du spectre, c'est-à-dire l'analyse quantitative, se résume donc à une simple multiplication matricielle ou un produit scalaire. L'appréciation des incertitudes sur les prédictions est détaillée dans le paragraphe §2.5.4.

Chapitre 4

Conception du logiciel de quantification par régression PLS : X-PLS v1.0

4.1. LE LOGICIEL X-PLS V1.0:

Actuellement, il existe plusieurs logiciels implémentant la méthode PLS. On peut citer : SIMCA-P (de Umetri AB), The Unscrambler (de Camo AS) comme exemples. Il y a aussi les modules PLS pour les logiciels comme MatLab (de Math Works Inc.) ou Microsoft Excel. Ces logiciels présentent une implémentation générale de la méthode PLS, c'est-à-dire qu'ils permettent la saisie à l'entrée d'un tableau de prédicteurs [X] quelconque et aussi des réponses [Y] correspondant. En spectroscopie XRF, ces entrées ont des formats spécifiques (SPE, MCA,…) et on a donc besoin d'un programme spécial pouvant convertir ces formats en données numériques qu'on peut utiliser.

La sélection des variables prédicteurs à utiliser lors de l'étalonnage est très importante. En spectroscopie XRF, cette sélection est principalement visuelle. Les logiciels susmentionnés n'ont pourtant pas cette capacité de visualisation des variables d'entrée.

Le logiciel que nous proposons est donc adapté aux spécificités de l'utilisation de la méthode PLS en spectroscopie XRF. Il utilise en effet le format SPE pour l'entrée des spectres. Chaque spectre pourra être visualisé pour pouvoir sélectionner la plage de canaux à utiliser. La pertinence de chaque échantillon pour l'étalonnage pourra être évaluée quant à elle par l'analyse visuelle de la graphe $\overline{T} = f(\overline{U})$. Enfin, comme l'utilisation du PLS en spectroscopie XRF est encore en plein développement, il est intéressant pour un centre de recherche de développer ses propres codes et de les adapter aux améliorations probables au sein de l'utilisation de cette méthode.

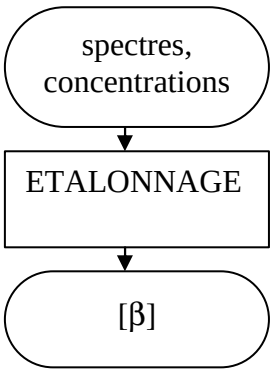
Nous avons entièrement utilisé l'outil de développement C++Builder v3 de Borland pour la réalisation du logiciel X-PLS v1.0.

4.2. CONCEPTION GENERALE :

Le logiciel est composé de deux modules principaux dont les rôles sont :

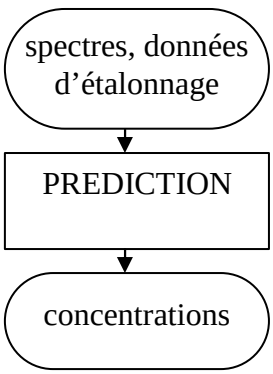
- l'étalonnage
- la prédiction

Etalonnage :



A l'entrée, l'utilisateur introduit les fichiers contenant les spectres XRF des échantillons standard et les concentrations des éléments d'intérêt. La sortie principale est le coefficient de régression [β].

Prédiction :



Lors de prédiction d'échantillons inconnus, à part les spectres XRF de ces échantillons, l'utilisateur charge les données d'étalonnage enregistrées dans des fichiers. Le résultat du processus de prédiction est bien sûr les concentrations des différents éléments d'intérêt.

4.3. CONCEPTION DETAILLEE :

4.3.1. Etalonnage :

Dans ce paragraphe, nous allons détailler tous les modules constituant la procédure d'étalonnage. Ces étapes (modules) sont montrées par le diagramme suivant.

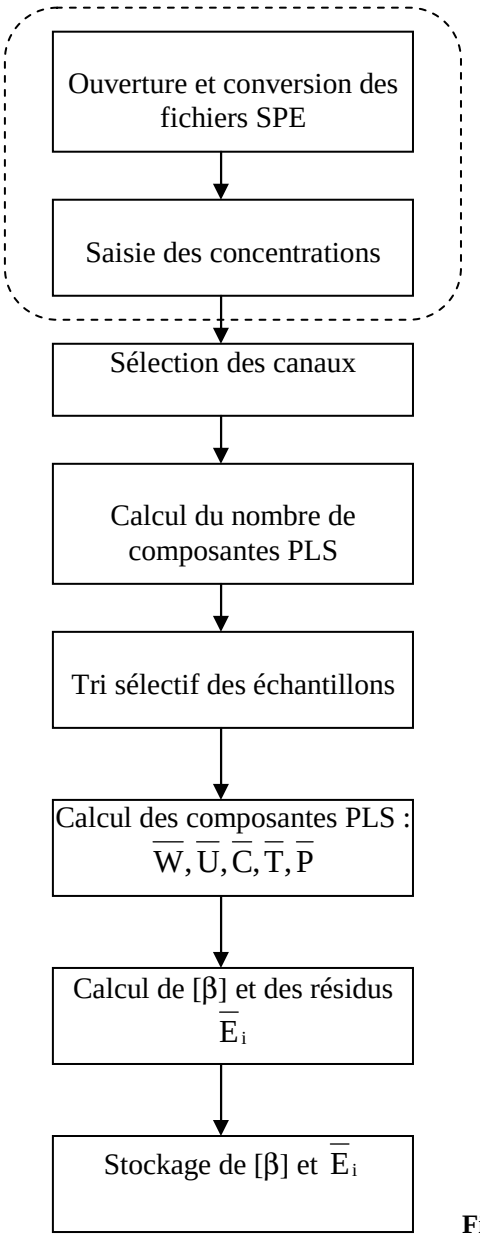


Figure 4.1 : Algorithme de l'étalonnage

4.3.1.1. Sélection des canaux :

La visualisation des spectres permet de délimiter la plage des canaux à utiliser lors de l'étalonnage. En effet, pour les éléments bien déterminés, on peut connaître les emplacements exacts de leurs pics caractéristiques sur le spectre. En ne sélectionnant que les zones contenant ces pics, on augmente la pertinence des variables prédictrices qui sont les contenus des canaux (les nombres de coups).

Dans notre logiciel, ce processus est effectué à l'aide de la fenêtre suivante :

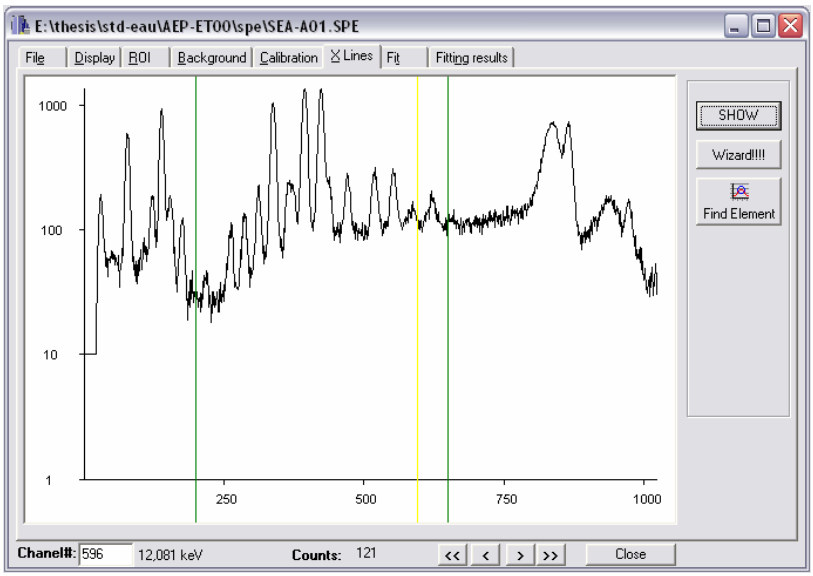


Figure 4.2 : Fenêtre de visualisation du spectre et sélection de plage de canaux

A l'aide du curseur, on définit les limites inférieure et supérieure de la plage de canaux à utiliser. Pour faciliter la visualisation, l'utilisateur peut avoir recours à un étalonnage en énergie. L'énergie correspondant à la position du curseur est ainsi affichée à côté du numéro de canal. En validant la sélection d'une plage par la touche APPLY, les numéros des canaux inférieur et supérieur sont affectés aux variables ch_beg et ch_end.

4.3.1.2. Réduction des variables :

Soit $[X] = [\overline{X}_1, ..., \overline{X}_p]$ où $\overline{X}_i = \begin{pmatrix} X_i^1 \\ ... \\ X_i^n \end{pmatrix}$

On forme le vecteur des moyennes comme suit : $\overline{X}_m = \begin{pmatrix} X_m^1 \\ ... \\ X_m^p \end{pmatrix}$ où $X_m^i = \frac{1}{n}\sum_{j=1}^{n} X_j^i$

On calcule par la suite les variances :

$$\sigma_{xi}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j^i - X_m^i)^2 \quad i=1,\ldots,p$$

Ainsi, au lieu d'utiliser la variable $\overline{X}_i$, on utilise la variable réduite $\frac{\overline{X}_i - X_m^i \bar{I}}{\sigma_{xi}^2}$

De la même manière, les $\overline{Y}_i$ sont réduites pour avoir $\frac{\overline{Y}_i - Y_m^i \bar{I}}{\sigma_{yi}^2}$

Une autre option qu'on utilise dans le cas des données spectroscopiques est la prise de la racine carrée au lieu de la division par la variance. Dans ce cas, les variables réduites sont de la forme :

$$\sqrt{\overline{X}_i - X_m^i \bar{I}} \text{ et } \sqrt{\overline{Y}_i - Y_m^i \bar{I}}$$

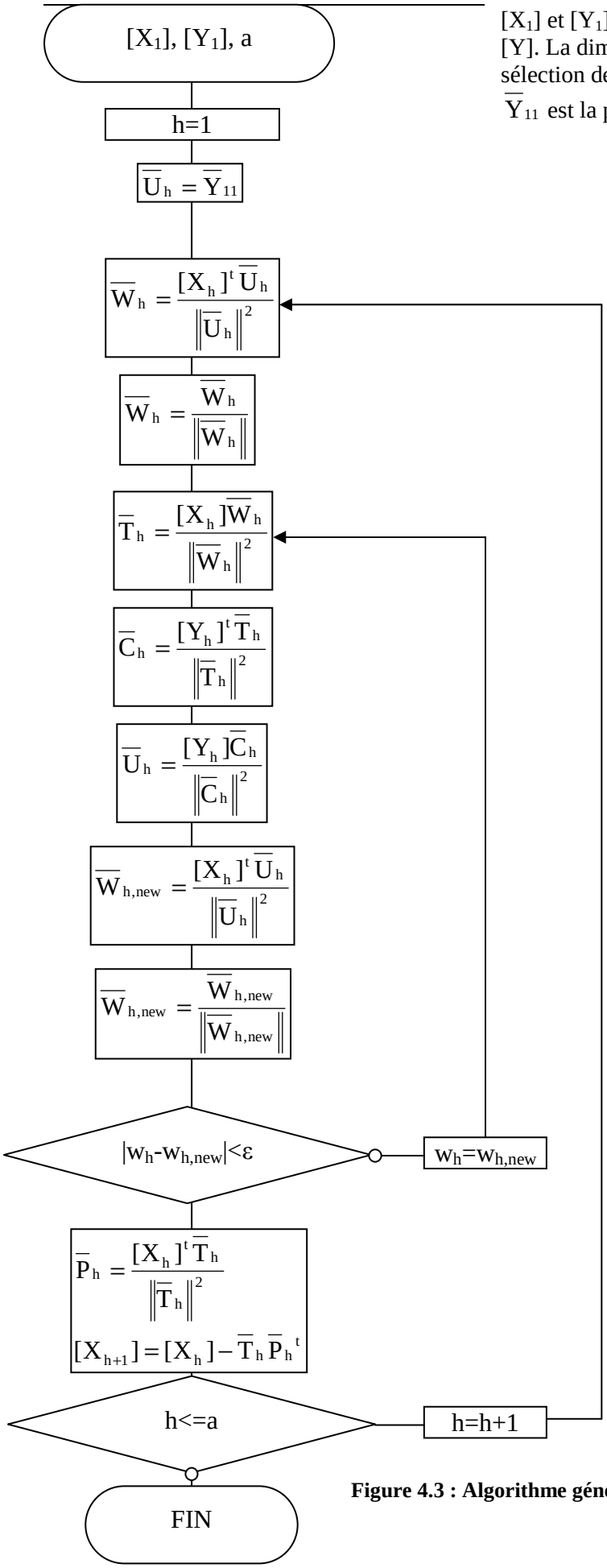
Ce sont donc ces deux variables qui seront utilisées pour le calcul des composantes dans l'étape suivante.

4.3.1.3. Calcul des composantes PLS :

Le calcul des composantes PLS constitue le cœur du processus d'étalonnage du modèle PLS. C'est en effet ces quantités qui détermineront :

- le coefficient de régression [β]
- la pertinence de chaque échantillon
- la fiabilité du modèle (valeur du PRESS)

Le diagramme suivant montre le processus suivi pour l'obtention de ces composantes.



$[X_1]$ et $[Y_1]$ sont les formes réduites de [X] et [Y]. La dimension de [X] étant définie par la sélection des canaux (ch_end – ch_beg)

$\overline{Y}_{11}$ est la première colonne de $[Y_{1]}$

Figure 4.3 : Algorithme général pour le calcul des composantes PLS

4.3.1.4. Calcul du nombre de composantes par validation croisée :

Le choix du nombre des composantes PLS est une étape cruciale. En effet, si trop peu de composantes sont prises, des changements importants au sein du spectre peuvent être négligés par le modèle. C'est par exemple le cas des changements apportés par les éléments à très basses concentrations. S'il est par contre trop élevé, le modèle tend à trop s'ajuster aux échantillons d'étalonnage et son pouvoir de prédiction pour les échantillons inconnus est ainsi faible.

La méthode que nous utilisons dans ce logiciel pour déterminer le nombre optimal de composantes est la validation croisée « leave one out » (LOO-CV). La figure suivante montre l'algorithme suivi.

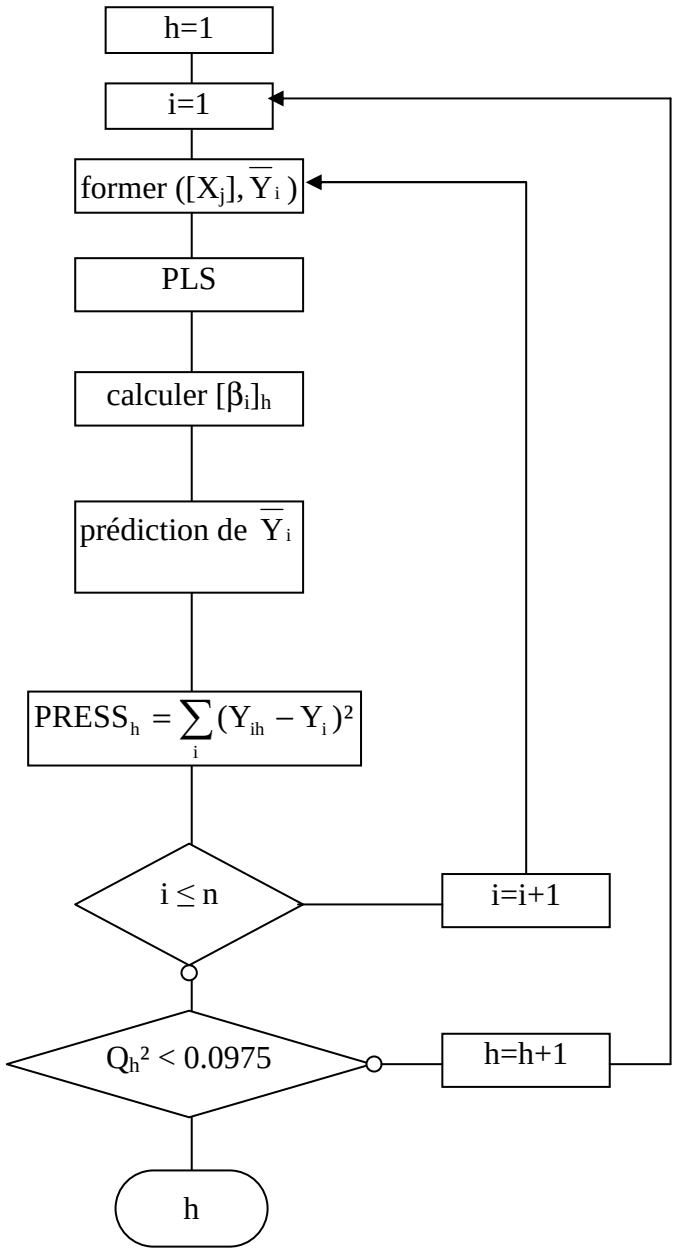


Figure 4.4 : Détermination automatique du nombre optimal de composantes PLS

4.3.1.5. Tri sélectif des échantillons :

Ce module a pour but d'éliminer les échantillons aberrants éventuels. De tels échantillons peuvent provenir :

- de la contamination de l'échantillon
- du comportement anormal des appareils lors des mesures
- de l'erreur lors de la préparation de l'échantillon

La méthode que nous utilisons dans ce logiciel pour éliminer ces échantillons est l'observation du graphe $\overline{T} = f(\overline{U})$. Normalement, si le modèle est fiable, une telle représentation graphique donnerait approximativement une ligne droite indiquant une relation de prédiction entre [X] et [Y]. Les échantillons qui dévient significativement de cette ligne sont donc identifiés comme aberrants et sont éliminés.

4.3.1.6. Calcul du coefficient de régression et des résidus :

La matrice représentant le coefficient de régression est calculée suivant l'équation (2.35)

Pour chaque échantillon i, le résidu est donné par $\overline{E}_i = \overline{Y}_i - \overline{X}_i[\beta]$ où $\overline{X}_i$ est le spectre de l'échantillon i.

4.3.1.7. Formats des données :

La plupart des données traitées par le programme sont des matrices. Ces données sont stockées dans des fichiers avant et après le traitement par le logiciel. Nous allons détailler dans ce paragraphe les structures des fichiers ainsi que les procédures pour leur extraction et leur stockage.

A l'entrée :

A l'entrée, à part les concentrations des éléments d'intérêt qui sont saisies à l'aide du clavier, on a comme données les spectres. Ils sont extraits à partir de fichiers de type SPE qui est la norme standard de l'Agence Internationale de l'Energie Atomique AIEA. Le fichier SPE est un fichier texte dont le contenu est le suivant :

```
$SPEC_ID:
Target
$DATE_MEA:
10-27-1993  12:49:34
$MEAS_TIM:
       200       209
```

Identificaton du spectre

```
$DATA:
          0      1023
```

Nombre de canaux. Ici: 1024 canaux.
0 à 1023

```
0     0     0     0     0     0     0     0     0     0
0     0     0     0     0     0     0     0     0     0
0     0     0     0     0     0     0     0     0     2
14    17    13    9     14    14    14    17    14    16
14    10    15    18    21    22    19    9     12    14
12    12    13    17    16    12    9     12    12    7
11    11    7     6     13    11    11    6     5     9
7     5     6     9     11    8     8     7     9     12
5     3     7     7     5     9     9     7     6     7
3     8     5     7     9     3     7     3     4     7
6     8     5     5     7     8     6     5     8     5
6     8     5     6     5     6     6     3     7     3
3     3     6     6     6     4     5     3     5     2
8     9     8     5     6     8     9     3     0     8
4     4     4     7     3     4     3     8     3     3
7     4     2     6     6     2     7     6     6     7
5     2     5     3     7     6     6     5     3 .....
```

Données: série de 1024 valeurs. Chaque valeur représente le nombre de coups et sa position dans la matrice représente le numéro de canal.

Figure 4.5 : Format d'un fichier SPE

Pour la lecture d'un tel fichier, nous avons écrit la procédure OpenSPEfile( ). Le fonctionnement de cette procédure est très simple. Elle lit en effet chaque ligne du fichier à partir de la ligne de données et convertit les textes en nombres entiers qui sont stockés dans une variable tableau tampon. Cette variable tampon est vidée à la fermeture du fichier et son contenu est transféré dans les récipients appropriés, en l'occurrence [X] (cf. Annexes A.2).

A la sortie :

A la fin d'un processus d'étalonnage, on obtient essentiellement les données suivantes : le coefficient de régression [β] et le nombre de composantes. Mais il existe aussi d'autres données importantes et qui sont propres à un étalonnage donné comme ces deux paramètres. Effectivement, ces deux paramètres ne suffisent pas pour la prédiction des échantillons inconnus. Comme les données ont été centrées et réduites, les moyennes et les variances de [X] et [Y] sont nécessaires pour le calcul des concentrations. Les informations suivantes sont aussi nécessaires pour les utilisations d'un modèle donné :

- nombre et noms des éléments d'intérêt
- le nombre d'échantillons standard
- le nombre de composantes PLS
- la plage de canaux utilisée
- les résidus e_i qui seront utilisés pour l'estimation des intervalles de prédiction

Pour faciliter la manipulation de ces données, nous les avons regroupées au sein d'une variable objet de type cCalib (cf. Annexes A.1). La définition d'un tel type d'objet est la suivante en C++ :

```
class nom_du_type
{
public :
liste des membres de l'objet ;
} ;
```

Dans notre cas, nous avons créé une instance de cette classe que nous avons nommé CalibData. C'est donc cet objet CalibData qui sera écrit dans un fichier de type binaire.

La lecture et l'écriture à partir et dans ces fichiers sont effectuées à l'aide des méthodes de **fstream.h** de C++.

4.3.2. Prédiction :

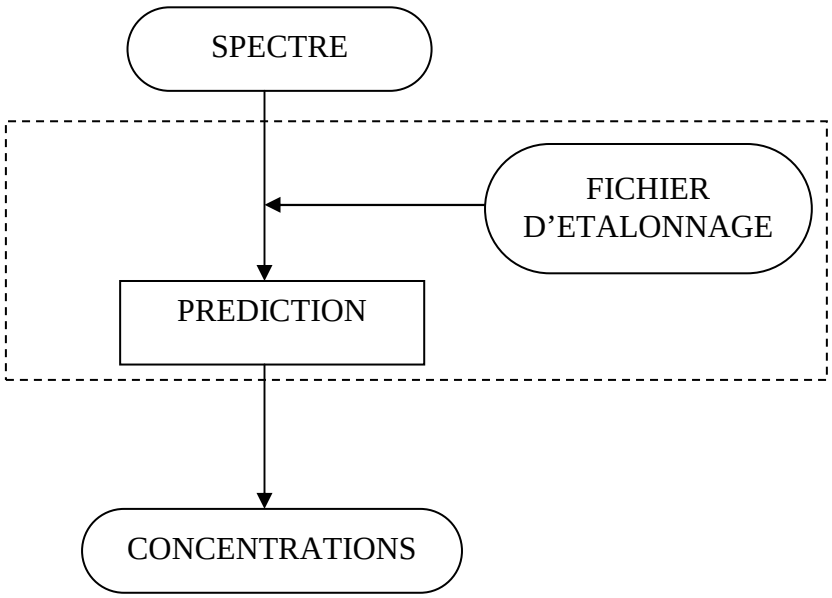


Figure 4.6 : Diagramme du processus de prédiction

La prédiction ou analyse quantitative peut être effectuée dans les cas suivants :

au cours d'un étalonnage : la procédure doit être dans ce cas exécutée avec succès

après le chargement d'un fichier d'étalonnage approprié

Dans tous ces cas, le logiciel utilise la variable booléenne CanPredict pour déterminer si la procédure d'étalonnage peut être effectuée ou non.

Pour le deuxième cas, l'utilisateur choisit tout d'abord le fichier d'étalonnage approprié. Ce choix est basé sur les éléments d'intérêt, les conditions de mesure et d'échantillonnage. Le logiciel ouvre alors le fichier choisi et charge ainsi les données dans l'objet CalibData.

Avant le traitement proprement dit du spectre, c'est-à-dire sa multiplication avec le coefficient de régression [β], on doit tout d'abord procéder à sa réduction. On utilise pour ceci les moyennes enregistrées dans CalibData à partir du fichier d'étalonnage. Chaque

colonne du spectre est diminuée de ces moyennes et divisée ensuite par les variances ou on prend la racine carrée. Si un spectre inconnu est représenté par le vecteur $\overline{\Xi}_i = \begin{pmatrix} \Xi_i^1 \\ ... \\ \Xi_i^p \end{pmatrix}$, un élément j devient après réduction :

$$\Xi_i^j = \frac{\Xi_i^j - \text{meanX}_j}{\text{VarX}_j} \text{ ou } \Xi_i^j = \sqrt{\Xi_i^j - \text{meanX}_j}$$

où meanXj et VarXj sont respectivement la moyenne et la variance de la colonne j de [X]

Pour avoir la valeur de la concentration, on doit encore tenir compte de la réduction qu'a subi $\overline{\Xi}_i$. Ainsi, une prédiction de la concentration est donnée par :

$$\hat{\psi} = \overline{\Xi}_i \overline{\beta}.\text{VarY} + \text{MeanY}$$

ou

$$\hat{\psi} = (\overline{\Xi}_i \overline{\beta})^2 + \text{MeanY}$$

avec VarY et MeanY représentant respectivement la variance et la moyenne des concentrations lors de l'étalonnage pour l'obtention de $\overline{\beta}$.

Une autre option est aussi présentée par le logiciel en ce qui concerne la prédiction. Cette option est appelée prédiction complète (full prediction) par opposition à la simple multiplication par le coefficient de régression qui est communément appelée prédiction rapide (short prediction). Cette méthode utilise les vecteurs PLS $\overline{W}$ et $\overline{Q}$. Nous pouvons en effet décomposer la matrice des spectres comme suit : $\hat{T} = \Xi\overline{W}$. La prédiction des concentrations est ensuite donnée par $\hat{\Psi} = \hat{T}\overline{Q}$ (cf. Chap 2.) en appliquant les mêmes réductions de variables que précédemment.

4.4. LES INTERFACES GRAPHIQUES :

Le logiciel X-PLS est une application avec interface à document multiple (MDI). Il est ainsi constitué d'une fenêtre principale comprenant un menu et une barre d'outils et des fenêtres filles pour les modules suivants :

- étalonnage
- validation du modèle
- prédiction

4.4.1. La fenêtre principale :

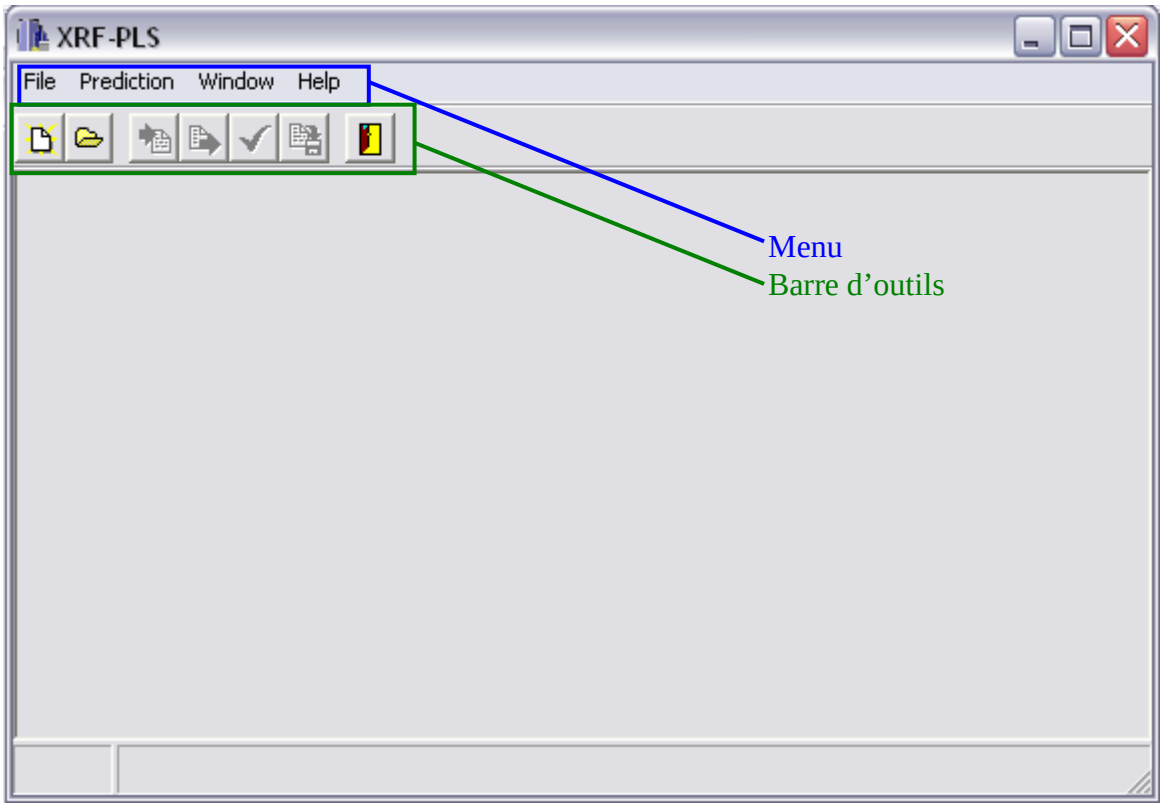
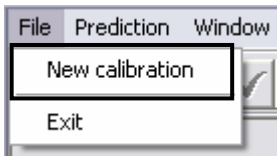


Figure 4.7 : La fenêtre principale du logiciel X-PLS

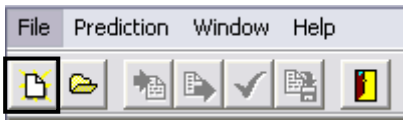
A partir de la fenêtre principale, on peut soit commencer une nouvelle série d'étalonnages soit effectuer la prédiction d'échantillons inconnus. Chacune de ces deux tâches peut être effectuée en utilisant soit le menu soit les boutons de la barre d'outils.

4.4.2. Etalonnage :

Menu :



Barre d'outils :



4.4.2.1. La fenêtre d'étalonnage :

En démarrant le processus d'étalonnage à l'aide de l'une des deux méthodes ci-dessus, on obtient à l'écran la fenêtre d'étalonnage. La fonction principale de cette fenêtre est l'entrée de toutes les données pour effectuer l'étalonnage : les spectres et les différentes concentrations des éléments étudiés pour les échantillons standard. On a pour cela un tableau dont le nombre de lignes et de colonnes varie suivant le nombre d'éléments et le nombre d'échantillons. Une ligne est formée par :

- le numéro de ligne (ou de l'échantillon)
- le nom du fichier contenant le spectre
- les concentrations des éléments (une colonne par élément)

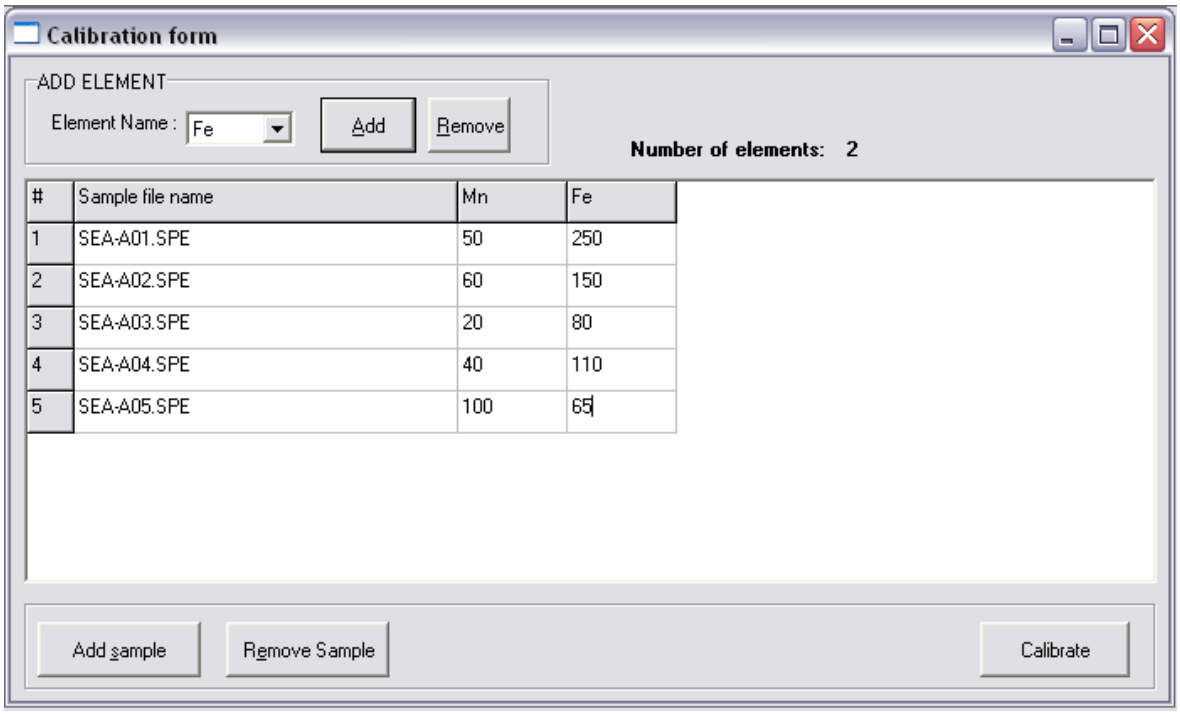
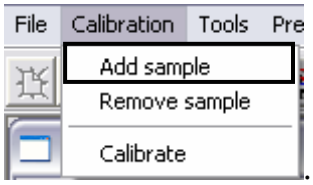


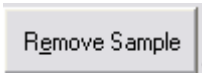
Figure 4.8 : La fenêtre d'étalonnage

4.4.2.2. Insertion d'échantillons :

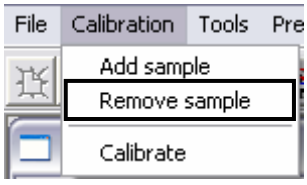
Un échantillon est inséré soit en utilisant le bouton 

, soit à l'aide du menu



. Une boîte de dialogue standard d'ouverture de fichiers est alors affichée. En sélectionnant un fichier (.SPE) puis en cliquant sur Ouvrir, le nom du fichier est ajouté au tableau. On peut enlever un échantillon en sélectionnant la ligne correspondante sur le tableau et en cliquant sur le bouton 

ou par le menu :



4.4.2.3. Les éléments à étudier :

On insère le nom d'un élément en le sélectionnant dans la boîte liste « Element name » et en cliquant sur 

. Une colonne est ajoutée au tableau pour chaque élément inséré. Un élément peut aussi être enlevé du tableau en sélectionnant la colonne correspondante et en pressant le bouton 

.

Quand tous les spectres sont insérés dans le tableau, les concentrations des différents éléments sont insérées dans le tableau à l'aide du clavier. Après avoir saisi toutes ces données, on procède à l'étalonnage proprement dit. Mais avant cela, on doit tout d'abord déterminer les conditions dans lesquelles cet étalonnage va être effectué. La condition la plus importante est la méthode de prétraitement des données. Ce paramètre peut être défini dans la fenêtre des Options qui est accessible via le menu : Tools->Options

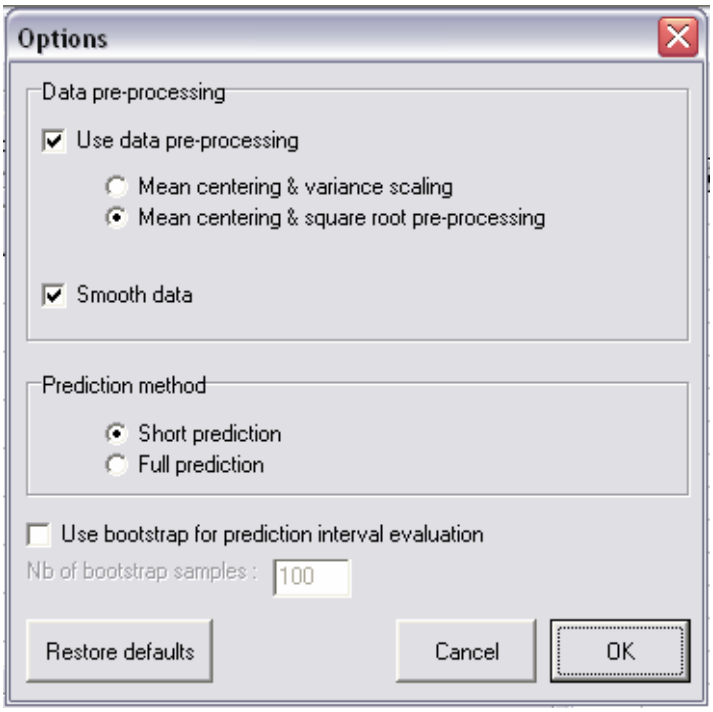
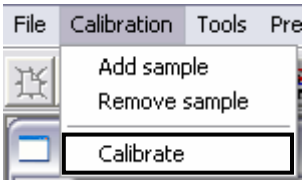


Figure 4.9 : La fenêtre des Options

On peut alors choisir de procéder ou non au prétraitement des données en cochant la case « Use data pre-processing » et les méthodes qu'on peut utiliser sont :

- le centrage par la moyenne suivi de la réduction par la variance

- le centrage par la moyenne et la prise de la racine carrée

Pour procéder à l'étalonnage, on utilise le menu 

ou le bouton 

.

Une autre fenêtre est alors affichée.

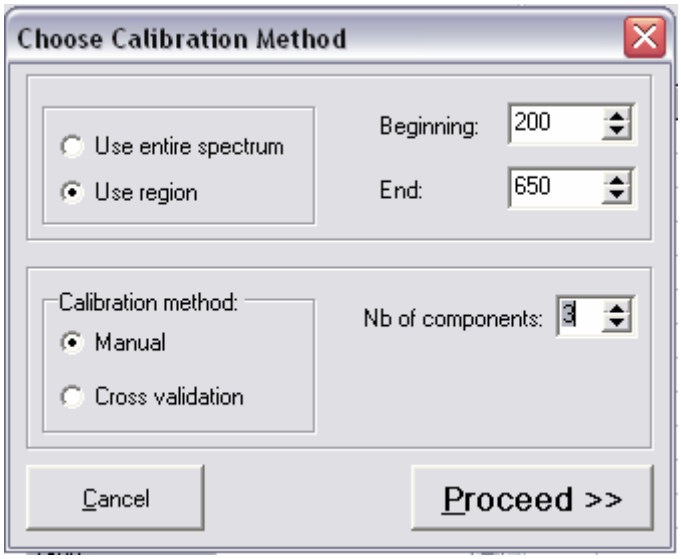


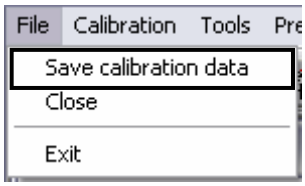
Figure 4.10 : La fenêtre des options d'étalonnage

Cette fenêtre permet de choisir la méthode à utiliser pour l'étalonnage. Les deux choix possibles sont : la méthode manuelle ou la méthode par validation croisée. La méthode manuelle permet de choisir arbitrairement le nombre de composantes PLS à utiliser. L'autre méthode quant à elle utilise la validation croisée et la comparaison des PRESS pour évaluer le nombre optimal de composantes. Le bouton 

lance le processus d'étalonnage qui est formé par les étapes suivantes :

- le calcul des composantes PLS : $\overline{W}, \overline{U}, \overline{C}, \overline{T}, \overline{P}$
- le calcul du coefficient de régression [β]
- le calcul des PRESS et RMSEC

Une boîte de dialogue indique la fin de ces processus. A ce stade, on peut passer à la prédiction d'échantillons inconnus. Il est tout de même intéressant d'enregistrer les données d'étalonnage qu'on vient d'obtenir pour qu'on puisse les utiliser ultérieurement. Le menu et la barre d'outils permettent d'effectuer cette sauvegarde :



ou


Le logiciel demande alors l'introduction des informations concernant les conditions de mesure et d'échantillonnage sans lesquelles l'enregistrement n'est pas possible.

A part les données qui permettent la prédiction ultérieure pour des échantillons inconnus comme le coefficient de régression, l'élément d'intérêt,…, les informations suivantes sont aussi enregistrées dans un fichier d'étalonnage :

- le nom de l'utilisateur et la date d'étalonnage
- les conditions de mesure et d'échantillonnage

Il faut en effet rappeler que les données d'étalonnage ne peuvent être utilisées que pour des échantillons préparés et mesurés dans les mêmes conditions que les échantillons d'étalonnage à partir desquels ces données ont été obtenues.

La procédure que nous venons de décrire suppose que tous les échantillons standard utilisés soient pertinents pour les éléments étudiés, c'est-à-dire qu'il n'y a aucun échantillon aberrant. La présence de tels échantillons détériorerait l'efficacité du modèle construit. Il est donc important de détecter la présence éventuelle de ces échantillons inutiles. L'outil utilisé par le logiciel X-PLS pour cela est le traçage du graphe $\overline{T} = f(\overline{U})$.

4.4.3. La prédiction :

La prédiction des concentrations d'un élément d'intérêt quelconque peut être effectuée dans les deux cas suivants :

- à la fin d'un processus d'étalonnage
- après le chargement d'un fichier d'étalonnage

4.4.3.1. Chargement d'un fichier d'étalonnage :

On peut procéder à une prédiction en chargeant les données enregistrées dans un fichier d'étalonnage. Rappelons que les plus importantes de ces données sont le coefficient de régression, le nom de l'élément d'intérêt, les erreurs d'étalonnage. Ce chargement se fait à l'aide du menu Prediction -> Load Calibration File ou du bouton de la barre d'outils  . Une boîte de dialogue standard d'ouverture de fichiers s'affiche alors et on peut sélectionner les fichiers d'étalonnage qui ont une extension .CAL.

Il est à remarquer que cette procédure de chargement de données d'étalonnage ne peut être effectuée dans le cas où un processus d'étalonnage est encore en cours. Il faut alors fermer la fenêtre d'étalonnage.

4.4.3.2. Options de prédiction :

Les options pouvant influencer la prédiction peuvent aussi être ajustées dans la fenêtre des Options. La première option est la méthode de prédiction à utiliser. On peut choisir soit la prédiction rapide (short prediction) soit la prédiction complète (full prediction) qui sont décrites dans le paragraphe §4.3.2. Il faut noter que la prédiction complète n'est pas accessible dans le cas où on utilise des données d'étalonnage à partir de fichier. Le deuxième paramètre est le choix d'utiliser ou non la méthode bootstrap pour l'estimation de l'intervalle de prédiction. On peut alors choisir le nombre d'échantillons bootstrap à utiliser qui est de 100 par défaut.

Il faut aussi noter que les paramètres concernant le prétraitement des données ne peuvent pas être changés lors de processus de prédiction.

4.4.3.3. Visualisation des données d'étalonnage :

La visualisation des données d'étalonnage est très importante surtout pendant le choix d'un fichier à charger. Il se peut en effet que le nom d'un tel fichier ne donne que très peu d'information concernant ces données.

Ces informations sont les suivantes :

- l'élément d'intérêt
- le nom de l'utilisateur et la date d'étalonnage

- les informations concernant les conditions de mesure et d'échantillonnage comme le type d'échantillons (liquide, solide,…), la méthode de mesure utilisée (TXRF, XRF conventionnel,…), les paramètres du tube à rayons X et du détecteur

Cette visualisation est faite par le menu Tools -> Show actual calibration data. La fenêtre suivante affiche alors ces données :

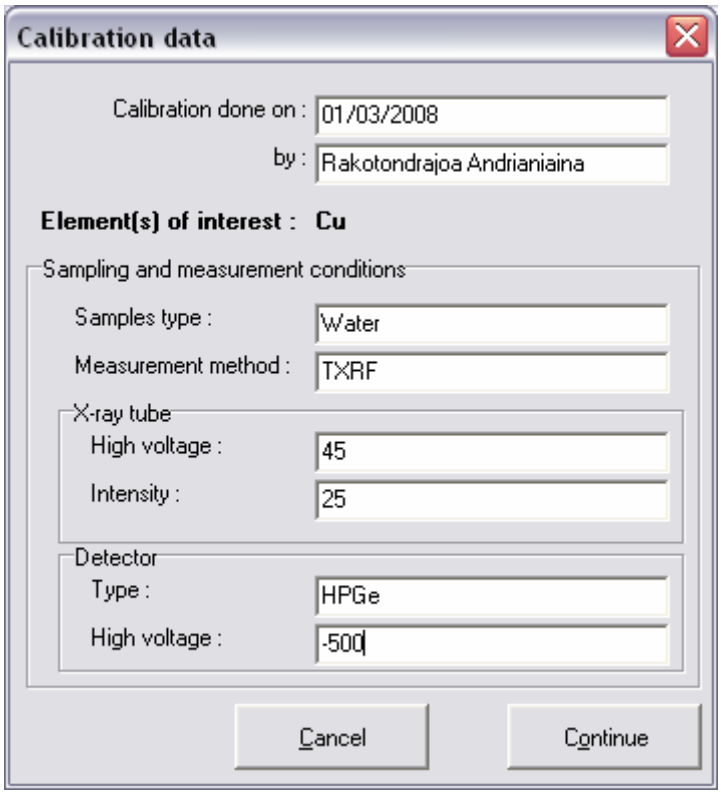
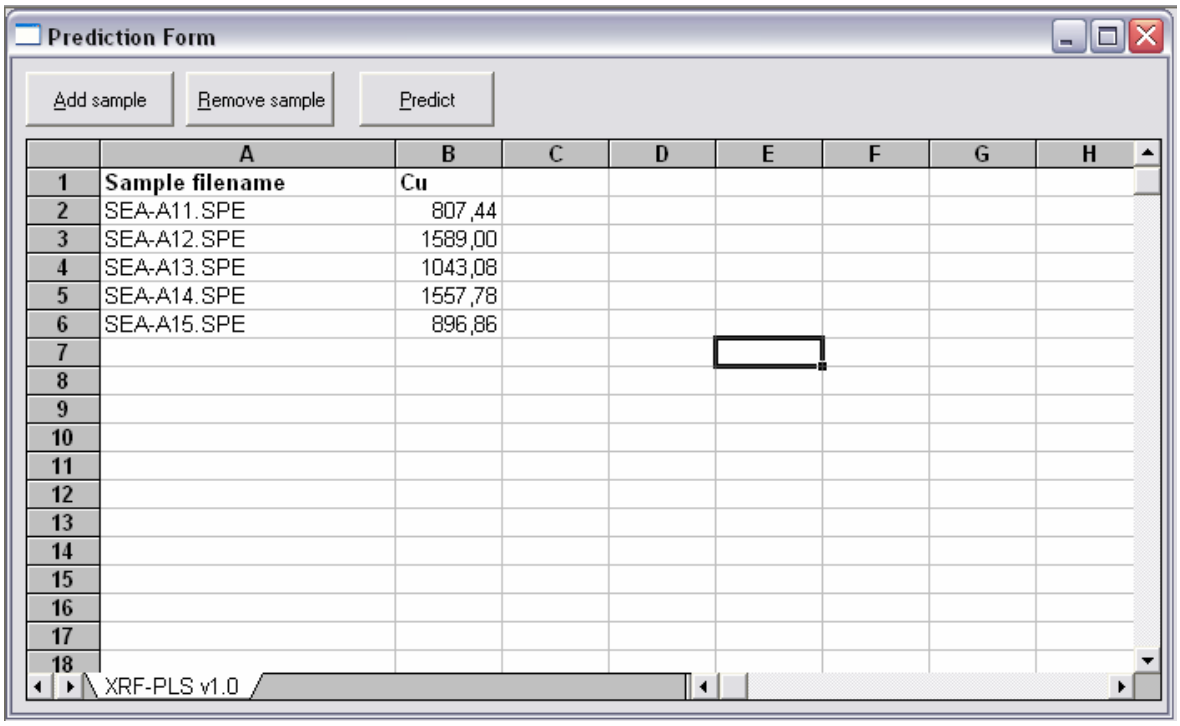


Figure 4.11 : Fenêtre des informations sur l'étalonnage

4.4.3.4. Prédiction :

La prédiction proprement dite s'effectue à l'aide de la fenêtre de prédiction qui est obtenue par le menu Prediction -> Predict Unknown Sample.

Cette fenêtre est composée d'un tableau qui contient les échantillons à étudier et les concentrations des éléments correspondants, de deux boutons pour l'insertion et la suppression d'un échantillon et un autre pour la prédiction des concentrations.



	A	B	C	D	E	F	G	H
1	**Sample filename**	**Cu**						
2	SEA-A11.SPE	807,44						
3	SEA-A12.SPE	1589,00						
4	SEA-A13.SPE	1043,08						
5	SEA-A14.SPE	1557,78						
6	SEA-A15.SPE	896,86						
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								

Figure 4.12 : La fenêtre de prédiction

En appuyant sur le bouton « Add sample », une boîte de dialogue d'ouverture de fichier s'affiche. On peut alors sélectionner le fichier SPE contenant le spectre XRF de l'échantillon à étudier. Le nom de chaque fichier sélectionné est ajouté en bas de la liste des fichiers déjà contenus dans le tableau. On peut enlever un fichier de cette liste en sélectionnant la ligne correspondante et en cliquant sur « Remove sample ». Les concentrations de l'élément d'intérêt pour ces échantillons sont affichées en appuyant sur le bouton « Predict ».

4.4.3.5. Enregistrement des résultats de prédiction :

Les résultats d'un processus de prédiction peuvent être enregistrés dans un fichier au format Microsoft Excel. Ceci est fait à l'aide du menu Prediction ->Save prediction results.

Chapitre 5

Applications à la fluorescence X à réflexion totale

5.1. EXPERIMENTATIONS :

Les séries de mesures que nous avons effectuées dans le cadre de ce travail ont pour buts de:

- tester le fonctionnement du logiciel X-PLS v1.0
- évaluer l'efficacité de la méthode de régression PLS en technique de la fluorescence X à énergie dispersive (ED-XRF), plus particulièrement en TXRF
- comparer les résultats obtenus pour différentes configurations des échantillons d'étalonnage
- déterminer les valeurs les plus appropriées des paramètres intervenant dans la régression PLS comme le nombre de composantes PLS, les plages de variables prédictrices à utiliser selon l'élément à étudier, les techniques de prétraitement des variables à utiliser,…

L'utilisation de la régression PLS en spectroscopie XRF nécessite la préparation de plusieurs échantillons standard pour l'étalonnage. Compte tenu des produits chimiques disponibles au sein du laboratoire, seuls les échantillons liquides peuvent être préparés en nombre important. C'est pour cette raison que nous avons choisi l'application de la régression PLS à des échantillons liquides et les mesures ont été effectuées en utilisant la méthode de la fluorescence X à réflexion totale qui est la plus appropriée pour l'analyse de ce type d'échantillons.

Chaque série de mesures aboutit à l'obtention de modèles PLS pour tous les éléments d'intérêt.

5.1.1. Préparations des échantillons :

Les échantillons d'étalonnage sont des échantillons standard, c'est-à-dire des échantillons dont les concentrations des éléments constitutifs sont bien déterminées. Dans notre cas, ces échantillons contiennent tous les éléments : Cu, Zn, As, Se, Ni. Nous avons utilisé des solutions mères à 10 ppm pour avoir les concentrations voulues pour chaque élément. Le volume total de chaque échantillon est de 5mL.

Pour la mesure en TXRF, 10µL de l'échantillon est prélevé à l'aide d'une micropipette puis déposée sur la surface d'un réflecteur soigneusement préparé à l'avance pour éviter les contaminations. Le tout est ensuite séché sous vide avant d'effectuer la mesure proprement dite à l'aide du spectromètre XRF.

5.1.2. Conditions de mesures :

Toutes les mesures sont effectuées avec la chaîne de spectrométrie XRF de Madagascar-INSTN. Elle comprend :

- le générateur de rayons X Kristalloflex
- le dispositif à réflexion totale qui comprend le collimateur, le réflecteur de séparation et le porte-échantillon
- le détecteur Si(Li) avec son système de refroidissement par azote liquide et le préamplificateur
- le module Canberra 1510 constitué de l'alimentation haute tension du détecteur, de l'amplificateur et du convertisseur analogique – digital (ADC)
- le terminal informatique qui comprend essentiellement le système S-100 pour l'analyseur multicanal (MCA)

Le tube à rayons X fonctionnait avec une tension de 45kV et une intensité de courant à 25mA. Le détecteur était quant à lui polarisé avec une tension de -500V.

5.1.3. Utilisation du logiciel X-PLS :

Nous avons choisi d'utiliser l'algorithme PLS1 pour tous les étalonnages que nous avons effectués dans le cadre de ce travail. Ceci nous permet en effet de nous pencher sur tous les détails des paramètres dont il faut prendre en compte pour l'étalonnage d'un élément.

5.2. SERIE D'ETALONNAGES N°1 :

Pour cette première série de mesures, nous avons préparé 15 échantillons que nous avons divisés en 2 lots. Le premier lot de 10 échantillons est utilisé comme lot d'étalonnage. Cela veut dire que ces 10 échantillons ont été utilisés pour construire les modèles PLS pour tous les éléments à étudier. Le deuxième lot qui comporte 5 échantillons sert quant à lui à tester les modèles obtenus. Le but est donc d'évaluer l'aptitude de ces modèles à prédire les concentrations des éléments d'intérêt dans le lot n°2. C'est en effet l'objectif général d'un étalonnage en PLS : construire des modèles qui pourraient par la suite prédire les concentrations pour des échantillons « inconnus ».

La constitution de chacun des 15 échantillons est donnée par le tableau suivant.

Tableau 5.1 : Constitution des 15 échantillons standard (ppb)

	STD01	STD02	STD03	STD04	STD05	STD06	STD07	STD08	STD09	STD10	STD11	STD12	STD13	STD14	STD15
Ni	240	140	180	250	150	110	190	120	130	200	230	220	170	210	160
As	180	150	220	130	110	240	140	200	190	160	210	230	170	120	250
Se	250	110	170	160	140	230	150	210	200	190	120	240	220	180	130
Cu	2000	700	1200	1900	1300	600	1500	1100	1400	1800	800	1600	1000	1700	900
Zn	1700	1300	600	1200	2000	1500	1400	1800	1000	1100	1600	900	800	1900	700

5.2.1. Etalonnage :

L'étalonnage consiste donc à déterminer le coefficient de régression [β] satisfaisant la relation de régression : $[Y] = [X][\beta]+[E]$

Comme nous utilisons la variante univariée de la régression PLS (PLS1), il nous est nécessaire de déterminer un coefficient de régression [β] pour chacun des éléments d'intérêt.

La matrice [X] est ainsi formée par les spectres XRF des 10 échantillons pour tous les éléments. Elle a donc 10 lignes et 1024 colonnes. [Y] qui est constitué de 10 lignes et d'une colonne contient les concentrations de l'élément à étudier pour les 10 échantillons.

5.2.2. Etalonnage du Cu :

Dans tout ce qui va suivre, pour chaque élément à étudier, nous allons déterminer les valeurs optimales des paramètres de la régression PLS. Cette optimisation est évaluée en général par le RMSEC et le RMSEP. Dans un premier temps, nous allons observer la variation du RMSEC en fonction du nombre de composantes PLS. L'effet de la délimitation des canaux à utiliser au lieu du spectre entier sera ensuite étudié à l'aide de la comparaison des coefficients de régression et encore des RMSEC. Ces mêmes études seront effectuées dans les cas des données qui ont subi un prétraitement qui est le centrage par la moyenne suivi de la prise des racines carrées.

5.2.2.1. Nombre de composantes PLS :

Nous utilisons la méthode de validation croisée Leave One Out (LOO-CV) pour calculer la valeur du RMSEC en fonction du nombre de composantes PLS. Cette méthode est détaillée dans le paragraphe §2.5.3.2. Les résultats obtenus sont présentés dans le tableau et la figure qui suivent.

Tableau 5.2 : Valeurs du RMSEC (ppb) pour un nombre de composantes de 1 à 9

Nb de composantes	RMSEC
1	427,7
2	161,8
3	127,4
4	118,7
5	100,2
6	99,5
7	99,1
8	99,1
9	99,1

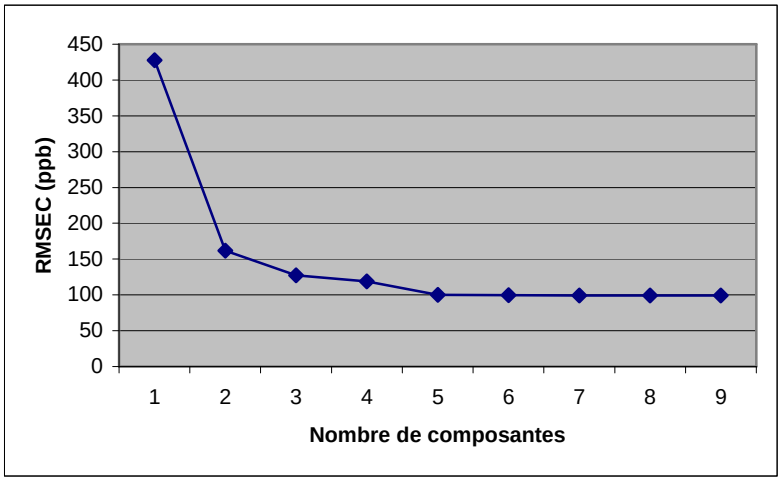
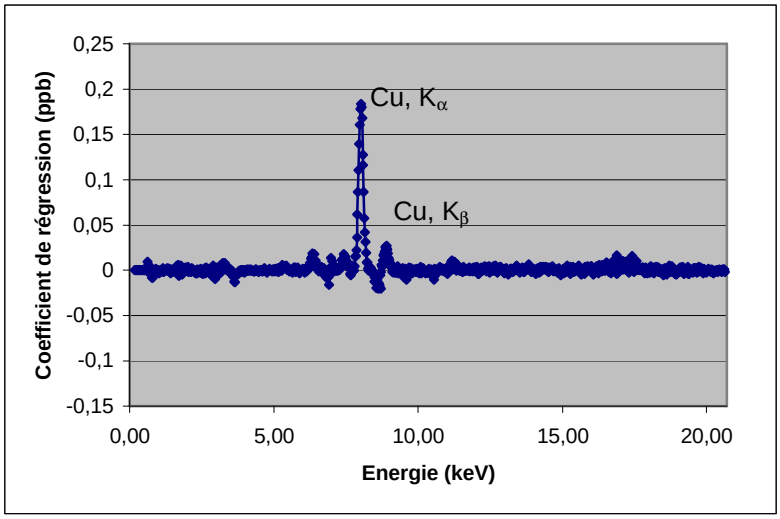
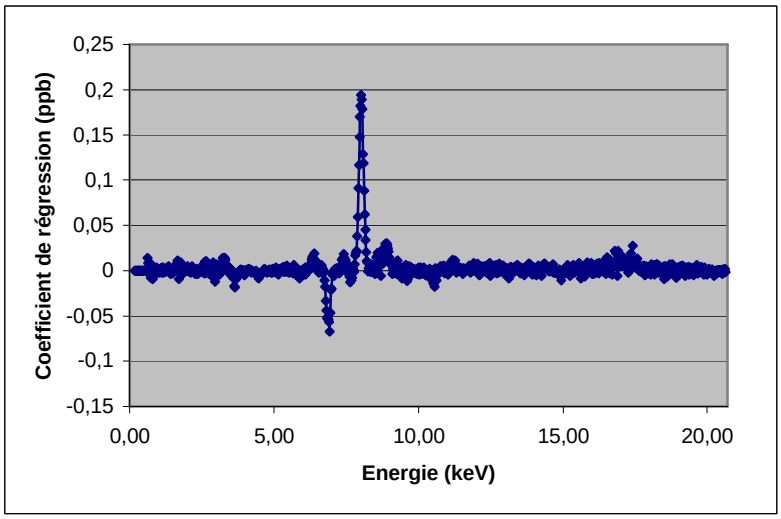


Figure 5.1 : Variation du RMSEC en fonction du nombre de composantes PLS pour le Cu

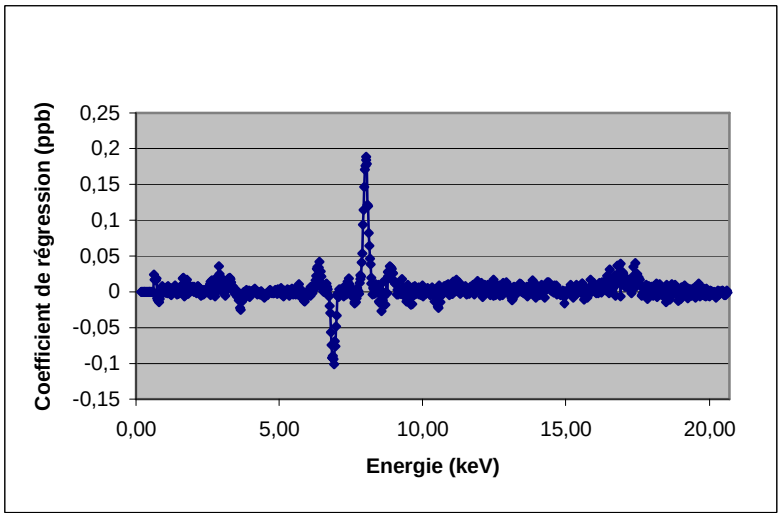
D'une façon générale, la valeur optimale du nombre de composantes PLS est celle pour laquelle la valeur du RMSEC commence à se stabiliser. Il ne faut pourtant pas oublier que le but est d'avoir un modèle PLS qui a le meilleur pouvoir de prédiction possible alors que le RMSEC ne renseigne que sur l'ajustage de ce modèle sur les échantillons d'étalonnage. Dans notre cas (Figure 5.1), ce nombre optimal peut être de 3 à 5. Trop de composantes risquent de nuire à la capacité de prédiction du modèle en modélisant les variations dans le spectre qui ne concernent pas l'élément étudié. L'observation des coefficients de régression pour ces différents nombres de composantes illustre très bien ce phénomène et donne ainsi des suppléments d'informations très utiles pour le choix du nombre de composantes à utiliser. Le coefficient de régression étant une matrice qu'on doit multiplier par la matrice du spectre, les valeurs de ses éléments doivent en général présenter des pics pour les canaux correspondant à l'élément d'intérêt et être approximativement nulles ailleurs (figure 5.2).


Coefficient de régression (ppb)
0,25
0,2
0,15
0,1
0,05
0
-0,05
-0,1
-0,15
Cu, Kα
Cu, Kβ
0,00
5,00
10,00
15,00
20,00
Energie (keV)

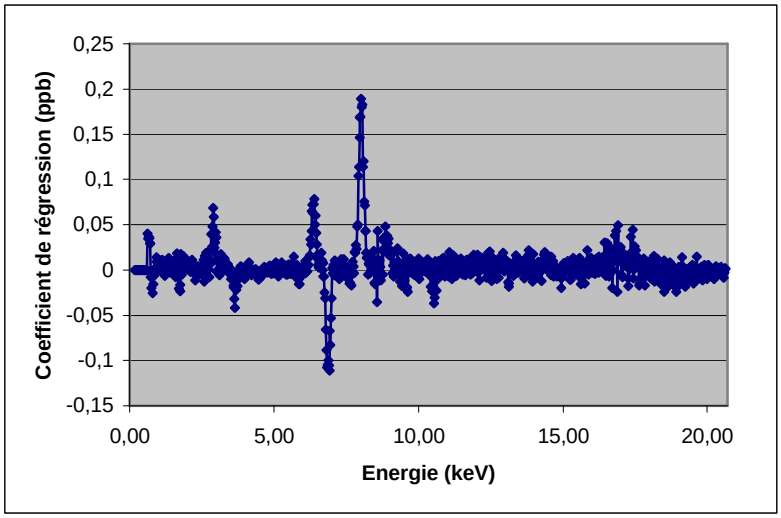
(a)


Coefficient de régression (ppb)
0,25
0,2
0,15
0,1
0,05
0
-0,05
-0,1
-0,15
0,00
5,00
10,00
15,00
20,00
Energie (keV)

(b)


Coefficient de régression (ppb)
0,25
0,2
0,15
0,1
0,05
0
-0,05
-0,1
-0,15
0,00
5,00
10,00
15,00
20,00
Energie (keV)

(c)



(d)

Figure 5.2 : Coefficient de régression obtenu pour 2(a), 3(b), 4(c), 5(d) composantes pour le Cu

Les graphes de la figure 5.2 montrent bien que le coefficient de régression présente des pics qui correspondent aux canaux (ou énergies) du Cu. Pour le modèle à 2 composantes, on ne détecte que les pics K_α et K_β du Cu. Pour 3 composantes, on peut distinguer un pic négatif qui correspond au Co. Ceci est dû à la présence d'une quantité assez importante de cet élément dans tous les échantillons. Le modèle élimine la contribution de cet élément pour qu'elle n'interfère pas celle du Cu qui est l'élément d'intérêt. A partir de 4 composantes PLS, on observe que le modèle prend en compte d'autres phénomènes qu'on peut qualifier de parasites. On remarque en effet la présence d'un pic correspondant à l'élément Fe qui est tout simplement de l'impureté lors de la préparation des échantillons. Pour 5 composantes, on peut même observer la modélisation du Mo.

On peut donc conclure à partir de l'observation de ces deux paramètres qui sont le RMSEC et le coefficient de régression que le nombre optimal de composantes PLS est de 3. Le RMSEC est en effet encore trop élevé pour 2 composantes alors qu'à partir de 4 composantes, des phénomènes parasites comme les impuretés et les bruits de fond commencent à être modélisés.

5.2.2.2. Délimitation de la plage de canaux à utiliser :

Nous avons pu confirmer à partir de l'observation de la variation du coefficient de régression en fonction de l'énergie que seule la plage contenant l'élément d'intérêt

contribue de façon significative à la construction du modèle PLS. Le reste du spectre ne constitue ainsi qu'une sorte de bruit de fond au dit modèle. L'élimination de cette partie inutile du spectre pourrait alors améliorer le pouvoir de prédiction de notre modèle. La figure suivante montre une visualisation du spectre d'un échantillon par le logiciel. Nous avons choisi la plage délimitée sur cette figure par les lignes vertes qui contiennent les éléments qui nous intéressent (les métaux lourds en général). Cette zone débute par le canal N°200 et finit au N°650, ce qui correspondent aux énergies de 4,16 à 13,16 keV. Nous allons utiliser cette plage d'énergies dans tout ce qui va suivre.

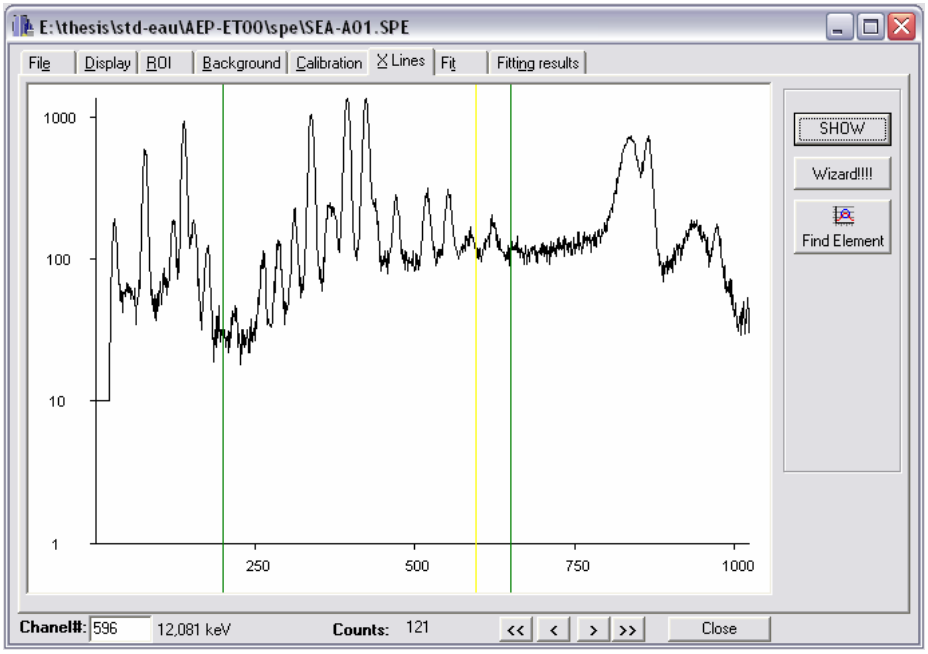


Figure 5.3 : Représentation graphique du spectre d'un échantillon d'étalonnage pour la sélection de la plage de canaux par le logiciel X-PLS

Comme dans le cas où le spectre entier a été utilisé dans le paragraphe précédent, les résultats obtenus sont présentés dans le tableau et les figures suivants pour le RMSEC et le coefficient de régression (tableau 5.3, figure 5.4 et 5.5).

Tableau 5.3 : Valeurs du RMSEC pour 1 à 9 composantes pour le Cu

Nb de composantes	RMSEC(ppb)
1	394,5
2	173,1
3	140,1
4	131,4
5	129,9
6	125,1
7	124,2
8	124,2
9	124,2

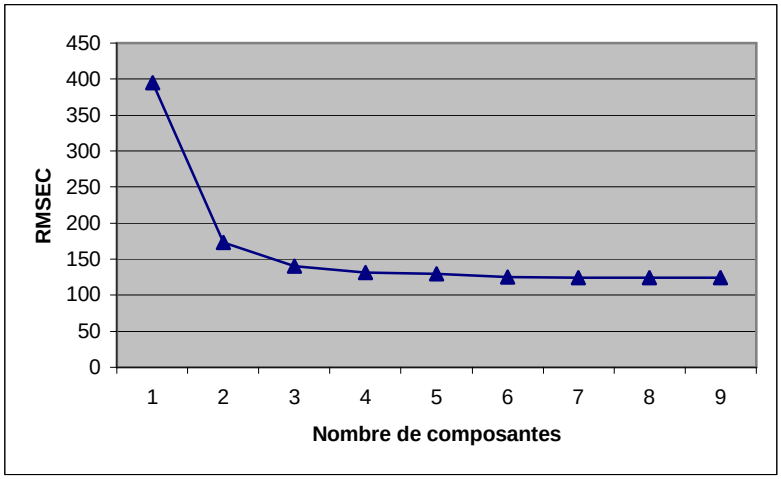
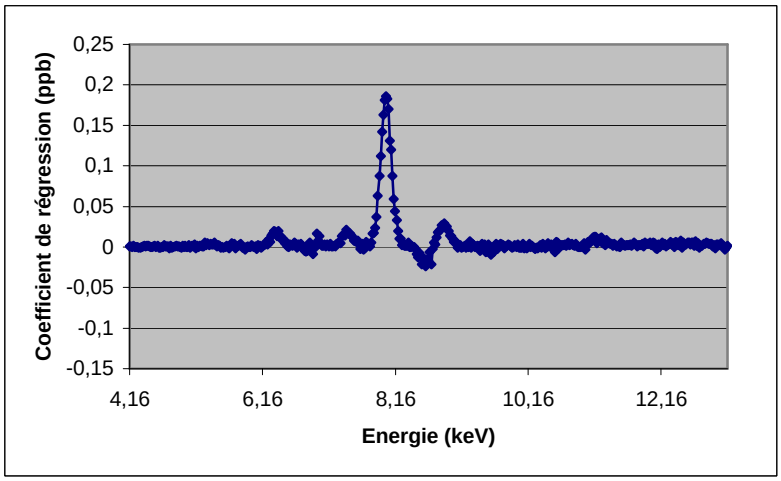
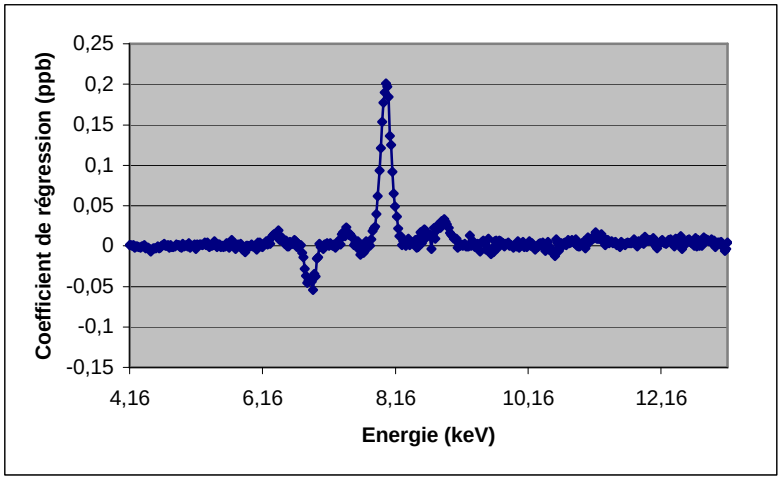


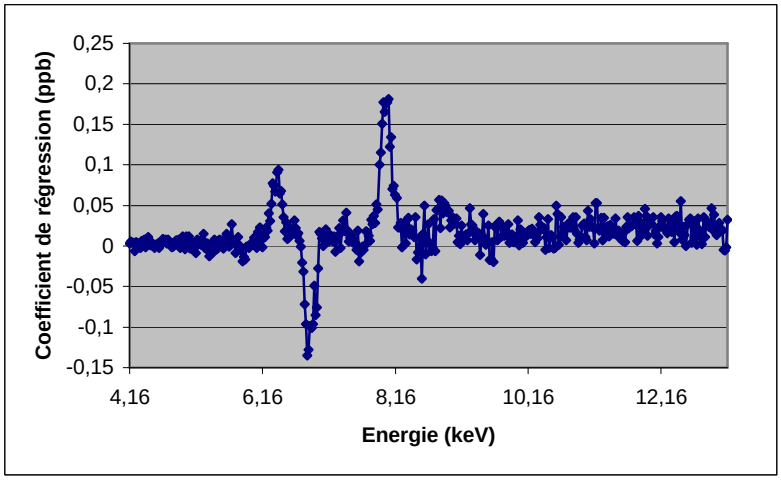
Figure 5.4 : Variation du RMSEC en fonction du nombre de composantes pour le Cu en utilisant la plage [200,650]



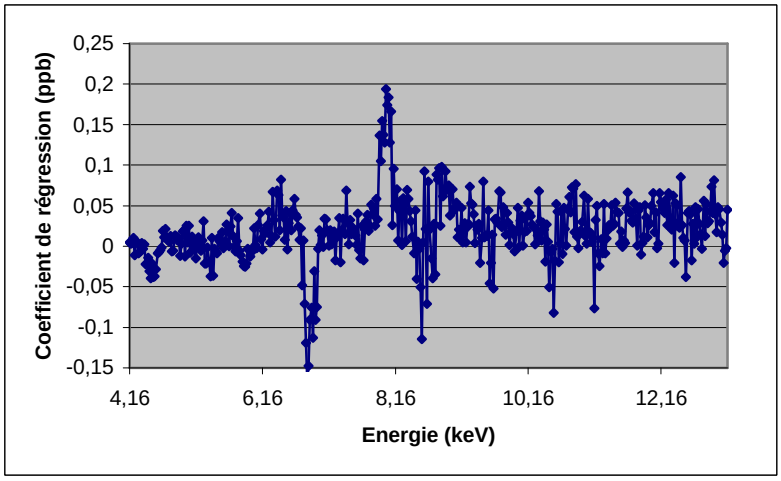
(a)



(b)



(c)



(d)

Figure 5.5 : Coefficient de régression obtenu pour 2(a), 3(b), 4(c), 5(d) composantes pour le Cu en utilisant la plage [200,650]

Ces résultats sont affichés ici dans le but de confirmer le choix du nombre de composantes que nous avons effectué dans le paragraphe précédent. Les mêmes observations que dans ce paragraphe s'appliquent en effet ici.

5.2.2.3. Prétraitement des données :

Les méthodes de prétraitement des données que nous avons développées dans le paragraphe §3.2.3 ont été appliquées à nos données.

5.2.2.4. Prédictions :

Les modèles obtenus vont maintenant être utilisés à la prédiction de la concentration du Cu dans les échantillons test SEA11 – SEA15. Nous allons observer pour cela le RMSEP qui quantifie en général l'erreur commise lors des prédictions. L'étude du RMSEP permet aussi une fois de plus de confirmer le choix du nombre optimal de composantes PLS qui est le paramètre le plus important dans la construction d'un tel modèle.

Nous avons ainsi utilisé un modèle construit avec des données ayant subi un centrage par la moyenne et la prise de la racine carrée et nous avons pris la plage de données 200-650. Les résultats obtenus pour les prédictions des échantillons test pour un nombre de composantes de 2 à 5 sont présentés dans le tableau et les graphiques suivants.

Tableau 5.4 : Valeurs des prédictions des concentrations (ppb) du Cu et du RMSEP en fonction du nombre de composantes PLS

	Valeurs réelles	Nombre de composantes			
		2	3	4	5
SEA11	800	799,5	810,0	785,4	784,9
SEA12	1600	1604,1	1560,8	1554,6	1550,4
SEA13	1000	1148,3	1084,8	1049,2	1046,3
SEA14	1700	1508,8	1558,2	1499,3	1506,8
SEA15	900	1029,8	951,3	872,4	876,7
	RMSEP	122,8	79,4	95,7	92,4

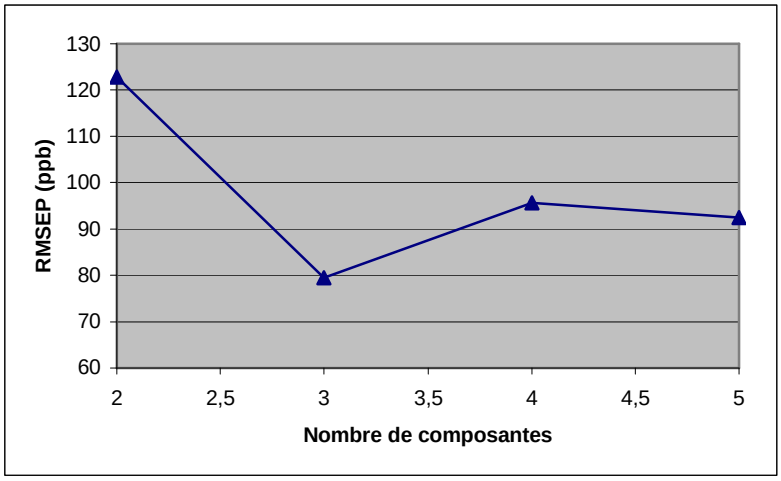


Figure 5.6 : Variation du RMSEP en fonction du nombre de composantes pour le Cu

On retrouve bien la justification du nombre optimal de composantes qui est de 3. Le RMSEP est en effet minimal pour ce nombre, ce qui veut dire que le pouvoir de prédiction du modèle est le meilleur. La figure suivante représente les prédictions des concentrations en fonction des valeurs réelles pour 3 composantes PLS.

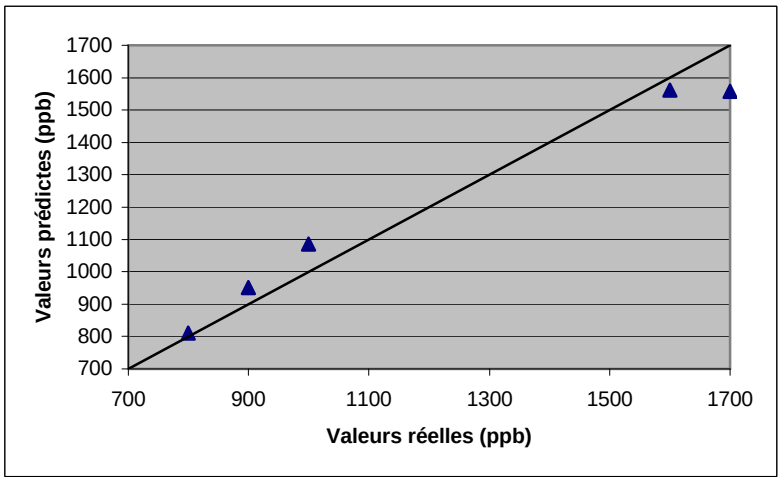


Figure 5.7 : Valeurs prédites des concentrations du Cu en fonction des valeurs réelles pour un modèle à 3 composantes

5.2.2.5. Intervalles de prédictions :

Pour la détermination de l'intervalle de prédiction, nous avons utilisé la méthode bootstrap décrite dans le paragraphe §2.5.4.3 avec les paramètres suivants :

- nombre d'échantillons bootstrap : 100
- type : intervalle de prédiction percentile
- niveau de confiance : 95%

Le tableau suivant récapitule les valeurs de ces intervalles pour le modèle obtenu précédemment, ainsi que l'évaluation de la précision sur les prédictions de la concentration du Cu.

Tableau 5.5 : Intervalles de prédiction pour un modèle à 3 composantes (cas du Cu)

	Valeurs réelles (ppb)	Valeurs prédites (ppb)	Intervalles de prédiction (ppb)	Incertitude (%)
SEA11	800	810,0	[794,3 ; 825,8]	1,9
SEA12	1600	1560,8	[1535,5 ; 1586,1]	1,6
SEA13	1000	1084,8	[1061,5 ; 1108,1]	2,1
SEA14	1700	1558,2	[1532,9 ; 1583,6]	1,6
SEA15	900	951,3	[923,1 ; 979,4]	2,9
			Moyenne	2,1 %

Application du lissage des données :

Après avoir appliqué la méthode de lissage de données décrite dans le paragraphe §3.2.4, les prédictions de la concentration du Cu pour les échantillons test ainsi que le RMSEP correspondant sont présentés dans le tableau et la figure qui suivent.

Tableau 5.6 : Prédictions de la concentration (ppb) du Cu après lissage des données

Valeurs réelles	Valeurs prédites
800	807,4
1600	1589,0
1000	1043,1
1700	1557,8
900	896,9
RMSEP	66,7

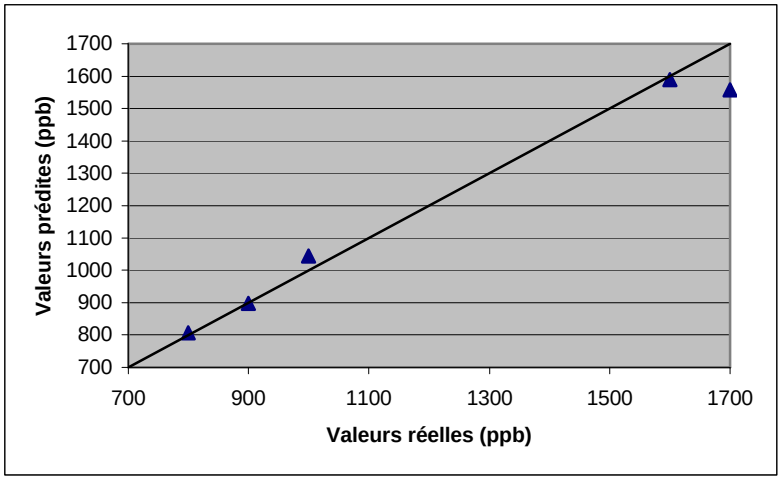


Figure 5.8 : Valeurs prédites des concentrations du Cu en fonction des valeurs réelles après lissage des données

On voit une nette amélioration du pouvoir de prédiction du modèle. Le RMSEP passe en effet de 79,4 à 66,7 ppb après le lissage.

Le tableau suivant récapitule les écarts relatifs commis pour les prédictions de la concentration du Cu dans les 5 échantillons test.

Tableau 5.7 : Valeurs des écarts relatifs des prédictions pour le Cu

	Ecart relatif (%)
SEA11	1,25
SEA12	2,45
SEA13	8,48
SEA14	8,34
SEA15	5,70
Moyenne	5,24%

5.2.3. Etalonnage du Zn :

Pour le cas du Zn, nous avons adopté la même méthodologie que pour le Cu.

Les spectres des 10 échantillons d'étalonnage sont utilisés pour construire le modèle PLS. Le spectre entier est d'abord utilisé pour définir le RMSEC à partir de la validation croisée LOO. On délimite par la suite la plage de canaux à utiliser. Après cela, on utilise les méthodes de prétraitement des données décrites dans le paragraphe précédent. Le modèle ainsi obtenu sera enfin utilisé pour la prédiction de la concentration du Zn dans les échantillons test.

5.2.3.1. Nombre de composantes PLS :

L'application du LOO-CV aux échantillons d'étalonnage a permis d'obtenir les résultats suivants.

Tableau 5.8 : Valeurs du RMSEC pour un nombre de composantes de 1 à 9

Nb de composantes	RMSEC
1	404,1
2	237,9
3	193,1
4	172,2
5	151,1
6	152,8
7	151,4
8	151,4
9	151,4

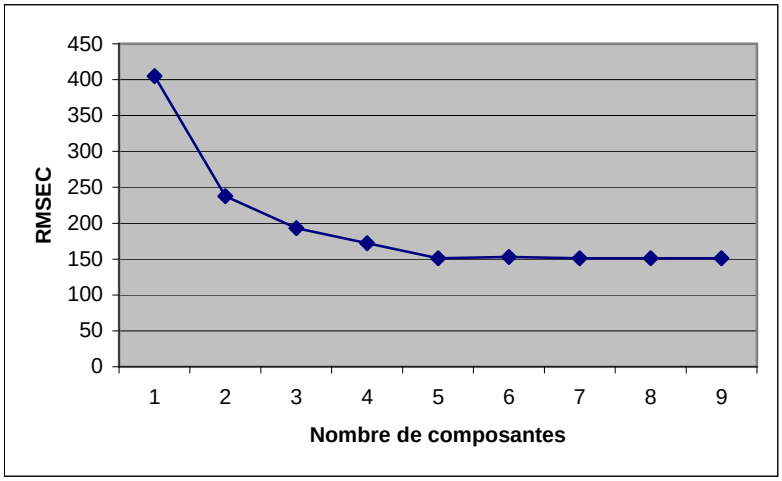
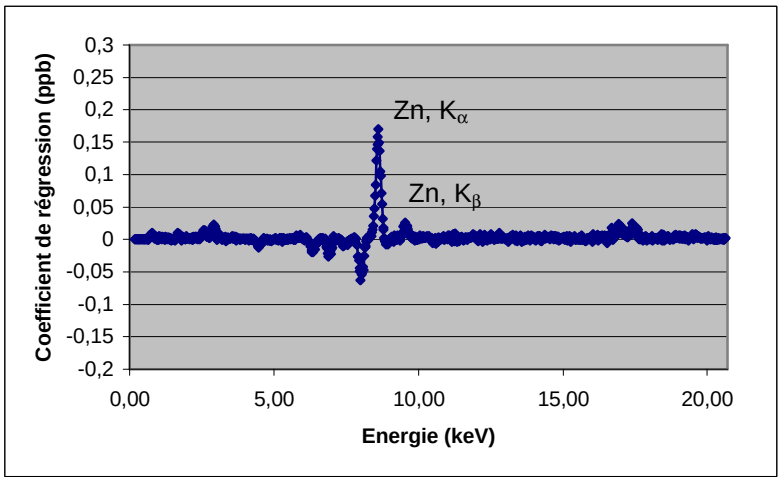
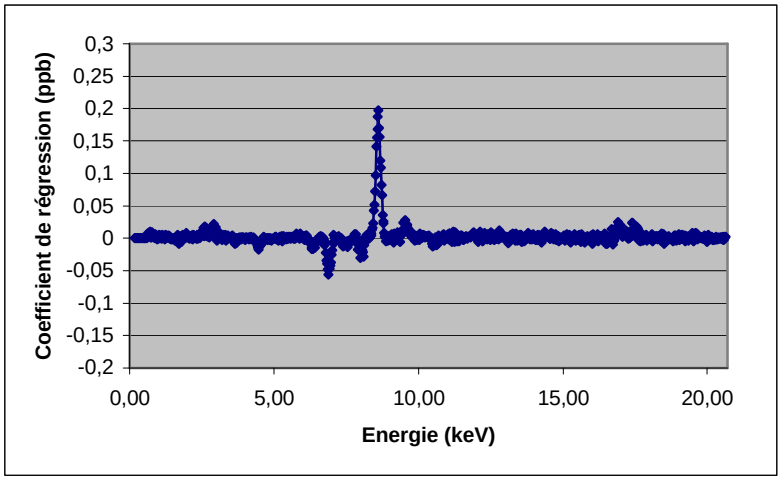


Figure 5.9 : Variation du RMSEC en fonction du nombre de composantes PLS pour le Zn

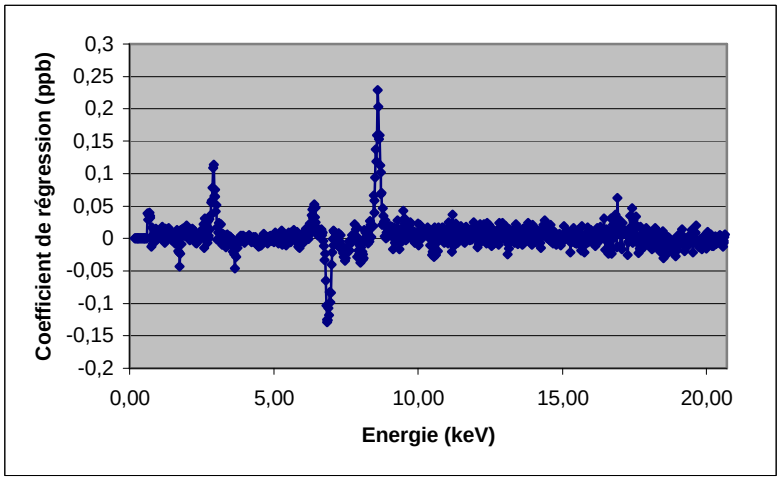
Comme pour le Cu, l'analyse du coefficient de régression est nécessaire pour avoir plus d'information pour le choix du nombre optimal de composantes PLS. Les figures 5.10 a,b,c,d suivantes montrent la variation de ce coefficient en fonction de l'énergie pour un nombre de composantes PLS de 2 à 5.



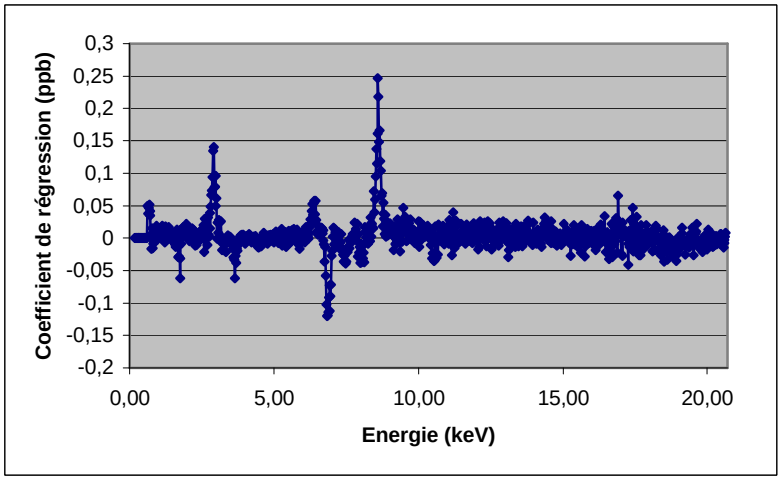
(a)



(b)



(c)



(d)

Figure 5.10 : Coefficient de régression obtenu pour 2(a), 3(b), 4(c), 5(d) composantes pour le Zn

On observe ainsi le phénomène de surévaluation (overfitting) à partir de 4 composantes. Un nombre de composantes égal à 3 est donc également pris pour le Zn.

5.2.3.2. Délimitation de la plage de canaux :

Comme pour le Cu, nous utilisons la plage d'énergies de 4,16 à 13,16 keV.

5.2.3.3. Prétraitement des données :

Les mêmes méthodes de prétraitement de données que dans le cas du Cu sont utilisées pour le Zn. La figure 5.11 suivante montre l'allure du coefficient de régression pour le modèle construit avec les conditions citées plus haut.

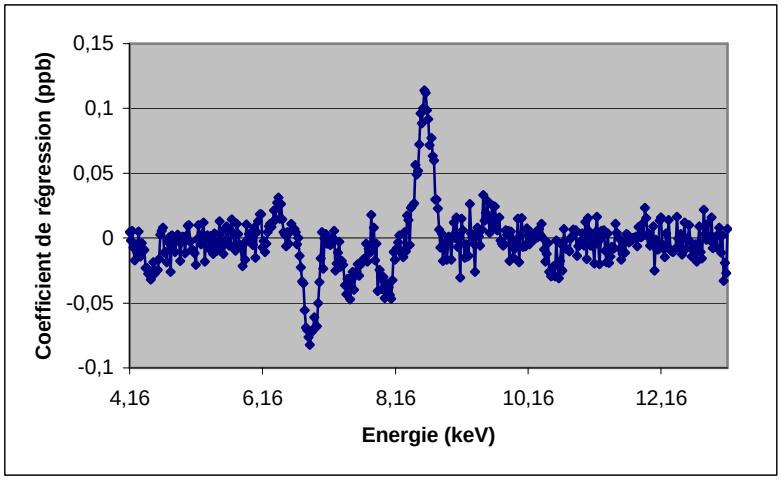


Figure 5.11 : Coefficient de régression du modèle PLS à 3 composantes pour le Zn en utilisant la plage [200,650]

5.2.3.4. Prédictions :

Le tableau suivant regroupe les valeurs obtenues pour les prédictions de la concentration du Zn dans les échantillons test SEA11 à SEA15 pour un nombre de composantes de 1 à 6. Ces mesures ont été effectuées dans le but de déterminer la variation du RMSEP en fonction du nombre de composantes afin de confirmer le choix du nombre optimal de composantes PLS utiliser.

Tableau 5.9 : Valeurs des prédictions des concentrations (ppb) du Zn et du RMSEP en fonction du nombre de composantes PLS

	Valeurs réelles	Nombre de composantes					
		1	2	3	4	5	6
SEA11	1600	1549,5	1414,1	1675,8	1686,4	1674,9	1673,2
SEA12	900	929,3	940,5	915,1	919,4	915,0	916,3
SEA13	800	1059,9	903,5	781,1	773,8	776,6	776,8
SEA14	1900	1893,7	1817,0	1583,2	1608,8	1607,0	1606,6
SEA15	700	1047,1	766,9	627,6	637,1	638,9	638,0
	RMSEP	195,7	107,7	149,6	139,5	138,5	138,6

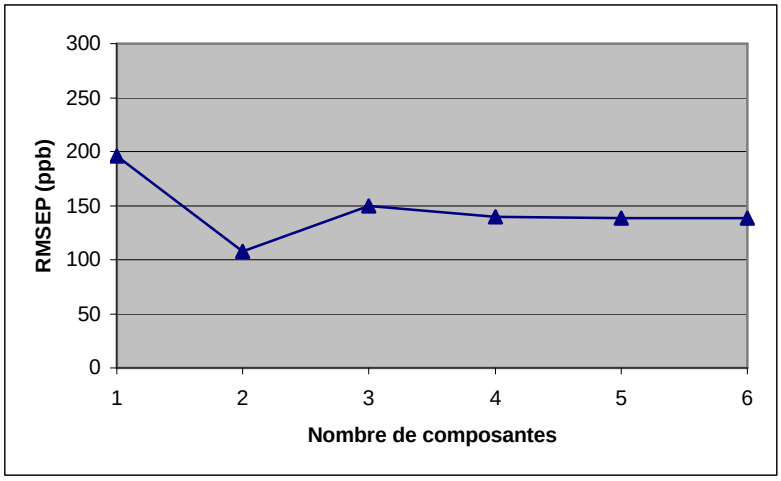


Figure 5.12 : Variation du RMSEP en fonction du nombre de composantes pour le Zn

On voit bien que le RMSEP commence à se stabiliser dès 2 composantes. Nous avons pourtant vu dans le paragraphe §5.2.3.1 (figure 5.9) que le RMSEC est encore trop élevé pour cette valeur du nombre de composantes PLS. Le choix de 3 composantes se justifie donc. La représentation de la variation des valeurs prédites en fonction des valeurs réelles de la concentration du Zn pour un modèle à 3 composantes est donnée par la figure suivante.

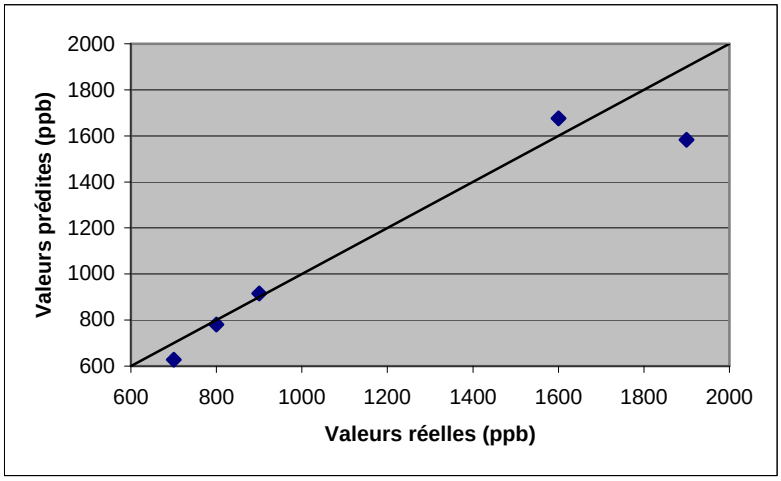


Figure 5.13 : Valeurs prédites des concentrations du Zn en fonction des valeurs réelles pour un modèle à 3 composantes

Les intervalles de prédiction sont regroupés dans le tableau suivant en utilisant les mêmes paramètres que dans le cas du Cu pour la méthode bootstrap.

Tableau 5.10: Intervalles de prédiction pour un modèle à 3 composantes (cas du Zn)

	Valeurs réelles (ppb)	Valeurs prédites (ppb)	Intervalles de prédiction (ppb)	Incertitude (%)
SEA11	1600	1675,8	[1657,8 ; 1693,8]	1,1
SEA12	900	915,1	[898,5 ; 931,6]	1,8
SEA13	800	781,1	[770,4 ; 791,8]	1,4
SEA14	1900	1583,2	[1561,6 ; 1604,8]	1,4
SEA15	700	627,6	[612,7 ; 642,5]	2,4
			Moyenne	1,6%

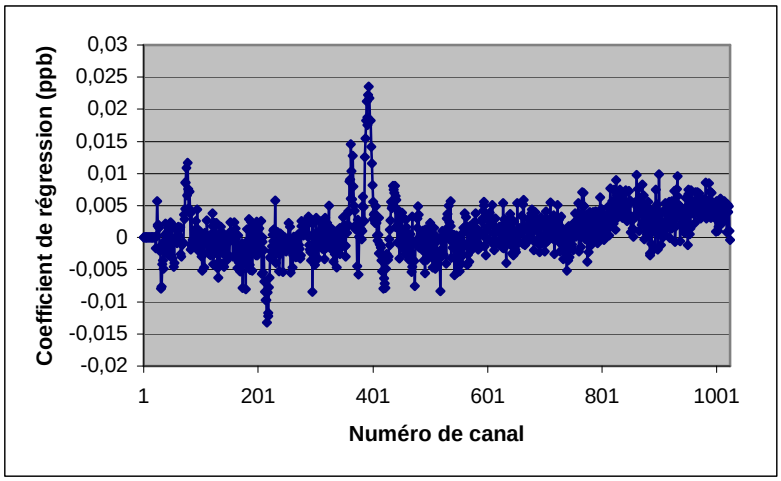
Les écarts relatifs commis lors des prédictions de la concentration du Zn dans les échantillons test SEA11-SEA15 sont présentés dans le tableau suivant.

Tableau 5.11: Valeurs des écarts relatifs des prédictions pour le Zn

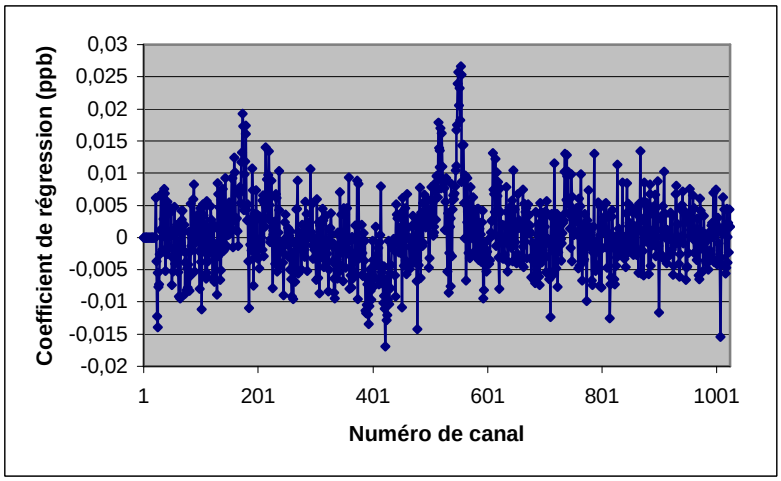
	Ecart relatif (%)
SEA11	4,74
SEA12	1,67
SEA13	2,37
SEA14	16,67
SEA15	10,35
Moyenne	7,15%

5.2.4. Etalonnage des autres éléments :

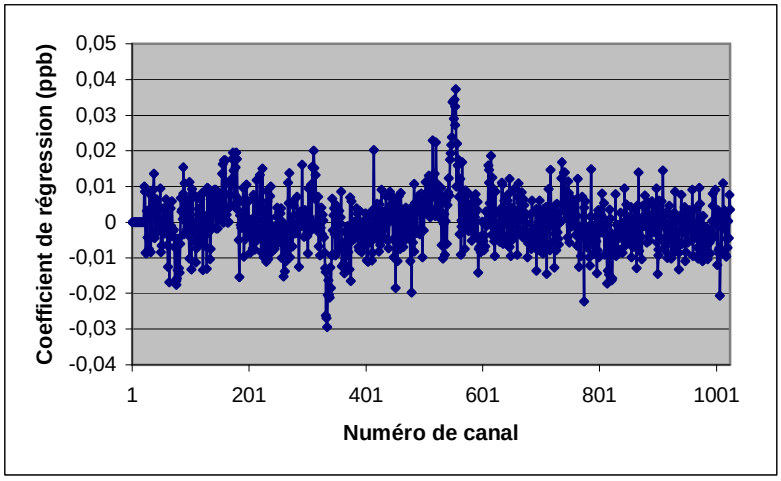
Les éléments qui n'ont pas encore été modélisés dans les échantillons d'étalonnage sont les suivants : Ni, As, Se. Les figures 5.14 a,b,c suivantes montrent le coefficient de régression pour chacun de ces éléments. Nous avons utilisé pour cela un modèle PLS à 3 composantes avec centrage à la moyenne et prise de la racine carrée.



(a)



(b)



(c)

Figure 5.14 : Coefficient de régression pour le Ni (a), As (b), Se (c)

Les échantillons que nous avons utilisés pour cette première série de mesures ont été préparés de manière à ce que les concentrations avoisinent les limites fixées par la norme du CEE pour l'eau potable. Ces valeurs limites sont données dans le tableau suivant :

Tableau 5.12 : Valeurs limites des concentrations pour l'eau potable pour le CEE

Element	Valeurs limites (ppb)
Ni	50
As	50
Se	10
Cu	1000
Zn	1000

C'est pour cela que les concentrations du Cu et du Zn sont à peu près 7 fois les concentrations des autres éléments. Comme la régression PLS maximise la covariance entre le spectre (variables prédictrices) et les concentrations (variables réponses), on s'attend à ce que seules les plus grandes variations au sein du spectre soient modélisées. Dans notre cas, seules les variations relatives au Cu et au Zn sont donc modélisées. Pour tous les cas de figures ci-dessus, en effet, on voit bien que le modèle n'arrive à modéliser aucune variation importante dans le spectre correspondant à une variation de la concentration en ce qui concerne ces autres éléments : Ni, As, Se.

5.2.5. Conclusion partielle :

Cette première série d'étalonnages nous a permis de déterminer la valeur optimale du nombre de composantes PLS pour le Cu et le Zn. Ce nombre est égal à 3. Les avantages de la délimitation de la plage de canaux ainsi que le lissage des données sur l'amélioration du pouvoir de prédiction du modèle ont aussi été démontrés lors de cette première expérimentation. Les résultats obtenus concernant la précision et l'exactitude des prédictions effectuées ont été satisfaisants. On a en effet une précision moyenne de 2,1% pour le Cu et 1,6% pour le Zn. En ce qui concerne l'exactitude, nous avons obtenu des valeurs moyennes des écarts relatifs de 5,24% pour le Cu et de 7,15% pour le Zn.

Une importante information concernant la préparation des échantillons d'étalonnage a aussi découlé de cette première partie de notre travail. Elle a en effet permis de conclure que les concentrations des éléments d'intérêt dans les échantillons d'étalonnage doivent être d'un même ordre de grandeur.

Cette dernière conclusion sera appliquée pour la préparation des échantillons d'étalonnage de la série d'expérimentations qui suit.

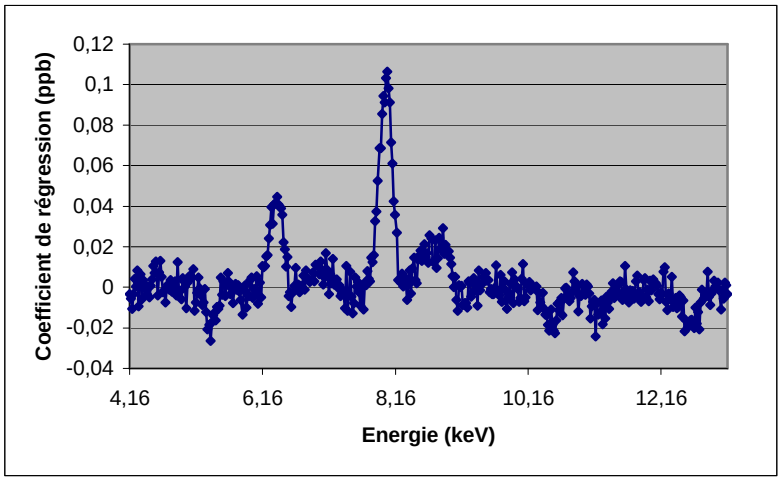
5.3. SERIE D'ETALONNAGES N°2 :

Cette deuxième série d'étalonnages a pour but de tester une configuration d'échantillons d'étalonnage dans laquelle les concentrations des éléments ont le même ordre de grandeur. 10 échantillons standard ont été préparés pour cela. Le tableau suivant représente les constitutions de tous ces échantillons.

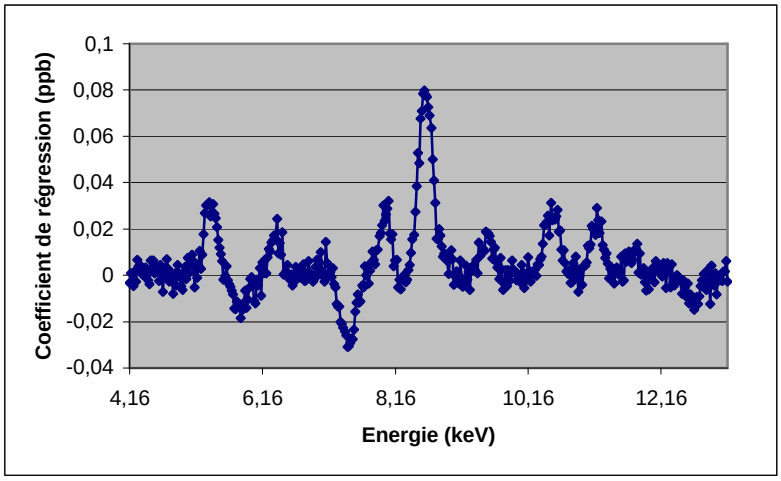
Tableau 5.13: Constitution des 10 échantillons standard pour l'étalonnage (ppb)

	STD01	STD02	STD03	STD04	STD05	STD06	STD07	STD08	STD09	STD10
Ni	450	400	300	150	0	250	350	200	500	50
Cu	250	200	350	50	100	300	450	200	0	150
Zn	0	50	140	200	500	450	300	350	100	250
As	50	250	0	300	200	150	450	100	400	500
Se	100	50	150	500	450	250	300	200	350	0

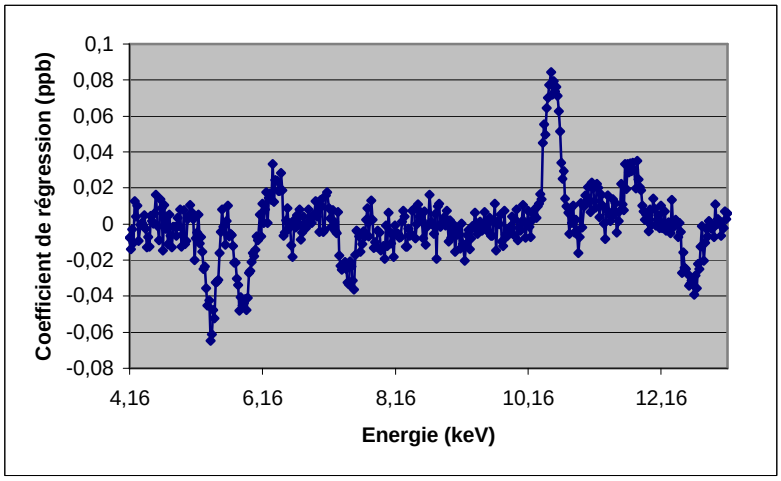
En utilisant les paramètres de la régression PLS que nous avons obtenus lors de la première série de mesures, nous obtenons les coefficients de régression que nous présentons sur les figures 5.15 a,b,c,d,e suivantes.



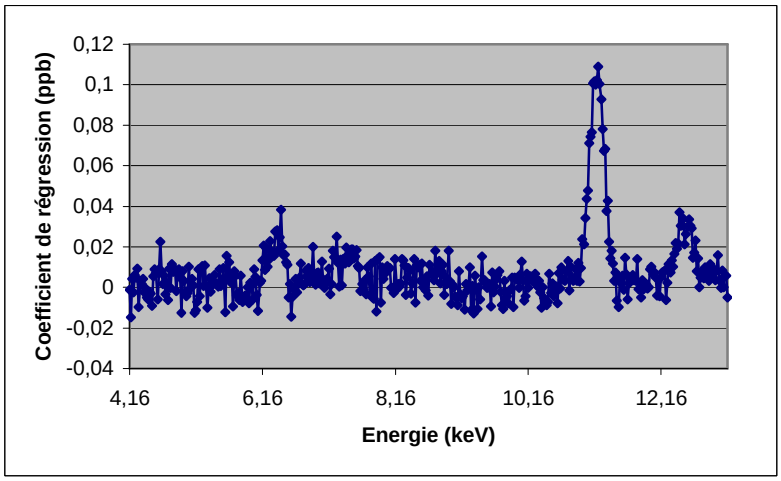
(a)


Coefficient de régression (ppb)
0,1
0,08
0,06
0,04
0,02
0
-0,02
-0,04
4,16
6,16
8,16
10,16
12,16
Energie (keV)

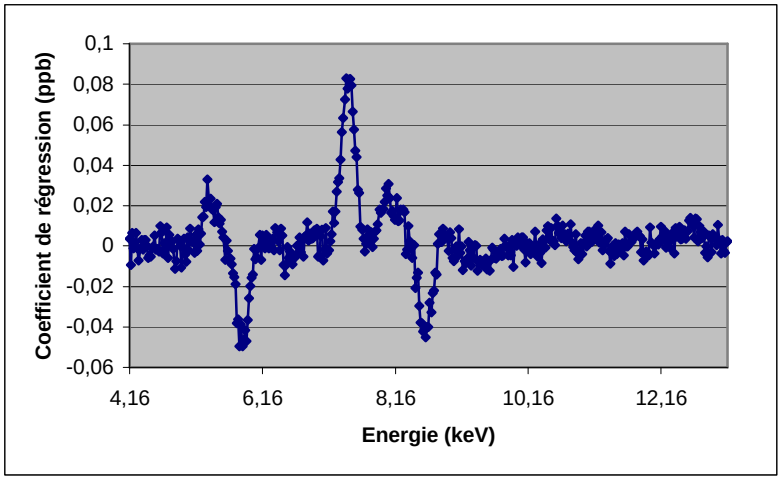
(b)


Coefficient de régression (ppb)
0,1
0,08
0,06
0,04
0,02
0
-0,02
-0,04
-0,06
-0,08
4,16
6,16
8,16
10,16
12,16
Energie (keV)

(c)


Coefficient de régression (ppb)
0,12
0,1
0,08
0,06
0,04
0,02
0
-0,02
-0,04
4,16
6,16
8,16
10,16
12,16
Energie (keV)

(d)



(e)

Figure 5.15 : Coefficient de régression pour le Cu (a), Zn (b), As (c), Se (d), Ni (e) pour la 2è série de mesures

Nous observons que pour chaque élément, les autres éléments présents dans les échantillons contribuent aussi au modèle. Ceci est prouvé par la présence de pics autres que ceux de l'élément d'intérêt. On peut expliquer cela par le fait que les concentrations de ces éléments ont le même ordre de grandeur et que ces concentrations ont été prises au hasard. Il se peut alors que des éléments présentent des variations de concentrations dans le même sens que l'élément d'intérêt pour des échantillons voisins. Ces variations sont alors modélisées en même temps que les variations relatives à l'élément d'intérêt. Ce phénomène peut nuire à la qualité de prédiction du modèle.

Pour résoudre ce problème, nous proposons de choisir des intervalles d'énergies différents pour chaque élément à étudier. Ces intervalles ne doivent contenir que les raies K_α et K_β de l'élément d'intérêt. En adoptant cette option, nous pouvons nous assurer que le modèle ainsi construit ne prend en compte que les informations concernant l'élément d'intérêt. En optant pour ce choix, on peut aussi déduire que les prétraitements des données sont facultatifs. Ces procédures ont en effet pour but d'éliminer les erreurs dues aux fluctuations statistiques dans les données. Or, ces fluctuations sont rendues négligeables car la taille des échantillons est très réduite (100 canaux au maximum).

Dans cette deuxième série d'expérimentations, nous nous intéressons directement à la capacité du modèle construit à prédire les concentrations dans des échantillons inconnus, c'est-à-dire, son pouvoir de prédiction. Nous avons préparé pour cela des échantillons test

sur lesquels les modèles PLS vont être utilisés pour la détermination des concentrations des éléments d'intérêt. Le tableau suivant regroupe les constitutions de ces échantillons test qui sont au nombre de 5.

Tableau 5.14 : Concentrations (ppb) des éléments dans les 5 échantillons test

	TE0101	TE0102	TE0103	TE0104	TE0105
Cr	100	200	400	100	300
Ni	0	100	300	300	100
As	200	400	200	500	0
Se	300	200	0	200	200
Mn	500	300	400	0	500
Cu	600	100	500	100	800
Zn	800	700	200	200	0
Pb	100	300	100	400	100
Co	200	500	300	300	400

Dans tout ce qui va suivre, nous n'allons préciser que les plages que nous avons utilisées pour chaque élément.

5.3.1. Cas du Cu :

Les raies K_α et K_β du Cu étant respectivement à 8,048 et 8,905 keV, nous avons choisi l'intervalle de canaux [375,450] qui correspond à la plage d'énergies : [7,66 ; 9,16] keV. Nous avons regroupé dans le tableau 5.15 suivant les valeurs des prédictions des concentrations du Cu dans les échantillons test ainsi que les RMSEP pour les nombres de composantes PLS de 1 à 5.

Tableau 5.15 : Prédictions des concentrations (ppb) du Cu et le RMSEP correspondant pour un nombre de composantes de 1 à 5

	Valeurs réelles	Nombre de composantes				
		1	2	3	4	5
TE0101	600	685,5	645,0	650,0	589,3	569,5
TE0102	100	120,7	99,4	73,0	70,5	56,1
TE0103	500	556,7	562,3	523,7	473,6	457,0
TE0104	100	72,2	73,3	66,9	70,6	68,5
TE0105	800	806,1	855,1	986,4	996,5	1011,3
	RMSEP	48,5	43,9	89,1	90,7	100,3

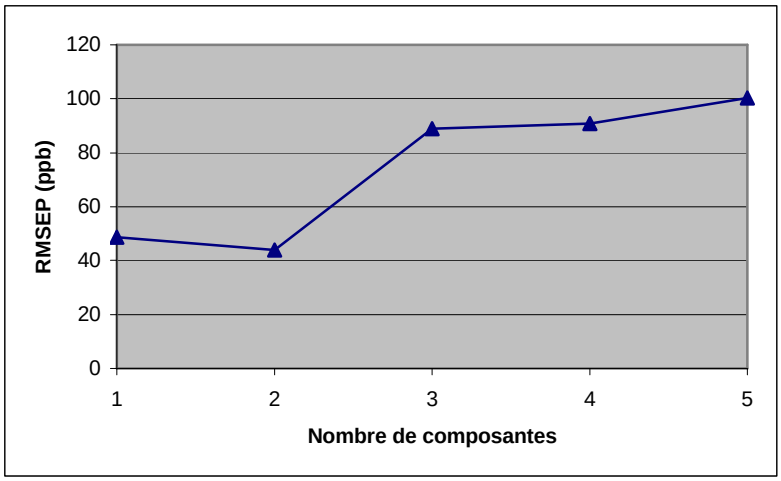


Figure 5.16 : Variation du RMSEP en fonction du nombre de composantes pour le Cu

Nous voyons que le pouvoir de prédiction du modèle est le meilleur pour un nombre de composantes égal à 2. La figure 5.17 suivante montre la variation des prédictions en fonction des valeurs réelles pour les concentrations du Cu.

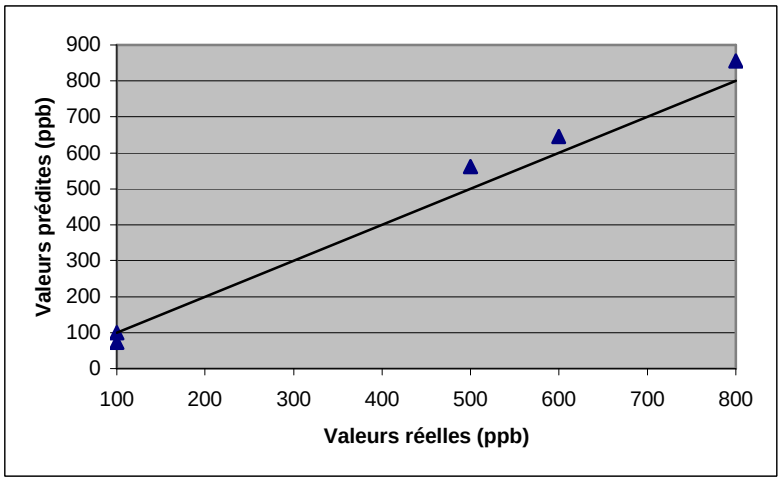


Figure 5.17 : Valeurs prédites des concentrations du Cu en fonction des valeurs réelles

Le tableau suivant contient les écarts relatifs par rapport aux valeurs attendues des concentrations ainsi que la valeur moyenne de ces écarts.

Tableau 5.16 : Valeurs des écarts relatifs des prédictions pour le Cu

	Ecarts (%)
TE0101	7,49
TE0102	0,63
TE0103	12,46
TE0104	26,75
TE0105	6,89
Moyenne	10,85%

5.3.2. Cas du Zn :

Pour le Zn, la plage que nous avons utilisée est [410-490] ([8,36 ; 9,96] keV) qui contient bien les raies K_α et K_β qui sont respectivement de 8,639 et 9,572 keV. Le tableau 5.17 et la figure 5.18 suivants représentent les valeurs des concentrations prédites pour les échantillons test pour 1 à 5 composantes ainsi que la variation des RMSEP correspondants en fonction du nombre de composantes.

Tableau 5.17 : Prédictions des concentrations (ppb) du Zn et le RMSEP correspondant pour un nombre de composantes de 1 à 5

	Valeurs réelles	Nombre de composantes				
		1	2	3	4	5
TE0101	800	490,7	690,7	714,5	648,6	630,9
TE0102	700	504,0	725,8	705,0	660,1	656,0
TE0103	200	237,8	197,5	224,5	173,5	163,6
TE0104	200	221,6	181,7	169,3	147,8	142,1
TE0105	0	136,2	0,2	62,3	56,1	29,0
	RMSEP	175,8	50,9	50,5	78,8	84,9

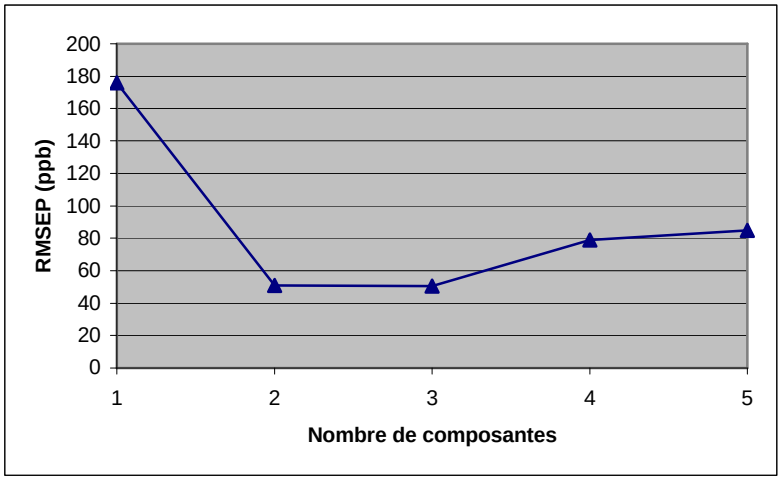


Figure 5.18 : Variation du RMSEP en fonction du nombre de composantes pour le Zn

Les valeurs du RMSEP sont approximativement les mêmes pour 2 et 3 composantes PLS. Nous avons ainsi pris 2 composantes car d'une façon générale, il est préférable de prendre le moins de composantes possibles pour éviter le phénomène de surévaluation ou overfitting. La figure suivante représente la variation des valeurs prédites en fonction des valeurs réelles des concentrations du Zn pour 2 composantes PLS.

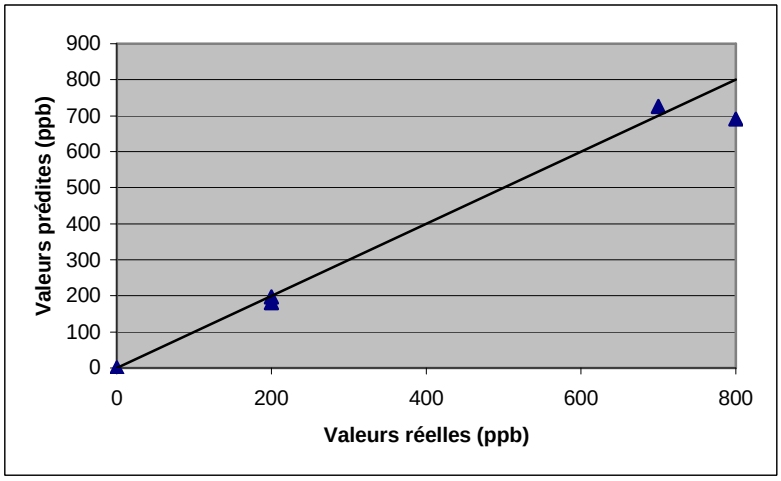


Figure 5.19 : Valeurs prédites des concentrations du Zn en fonction des valeurs réelles

Les écarts relatifs pour les prédictions sont présentés dans le tableau suivant.

Tableau 5.18 : Valeurs des écarts relatifs des prédictions pour le Zn

	Ecarts (%)
TE0101	13,67
TE0102	3,68
TE0103	1,23
TE0104	9,12
TE0105	N D
Moyenne	6,93%

5.3.3. Cas de l'As :

Les raies K_α et K_β de l'As sont à 10,54 et 11,72 keV, nous avons ainsi utilisé la plage de canaux [500-600] qui correspond à la plage d'énergies [10,16 ; 12,16] keV. Comme pour les éléments précédents, nous présentons dans le tableau suivant les valeurs des prédictions des concentrations de l'As pour les échantillons test. La figure 5.20 montre quant à elle la variation du RMSEP en fonction du nombre de composantes PLS.

Tableau 5.19 : Prédictions des concentrations (ppb) de l'As et le RMSEP correspondant pour un nombre de composantes de 1 à 5

	Valeurs réelles	Nombre de composantes				
		1	2	3	4	5
TE0101	200	142,0	113,7	77,9	82,5	79,8
TE0102	400	494,5	550,2	353,8	339,1	331,7
TE0103	200	122,4	198,8	221,1	251,4	248,0
TE0104	500	560,3	641,3	530,9	556,8	561,0
TE0105	0	3,7	0,2	0,1	0,3	0,0
	RMSEP	66,3	100,0	60,7	68,4	70,9

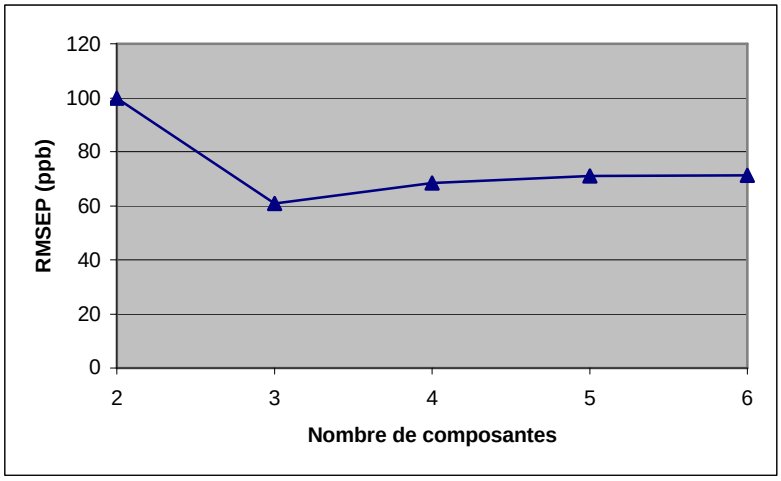


Figure 5.20 : Variation du RMSEP en fonction du nombre de composantes pour l'As

La valeur minimale du RMSEP est obtenue pour 3 composantes. La variation des valeurs obtenues pour les prédictions des concentrations de l'As en fonction des valeurs réelles est représentée sur la figure 5.21 suivante.

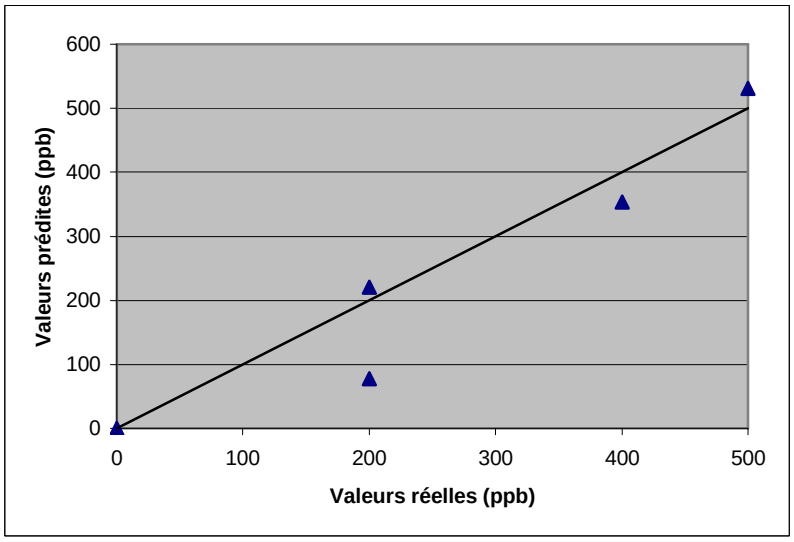


Figure 5.21 : Prédictions des concentrations de l'As en fonction des valeurs réelles

Nous remarquons un échantillon qui est aberrant. Ceci est sûrement dû à une erreur dans la préparation de l'échantillon. En éliminant cet échantillon, nous avons dans le tableau suivant les écarts relatifs pour les prédictions.

Tableau 5.20 : Valeurs des écarts relatifs des prédictions pour l'As

	Ecart relatif (%)
~~***TE0101***~~	~~*61,06*~~
TE0102	11,55
TE0103	10,52
TE0104	6,18
TE0105	N.D[3].
Moyenne	9,42%

5.3.4. Cas du Se :

Le Se possède deux raies K_α et K_β à 11,18 keV et 12,65 keV. Nous avons ainsi choisi la plage de canaux [530,650] qui correspond à [10,76 ; 13,01] keV.

Les résultats obtenus lors des prédictions des concentrations du Se sont représentés dans le tableau suivant.

Tableau 5.21 : Prédictions des concentrations (ppb) du Se et le RMSEP correspondant pour un nombre de composantes de 1 à 5

	Valeurs réelles	Nombre de composantes				
		1	2	3	4	5
TE0101	300	262,8	260,6	282,3	254,3	247,8
TE0102	200	201,8	201,9	185,7	202,3	233,2
TE0103	0	8,9	8,1	6,7	6,7	6,1
TE0104	200	192,4	191,2	163,4	184,1	199,5
TE0105	200	154,7	170,0	196,7	157,9	144,7
	RMSEP	26,7	22,8	19,6	28,8	37,2

La figure suivante illustre la variation du RMSEP en fonction du nombre de composantes.

[3] *Non disponible. C'est le cas où la valeur réelle de la concentration est nulle.*

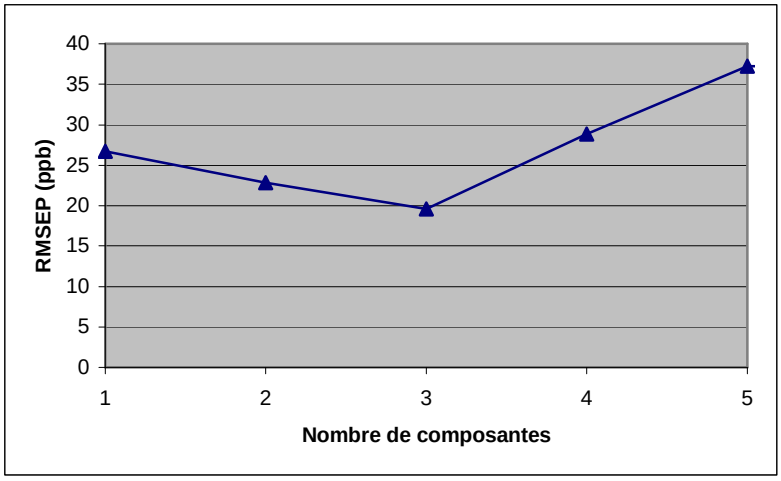


Figure 5.22 : Variation du RMSEP en fonction du nombre de composantes pour le Se

On obtient le meilleur pouvoir de prédiction du modèle avec 3 composantes PLS. La variation des valeurs de prédiction en fonction des valeurs réelles des concentrations du Se est représentée sur la figure suivante, pour un modèle à 3 composantes.

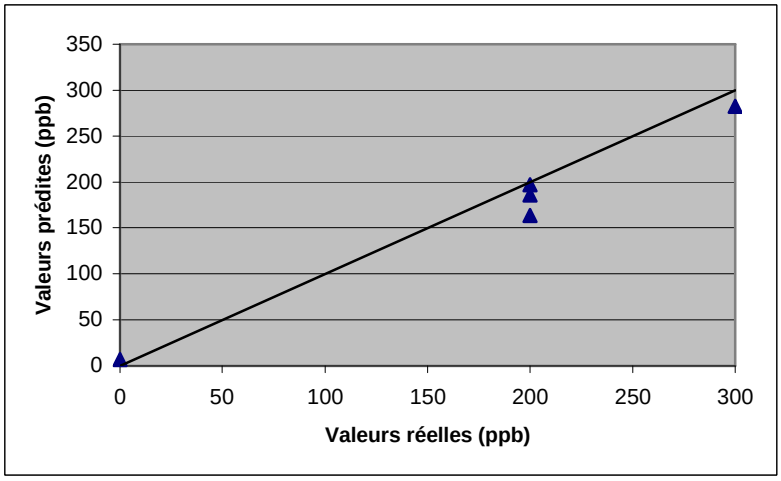


Figure 5.23 : Prédictions des concentrations du Se en fonction des valeurs réelles

Les écarts relatifs commis pendant les prédictions de la concentration du Se pour le modèle à 3 composantes sont regroupés dans le tableau qui suit.

Tableau 5.22 : Valeurs des écarts relatifs des prédictions pour le Se

	Ecart relatif (%)
TE0101	5,89
TE0102	7,14
TE0103	N.D.
TE0104	18,32
TE0105	1,65
Moyenne	8,25%

5.3.5. Cas du Ni :

Pour le Ni, nous avons choisi la plage de données [345-415] qui contient les raies K_α et K_β du Ni qui sont de 7,478 et de 8,265 keV. Nous présentons dans le tableau 5.23 suivant les valeurs des prédictions de la concentration du Ni pour un nombre de composantes de 1 à 5 ainsi que les valeurs du RMSEP.

Tableau 5.23 : Prédictions des concentrations (ppb) du Ni et le RMSEP correspondant pour un nombre de composantes de 1 à 5

	Valeurs réelles	Nombre de composantes				
		1	2	3	4	5
TE0101	0	60,1	23,8	11,9	49,3	74,4
TE0102	100	115,6	121,9	61,4	34,8	21,2
TE0103	300	369,9	305,9	284,9	109,2	96,2
TE0104	300	232,2	258,6	229,9	164,7	161,4
TE0105	100	203,1	121,9	152,0	1,8	8,2
	RMSEP	69,2	25,6	43,5	119,2	127,2

Comme pour les cas précédents, la variation du RMSEP en fonction du nombre de composantes est représentée par la figure suivante.

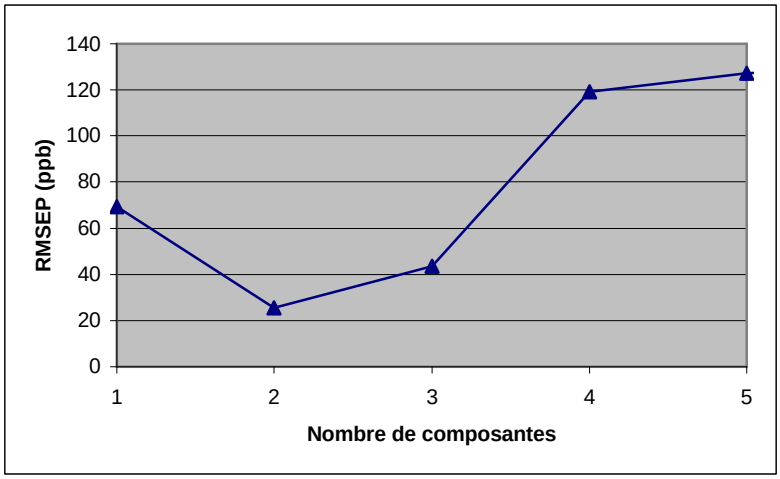


Figure 5.24 : Variation du RMSEP en fonction du nombre de composantes pour le Ni

Nous observons la valeur minimale du RMSEP pour 2 composantes PLS. Nous présentons dans la figure suivante la variation des valeurs des prédictions de la concentration du Ni en fonction des valeurs réelles.

Il faut remarquer que le point qui est entouré d'un cercle sur la figure représente deux échantillons qui ont la même valeur (121,9 ppb) : TE0102 et TE0105.

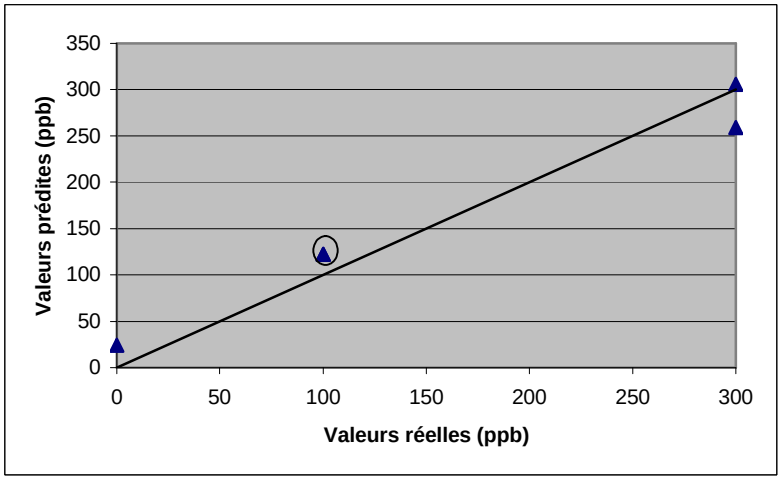


Figure 5.25 : Prédictions des concentrations du Ni en fonction des valeurs réelles

Les écarts relatifs commis pour ces prédictions sont représentés dans le tableau suivant.

Tableau 5.24 : Valeurs des écarts relatifs des prédictions pour le Ni

	Ecart relatif (%)
TE0101	N.D.
TE0102	21,95
TE0103	1,97
TE0104	13,79
TE0105	21,91
Moyenne	14,90%

5.3.6. Conclusion partielle :

Cette série de mesures nous a permis de démontrer qu'on peut modéliser plusieurs éléments si les concentrations de ces éléments ont le même ordre de grandeur dans les échantillons d'étalonnage. On a aussi remarqué que le nombre optimal de composantes PLS peut changer d'un élément à un autre mais aussi d'une configuration d'étalonnage à une autre pour un même élément. On peut donner comme exemple le cas du Cu. Ce nombre optimal était de 3 dans la première série d'étalonnages alors qu'il était de 2 dans la deuxième. Nous pouvons en déduire qu'on doit toujours procéder à l'analyse du RMSEP et/ou du RMSEC pour déterminer le nombre de composantes à utiliser, au lieu d'utiliser les valeurs obtenues lors d'autres étalonnages.

Les écarts relatifs des prédictions pour les éléments que nous avons modélisés dans cette partie sont regroupés dans le tableau suivant.

Tableau 5.25 : Récapitulation des écarts relatifs des prédictions

Elément	Ecart relatif (%)
Ni	14,90
Cu	10,85
Zn	6,93
As	9,42
Se	8,25

Ces valeurs sont encore relativement élevées. Ceci est dû à l'insuffisance du nombre d'échantillons utilisés pour l'étalonnage. On n'a pu utiliser en effet que 10 échantillons pour la modélisation de ces 5 éléments. Comme la régression PLS cherche le maximum de

covariance entre le spectre et les concentrations, la qualité du modèle obtenu peut être améliorée en augmentant le nombre des échantillons d'étalonnage utilisés.

Conclusion générale

La méthode de régression PLS peut être utilisée comme méthode de quantification en spectrométrie XRF. L'étalonnage est l'étape la plus cruciale de cette technique. C'est en effet lors de cette étape qu'on détermine les valeurs optimales des paramètres qui entrent en jeu dans la construction d'un modèle PLS. La sélection des variables à utiliser (délimitation de la plage d'énergies ou de canaux) lors de l'étalonnage est aussi très importante. En ne prenant que les variables les plus pertinentes pour chacun des éléments à étudier, on améliore la qualité du modèle, c'est-à-dire, son pouvoir de prédiction. Les avantages des prétraitements des données comme le centrage, la prise de la racine carrée et le lissage ont aussi été démontrés.

Les résultats que nous avons obtenus lors de l'application de cette méthode en TXRF confirment la potentialité de la régression PLS dans le domaine de l'analyse quantitative en spectroscopie XRF. La visualisation des coefficients de régression illustre très bien l'existence de la relation de régression entre le spectre et les concentrations. Nous avons aussi pu prouver qu'on peut avoir une exactitude satisfaisante des prédictions dans des bonnes conditions d'étalonnage. Une de ces conditions est que la concentration d'un élément doit changer le plus possible dans le plus d'échantillons d'étalonnage possible. Cette condition a été satisfaite dans la première série de mesures que nous avons effectuées où on n'a eu que le Cu et le Zn qui avaient des proportions majeures dans 10 échantillons. Les résultats numériques ont donné des écarts relatifs de prédiction de 5,24% pour le Cu et de 7,16% pour le Zn en ce qui concerne la première série de mesures. Pour la deuxième série, ces écarts sont les suivants : 10,85% pour le Cu, 6,93% pour le Zn, 9,42% pour l'As, 8,25% pour le Se et 14,90% pour le Ni. Ces résultats ont tous été obtenus à partir de 10 échantillons d'étalonnage qui s'avèrent insuffisants. On peut donc améliorer ces résultats en utilisant plus d'échantillons pour l'étalonnage. On peut trouver par exemple dans la littérature l'utilisation de 52 échantillons pour un étalonnage dans le domaine de la spectrophotométrie [24], 131 dans le domaine de la fluorométrie [8] ou encore 69 échantillons dans un travail sur la spectrométrie proche infrarouge [3].

Tous les résultats présentés ici ont été obtenus en utilisant le logiciel X-PLS v1.0 que nous avons développé dans le cadre de ce travail. Ce logiciel permet d'effectuer toutes les tâches

requises pour l'étalonnage et aussi la prédiction des concentrations dans des échantillons inconnus. Il offre en outre la possibilité d'enregistrer dans un fichier les résultats d'un étalonnage pour qu'on puisse les utiliser ultérieurement pour la détermination des concentrations des éléments d'intérêt dans les conditions adéquates. L'utilisation de ce logiciel illustre très bien l'atteinte des buts principaux de ce travail qui sont le gain de temps et la facilité d'utilisation. L'analyse quantitative se résume en effet aux étapes suivantes : sélection du fichier contenant le modèle PLS adéquat, sélection des fichiers contenant les spectres des échantillons à analyser et enfin lancement de la prédiction par un simple clic.

Ce travail est le premier dans l'application de la régression PLS en TXRF. Le logiciel XPLS v1.0 est aussi le premier logiciel de quantification par régression PLS consacré spécialement à la spectroscopie XRF. Plusieurs travaux peuvent encore être effectués dans le but d'améliorer les résultats obtenus.

Dans le domaine de l'étalonnage par exemple, on peut utiliser des échantillons standard dans lesquels on ne fait varier que la concentration de l'élément d'intérêt tout en gardant les autres constantes. Ceci nécessite la préparation de plusieurs échantillons. Nous proposons au moins 10 échantillons par élément. Les résultats que nous avons obtenus sont relatifs à un seul type de matrice. D'autres types de matrices doivent aussi être étudiées.

Une autre perspective plus intéressante est l'utilisation de la méthode de Monte Carlo pour simuler des spectres XRF. Ceci peut très bien résoudre les problèmes liés aux préparations d'un nombre important d'échantillons : insuffisance des produits chimiques, erreurs commises lors des préparations, erreurs introduites par les appareils de mesures (spectromètre surtout). On peut en effet, en utilisant cette méthode, générer par ordinateur des spectres avec des concentrations bien déterminées de tous les éléments d'intérêt. On peut ainsi simuler toutes les conditions de mesure et d'échantillonnage possibles et effectuer les étalonnages correspondants. Les résultats ainsi obtenus peuvent être par la suite compilés pour former une base de données d'étalonnage qu'on peut utiliser pour déterminer les concentrations des éléments d'intérêt dans tout échantillon préparé et mesuré dans n'importe quelle condition.

Enfin, on peut aussi envisager l'application du présent travail pour d'autres méthodes d'analyse en spectroscopie XRF mais aussi et surtout dans d'autres domaines comme la spectroscopie gamma ou la Résonance Magnétique Nucléaire (RMN) par exemple.

Références

[1] Hervé ABDI, **Partial Least Squares (PLS) Regression**, In: Lewis-Beck M., Bryman, A., Futing T. (Eds.). Encyclopedia of Social Sciences Research Methods. (2003)

[2] M. J. ADAMS and J. R. ALLEN, **Quantitative X-ray fluorescence analysis of geological materials using partial least squares regression**, Journal Analyst, Vol. 123, 537-541, (April 1998)

[3] S. AJI et al, **Apport du bootstrap à la régression PLS : application à la prédiction de la qualité des gazoles**, Oil & Gas Science and Technology – Rev. IFP, Vol. 58, No. 5, 599-608, (2003)

[4] Shaul BARKAN et al, **VORTEX™ – A new high performance silicon drift detector for XRD and XRF applications**, JCPDS - International Centre for Diffraction Data, Advances in X-ray Analysis, Volume 46, 332-337, (2003)

[5] Tathagata BATTACHARYA, **Prediction of Silicon Content in Blast Furnace Hot Metal Using Partial Least Squares (PLS)**, ISIJ International, Vol. 45, No. 12, 1943-1945, (2005)

[6] Philip R.BEVINGTON, **Data reduction and Error Analysis for the Physical Sciences**, McGraw-Hill Inc., (1969)

[7] Karl S. BOOKSH, **Chemometric Methods in Process Analysis**, Encyclopedia of Analytical Chemistry, R.A. Meyers (Ed.), 8145–8169, (2000)

[8] Rasmus BRO et al, **Standard error of prediction for multilinear PLS: 2. Practical implementation in fluorescence spectroscopy** , Chemometrics and Intelligent Laboratory Systems 75, 69–76, (2005)

[9] A.M.C. DAVIES, **Uncertainty Testing in PLS regression**, Spectroscopy Europe 13/2 2001, 16-19 (2001)

[10] Michael C. DENHAM, **Prediction intervals in partial least squares**, Journal of Chemometrics, Vol 11, 39-52 (1997)

[11] Jan S. IWANCZYK & Bradley E. PATT, **New X-ray detectors for XRF analysis**, JCPDS-International Centre for Diffraction Data 1999, 951-957, (1999)

[12] N. MAJCEN et al., **Linear and non-linear multivariate analysis in the quality control of industrial titanium dioxide white pigment**, Analytica Chimica Acta 348, 87-100, (1997)

[13] R. MANNE, **Analysis of two partial-least-squares algorithms for multivariate calibration**, Chemometrics and Intelligent Laboratory Systems 2, 187-197, (1987)

[14] Brian D. MARX and Paul H. C. EILERS, **Multivariate calibration stability: a comparison of methods**, Journal of Chemometrics, 129-140, (2002)

[15] Boaz NADLER and Ronald R. COIFMAN, **The prediction error in CLS and PLS: the importance of feature selection prior to multivariate calibration**, J. Chemometrics 19, 107–118, (2005)

[16] Alejandro C. OLIVIERI et al. , **Uncertainty estimation and figures of merit for multivariate calibration**, Pure Appl. Chem., Vol. 78, No. 3, 633–661, (2006)

[17] Sven SERNEELS, Christophe CROUX, P.J. VAN ESPEN, **Influence properties of partial least squares Regression**, Chemometr. Intell. Lab. Syst., 71, 13-20, (2004)

[18] S. SERNEELS, P. LEMBERGE and P. J. Van Espen, **Sample specific prediction intervals in SIMPLS**, PLS and related methods, DECISIA, Levallois Perret, France, 219-233, (2003)

[19] A. SOLTANI et al., **Prediction of viability of oriental beechnuts using NIR spectroscopy and PLS regression**, J. Near Infrared Spectroscopy 11, 357 – 364, (2003)

[20] M. TENENHAUS, **La régression PLS: théorie et pratique**, Editions TECHNIP, (1998)

[21] Volker THOMSEN and Debbie SCHATZLEIN, **Advances in Field-Portable XRF**, Spectroscopy 17(7), 14-21, (July 2002)

[22] P. VAN ESPEN, P. LEMBERGE, **ED-XRF spectrum evaluation and quantitative analysis using multivariate and nonlinear techniques**, JCPDS-International Centre for Diffraction Data, Advances in X-ray Analysis,Vol.43, (2000)

[23] René E. VAN GRIEKEN, Andrzej A. MARKOWICZ (eds), **Handbook of X-Ray Spectrometry**, Practical spectroscopy series Vol 14, 1993

[24] Svante WOLD, **The PLS method – partial least squares projections to latent structures – and its applications in industrial RDP(research, development and production),** Proceedings of COMPSTAT 2004, PRAGUE, (2006)

[25] **Eclipse III: X-Ray Tube System for Portable XRF**, Amptek Inc., (Avril 2006)

[26] **Numerical recipes in C: The art of Scientific Computing**, Cambridge University Press, (1988)

[27] **X-123: Complete X-ray spectrometer**, Amptek Inc., (Novembre 2006)

[28] **XR-100CR technical specifications**, Amptek Inc., (Octobre 2006)

Annexes

A.1. Les variables principales :

La variable objet (classe) cCalib contient toutes les données concernant un étalonnage en cours. Nous utilisons une instance nommée CalibData de cette classe dans tout le programme. Cette variable a été construite dans le but de regrouper tous les paramètres relatifs à un étalonnage dans une seule variable et de pouvoir les stocker dans un fichier par la suite pour qu'on puisse les utiliser ultérieurement.

```
class cCalib
{
public:

char ElementList[20][2],      //noms des éléments d'intérêt
UserName[50],                 //nom de l'utilisateur
Date[12],                     //date de l'étalonnage
SampleType[20],               //type d'échantillons utilisés (liquide,…)
DetType[10],                  //type de détecteur
Method[40];                   //méthode de mesure (TXRF par ex.)
int nbSamp,                   //nombre d'échantillons
nbCons,                       //nombre d'éléments
nbChan,                       //nombre de canaux pour le spectre
nbComp,                       //nombre de composantes PLS
nbBoot,                       //nombre d'échantillons bootstrap
ch_beg,ch_end,                // delimitation de la plage de canaux
TubeHV, TubeCurrent,          //information sur le tube R-X
DetHV,                        //alimentation du détecteur
PreProcMethod;                //méthode de prétraitement des données
bool UsePreProc;              //utiliser ou non le prétraitement
double meanY[20],             //moyennes des Y
BCoeff[2000][20],             //coefficient de régression
meanX[2000],                  //moyennes des X
VarX[2000],                   //variances des X
VarY[20],                     //variances des Y
e[100][20];                   //résidus

};
```

Les spectres et les concentrations sont quant à eux stockés dans les variables tableaux déclarées de la manière suivante :

```
double
X[100][2000][10],             //X[nombre d'échantillons][nombre de
                              canaux][nombre de composantes]
Y[100][20][10] ;              //Y[nombre d'échantillons][nombre
                              d'éléments][nombre de composantes]
```

A.2. Obtention de [X] à partir d'un fichier:

La fonction OpenSPEfile( ) extrait les données d'un spectre contenues dans un fichier de type SPE. Elle stocke ces données dans une variable tampon Xtmp. Elles seront par la suite transférées dans les variables [X]. Cette fonction nécessite l'utilisation de fstream.h de C++. Les paramètres sont le nom du fichier (NameFile) et un numéro identifiant l'échantillon en cours.

```
void OpenSPEfile(AnsiString NameFile, int SampleNum)

{
char buf[6]="",name[50]="",ch=' ',ch1=' ',*endptr;

AnsiString Val="";
int i,LineNum=0,nbChan1=1;
double *tmp;


tmp=(double*)malloc(2000*sizeof(double));

for (i=1;i<= NameFile.Length();i++)
    name[i-1]=NameFile[i];
ifstream fin(name);
if (fin.good())
   {
   char ptrch[6]="";
    while(fin.get(ch))
    {
      *ptrch=ch;
      if (ch=='\n')
          {
          LineNum++;
          if (LineNum==9)
             ch1=' ';
          }
      if (LineNum>=9)
         {
          if (ch!=' ')
             if (ch!='\n')
                {
                *ptrch=ch;
                strcat(buf,ptrch);
                }
          if (ch==' ')
             if (ch1!=' ')
                if (LineNum==9)
                   {
                   if (ch1!='\n')
                      {
                      tmp[nbChan1] = strtod(buf, &endptr);
                      nbChan1++;
                      *ptrch=' ';
                      strcpy(buf,ptrch);
                      }
```

```
                    }
                else
                    {
                    tmp[nbChan1] = strtod(buf, &endptr);
                    nbChan1++;
                    *ptrch=' ';
                    strcpy(buf,ptrch);
                    }
          }
      ch1=ch;
    }
   strcat(buf,ptrch);
   tmp[nbChan1] = strtod(buf, &endptr);

   }
fin.close();


// check for all samples if the number of channels are the same (as the
first)

if (SampleNum==1)
    {
    CalibData.nbChan = nbChan1;
    }


for (i=1;i<=CalibData.nbChan;i++)
    Xtmp[SampleNum][i]=tmp[i];
free(tmp);
}
```

A.3. Décomposition des données :

La fonction pls2 suivante constitue le cœur du logiciel X-PLS v1.0. Elle calcule toutes les composantes PLS à partir desquelles le coefficient de régression est obtenu. L'algorithme suivi est présenté dans le paragraphe §2.5.1.2 (p. 34). Cette fonction prend en paramètres les matrices [X] (spectres) et [Y] (concentrations) et le nombre d'échantillons d'étalonnage.

```
void pls2(double X1[20][2000][10],double Y1[20][20][10],int SampNb)
{

for (int h=1;h<=CalibData.nbComp;h++)
    {
    bool conv;
    double uu=0,wh,wh1, dw;
    for (int i=1;i<=SampNb;i++)
        {
        u[i][h]=Y1[i][1][h];
        uu+=pow(u[i][h],2);
        }
    do
        {
        conv=true;
        for (int i=1;i<=CalibData.nbChan;i++)
            {
            w[i][h]=0;
            for (int j=1;j<=SampNb;j++)
                w[i][h]+=(X1[j][i][h]*u[j][h])/uu;
            }
        // ----Norm w
        norm=0;
        for (int i=1;i<=CalibData.nbChan;i++)
            norm+=pow(w[i][h],2);
        norm=sqrt(norm);
        for (int i=1;i<=CalibData.nbChan;i++)
            w[i][h]/=norm;
        //-----

        //----- t components
        uu=0;
        for (int i=1;i<=CalibData.nbChan;i++)
            uu+=pow(w[i][h],2);

        for (int i=1;i<=SampNb;i++)
            {
            t[i][h]=0;
            for (int j=1;j<=CalibData.nbChan;j++)
                t[i][h]+=(X1[i][j][h]*w[j][h])/uu;
            }

        //----- c components
        uu=0;
        for (int i=1;i<=SampNb;i++)
            uu+=pow(t[i][h],2);

        for (int i=1;i<=CalibData.nbCons;i++)
```

```
            {
            c[i][h]=0;
            for (int j=1;j<=SampNb;j++)
                c[i][h]+=(Y1[j][i][h]*t[j][h])/uu;
            }

        //----- u components
        uu=0;
        for (int i=1;i<=CalibData.nbCons;i++)
            uu+=pow(c[i][h],2);

        for (int i=1;i<=SampNb;i++)
            {
            u[i][h]=0;
            for (int j=1;j<=CalibData.nbCons;j++)
                u[i][h]+=(Y1[i][j][h]*c[j][h])/uu;
            }
        //----- new w components
        uu=0;
        for (int i=1;i<=SampNb;i++)
            uu+=pow(u[i][h],2);

        for (int i=1;i<=CalibData.nbChan;i++)
            {
            wnew[i]=0;
            for (int j=1;j<=SampNb;j++)
                wnew[i]+=(X1[j][i][h]*u[j][h])/uu;
            }
        // ----Norm wnew
        norm=0;
        for (int i=1;i<=CalibData.nbChan;i++)
            norm+=pow(wnew[i],2);
        norm=sqrt(norm);
        for (int i=1;i<=CalibData.nbChan;i++)
            wnew[i]/=norm;

        //------ Convergence

        wh=0;
        wh1=0;
        if (CalibData.nbCons>1)
            {
            for (int i=1;i<=CalibData.nbChan;i++)
                {
                dw=wnew[i]-w[i][h];
                wh+=pow(dw,2);
                wh1+=pow(wnew[i],2);
                }
            if (fabs(sqrt(wh)/sqrt(wh1))>eps)
                conv=false;
            }

    } while (!conv);

    //----- p components
    uu=0;
    for (int i=1;i<=SampNb;i++)
        uu+=pow(t[i][h],2);

    for (int i=1;i<=CalibData.nbChan;i++)
        {
```

```
        p[i][h]=0;
        for (int j=1;j<=SampNb;j++)
            p[i][h]+=(X1[j][i][h]*t[j][h])/uu;
        }

    //---- Deflations

    for (int i=1;i<=SampNb;i++)
        for (int j=1;j<=CalibData.nbChan;j++)
            X1[i][j][h+1]=X1[i][j][h]-(t[i][h]*p[j][h]);

    for (int i=1;i<=SampNb;i++)
        for (int j=1;j<=CalibData.nbCons;j++)
            Y1[i][j][h+1]=Y1[i][j][h]-(t[i][h]*c[j][h]);
    }
}
```

A.4. Calcul du coefficient de régression :

La fonction GetBCoeff( ) calcule le coefficient de régression à partir des composantes PLS $\overline{W}, \overline{P}, \overline{C}$. On utilise pour cela la relation (2.35 p. 35). Le coefficient obtenu est stocké dans la variable tableau BStar. Ce coefficient peut être le coefficient du modèle (cas où BootNb = 0) ou le coefficient d'un échantillon bootstrap (BootNb > 0). La fonction gaussj( ) utilise la méthode d'élimination de Gauss-Jordan [26] pour la résolution d'équations algébriques linéaires. Dans notre cas, elle est utilisée pour obtenir l'inverse de la matrice $\overline{P}^t \overline{W}$.

```
void GetBCoeff(int BootNb)

{
double Wstar[2000][10],temp[10];

if (CalibData.nbComp > 1)
    {
    for (int i = 1; i <= CalibData.nbComp; i++)
        for (int j = 1; j <= CalibData.nbComp; j++)
            {
            if (i == j) PW[i][j] = 1;
            else if (j < i) PW[i][j] = 0;
            else if (j > i + 1) PW[i][j] = 0;
            else if (j == i + 1)
                {
                PW[i][j] = 0;
                for (int k = 1; k <= CalibData.nbChan; k++)
                    PW[i][j] += p[k][i] * w[k][j];
                }
            }
    for (int i = 1; i <= CalibData.nbComp; i++)
        temp[i] = t[1][i];
    gaussj(PW,CalibData.nbComp,temp); //get inverse of p'w
    for (int i = 1; i <= CalibData.nbChan; i++)
        for (int j = 1; j <= CalibData.nbComp; j++)
            {
            Wstar[i][j] = 0;
            for (int k = 1; k <= CalibData.nbComp; k++)
                Wstar[i][j] += w[i][k] * PW[k][j];
            }
    for (int i = 1; i <= CalibData.nbChan; i++)
        for (int j = 1; j <= CalibData.nbCons; j++)
            {
            Bstar[i][j][BootNb] = 0;
            for (int k = 1; k <= CalibData.nbComp; k++)
                Bstar[i][j][BootNb] += Wstar[i][k] * c[j][k];
            }
    }
else
    {
    for (int i = 1; i <= CalibData.nbChan; i++)
        for (int j = 1; j <= CalibData.nbCons; j++)
            Bstar[i][j][BootNb] = w[i][1] * c[j][1];
    }
}
```

A.5. Prédiction:

La fonction ShortPredict( ) assure la prédiction en effectuant la multiplication du spectre de l'échantillon inconnu par le coefficient de régression.

```
void ShortPredict()
{
for (int k = 1; k <= CalibData.nbCons; k++)
    for (int i = 1; i <= nbUSamp; i++)
        {
        UYe[i][k] = 0;
        for (int j = 1; j <= CalibData.nbChan; j++)
            UYe[i][k] += Xtmp[i][j] * CalibData.BCoeff[j][k];
        }
}
```

A.6. Estimation de l'intervalle de prédiction par la méthode bootstrap :

La fonction bootstrap implémente la méthode bootstrap pour l'évaluation de l'intervalle de prédiction. Elle utilise essentiellement les fonctions randomize( ) et random( ) de time.h pour la génération de nombre aléatoire nécessaire pour le tirage avec remise de la méthode bootstrap.

```
void bootstrap()
{
double tmp;
//int p;
Screen->Cursor = crHourGlass;
randomize();
//--Generate the bootstrap samples: Y*=XB+y*
for (int k = 1; k <= CalibData.nbBoot; k++)
    for (int i = 1; i <= CalibData.nbSamp; i++)
        for (int j = 1; j <= CalibData.nbCons; j++)
            Ystar[i][j][k] = Ye[i][j][CalibData.nbComp] +
Y[random(CalibData.nbSamp) + 1][j][CalibData.nbComp + 1];

for (int k = 1; k <= CalibData.nbBoot; k++)
    {
    //--transfer the k-th sample to the buffer (Ytmp)
    for (int i = 1; i <= CalibData.nbSamp; i++)
        for (int j = 1; j <= CalibData.nbCons; j++)
            Ytmp[i][j][1] = Ystar[i][j][k];
    pls2(X,Ytmp,CalibData.nbSamp);
    GetBCoeff(k); //k-th bootstrap regression coeff.
    }

//Generate bootstrap replications for unknown samples
randomize();
for (int k = 1; k <= CalibData.nbBoot; k++)
    for (int i = 1; i <= nbUSamp; i++)
        for (int j = 1; j <= CalibData.nbCons; j++)
            UYeStar[i][j][k] = UYe[i][j] + Y[random(CalibData.nbSamp) +
1][j][CalibData.nbComp + 1];
```

```
//Generate bootstrap replications of the predictions
// use btemp to hold these variables
for (int l = 1; l <= CalibData.nbBoot; l++)
    for (int k = 1; k <= CalibData.nbCons; k++)
        for (int i = 1; i <= nbUSamp; i++)
            {
            btemp[i][k][l] = 0;
            for (int j = 1; j <= CalibData.nbChan; j++)
                btemp[i][k][l] += Xtmp[i][j] * Bstar[j][k][l];
            }

//Adjust UYeStar and btemp according to the preprocessing method used
if (CalibData.UsePreProc)
    {
    for (int k = 1; k <= CalibData.nbBoot; k++)
        {
        for (int i = 1; i <= nbUSamp; i++)
            {
            for (int j = 1; j <= CalibData.nbCons; j++)
                {
                if (CalibData.PreProcMethod == 0)
                    {
                    UYeStar[i][j][k] = UYeStar[i][j][k] *
CalibData.VarY[j] + CalibData.meanY[j];
                    btemp[i][j][k] =  btemp[i][j][k] * CalibData.VarY[j]
+ CalibData.meanY[j];
                    }
                else if (CalibData.PreProcMethod == 1)
                    {
                    UYeStar[i][j][k] = pow((UYeStar[i][j][k] +
CalibData.meanY[j]),2);
                    btemp[i][j][k] =  pow((btemp[i][j][k] +
CalibData.meanY[j]),2);
                    }
                }
            }
        }
    }

//Get the bootstrap replications of the prediction errors

for (int k = 1; k <= CalibData.nbBoot; k++)
    for (int i = 1; i <= nbUSamp; i++)
        for (int j = 1; j <= CalibData.nbCons; j++)
            eStar[i][j][k] = UYeStar[i][j][k] - btemp[i][j][k];

//Sort eStar
for (int j = 1; j <= nbUSamp; j++)
    for (int k = 1; k <= CalibData.nbCons; k++)
        for (int p = CalibData.nbBoot - 1; p > 1; p--)
            for (int i = 1; i <= p; i++)
                if (eStar[j][k][i] >= eStar[j][k][i + 1])
                    {
                    tmp = eStar[j][k][i + 1];
                    eStar[j][k][i + 1] = eStar[j][k][i];
                    eStar[j][k][i] = tmp;
                    }
Screen->Cursor = crDefault;
}
```

Utilisation de la régression PLS en spectroscopie XRF : développement d'un logiciel d'analyse quantitative et applications à la spectrométrie XRF a réflexion totale.

Résumé : La plupart des applications et des nouvelles générations d'appareils en spectroscopie par fluorescence X nécessitent que les résultats des mesures effectuées soient connus le plus rapidement possible. Le but du présent travail est de proposer une solution alternative pouvant convertir directement le spectre obtenu par le spectromètre en concentrations des éléments à étudier. Cette solution est l'utilisation de la régression PLS qui consiste à établir une relation de régression [Y]=[X][β]+[E] entre les concentrations et les spectres avec [E] le résidu de la régression. Le principe est de déterminer le coefficient [β] en effectuant un étalonnage à l'aide de plusieurs échantillons standard. Ce coefficient sert par la suite à prédire les concentrations dans les échantillons inconnus par simple multiplication matricielle. Pour pouvoir utiliser cette méthode, nous avons développé le logiciel X-PLS v1.0 qui implémente la régression PLS. Ce logiciel permet d'effectuer toutes les tâches requises pour l'étalonnage et aussi la prédiction des concentrations dans des échantillons inconnus. Il offre en outre la possibilité d'enregistrer dans un fichier les résultats d'un étalonnage pour qu'on puisse les utiliser ultérieurement dans des conditions adéquates.
Les résultats que nous avons obtenus lors de l'application de cette méthode en spectroscopie XRF à réflexion totale permettent d'affirmer que la régression PLS peut être utilisée comme méthode de quantification en spectroscopie XRF. La visualisation des coefficients de régression illustre la relation de régression entre le spectre et les concentrations. Nous avons aussi pu prouver qu'on peut avoir une exactitude satisfaisante des prédictions dans des bonnes conditions d'étalonnage. Une de ces conditions est que la concentration d'un élément doit changer le plus possible dans le plus d'échantillons d'étalonnage possible. Cette condition a été satisfaite dans la première série de mesures que nous avons effectuées où on n'a eu que le Cu et le Zn qui avaient des proportions majeures dans 10 échantillons. On a alors obtenu des écarts relatifs de prédiction de 5,24% pour le Cu et 7,16% pour le Zn. Dans la deuxième série de mesures, nous avons eu 5 éléments (Cu, Zn, As, Se, Ni) dans 10 échantillons d'étalonnage. Les écarts relatifs de prédiction ont été les suivants : 10,85% pour le Cu, 6,93% pour le Zn, 9,42% pour l'As, 8,25% pour le Se et 14,90% pour le Ni.

Mots clés: *Régression PLS, spectroscopie XRF, TXRF, méthode de quantification*

Abstract: Most of the applications and new generation of instruments in X ray Fluorescence spectroscopy require that the results of measurements be known as soon as possible.
The present work aims to introduce an alternative solution allowing the direct conversion of spectra obtained by the XRF spectrometer to elements concentrations. This solution is the use of PLS regression which consists of establishing a regression relation [Y]=[X][β]+[E] between concentrations and spectra where [E] is the residual of the regression. The principle is to determine the coefficient [β] by mean of calibration with standard samples. This coefficient is used after to predict elements concentrations in unknown samples by a simple matrix multiplication. In order to use this method, we have developed the X-PLS v1.0 software which implements the PLS regression. This software allows to perform all required tasks for the calibration and also the prediction of concentrations in unknown samples. Besides, it gives the possibility to save the calibration results in a file so that they can be used later in adequate conditions.
The results obtained during the application of this method in total reflection X ray spectrometry can confirm that the PLS regression can be used as quantification method in XRF spectroscopy. The graphical representation of the regression coefficient illustrates the existence of a regression relation between spectra and concentrations. We could also prove that satisfactory accuracy can be obtained for the prediction in good conditions. One of these conditions is that the concentration of the element of interest must change as much as possible in as much samples as possible. This condition has been satisfied during the first set of measurements in which only the elements Cu and Zn had the major proportions within 10 samples. Relative errors obtained for prediction were 5,24% for the Cu and 7,16% for the Zn. For the second set, we used 5 elements (Cu, Zn, As, Se, Ni) in 10 samples. Results obtained for the predictions have had the following relative errors: Cu: 10,85% , Zn : 6,93%, As : 9,42% , Se : 8,25% and 14,90% for the Ni.

Keywords: *PLS regression, XRF spectroscopy, TXRF, quantification method*

Directeurs de thèse :	**Impétrant :** RAKOTONDRAJOA Andrianiaina
RABOANARY Roland	Tél : (261) 32 04 931 73
	e-mail : andry_gasy@yahoo.fr
RAOELINA ANDRIAMBOLOLONA	Lot II O 9 bis Anjanahary, Antananarivo 101 Madagascar