Chimiométrie appliquée à la spectroscopie de plasma induit par laser (LIBS) et à la spectroscopie terahertz

Josette El Haddad

THÈSE

PRÉSENTÉE A

L'UNIVERSITÉ BORDEAUX 1

ÉCOLE DOCTORALE DES SCIENCES PHYSIQUES ET DE L'INGÉNIEUR

Par Josette, EL HADDAD

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : Doctorat en laser, matière, nanosciences

CHIMIOMETRIE APPLIQUÉE À LA SPECTROSCOPIE DE PLASMA INDUIT PAR LASER (LIBS) ET À LA SPECTROSCOPIE TERAHERTZ

Directeurs de recherche : Bruno Bousquet et Lionel Canioni

Soutenue le : 13 Décembre 2013

Devant la commission d'examen formée de :

M. GIAMARCHI, Philippe	Professeur, Université de Bretagne Occidentale	Rapporteur
M. FORNI, Olivier	Chargé de recherche, CNRS	Rapporteur
M. DOUCET, François	Agent de recherche, CNRC, Québec	Examinateur
M. SARACCO, Jérôme	Professeur, Université Bordeaux 1	Examinateur
M. CANIONI, Lionel	Professeur, Université Bordeaux 1	Co-directeur de thèse
M. BOUSQUET, Bruno	Maîtres de conférences, Université Bordeaux 1	Co-directeur de thèse
M. GALLOU, Dominique	Docteur, IVEA solutions	Invité

CHIMIOMETRIE APPLIQUEE A LA SPECTROSCOPIE DE PLASMA INDUIT PAR LASER (LIBS) ET A LA SPECTROSCOPIE TERAHERTZ

L'objectif de cette thèse était d'appliquer des méthodes d'analyse multivariées au traitement des données provenant de la spectroscopie de plasma induit par laser (LIBS) et de la spectroscopie térahertz (THz) dans le but d'accroître les performances analytiques de ces techniques.

Les spectres LIBS provenaient de campagnes de mesures directes sur différents sites géologiques. Une approche univariée n'a pas été envisageable à cause d'importants effets de matrices et c'est pour cela qu'on a analysé les données provenant des spectres LIBS par réseaux de neurones artificiels (ANN). Cela a permis de quantifier plusieurs éléments mineurs et majeurs dans les échantillons de sol avec un écart relatif de prédiction inférieur à 20% par rapport aux valeurs de référence, jugé acceptable pour des analyses sur site. Dans certains cas, il a cependant été nécessaire de prendre en compte plusieurs modèles ANN, d'une part pour classer les échantillons de sol en fonction d'un seuil de concentration et de la nature de leur matrice, et d'autre part pour prédire la concentration d'un analyte. Cette approche globale a été démontrée avec succès dans le cas particulier de l'analyse du plomb pour un échantillon de sol inconnu. Enfin, le développement d'un outil de traitement par ANN a fait l'objet d'un transfert industriel.

Dans un second temps, nous avons traité des spectres d'absorbance terahertz. Ce spectres provenaient de mesures d'absorbance sur des mélanges ternaires de Fructose-Lactose-acide citrique liés par du polyéthylène et préparés sous forme de pastilles. Une analyse semi-quantitative a été réalisée avec succès par analyse en composantes principales (ACP). Puis les méthodes quantitatives de régression par moindres carrés partiels (PLS) et de réseaux de neurons artificiels (ANN) ont permis de prédire les concentrations de chaque constituant de l'échantillon avec une valeur d'erreur quadratique moyenne inférieure à 0.95 %. Pour chaque méthode de traitement, le choix des données d'entrée et la validation de la méthode ont été discutés en détail.

Mots clés: chimiométrie, spectroscopie de plasma induit par laser (LIBS), spectroscopie THz, analyse multivariée, analyse en composantes principales (ACP), régression par moindres carrés partiels (PLS), réseaux de neurones artificiels (ANN), analyse quantitative et semi-quantitative, classement.

CHEMOMETRIC APPLIED TO LASER-INDUCED BREAKDOWN SPECTROSCOPY (LIBS) AND TERAHERTZ SPECTROSCOPY

The aim of this work was the application of multivariate methods to analyze spectral data from laser-induced breakdown spectroscopy (LIBS) and terahertz (THz) spectroscopy to improve the analytical ability of these techniques.

In this work, the LIBS data were derived from on-site measurements of soil samples. The common univariate approach was not efficient enough for accurate quantitative analysis and consequently artificial neural networks (ANN) were applied. This allowed quantifying several major and minor elements into soil samples with relative error of prediction lower than 20% compared to reference values. In specific cases, a single ANN model didn't allow to successfully achieving the quantitative analysis and it was necessary to exploit a series of ANN models, either for classification purpose against a concentration threshold or a matrix type, or for quantification. This complete approach based on a series of ANN models was efficiently applied to the quantitative analysis of unknown soil samples. Based on this work, a module of data treatment by ANN was included into the software Analibs of the IVEA company.
The second part of this work was focused on the data treatment of absorbance spectra in the terahertz range. The samples were pressed pellets of mixtures of three products, namely fructose, lactose and citric acid with polyethylene as binder. A very efficient semi-quantitative analysis was conducted by using principal component analysis (PCA). Then, quantitative analyses based on partial least squares regression (PLS) and ANN allowed quantifying the concentrations of each product with a root mean square error (RMSE) lower than 0.95 %. All along this work on data processing, both the selection of input data and the evaluation of each model have been studied in details.

Keywords: chemometrics, laser-induced breakdown spectroscopy (LIBS), Terahertz (THz) spectroscopy, multivariate analysis, quantification, principal component analysis (PCA), partial least square regression (PLS), artificial neural networks (ANN), quantitative, semi-quantitative, classification analysis

Remerciements

Cette thèse a été menée au Laboratoire Ondes et Matières d'Aquitaine (LOMA). Je tiens tout d'abord à remercier l'ADEME pour avoir cofinancé mon travail de thèse. Je tiens à exprimer ma profonde gratitude à Monsieur Dominique GALLOU pour la confiance qu'il m'a accordée en acceptant de cofinancer via la société IVEA Solutions, ma troisième année de thèse. Je remercie le directeur du LOMA, Monsieur Jean-Pierre DELVILLE de m'accueillir dans son laboratoire.

Je suis très sensible à l'honneur que m'ont fait les membres du jury en acceptant de juger ce travail. Je remercie tout d'abord Messieurs Philippe GIAMARCHI et Olivier FORNI d'avoir analysé mon mémoire de thèse en qualité de rapporteurs. Je remercie également Monsieur Jérôme SARACCO pour avoir évalué mon travail et présidé le jury lors de ma soutenance de thèse. Enfin j'adresse ma profonde gratitude à Monsieur François DOUCET pour avoir accepté de juger mon travail, et de participer à jury par visioconférence.

Je remercie très chaleureusement mon directeur Monsieur Bruno BOUSQUET pour la confiance, le soutien et l'autonomie qu'il m'a accordés. P.S. Le réseau de neurone a peur de lui car l'ANN hésite de se bloquer devant lui. Votre aide pour la rédaction du mémoire a été précieuse. Durant ces trois années, il y avait des moments inoubliables ! Vous avez été sans cesse motivant et souriant même dans les moments de stress, incroyable !!

Je remercie également Monsieur Lionel CANIONI de m'avoir accueillie dans son groupe de recherche. Je salue sa grande culture scientifique et sa disponibilité malgré un emploi de temps surchargé. J'exprime aussi ma reconnaissance à Monsieur Patrick MOUNAIX pour l'attention qu'il a portée à mon travail et pour sa confiance qui m'a conduite à appliquer certaines méthodes de chimiométrie à la spectroscopique térahertz. Il a apporté une grande richesse à ce mémoire et je le remercie par ailleurs pour sa sympathie. Il était le gardien de l'équipe, toujours présent à côté de nous. Je remercie également Monsieur Jean-Baptiste SIRVEN. J'ai en effet commencé ma thèse en poursuivant les travaux qu'il avait initiés au cours de sa thèse et en particulier en m'appuyant sur l'algorithme ANN qu'il avait programmé sous Igor et qui s'est avéré plus performant que celui d'un logiciel commercial.

Je tiens encore à remercier toute l'équipe d'IVEA. Merci à Guillaume GALLOU, chef du projet CALIPSO, pour son aide et pour les mesures LIBS exploitées au cours de ce travail de thèse. Merci à Christopher FORGERON, pour l'implémentation de l'algorithme ANN dans le logiciel Analibs. Je remercie vivement Amina ISMAEL pour toutes les informations qu'elle m'a transmises concernant le traitement des données, les manipulations LIBS ainsi que pour son accompagnement au début de ma thèse.

Je tiens aussi à remercier l'équipe du BRGM, Valérie LAPERCHE, Delphine BRUYERE et Karine MICHEL. Merci pour votre contribution à ce travail de recherche à travers des prélèvements et des analyses sur site ainsi que pour les informations géologiques que vous m'avez transmises.

J’ai par ailleurs la chance de connaître Myriam BOUERI, ma sœur en France. Elle m’a appris que l’obstacle n’existe plus devant moi GRAZIE. Quant à Jean-Paul GUILLET ☺, mon frère (LOCA), merci pour tout le temps que tu as perdu en m’aider 谢谢.

Je remercie enfin tous les membres du laboratoire, en particulier les secrétaires (Isabelle, Laurette, Suzanne, Annie, et Sophie) ainsi que les informaticiens Hassan et Richard qui m’ont témoigné de la sympathie. De plus, je tiens à saluer Richard, Fabien et Rokhaya pour notre amitié. On a partagé ensemble nos repas du midi et grâce à eux, j’ai oublié le stress du travail et même le riz sec avait du goût. Merci Richard, tu as été le père de tous les étudiants et surtout le mien.

Mes remerciements vont à tous les membres de l’équipe SLAM : Inka, Yannick (x2), Nicolas, Gautier, Riad, Ayesha, Arnaud, Nadezda, Marie, Hugo, Konstantin, Jean-Baptiste, Frederick, Léna et Joyce. Inka, ta gentillesse va me manquer ; Yannick Petit m’a appris comment travailler en silence et patience ; Nicolas m’a appris à être dynamique ; Arnaud, merci pour ton aide pour apprendre à utiliser le logiciel « Igor j’adore » ; Gautier, c’est de te voir travailler qui m’a encouragé à travailler aussi. Frederick je te remercie pour toutes tes mesures Téra et la préparation des échantillons. Et la sweet Joyce :D est mieux que tous les médicaments anti-stress du monde, Thanks Joyci !! Pour tous les efforts et les préparations pour le jour de la soutenance. Slams, vous avez réussi à faire du 13/12/13 un jour inoubliable dans ma vie.

Je remercie aussi tous mes amis qui, par leur aide ou leur sympathie ont participé à la réalisation de cette thèse. Et plus particulièrement merci à Georges Nader.

Merci maman pour toutes tes prières, merci mon père pour m’avoir encouragée à venir en France et à découvrir le monde de la recherche. Merci Elie et Eliane pour le soutien que vous m’avez apporté tout au long de ma thèse.

Enfin, Merci Dieu !! Ta présence à côté de moi, que je ne mériter pas, est une force, ton Esprit est une lumière de joie et de paix. Et Oh ! Toute belle, ma mère (N.D. de Lourdes ; N.D. du Liban) je te salue pour ta protection dans tes larges bras contre la désespérance surtout au début et à la fin de la thèse.

A mon Père,

« Eh bien, moi, je vous dis : Demandez, vous obtiendrez ; cherchez, vous trouverez ; frappez, la porte vous sera ouverte » Jésus [Saint Luc, 11,9]

Table des matières

Table des figures

Liste des tableaux

Liste des acronymes

A	Nombre des composantes principales
ACP	Analyse en composantes principales
ADEME	Agence de l'Environnement et de la Maîtrise de l'énergie
ANN	« Artificial neurals networks » réseaux de neurones artificiels
BRGM	Bureau de Recherches Géologiques et Minières
CCD	« Charge-Coupled Device »
ER	Erreur relative moyenne
ERC	Erreur relative moyenne de calibration
ERT	Erreur relative moyenne de test
ERV	Erreur relative moyenne de validation
ICP-AES	« Inductively coupled plasma - atomic emission spectroscopy »
ICP-MS	« Inductively coupled plasma - mass spectrometry »
LA-ICP-AES	« Laser ablation - inductively coupled plasma - atomic emission spectroscopy »
LIBS	« Laser induced breakdown spectroscopy »
LOMA	Laboratoire Ondes et Matière d'Aquitaine
LOO	« leave one out »
ME	Metaleurope
NIST	National institute of standards and technology
p	«Loadings »
PE	Polyéthylène
PLS	« Partial least square » régression aux moindres carrés partiels
R^2	Coefficient de corrélation
RMSE	« Root mean square error » Erreur quadratique moyenne
RMSEC	Erreur quadratique moyenne de calibration
RMSET	Erreur quadratique moyenne de test
RMSEV	Erreur quadratique moyenne de validation
SEB	Saint-Sébastien d'Aigrefeuille
SLAM	« Short Laser Applications and Materials »

SLM	Saint-Laurent le minier
t	Scores
THz	Térahertz
XRF	Fluorescence X

Introduction

Au cours de ces dernières années, la chimie analytique s'est naturellement orientée vers le développement de techniques de plus en plus rapides et qui demandent un minimum de préparation des échantillons. Au-delà de l'instrumentation, les techniques modernes d'analyse s'appuient sur différents domaines scientifiques tels que l'informatique, les mathématiques et les statistiques dans le but d'accroître les performances analytiques, notamment dans le cas de mesures sur site. Parmi les différentes techniques analytiques, les méthodes de spectroscopie sont parmi les plus prometteuses pour des analyses rapides directement sur site mais les spectres enregistrés contiennent souvent des milliers de valeurs. Par conséquent, afin de tirer le meilleur parti de ce grand nombre de données, l'utilisation de méthodes sophistiquées de calculs mathématiques et statistiques est nécessaire, et c'est là que l'on fait appel à la chimiométrie.

La chimiométrie est définie comme étant le développement et l'application de méthodes mathématiques et statistiques pour extraire un maximum d'informations de mesures chimiques. La première société de chimiométrie fut fondée à Seattle en 1974. Les premières approches de chimiométrie s'appuyaient sur des modèles de reconnaissance non supervisés pour résoudre des problèmes multivariés et ce bien avant l'invention de l'ordinateur et des modèles de prédiction [1]. Notons d'ailleurs que le besoin de faire appel à des méthodes statistiques et mathématiques pour l'extraction des données utiles n'est pas réservé au domaine de la chimie mais au contraire, on le retrouve aussi dans les domaines de la biologie (biométrie), de l'économie (économétrie), ou encore de la psychologie (psychométrie).

La chimiométrie est devenue aujourd'hui une branche à part entière de la chimie. Appliquée surtout à la spectroscopie, elle permet d'établir une relation entre la concentration de l'analyte d'une part et les valeurs d'intensité de quelques raies spectrales préalablement sélectionnées [2]. Le chimiométricien doit avant tout maîtriser les principes et les algorithmes de calcul des différents outils de chimiométrie, parmi lesquels les modèles de compression des données ou de reconnaissance de formes, ainsi que les modèles de régression incluant des méthodes linéaires et non-linéaires. Mais il doit aussi comprendre le problème de chimie posé afin de choisir les outils de traitement les plus adaptés et les plus performants. La chimiométrie est aujourd'hui appliquée à de nombreux domaines de la chimie analytique, notamment aux méthodes séparatives telles que la chromatographie [3, 4] et l'électrophorèse [5] et aux

méthodes spectroscopiques telles que Raman [6], FT-IR [7], NIR [8], VIS-NIR [9], térahertz (THz) [10] ou la spectroscopie de plasma induit par laser (LIBS) [11]. Plusieurs domaines d'application bénéficient déjà des avantages apportés par la chimiométrie comme l'agro-alimentaire [12, 13], l'environnement et les géosciences [14], la pharmaceutique [15] le contrôle réactionnel industriel en temps réel [16], la criminologie [17], l'histoire et l'art [18, 19], ainsi que le domaine militaire et la sûreté nationale via la détection des explosifs [20].

Ces travaux de thèse ont été réalisés au Laboratoire Ondes et Matière d'Aquitaine (LOMA). Ce laboratoire est impliqué dans deux domaines de spectroscopie qui connaissent une réelle effervescence au niveau international ces dernières années, à savoir la LIBS et la spectroscopie THz. Il est important de signaler ici que ce travail de thèse a fait l'objet de deux contrats de recherche. Le premier est un contrat signé avec l'ADEME dans le cadre d'un projet à finalité de transfert industriel intitulé CALIPSO et dédié à l'analyse LIBS sur site de sols pollués en métaux lourds (2011-2012). Un consortium entre le BRGM, spécialiste en géosciences, IVEA entreprise privée commercialisant des systèmes LIBS de terrain et le LOMA pour l'analyse des données a ainsi été constitué pour une durée de deux ans. Le travail engagé a ensuite été complété par un contrat de recherche d'un an (2013) entre le LOMA et l'entreprise IVEA.

- La LIBS est une technique de spectroscopie d'émission atomique qui fournit un spectre dans le domaine UV-visible contenant des centaines de raies atomiques. Il s'agit pour notre équipe de parvenir à analyser des échantillons de sols directement sur site. Plus précisément, on cherche à quantifier des éléments pour le diagnostic des sols pollués.
  Or, devant la diversité des échantillons de sols et les effets de matrice qui empêchent toute analyse univariée, il a été nécessaire de faire appel aux outils de chimiométrie pour atteindre notre objectif. De plus, étant donné le comportement non-linéaire de certaines raies spectrales en fonction de la concentration de l'analyte, nous avons décidé de développer des modèles prédictifs à base de réseaux de neurones artificiels (ANN). On considèrera que les analyses sur site sont acceptables pour un écart relatif de prédiction des concentrations autour de 20%.

- La spectroscopie THz qui permet de sonder le couplage des molécules avec leur environnement, fournit quant à elle un spectre dans le domaine THz souvent avec peu de contraste au niveau des bandes d'absorption. Aucune analyse n'avait encore été menée par chimiométrie dans le cadre de cette spectroscopie avant mon arrivée au laboratoire et il a donc été question de faire une première démonstration de l'intérêt des outils de chimiométrie pour ce type de spectroscopie. C'est pourquoi nos travaux sur le sujet concernent des échantillons d'étude spécialement préparés pour cette démonstration, à savoir des mélanges ternaires de fructose-lactose-acide citrique. Il s'agissait alors de mettre en œuvre des analyses semi-quantitatives et quantitatives de chaque constituant du mélange par des approches multivariées et d'en évaluer les performances.

Le mémoire de thèse présenté ici est donc organisé de la façon suivante :

Dans le chapitre 1, les principes des différentes méthodes de chimiométrie sont présentés. On y aborde notamment l'analyse en composantes principales (ACP), la régression aux moindres carrés partiels (PLS) et les réseaux de neurones artificiels (ANN) non seulement pour du tri mais aussi pour des analyses quantitatives. La méthodologie complète y est détaillée en allant de la sélection des échantillons en lots de calibration, validation et test, à la sélection des variables d'entrée, puis à la validation croisée et enfin à la vérification de la signification statistique des modèles par la procédure de mélange aléatoire des données de sortie (Y-randomization).

Dans le chapitre 2, la chimiométrie est appliquée à la LIBS. On introduit dans un premier temps le principe de cette technique de spectroscopie, puis les campagnes de mesures sur site dédiées à l'analyse des sols, incluant la préparation des échantillons et le protocole de mesure. Les spectres LIBS seront dans un premier temps analysés les uns par rapport aux autres de manière exploratoire à l'aide de la technique non-supervisée d'ACP. L'objectif à ce stade est d'observer si l'on peut distinguer différentes matrices de sols. Le traitement des données par ANN est ensuite discuté pour les constituants majeurs et mineurs des sols tels que l'aluminium, le calcium, le cuivre ou le plomb. Le cas du plomb sera traité avec une attention tout à fait particulière car il permet de discuter non seulement des effets de matrice mais aussi des effets liés à la grande étendue des valeurs de concentrations pour cet élément. Finalement, une stratégie complète d'analyse d'un échantillon inconnu sera présentée, incluant classification et quantification. Nous décrivons aussi dans ce chapitre le transfert Recherche-Industrie qui a été réalisé dans le cadre d'un projet collaboratif et qui a conduit à la conception et au développement d'un module « ANN » intégré dans le logiciel Analibs commercialisé par la société IVEA, spécialiste des analyses LIBS.

Dans le chapitre 3, la chimiométrie est appliquée à la spectroscopie térahertz. Le principe de la mesure du spectre d'absorbance dans le domaine THz est donné dans un premier temps puis on discute d'une méthode de préparation d'une série d'échantillons composés de mélanges ternaires de lactose, fructose et acide citrique. On utilise ensuite l'ACP pour localiser chaque échantillon dans le diagramme ternaire. Il s'agit alors d'une analyse semi-quantitative. Enfin, on applique les méthodes quantitatives PLS et ANN pour déterminer la concentration de chaque constituant dans l'échantillon.

Enfin, dans la conclusion de ce mémoire, nous proposons un résumé des principaux résultats obtenus en LIBS pour des analyses sur site d'échantillons de sols puis nos premiers résultats obtenus en spectroscopie térahertz. Nous discutons enfin des limitations et des perspectives envisagées.

Chapitre 1. **Application de la chimiométrie à la spectroscopie**

1.1 **Introduction**

Le terme chimiométrie a été proposé il y a déjà plus de 30 ans pour décrire les opérations mathématiques associées à l'interprétation des données chimiques [21]. La chimiométrie est basée sur le calcul statistique, aussi bien pour les traitements les plus simples, dits uni-variés car mettant en jeu une seule variable, que pour les traitements multi-variés qui, eux, mettent en jeu plusieurs variables et qui nécessitent des moyens de calcul adaptés.

Par définition, une variable est une grandeur physique ou chimique qui sera mesurée pour chaque échantillon. En spectroscopie, ce sera par exemple l'intensité d'un pic sur un spectre d'émission (cas de la LIBS) ou d'absorption (cas de la spectroscopie THz). Dans le cas d'un ensemble de N échantillons, les N valeurs x_i de la variable X seront caractérisées par la valeur moyenne ainsi que par la variance observée sur l'ensemble des N points et dont les définitions sont rappelées par les équations (**1-1**) et (**1-2**), respectivement.

Moyenne : $$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \quad (1\text{-}1)$$

Variance : $$s_x^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1} \quad (1\text{-}2)$$

Dans le cadre de l'analyse uni-variée, on établit une relation entre une variable d'entrée unique (x) et une variable de sortie unique (y). Dans le cas de la spectroscopie, la variable de sortie est la concentration alors que la variable d'entrée est l'intensité d'une raie d'émission ou d'absorption. Notons que la variable d'entrée peut aussi être la valeur de l'aire sous le pic d'une raie d'intérêt. La courbe qui permet de visualiser la variation de la variable de sortie (y) en fonction de la variable d'entrée (x) est tout simplement la courbe d'étalonnage. C'est la méthode d'analyse la plus simple et aussi la plus couramment utilisée. Par exemple, en LIBS, lorsqu'il s'agit d'analyser la concentration en plomb, on s'appuie sur un étalonnage donnant

la concentration en plomb en fonction de l'intensité de la raie de Pb I à 405,7 nm. Notons qu'il est fréquent d'appliquer un prétraitement aux données, notamment en normalisant par rapport à un étalon interne afin de réduire les fluctuations expérimentales et d'améliorer les performances analytiques. Idéalement, on souhaite obtenir une relation linéaire entre la variable d'entrée (x) et la variable de sortie (y) mais ce n'est pas toujours le cas et certaines courbes détalonnages sont polynomiales [22, 23].

On peut caractériser la relation entre deux variables en calculant la covariance donnée par l'équation (**1-3**). Si la covariance est nulle, on peut en déduire que les variables (x) et (y) ne sont pas corrélées. A l'inverse, plus la covariance augmente et plus les variables sont corrélées. Notons cependant que la covariance souffre des effets d'unités puisque chaque variable (x) et (y) est exprimée dans sa propre unité. Pour contourner ce problème, on calcule plutôt le coefficient de corrélation [21] donné par l'équation (**1-4**). Le coefficient de corrélation est ainsi sans unité et normalisé à 1. Lorsqu'il vaut 0, il y a absence de corrélation et lorsqu'il vaut 1, la corrélation est maximale.

Covariance:
$$cov(x,y) = \sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \tag{1-3}$$

Coefficient de corrélation:
$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \tag{1-4}$$

Il est important de signaler ici que dans le cas d'une analyse univariée de N échantillons, les variables (x) et (y) sont toutes les deux représentées par des vecteurs colonnes de N lignes. Lorsqu'on passe à une analyse multi-variée, il y a plusieurs variables en entrée, ce qui signifie que l'on a alors affaire non plus à un vecteur (x) de dimension (N,1) mais à une matrice X de dimension (N,k) où k est le nombre de variables d'entrée prises en compte pour l'analyse. Il est alors intéressant de constater que, même sans aucune connaissance de la variable (y), on peut s'intéresser à rechercher d'éventuelles corrélations entre les variables d'entrée grâce au calcul de la covariance ou du coefficient de corrélation. Dans le cas de deux variables d'entrée x_j et x_k on peut alors réécrire les relations (1-3) et (1-4) de la façon suivante :

Covariance:
$$cov(x_j, x_k) = \sum_{i=1}^{N}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \tag{1-5}$$

Coefficient de corrélation:
$$r_{jk} = \frac{\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{s_j s_k} \tag{1-6}$$

Grâce à ce calcul de covariance, on détermine les variables d'entrée qui sont corrélés, non-corrélés ou anti-corrélés. Cette approche qui consiste à décrire les relations entre les variables d'entrée (matrice X) est au cœur de la technique d'analyse en composantes principales qui sera discutée plus tard. Par extension, on peut étudier la corrélation entre les variables d'entrée X et la variable de sortie (y), et c'est cette approche qui est à la base de l'algorithme

de régression aux moindres carrés partiels (PLS-1) qui sera discuté plus loin. Enfin, de même que l'on peut avoir plusieurs variables d'entrée qui forment une matrice X de dimension (N,k), on peut aussi traiter le cas de plusieurs variables de sortie, par exemple les concentrations de m analytes, et l'on a alors affaire à une matrice Y de dimension (N, m). Le calcul de covariance est là encore à l'origine de l'algorithme de calcul PLS-2 qui s'applique dans ce cas-là.

Lorsqu'on passe en revue les méthodes multi-variées les plus couramment utilisées en chimiométrie, il est intéressant de les classer en trois catégories :

- Les techniques exploratoires : PCA, ICA [24]
- Les méthodes de classification supervisées : SIMCA [25], LDA [26], PLS-DA [27], KNN [28], SVM [29, 30], ANN [31].
- Les méthodes de régression ou d'analyse quantitative : PCR [32], PLS [33], LS-SVM [34], ANN [35].

Nous n'allons pas détailler ici chacune de ces méthodes mais simplement approfondir la discussion pour les méthodes que nous avons sélectionnées pour analyser les spectres LIBS et les spectres térahertz et dont les résultats seront donnés dans les chapitres suivants. Il s'agit donc de comprendre et de maîtriser parfaitement les algorithmes de calcul dans le but de les implanter dans le logiciel commercialisé par la société IVEA, partenaire de ce projet. Nous allons commencer par une discussion sur l'ACP qui est sans doute la technique exploratoire la plus répandue pour décrire les données d'entrée, rechercher d'éventuels échantillons aberrants et connaître les corrélations entre les variables d'entrée. Nous discuterons notamment de l'intérêt de l'ACP pour compresser les données. Ensuite, nous poursuivrons avec la régression PLS pour l'analyse quantitative. Puis, nous décrirons en détail l'algorithme de réseau de neurones à trois couches que nous avons utilisé pour analyser les spectres LIBS et THz, tant pour la classification que pour l'analyse quantitative. Enfin, la notion de performance étant essentielle, nous donnerons dans la dernière partie de ce chapitre la méthodologie complète que nous avons adoptée afin d'évaluer les performances des différentes approches chimiométriques. Remarquons que seules quelques techniques de chimiométrie sont discutées dans ce travail de thèse, sachant qu'une approche exhaustive n'était pas notre objectif.

1.2 **Analyse en composantes principales**

L'analyse en composantes principales (ACP ou PCA en anglais) est la méthode de classification la plus courante. Elle est non supervisée, c'est-à-dire sans phase d'apprentissage et elle vise simplement à décrire les données d'entrée, à savoir la matrice X discutée plus haut de dimension (N,k) où N est le nombre de mesures et k le nombre de variables considérées. Cette technique exploratoire permet de constater si les échantillons forment des groupes distincts, si un échantillons est anormalement éloigné des autres et peut donc être considéré comme aberrant.

L'ACP est entièrement basée sur le calcul de la matrice des covariances qui s'écrit simplement comme le produit de la matrice X des variables d'entrée par sa matrice transposée

X^T. La matrice de covariances de X est donc une matrice carrée de dimension (k,k), symétrique et les valeurs situées sur la diagonale de la matrice sont les valeurs des variances de chaque variable. Enfin, le signe des termes non diagonaux informe sur la nature de la corrélation : un signe plus indique que les deux variables évoluent dans le même sens, alors qu'un signe moins indique que lorsqu'une variable croît, l'autre décroît. Nous devons ensuite calculer les valeurs propres et les vecteurs propres de cette matrice de covariances, en prenant soin de normaliser à 1 la norme des vecteurs propres. Ce calcul complet de diagonalisation de la matrice de covariance ne peut malheureusement pas être mis en œuvre dans le cas de matrices de grandes dimensions et l'on a alors recours à un calcul itératif couramment utilisé et connu sous le nom de NIPALS, pour « nonlinear iterative partial least squares » [36] et décrit ci-après.

Matrice d'origine : $$X = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nk} \end{bmatrix} \quad (1\text{-}7)$$

Notons qu'un prétraitement des données est nécessaire. Il faut en effet que les données soient centrées (la valeur moyenne est soustraite à toutes les données) ou bien centrées et réduites, ce qui revient à normaliser les données centrées par rapport à leur variance.

Par une approche itérative, modèle en composantes principales s'écrit:

$$X = TP^T + E \quad (1\text{-}8)$$

où X est la matrice des données initiales de dimension (N,k), T est la matrice des *scores* de dimension (N,A), P la matrice des *loadings* de dimension (k,A) et E celle des résidus de dimension (N,k). Dans ces conditions, A est le nombre de composantes principales que l'on va prendre en compte pour réécrire la matrice X. Notons que P est la matrice des vecteurs propres de la matrice des covariances.

On démarre le calcul à l'itération 0 avec:

$$E_0 = X \quad (1\text{-}9)$$

Donc E_0 a pour dimension (N,k).

La matrice des scores s'écrit : $$T = \begin{bmatrix} t_{11} & \cdots & t_{1A} \\ \vdots & \ddots & \vdots \\ t_{N1} & \cdots & t_{NA} \end{bmatrix} = [t_1\ t_2 \dots t_A] \quad (1\text{-}10)$$

La matrice des loadings s'écrit : $$P = \begin{bmatrix} p_{11} & \cdots & p_{1A} \\ \vdots & \ddots & \vdots \\ p_{k1} & \cdots & t_{kA} \end{bmatrix} = [p_1\ p_2 \dots p_A] \quad (1\text{-}11)$$

Dans ce modèle itératif, le processus peut être interrompu à tout moment, ce qui signifie que le nombre A de composantes peut être sélectionné. Cela permet de bâtir un modèle avec seulement deux composantes si cela suffit, ou bien trois ou plus selon un critère de choix qui sera discuté ultérieurement. La relation (1-8) s'écrit alors ;

$$E_a = E_{a-1} - t_a p_a^T \tag{1-12}$$

Et le vecteur propre p_a^T est tel que :

$$E_{a-1} = t_a p_a^T \tag{1-13}$$

On en déduit que :

$$t_a^T E_{a-1} = t_a^T t_a p_a^T \tag{1-14}$$

Donc :

$$p_a^T = \frac{t_a^T E_{a-1}}{t_a^T t_a} \tag{1-15}$$

Soit encore :

$$p_a = \frac{E_{a-1}^T t_a}{t_a^T t_a} \tag{1-16}$$

Comme première colonne de la matrice T, que l'on notera t_1, de dimension (N,1), on choisit la colonne de la matrice X qui présente la variance la plus élevée. On peut alors calculer le premier vecteur propre p_1 de dimension (k,1) par la relation :

$$p_1^T = \frac{t_1^T E_0}{t_1^T . t_1} \tag{1-17}$$

Puis, on normalise afin d'éviter tout problème d'interprétation des résultats.

p_1 devient alors :

$$\frac{p_1}{\sqrt{p_1^T p_1}} \tag{1-18}$$

On calcule alors le résidu à l'itération n°1 :

$$E_1 = E_0 - t_1 p_1^T \tag{1-19}$$

Ce nouveau résidu permet de calculer la seconde colonne de la matrice T des scores via la relation :

$$t_2 = E_1 \frac{p_1}{p_1^T\ p_1} \tag{1-20}$$

On obtient alors $p_2^T = \frac{t_2^T E_1}{t_2^T . t_2}$ que l'on normalise avant de calculer le résidu à l'itération n°2 :

$$E_2 = E_1 - t_2 p_2^T \tag{1-21}$$

Et ainsi de suite.

On calcule les valeurs propres $\hat{\tau}_a$ définies par :

$$\hat{\tau}_a = t_a^T t_a \tag{1-22}$$

On considère que le calcul itératif a convergé lorsque $(\hat{\tau}_a - \hat{\tau}_{a-1}) \leq 0.00001$

Grâce à cet algorithme itératif, on contourne le problème de la diagonalisation d'une matrice de très grande dimension et on décrit la matrice initiale X de dimension (N,k) dans une nouvelle base de dimension A donnée par le nombre de composantes. Autrement dit, la matrice T aura pour dimension (N, A) avec A<<k.

Il s'ensuit une des propriétés les plus importantes de l'ACP qui est la compression des données. Ainsi, un jeu de plusieurs variables (par exemples les longueurs d'ondes dans le spectre auxquelles sera prélevé le signal LIBS ou THz) peut être réduit à deux dimensions (ou trois dimensions maximum) sans pour autant perdre une trop grande quantité d'information. Pour illustrer cette propriété extrêmement intéressante de l'ACP, prenons l'exemple d'une image d'avion tel que montré sur la Figure 1-1. En choisissant trois axes comme sur la figure, on constate que le jeu des données qui décrit l'avion peut facilement être projeté sur des plans, certes avec une perte d'information, mais avec tout de même beaucoup de renseignements sur le jeu de données de départ. Par exemple, la projection dans le plan (1,3) présentée sur la Figure 1-1(b) ne permet pas de comprendre que l'avion a des ailes alors que les deux autres projections le permettent. On s'accordera cependant sur le fait que 3 composantes suffisent à décrire l'avion avec une très bonne connaissance. Par analogie avec des données spectrales, les premières composantes étant celles qui expliquent le maximum de variance du jeu de données, on pourra compresser les données avec la même efficacité que ce qui a été montré dans le cas de l'avion.

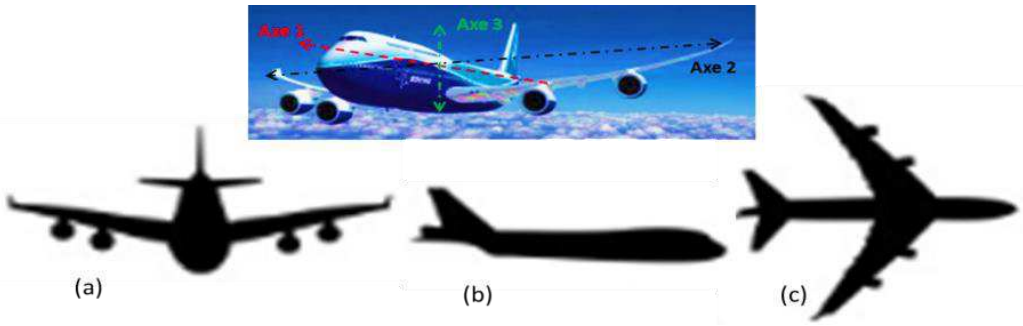


Figure 1-1 Projections des données initiales sur différents plans. (a) projection sur le plan (2,3), (b) projection sur le plan (1,3), (c) projection sur le plan (1,2).

Ainsi, un spectre d'émission ou d'absorption composé initialement de centaines ou de milliers de variables pourra être compressé dans un espace à 2 ou 3 dimensions. C'est l'utilisation que nous avons faite de l'ACP. En effet, même si on peut construire des modèles avec un nombre A de composantes plus élevées, l'expérience montre que ce sont souvent les trois premières

composantes qui offrent une très grande quantité d'information et que les composantes supplémentaires n'apportent finalement que peu d'information supplémentaire.

En conclusion, la première composante correspond à un axe qui décrit la plus grande quantité de la variance du jeu de données, puis la seconde composante, par définition orthogonale à la première explique le maximum de variance de la partie résiduelle et ainsi de suite jusqu'à ce que plus aucune information significative ne soit ajoutée [21, 37]. Afin de donner un exemple concret d'utilisation de l'ACP en spectroscopie, citons les travaux effectués dans notre groupe de recherche pour traiter des spectres LIBS d'échantillons complexes de sols pollués [38]. La Figure 1-2 donne le diagramme des scores dans le plan des deux premières composantes principales sur lequel on voit une séparation entre deux classes de sols. D'après les loadings montrés sur la Figure 1-3, on comprend que l'axe 1 horizontal est principalement lié au calcium tandis que l'axe 2 vertical est lié à l'aluminium et au silicium. On vérifie aussi que les raies d'aluminium et de silicium sont corrélées entre elles, ce qui est normal puisqu'on fait référence à des alumino-silicates dans le cas de ces sols. On apprend aussi que l'aluminium et le silicium ne sont absolument pas corrélés au calcium.

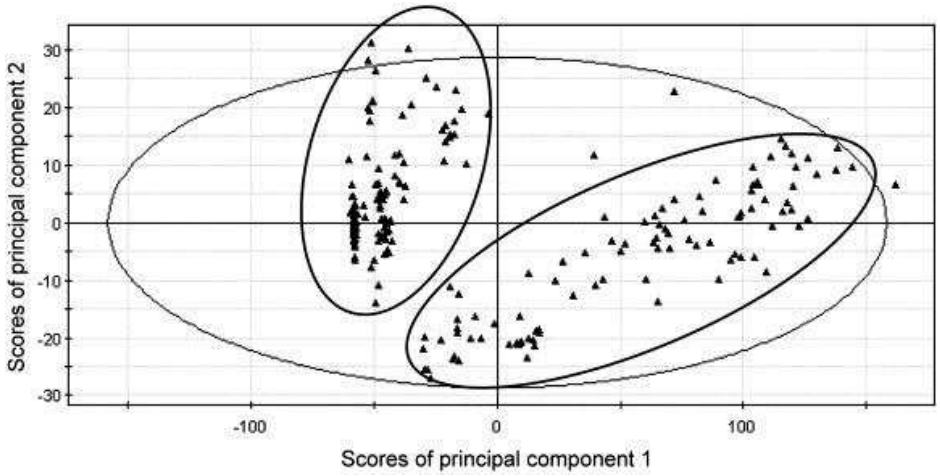


Figure 1-2 Résultat de calcul ACP appliqué des données LIBS relatives à des échantillons de sol. Scores dans le plan des deux premières composantes principales. Extrait de [38]

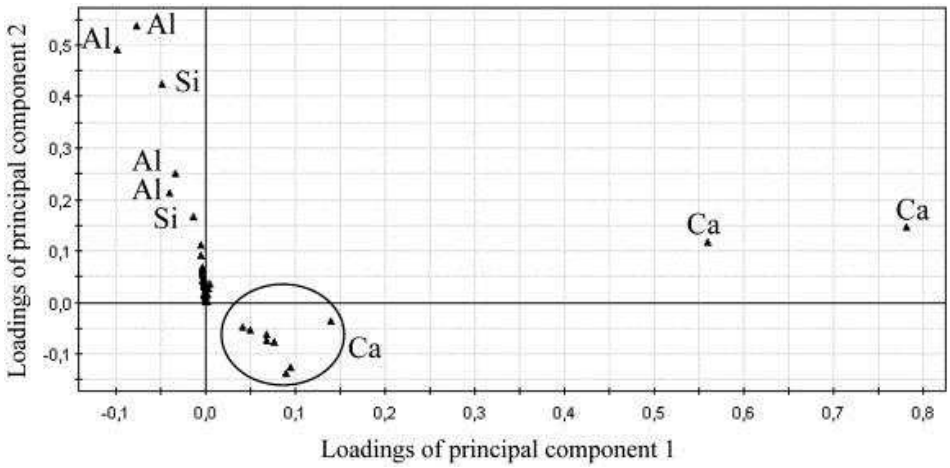


Figure 1-3 Résultat de calcul ACP appliqué des données LIBS relatives à des échantillons de sol. Loadings dans le plan des deux premières composantes principales (1,2). Extrait de [38]

Plus généralement, la méthode ACP a été utilisée pour la discrimination des spores bactériennes des moisissures, des pollens, et des protéines par LIBS [39]. Dans ce cas, la première composante représentait 91.53% de la variance totale mais les loadings de la première composante ne représentaient aucune raie atomique spécifique et il en a été déduit que la première composante représentait en fait la variance entre chaque tir de mesure. La deuxième composante représentait quant à elle plusieurs raies significatives (Mg II, Ca II, Ca I, H I, N I et O I). Et par la présentation des deux composantes (1 et 2) trois groupes de points ont été distingués, un pour les pollens et les ovalbumines, le deuxième pour les moisissures, et le troisième pour les bactéries [40]. Une discrimination des sols sans aucun chevauchement entre deux sites géologiques distincts a aussi été obtenue par application du calcul ACP à des données LIBS en limitant l'analyse aux deux premières composantes pour lesquelles la variance cumulée était de 92.9% [41]. Notons encore que la méthode ACP a également été utilisée en spectroscopie THz pour le diagnostic des tissus cancérigènes [42] et pour la détection des matériaux explosifs [43].

Plus généralement, l'ACP peut aussi être utilisée de manière exploratoire pour identifier des différences de propriétés physiques entre les échantillons comme la température et le taux d'humidité. Elle est considérée efficace pour décrire rapidement un jeu de données de façon globale. Dans les cas où il y a beaucoup de variables cette méthode permet de connaître la corrélation entre les variables et de savoir quelles variables apportent la plus grande variabilité à la matrice X. De plus, elle permet également de détecter d'éventuels points aberrants qui pourront être ensuite exclus de l'analyse.

Sur la base de tous ces avantages, il a été montré qu'une classification par ACP de matériaux de référence certifiés (MRC) comme étape préliminaire avant l'analyse quantitative était une façon de s'affranchir des effets de matrices en LIBS [44]. Dans ce travail, les scores dans le plan des deux premières composantes principales présentés sur la Figure 1-4 ont permis de distinguer quatre classes d'échantillons.

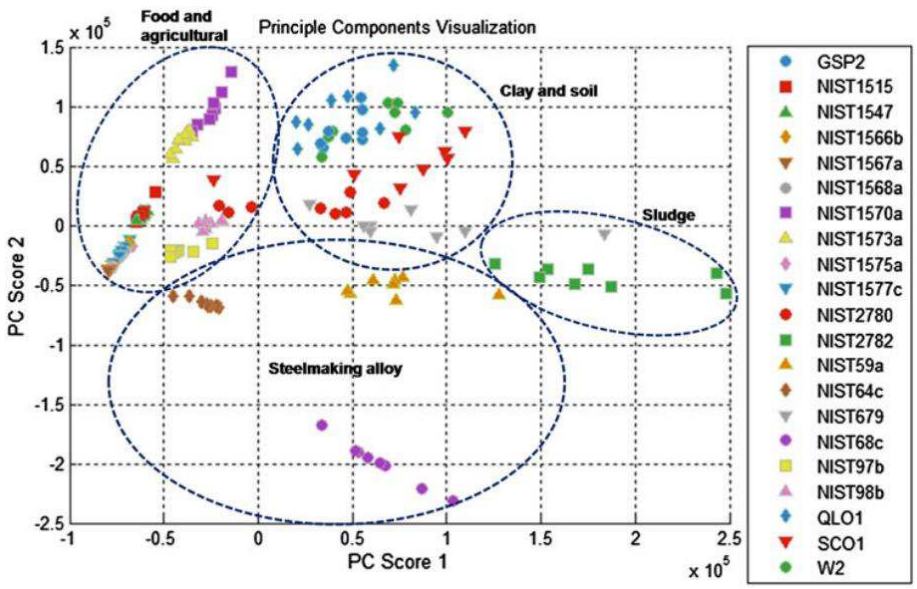


Figure 1-4 Scores dans le plan des deux premières composantes principales obtenus par calcul ACP de données LIBS issues de 21 échantillons décrits par le code couleur indiqué sur la droite. D'après la référence [44].

En conclusion, la technique d'ACP peut être appliquée à chaque type de spectroscopie, que ce soit en LIBS, THz, Raman, NIR, MIR ou autres. Dans tous les cas, elle offre une approche exploratoire très efficace qui permet de comprendre des similitudes et des corrélations et de rejeter des données aberrantes. Pour l'ensemble des études ACP présentées dans ce mémoire, nous avons utilisé le logiciel commercial SIMCA-P+ Version 12.0.1.0. Nous montrerons plus particulièrement dans le chapitre 3 que l'ACP est très efficace pour conduire des analyses semi-quantitatives de spectres THz de mélanges ternaires.

1.3 **Analyse par régression aux moindres carrés partiels**

La régression par moindres carrés partiels, qui sera désormais notée PLS, est une méthode d'analyse quantitative. Elle est supervisée car elle s'appuie sur une phase d'étalonnage. Elle fonctionne sur le même principe que l'ACP dans la mesure où là encore, il est question de calculer les covariances entre les variables. La différence importante est que, alors qu'en ACP on ne s'intéressait qu'à la matrice des covariances des données d'entrée X, ici on s'intéresse aux variables x qui présentent la plus grande variance corrélée à la concentration y. Toutes les propriétés de l'ACP sont sauvegardées : compression des données dans un espace de faible dimensionnalité (le plus souvent à 2 ou 3 dimensions en ce qui concerne nos études de spectroscopie) et composantes orthogonales. Notons que la sortie du modèle PLS est généralement une seule variable, à savoir la concentration de l'analyte en ce qui concerne nos travaux de spectroscopie. On parle dans ce cas de modèle PLS-1 par opposition au modèle PLS-2 qui permet de traiter en sortie non pas un vecteur mais une matrice et qui permet notamment de prédire les concentrations de plusieurs éléments simultanément.

Les chimistes se sont intéressés aux méthodes multivariées dès lors que l'approche univariée ne permettait pas de traiter des spectres complexes. La régression multi-linéaire (MLR) s'est

avérée efficace uniquement dans le cas de variables non corrélées entre elles [45, 46] ce qui est finalement rarement le cas dans les spectres d'émission ou d'absorption. C'est pour cela qu'il a fallu se tourner vers une méthode alternative, la régression PLS [47]. Par ailleurs, la PLS permet de prendre en compte des effets de normalisation par étalon interne sans avoir besoin d'identifier au préalable quelle raie doit être utilisée pour cet étalonnage [48]. Elle permet aussi de trouver rapidement quelles sont les données d'entrée (matrice X) corrélées à la concentration de l'analyte (vecteur y), et qui seront naturellement sélectionnées pour l'étalonnage. Notons que pour la LIBS, la PLS a permis d'analyser des spectres aux premiers instants de l'émission du plasma en contournant le problème d'écrantage du signal caractérisé par les raies atomiques par le spectre continu provenant du bremsstrahlung [49].

Comme pour l'ACP, un prétraitement des données spectrales est nécessaire afin de disposer de données centrées sur la valeur moyenne ou même centrées réduites, ce qui signifie qu'elles sont en plus normalisées par l'écart-type. Le choix entre ces deux prétraitements n'est pas évident et dépend de la nature des données. Il n'y a pas de prétraitement qui soit meilleur que l'autre a priori et finalement un test entre les deux prétraitements est la seule façon pour choisir le meilleur au cas par cas. Dans le cas de l'analyse de données spectrales, considérons comme précédemment une matrice X de dimension (N,k) avec N le nombre de spectres et k le nombre de variables. En sortie du modèle, nous aurons le vecteur y de longueur N correspondant à la concentration de l'analyte pour chaque échantillon.

On propose de décrire ici l'algorithme original de PLS [37, 50]. La matrice X et le vecteur y doivent être décrits par les combinaisons linéaires suivantes:

$$X = t_1 p_1^T + t_2 p_2^T + \cdots + t_A p_A^T + E_A \tag{1-23}$$

$$y = t_1 q_1 + t_2 q_2 + \cdots + f_A \tag{1-24}$$

On remarque que les deux quantités X et y sont exprimées en fonction des mêmes scores t_i tandis que chacune a ses propres vecteurs propres à savoir p_i pour X et q_i pour y. De même, chaque quantité a son propre résidu à savoir E_i pour X et f_i pour y. Grâce à ce système de deux équations couplées, on maximise la corrélation entre les N spectres (X) et les N valeurs de concentration de l'analyte (y).

On prend comme conditions initiales : E_0= X et f_0= y. Puis on relie les deux résidus par un paramètre de poids W tel que :

$$W_1 = \mathrm{E}_0^T f_0 = X^T y \tag{1-25}$$

On peut ainsi calculer le premier score qui s'écrit :

$$t_1 = E_0 W_1 = X W_1 \tag{1-26}$$

Ce qui permet de calculer les vecteurs propres p_1 et q_1 :

$$p_1 = \frac{E_0^T t_1}{t_1^T t_1} \tag{1-27}$$

$$q_1 = \frac{f_0^T\, t_1}{t_1^T t_1} \tag{1-28}$$

On peut finalement calculer les résidus à l'itération suivante :

$$E_1 = E_0 - t_1 P_1^T \tag{1-29}$$

$$f_1 = f_0 - t_1 q_1^T \tag{1-30}$$

A ce stade, on peut calculer le second poids et ensuite le second score :

$$W_2 = \mathrm{E}_1^T\, f_1 \tag{1-31}$$

$$t_2 = E_1 W_2 \tag{1-32}$$

Et ainsi de suite. Plus généralement, on peut donc écrire :

$$E_a = E_{a-1} - t_a p_a^T \tag{1-33}$$

$$f_a = f_{a-1} - t_a q_a \tag{1-34}$$

$$w_a = E_{a-1}^T f_{a-1} \tag{1-35}$$

$$t_a = E_{a-1} w_a \tag{1-36}$$

$$p_a = \frac{E_{a-1}^T t_a}{t_a^T t_a} \tag{1-37}$$

$$q_a = \frac{f_{a-1}^T t_a}{t_a^T t_a} \tag{1-38}$$

Une fois que le modèle PLS est construit, pour calculer les scores d'un échantillon inconnu, il faut avant tout lui appliquer la même normalisation que celle qui a servi à l'étalonnage. Ainsi, si l'étalonnage a été effectué avec des variables centrées, on devra soustraire au spectre inconnu (vecteur x_0) le vecteur $\bar{X}$ des moyennes de la matrice X : $e_0 = x_0 - \bar{X}$

Pour un échantillon inconnu, les scores sont également calculés d'une façon itérative si bien que pour chaque composante a, on peut écrire :

$$t_{a0} = e_{a-1}\, w_a \tag{1-39}$$

Et on calcule le nouveau résidu pour le prochain calcul de score :

$$e_a = e_{a-1} - t_{a0} p_a \tag{1-40}$$

Finalement, lorsqu'on fixe le nombre A de composantes, la valeur prédite pour y s'écrit :

$$\hat{y}_{A0} = \sum_{a=1}^{A} t_{a0} q_a \tag{1-41}$$

Notons que si les données sont centrées, il faudra finalement ajouter à toutes les valeurs prédites de y la valeur moyenne des concentrations ($\bar{y}$) du modèle PLS [50], à savoir la moyenne des concentrations des échantillons du lot de calibration. Notons ici que ce que l'on appelle le lot de calibration est le sous-ensemble des échantillons de départ qui sont utilisés pour construire le modèle de régression. De même, le lot de validation décrit le sous-ensemble des échantillons qui sont utilisés pour valider le modèle. Il permet de choisir parmi plusieurs modèles lequel offre les meilleurs résultats de prédiction. Enfin, le lot de test décrit quant à lui le sous-ensemble des échantillons qui ne sont utilisés qu'a posteriori, une fois que le meilleur modèle a été déterminé. Pour choisir le nombre A de composantes, on applique généralement la méthode de validation croisée interne ou externe. Le logiciel SIMCA-P+ que nous avons utilisé pour nos études s'appuie sur la validation croisée interne. Celle-ci consiste rejeter un échantillon du lot de calibration afin de prédire sa concentration par PLS et à recommencer le calcul pour tous les échantillons. On choisira le nombre A de manière à minimiser l'erreur de prédiction moyenne.

La PLS a été exploitée en LIBS notamment pour analyser la concentration de carbone et distinguer la nature organique/inorganique [51]. En effet l'analyse de la raie du carbone à 247.8 nm seule n'est pas suffisante pour conclure. A l'inverse, l'analyse par PLS de tout le spectre compris entre 200 et 800 nm a permis de réaliser la distinction organique/inorganique avec succès et la raie du magnésium à 280.3 nm a révélé une influence particulièrement importante. Les auteurs ont montré qu'au-delà de la 4ème composante principale, la contribution des raies spectrales n'est pas significative comparée au bruit et qu'il est donc préférable de ne pas tenir compte des composantes suivantes. Le choix du nombre optimal de composantes principales est déterminé par validation croisée [52]. Notons enfin que la méthode PLS a également été exploitée dans le cadre de l'analyse de données LIBS dans le contexte de l'exploration spatiale, avec en particulier l'analyse du sol lunaire [53] et celles menées actuellement sur le sol martien via les mesures réalisées par le rover Curiosity [54, 55].

1.4 **Réseaux de neurones artificiels - ANN**

Les réseaux de neurones artificiels ou ANN pour « artificial neural networks » constituent l'une des méthodes bien connues d'intelligence artificielle pour résoudre des problèmes complexes. Les applications les plus courantes sont la reconnaissance de formes et plus généralement l'analyse des données pour trier, prendre une décision ou contrôler un procédé. L'ANN est évidemment inspiré du cerveau biologique. On peut lui apprendre à fournir certaines réponses en lui proposant une série d'exemples. L'ANN fera alors intervenir des paramètres tels que la vitesse d'apprentissage et la mémoire tout comme le cerveau biologique. Une fois l'apprentissage effectué, il sera à même de résoudre de nouveaux problèmes.

Les avantages d'un réseau de neurones sont :

- La possibilité de fournir une réponse soit linéaire soit non linéaire selon le contexte. Ceci est très utile car dans le cadre de cas réels, les données évoluent souvent de façon non linéaire et les modes de calcul linéaires peuvent s'avérer insuffisants. La Figure 1-5 donne un exemple de séparation d'un jeu de données en deux classes A et B. En (a) une séparation des données est possible à l'aide d'un modèle linéaire tandis qu'en (b) elle ne l'est pas et un modèle non-linéaire est indispensable.
- La possibilité de construire un modèle uniquement basé sur la connaissance des données d'entrée et de sortie, sans nécessité de connaître le type de relation mathématique qui les relie.
- La possibilité d'utiliser l'ANN dans un environnement différent de celui qui a servi à la phase d'apprentissage grâce à sa capacité d'adaptation et sa faible sensibilité aux données aberrantes. En effet, chaque neurone est sensible à l'activité globale générée par tous les autres neurones collectivement.
- La possibilité de pouvoir interpréter un résultat tel qu'un tri avec un taux de confiance.
- La possibilité de continuer à fonctionner malgré une panne ou un dysfonctionnement de l'un des neurones. Dans ce cas, la réponse de l'ANN est légèrement dégradée mais reste viable car l'information a été transmise de façon distribuée sur tous les neurones. L'ANN est donc particulièrement robuste.

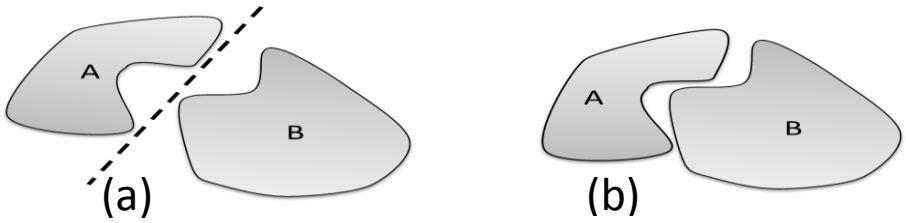


Figure 1-5 Les deux classes A et B sont séparables soit linéairement (a), soit non-linéairement (b)

1.4.1 Modèle du perceptron simple

La brique élémentaire d'un réseau de neurones artificiel est le perceptron simple. Il s'agit d'un système qui reçoit des données en entrée et qui fournit une réponse en sortie. Les données d'entrée ne sont pas toutes connectées au perceptron avec le même poids mais au contraire chacune d'entre elle est affectée d'un poids différent, indiquant son importance (Figure 1-6).

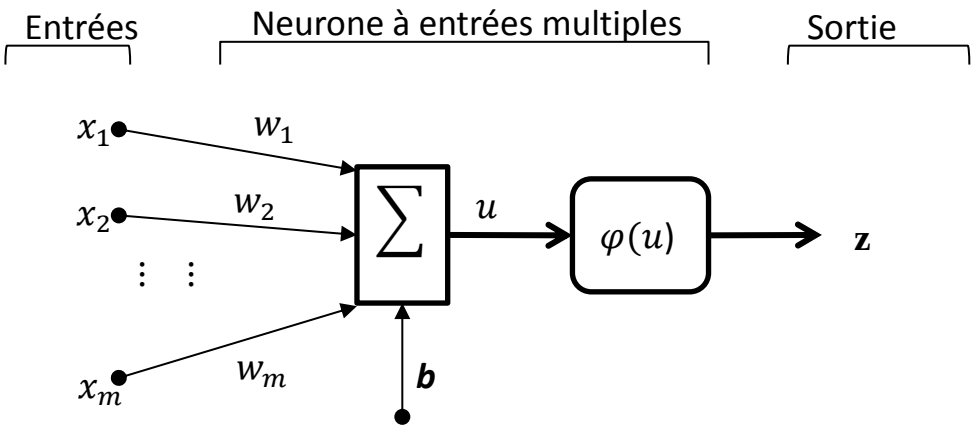


Figure 1-6 Schéma de principe d'un perceptron simple

La réponse fournie par le perceptron est basée sur un calcul simple mettant en jeu une fonction d'activation, la plus fréquemment utilisée étant la fonction sigmoïde définie par :

$$\varphi(u) = \frac{1}{1 + e^{-u}} \tag{1-42}$$

Afin de prendre en compte les données d'entrée la variable u est définie par :

$$u = \sum_{i=1}^{m} w_i x_i - b \tag{1-43}$$

Notons que la somme pondérée des données d'entrée est comparée à un seuil (ou biais) noté b. Il faut comprendre ici que si la somme pondérée est supérieure au seuil d'activation, le perceptron fournit une réponse proche de 1 et dans le cas contraire une réponse proche de 0. Le biais peut être ajusté au même titre que les poids dans le but d'optimiser la réponse du perceptron. Cela permet de traiter le biais comme n'importe quel autre poids lorsqu'on écrit un algorithme de calcul, en introduisant simplement une (m+1)ième donnée d'entrée égale à 1. La valeur de sortie du perceptron est toujours comprise entre 0 et 1 d'après la fonction d'activation sigmoïde. Notons que dans le cas d'un neurone biologique, la valeur de sortie est égale à 0 ou à 1 car la fonction d'activation est binaire, en tout ou rien. L'avantage de la fonction sigmoïde est la possibilité de calculer une fonction dérivée utile pour faire converger le calcul itératif qui consiste à rechercher les meilleurs poids.

1.4.2 Algorithme de calcul ANN

Un réseau de neurones artificiels (ANN) est une association de plusieurs perceptrons interconnectés. Plusieurs architectures existent selon le nombre de couches de neurones dans la chaîne de transmission, le nombre de neurones dans chaque couche et aussi selon les règles de propagation du signal dans le réseau. Dans l'étude présentée ici, nous ne traiterons que le cas d'un réseau à trois couches : une couche d'entrée, une couche cachée et une couche de sortie. Le signal se propage uniquement de l'entrée vers la sortie mais l'erreur entre le signal de sortie et la valeur de référence se propage en sens opposé, de la sortie vers l'entrée, afin

d'ajuster les poids de manière itérative. La couche d'entrée reçoit tout simplement les données d'entrée provenant des expériences de spectroscopie. La couche de sortie fournit une réponse qui sera ici une valeur prédite de la concentration de l'analyte. La couche cachée permet d'interconnecter l'entrée et la sortie du réseau. Elle permet en quelque sorte d'extraire et de trier les informations de la couche d'entrée dans un nouvel espace. Il a été démontré qu'une seule couche cachée était suffisante dans le cas du traitement de données spectrales [56] et c'est pourquoi notre étude se limitera à une architecture de réseau présentant une seule couche cachée.

Un réseau à 3 couches (couche d'entrée, couche cachée, couche de sortie) est présenté sur la Figure 1-7. Les données d'entrée x_i sont tout simplement les données provenant des spectres. Cela peut être l'intensité de quelques raies spectrales bien choisies ou encore des données issues de la compression - par ACP ou par PLS - des spectres d'origine comme nous le verrons plus loin. La couche cachée contient plusieurs perceptrons qui fonctionnent comme des relais du signal et la couche de sortie peut contenir soit un seul neurone s'il s'agit de focaliser l'analyse sur un seul composé ou au contraire plusieurs neurones si l'on souhaite mener une analyse multiéléments en parallèle.

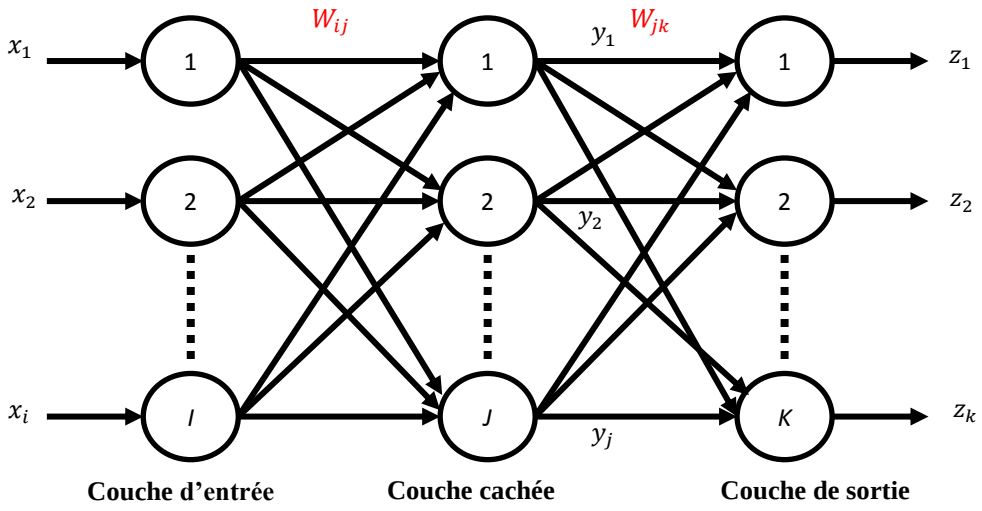


Figure 1-7 Architecture d'un réseau de neurones artificiels à 3 couches.

Notons que la couche cachée apporte un certain bénéfice au système mais que d'un autre côté, cela peut rendre l'interprétation plus compliquée. En effet, l'existence de signaux intermédiaires créés dans la couche cachée peut masquer la relation de cause à effet qui existe entre l'entrée et la sortie de l'ANN. L'algorithme de calcul d'ANN fonctionne sur un mode itératif. Initialement, les poids et biais sont choisis de façon aléatoire si bien que la valeur en sortie est très éloignée de la valeur cible fournie par une mesure de référence. On calcule alors l'erreur entre la valeur prédite et la valeur de référence et on fait propager cette erreur en contre-sens (Figure 1-8), soit de la sortie vers l'entrée de l'ANN afin de calculer de nouvelles valeurs de poids et de biais destinées à minimiser l'erreur.

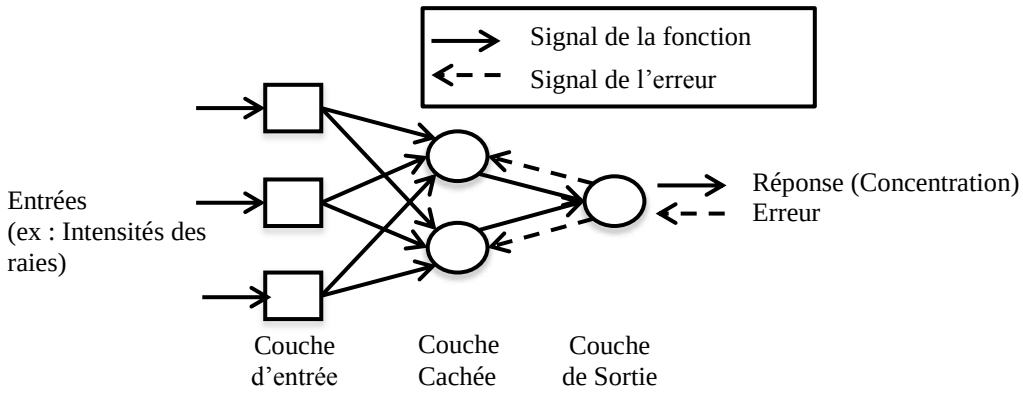


Figure 1-8 Propagation du signal de l'entrée vers la sortie et rétro-propagation de l'erreur

Les meilleurs poids ayant été trouvés pour l'échantillon 1, on introduit les données de l'échantillon 2 en entrée de l'ANN et on recommence. On passe ainsi en revue tous les échantillons du lot de calibration et cela constitue la première itération du calcul. On applique ensuite un grand nombre d'itérations afin que le modèle ait vu tous les échantillons un grand nombre de fois. Les valeurs des poids et biais que l'on obtient correspondent ainsi au meilleur compromis pour relier les données d'entrée aux données de sortie.

Voici à présent le détail du calcul étape par étape.

1.4.2.1 Prétraitement et initialisation

Avant tout, il est très important de normaliser les données d'entrée afin de faciliter la convergence de l'apprentissage. Une normalisation des données d'entrées par rapport à la valeur maximale (pour le lot de calibration) permet de disposer de valeurs en entrée normalisées à 1. Par ailleurs, les données d'entrée doivent être toutes positives pour que les poids de la couche cachée soient significatifs et que l'algorithme converge. Lorsque les données d'entrée sont les intensités de quelques raies spectrales, il n'y a aucune difficulté. En revanche, lorsqu'elles proviennent d'un calcul préalable par ACP ou PLS visant à compresser les données spectrales, elles correspondent à des scores dont les valeurs peuvent être aussi bien positives que négatives. Dans ce cas, on envisagera d'appliquer un offset à l'ensemble des données afin de n'introduire que des valeurs positives dans l'ANN. D'autre part, nous savons que pour obtenir un apprentissage efficace, les données d'entrée ne doivent pas être corrélées [57]. Cette condition est satisfaite dès lors que l'on sélectionne des raies spectrales relatives à des éléments chimiques différents, bien que certaines corrélations de type Al-Si ou Ca-Mg peuvent subsister notamment dans le cadre de l'analyse des sols. Et, lorsque les données d'entrée proviennent des scores d'un calcul par ACP, elles sont naturellement non-corrélées par définition.

Comme pour tout calcul itératif, la convergence de l'ANN dépend du choix des valeurs initiales des poids et des biais. En effet, un choix inapproprié peut entrainer une saturation de la réponse de l'ANN ou une convergence trop lente [58]. Les valeurs initiales sont choisies de façon aléatoire mais cependant dans un intervalle bien précis. Il y a plusieurs méthodes pour ce choix des conditions initiales [59]. En ce qui concerne le logiciel scientifique Igor Pro que

nous avons utilisé au laboratoire, nous ne savons pas selon quels critères sont fixées les valeurs initiales mais en revanche, en ce qui concerne l'algorithme d'ANN que nous avons développé nous même pour la société IVEA, nous avons choisi des valeurs initiales dans l'intervalle [-0,77 ;+0,77] car ces conditions initiales permettent d'atteindre de bonnes performance dans le cas d'un ANN à une seule couche cachée comme c'est le cas pour notre étude [60].

1.4.2.2 Transmission du signal - algorithme *Feed-Forward*

Le signal est transmis de la couche d'entrée vers la couche de sortie. Rappelons que pour simplifier l'algorithme, les biais sont considérés comme des poids reliés à des entrées égales à 1. On peut dans un premier temps écrire le signal sortant de chaque neurone j de la couche cachée.

$$y_j = \frac{1}{1 + e^{-u_j}} \tag{1-44}$$

Avec

$$u_j = \sum_i w_{ji}.x_i \tag{1-45}$$

Le calcul démarre avec les valeurs (x_i) provenant de l'échantillon 1 (du lot de calibration) et des valeurs de poids w_{ji} aléatoires.

On peut alors calculer le signal sortant du neurone k de la couche de sortie de l'ANN.

$$z_k = \frac{1}{1 + e^{-v_k}} \tag{1-46}$$

Avec

$$v_k = \sum_j w_{kj}.y_j \tag{1-47}$$

Les valeurs v_k obtenues par le calcul sont à comparer aux valeurs de référence. On aura pris soin de normaliser les valeurs de référence à 1 afin de pouvoir les comparer aux valeurs calculées par l'ANN qui sont comprises entre 0 et 1 par définition de la fonction d'activation sigmoïde.

1.4.2.3 Retro-propagation de l'erreur

Il existe deux approches pour prendre en compte l'erreur entre la valeur calculée et la valeur de référence. La première, dite *online* consiste à effectuer une rétro-propagation de l'erreur après chaque nouveau calcul, pour chaque échantillon. La seconde approche connue sous le terme de *batch learning*, consiste à faire un premier calcul pour tous les échantillons du lot de calibration puis à calculer une erreur moyenne qui sera exploitée en rétro-propagation. Dans cette étude, nous avons choisi d'exploiter la première approche qui est aussi la plus répandue,

non seulement à cause de sa simplicité de mise en œuvre comparée à la seconde mais aussi parce qu'elle permet de réaliser a priori des corrections beaucoup plus fines.

Pour simplifier la discussion, considérons un ANN avec un seul neurone dans la couche de sortie. On calcule pour commencer l'erreur absolue entre la valeur de sortie de l'ANN et la valeur de référence pour le premier échantillon. La retro-propagation de l'erreur s'effectue en plusieurs étapes :

Etape 1 : correction des poids du neurone de sortie

- Calcul de la dérivée de la fonction d'activation du perceptron de la couche de sortie

En notant la fonction d'activation $F(v) = \frac{1}{1+e^{-v}}$ on trouve :

$$\boldsymbol{F'(v) = (e^{-v})/(1 + e^{-v})^2} \quad (1\text{-}48)$$

Avec $v = \sum_j w_j . y_j$

- Calcul du terme de correction δ défini par :

$$\delta = (t - z)F'(v) \quad (1\text{-}49)$$

Où t désigne la valeur de référence et z la valeur calculée par l'ANN.

- Calcul des nouveaux poids du neurone de la couche de sortie

$$w_j(n+1) = w_j(n) + \eta . \delta . y_j + \alpha\left[w_j(n) - w_j(n-1)\right] \quad (1\text{-}50)$$

Dans cette expression, η est la vitesse d'apprentissage, α le terme de mémoire et y_j les données d'entrée de la couche de sortie, c.-à-d. les sorties de chaque neurone de la couche cachée. L'indice n indique le nombre d'itérations. Notons que pour la première itération (n=0), le terme $w_j(n-1)$ n'existe pas et l'on a alors recours à une autre définition valable uniquement pour n=0:

$$w_j(1) = w_j(0) + \eta . \delta . y_j \quad (1\text{-}51)$$

Notons que les paramètres η de vitesse d'apprentissage et α de mémoire sont des paramètres ajustables qu'il conviendra de choisir pour optimiser les performances de l'ANN.

Etape 2 : correction des poids des neurones de la couche cachée

- Calcul des dérivées des fonctions d'activation de chaque perceptron dans la couche cachée

Par analogie avec la relation (1-48) on peut écrire :

$$F'(u) = (e^{-u})/(1 + e^{-u})^2 \quad (1\text{-}52)$$

Avec $u = \sum_j w_j . x_j$

- Calcul des termes de correction δ_j pour chaque perceptron j de la couche cachée, définis par :

$$\delta_j = F'(u).\delta.w_j \tag{1-53}$$

Notons que dans le cas où l'ANN aurait plusieurs neurones dans sa couche de sortie, il faudrait calculer autant de termes de correction δ_k que de neurones dans la couche de sortie. Dans ce cas, les poids w_j deviendraient w_{jk} et les termes de corrections δ_j s'écriraient:

$$\delta_j = F'(u).\sum_k \delta_k w_{jk} \tag{1-54}$$

- Calcul des nouveaux poids pour chaque neurone de la couche cachée

$$w_{ij}(n+1) = w_{ij}(n) + \eta.\delta_j.x_i + \alpha\left[w_{ij}(n) - w_{ij}(n-1)\right] \tag{1-55}$$

Dans cette expression, η est toujours la vitesse d'apprentissage et α le terme de mémoire, les mêmes que précédemment. Par ailleurs, x_i représente les données d'entrée de chaque neurone de la couche cachée. L'indice n indique le nombre d'itérations. Notons que là encore, pour la première itération (n=0), le terme $w_{ij}(n-1)$ n'existe pas et l'on a alors recours à une autre définition valable uniquement pour n=0:

$$w_{ij}(1) = w_{ij}(0) + \eta.\delta_j.x_i \tag{1-56}$$

Le choix des valeurs de la vitesse d'apprentissage η et du terme de mémoire α est très important pour la convergence de l'algorithme. En effet, si la vitesse est trop élevée, il est possible d'observer des instabilités avec un modèle qui oscille sans arrêt sans se stabiliser. Le terme de mémoire permet dans tous les cas de stabiliser le modèle et de maintenir une même direction de convergence. En pratique, ces deux termes sont positifs et généralement compris entre 0,01 et 0,8 [58, 61]. Cependant le choix des valeurs de ces deux paramètres dépend des données à traiter et par conséquent, l'analyste doit tester différentes valeurs à tâtons avant de choisir celles qui conviennent le mieux. Ajoutons enfin que pour qu'un modèle ANN soit répétable et donc robuste, il faut choisir des valeurs faibles pour la vitesse d'apprentissage et pour le terme de mémoire.

La Figure 1-9 donne un exemple de valeurs de l'erreur quadratique moyenne que l'on notera RMSE, dont la définition mathématique est donnée dans l'équation (1-59) de ce mémoire, en fonction du nombre d'itérations pour différentes valeurs de la vitesse d'apprentissage et du terme de mémoire. La comparaison des courbes verte et bleue permet de saisir l'influence du terme de mémoire. Dans ce cas précis, plus le terme de mémoire a une valeur élevée, plus le modèle converge vite et conserve ensuite des valeurs de RMSE plus faibles. Notons cependant que les courbes présentées ici correspondent aux données du lot de calibration uniquement et peuvent donc ne pas être représentatives des performances de prédiction du modèle pour des échantillons inconnus. La comparaison des courbes bleue et rouge permet quant à elle de comprendre le rôle de la vitesse d'apprentissage. On constate ici que pour une vitesse de 0,3 (rouge) la convergence est beaucoup plus rapide que pour une vitesse de 0,02 (bleu). De plus, même après 50000 itérations, la valeur de RMSE est beaucoup plus faible (4 à

5 fois) pour la vitesse d'apprentissage 0,3 que pour 0,02. Là encore, une mise en garde s'impose puisqu'il ne s'agit ici que des données provenant d'un lot de calibration. Il est ensuite très important d'observer l'évolution de RMSE pour des données de validation afin de déterminer les meilleures valeurs pour ces deux paramètres ajustables que sont la mémoire et la vitesse d'apprentissage. Cette seconde évolution de RMSE est présentée en Figure 1-10.

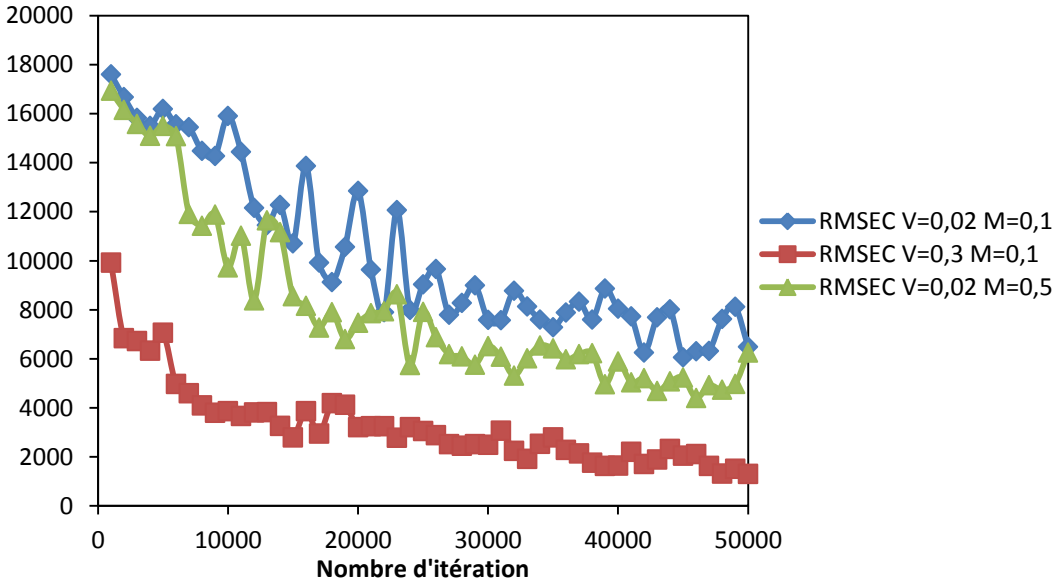


Figure 1-9 Exemple de convergence d'un modèle ANN caractérisée par la variation de RMSE en fonction du nombre d'itérations. Bleu : pour une vitesse d'apprentissage V=0,02 et un terme de mémoire M=0,1. Rouge pour V=0,3 et M=0,1. Vert pour V=0,02 et M=0,5.

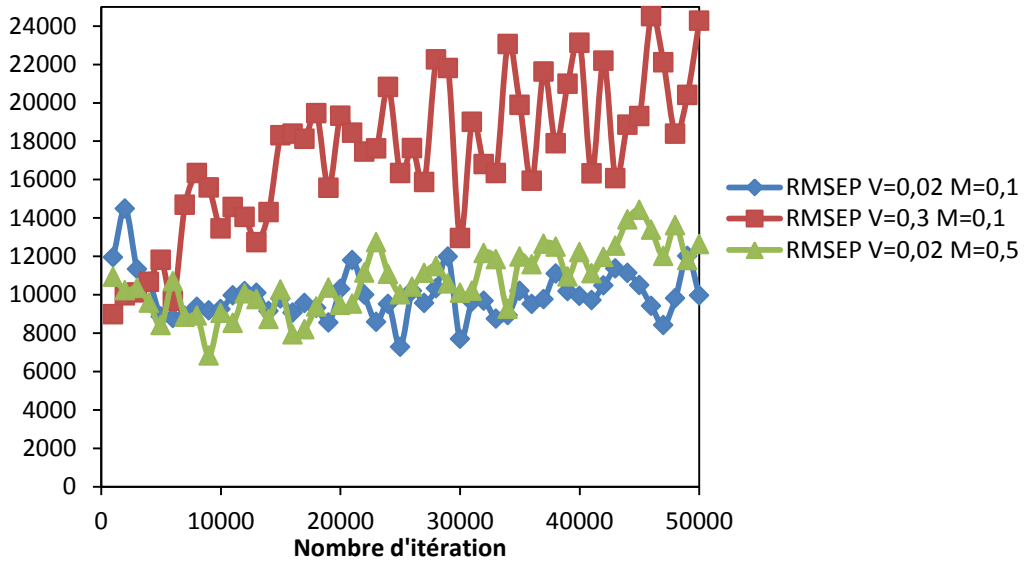


Figure 1-10 Variation de RMSE en fonction du nombre d'itérations lorsque le modèle présenté en Figure 1-9 est appliqué aux échantillons du lot de validation. Code couleur identique à celui de la figure 1-9 ; Bleu : pour une vitesse d'apprentissage V=0,02 et un terme de mémoire M=0,1. Rouge pour V=0,3 et M=0,1. Vert pour V=0,02 et M=0,5.

En comparant les courbes verte et bleue, on constate que le choix d'un terme de mémoire plus élevé n'offre finalement pas d'avantage ici pour le lot de validation. Mais en comparant les

courbes bleue et rouge, on découvre clairement que la vitesse d'apprentissage plus élevée que l'on considérait jusqu'ici comme un avantage peut être en fait très néfaste aux prédictions des valeurs de sortie de l'ANN pour le lot de validation. Les courbes rouges des figures 1-9 et 1-10 illustrent à merveille le phénomène bien connu de sur-apprentissage. En effet, RMSE diminue fortement pour le lot de calibration alors qu'elle augmente pour le lot de validation. Cela revient à dire que seuls les échantillons du lot de calibration peuvent être correctement prédits mais que la généralisation à d'autres échantillons n'est pas efficace. On préfèrera donc utiliser une vitesse d'apprentissage plus faible.

1.4.2.4 Critère d'arrêt de l'apprentissage

Rappelons que dans la phase d'apprentissage, on introduit dans l'ANN d'abord les données de l'échantillon 1. Le calcul de l'erreur puis des termes de correction permet de calculer de nouveaux poids. Ces nouveaux poids sont utilisés avec les données de l'échantillon 2 et le calcul se poursuit de manière itérative. Lorsque tous les échantillons ont été introduits, cela termine l'itération n°1 du calcul et on recommence au début en repassant tous les échantillons une $2^{ème}$ fois et ainsi de suite. Le critère d'arrêt de l'apprentissage de l'ANN s'appuie sur la méthode de validation croisée externe. Celle-ci nécessite la séparation préalable des données à traiter en deux lots : un lot d'apprentissage et un lot de validation. Le lot d'apprentissage sert à calculer le modèle, c'est-à-dire à rechercher les meilleurs poids qui permettent de minimiser l'erreur. A chaque itération du calcul, on utilise le modèle ANN avec ses paramètres du moment afin de traiter les données du lot de validation. On peut donc calculer pour chaque itération l'erreur moyenne non seulement pour les données du lot d'apprentissage mais aussi pour celles du lot de validation et suivre l'évolution de ces deux erreurs en fonction du nombre d'itérations. Le meilleurs choix consiste à arrêter le calcul lorsque l'ANN commence à passer en régime de sur-apprentissage. Concrètement, cela correspond à une augmentation de l'erreur pour le lot de validation tandis que celle du lot d'apprentissage continue à baisser. La Figure 1-11 décrit l'évolution typique des erreurs en fonction du nombre d'itérations. Le critère d'arrêt consiste donc à trouver le minimum de la courbe qui décrit l'erreur du lot de validation (en rouge). On choisit alors le nombre d'itérations trouvé selon ce critère pour lancer tous les futurs calculs basés sur l'utilisation de ce modèle ANN. Nous retiendrons que la méthode de validation croisée est rapide et fiable. Elle permet non seulement de choisir le nombre optimum d'itérations mais aussi les valeurs des paramètres ajustables que sont la vitesse d'apprentissage, le terme de mémoire (cf. Figure 1-11) et le nombre de neurones dans la couche cachée [58, 62, 63].

L'algorithme de calcul ANN présenté ici a été appliqué à l'analyse quantitative des spectres LIBS de sols pollués par J.-B. Sirven et al. [56]. Plus récemment, ce même algorithme a été utilisé par D. Diego-Vallejo et al. [64] pour la classification des cellules solaires.

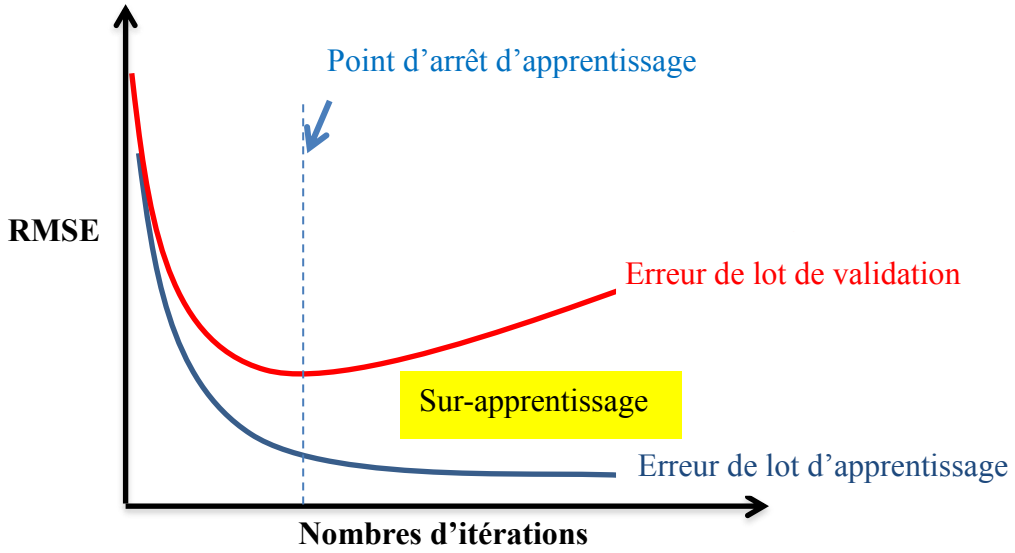


Figure 1-11 Evolution des erreurs (RMSE) en fonction du nombre d'itérations pour le lot d'apprentissage (bleu) et le lot de validation (rouge). Le point d'arrêt de l'apprentissage correspond au minimum de la courbe rouge. [58]

1.4.2.5 Architectures des réseaux

Dans l'étude présentée ici, nous avons uniquement exploité des réseaux de neurones ou ANNs à trois couches contenant une couche d'entrée recevant les données expérimentales, une couche cachée contenant un nombre de neurones à choisir et une couche de sortie fournissant la réponse de l'ANN. Dans la plupart des cas, nous avons utilisé l'architecture présentée sur la Figure 1-12-a selon laquelle un seul neurone se trouve dans la couche de sortie. La réponse fournie par un tel ANN peut être quantitative, par exemple la valeur de concentration de l'analyte mais elle peut aussi être juste un nombre utilisé pour déclencher une prise de décision, par exemple 1 pour oui et 0 pour non.

La seconde architecture que nous avons utilisée est présentée sur la Figure 1-12-b. Dans ce cas, il y a plusieurs neurones dans la couche de sortie. On utilisera cette architecture pour obtenir plusieurs informations en parallèle. En analyse quantitative, cela permet de prédire les concentrations de plusieurs analytes en un même calcul. Dans le cadre d'un tri, ça permet de trier les données non pas juste en deux classes (cas d'un seul neurone en sortie) mais en plusieurs.

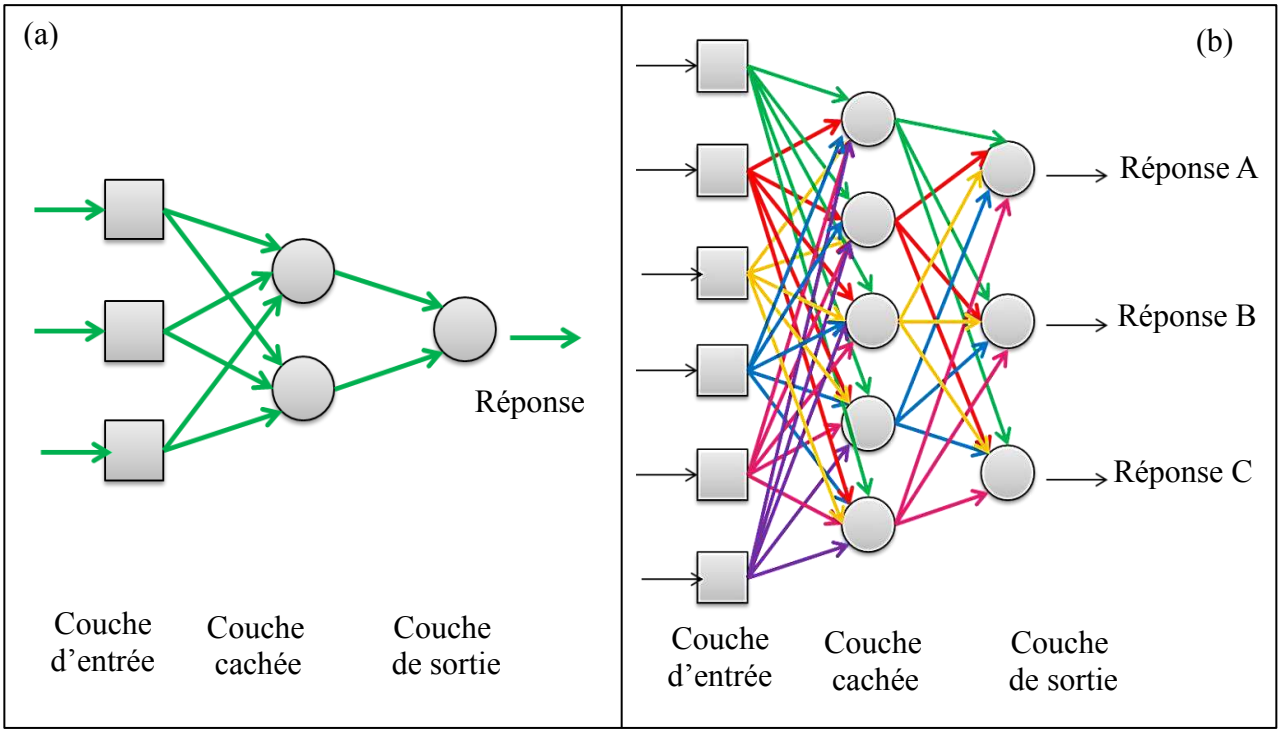


Figure 1-12 Architectures d'ANN à trois couches avec (a) un seul neurone, (b) plusieurs neurones dans la couche de sortie.

1.5 **Fiabilité d'un modèle**

Tout modèle de calcul, quel qu'il soit, doit être évalué. Pour que l'évaluation ait du sens, avant tout les données d'entrée doivent être sélectionnées avec soin. En spectroscopie, il faudra évidemment choisir les raies ou bandes spectrales les plus pertinentes, c'est-à-dire celles qui sont fortement corrélées aux valeurs des concentrations de l'analyte. Par ailleurs, il faut systématiquement répartir les données disponibles en trois lots : calibration, validation et test. Cette répartition permet de disposer de données indépendantes ; les données du lot de calibration servent à bâtir le modèle de calcul, celles de validation sont utilisées pour tester chaque modèle pendant la phase de validation croisée externe et enfin les données du lot de test sont simplement utilisées a posteriori à l'aide du modèle qui a été retenu grâce aux deux premiers lots.

Le choix des données d'entrées et la répartition dans les lots influent sur les performances du modèle de calcul. On veillera donc à choisir des données qui couvrent si possible l'ensemble de la gamme des concentrations pour chacun des trois lots. De plus, tout modèle doit être accompagné de son domaine d'applicabilité. Ainsi, on ne pourra pas prédire de concentration plus élevée que la valeur maximale exploitée au cours de l'étalonnage. A l'inverse, du côté des basses concentrations, les procédures de calcul des limites de détection et de quantification sont bien établies [65] même s'il reste encore des discussions à ce sujet dès lors que l'on a affaire à des analyses multi-variées. Enfin, le modèle retenu doit être stable, sans sur-apprentissage et avoir une véritable signification statistique. Cela signifie que le résultat

d'un modèle ne doit pas être le fruit de la chance mais au contraire d'une réelle combinaison entre les données d'entrée et la valeur de sortie.

1.5.1 Sélection des données d'entrée

La première règle consiste à sélectionner des données justes, précises et fiables [65-67]. En spectroscopie, il faut choisir des raies ou des bandes spectrales avec un bon rapport signal sur bruit et des propriétés favorables. De façon générale, cela signifie que dans un spectre, on évitera les raies susceptibles de présenter des saturations, des élargissements ou encore des interférences spectrales entre plusieurs espèces chimiques. Dans le cas particulier de la LIBS, on choisira ainsi des raies persistantes d'après la base de données du NIST, sans auto-absorption et sans interférences [68].

Notons aussi que l'échantillon analysé doit être homogène du moins à l'échelle de la zone d'analyse. En LIBS, les alliages métalliques et les solutions liquides sont considérés homogènes. Par contre les échantillons de sols sont clairement hétérogènes à l'échelle du spot laser. Dans ce cas, plutôt que de traiter chaque spectre expérimental séparément, on préfèrera traiter un spectre moyen, résultats d'une série de tirs laser en plusieurs points de la surface de l'échantillon. Les précautions expérimentales ne sont pas les mêmes en spectroscopie THz si l'on considère que les échantillons analysés peuvent être considérés homogènes à l'échelle du spot THz. Par ailleurs, il est souhaitable de systématiquement chercher à éliminer les données aberrantes afin de concentrer la construction du modèle sur des données fiables et donc de converger vers un modèle plus robuste.

Pour les analyses quantitatives, idéalement il faut que les échantillons proviennent de la même matrice. En effet, des effets de matrice ont été observés très tôt en LIBS [69-71], ce qui limite à priori le champ d'utilisation de cette technique. Malheureusement, les sols correspondent à une infinité de matrices différentes mais on cherchera cependant à constituer de grandes familles afin de pouvoir ensuite réaliser des analyses quantitatives. Notons que pour une classe de sols donnée, la phase d'apprentissage peut être longue pour prendre en compte les multiples spécificités des sols et ne pas restreindre le modèle à un sous-ensemble choisi de façon aléatoire car la sélection des données optimisée donne toujours un modèle plus performant qu'une sélection aléatoire [65].

On ne peut évaluer un modèle quantitatif que si les données sont préalablement séparées en trois lots : le lot d'apprentissage, le lot de validation et enfin le lot de test. Le premier lot sert à construire des modèles, le second à choisir le meilleur modèle en évitant le sur-apprentissage. Le dernier lot permet de connaître les performances du modèle lorsqu'il est utilisé a posteriori sur un jeu de données indépendantes. On propose de séparer, dans la mesure du possible, les données initiales de la façon suivante : 60% pour le lot d'apprentissage, 30% pour le lot de validation et 10% pour le lot de test. Chaque lot doit contenir des échantillons représentatifs de toute la gamme de concentrations et plus de 5 échantillons pour une raison de fiabilité du calcul statistique.

1.5.2 Evaluation du modèle

Un modèle peut être évalué par la méthode de validation croisée interne (cas de la PLS) ou par la méthode de validation croisée externe (cas de l'ANN). Lorsqu'on fait appel à la validation croisée interne, on ne se sert que des données du lot de calibration. On exclut du modèle à tour de rôle une donnée (méthode LOO pour leave one out) ou plusieurs (méthode LMO pour leave many out) [65]. On construit à chaque fois un modèle avec toutes les données d'entrée moins celle(s) qui a (ont) été écartée(s) puis on teste le modèle sur cette (ces) donnée(s). La validation-croisée peut être aussi externe en utilisant pour l'évaluation une série des données complètement indépendantes de celles du lot d'apprentissage. Dans le cas de l'ANN, la méthode externe est recommandée pour l'apprentissage rapide et l'optimisation du modèle. Une fois construit le meilleur modèle, les échantillons qui composent le lot de test ne doivent pas être choisis aléatoirement mais au contraire couvrir toute la gamme des valeurs de concentrations. Il est souvent annoncé que la seule méthode efficace pour valider un modèle ANN est l'approche apprentissage-validation-test [72]. Les données du lot de test devant être totalement indépendantes, elles peuvent provenir d'une campagne de mesure indépendante réalisée un autre jour par exemple.

1.5.2.1 Cas d'un modèle de Classification

Lorsque le but des analyses est la classification ou le tri devant entraîner une prise de décision, il est important d'éviter les erreurs de classification qui pourraient avoir des conséquences graves. On peut alors adopter une approche statistique semblable à celle pratiquée en médecine lorsqu'on cherche à déterminer si une maladie est présente ou pas [73]. La Figure 1-13 définit les figures de mérite qui doivent être utilisées pour déterminer les performances de la classification. En spectroscopie, le résultat du test est obtenu par l'expérience (LIBS, THz, Raman, NIR, MIR, etc…) et comparé à une valeur de référence. Dans le cadre des analyses LIBS présentées au chapitre 2, les valeurs de référence sont données par des analyses ICP-AES. Nous prendrons cet exemple LIBS vs. ICP-AES pour illustrer nos propos. Quand un échantillon passe le test (test positif) cela signifie que le résultat obtenu en traitant les données LIBS est conforme à celui donné par la technique ICP-AES. Dans ce cas, on a affaire à un vrai positif et le nombre de cas de ce type est noté A dans la figure 1-13. En revanche, si le test est positif alors que d'après la technique de référence on aurait dû obtenir un résultat négatif, on dit que l'on a affaire à un faux positif, et B désigne le nombre de ces cas dans la Figure 1-13. En outre, quand le résultat du test est négatif, conformément à ce qu'on attendait d'après la technique de référence, on dit que l'on a affaire à un vrai négatif (valeur D) et lorsque le résultat du test est négatif alors qu'on s'attendait à l'inverse d'après la technique de référence, on dit que l'on a affaire à un faux négatif (valeur C). Grâce aux quatre valeurs A, B, C et D, il est possible de définir quatre facteurs de mérite qui traduisent la capacité du test à répondre correctement à la question posée.

		Valeur de référence		
		Positive	Negative	
Résultat de test	Positif	A vrai Positif	B Faux Positif	Valeur de prédiction positive = A/(A+B)
	Négatif	C Faux Négatif	D Vrai Négatif	Valeur de prédiction négatif= D/(C+D)
		Sensitivité = /(A+C)	Spécificité = D/(B+D)	

Figure 1-13 Définition de figure de mérite reliée à une classification d'un test.

- La sensibilité définie par la quantité A/(A+C). Ce facteur mesure le taux des résultats positifs par rapport à tous les échantillons réellement positifs. En médecine ce facteur permet de mesurer le taux de patients déclarés malades à l'aide du test par rapport au nombre total de patients malades. En analyse environnementale cela permet de mesurer par exemple le taux d'échantillons déclarés pollués à l'aide du test par rapport au nombre total d'échantillons pollués. La sensibilité doit être égale à 100% pour garantir que tous les patients malades seront déclarés comme tels par le test ou encore que tous les échantillons pollués seront déclarés comme tels par le test. Donc, dans le cas idéal, C=0. A l'inverse, si la sensibilité est inférieure à 100 % cela traduit le fait que C≠0 et donc qu'il existe des faux négatifs. Dans ce cas, des patients malades seront déclarés en bonne santé ou encore des échantillons de pollués seront déclarés non pollués.
- La spécificité est définie par la quantité D/(B+D). Ce facteur mesure la proportion d'échantillons déclarés négatifs par le test par rapport à tous les échantillons vraiment négatifs. Dans le cas idéal, on souhaite que le nombre de faux positifs B soit égal à zéro. En médecine, B=0 signifie que tous les patients sains sont considérés comme tels grâce au test et en environnement, que tous les échantillons non pollués sont considérés comme tels par le test. A l'opposé, lorsque B ≠0, cela signifie que certains patients en bonne santé seront déclarés malades par le test, ou encore que certains échantillons non pollués seront considérés pollués. Cette situation correspond à une spécificité inférieure à 100%.

Il est facile de comprendre que la sensibilité et la spécificité sont des facteurs de mérite complémentaires. L'idéal est d'obtenir 100% pour les deux. De manière générale, le meilleur test sera celui qui donne les valeurs les plus élevées pour ces deux facteurs simultanément. Mais selon les cas à traiter, on peut préférer favoriser soit la sensibilité soit la spécificité par rapport aux conséquences que cela entraînera.

- La valeur de prédiction positive est définie par la quantité A/(A+B). Ce facteur indique par exemple la proportion de patients réellement malades par rapport à tous les patients déclarés malades par le test. Là encore, ce facteur doit être égal à 100% dans l'idéal. Quand sa valeur est inférieur à 100%, cela signifie que B≠0 et donc qu'il existe

des faux positifs. Dans l'exemple précédent, il y a alors des patients en bonne santé qui sont déclarés malades par le test.

- La valeur de prédiction négative est définie par D/(C+D). Idéalement cette valeur doit aussi être égale à 100%. Quand C≠0, cela signifie qu'il existe des faux négatifs. Dans ce cas, certains patients malades sont déclarés en bonne santé par le test ou encore des échantillons pollués sont déclarés comme non-pollués par le test.

En résumé, quelle que soit la technique d'analyse, un test de classification doit être évalué soit par les quatre valeurs A, B, C et D soit par les quatre facteurs de mérite : sensibilité, spécificité, valeur de prédiction positive et valeur de prédiction négative [73]. Ces deux approches donnent évidemment la même information, même si la deuxième est préférée car elle est plus facile à interpréter pour évaluer les performances d'un test. Notons que les conséquences d'un test ayant des performances insuffisantes peuvent être dramatiques. En effet, B≠0 peut conduire à administrer un traitement médical ou chirurgical à un patient en bonne santé et à l'opposé lorsque C≠0 on risque de ne pas donner de traitement à un patient malade. De la même façon, B≠0 peut conduire à déclencher un traitement de dépollution pour un site non-pollué, alors que C≠0 conduit à ne pas traiter un site qui nécessite une dépollution.

Pour réduire les risques de classification fausse, il est recommandé d'introduire une tolérance lorsqu'on donne les résultats du test. Cela se comprend aisément dans le cas d'un tri par rapport à un seuil. Si les valeurs supérieures à la valeur seuil sont à classer dans la classe 1 et celles inférieures au seuil dans la classe 2, toutes les valeurs qui se situent à proximité du seuil peuvent basculer d'un côté ou de l'autre pour de simples raisons de fluctuations expérimentales. Une discussion poussée sur les performances de classification d'un modèle chimiométrique est proposée dans les références [30, 74], plus spécialement dans le cas où le tri porte sur plus de deux classes.

1.5.2.2 Cas d'un modèle quantitatif

Dans le cas d'une analyse quantitative, les performances d'un modèle sont directement reliées à sa capacité à prédire correctement les valeurs de concentrations de l'analyte. Les facteurs de mérite les plus couramment utilisés sont les coefficients R^2 et Q^2 définis par :

$$R^2 = \left(\frac{\Sigma(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\Sigma(y_i - \bar{y})^2 \, \Sigma(\hat{y}_i - \bar{\hat{y}})^2}} \right)^2 \qquad (1\text{-}57)$$

$$Q^2 = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2} \qquad (1\text{-}58)$$

Dans ces expressions, y_i désigne la valeur de référence de la concentration de l'échantillon i, $\bar{y}$ la valeur moyenne des concentrations de référence, $\hat{y}_i$ est la valeur de la concentration prédite par le modèle pour l'échantillon i et $\bar{\hat{y}}$ est la valeur moyenne des concentrations prédites par le modèle [67]. R^2 est le coefficient de corrélation entre les valeurs calculées par le modèle et les valeurs de référence. Q^2 permet de mesurer l'aptitude du modèle à calculer

des valeurs proches des valeurs de référence. Notons que les facteurs R^2 et Q^2 peuvent être calculés séparément pour les lots d'apprentissage, de validation et de test. Idéalement, les valeurs de R^2 et Q^2 doivent être égales à 1 lorsque la corrélation est parfaite et que les valeurs calculées par le modèle sont exactement égales aux valeurs de référence. Certains auteurs considèrent qu'un modèle est acceptable lorsque $R^2>0.6$ et $Q^2 >0.5$ [65-67]. La valeur de R^2 peut être égale à 1 dans le cas de faible degré de liberté ou des variables multi-colinéaires [65]. Par ailleurs la valeur de Q^2 peut être très proche de 1 pour le lot de calibration mais très faible pour le lot de validation spécialement dans le cas d'un modèle non-linéaire [67].

Par conséquent, les coefficients R^2 et Q^2 ne suffisent pas pour caractériser complètement un modèle et d'autres facteurs de mérite sont utilisés dont notamment la moyenne des erreurs quadratiques ou RMSE [75] définie par la relation (1-59).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{1-59}$$

Dans cette expression, N est le nombre de données dans le lot de données considéré. Ce facteur peut lui aussi être calculé pour le lot de calibration (RMSEC) tout comme pour le lot de validation (RMSEV) ou de test (RMSET). Dans tous les cas, on cherche à obtenir une valeur de RMSE aussi faible que possible pour chacun des trois lots. Notons que des valeurs très différentes pour RMSEC et RMSEV révèlent un risque de modèle instable ou encore de sur-apprentissage. La cause peut aussi se situer dans une mauvaise répartition des données dans les différents lots. Remarquons que le calcul de RMSE est fortement influencé par les valeurs de concentrations les plus élevées. Pour contourner ce problème, il est possible de calculer l'erreur relative moyenne pour laquelle les données ont toutes la même contribution et qui est définie par la relation (1-60).

$$ER\ (\%) = \frac{\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i}}{N} \times 100 \tag{1-60}$$

Comme pour RMSE, l'erreur relative moyenne peut être calculée pour le lot de calibration (ERC), le lot de validation (ERV) et le lot de test (ERT). Là encore, si les valeurs pour les trois lots sont très différentes, ce peut être le signe d'un modèle instable ou se trouvant dans un régime de sur-apprentissage.

1.5.3 Signification statistique d'un modèle quantitatif

Il est possible qu'un modèle donne des résultats tout à fait satisfaisants alors que pourtant, il n'existe pas de réelle corrélation entre les données d'entrée X et les valeurs calculées en sortie Y. Dans ce cas, les performances acceptables sont le fruit du hasard. Pour être sûr qu'une vraie corrélation existe entre les données d'entrée X et de sortie Y, on a recours à une procédure connue sous le nom de Y-Randomization [76]. Elle a pour but de vérifier que le modèle est réellement pertinent et elle est reconnue comme étant la plus puissante procédure de validation [77]. En effet, si la méthode de validation croisée permet d'évaluer la capacité prédictive d'un modèle, elle ne permet pas en revanche de déterminer si les résultats du

modèle ont une réelle signification statistique. Seule la procédure de Y-randomization permet de faire cette validation. Elle consiste à mélanger aléatoirement les données Y lors de la phase de calibration afin de détruire toute corrélation éventuelle entre les données d'entrée X et les données Y. Une fois le mélange réalisé, il n'y a aucune raison que le modèle soit capable de prédire correctement les valeurs Y puisqu'il n'existe plus aucune corrélation de ces valeurs avec les données d'entrée X. Par conséquent, on s'attend à ce que les performances de prédiction du modèle soient complètement dégradées et ceci peut être vérifié grâce au calcul des facteurs de mérite introduits précédemment, à savoir R^2, Q^2 et RMSE. On répète la procédure de mélange aléatoire 25 fois afin de disposer d'un résultat statistique [76]. Évidemment, les valeurs de R^2 et Q^2 doivent être plus élevées dans le cas du modèle initial qu'après un mélange aléatoire des données Y. De même, il faut que la valeur de RMSE soit inférieure dans le cas normal qu'après un mélange aléatoire. Si c'est le cas, on peut dire que le modèle est statistiquement significatif. Notons que certains auteurs ont proposé comme critère objectif $R^2 < 0.3$ et $Q^2 < 0.05$ après application de la procédure de mélange aléatoire pour conclure à la réussite de la procédure de Y-randomization [66].

1.5.4 Domaine d'applicabilité d'un modèle quantitatif

Un modèle quantitatif est construit à partir d'une procédure d'étalonnage basée sur la connaissance des concentrations d'une série d'échantillons connus. En conséquence, le modèle présente des valeurs limites provenant des valeurs minimale et maximale des concentrations du lot de calibration. Concernant la valeur maximale, il est de coutume de considérer qu'aucune valeur de concentration supérieure à la valeur maximale du lot de calibration ne pourra être prédite. Ce qui motive l'idée de cette limite supérieure est que le modèle pourrait éventuellement devenir non-linéaire pour des raisons de saturation et dans ce cas, l'application du modèle quantitatif au-delà de la limite maximale n'a pas de sens. En ce qui concerne la limite inférieure, les choses sont différentes dans la mesure où tous les modèles sont considérés linéaires pour les basses concentrations. Ainsi la droite d'étalonnage peut raisonnablement être prolongée vers les valeurs de concentrations plus basses que la limite inférieure fixée par le lot de calibration. Dans ce cas précis, il est de coutume de définir la limite de détection et plus important encore la limite de quantification. Cette dernière donne la valeur minimale en dessous de laquelle il sera impossible de prédire une valeur de concentration. Dans le cas d'un modèle uni-varié, la définition de la limite de quantification est bien établie [78]. En revanche, dans le cas d'une analyse multi-variée, plusieurs définitions de la limite de détection ont été proposées [79, 80], ce qui montre qu'il existe encore des discussions sur ce sujet. Notons cependant que les limites de détection et de quantification n'ont pas été étudiées dans le cadre de ce travail de thèse. Cependant le domaine d'applicabilité du modèle quantitatif est non seulement déterminé par les limites décrites juste avant mais aussi par les matrices des échantillons. Plus précisément la matrice d'un échantillon inconnu doit être similaire aux matrices caractérisant les échantillons du lot d'apprentissage afin que le modèle puisse être utilisé.

1.6 **Conclusion**

Dans ce chapitre, nous avons présenté les méthodes de chimiométrie (ACP, PLS, ANN) qui ont ensuite été appliquées à deux types de spectroscopie, LIBS et THz. Un accent particulier a été mis sur la sélection et la préparation des données ainsi que sur l'évaluation des performances d'un modèle, que celui-ci soit quantitatif ou destiné à de la classification. Nous avons plus particulièrement détaillé le cas d'un réseau de neurones artificiel à 3 couches en discutant des conditions de convergence et du risque de sur-apprentissage. La pertinence et la fiabilité d'un modèle en général ont été également traitées au travers des procédures de validation croisée et de Y-randomization.

Ce chapitre a permis d'introduire les outils de chimiométrie ainsi que les bonnes pratiques nécessaires à leur utilisation de façon sérieuse. Les deux chapitres qui suivent concernent la mise en application des outils de chimiométrie à deux techniques de spectroscopie : la LIBS et la spectroscopie THz.

Chapitre 2. **Chimiométrie appliquée à la LIBS**

La LIBS est une technique de spectroscopie d'émission atomique à partir d'un plasma induit par laser. Elle permet des analyses sur site, rapide et quasiment sans endommagement de l'échantillon. C'est grâce à ces avantages mais aussi au fait qu'elle permette des mesures à plusieurs mètres de distance que la LIBS est à l'heure actuelle l'une des techniques exploitées par la NASA pour analyser le sol martien. Un système LIBS est en effet embarqué sur le rover Curiosity qui s'est posé sur le sol martien en août 2012. Les résultats présentés dans ce chapitre concernent des analyses LIBS sur site de sols pollués, sur Terre. Les campagnes de mesure ont été coordonnées par l'ADEME et le BRGM et réalisées avec un instrument LIBS mobile commercialisé par la société IVEA, spécialiste de la LIBS. Le LOMA est venu compléter ce consortium en apportant son expertise dans le domaine du traitement des données dans le cadre du projet CALIPSO, projet industriel financé par l'ADEME.

Ce chapitre va donc aborder brièvement le principe de la LIBS puis la description de l'instrument utilisé pour cette étude. Nous discuterons ensuite de la sélection des échantillons sur le site par fluorescence X, puis de la préparation des échantillons et nous verrons enfin le protocole d'analyse LIBS. En complément, nous présenterons les analyses de laboratoire ICP-AES qui ont fourni les valeurs de référence des concentrations et tout ceci nous permettra d'analyser les performances des analyses LIBS lorsqu'on traite les données avec des réseaux de neurones. Nous présenterons ensuite le travail concernant le traitement des données LIBS par chimiométrie avec pour objectif l'analyse quantitative sur site d'échantillons de sol. L'ACP permet de décrire les données et d'identifier les quelques éléments qui caractérisent au mieux les différentes matrices. Cela permet de sélectionner les données d'entrée qui seront injectées dans le réseau de neurones pour les analyses quantitatives. L'objectif à atteindre est une erreur relative moyenne inférieure à 20% pour les mesures LIBS réalisées directement sur site. Et nous verrons qu'un seul modèle ANN peut s'avérer insuffisant dans certains cas liés à des matrices très différentes ou à des concentrations qui se répartissent sur une très grande gamme. Dans ce cas, on fait appel à plusieurs modèles ANN, chacun étant optimisé pour une matrice donnée et/ou pour une gamme de concentration donnée. On présentera enfin le

transfert industriel qui a permis la conception et la fabrication d'un module logiciel implanté dans le logiciel Analibs commercialisé par la société IVEA.

2.1 **Généralités sur la LIBS**

En 2012, la LIBS (Laser-Induced Breakdown Spectroscopy) a fêté son 50ème anniversaire et rappelons que c'est une équipe française qui est à l'origine de la première publication scientifique sur ce sujet en 1963 [81]. Dès 1968, des effets de matrices ont été observés sur les spectres LIBS [69], conduisant à penser que l'analyse quantitative des données LIBS allait être compliquée dans un certain nombre de cas. En 1996, Eppler et al. [82] ont démontré l'influence néfaste des effets de matrices sur la limite de détection du plomb et du baryum dans des échantillons de sol. Malgré ces bémols, la LIBS reste très attractive car utilisable sur site [83] à l'opposé d'autres techniques telles que l'ICP-AES ou l'ICP-MS. Elle est en constant développement, que ce soit sur le plan de l'instrumentation, du traitement de données ou des applications. Cela se traduit par un nombre toujours croissant de publications et la parution de plusieurs livres de référence [22, 23, 84] ces dernières années.

A l'instar des autres techniques de spectroscopie d'émission atomique, la LIBS est basée sur l'atomisation et la vaporisation des échantillons [23], l'excitation des atomes et la détection de la lumière émise. De même, la LIBS doit faire l'objet d'un étalonnage avant de permettre la quantification d'un élément dans un échantillon inconnu. En LIBS, la vaporisation, l'atomisation et l'excitation sont produites par une forte irradiance laser, ce qui signifie une forte puissance par unité de surface. Pour réaliser ces conditions, on utilise un laser impulsionnel, ce qui fait croître la valeur de la puissance crête et un dispositif de focalisation afin de faire converger le faisceau laser sur une petite surface. Si l'irradiance est supérieure au seuil d'ablation du matériau analysé, il y a vaporisation et formation d'un plasma contenant des espèces (ions, atomes et molécules) excitées et des électrons libres. La désexcitation des ions et des atomes est à l'origine d'un rayonnement caractérisé par un spectre des raies atomiques détecté à l'aide d'un spectromètre UV-Visible. L'analyse du spectre LIBS consiste donc d'une part à reconnaître les raies atomiques à partir d'une base de données et d'autre part à mesurer leurs intensités à des fins d'analyse quantitative. La Figure 2-1 décrit un montage LIBS typique.

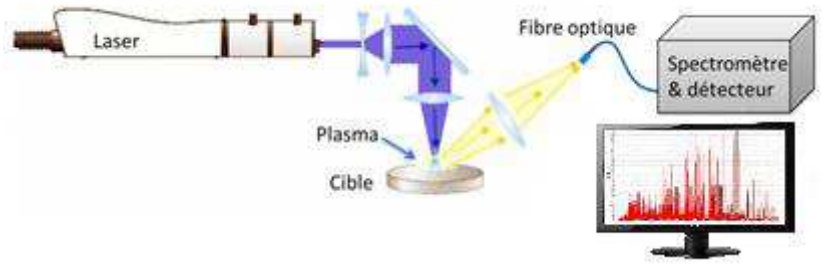


Figure 2-1 Montage LIBS typique

Notons qu'un spectre LIBS est détectable pendant une durée très courte allant de quelques nanosecondes à quelques microsecondes maximum selon les conditions expérimentales. Aux temps cours, un spectre continu se superpose au spectre de raies atomiques qu'il peut

masquer. Il est dû au rayonnement de freinage des électrons plus connu sous le nom de bremsstrahlung. Cela impose de démarrer la détection du spectre LIBS après un certain délai, contrôlé et optimisé, après l'impact du laser. De plus, une durée d'acquisition astucieusement choisie permet d'optimiser le rapport signal/fond [23, 63].

La Figure 2-2 représente un spectre LIBS typique sur la bande spectrale 200-900 nm. On distingue clairement les raies spectrales de plusieurs éléments tels que par exemple Al, Ca, Ba, Fe, Si.

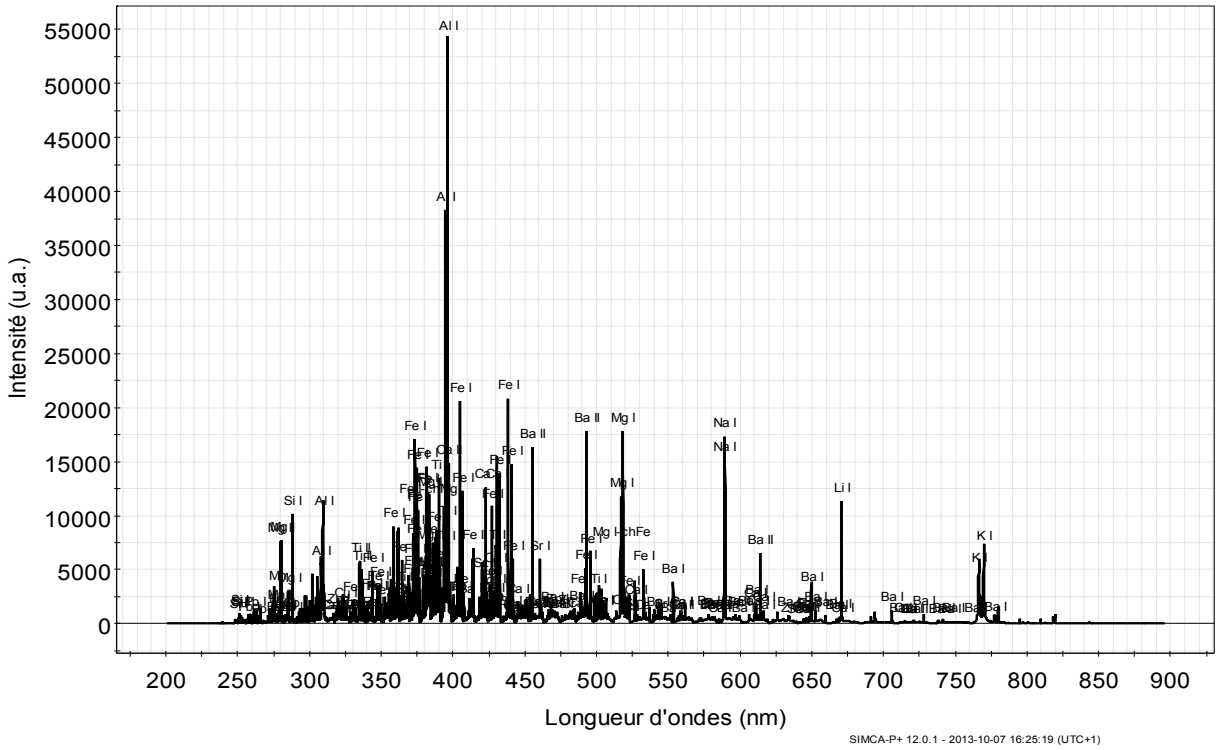


Figure 2-2 Spectre LIBS typique d'un échantillon de sol

Le Laser le plus utilisé dans le domaine de la LIBS est le Nd:YAG à la longueur d'onde de 1064 nm, et parfois aux longueurs d'ondes harmoniques car il est connu que l'ablation à l'aide d'un rayonnement UV est plus efficace que celle issue d'un rayonnement infrarouge [63]. Typiquement, les énergies des impulsions laser vont de 1 à 1000 mJ, et la durée des impulsions est de l'ordre de 10 ns [85] alors que le taux de répétition est classiquement de 10 à 100 Hz.

Les spectromètres les plus couramment utilisés pour les mesures sur site sont de type Czerny-Turner compacts avec capteur CCD. Notons cependant que les expériences LIBS de laboratoire sont basées assez souvent sur le spectromètre de type Echelle qui, grâce à sa capacité à disperser la lumière selon deux dimensions, offre à la fois une très grande résolution spectrale ($\lambda/\Delta\lambda$ = 10000) et une large gamme spectrale (200-1000 nm). Ces spectromètres sont le plus souvent équipés d'un détecteur avec intensificateur afin de tirer profil d'un bon contrôle temporel et d'une amplification du signal. Notons enfin que le spectromètre de type Paschen-Rünge associé à plusieurs photomultiplicateurs permet une

analyse multi-voies à haute cadence (plusieurs kHz), ce qui ouvre potentiellement la voie à des applications industrielles [86].

Dans le cadre de ce travail, les mesures LIBS sur site ont été effectuées à l'aide de deux systèmes LIBS commercialisés par la société IVEA : EasyLIBS qui est un système portable de la taille d'une perceuse et MobiLIBS qui est un système mobile plus lourd mais aussi de meilleure résolution spectrale. On ne présentera dans ce mémoire de thèse que les résultats obtenus avec le système MobiLIBS. Les données provenant de l'autre système ont cependant été traitées selon la même méthodologie et ont également donné des résultats très satisfaisants. Le système MobiLIBS a été transporté dans une camionnette sur les différents sites d'analyse. Les éléments principaux qui le composent sont : un laser Nd:YAG à 266 nm/20 Hz/2,5 mJ par impulsion, une tête optique pour la mise en forme et la focalisation du faisceau, pour le contrôle de l'énergie et du nombre de tirs laser ainsi que pour la collecte du signal à analyser et enfin un spectromètre de type Echelle équipé d'une caméra CCD intensifiée, le tout piloté par ordinateur. Ce dispositif est par ailleurs équipé d'une chambre d'analyse qui contient une platine de translation 3-axes motorisée pour déplacer l'échantillon. La taille du spot laser est de 50 µm de diamètre sur l'échantillon.

Après quelques essais préliminaires effectués en laboratoire sur des pastilles de sols sélectionnées de façon arbitraire, les paramètres qui ont été retenus pour les mesures LIBS sont :

- 2 pré-tirs pour nettoyer la surface de l'échantillon
- 20 tirs accumulés sur un même point de mesure
- 25 points de mesures à la surface de l'échantillon
- Utilisation du spectre moyen calculé à partir des spectres obtenus avec ces 25 points et considéré comme représentatif de l'échantillon pour l'analyse des données.

Pour les analyses LIBS sur site nous avons systématiquement appliqué ce protocole. De plus, avant chaque mesure sur site, le bon fonctionnement de l'instrument était vérifié à l'aide d'un échantillon de référence en acier.

Notons que pour un laser cadencé à 20 Hz, la mesure ne dure qu'une seconde en chaque point, ce qui entraîne, en tenant compte des temps de déplacement des platines de translation, moins de deux minutes par échantillon.

Typiquement, la masse ablatée lors d'une analyse LIBS est de l'ordre du microgramme par impulsion et dépend fortement de l'énergie de l'impulsion laser [87]. Le plasma LIBS dont la taille est de quelques millimètres au maximum, peut être caractérisé par ses paramètres physiques intrinsèques à savoir la température et de la densité électronique. En faisant l'hypothèse qu'il existe au sein du plasma un équilibre thermodynamique local, on considère que la distribution statistique de Boltzmann est vérifiée, ce qui permet de retrouver expérimentalement la température du plasma à l'aide d'un diagramme de Boltzmann [88]. Cette pratique est très répandue et permet d'estimer la température initiale d'un plasma LIBS à une valeur de l'ordre de 10000 K. La densité électronique est quant à elle mesurée à partir de l'élargissement Stark de certaines raies spectrales. Elle est de l'ordre de 10^{18} cm^{-3} aux premiers instants après le tir laser. Il est important de noter que les caractéristiques d'un plasma ont une très grande influence sur les spectres LIBS et par conséquent sur l'analyse

quantitative [23]. En effet, l'intensité d'une raie spectrale associée à l'analyte est reliée non seulement à la concentration de l'analyte mais aussi aux propriétés du plasma et par conséquent, à la nature de l'échantillon et du gaz ambiant ainsi qu'aux caractéristiques de la source laser.
Finalement l'analyse LIBS quantitative n'a de sens que si la composition du volume ablaté est représentative de l'ensemble de l'échantillon et si la composition du plasma est identique à celle de l'échantillon solide. On dit alors que l'ablation est stœchiométrique. Cette condition est nécessaire dans le cas d'une analyse auto-calibrée [89] et bien évidemment souhaitable dans tous les cas. Cependant, notons qu'une analyse s'appuyant sur un étalonnage peut être parfaitement menée même si la condition de stœchiométrie n'est pas totalement respectée. Par ailleurs, la méthode auto-calibrée s'appuie sur deux autres hypothèses : le plasma doit être à l'équilibre thermodynamique local et optiquement fin. Or, il est facile de constater que certaines raies présentent un élargissement, un aplatissement du pic, et parfois même un creusement de trou qui traduisent clairement l'existence d'un phénomène d'auto-absorption au sein du plasma. Une procédure d'étalonnage permet de prendre en compte ces effets mais on comprend ici tout l'intérêt d'une analyse multivariée afin de traiter non pas une raie unique mais un ensemble de données. Enfin, le phénomène d'auto-absorption est très nuisible à la linéarité de la courbe d'étalonnage et on voit donc apparaître clairement l'intérêt d'utiliser le réseau de neurones artificiels qui permet de modéliser des comportements non-linéaires, contrairement aux techniques matricielles telles que la PLS qui s'appuient, elles, sur de l'algèbre linéaire.

2.2 **Campagnes de mesures**

Plusieurs campagnes de mesures LIBS ont été réalisées sur des sites géologiques différents dans le cadre du projet CALIPSO financé par l'ADEME. Les échantillons de sols collectés lors de ces campagnes ont révélé des différences importantes de concentrations et donc de matrices. Les trois sites étudiés sont les suivants :

- **Site Saint-Laurent le minier (SLM)**

Les échantillons proviennent d'une part des bassins de décantation et des stériles d'une ancienne mine et d'autre part de terrains situés dans le village. Ainsi, les échantillons se répartissent entre anthroposols, sols, roches et sédiments alluvionnaires. Ils sont généralement riches en carbonates et en minerais.

- **Site Metaleurope (ME)**

Les échantillons collectés sur ce site sont tous considérés comme des sols au sens géologique du terme, contrairement à ceux provenant du site SLM. On peut donc s'attendre à obtenir une matrice mieux définie liée à une variabilité des concentrations bien moindre que pour SLM.

- **Site Saint-Sébastien d'Aigrefeuille (SEB)**

Cet ancien site minier est caractérisé par une matrice alumino-silicate et comme pour le site SLM, les échantillons proviennent soit de stériles, soit de bassins de décantation (fine granulométrie de sol), soit encore du terrain naturel non lié à l'activité minière.

Notons que les trois sites retenus pour cette étude sont connus comme étant pollués. Ils sont par conséquent placés sous le contrôle de l'ADEME. Dans ce qui suit, nous allons tout d'abord exposer comment les échantillons ont été sélectionnés sur les différents sites géologiques à l'aide d'une mesure XRF directe, puis nous présenterons les analyses ICP-AES de laboratoire effectuées a posteriori et qui ont valeur de référence, et nous décrirons enfin le protocole de mesure LIBS.

2.2.1 Echantillonnage des sites par mesures XRF

Les géologues utilisent couramment la fluorescence X (XRF) pour les analyses sur site, considérée comme étant une méthode simple, rapide et recommandée en priorité pour faire des cartographies de terrain de large surface, pour des quartiers, des régions, et des pays [90]. Les échantillons de sols prélevés sur site ont donc été immédiatement analysés par XRF à l'aide d'un instrument portable semblable à un pistolet (Niton) (cf. Figure 2-3). La fluorescence X est similaire à la fluorescence que l'on connaît dans le domaine du visible. Une source de rayons X excite les atomes de l'échantillon, si bien que des électrons des couches internes, après avoir été excités par la source X, regagnent la couche K en cédant des photons X caractéristiques de chaque atome. L'intensité d'une raie dépend de la probabilité de transition entre les deux niveaux électroniques et de la concentration de l'atome dans l'échantillon de telle sorte qu'une analyse quantitative directe est tout à fait possible [91, 92]. Il faut savoir que les photons de la source ne pénètrent que jusqu'à 2 mm de profondeur dans les échantillons de sols. Au-delà, l'élément chimique recherché ne sera pas détecté.

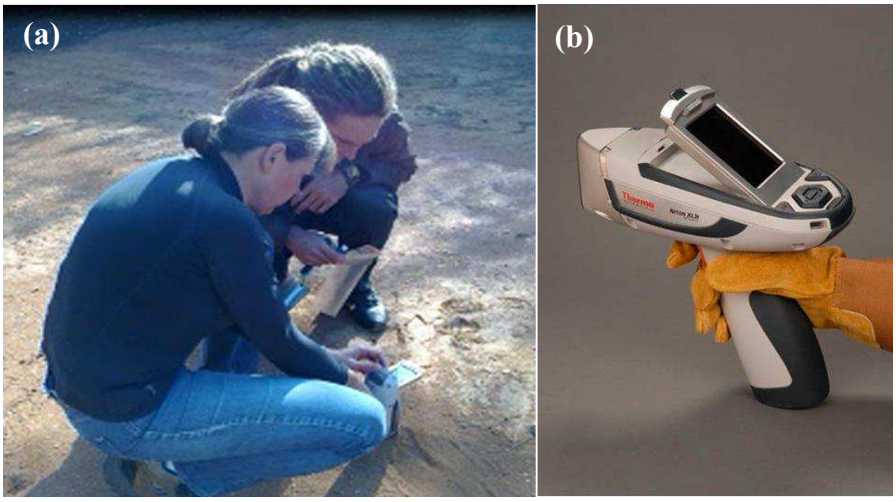


Figure 2-3 (a) Utilisation du système de fluorescence X portable lors d'une campagne sur site (SLM-bassin de décantation). (b) Photo commerciale de l'analyseur XRF portable (Niton XL3p)

Chaque échantillon de sol prélevé est préalablement tamisé à 2 mm pour éliminer les cailloux puis homogénéisé avant d'être analysé sur site par XRF et enfin divisé en deux portions, la première étant destinée à l'analyse LIBS sur site tandis que la seconde est destinée à l'analyse

par ICP-AES en laboratoire. Précisons que malgré son utilisation très répandue aujourd'hui, la fluorescence X portable ne doit pas être considérée comme une technique analytique suffisamment fiable pour des analyses quantitatives [93]. C'est pourquoi nous l'avons exploitée ici uniquement à des fins d'aide à l'échantillonnage. La mesure XRF directe dure typiquement 60 secondes en réalisant plus précisément trois mesures successives (30s, 15s, 15s). Le mode de calibration choisi est optimal pour la quantification des sols. Notons enfin que la XRF est utilisée non pas pour fournir une véritable analyse quantitative mais plutôt pour évaluer sur site la variabilité des concentrations des échantillons. Ainsi, d'après les mesures XRF, nous avons collecté sur les trois sites présentés précédemment des échantillons de sols ayant des gammes de concentration du type :

- [Pb] : 50-170 000 ppm
- [Zn] : 50-450 000 ppm

Les valeurs des concentrations sont mesurées pour les autres éléments chimiques comme (Cd, As, Sb, Fe, Sn,…).

La technique XRF n'étant pas considérée comme technique de référence, nous ne comparerons pas les résultats LIBS aux résultats XRF car cette comparaison n'entre pas dans le champ de nos travaux.

2.2.2 Mesures ICP-AES en laboratoire pour les valeurs de référence

L'ICP-AES est aujourd'hui considérée comme une méthode de référence pour l'analyse élémentaire. Avant de décrire la technique elle-même, nous mettons l'accent sur l'importance de la préparation des échantillons par minéralisation.

Minéralisation des échantillons

Pour des analyses par ICP-AES, la dissolution et la minéralisation sont des étapes préalables très importantes. La dissolution permet généralement d'obtenir l'analyte et les autres éléments sous forme ionique simple mais ce n'est malheureusement pas efficace pour les échantillons de sols, sédiments, plantes, et les tissus biologiques. Dans le cas précis des sols, la dissolution ne conduit pas à la décomposition des matrices et l'analyte peut ainsi être masqué s'il est incorporé dans une molécule organique. On doit alors avoir recours à une minéralisation. Les deux méthodes de minéralisation qui sont couramment utilisées sont :

- La minéralisation par voie sèche
  Il s'agit d'une calcination à haute température pendant plusieurs heures dans un four à moufle (450-550°C). Cette méthode est particulièrement adaptée pour décomposer la matrice organique, mais il y a cependant un risque d'évaporation des éléments volatiles comme Hg, As, ou Se ou encore Pb, Cd, ou Tl volatiles à plus haute température. On peut cependant éviter cette évaporation indésirable en ajoutant de l'oxyde de magnésium ou du nitrate de magnésium.

- La minéralisation par voie humide
  Il s'agit d'une oxydation de la matrice organique dans une phase aqueuse à haute température en présence d'un mélange d'acides et de peroxyde d'hydrogène ou bien d'un mélange d'acide perchlorique et d'acide fluorhydrique. Cette technique est largement répandue pour la minéralisation des sols et des sédiments [94, 95] mais ne convient pas à l'analyse du tellure, du sélénium, du mercure et de l'arsenic.

Dans le cadre de cette étude, le BRGM - laboratoire partenaire du projet - a utilisé la méthode de fusion décrite par la norme internationale ISO 14869-2 :2002 et citée dans les références [96, 97]. Les échantillons de sols sont broyés à 80 µm. Dans un creuset, à 1 g d'échantillon on ajoute 3 g de peroxyde de sodium et on place le mélange une heure dans un four à 450°C. On verse ensuite le contenu du creuset dans un tube, et on ajoute de l'acide chlorhydrique et de l'eau jusqu'à 100 ml ; la solution est prête pour l'analyse. Enfin, 10 % des échantillons sont dupliqués pour évaluer la fiabilité de l'analyse.

Mesures par ICP-AES

L'ICP-AES est une spectroscopie d'émission atomique qui est devenue aujourd'hui une méthode de référence. L'échantillon est préparé sous forme d'un aérosol qui est ensuite excité à l'aide d'une torche à plasma. La Figure 2-4 décrit le montage classique associé à cette technique de spectroscopie, depuis la nébulisation de l'échantillon, vers la torche à plasma et jusqu'à la détection. Une torche est constituée de tubes en quartz coaxiaux dans lesquels circule un gaz d'argon. Une bobine fournit au tube une puissance d'excitation de l'ordre du kW à une fréquence de 27 ou 41 MHz. L'ionisation du flux d'argon est amorcée par la décharge d'une bobine Tesla puis les ions et les électrons interagissent avec le champ magnétique variable produit par la bobine d'induction. Cette interaction contraint les ions et les électrons à décrire une trajectoire circulaire et donc un échauffement par effet Joule qui résulte de la résistance à ce mouvement. La température de ce plasma est typiquement de ~ 10 000 K. Ce plasma est stationnaire contrairement à celui exploité en LIBS et résultant d'une ablation laser. Un spectromètre UV-visible détecte la lumière émise par le plasma. L' échantillon est nébulisé par un nébuliseur à flux croisés d'argon ou encore par un nébuliseur à ultrasons et les fines gouttelettes qui en résultent sont transportées dans la torche par le courant d'argon qui traverse le tube central en quartz à raison de 0,3 à 5 l/min [98]. Notons que l'injection de l'échantillon peut aussi se faire par ablation laser LA-ICP-AES mais ce n'est pas la technique qui a été utilisée ici. L'avantage de la technique est que, avant d'atteindre la zone d'observation dans le plasma, les atomes de l'échantillon passent environ 2 ms dans une région où la température est comprise entre 4000 et 8000 K. il en résulte une atomisation plus complète et une diminution des interférences chimiques. De plus, les interférences d'ionisation sont faibles, voire inexistantes probablement parce que l'ionisation d'argon maintient une concentration en électrons élevée et sensiblement constante. Un autre avantage est que l'atomisation se produit dans un milieu chimiquement inerte, ce qui allonge la durée de vie de l'analyte. De plus le profil de température du plasma est uniforme de sorte que l'on n'observe pas de phénomène d'auto-absorption comme c'est le cas en LIBS. Il s'ensuit que les courbes d'étalonnage sont généralement bien linéaires sur un domaine de

concentrations s'étendant sur plusieurs ordres de grandeurs. Pour ces raisons, la technique d'ICP-AES a été retenue comme méthode de référence pour les mesures de concentrations des échantillons de sol. Les mesures LIBS peuvent donc être confrontées à des valeurs de référence, ce qui permet d'éviter d'avoir recours à des échantillons de sol de référence comme ceux du NIST. Notons cependant que l'ICP-AES nécessite une préparation très lourde de l'échantillon et qu'elle est totalement destructive. Cependant, suite à sa robustesse, ce système sera utilisé comme méthode de référence pour comparer et valider les autres méthodes moins robustes comme la LIBS et la fluorescence X [99-102].

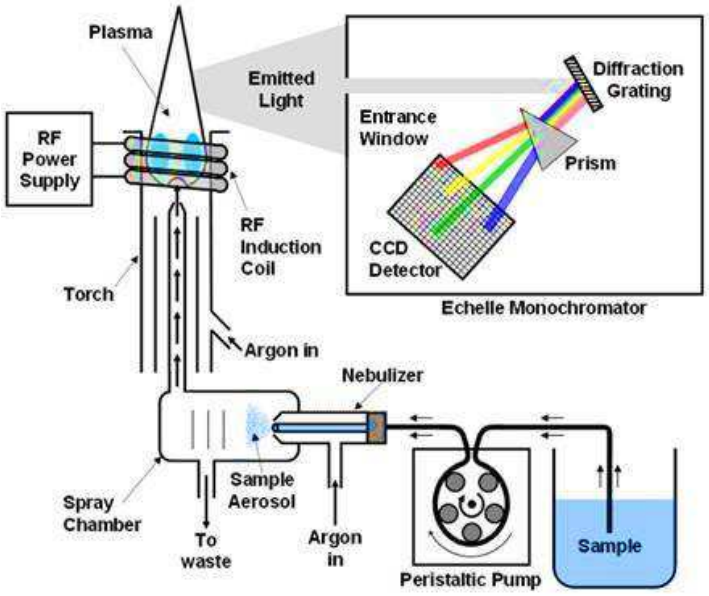


Figure 2-4 Montage typique d'un dispositif ICP-AES. Extrait de [103]

2.2.3 Préparation des échantillons pour les mesures LIBS

Les échantillons de sols nécessitent une préparation simple préalable aux mesures LIBS [104]. En effet, le taux d'humidité ou encore la compacité sont des facteurs qui influencent grandement le signal LIBS et qui peuvent donc dégrader les performances analytiques de la technique [104]. Dans le cadre de la technique de LA-ICP-MS où l'échantillon solide est vaporisé par ablation laser, il a été recommandé un broyage dans un mortier puis un tamisage à moins de 1 µm pour avoir une pastille homogène [105]. De plus, l'ajout d'un liant a aussi été recommandé dans le cas où la pastille se dégrade rapidement lors de l'ablation laser [106]. La méthode de préparation utilisée pour cette étude est beaucoup plus simple et rapide. Il n'y a ni broyage dans un mortier ni ajout de liant dans un souci de mesures rapides sur site. En revanche, les échantillons ont été tamisés à 2 mm pour éliminer les brindilles et les cailloux mais sans pour autant privilégier une taille de grain en particulier. Notons que c'est à ce stade de la préparation qu'ont été effectuées les mesures par XRF. Puis, quelques grammes de l'échantillon sont séchés au four micro-ondes pour finalement fabriquer une pastille de 300 mg à l'aide d'une presse manuelle appliquant une pression de 8000 psi pendant une minute. Remarquons aussi qu'une bonne homogénéisation de la poudre de sol avant la fabrication de la pastille permet de considérer que l'analyse LIBS, bien que surfacique par principe, donne

dans ce cas une information représentative du volume. Grâce à ces précautions expérimentales, la comparaison des mesures LIBS et ICP-AES a du sens.

La Figure 2-5 montre une photo des instruments utilisés à bord de la camionnette pour la préparation sur site des échantillons de sol.

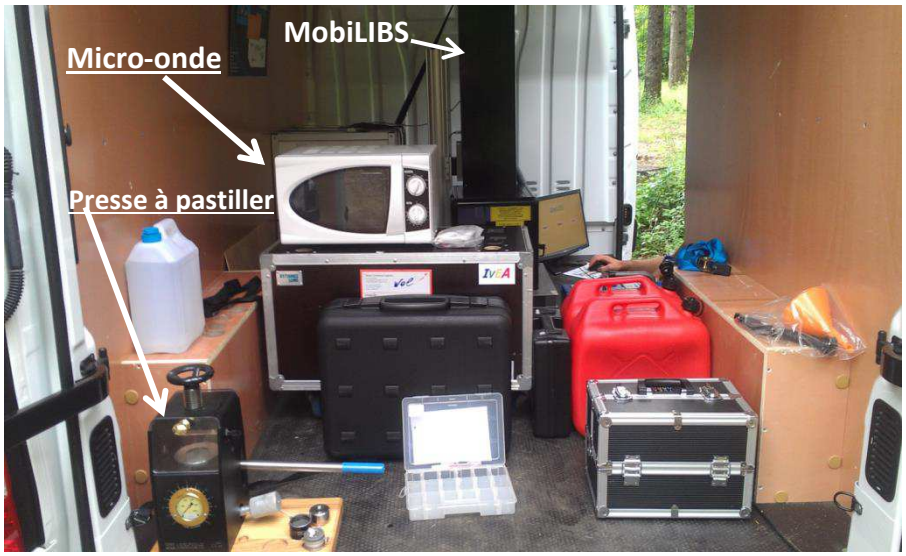


Figure 2-5 Appareillages de préparation des échantillons utilisés lors des campagnes de mesures LIBS sur site

2.3 **Résultats des analyses LIBS**

2.3.1 Spectres LIBS

Les spectres LIBS sont systématiquement corrigés du fond continu (offset) présent sur les spectres bruts. Un spectre LIBS typique de sol est présenté sur la Figure 2-6. Chaque raie peut être reliée à un élément chimique (atome ou ion simplement ionisé) en faisant référence à une base de données comme celle du NIST [107]. Dans cet échantillon de sol en particulier, nous avons détecté des éléments majeurs (Al, Ca, Si, K, Na, Mg, Ti, Fe, Mn) et des éléments mineurs (Pb, Cu, Ni, Zn, V, Cr, Ag, Sr, Sn). Plus généralement, les éléments détectés reflètent les caractéristiques du site analysé.

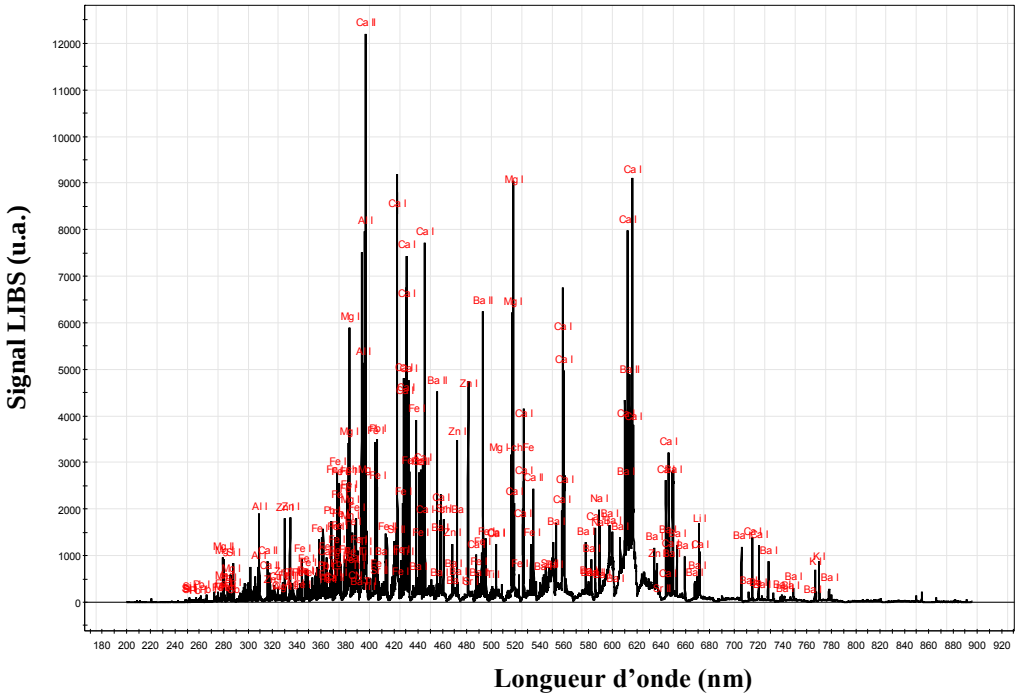


Figure 2-6 Spectre LIBS typique d'un échantillon de sol enregistré sur site

La Figure 2-7 représente un zoom du spectre LIBS donné sur la Figure 2-6 dans la bande 396-434 nm. On remarque les raies intenses de Ca I et plus particulièrement la raie à 422 nm qui présente un profil très élargi avec même un trou en son centre. Ceci est la manifestation spectaculaire du phénomène d'auto-absorption lié au fait que cette raie de Ca I est résonante, c'est-à-dire avec le niveau fondamental comme niveau inférieur de la transition associée. On remarque aussi la présence d'une raie de l'aluminium, de plusieurs raies du fer, ainsi que des raies du plomb, du strontium, du baryum et du titane.

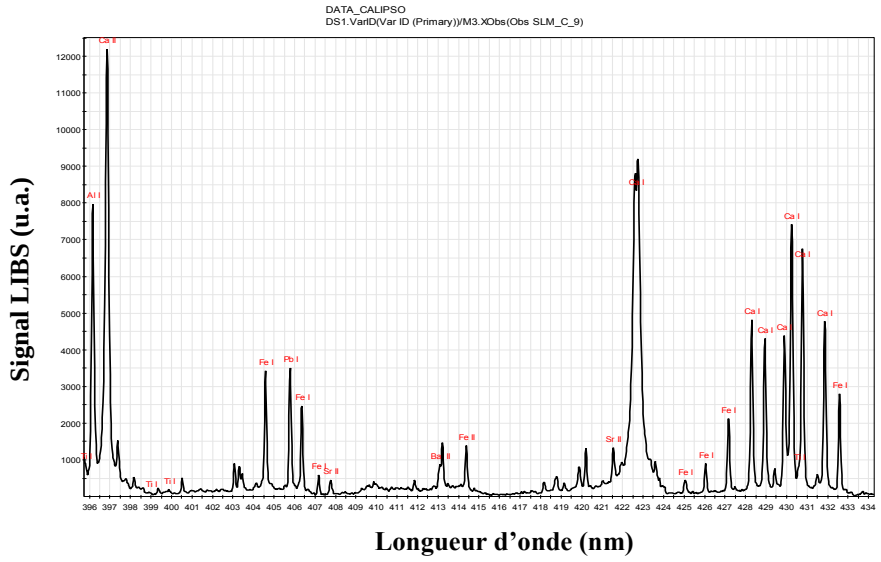


Figure 2-7 Zoom du spectre de sol de la Figure 2-6 dans la fenêtre 396-434 nm

L'étalonnage du spectromètre est effectué à l'aide d'une lampe spectrale Hg-Ar [108]. Cependant, d'une campagne de mesure à l'autre, il peut se produire un léger décalage en longueur d'onde. Evidemment, lors d'une analyse statistique de nombreux spectres, un décalage peut avoir des conséquences graves et conduire à de mauvaises conclusions. On veille donc à recaler les différents spectres sur une échelle unique de longueurs d'onde avant de procéder à tout traitement statistique. Les spectres ainsi recalés sont rangés sous forme d'une matrice X où l'on aura sur chaque ligne le spectre d'un échantillon. Ainsi, chaque valeur de la matrice est en fait l'intensité du spectre LIBS à une longueur d'onde donnée et pour un échantillon donné. Notons aussi que les échantillons de sols ont par ailleurs été analysés par ICP-AES en laboratoire après chaque campagne sur site. Ceci a permis de connaître les valeurs des concentrations des différents éléments chimiques, considérées comme valeurs de référence pour cette étude. Ainsi, si par exemple, on souhaite analyser par LIBS la concentration en plomb, on fabriquera un modèle avec en entrée la matrice X déjà décrite et en sortie un vecteur colonne Y contenant les valeurs des concentrations en plomb de chaque échantillon.

2.3.2 Description graphique des échantillons par ACP

L'ACP permet de visualiser des données de façon non-supervisée. Nous avons donc utilisé cet outil pour comprendre les principales caractéristiques de données LIBS collectées lors des campagnes de mesure sur site. Puis, dans un second temps, afin de pouvoir interpréter plus finement les résultats, nous avons appliqué la même analyse aux données de référence provenant des mesures ICP-AES de laboratoire.

2.3.2.1 ACP appliquée aux données LIBS

Les données LIBS ont été collectées au cours de plusieurs campagnes :

- 2 campagnes sur le site Saint-Laurent-le-Minier, notées SLM_C1 et SLM_C2
- 1 campagne sur le site Metaleurope, notée ME
- 1 campagne sur le site Saint Sébastien, notée SEB

Ces campagnes ont permis de collecter 181 spectres, donc 181 lignes dans la matrice X. Le nombre de variables est donné par le nombre de points sur un spectre LIBS. Avec le système déployé sur site, ce nombre était 23988, qui correspond au nombre de colonnes de la matrice X. Rappelons que les spectres sont tous corrigés d'un fond continu. De plus, un prétraitement est appliqué avant l'analyse. Il consiste à centrer toutes les variables, ce qui revient pour une colonne donnée à calculer la valeur moyenne pour tous les échantillons et à soustraire cette valeur moyenne à toutes les valeurs de la colonne. On injecte alors la matrice X ainsi préparée dans un calcul d'ACP (logiciel SIMCA-P+) et on obtient le résultat présenté dans le Tableau 2-1. Dans ce tableau, A indique le numéro de la composante principale sachant que celles-ci sont rangées de sorte que la composante A=1 contient toujours la plus grande partie de l'information permettant d'expliquer la covariance de la matrice X. La valeur correspondante de R2X n'est autre que la première valeur propre de la matrice de covariance de X normalisée à 1. Ainsi, dans le cas présent, la première valeur propre étant de 0,522 on

peut dire que la première composante permet de décrire à elle seule 52,2% de la variance totale du jeu de données. On comprend alors que puisque la deuxième composante permet d'expliquer 22,7% de la variance, on peut donner le cumul des deux premières composantes en ajoutant simplement les deux premières valeurs propres. C'est ce qui est indiqué dans la colonne R2X(cum). Dans le cas présent, 6 composantes principales ont permis d'expliquer 96% de la variance du jeu de données. On notera cependant que les trois premières composantes suffisent déjà à expliquer 90% de la variance, ce qui offre de belles perspectives de compression des données et qui laisse penser qu'on pourra observer la quasi-totalité de l'information dans un espace à trois dimensions.

A	**R2X**	**R2X (cum)**
1	0,522	0,522
2	0,227	0,749
3	0,158	0,907
4	0,03	0,937
5	0,0147	0,951
6	0,00863	0,96

Tableau 2-1 Résultat d'ACP sur les données LIBS (181 spectres de 13988 points ; voir texte) ; (A) l'indice de la composante; R2X : fraction de la variation de X expliquée par composant; R2X (cum)- somme des valeurs de R2X jusqu'à la composante étudiée.

La Figure 2-8 présente les scores des différentes observations (les 181 échantillons) dans le plan des deux premières composantes principales (1 et 2). On remarque que les échantillons SEB (rose) sont les moins dispersés et tous groupés sur la gauche du graphique. On note ensuite que les échantillons de la campagne ME (bleu) sont très dispersés le long de l'axe 1 mais peu le long de l'axe 2. Enfin, les échantillons de la campagne SLM sont quant à eux dispersés selon les deux axes (1 et 2). On peut remarquer enfin que certains points se trouvent en dehors de l'ellipse qui décrit un critère de confiance à 95% donnant généralement la limite d'un modèle. Les points extérieurs sont généralement considérés comme aberrants et finalement exclus de l'analyse. Cependant ici, nous avons affaire à des sols qui sont des échantillons très hétérogènes et pour lesquels des concentrations anormalement élevées pourraient conduire à trouver des points à l'extérieur de l'ellipse. On préfère donc ici ne pas exclure ces points qui peuvent potentiellement fournir des informations utiles sur le site.

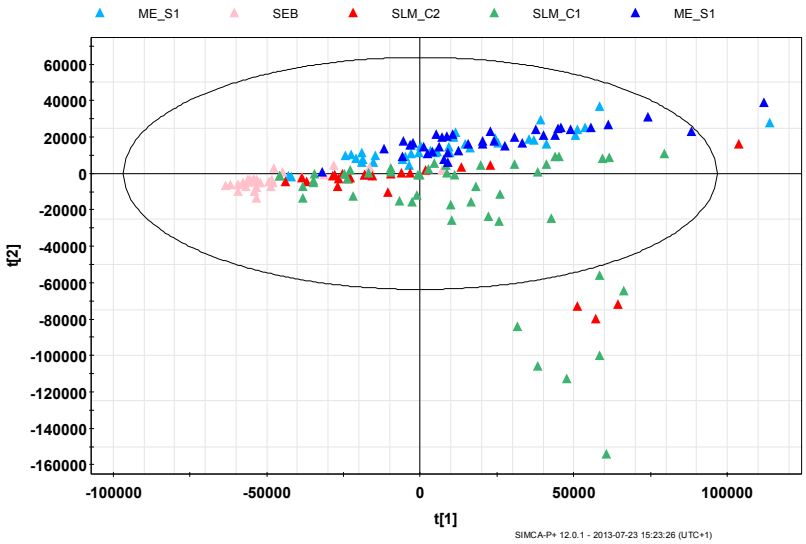


Figure 2-8 Scores ACP des deux premières composantes principales (1,2) obtenus sur les données LIBS des 181 échantillons de sol provenant de différents sites.

La Figure 2-9 présente la projection des vecteurs propres (loadings) dans le plan des deux premières composantes principales. Les raies intenses sont naturellement les plus éloignées de l'origine des axes. C'est le cas des raies de Ca (rouge), Al (bleu foncé) et Ba (vert) suivant l'axe 1, par contre suivant l'axe 2 seules les raies d'Al et de Ba restent éloignées alors que les raies de Ca sont plus proches du zéro. En analysant simultanément les Figure 2-8 et Figure 2-9, on peut en déduire que les échantillons représentés les plus à droite sur le graphe des scores (2-8) sont associés aux plus fortes concentrations en Ca et/ou Al alors que ceux qui sont représentés en bas du graphe sont associés aux plus fortes concentrations en Ba. A l'inverse, on peut conclure que les échantillons SEB ont les plus faibles concentrations en Ba, Al et Ca. Par ailleurs, on sait que Al, Ca et Ba sont trois éléments majeurs d'après les données de référence ICP-AES. Ils sont représentatifs de trois matrices différentes qualifiées d'aluminosilicate, calcaire, et minerai. Afin de mieux distinguer les différences entre les échantillons provenant de différents sites et associés à différentes matrices, nous avons observé le rôle joué par la composante 3 du traitement d'ACP.

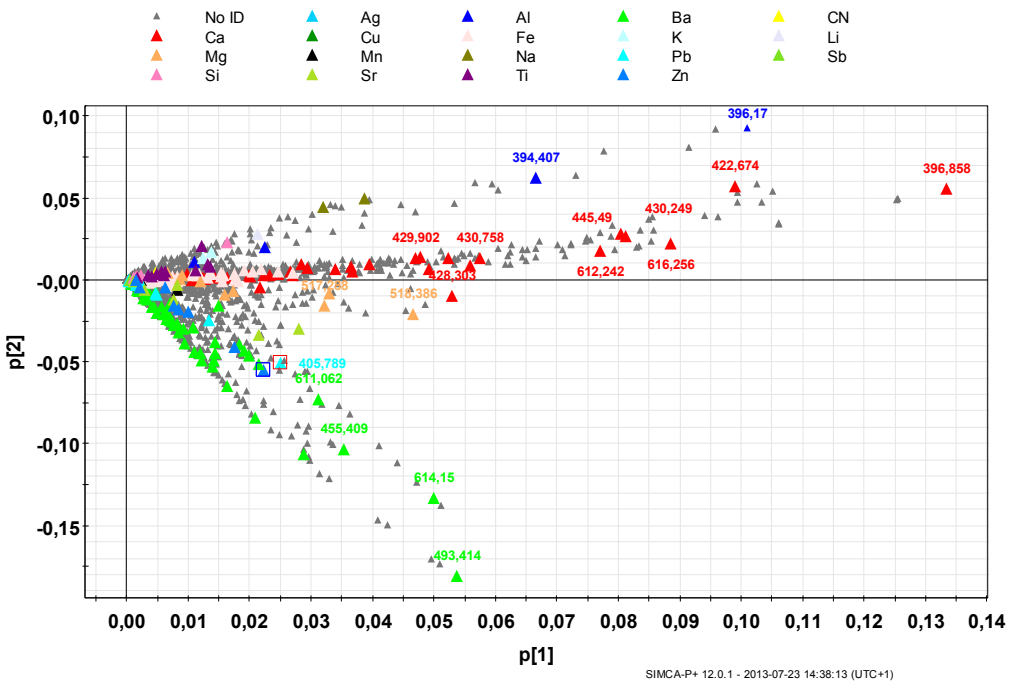


Figure 2-9 Vecteurs propres (loadings) ACP des deux premières composantes principales (1,2) obtenus sur les données LIBS des 181 échantillons de sol provenant de différents sites.

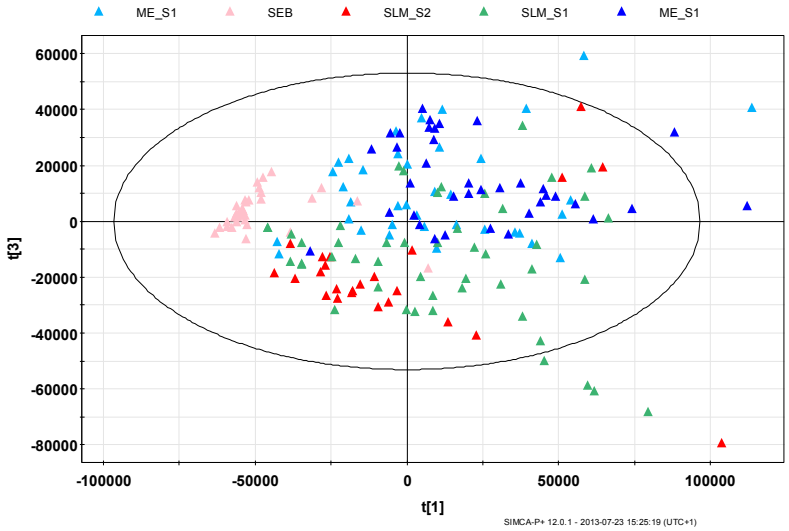


Figure 2-10. Scores ACP des composantes principales (1,3) obtenus sur les données LIBS des 181 échantillons de sol provenant de différents sites.

La Figure 2-10 décrit les scores dans le plan (1,3). On remarque que la dispersion des échantillons dans la direction verticale (axe 3) est plus prononcée que suivant l'axe 2. D'après la Figure 2-11 qui décrit les loadings dans le plan (1,3), on a la confirmation que l'axe 3 présente une meilleure séparation d'Al et Ca que l'axe 2, même si cela ne permet pas de trier facilement les échantillons sur la Figure 2-10.

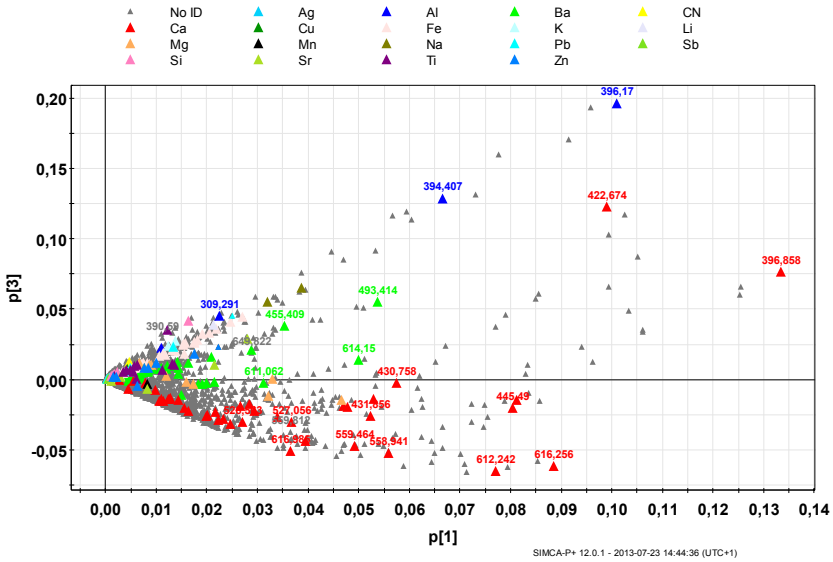


Figure 2-11 Vecteurs propres (loadings) ACP dans le plan (1,3) obtenus sur les données LIBS des 181 échantillons de sol provenant de différents sites.

La Figure 2-12 et la Figure 2-13 montrent enfin ces projections dans le plan (2,3). Le calcul des valeurs propres montre que ce plan permet de représenter 39% de la variance du jeu de données. La Figure 2-13 révèle la présence de trois pôles : un pôle alumino-silicaté dans la direction à droite et vers le haut, un pôle carbonaté riche en Ca et Mg dans la direction à droite vers le bas et enfin un pôle minerais vers la gauche dans lequel on trouve de fortes concentrations de Ba et de Zn. Ainsi, d'après la Figure 2-12, on peut dire que les échantillons ME se répartissent vers le pôle alumino-silicaté, tandis que bon nombre d'échantillons de SLM se répartissent en direction du pôle carbonaté avec cependant quelques-uns en direction du pôle minerai. Enfin, on remarque que la projection (2,3) n'apporte aucune information sur les échantillons SEB car tous les échantillons sont placés à proximité de l'origine. Les données SEB ont été cependant bien séparées à l'aide de la composante 1, probablement à cause des faibles concentrations en Ca.

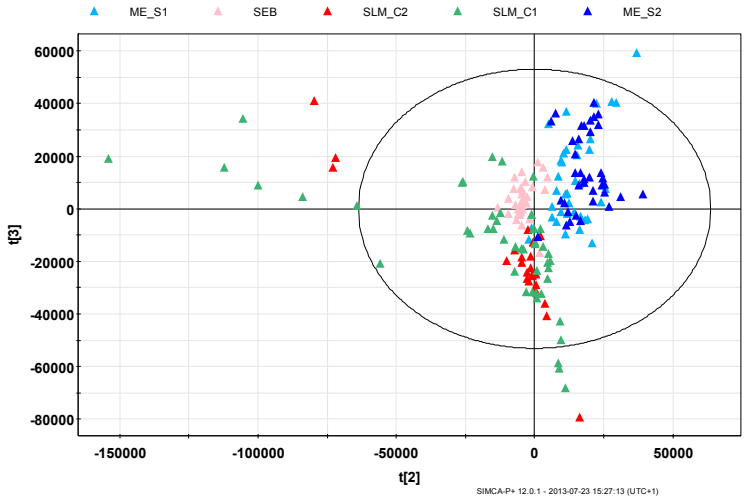


Figure 2-12 Scores ACP composantes principales (2,3) obtenus sur les données LIBS des 181 échantillons de sol provenant de différents sites.

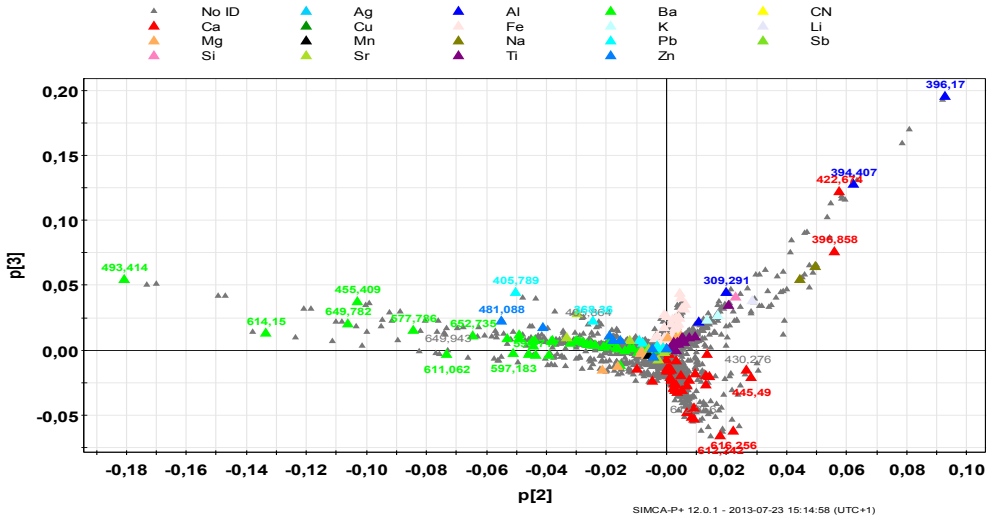


Figure 2-13 Vecteurs propres (loadings) ACP dans le plan (2,3) obtenus sur les données LIBS des 181 échantillons de sol provenant de différents sites.

Comme attendu, l'ACP peut servir de méthode de classification non-supervisée, même si on remarque ici qu'il est difficile de séparer les échantillons de sols en classes bien distinctes et que les points se répartissent de manière continue. Par ailleurs, la Figure 2-13 montre clairement que la raie de Ca I à 422.674 nm ne se comporte pas du tout comme les autres raies du calcium (rouge). Ceci n'est pas très surprenant car nous avons montré sur la Figure 2-7 qu'il s'agissait d'une raie profondément perturbée par le phénomène d'auto-absorption. Il est donc intéressant de noter ici que pour un élément donné, les raies qui présentent des anomalies peuvent être facilement identifiées par ACP. Ainsi, la raie du Ca II à 396.358 nm est également identifiée comme non-corrélée aux autres raies du calcium alors qu'elle est corrélée à une raie intense voisine du Al à 396.17 nm. On met ainsi à jour un problème d'interférence spectrale. Ceci nous amène à conclure d'une part que les méthodes de

chimiométrie permettent de prendre en compte des interférences spectrales et d'autre part que l'ACP peut être un outil extrêmement précieux pour mener des analyses LIBS et pour avoir une meilleure connaissance de la nature des raies spectrales considérées pour une future analyse quantitative. Par ailleurs, l'ACP nous a aussi permis d'observer que lorsqu'on mène deux campagnes de mesures LIBS sur un même site géologique, on retrouve des caractéristiques très semblables entre les différentes données. Dans le cadre de cette étude, le site SLM a ainsi été analysé à deux reprises et il y a un très bon recouvrement entre les deux séries de mesures. Notons cependant que le nombre de mesures n'est pas suffisant ici pour trancher cette question définitivement mais que nous avons démontré que les caractéristiques géochimiques d'un site étaient finalement bien contenues dans les données LIBS.

2.3.2.2 ACP appliquée aux données ICP-AES de référence

Afin d'approfondir notre compréhension des corrélations qui ont été observées en appliquant l'ACP aux données LIBS, nous avons analysé également par ACP les données de références obtenues à partir de mesures ICP-AES. Dans ce cas, la matrice X est constituée de 181 lignes qui sont les 181 spectres LIBS enregistrés et de 12 colonnes qui sont les concentrations mesurées par ICP-AES de : Al, Si, Ti, K, Mg, Ca, Ba, Zn, Pb, Ag, Cu, Fe. Le Tableau 2-2 fournit les résultats relatifs aux deux premières composantes et révèle que le plan (1,2) permet d'expliquer 64,2% de la variance des données ICP-AES.

A	R2X	R2X(cum)
1	0,412	0,412
2	0,23	0,642

Tableau 2-2 Résultat d'ACP sur les données ICP-AES; (A) l'indice de la composante; R2X : fraction de la variation de X expliquée par composant; R2X (cum)- somme des valeurs de R2X jusqu'à la composante étudiée.

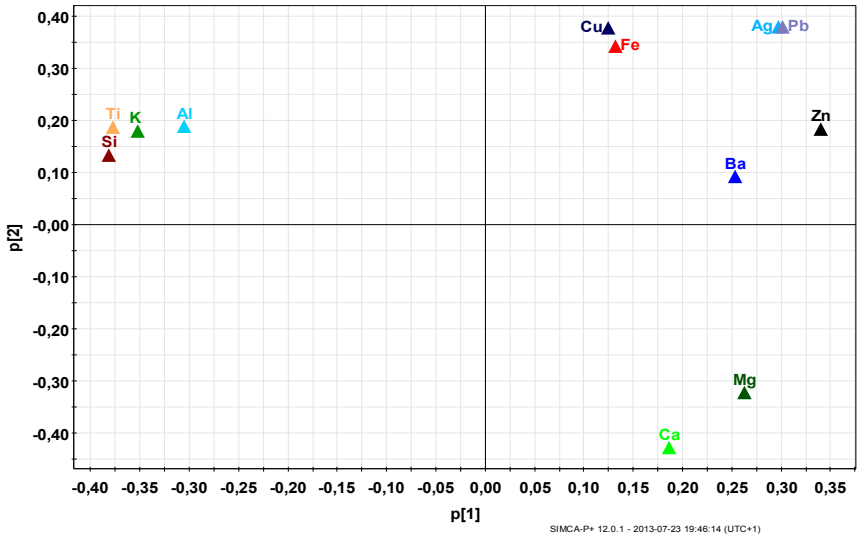


Figure 2-14 Vecteurs propres (loadings) ACP des deux premières composantes principales (1,2) obtenus sur les données ICP-AES des 181 échantillons de sol provenant de différents sites.

La Figure 2-14 représente les loadings dans le plan (1,2) et révèle qu'il existe une première corrélation entre les concentrations de Al, Si, Ti, et K (à gauche), puis une seconde corrélation entre les concentrations de Mg, et Ca (en bas), et enfin une troisième corrélation entre les concentrations de Ba, Zn, Pb, Ag, Cu, Fe. Cette étude ACP sur des données ICP-AES confirme bien les résultats obtenus avec des données LIBS, à savoir le rôle particulier de trois pôles :

- Sols alumino-silicates riches en Al et Si (avec une corrélation avec K et Ti)
- Sols calcaires riches en Ca et Mg
- Sols de type minerai riches en Ba, Zn, Pb, Ag.

Il faudra donc prendre en compte l'existence de ces trois pôles si cela s'avère nécessaire pour mener à bien des analyses quantitatives.

2.3.3 Analyse quantitative par ANN

Nous avons démontré que l'analyse LIBS quantitative de sols était impossible en utilisant une méthode univariée, ceci à cause d'importants effets de matrice[35]. D'autres ont également conclu que des approches multivariées étaient nécessaires pour mener à bien des analyses LIBS quantitatives pour les sols [109]. Dans ce contexte, la PLS étant une méthode multi-linéaire, elle ne permet pas de prendre en compte des comportements non-linéaires et elle présente donc un intérêt limité lorsqu'elle est appliquée à l'analyse LIBS quantitative d'un métal dans des échantillons de sols [110]. A l'inverse, la méthode ANN a montré lors de travaux précédents une certaine efficacité pour traiter des données issues d'échantillons géologiques comme les sols et les roches et présentant des comportements non-linéaires [63, 111-113]. Nous avons donc opté pour une analyse mettant en œuvre des réseaux de neurones artificiels (ANN) qui permettent a priori de prendre en compte les non-linéarités et par conséquent de prendre en compte le problème des effets de matrice.

2.3.3.1 Choix des données d'entrée

Comme nous l'avons vu dans le chapitre 1, le choix des données d'entrée influence grandement les performances d'un modèle de quantification. Lorsqu'on observe un spectre LIBS dans le cas particulier d'un sol, on constate que le calcium donne plus de 40 raies spectrales entre 250 et 800 nm et que le fer en donne encore plus. Il est donc difficile de se décider sur le choix des raies qui serviront à l'analyse. Nous avons donc établi notre propre recette sur la base de considérations qui nous semblaient les plus pertinentes. Cette recette a été décrite dans un article qui a été publié dans Spectrochimica Acta Part B en 2013 [35] et que nous présentons ci-dessous en 4 étapes.

- Etape 1 : pour chaque élément, établir la liste des raies persistantes d'après la base de données atomiques du NIST
- Etape 2 : Eliminer de la liste les raies d'émission pour lesquelles le niveau inférieur de la transition est le niveau fondamental afin d'éviter l'auto-absorption
- Etape 3 : Eliminer de la liste les raies qui ont un faible contraste, c'est-à-dire un faible rapport signal/fond. On veille ainsi à supprimer des longueurs d'onde qui ne

contiennent pas d'information sur l'élément et ce en particulier pour les éléments en faible concentration.

- Etape 4 : Cette recette n'est pas figée mais elle doit être considérée comme un outil d'aide à la sélection des données. En effet, pour un élément mineur par exemple, il est possible que l'on ne détecte que les raies résonnantes et dans ce cas on les gardera pour l'analyse en sachant que les risques d'auto-absorption sont faibles lorsqu'il s'agit d'un élément mineur.

Cette méthode a été appliquée pour l'analyse des sols. Elle a permis de choisir non seulement les raies de l'analyte mais aussi celles des éléments qui caractérisent la matrice et qui ont été déterminés par ACP dans la partie précédente. Les raies sélectionnées d'après cette recette sont présentées dans le Tableau 2-3.

Eléments et longueurs d'onde (nm) des raies sélectionnées	
Al	308.215, **309.271**, 394.400, 396.152
Ba	**652.731**, 659.532, 669.384, 705.994, 728.029
Ca	442.544, 558.875, 610.272, **612.221**, 616.217,643.907, 646.256
Cu	324.754, 327.396
Fe	278.81, 358.119, 373.486, 374.556, 374.826, 374.948, 375.823,**382.042**, 388.628, 404.581, 438.354
Ti	365.349, 375.285, 395.633, 395.82, **399.863**, 498.173, 499.106
Pb	261.418, 283.305, 363.957, 368.346, 405.781

Tableau 2-3 Raies spectrales sélectionnées pour fournir les données d'entrée de l'ANN.

En exploitant l'ensemble des raies présentées dans le Tableau 2-3, nous avons observé un risque de sur-apprentissage du modèle ANN sans doute dû à un nombre trop important de données d'entrée. Par conséquent, nous avons décidé de réduire de façon significative le nombre de données d'entrée et de ne garder que les raies spectrales en gras dans le Tableau 2-3, à savoir une seule raie par élément de la matrice du sol soit la raie à 309,271 nm de Al, la raie à 652,731 nm du Ba, celle à 612,221 nm du Ca, celle à 382,042 nm pour Fe et celle à 399,863 nm pour Ti. En revanche, dans le cadre de l'analyse quantitative du plomb, l'élément que nous avons choisi d'analyser en détails pour montrer les potentialités de la technique, nous avons conservé les 5 raies spectrales de cet élément données dans le Tableau 2-3.

Les résultats présentés ici concernent le site SLM seul car il présente une grande diversité des matrices d'un point de mesure à l'autre (cf. Figure 2-8) et que par ailleurs, deux campagnes de mesures ont été effectuées sur ce site à plusieurs mois d'intervalle. On notera SLM_C1 et SLM_C2 les deux lots d'échantillons résultants de ces deux campagnes. Il va sans dire que l'analyse univariée ne peut pas convenir dans ces conditions et qu'une approche multivariée est nécessaire. En revanche, la PLS n'a pas permis d'atteindre des résultats satisfaisants probablement à cause du caractère non-linéaire des relations entre les données d'entrée X et les concentrations Y. Ceci nous a conduit à exploiter les réseaux de neurones artificiels et nous présentons ici les résultats pour la quantification de différents éléments chimiques majeurs comme le calcium, l'aluminium ou le fer ainsi que pour le cuivre qui est un élément

mineur et le plomb dont la concentration s'étale sur une très grande gamme de concentrations d'un échantillon à l'autre.

2.3.3.2 Quantification des éléments majeurs des échantillons de sol par ANN

2.3.3.2.1 Analyse quantitative du calcium

Le calcium est un élément majeur des sols ramassés sur le site de Saint-Laurent-Le-Minier, et il témoigne que l'on a affaire ici à une matrice calcaire. Avant tout, nous avons sélectionné les données d'entrée de l'ANN. Dans ce cas, il s'agit des 7 raies du Ca indiquées dans le Tableau 3-4 à savoir 442.544, 558.875, 610.272, 612.221, 616.217,643.907 et 646.256 nm, et les 4 raies qui caractérisent le reste de la matrice : Fe (382.042 nm), Ti (399.863 nm), Ba (652.731 nm) et Al (309.271 nm). On utilise donc 11 entrées en tout, 7 relatives au calcium et nécessairement corrélés entre elles, et les 4 autres qui sont associées aux pôles dont nous avons discuté précédemment : les raies de Al et Ti corrélées entre elles sont relatives au pôle alumino-silicaté tandis que celles du Ba et du Fe représentent le pôle minerai. La deuxième étape consiste à répartir les échantillons en différents lots. Ainsi, sur un total de 71 spectres en tout pour les deux campagnes SLM_C1 et SLM-C2, les spectres sont répartis de la façon suivante : 44 spectres dans le lot d'apprentissage, 17 dans le lot de validation, 10 dans le lot de test. Rappelons ici que le lot d'apprentissage permet de fabriquer les modèles, le lot de validation permet ensuite de choisir le meilleur modèle et finalement le lot de test permet de tester le meilleur modèle a posteriori. Enfin, chaque modèle a été recalculé 5 fois pour les mêmes paramètres avec simplement des poids initiaux différents choisis aléatoirement afin de tester la robustesse.

Afin de trouver le meilleur modèle ANN, nous avons fait varier les quatre paramètres suivants : le nombre de neurones dans la couche cachée, la vitesse d'apprentissage, le terme de mémoire et le nombre d'itérations. Nous avons choisi de calculer les deux types d'erreurs présentées dans le chapitre 1 comme critère de choix, à savoir l'erreur relative moyenne et l'erreur quadratique moyenne pour le lot de calibration (ERC et RMSEC) et pour le lot de validation (ERV, RMSEV). Ces erreurs sont calculées par validation croisée. Rappelons que l'on cherche à obtenir un compromis donnant la valeur minimale d'erreur pour chaque lot de données séparément. Le lot de test n'intervient pas du tout dans ce processus, il sera juste utilisé à la fin et seulement pour le modèle ANN qui aura été optimisé à l'aide des deux autres lots de données. Les paramètres ajustables sont testés dans les gammes suivantes :

- le nombre de neurones dans la couche cachée : entre 1 et 10 par pas de 1
- la vitesse d'apprentissage : entre 0.05 et 0.5
- le terme de mémoire : entre 0.01 et 0.5
- le nombre d'itérations : entre 1000 et 50000 par pas de 1000

Il n'est pas très intéressant de montrer l'ensemble des résultats de cette phase d'optimisation, aussi nous avons choisi d'illustrer le processus d'optimisation en ne montrant, à titre d'exemple, que la recherche du nombre optimal d'itérations dans le cas où le nombre de

neurones de la couche cachée est choisi égal à 5, la vitesse d'apprentissage est choisie égale à 0.2 et le terme de mémoire est fixé à 0.05. La Figure 2-15 donne les résultats des ERC et ERV en fonction du nombre d'itérations tandis que la Figure 2-16 donne les valeurs de RMSEC et RMSEV dans les mêmes conditions. On recherche sur ces courbes le nombre d'itérations optimal qui permet de minimiser les signaux simultanément et pour cela on répète 5 fois chaque calcul avec des valeurs aléatoires de départ différentes. On constate tout d'abord que pour le lot de calibration, ERC et RMSEC ne font que décroître au fur et à mesure que le nombre d'itérations augmente pour se stabiliser vers 45000-50000 itérations. En revanche, on remarque que ERV atteint un minimum pour 8000 itérations et remonte ensuite. On en déduit que le nombre optimum d'itération est alors de 8000 car après, ERC diminue mais ERV augmente, ce qui signifie que le modèle est passé en sur-apprentissage. Ce choix est beaucoup moins facile à faire avec les valeurs RMSEC et RMSEV même si on constate que c'est autour de 10000 itérations que les deux courbes se croisent.

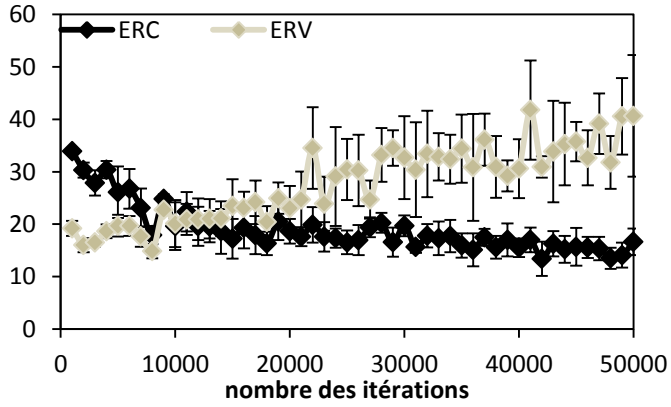


Figure 2-15 ERC et ERV (en %) pour des modèles ANN construits pour quantifier le calcium sur la base de 5 neurones dans la couche cachée, 0.2 de vitesse d'apprentissage et 0.05 de terme de mémoire. Les barres d'erreur représentent l'écart-type des 5 répétitions.

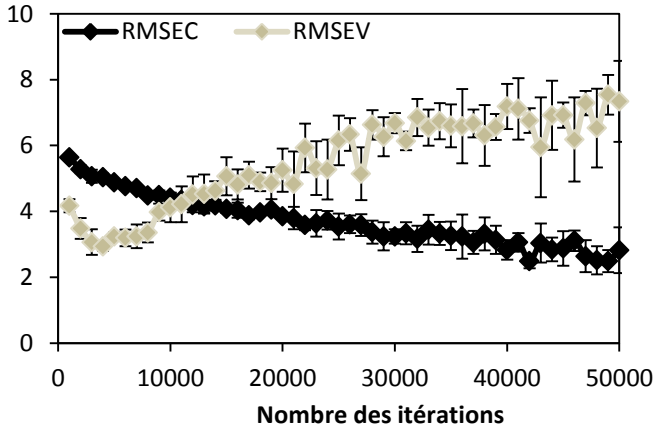


Figure 2-16 RMSEC et RMSEV (en g pour 100 g) pour des modèles ANN construits pour quantifier le calcium sur la base de 5 neurones dans la couche cachée, 0.2 de vitesse d'apprentissage et 0.05 de terme de mémoire. Les barres d'erreur représentent l'écart-type des 5 répétitions.

On refait ces calculs pour les autres valeurs des paramètres ajustables et on trouve finalement que le meilleurs compromise correspond à:

- Nombre des neurones dans la couche cachée : 5
- Vitesse d'apprentissage : 0.1
- Terme de mémoire : 0.05
- Nombre des itérations : 8000

Le Tableau 2-4 donne les performances générales du meilleur ANN pour quantifier le calcium telles qu'elles ont été définies dans le chapitre 2. Ici, le lot de test est particulièrement utile pour savoir si le caractère prédictif de l'ANN retenu est généralisable. On remarque ici que les erreurs relatives moyennes sont inférieures à 20% pour chacun des trois lots de données, ce qui est tout à fait satisfaisant pour une analyse réalisée sur site.

	Lot d'apprentissage	Lot de validation	Lot de test
R^2	0,84	0,89	0,83
Q^2	0,82	0,87	0,82
ER (%)	17	14	13
RMSE (g/100g)	4	3	4

Tableau 2-4 Performances du meilleur ANN pour l'analyse quantitative du Ca

Il reste cependant à vérifier que les performances obtenues pour ce modèle ANN sont bien significatives et cela peut être fait en appliquant la technique de Y-randomization décrite au chapitre 1. En pratique, on conserve toujours les mêmes données d'entrée et on fixe les paramètres de l'ANN. Puis on permute de façon aléatoire les valeurs des concentrations en sortie lors de la phase d'étalonnage et on calcule les différents facteurs de mérite, à savoir R^2, Q^2 et RMSE. On répète cette opération 25 fois afin de disposer d'un résultat moyen. Les facteurs de mérite obtenus après cette procédure de vérification sont donnés dans le Tableau 2-5. Ils montrent clairement que le modèle d'ANN que nous avions retenu pour quantifier le Ca n'avait pas été optimisé par chance mais sur la base d'une véritable optimisation des poids appliqués aux variables d'entrée.

Y-Random	Moyenne	écart-type
R^2c	0,06	0,09
R^2v	0,10	0,14
R^2t	0,30	0,15
Q^2r	0,53	0,15
Q^2v	-1,09	0,83
Q^2t	-2,51	3,85
RMSEC (g/100g)	7	1
RMSEV (g/100g)	12	3
RMSET (g/100g)	12	3
ERC (%)	46	13
ERV (%)	98	53
ERT (%)	91	69

Tableau 2-5 Moyenne et écart-type des facteurs de mérites calculés pour 25 permutations aléatoires des données de sortie (ici, la concentration de Ca) pour le meilleur ANN optimisé pour une analyse quantitative de Ca.

Le test Y-randomization étant passé avec succès, on peut enfin conclure que l'ANN optimisé ici est capable de prédire les concentrations en calcium des sols provenant du site SLM avec une erreur relative moyenne inférieure à 20%. Le seuil de 20% correspond à la valeur jugée acceptable de l'erreur relative pour des mesures sur site en accord avec la proposition de Essignton et al. [110]. On peut cependant s'intéresser aux performances du modèle ANN au cas par cas pour différentes gammes de concentrations du calcium. La Figure 2-17 donne les concentrations de Ca (en %) obtenues par ICP-AES (valeurs de référence, bleu) et par LIBS puis traitement ANN (rouge) pour les échantillons du lot de validation présentés par concentrations croissantes. Pour faciliter la lecture de ce diagramme, on a reporté sur les valeurs ICP-AES des barres d'erreur correspondant à une erreur relative de 20% pour chaque concentration. Ainsi, si la concentration prédite par LIBS/ANN se trouve à l'intérieur de cette barre d'erreur, on peut considérer que le résultat est satisfaisant. Notons aussi que les résultats LIBS/ANN présentent eux aussi des barres d'erreurs qui sont déterminées par la valeur de l'écart-type sur 5 répétitions du calcul d'ANN à partir de poids initiaux aléatoires. On remarque que pour une grande majorité d'échantillons, la concentration de Ca est correctement prédite avec cependant une exception notable pour l'échantillon C2_11, largement sous-évalué en LIBS par rapport à la valeur de référence. La Figure 2-18 décrit le même type de résultat mais pour les échantillons du lot de test. Là encore, on constate que pour la grande majorité des échantillons, la concentration prédite est correcte, excepté pour l'échantillon C2_1 qui est sous-évalué en LIBS par rapport à la valeur de référence. En conclusion, sur les 27 échantillons analysés, seuls deux ont présenté des anomalies graves lors de l'analyse LIBS par ANN. Notons que ce résultat est assez remarquable pour des analyses sur site d'échantillons aussi complexes que des sols.

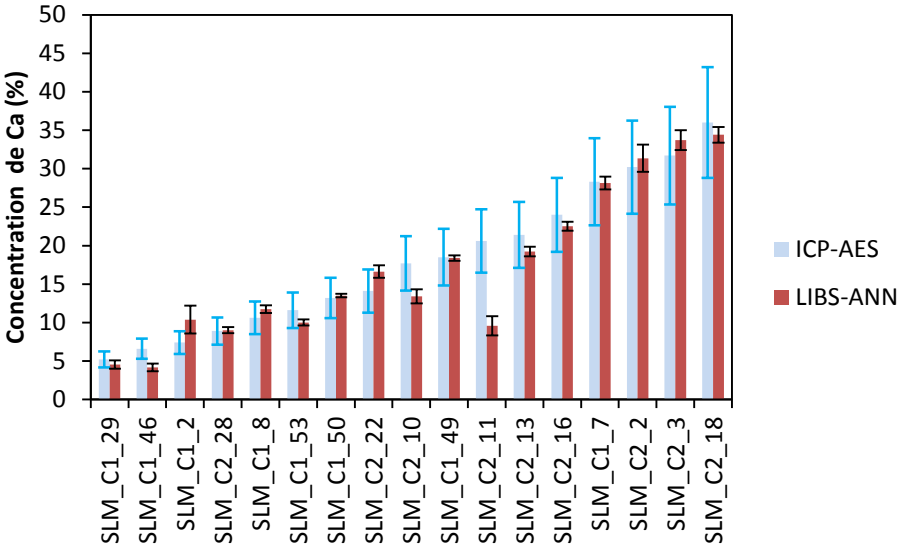


Figure 2-17 Comparaison entre la concentration de référence et la concentration prédite en LIBS-ANN pour les échantillons du lot de validation. Les barres bleues représentent les écarts de 20% autour de la concentration de référence, et les barres noires représentent l'écart-type sur 5 répétitions des concentrations prédites par ANN.

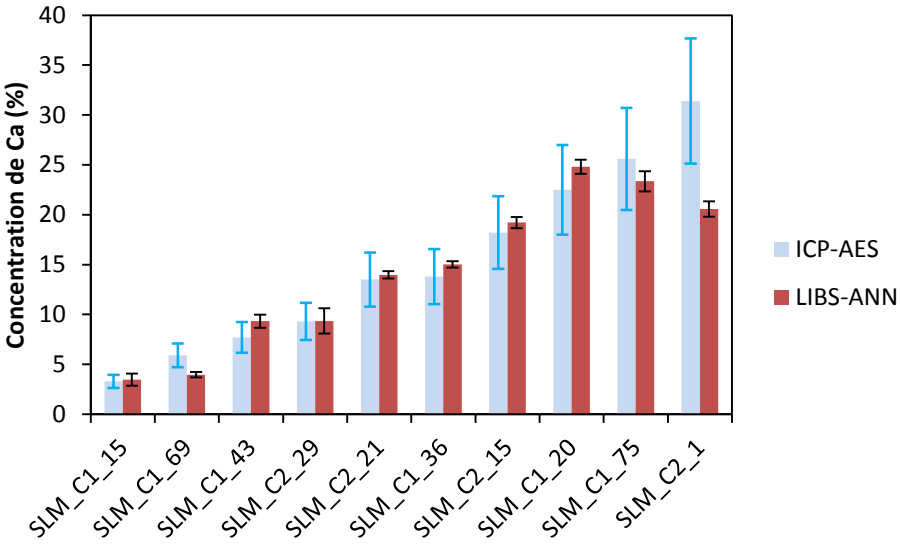


Figure 2-18 Comparaison entre la concentration de référence et la concentration prédite en LIBS-ANN pour les échantillons du lot de test. Les barres bleues représentent les écarts de 20% autour de la concentration de référence, et les barres noires représentent l'écart-type sur 5 répétitions des concentrations prédites par ANN

2.3.3.2.2 Analyse quantitative de l'aluminium

L'aluminium peut facilement exister en tant qu'élément majeur (g/100g) dans le sol car il est l'un des principaux constituants des argiles, toujours associé à du SiO_4. En revanche, on ne le trouve à l'état pur ou même à l'état d'oxyde ou d'hydroxyde que dans les cas de dégradation de l'argile par attaque acide par exemple, et il se comporte alors comme un véritable poison

pour les plantes. On comprend ici que la LIBS qui est une technique d'analyse élémentaire ne permet pas de quantifier la pollution ou la toxicité à l'aluminium d'un sol et que des analyses complémentaires restent nécessaire, par exemple pour savoir si le sol est acide (pH<5.8). Il n'est donc pas facile de trouver des normes donnant les valeurs limites de concentration en aluminium qui permettent de déterminer si le sol est pollué et seul un expert capable de prendre en compte la nature du sol pourra utiliser les valeurs de concentrations correctement. En appliquant la recette énoncée plus haut, nous avons recensé 4 raies persistantes pour l'aluminium à partir de la base de données du NIST : 308.215, 309.271, 394.400, et 396.152 nm. Ici, nous avons décidé de garder les raies résonnantes car le nombre total de raies est faible. En plus de ces 4 raies de l'aluminium, nous avons ajouté 4 raies caractérisant la matrice : une raie du Fe à 382.042 nm, une raie du Ti à 399.863 nm, une raie du Ba à 652.731 nm et une raie du Ca à 612.221 nm, soit un total de 8 entrées. Dans le cas présent, nous avons répartis les 67 échantillons pour lesquels l'aluminium était détecté en trois lots : 39 dans le lot d'apprentissage, 19 dans le lot de validation, et 9 dans le lot de test.

Les paramètres optimums pour l'ANN obtenus par validation croisée sont:

Nombre de neurones dans la couche cachée : 4
Vitesse d'apprentissage : 0.05
Terme de mémoire : 0.01
Nombre d'itérations : 8000

Grâce à cette optimisation, l'ANN a permis de prédire les concentrations en aluminium dans les échantillons de sol avec les performances mentionnées dans le Tableau 2-6. Là encore, on peut constater que l'erreur relative moyenne est inférieure à 20% pour les trois lots de données.

	Lot d'apprentissage	Lot de validation	Lot de test
R^2	0,90	0,87	0,96
Q^2	0,90	0,86	0,92
ER (%)	17	15	12
RMSE (g/100g)	1	1	1

Tableau 2-6 Performances du meilleur ANN pour l'analyse quantitative de Al

Pour achever l'évaluation du modèle, on applique la technique de Y-randomization pour une série de 25 calculs afin d'établir la moyenne et l'écart-type des facteurs de mérites R^2, Q^2 et RMSE. Les résultats présentés dans le Tableau 2-7 montrent clairement que le modèle d'ANN que nous avions retenu pour quantifier l'aluminium n'avait pas été optimisé par chance mais sur la base d'une véritable optimisation des poids aux variables d'entrée.

Y-random	Moyenne	écart-type
R^2c	0,04	0,05
R^2v	0,04	0,05
R^2t	0,27	0,15
Q^2r	0,50	0,15
Q^2v	-0,84	0,67
Q^2t	-0,65	0,62
RMSEC (g/100g)	2	0,4
RMSEV (g/100g)	4	0,7
RMSET (g/100g)	4	0,6
ERC (%)	48	17
ERV (%)	84	31
ERT (%)	79	27

Tableau 2-7 Moyenne et écart-type des facteurs de mérites calculés pour 25 permutations aléatoires des données de sortie (ici, la concentration de Al) pour le meilleur ANN optimisé pour une analyse quantitative de Al

La vérification par Y-randomization étant faite, on peut définitivement conclure que le modèle ANN ainsi optimisé permet de prédire les concentrations de Al avec une erreur relative moyenne inférieure à 20%. Mais observons néanmoins les performances détaillées du modèle pour chaque valeur de concentration. Sur la Figure 2-19, on a reporté les valeurs de concentrations données par ICP-AES (bleu) et par LIBS_ANN (rouge) pour le lot de validation. Les barres d'erreur des données ICP-AES correspondent simplement à la valeur relative de 20% calculée pour chaque valeur de concentration. Celles des donnes LIBS_ANN correspondent à l'écart-type de la valeur prédite après 5 répétitions du calcul d'ANN pour lesquelles on démarre à chaque fois de nouvelles valeurs aléatoires. On remarque que les concentrations prédites sont souvent dans les 20% de tolérance ou presque exceptées celles des échantillons SLM_C1_11 et SLM_C1_69 qui sont largement surestimées par rapport aux valeurs de référence. On notera cependant que ces deux valeurs de concentrations en aluminium autour de 8% auraient dû être correctement prédites puisque le modèle d'ANN a permis de prédire correctement les concentrations des échantillons du lot de test jusqu'à plus de 12% comme on peut le voir sur la Figure 2-20. Par conséquent, les deux échantillons SLM_C1_11 et SLM_C1_69 présentent probablement des caractéristiques de matrice qui rendent moins précise l'estimation de leur concentration en aluminium.

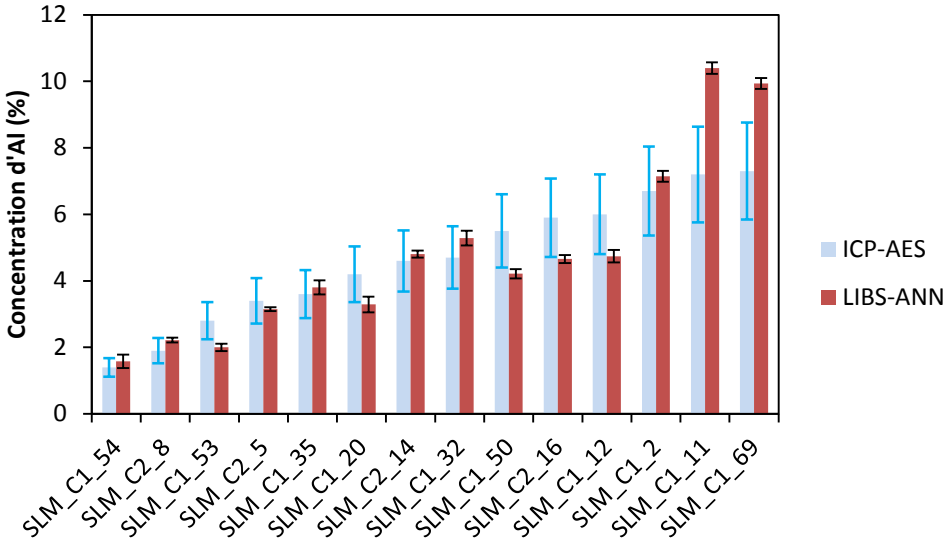


Figure 2-19 Comparaison entre la concentration d'aluminium de référence et la concentration prédite en LIBS-ANN pour les échantillons du lot de validation. Les barres bleues représentent les écarts de 20% autour de la concentration de référence, et les barres noires représentent l'écart-type sur 5 répétitions des concentrations prédites par ANN.

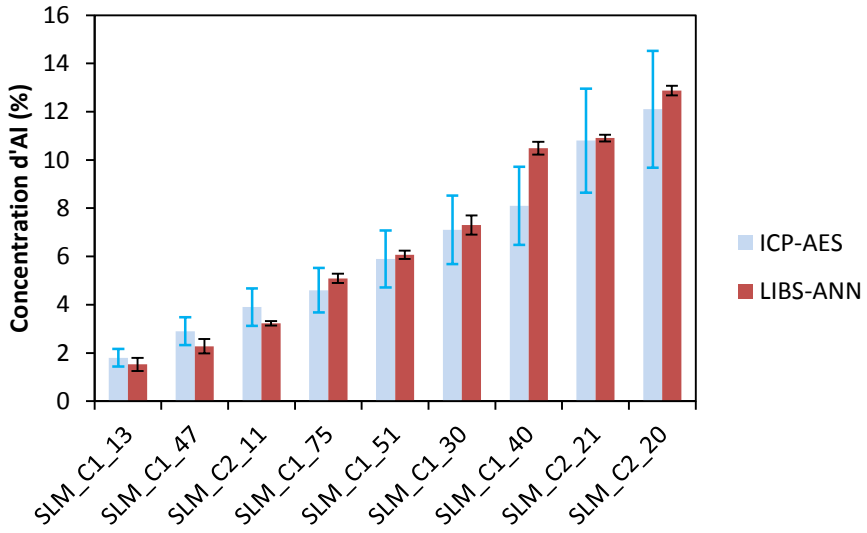


Figure 2-20 Comparaison entre la concentration d'aluminium de référence et la concentration prédite en LIBS-ANN pour les échantillons du lot de test. Les barres bleues représentent les écarts de 20% autour de la concentration de référence, et les barres noires représentent l'écart-type sur 5 répétitions des concentrations prédites par ANN.

2.3.3.2.4 Analyse quantitative du cuivre

Le cuivre est un élément mineur qui se trouve dans le sol grâce à des procédés naturels et humains. On le trouve souvent près des mines, des installations industrielles, des décharges et des broyeurs d'ordure. Dans le sol, il se lie fortement aux matières organiques et aux minéraux, et il peut finir par s'accumuler dans les plantes et les animaux. Du fait de son effet néfaste sur les plantes, le cuivre est une sérieuse menace pour la production des terres agricoles. On notera ici que dans la région bordelaise, les viticulteurs traitent la vigne contre les maladies à l'aide d'un mélange fortement concentré en cuivre appelé la bouillie bordelaise, qui est potentiellement nuisible à l'environnement. C'est pour cela que la concentration du cuivre dans le sol mérite d'être analysée. Sur les spectres LIBS, nous n'avons détecté que deux raies de cuivres à 324.754 et 327.396 nm que nous avons injectées dans l'ANN en ajoutant 5 raies de la matrice : une raie du Fe à 382.042 nm, une raie du Ti à 399.863 nm, une raie du Ba à 652.731 nm, une raie de Al à 309.271 nm et une raie du Ca à 612.221 nm, soit un total 7 entrées pour l'ANN. Nous avons séparé les 33 échantillons dans lequel le cuivre était détecté en 3 lots : 20 échantillons dans le lot d'apprentissage, 7 dans le lot de validation, et 6 dans le lot de test.

Les paramètres de l'ANN optimisés d'après la validation croisée sont:

Nombre de neurones dans la couche cachée : 3
Vitesse d'apprentissage : 0.1
Terme de mémoire : 0.01
Nombre d'itérations : 12000

Ces paramètres ont permis d'obtenir les résultats concernant le cuivre qui sont présentés dans le Tableau 2-8.

	Lot d'apprentissage	Lot de validation	Lot de test
R^2	0,93	0,95	0,88
Q^2	0,93	0,95	0,86
ER (%)	19	13	17
RMSE (ppm)	32	28	36

Tableau 2-8 Performances du meilleur ANN pour l'analyse quantitative de Cu

Comme précédemment, on adopte la technique de Y-randomization répétée 25 fois afin de tester si les résultats du modèle ont du sens. Les résultats de cette vérification sont donnés dans le Tableau 2-9.

Y-random	Moyenne	écart-type
R^2c	0,19	0,20
R^2v	0,24	0,27
R^2t	0,60	0,21
Q^2r	0,76	0,15
Q^2v	-19,00	83,12
Q^2t	-4,88	9,66
RMSEC (ppm)	53	18
RMSEV (ppm)	170	49
RMSET (ppm)	161	47
ERC (%)	34	12
ERV (%)	130	75
ERT (%)	105	54

Tableau 2-9 Moyenne et écart-type des facteurs de mérites calculés pour 25 permutations aléatoires des données de sortie (ici, la concentration de Cu) pour le meilleur ANN optimisé pour une analyse quantitative de Cu

On en déduit que le modèle ANN qui relie les données d'entrée X à la concentration de cuivre (Y) a du sens et n'a pas donné les résultats observés par chance. De plus, la Figure 2-21 montre que la prédiction de la concentration en cuivre fonctionne pour presque tous les échantillons sauf pour SLM_C2_1 qui est un peu sur estimé par LIBS_ANN. Tous les autres échantillons sont prédits avec une erreur relative moyenne de 20%

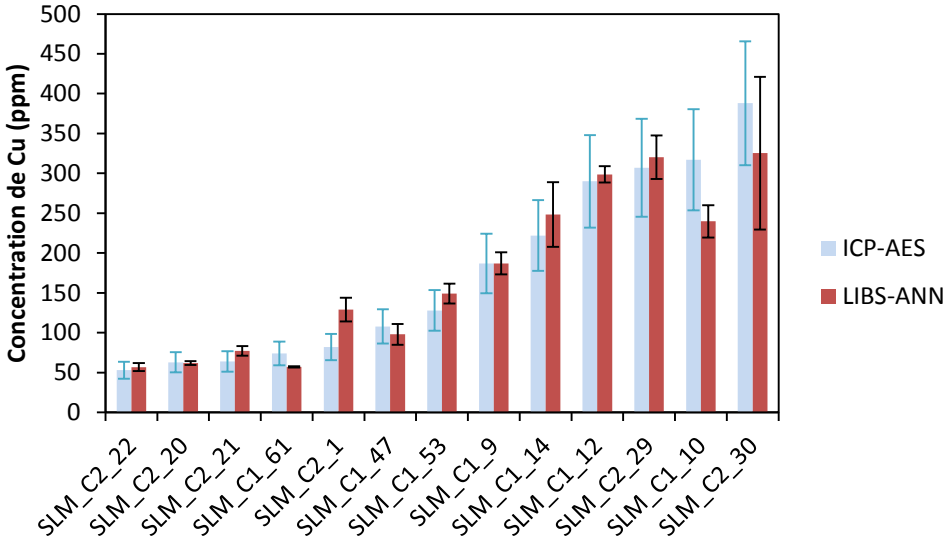


Figure 2-21 Comparaison entre la concentration de cuivre de référence et la concentration prédite en LIBS-ANN pour les échantillons des lots de validation et de test. Les barres bleues représentent les écarts de 20% autour de la concentration de référence, et les barres noires représentent l'écart-type sur 5 répétitions des concentrations prédites par ANN.

2.3.3.2.5 Analyse quantitative du fer

Le fer est le 4ème élément de la croûte terrestre, issu d'altérations (carbonates, silicates, phosphates). On le rencontre souvent sous la forme d'oxyde ferrique Fe_2O_3 de couleur rouille

ou d'hydroxyde ferrique $Fe(OH)_3$ dans les sols riches en oxygène. A l'inverse, dans les sols pauvres en oxygène, on le trouve sous forme d'oxyde ferreux FeO, d'hydroxyde ferreux $Fe(OH)_2$, de carbonate ferreux $FeCO_3$, bicarbonates $Fe(CO_3H)_2$ et de sulfure de fer FeS. Le fer joue un rôle majeur dans les activités microbiennes. En appliquant la recette présentée plus haut aux spectres LIBS des sols, on retient finalement 5 raies pour le Fer à 278.81, 358.119, 375.823, 382.042 et 438.354 nm. En entrée de l'ANN, on ajoute aussi les quatre raies de la matrice Ca (612.221 nm), Ti (399.863 nm), Ba (652.731 nm) et Al (309.271 nm) pour arriver à un total de 9 entrées. Puis on sépare en trois lots les 71 échantillons de sol : 44 pour le lot d'apprentissage, 17 pour le lot de validation et10 pour le lot de test.

Les paramètres de l'ANN identifiés comme étant optimum par validation croisée sont:

Nombre de neurones dans la couche cachée : 2
Vitesse d'apprentissage : 0.1
Terme de mémoire : 0.01
Nombre d'itérations : 15000

Notons ici que le nombre de neurones dans la couche cachée est seulement de deux. Cependant la Figure 2-22 montre clairement que l'erreur de prédiction est minimale pour ce nombre de neurones.

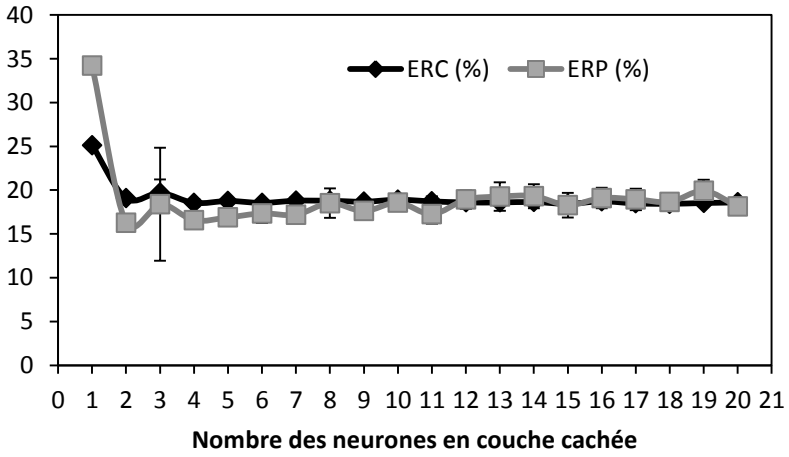


Figure 2-22 Evolution de ERC et ERP en fonction du nombre de neurones dans la couche cachée.

Finalement, les performances de ce réseau dédié à l'analyse quantitative du fer sont données dans le Tableau 2-10.

	Lot d'apprentissage	Lot de validation	Lot de test
R^2	0,97	0,82	0,90
Q^2	0,97	0,79	0,75
ER (%)	19	15	18
RMSE (g/100g)	0,8	1,0	1,1

Tableau 2-10 Performances du meilleur ANN pour l'analyse quantitative de Fe

Là encore, pour compléter l'étude, on procède à la vérification du modèle par la procédure de Y-randomization. Les résultats après 25 répétitions de la procédure sont indiqués dans le Tableau 2-11

Y-random	Moyenne	écart-type
R^2c	0,07	0,07
R^2v	0,19	0,18
R^2t	0,27	0,21
Q^2r	0,48	0,22
Q^2v	-9,17	14,70
Q^2t	-9,59	34,71
RMSEC (g/100g)	2	1
RMSEV (g/100g)	4	2
RMSET (g/100g)	5	3
ERC (%)	42	7
ERV (%)	107	68
ERT (%)	95	76

Tableau 2-11 Moyenne et écart-type des facteurs de mérites calculés pour 25 permutations aléatoires des données de sortie (ici, la concentration de Fe) pour le meilleur ANN optimisé pour une analyse quantitative de Fe

L'erreur relative moyenne est là encore inférieure à 20%, ce qui est tout à fait satisfaisant pour des mesures sur site. Cependant, on peut analyser les performances du modèle ANN non pas en moyenne mais pour chaque gamme de concentration. On obtient les résultats donnés sur les Figure 2-23 pour le lot de validation et Figure 2-24 pour le lot de test.

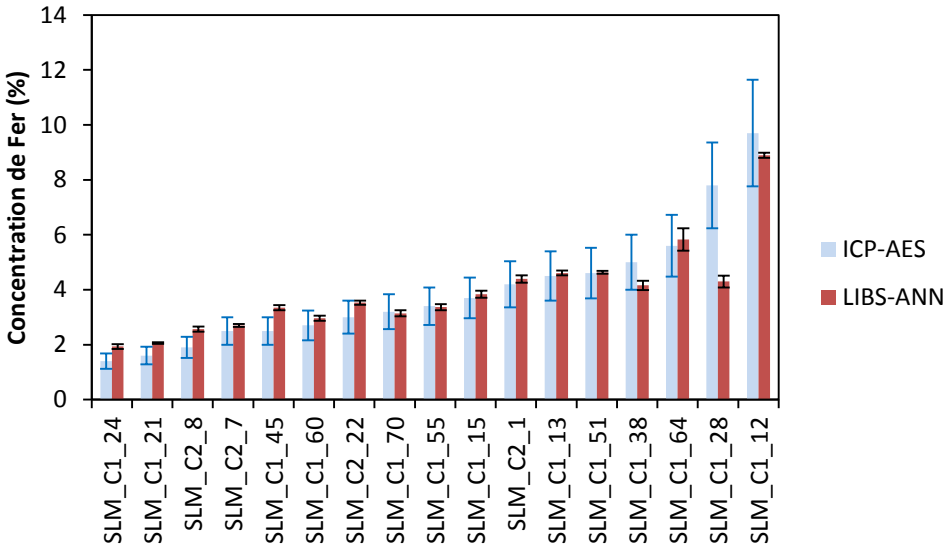


Figure 2-23 Comparaison entre la concentration de fer de référence et la concentration prédite en LIBS-ANN pour les échantillons du lot de validation. Les barres bleues représentent les écarts de 20% autour de la concentration de référence, et les barres noires représentent l'écart-type sur 5 répétitions des concentrations prédites par ANN.

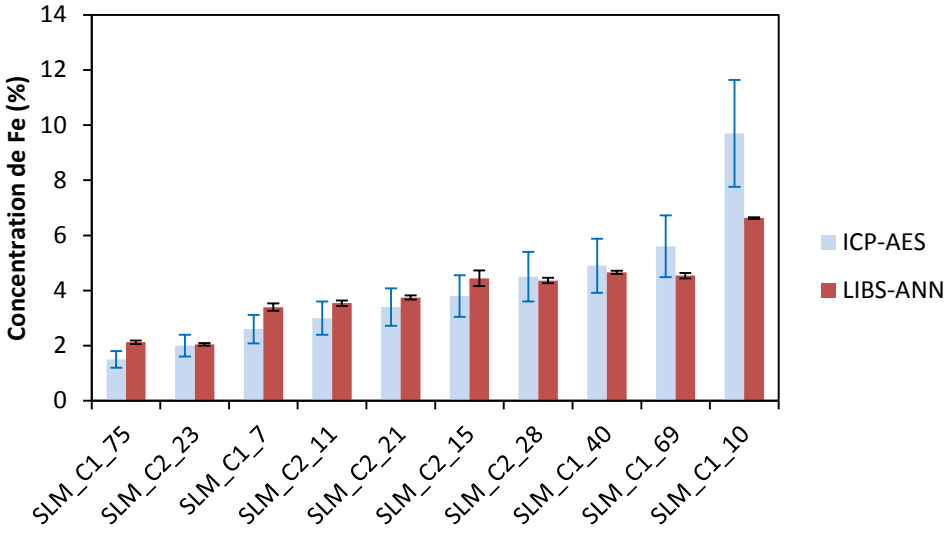


Figure 2-24 Comparaison entre la concentration de fer de référence et la concentration prédite en LIBS-ANN pour les échantillons du lot de test. Les barres bleues représentent les écarts de 20% autour de la concentration de référence, et les barres noires représentent l'écart-type sur 5 répétitions des concentrations prédites par ANN.

Ces résultats montrent que le modèle ANN appliqué aux données LIBS permet de prédire généralement bien la plupart des concentrations en fer. Il semblerait cependant que les concentrations les plus élevées, c'est-à-dire au-delà de 8%, soient sensiblement sous-estimées par l'ANN. Pour tenter d'expliquer cela, nous avons observé de plus près les spectres correspondants. Par exemple le spectre moyen associé à l'échantillon SLM_C1_28 présente un signal d'émission pour le fer très faible comparé aux autres échantillons dans la même gamme de concentrations d'après les données ICP-AES. Cet échantillon pourrait donc être considéré comme aberrant et exclu de l'analyse quantitative. Les raisons de l'écart entre le résultat LIBS et le résultat ICP-AES n'ont cependant pas été identifiées dans le cadre de ce travail de thèse.

2.3.3.3 Quantification du plomb dans le cas d'une large gamme de concentrations

Le plomb est l'un des métaux lourds que l'on trouve de façon naturelle dans notre environnement mais dont l'origine est le plus souvent anthropique lorsqu'on atteint des concentrations qui en font un polluant. Il peut notamment provenir de fumées d'échappement, de procédés industriels, de la combustion des déchets solides, de la corrosion des tuyauteries, et on le trouve également en forte concentration sur les anciens sites miniers. Les effets du plomb sur la santé sont bien connus avec en particulier le saturnisme infantile si bien que le plomb est l'un des éléments de choix lorsqu'il s'agit de démontrer le potentiel d'une nouvelle technique de mesure environnementale. Dans le cadre de notre étude LIBS, la particularité est que nous avons collecté lors de la campagne de mesure SLM des échantillons de sols avec des concentrations en plomb extrêmement différentes et s'étalant sur une gamme de concentration très étendue allant de 200 à 100 000 ppm. Nous insistons ici sur le fait que ces valeurs de

concentrations ont été rencontrées pour des sols directement prélevés sur le terrain sans aucun ajout ultérieur.

2.3.3.3.1 Analyse du plomb par un seul modèle ANN

Dans le spectre LIBS, nous avons sélectionné 5 raies persistantes de Pb I à : 261.418, 283.305, 363.957, 368.346 et 405.781 nm. Pour les raisons déjà évoquées, nous avons également pris en compte 5 raies de la matrice (Fe (382.042 nm), Ca (612.221 nm), Ti (399.863 nm), Ba (652.731 nm) et Al (309.271 nm)), de sorte que nous avons exploité en tout 10 données d'entrée pour le calcul ANN. Les 62 échantillons prélevés sur le site SLM ont été répartis en trois lots avec 60% en apprentissage, 30% en validation et 10% en test.

Les paramètres optimums de l'ANN obtenus par validation croisée sont:

Nombre de neurones dans la couche cachée : 3
Vitesse d'apprentissage : 0.3
Terme de mémoire : 0.1
Nombre d'itérations : 10000

Les résultats de ce modèle ANN sont donnés dans le Tableau 2-12. On remarque que malgré l'existence d'une corrélation entre les données LIBS et les valeurs de concentrations en plomb, l'erreur relative calculée en moyenne sur tous les échantillons de chaque lot est supérieure à 150%. Ce résultat nous permet de conclure que dans ce cas précis, l'ANN est incapable de prédire correctement les valeurs des concentrations en plomb.

62 échantillons	R^2	Q^2	ER (%)	RMSE (ppm)
Lot d'apprentissage	0.98	0.98	161	4 600
Lot de validation	0.82	0.75	160	12 600
Lot de test	0.80	0.69	157	13 400

Tableau 2-12 Résultats pour la quantification du plomb des échantillons de sols du site SLM par un modèle ANN unique

La raison des mauvaises performances analytiques observées ici semble être la très vaste gamme de concentrations. En LIBS et lorsqu'on adopte une approche univariée, il est fréquent de quantifier un élément faiblement concentré à l'aide d'une première raie spectrale et de changer de raie spectrale pour traiter les cas des fortes concentrations. En effet, la seule raie spectrale détectable aux faibles concentrations peut présenter un comportement indésirable de saturation à des concentrations plus élevées, notamment à cause du phénomène d'auto-absorption. Dans le Tableau 2-12, les valeurs de RMSE pour les lots de validation et de test étant supérieures à 12 500 ppm, nous avons choisi de fixer comme valeur seuil la valeur de concentration de 10 000 ppm et de développer deux modèles ANN. Ainsi, les concentrations inférieures à 10 000 ppm seront prédites par le modèle ANN1 tandis que celles supérieures à 10 000 ppm par le modèle ANN2.

2.3.3.3.2 Analyse du plomb par 2 modèles ANN

Les 62 échantillons prélevés sur le site SLM ont été divisés en deux groupes en fonction de leur concentration sur la base des valeurs fournies par ICP-AES. Ainsi le premier groupe de 36 échantillons correspond à des valeurs de concentration en plomb inférieures à 10 000 ppm tandis que le second groupe de 26 échantillons correspond à des valeurs de concentrations en plomb supérieures à ce seuil. L'idée est de rechercher le modèle ANN optimum pour chaque groupe :

- ANN1 est donc bâti pour des prédire des concentrations en plomb inférieures à 10 000 ppm : 21 échantillons en apprentissage, 9 en validation et 6 en test.
- ANN2 est quant à lui conçu pour prédire des concentrations en plomb supérieures à 10 000 ppm : 14 échantillons en apprentissage, 6 en validation et 6 en test.

Le Tableau 2-13 donne les paramètres des modèles ANN1 et ANN2 après recherche des conditions optimales par la méthode de validation croisée.

Modèle	ANN1	ANN2
Nombre de neurones dans la couche cachée	5	5
Vitesse d'apprentissage	0.2	0.02
Terme de mémoire	0.1	0.1
Nombre d'itérations	38 000	18 000

Tableau 2-13 Paramètres des modèles ANN permettant de quantifier le plomb pour des échantillons de sols prélevés sur le site SLM

Les performances de ces deux modèles ANN sont données dans le Tableau 2-14. On constate que les erreurs relatives se situent autour de 20% ce qui est conforme au critère discuté précédemment et qui rend cette analyse du plomb tout à fait acceptable. Notons cependant que lorsqu'on regarde de plus près les résultats du modèle ANN1, l'erreur relative étant beaucoup plus faible pour le lot de calibration que pour les lots de validation et de test, on pourrait craindre un sur-apprentissage du modèle. Or, les résultats de validation croisée pour le modèle ANN1 présentés sur la Figure 2-25 et qui donnent l'évolution des paramètres RMSEC et RMSEV d'une part et REC et REV d'autre part, montrent clairement que le choix d'un modèle ANN1 optimisé pour 38 000 itérations est tout à fait justifié. En effet, non seulement les valeurs des erreurs pour le lot de validation sont les plus basses mais en plus, les écarts-type sur ces erreurs, obtenus à partir de 5 calculs successifs, sont également réduits pour ce nombre d'itérations, indiquant que le modèle ANN est très stable dans ces conditions.

Modèle	ANN1	ANN2
Gamme de concentration	<10 000 ppm	>10 000 ppm
R^2c	0.99	0.91
R^2v	0.98	0.78
R^2 t	0.97	0.92
Q^2c	0.99	0.90
Q^2v	0.96	0.79
Q^2t	0.95	0.84
ERC (%)	9	19
ERV (%)	24	22
ERT (%)	21	20
RMSEC (ppm)	80	9520
RMSEV (ppm)	450	9250
RMSET (ppm)	410	8520

Tableau 2-14 Performances des modèles ANN1 et ANN2 associés aux paramètres donnés dans le Tableau 2-13

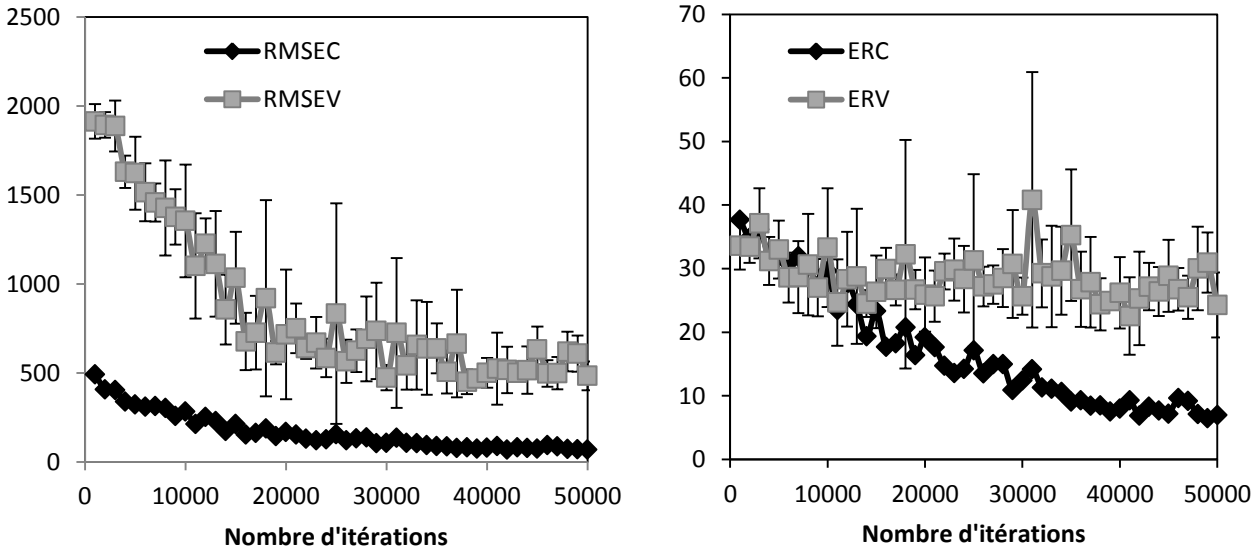


Figure 2-25 Evolution des erreurs RMSEC et RMSEV ainsi que ERC et ERP en fonction du nombre d'itérations pour le modèle ANN1.

Cette étude sur l'analyse quantitative du plomb ne serait pas complète sans la dernière vérification basée sur la procédure de Y-randomization. Les résultats obtenus pour ANN1 et ANN2 dans le cadre de cette procédure sont présentés dans le Tableau 2-15. Ils montrent que les performances des modèles sont considérablement dégradées après Y-randomization ce qui confirme que les résultats présentés dans le Tableau 2-14 ont une réelle signification statistique.

Modèle Y-randomization	ANN1 <10 000 ppm	ANN2 >10 000 ppm
R^2c	0.73	0.79
R^2v	0.01	0.00
R^2t	0.00	0.00
Q^2c	0.7	0.6
Q^2v	-3.8	-1.5
Q^2t	-34.9	-2.4
ERC (%)	72	38
ERV (%)	221	89
ERT (%)	221	86
RMSEC (ppm)	1070	15190
RMSEV (ppm)	3110	33890
RMSET (ppm)	3040	34930

Tableau 2-15 Performances des modèles ANN1 et ANN2 résultant de la procédure Y-randomization.

Toutes les vérifications ayant été faites, il est à présent possible d'analyser les performances des modèles ANN1 et ANN2 non pas à travers des quantités moyennes mais au contraire pour chaque valeur de concentration des échantillons des lots de validation et de test. Les résultats concernant le modèle ANN1 sont donnés dans la **Figure 2-26** et ceux concernant le modèle ANN2 dans la figure 2-27. Dans les deux cas, nous avons comparé les valeurs prédites par l'approche LIBS-ANN (rouge) à celles fournies par la technique ICP-AES et considérées comme étant les valeurs de référence (bleu). Les barres d'erreur portant sur les données LIBS-ANN correspondent à l'écart-type sur 5 répétitions du calcul ANN et celles portant sur les données ICP-AES indiquent simplement une tolérance de 20% pour chaque valeur de concentration dans le but d'interpréter plus facilement les résultats des analyses sur site. On constate que pour le modèle ANN1, les prédictions sont généralement satisfaisantes à l'exception de quelques échantillons. De façon générale, le modèle ANN1 a tendance à fournir des valeurs surestimées bien que pour l'échantillon T_4, la valeur prédite soit étrangement basse comparée à la valeur de référence avec pourtant un écart-type très faible. De même, le modèle ANN2 a tendance à fournir des valeurs supérieures aux valeurs de référence (**Figure 2-27**) mais le plus souvent dans l'intervalle de tolérance de 20%. On vérifie ainsi que les performances des modèles ANN1 et ANN2 sont tout à fait satisfaisantes pour chaque gamme de concentration en plomb.

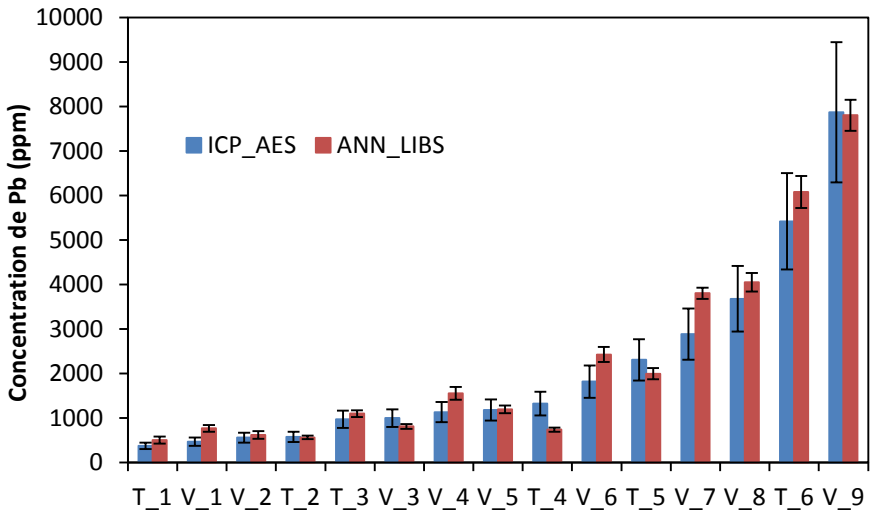


Figure 2-26 Comparaison entre les concentrations de plomb de référence données par ICP-AES (bleu) et les concentrations prédites en LIBS-ANN (rouge) à l'aide du modèle ANN1 pour les échantillons des lots de validation et de test. Les barres d'erreur représentent, pour les données ICP-AES, un écart de 20% autour de la concentration de référence, et pour les données LIBS-ANN, l'écart-type sur 5 répétitions du calcul.

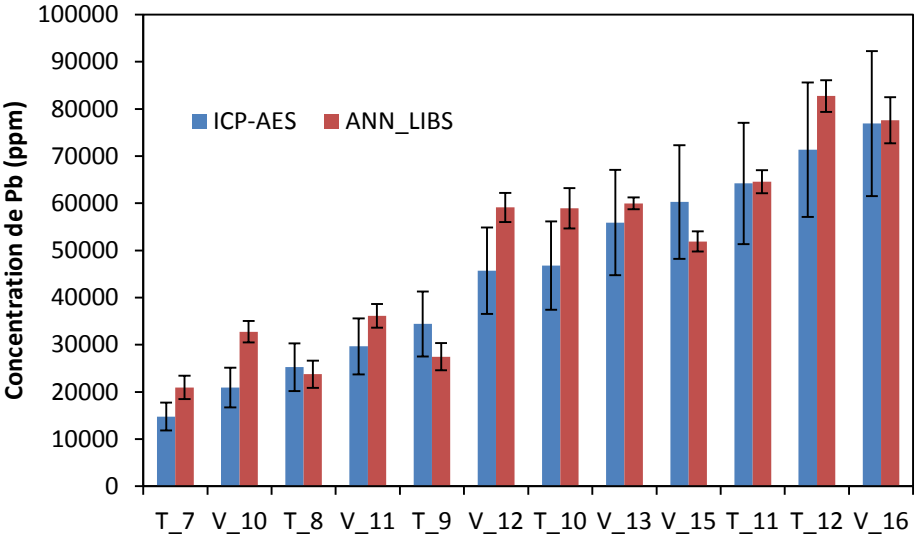


Figure 2-27 Comparaison entre les concentrations de plomb de référence données par ICP-AES (bleu) et les concentrations prédites en LIBS-ANN (rouge) à l'aide du modèle ANN2 pour les échantillons des lots de validation et de test. Les barres d'erreur représentent, pour les données ICP-AES, un écart de 20% autour de la concentration de référence, et pour les données LIBS-ANN, l'écart-type sur 5 répétitions du calcul.

Reste à savoir lequel des deux modèles on doit appliquer à un échantillon inconnu. En effet le modèle ANN1 n'a de sens que jusqu'à la valeur maximale de 10 000 ppm. A l'inverse, si on

applique le modèle ANN2 à un échantillon de concentration en plomb inférieure à 10 000 ppm, on sait d'avance que le résultat sera médiocre puisqu'un seul modèle ANN ne nous a pas permis de quantifier tous les échantillons. Pour choisir le modèle approprié, on procède à un tri des échantillons de telle sorte que les échantillons présentant une concentration en plomb inférieure à 10 000 ppm seront dans la classe 0 et les autres dans la classe 1. Ce tri a été réalisé à l'aide d'un réseau de neurone dont le seul neurone de la couche de sortie fournit comme réponse la valeur 0 pour indiquer de classer l'échantillon dans la classe 0 et la valeur 1 dans l'autre cas. On conserve les mêmes données d'entrée que pour une analyse quantitative, à savoir 5 raies de plomb et 5 raies d'éléments de la matrice. Les 62 échantillons de sols prélevés sur le site SLM ont été répartis en 3 lots avec 36 échantillons en apprentissage, 19 en validation et 7 en test.

Les paramètres de l'ANN identifiés comme étant optimum par validation croisée sont:

Nombre de neurones dans la couche cachée : 3
Vitesse d'apprentissage : 0.05
Terme de mémoire : 0.1
Nombre d'itérations : 21000

La fonction d'activation étant une sigmoïde, la valeur de sortie n'est pas simplement binaire 0/1 mais peut aussi prendre des valeurs intermédiaires. Ceci est illustré sur la Figure 2-28 pour les échantillons du lot d'apprentissage. D'après les données ICP-AES (bleu), on s'attend à classer dans la classe 0 tous les échantillons de C_1 à C_21 et dans la classe 1 les échantillons de C_22 à C_36. Les résultats fournis par l'ANN sont assez proches mais il sera nécessaire d'introduire une tolérance afin de minimiser les erreurs de classement. On remarque enfin que le changement de valeur 0->1 est bien obtenu pour des valeurs de concentrations autour de 10 000 ppm, l'échantillon C_21 étant concentré à 9380 ppm alors que les échantillons C_22 et C_23 sont respectivement à 11420 et 12150 ppm.

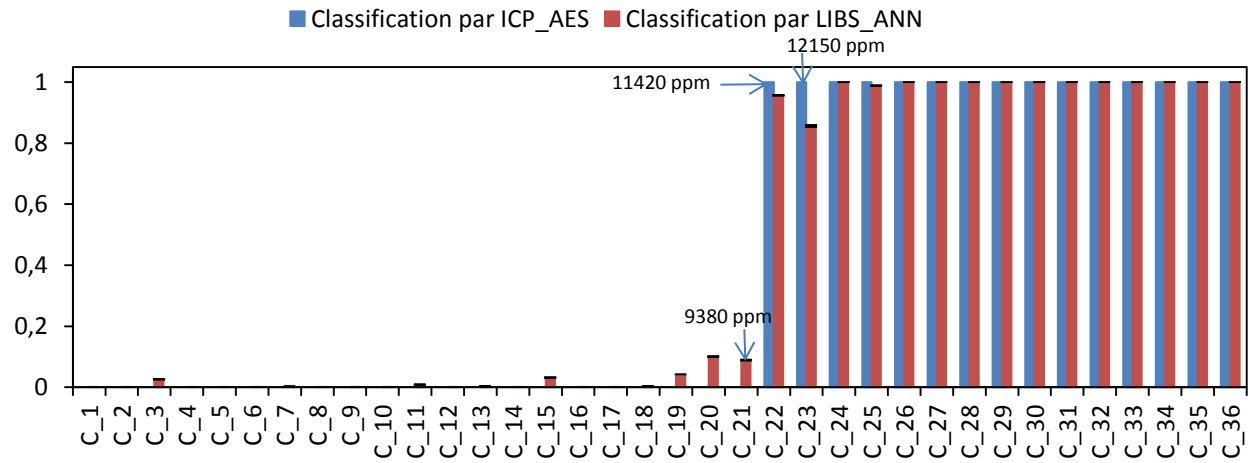


Figure 2-28 Comparaison de la classification LIBS-ANN et de la classification par ICP-AES pour les échantillons de sols du site SLM.

Afin d'évaluer proprement les performances du tri, on fait appel aux indicateurs qui ont été présentés dans le chapitre 1 de ce mémoire. Le Tableau 2-16 donne ainsi les performances globales du tri par ANN pour le lot de validation dans le cas où l'on n'introduit aucune tolérance. Il faut comprendre que dans ce cas, toute valeur numérique différente de 0 ou de 1 donnera un résultat « faux ». Le test qui est décrit dans le Tableau 2-16 consiste à déterminer si un échantillon appartient à la classe 1, étant entendu que tous ceux qui n'appartiennent pas à la classe 1 sont rangés dans la classe 0. Pour être rangé dans la classe 1, l'échantillon doit avoir des données spectrales telles que le calcul d'ANN sera exactement égal à 1 car on n'a pas introduit de tolérance.

Valeur de tolérance = 0 36 échantillons / Lot d'apprentissage	Classe 1 d'après l'analyse ICP-AES	Classe 0 d'après l'analyse ICP-AES
Classe 1 d'après l'analyse ANN des données LIBS	12 vrais positifs	0 faux positifs
Classe 0 d'après l'analyse ANN des données LIBS	3 faux négatifs	21 vrais négatifs

Tableau 2-16 Performances de la classification par ANN-LIBS pour les 36 échantillons du lot de validation.

On constate qu'il n'y a aucun faux positif donc aucun échantillon qui donnerait une valeur de sortie d'ANN égale à 1 alors qu'il aurait une concentration en Pb inférieure à 10 000 ppm. En revanche il y a 3 faux négatifs, c'est-à-dire trois échantillons classés en 0 alors qu'ils auraient dû être en classe 1. Cela est dû à des valeurs de sortie proches de 1 mais tout de même inférieures à cette valeur. L'absence de tolérance sur ce test induit ces faux négatifs et il est donc recommandé d'introduire une tolérance afin de réduire le nombre de mauvais classements. En l'absence de tolérance, la sensibilité est égale à 12/15=80% et la spécificité à 100%.

Lorsqu'on passe à une tolérance égale à 0.05, cela signifie que toute valeur de sortie de l'ANN supérieure à 0.95 sera assimilée égale à 1 si bien que l'échantillon sera rangé dans la classe 1. Dans ce cas, la sensibilité s'améliore et on atteint 93%. Enfin, lorsqu'on fixe la tolérance à 0.2, cela signifie qu'on range dans la classe 1 tout échantillon donnant une valeur de sortie d'ANN supérieure à 0.8. Dans ce cas, on obtient une sensibilité de 100% avec toujours une spécificité de 100%. On en conclut qu'une tolérance égale à 0.2 permet de réaliser le tri des échantillons en deux classes avec 100% de succès. Ceci a été vérifié pour les deux lots de validation et de test.

Mais rappelons ici que ces résultats de tri ont été obtenus en injectant 10 données en entrée de l'ANN. Il est intéressant de tester les performances d'un tel tri dans le cas où l'on ne garde que les 5 raies du plomb en entrée de l'ANN. Les résultats pour les échantillons du lot de calibration sont présentés dans le Tableau 2-17 pour différentes valeurs de tolérance.

valeur de tolérance	Sensibilité (%)	Spécificité (%)
0.05	67	100
0.2	80	100
0.3	80	100
0.4	81	100
0.5	93	95

Tableau 2-17 Performances de la procédure de classification par ANN pour le lot d'apprentissage avec uniquement 5 raies du plomb en données d'entrée de l'ANN pour différentes valeurs de tolérance.

On remarque que la sensibilité n'est que de 67% pour une tolérance de 0.05 et qu'elle augmente jusqu'à 93% pour une tolérance de 0.5. Cependant la spécificité baisse dans ce dernier cas. Ceci illustre une nouvelle fois l'intérêt de prendre en compte dans les calculs par ANN des données supplémentaires qui ne proviennent pas de l'analyte mais plutôt de la matrice.

2.3.4 Présentation des données ICP-AES dans un diagramme ternaire

Nous avons démontré précédemment la possibilité de quantifier par ANN-LIBS des éléments majeurs ou mineurs constituants des échantillons de sol. Nous avons aussi présenté la capacité de l'ANN à classer les échantillons d'un même site en différentes classes selon une valeur seuil de concentration préalablement fixée. Reste donc à démontrer que l'ANN peut aussi classer les sols provenant de différents sites en fonction de la matrice qui les compose. Pour cela, commençons par observer la nature des sols prélevés sur la base des données de référence ICP-AES. Sur un plan géologique, on distingue généralement trois pôles permettant de décrire les sols dans un diagramme ternaire. Il s'agit du pôle silicaté lié à la présence d'aluminium et de silicium puis du pôle carbonaté associé à la présence de calcium et de magnésium et enfin le pôle minerai lié à la présence de métaux tels que le zinc. Dans le cadre de cette étude, le pôle minerai sera plus précisément caractérisé par le zinc, le baryum et le plomb. On a donc choisi de fabriquer trois nombres censés représenter les pourcentages de chaque pôle dans la composition d'un sol.

$$\%\ \text{pôle silicaté} = \frac{[Si]+[Al]}{[Si]+[Al]+[Ca]+[Mg]+[Ba]+[Zn]+[Pb]}$$

$$\%\ \text{pôle carbonaté} = \frac{[Ca]+[Mg]}{[Si]+[Al]+[Ca]+[Mg]+[Ba]+[Zn]+[Pb]}$$

$$\%\ \text{pôle minerai} = \frac{[Ba]+[Zn]+[Pb]}{[Si]+[Al]+[Ca]+[Mg]+[Ba]+[Zn]+[Pb]}$$

Notons que la somme des trois nombres est toujours égale à 100% par définition et que les valeurs des concentrations proviennent des analyses ICP-AES.

Les échantillons de sols de trois campagnes sur trois sites différents sont présentés dans le diagramme ternaire en Figure 2-29. Ceux du site SLM sont en bleu, ceux du site SEB en rouge et enfin ceux du site ME en vert. On remarque que les échantillons des sites SEB et ME sont très silicatés et très peu carbonatés. Ils sont également très peu riches en minerais. A l'inverse, on remarque que les échantillons SLM présentent une grande variabilité dans le diagramme, certains étant très riches en minerais, d'autres étant très silicatés et d'autres encore très carbonatés. Une telle diversité obtenue à l'aide des mesures ICP-AES confirme que les analyses LIBS ne peuvent pas être simples.

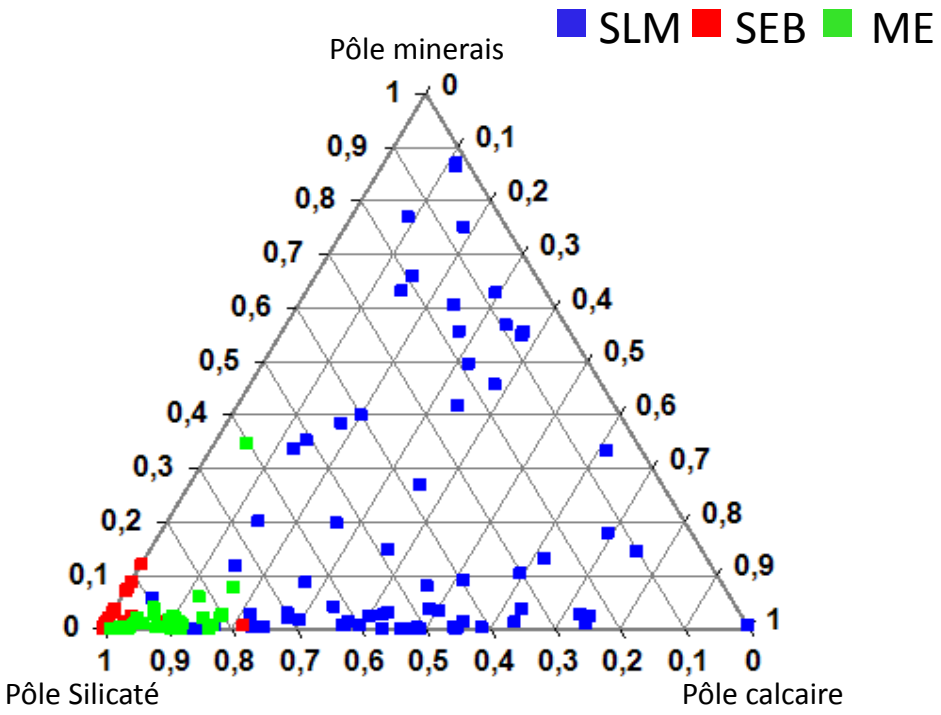


Figure 2-29 Diagramme ternaire indiquant les concentrations relatives des trois matrices pour les échantillons de sols provenant de 3 campagnes de mesure, d'après les données ICP-AES.

2.3.5 Présentation des données LIBS-ANN dans un diagramme ternaire

On cherche à savoir si les analyses LIBS permettent de retrouver une répartition des échantillons de sols dans un diagramme ternaire semblable à celle obtenue à partir des données ICP-AES. On utilise donc ici un modèle ANN qui possède 3 neurones dans la couche de sortie. Les valeurs de sortie sont les concentrations relatives (%) de chacune des trois matrices. Les données d'entrée de l'ANN sont dans ce cas les 35 raies présentées dans le Tableau 2-18. Elles permettent de prendre en compte les trois pôles décrits précédemment et ont été sélectionnées à partir de la base de données du NIST.

Eléments et longueurs d'onde (nm) des raies sélectionnées	
Si	250.691; 251.625; 251.927 ; 252.42 ; 252.849; 288.151
Zn	307.558; 319.675; 330.262; 334.487; 472.211; 481.058; 636.276
Ca	612.242; 442.563; 558.871; 610.301; 616.217; 643.914; 646.279
Al	309.291; 308.223; 394.407; 396.17
Pb	261.407; 283.298; 363.969; 368.36; 405.789
Ba	652.735; 659.538; 669.393; 706.028; 728.009
Mg	285.219

Tableau 2-18 Raies spectrales sélectionnées pour fournir les données d'entrée de l'ANN dédié à l'analyse des matrices de sols.

L'architecture du modèle ANN dédié à l'analyse des matrices est basée sur 3 couches et sur 3 neurones dans la couche de sortie (Figure 1-12). Chaque neurone de la couche de sortie donne directement le pourcentage du pôle silicaté (1er neurone), celui du pôle carbonaté (2ème neurone) et enfin celui du pôle minerai (3ème neurone). Dans un souci d'évaluation de la méthode, les échantillons sont répartis en trois lots, comme indiqué dans le Tableau 2-19.

Nombre des échantillons	SLM_C1	SLM_C2	SEB	ME_S2
Lot d'apprentissage	27	14	20	23
Lot de validation	14	7	11	11
Lot de test	5	3	4	3

Tableau 2-19 Répartition en trois lots des échantillons provenant de quatre campagnes de mesures

Les paramètres choisis pour ce nouvel ANN à 3 sorties, sur la base de la validation croisée, sont :

Nombre de neurones dans la couche cachée : 10
Vitesse d'apprentissage : 0.01
Terme de mémoire : 0.1
Nombre d'itérations : 32000

Avec ces paramètres d'optimisation, on obtient pour les trois neurones de sortie les performances présentées dans le Tableau 2-20.

	Pôle silicaté	Pôle carbonaté	Pôle minerai
R^2			
Lot d'apprentissage	0.95	0.92	0.99
Lot de validation	0.97	0.93	0.90
Lot de test	0.94	0.94	0.97
RMSE (%)			
Lot d'apprentissage	7	6	2
Lot de validation	5	6	5
Lot de test	7	5	4

Tableau 2-20 Performances de l'ANN à trois sorties pour une analyse semi-quantitative

Grâce à ce calcul d'ANN, on peut établir le diagramme ternaire présenté sur la Figure 2-30.

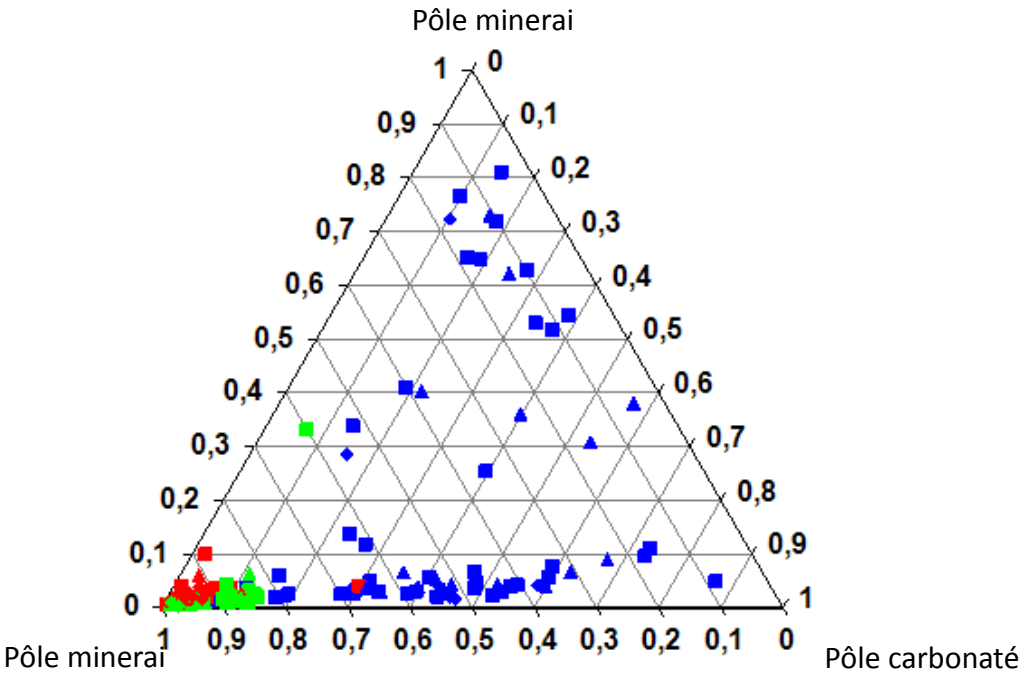


Figure 2-30 Diagramme ternaire donnant les concentrations relatives des échantillons de sols calculées par LIBS-ANN. Rouge : SEB, vert : ME, Bleu : SLM. Les carrés représentent le lot d'apprentissage, les triangles le lot de validation, les losanges le lot de test.

On obtient une répartition des échantillons dans le diagramme ternaire très semblable à celle établie par les données ICP-AES et des valeurs de RMSE pour les concentrations prédites inférieures à 10 %, pour les trois lots et pour les trois matrices. Ce résultat est très intéressant car il démontre que les données LIBS traitées par ANN permettent d'établir un diagramme ternaire conforme à celui donné par les données de référence. Par conséquent, il est possible de mener une analyse LIBS-ANN semi-quantitative qui permet de connaître à quel type de matrice est associé chaque échantillon de sol. L'intérêt de cette analyse semi-quantitative est de permettre de connaître la matrice caractéristique d'un sol inconnu afin de pouvoir ensuite appliquer le modèle quantitatif le mieux adapté.

2.3.6 Analyse quantitative multi-ANN

La différence entre les matrices des sols provenant du site SLM d'une part et des sites SEB et ME d'autre part a été établie par ACP ainsi que par le biais d'une analyse par ANN à trois sorties correspondant aux trois pôles : silicaté, carbonaté et minerai. La question qui se pose à présent est de savoir si on peut bâtir un modèle ANN quantitatif à partir d'échantillons provenant de différents sites.

2.3.6.1 Analyse quantitative de l'aluminium

L'aluminium a été choisi comme premier élément d'étude. Les données d'entrée sont les 8 raies qui ont servi précédemment à l'analyse de l'aluminium sur le site SLM seul. Les échantillons de sols des trois sites SLM, ME et SEB ont été répartis comme suit : 91 échantillons dans le lot d'apprentissage, 34 dans le lot de validation et 12 dans le lot de test.

Le meilleur modèle ANN a été trouvé pour les paramètres suivants : 5 neurones dans la couche cachée, vitesse d'apprentissage = 0.1, terme de mémoire = 0.1 et nombres d'itérations = 8000. Dans ces conditions, les résultats obtenus sont présentés dans le Tableau 2-21.

	Lot d'apprentissage	Lot de validation	Lot de test
R^2	0,79	0,68	0,77
Q^2	0,79	0,67	0,77
ER (%)	19	19	20
RMSE (%)	2	2	2

Tableau 2-21 Performances du modèle ANN pour quantifier l'aluminium et basé sur un apprentissage à partir d'échantillons provenant des trois sites SLM, ME et SEB

Concernant les 34 échantillons du lot de validation, la Figure 2-31 montre que 27 d'entre eux sont quantifiés avec une erreur relative inférieure à 20%. En ce qui concerne le lot de test, ce sont 9 échantillons sur 12 qui sont quantifiés avec une erreur relative inférieure à 20% comme le montre la Figure 2-32.

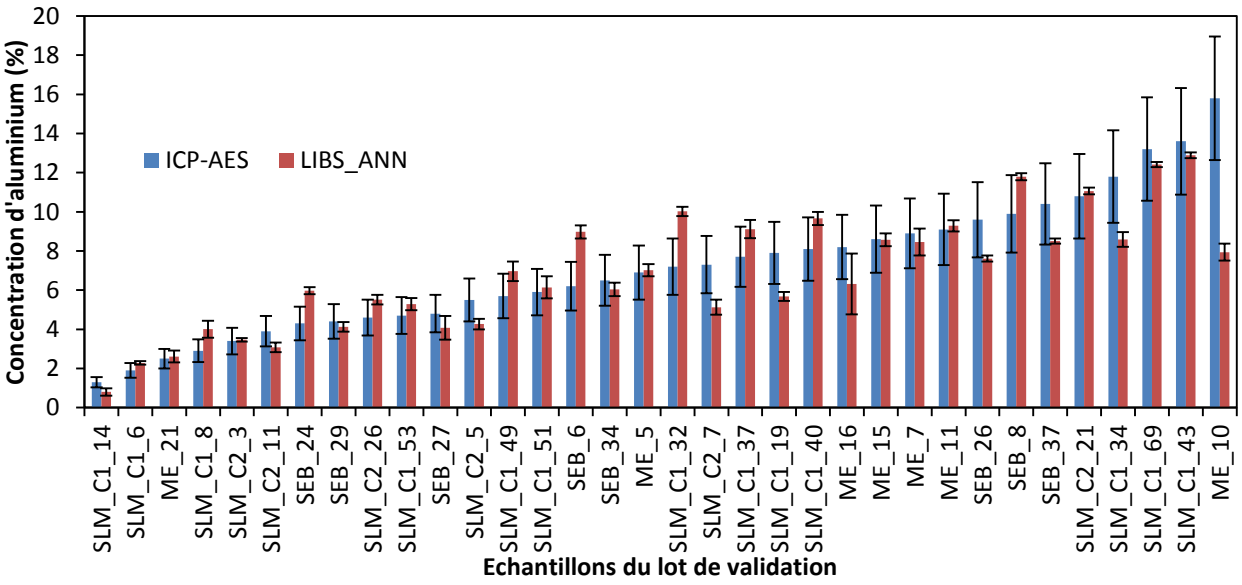


Figure 2-31 Concentrations d'aluminium pour les échantillons du lot de calibration. Bleu : ICP-AES, rouge : LIBS-ANN. Les barres d'erreurs sur les concentrations ICP-AES représentent l'erreur relative de 20 % et celles des concentrations prédites par LIBS-ANN correspondent aux écarts-types résultant de 5 répétitions du modèle ANN.

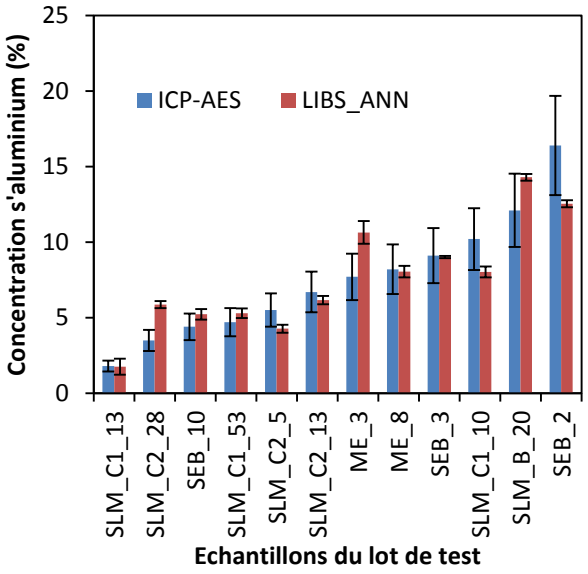


Figure 2-32 Concentrations d'aluminium pour les échantillons du lot de test. Bleu : ICP-AES, rouge : LIBS-ANN. Les barres d'erreurs sur les concentrations ICP-AES représentent l'erreur relative de 20 % et celles des concentrations prédites par LIBS-ANN correspondent aux écarts-types résultant de 5 répétitions du modèle ANN.

En conclusion, le modèle ANN construit pour la quantification de l'aluminium a permis de prédire 80% des échantillons dans la limite de tolérance des 20% d'erreur relative. Par conséquent, il est tout à fait capable de prédire à l'aide d'un seul modèle ANN la concentration en aluminium d'un échantillon inconnu, que celui-ci soit associé à une matrice silicatée, carbonatée ou minerai. Notons que ce résultat très positif a été établi pour un élément majeur des sols mais que pourtant une analyse univariée basée sur la raie de Al I à 309.271 nm donne une valeur de R^2= 0.058 seulement qui traduit le fait d'une absence de corrélation entre cette raie de Al I et les concentrations (cf. Figure 2-33). On en déduit que les résultats obtenus par l'ANN sont remarquables.

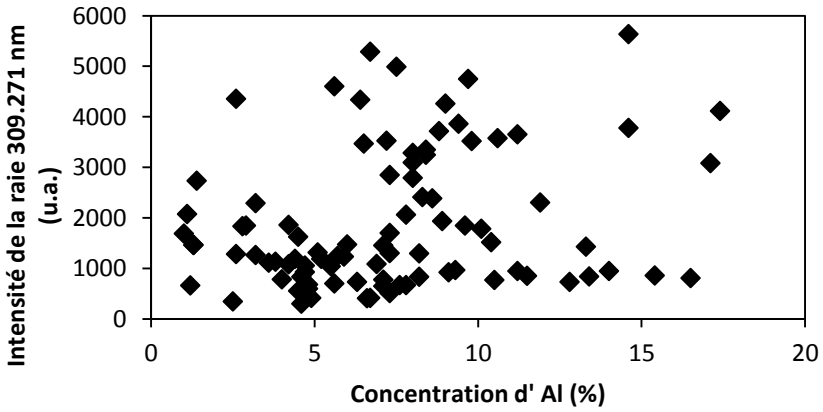


Figure 2-33 Intensité de la raie d'Al I à 309.271 nm en fonction de la concentration en aluminium pour le lot d'apprentissage (analyse univariée).

2.3.6.2 Analyse quantitative du plomb

Le plomb est un élément que l'on trouve selon des concentrations très diverses comme on l'a vu précédemment. Nous avons montré dans le cas de l'analyse des échantillons du site SLM qu'il était nécessaire de séparer les échantillons en deux groupes, l'un correspondant aux concentrations en plomb inférieures à 10 000 ppm et l'autre à celles supérieures à 10 000 ppm.

En généralisant le travail précédent (un seul site géologique) à l'analyse de sols provenant de différents sites, un premier modèle ANN a été construit pour les échantillons ayant une concentration en plomb inférieure à 10 000 ppm, d'après les valeurs de référence. Parmi les 89 échantillons prélevés sur les trois sites SLM, ME et SEB, 55 ont été utilisés pour le lot de calibration, 26 pour celui de validation et 8 pour celui de test. Les 10 raies (5 relatives au plomb et 5 à la matrice) déjà exploitées dans l'étude précédente ont été sélectionnées pour cette analyse. L'ANN a été optimisé par validation croisée avec les paramètres suivants : 8 neurones dans la couche cachée, vitesse d'apprentissage =0.1, terme de mémoire =0.1 et nombre d'itérations = 4000. Dans ces conditions, les performances sont : ERC= (35.73 ± 0.79) %; ERP= (33.73 ± 0.92) %; RMSEC= (773.46 ± 4.02) ppm; RMSEP = (1341.44 ± 34.55) ppm. La valeur de ERP obtenue ici n'est pas suffisante pour une quantification même dans le cas présent de mesures sur site. On remarque sans surprise que les performances de l'ANN pour quantifier le plomb à des concentrations inférieures à 10 000 ppm étaient meilleures dans le cas de l'analyse du site SLM seul. La diversité des matrices influence donc les résultats des analyses.

Si on s'intéresse à présent à la quantification du plomb pour les échantillons des sites SEB et ME seulement (on exclut le site SLM), on a affaire à 50 échantillons répartis de la façon suivante : 34 en apprentissage, 10 en validation et 6 en test. Dans ce cas, le meilleur modèle ANN pour quantifier le plomb est obtenu pour : nombre de neurones dans la couche cachée =3, vitesse d'apprentissage =0.1, terme de mémoire = 0.1 et 9000 itérations. Les performances du modèle retenu sont rassemblées dans le Tableau 2-22:

	Lot d'apprentissage	Lot de validation	Lot de test
R^2	0,96	0,97	0,85
Q^2	0,96	0,94	0,84
ER (%)	22	22	14
RMSE (ppm)	550	510	820

Tableau 2-22 Performances du modèle ANN pour quantifier le plomb sur les sites ME et SEB.

La Figure 2-34 donne le détail des performances du modèle ANN pour chaque valeur de concentration en plomb. On remarque que pour les sites ME et SEB la concentration en plomb reste inférieure à 8 000 ppm.

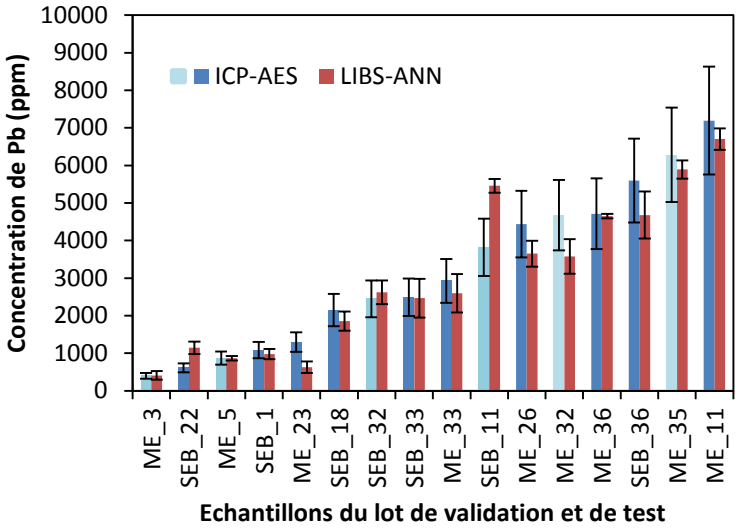


Figure 2-34 Concentration du plomb (ppm) obtenue par ICP-AES pour le lot de validation (bleu foncé) et pour le lot de test (bleu clair) et valeurs obtenues par LIBS-ANN (rouge). Les barres d'erreur sur les valeurs de l'ICP-AES représentent les 20% de tolérance et celles des valeurs LIBS-ANN indiquent les écarts-types calculés sur 5 répétitions du calcul ANN.

On en déduit que pour ces deux sites (ME et SEB), l'analyse du plomb par LIBS-ANN est tout à fait satisfaisante car les matrices sont assez voisines d'après le diagramme ternaire. Cependant, le modèle ANN développé sur la base de ces échantillons présente une valeur limite de concentration de plomb qui est telle qu'il ne sera pas possible de quantifier un échantillon ayant une concentration en plomb supérieure à 8 000 ppm.

En conclusion, en ce qui concerne l'analyse du plomb, nous avons démontré que d'une part, il était nécessaire de trier les échantillons en fonction de leur matrice via leur représentation dans un diagramme ternaire afin d'appliquer un modèle spécifique ME+SEB indépendant puis de trier les échantillons du site SLM en fonction d'une concentration seuil fixée ici à 10 000 ppm. Par ailleurs, en examinant de plus près les échantillons du site SLM sur le diagramme ternaire, on remarque que tous les échantillons qui correspondent à un pourcentage en minerai inférieur à 10% ont systématiquement une concentration en plomb inférieure à 10 000 ppm comme on peut le voir sur la Figure 2-35. Par conséquent, on peut choisir comme critère de tri des échantillons soit la valeur de 10% pour la concentration en minerai, soit la valeur de 10 000 ppm pour la concentration en plomb.

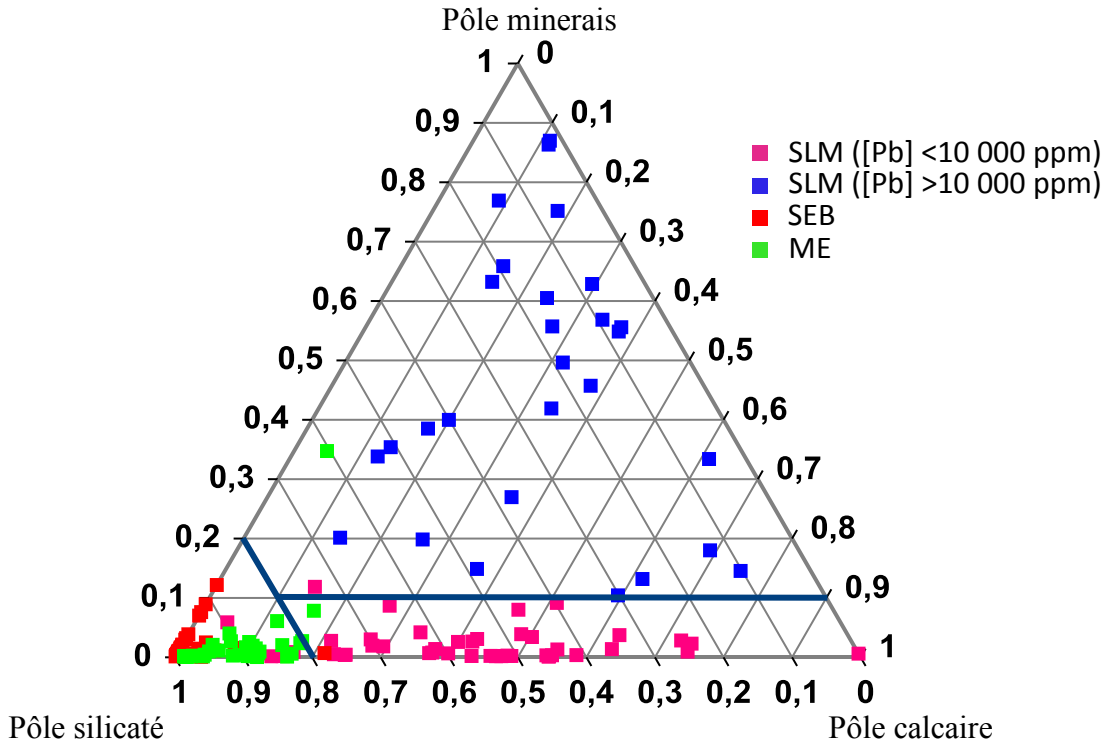


Figure 2-35 Représentation des échantillons de sol provenant de trois sites – SLM, ME et SEB – dans un diagramme ternaire.

En résumé, pour quantifier le plomb dans un échantillon inconnu, il est nécessaire de travailler en plusieurs étapes avec plusieurs modèles ANN.

- Etape 1 : Un premier modèle ANN que l'on notera ANN(1) à trois sorties permet de placer l'échantillon dans le diagramme ternaire.

- Etape 2 : Si l'échantillon se trouve à proximité du pôle silicaté (plus de 80% de sa composition due à la caractéristique silicatée), on applique alors un modèle ANN quantitatif à une seule sortie, la concentration en plomb. Ce modèle sera noté ANN(2) et correspond sur le diagramme de la Figure 2-35 au petit triangle qui englobe les échantillons de ME et de SEB.

- Etape 3 : Si l'échantillon n'est pas proche du pôle silicaté (au-delà de la limite des 80%) on vérifie d'après le diagramme ternaire si on doit le considérer plutôt riche en minerai (%minerai>10%) ou l'inverse.

- Etape 4 : Si l'échantillon se situe dans le diagramme ternaire au-delà de la limite des 10% de minerai, on applique alors un modèle ANN que l'on notera ANN(3) à une seule sortie donnant la concentration en plomb dans une gamme de concentrations supérieures à 10 000 ppm.

- Etape 5 : Si l'échantillon se situe dans le diagramme ternaire en dessous de la limite des 10% de minerai, on applique alors un modèle ANN que l'on notera ANN(4) à une seule sortie donnant la concentration en plomb dans une gamme de concentrations inférieures à 10 000 ppm.

On a donc établi un protocole complet d'analyse du plomb qui met en jeu 4 modèles ANN selon les cas de figure. Cette démonstration d'une analyse quantitative à partir de plusieurs sites a été faite pour le plomb à titre d'exemple. Signalons enfin que l'ordre des étapes citées précédemment peut être inversé ; on peut ainsi commencer par un classement des échantillons en fonction de la concentration en minerai puis classer les échantillons très peu riches en minerais selon leurs concentrations relatives en matrices silicatée et carbonatée. Les résultats obtenus en modifiant l'ordre des étapes de traitement sont tout à fait comparables à ceux qui ont été présentés ici. La même méthodologie peut bien évidemment être appliquée pour tous les autres éléments à analyser. En particulier, elle a donné des résultats – non présentés ici – tout à fait satisfaisants pour l'analyse quantitative du silicium, élément majeur de la plupart des sols.

2.3.7 PLS-ANN

Nous venons de démontrer la pertinence de l'ANN pour l'analyse quantitative de données LIBS dans le cas de mesures d'échantillons de sol sur site. Cependant, la sélection des données d'entrée reste un point qui n'a pas encore été complètement traité et celle-ci est faite au cas par cas selon l'appréciation de l'analyste. Pour aller plus loin dans cette discussion du choix des données d'entrée de l'ANN, nous avons exploré l'intérêt de compresser les données LIBS à l'aide d'une régression PLS puis d'injecter un très petit nombre de données résultant de la PLS dans l'ANN. Le couplage des techniques PLS et ANN sera appelé PLS-ANN.

2.3.7.1 Quantification d'un élément majeur : le calcium

Dans le cadre de nos travaux, chaque spectre LIBS contient 23988 points dans la gamme spectrale 200-896 nm sachant que chaque spectre est la moyenne de 25 spectres enregistrés lors de 25 répétitions de la mesure sur la surface de l'échantillon. Nous donnons ici des résultats de traitements PLS d'une part et PLS suivi d'un calcul d'ANN d'autre part.

- Pour le traitement PLS, on a divisé les échantillons en deux lots que l'on appellera des classes : classe 1 pour l'apprentissage et classe 2 pour le test.
- Pour l'étude PLS-ANN. les échantillons de la classe 1 définie précédemment sont divisés entre un lot d'apprentissage et un lot de validation. Ainsi, pour le lot de validation de l'ANN, il s'agit d'échantillons qui ont servi à l'apprentissage du modèle PLS mais pas à celui de l'ANN. Enfin, les échantillons de la classe 2 définie précédemment constituent le lot de test de l'ANN.

La répartition des échantillons est résumée dans le Tableau 2-23.

PLS	PLS_ANN	SLM_2011	SLM_2012	ME	SEB
Classe 1	Lot d'apprentissage	32	15	20	24
	Lot de validation	5	4	10	5
Classe 2	Lot de test	8	4	5	6

Tableau 2-23 Répartition en trois lots des échantillons de sol provenant des différentes campagnes

2.3.7.1.1 Mesures quantitatives du calcium par PLS

On construit un modèle PLS à partir des échantillons de la classe 1. Par la méthode de validation croisée interne (LOO), on obtient un modèle à 7 composantes dont les paramètres sont indiqués dans le Tableau 2-24. Les données spectrales sont centrées et réduites (on soustrait la moyenne et on divise par la variance). Cette normalisation est effectuée à l'aide de l'option UV du logiciel SIMCA-P+.

A	Valeurs propres	R2Y	R2Y(cum)	Q^2(LOO)	Q2(cum)
1	32,2	0,507	0,507	0,466	0,466
2	43,4	0,176	0,683	0,338	0,647
3	6,37	0,0857	0,769	0,18	0,711
4	5,31	0,0488	0,818	0,0725	0,732
5	1,37	0,118	0,936	0,498	0,865
6	0,944	0,0363	0,972	0,262	0,901
7	0,469	0,0186	0,991	0,0104	0,902

Tableau 2-24 Résultats du calcul PLS à partir des 115 spectres LIBS du lot de calibration (classe 1) contenant 23988 valeurs chacun.

Une fois le modèle construit à l'aide de la classe 1 uniquement, on calcule les concentrations en calcium pour les 2 classes (1 et 2), et on obtient les résultats présentés dans le Tableau 2-25.

	Classe 1	Classe 2
Erreur relative (%)	13	34
RMSE (%)	0,9	3,0

Tableau 2-25 Performances de la PLS dédiée à l'analyse du calcium.

On constate que l'erreur relative de prédiction pour les échantillons de la classe 2 est supérieure à 34%, ce qui est très supérieur à la valeur de 20% acceptable. C'est pour cela que la PLS n'a pas été retenue pour l'analyse LIBS quantitative des sols et que nous lui avons préféré le modèle ANN, apte à prendre en compte des non-linéarités. Cependant, le tableau 2-24 révèle que les 7 premières composantes principales permettent d'expliquer plus de 99% de la variance de la matrice de données. Il semble donc tout à fait envisageable de compresser les 23988 valeurs initiales de chaque spectre LIBS à seulement 7 valeurs qui correspondent simplement aux scores calculés en PLS pour les 7 premières composantes principales.

2.3.7.1.2 Mesures quantitatives du calcium par PLS-ANN

Les échantillons de la classe 2 constituent à présent le lot de test (cf. tableau 2-23). Ils sont donc inconnus et exploités uniquement a posteriori. La classe 1 est quant à elle divisée en deux lots afin de constituer le lot de calibration et le lot d'étalonnage de l'ANN comme décrit dans le Tableau 2-23. On utilise les scores des 7 premières composantes principales du calcul PLS précédent comme données d'entrées du calcul ANN. Le modèle est conçu à partir des échantillons du lot de calibration puis validé à l'aide des échantillons du lot de validation. Tous ces échantillons étant dans la classe 1 qui a servi à calculer le modèle PLS, les 7 scores de chaque échantillon dont connus. En revanche, pour les échantillons du lot de test (classe 2), il est nécessaire de les projeter dans le nouvel espace à 7 dimensions défini par le calcul PLS afin de pouvoir ensuite les exploiter (via leurs 7 scores chacun) pour évaluer a posteriori le modèle ANN.

On construit donc un réseau de neurones à 3 couches.

- Couche d'entrée : 7 entrées, les scores des 7 premières composantes principales
- Couche de sortie : 1 seul neurone donnant la concentration en calcium

Après validation croisée, les paramètres optimums pour ce modèle ANN sont:

- Nombres de neurones dans la couche cachée = 4
- vitesse d'apprentissage = 0.05
- terme de mémoire = 0.1
- nombres d'itérations = 14000

Les performances du modèle ANN optimal sont données dans la Tableau 2-26 ainsi que les résultats obtenus après application de la procédure de Y-randomization. On constate que le modèle retenu a une réelle signification statistique car les performances sont nettement supérieures pour le modèle optimisé que lorsqu'on applique la procédure de Y-randomization. De plus, le modèle optimisé est tout à fait satisfaisant puisqu'il fournit des prédictions de concentration avec des erreurs relatives inférieures à 15% pour les échantillons des lots de calibration et de validation.

	Modèle optimum	Y_Random
R^2c	0,98	0,55
R^2v	0,83	0,01
Q^2c	0,88	0,52
Q^2v	0,81	-1,96
ERC (%)	10	77
ERV (%)	15	284
RMSEC (%)	1	7
RMSEV(%)	1	12

Tableau 2-26 Performances du meilleur modèle ANN pour l'analyse du calcium et celles issues de la procédure de Y-Randomization. Résultats donnés pour les lots de calibration et de validation.

La Figure 2-36 montre une comparaison entre les résultats du calcul ANN basé sur les scores de la PLS et les valeurs de référence (ICP-AES) pour chaque valeur de concentration pour les

échantillons du lot de test (échantillons de la classe 2). Les performances moyennes pour le lot de test sont données quant à elles dans le **Tableau 2-27**.

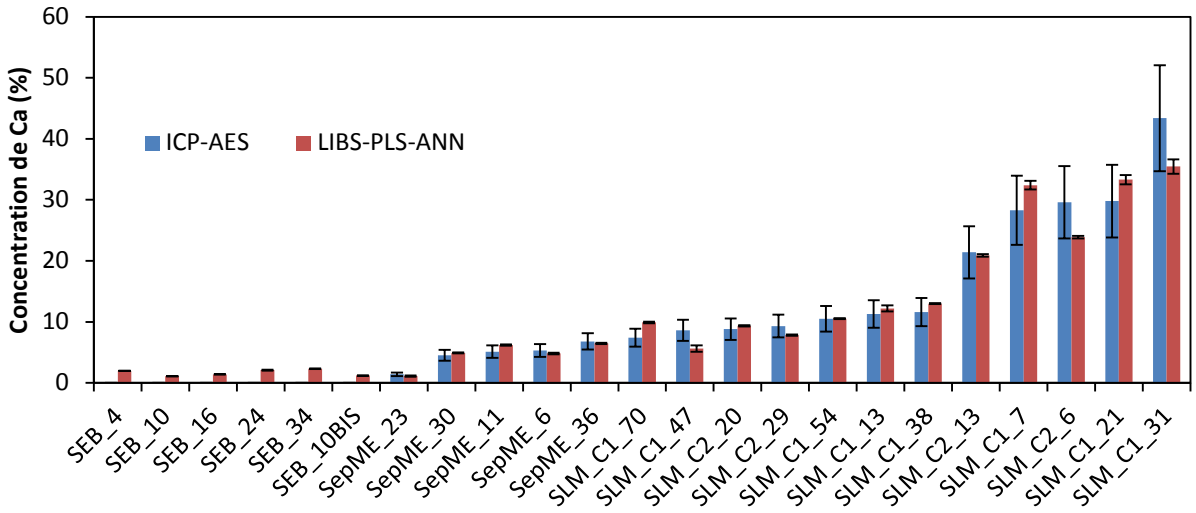


Figure 2-36 Concentrations en calcium des échantillons du lot de test. Bleu : ICP-AES, rouge : LIBS-PLS-ANN. Les barres d'erreur sur les résultats ICP-AES représentent l'erreur relative à 20% prise comme tolérance et celles sur les résultats LIBS-PLS-ANN représentent l'écart-type sur 5 répétitions du calcul ANN.

R^2T	Q^2T	ERT (%)	RMSET (%)
0.98	0.98	14	3

Tableau 2-27 Performances du meilleur modèle ANN pour l'analyse du calcium pour les échantillons du lot de test.

En conclusion, pour le calcium qui est un élément majeur de la matrice de sol, une compression des spectres LIBS a été réalisée par PLS permettant de passer de 23988 à seulement 7 valeurs par spectre. Le calcul ANN appliqué à ces 7 valeurs d'entrée par échantillon a donné des performances très satisfaisantes avec une erreur relative de prédiction inférieure à 15% pour les échantillons du lot de test. On a ainsi démontré que la sélection des données d'entrée de l'ANN pouvait, dans ce cas précis, être effectuée de façon mathématique, sans nécessiter d'intervention de l'analyste ni de référence à une base de données.

2.3.7.2 Quantification d'un élément mineur, le cuivre par PLS-ANN

Transposons à l'analyse quantitative d'un élément mineur la méthode introduite juste avant dans le cadre de l'analyse du calcium. Là encore, l'objectif de la compression par PLS est de ne pas avoir à effectuer une sélection arbitraire des raies dans le spectre LIBS. Notons au passage que le cuivre se trouvant en faible concentration, on ne détecte dans le spectre LIBS que les raies à 324.7 nm et à 327.4 nm. Reste donc à voir si une compression des données par PLS peut conduire à de bonnes performances.

Les échantillons de sols ont été répartis en 3 lots : 82 en apprentissage, 15 en validation et 7 en test. Le nombre de variables est 23988 correspondant à l'ensemble du spectre LIBS sur la

gamme (200.33-895.717) nm. Le Tableau 2-28 présente les résultats relatifs aux 5 premières composantes principales. On note que les composantes 2 et 3 ne sont pas significatives (Q^2 (LOO) < 0,05) et qu'il existe une faible corrélation entre le modèle PLS et la concentration du cuivre.

A	Valeurs propres	R2Y	R2Y(cum)	Q2 (LOO)	Q2(cum)
1	43,5	0,142	0,142	0,0901	0,0901
2	8,17	0,125	0,267	0,0113	0,1
3	4,47	0,241	0,507	-0,0342	0,0696
4	2,74	0,327	0,834	0,483	0,519
5	4,93	0,048	0,882	0,176	0,604

Tableau 2-28 Résultats du calcul de PLS à partir de 82 spectres du lot de calibration comptant 23988 valeurs chacun.

En observant la projection des vecteurs propres (loadings) des deux premières composantes (1 et 2) sur la Figure 2-37, on vérifie bien les deux raies du cuivre jouent un rôle essentiel dans le calcul de régression. Ceci confirme que la PLS permet de révéler efficacement les variables du spectre LIBS qui sont corrélées à la variable de sortie Y, à savoir ici à la concentration en cuivre, et ce, sans connaissance a priori des raies spectrales à sélectionner. Par contre on remarque aussi une corrélation des raies du plomb avec les raies du cuivre, et ceci peut conduire à une interférence sur la prédiction de la concentration de cuivre.

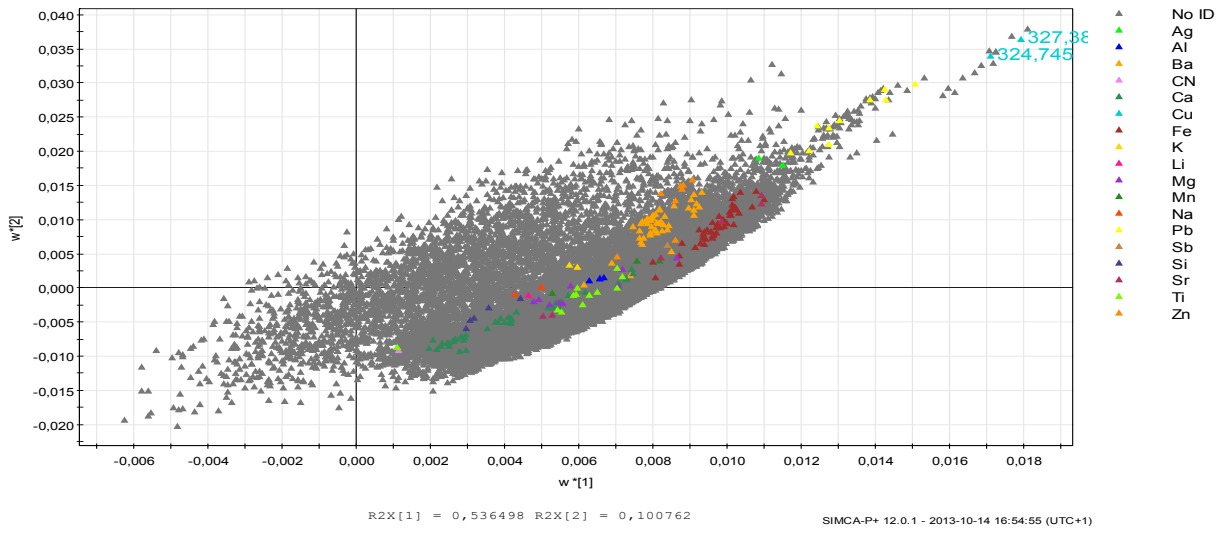


Figure 2-37 Loadings dans le plan (1,2) du modèle PLS dédié à l'analyse du cuivre.

Pour la suite, on utilise les 5 scores du Tableau 2-28 comme entrées du modèle ANN. Le Tableau 2-29 présente le résultat des erreurs de prédiction par PLS-ANN. On note que la compression du spectre total n'a pas abouti à une prédiction satisfaisante (ERV>>20%). Peut-être faut-il réduire la bande spectrale avant la compression par PLS pour que le couplage PLS-ANN soit plus significatif. En conclusion, PLS-ANN est plus performant pour les éléments majeurs que pour les mineurs et les éléments peu émissifs.

	Lot d'apprentissage	Lot de validation
ER (%)	33	60
RMSE (ppm)	37	182

Tableau 2-29 Résultats d'un modèle PLS-ANN dédié à l'analyse quantitative du cuivre. Nombre de neurones dans la couche cachée = 3, vitesse d'apprentissage = 0.075, terme de mémoire = 0.1, nombre d'itérations = 12000.

Pour une comparaison plus approfondie entre les deux approches, i.e. le calcul ANN à partir de données sélectionnées de façon experte à partir du NIST et l'approche PLS-ANN qui consiste à injecter dans l'ANN les scores d'un calcul PLS préalable, nous avons choisi de traiter les données d'un seul site géologique. Il s'agit du site SLM pour lequel les résultats du calcul ANN direct sont présentés au paragraphe 2.3.3 du chapitre 2. Une comparaison stricte est possible en utilisant dans les deux calculs – ANN direct et PLS ANN – les mêmes échantillons dans chaque lot. Ainsi avec 18 échantillons dans le lot d'apprentissage, le calcul PLS donne 3 premières composantes significatives et les résultats présentés dans le Tableau 2-30. Les données sont plus homogènes et donc la compression est mieux que lors de l'étude des données sur les différents sites.

A	**Valeurs propres**	**R2Y**	**R2Y(cum)**	**Q2(LOO)**	**Q2(cum)**
1	10,3	0,338	0,338	0,184	0,184
2	2,21	0,313	0,651	0,187	0,336
3	0,565	0,318	0,969	0,431	0,623

Tableau 2-30 Résultats du calcul de PLS à partir de 18 spectres du lot de calibration comptant 23988 valeurs chacun.

En utilisant les scores des trois premières composants comme entrées de l'ANN on obtient les résultats présentés dans le tableau 2-31.

	Lot d'apprentissage	Lot de validation	Lot de test
ERC (%)	18	28	28
RMSE (ppm)	30	57	41

Tableau 2-31 Performances de modèles ANN dédié à l'analyse quantitative du cuivre du site SLM. Nombre de neurones dans la couche cachée = 3,vitessed'apprentissage = 0.075, terme de mémoire = 0.1, nombre d'itérations = 2000.

D'après le Tableau 2-31 et le Tableau 2-30, on remarque sans surprise que la procédure PLS-ANN est plus performante dans le cas d'un seul site géologique que lorsque trois sites différents sont à prendre en compte. De plus, d'après la Figure 2-38, si on tient compte de la tolérance de 20% qui a été fixée pour juger une prédiction acceptable, on remarque que le traitement PLS-ANN révèle 6 échantillons qui ne respectent pas le critère de tolérance de 20% tandis qu'il n'y en a qu'un seul lorsqu'on applique l'ANN directement aux données sélectionnées d'après le NIST. Ce résultat nous conduit à conclure que, dans ce cas précis de l'analyse du cuivre et pour la gamme de concentrations présentée ici, la méthode basée sur une sélection experte de données significatives dans le spectre LIBS semble plus performante que la méthode PLS-ANN.

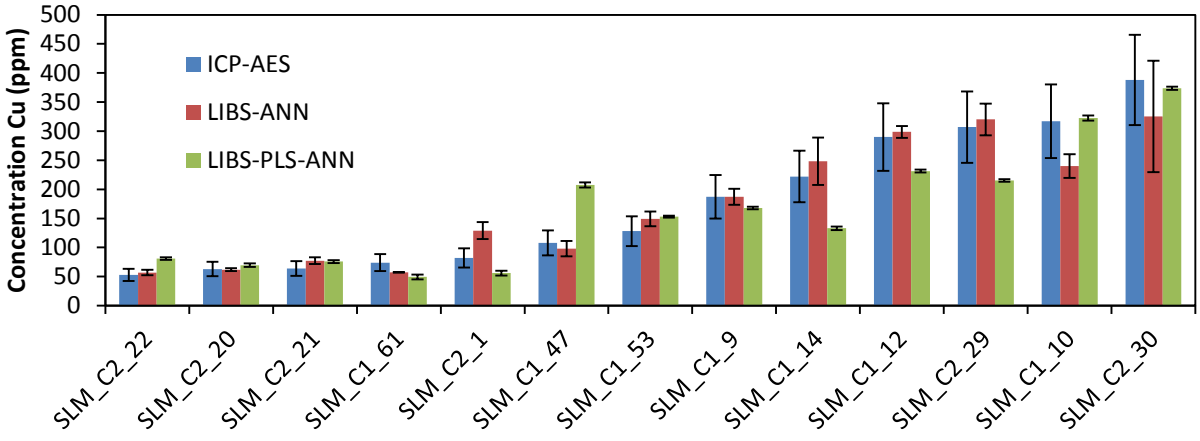


Figure 2-38 Comparaison entre la concentration de cuivre de référence, la concentration prédite en LIBS-ANN et la concentration par LIBS-PLS-ANN pour les échantillons des lots de validation et de test. un écart de 20% autour de la concentration de référence, et pour les données LIBS-ANN et LIBS-PLS-ANN, l'écart-type sur 5 répétitions du calcul d'ANN.

2.4 **Transfert industriel**

Ce travail de thèse s'inscrit dans le cadre d'un projet industriel du nom de Calipso. Ce projet, porté par l'entreprise IVEA a été lauréat de l'appel à projets éco-industries en 2010 et financé par l'ADEME de janvier 2011 à janvier 2013. Il impliquait également le BRGM en tant que spécialiste en géosciences. Ce projet de recherche a ensuite été prolongé, financé par la société IVEA pour l'année 2013. L'objectif du projet Calipso était la mise au point et la validation d'un protocole d'analyse quantitative élémentaire sur site par le biais d'une mesure LIBS et d'un traitement des données par réseaux de neurones artificiels (ANN). Pour atteindre cet objectif, nous avons tout d'abord exploité le logiciel de calcul scientifique Igor Pro afin de réaliser les calculs d'ANN au laboratoire et, dans un second temps, nous avons transféré l'intégralité de l'algorithme de calcul ANN à l'entreprise IVEA. Cet algorithme a ainsi été implanté dans le logiciel commercial Analibs, en langage C. Pour cela, nous avons travaillé avec IVEA par itérations successives afin de tester le logiciel par des comparaisons répétées entre les résultats obtenus et ceux provenant du logiciel scientifique Igor Pro. Nous avons dû rechercher des erreurs de programmation et les corriger et définir aussi des conditions initiales et des critères d'arrêt. Nous avons finalement obtenu un module offrant des performances similaires à celles du logiciel scientifique Igor Pro mais implanté dans le logiciel Analibs et dédié à l'analyse LIBS. Nous avons conçu le module ANN du logiciel Analibs de manière à ce que celui-ci puisse être utilisé en routine pour des analyses LIBS et de façon simple mais aussi de manière à ce qu'un expert en ANN puisse s'en servir pour rechercher les meilleures optimisations. Le module ANN se décline donc en plusieurs parties qui sont décrites ci-après.

- Partie 1 : pré-apprentissage-ANN

Cette partie concerne tout d'abord la sélection des spectres LIBS enregistrés. On choisit aussi l'analyte et on associe à chaque spectre LIBS la valeur de concentration de référence obtenue par ICP-AES. On classe les concentrations par ordre croissant afin de faciliter le choix des échantillons à répartir dans les lots de calibration, validation et test. Cette répartition dans les trois lots est réalisée manuellement, à l'appréciation de l'utilisateur. La dernière étape consiste à choisir les variables d'entrée. On fait appel aux raies persistantes non seulement pour l'analyte mais aussi pour des éléments majeurs de la matrice. Notons cependant que le choix des variables d'entrée n'est pas simple ni totalement figé et fera sans aucun doute l'objet d'améliorations lors de travaux futurs. Une fois les variables sélectionnées et les échantillons répartis par lots, on procède à des normalisations à la fois pour les données d'entrées et pour les valeurs de concentration en sortie sachant que la fonction d'activation sigmoïde utilisée au niveau de chaque neurone ne peut fournir qu'une réponse comprise entre 0 et 1.

- Partie 2 : apprentissage-ANN

Les poids de l'ANN ont des valeurs initiales fixées de façon aléatoire et la première approche consiste à fixer les paramètres de l'ANN, à savoir le nombre de neurones dans la couche cachée, la vitesse d'apprentissage, le terme de mémoire et le nombre d'itérations. On lance alors de façon itérative l'algorithme « feed-forward » qui permet de calculer la valeur de la concentration à partir des données d'entrée puis l'algorithme de « backward propagation of error » qui permet de modifier les valeurs de poids de l'ANN à partir du calcul de l'erreur entre la concentration prédite et la valeur de référence. A la fin du calcul, on affiche les performances, à savoir les erreurs relatives moyennes et les valeurs de RMSE pour les lots de calibration et de validation. Notons que seuls les échantillons du lot de calibration sont utilisés pour fabriquer le modèle ; ceux du lot de validation sont exploités a posteriori pour évaluer les performances du modèle.

Pendant la phase d'étude, on recherche les meilleurs paramètres de l'ANN et on est donc amené à faire varier tour à tour la valeur du nombre de neurones dans la couche cachée, la vitesse d'apprentissage, le terme de mémoire et le nombre d'itérations. Ce calcul systématique est nécessaire pour rechercher les conditions optimales. Chaque paramètre peut ainsi varier entre une valeur minimum et une valeur maximum prédéfinies par l'analyste avec un pas lui aussi sélectionné par l'analyste. On affiche sur un graphe pour chaque pas le résultat des erreurs pour les deux lots de calibration et de validation. Cette procédure permet de trouver les meilleurs paramètres et de stopper l'apprentissage au bon moment. En effet, on sait que trop d'itérations peuvent entraîner un sur-apprentissage et il est donc essentiel de rechercher le nombre d'itérations permettant de réduire à la fois les erreurs de calibration et de prédiction. Finalement, le meilleur modèle peut être sauvegardé pour une utilisation ultérieure. Il contient la sélection des données d'entrée et leur normalisation, la valeur de concentration maximale qui permet de convertir la réponse de l'ANN entre 0 et 1 en valeur de concentration ainsi que tous les poids qui permettent une utilisation directe pour prédire la concentration d'un échantillon inconnu.

- Partie 3 : utilisation directe de l'ANN

On introduit les données d'un échantillon inconnu dans l'ANN qui a été sélectionné. Le calcul direct permet de fournir une valeur de concentration. Ainsi, l'ANN est directement utilisable pour fournir sur site des valeurs de concentration en temps réel.

Aujourd'hui, la société IVEA dispose d'un logiciel qui permet de rechercher le meilleur modèle ANN pour un problème posé sans nécessité d'avoir recours à un autre logiciel scientifique. Le logiciel a été validé et le calcul peut naturellement être appliqué à n'importe quel type d'échantillon, moyennant des ajustements mineurs. Il représente donc une plus-value importante pour les produits commercialisés par IVEA.

2.5 **Conclusion**

L'analyse LIBS quantitative des échantillons de sols sur site s'est avérée extrêmement complexe. Nous avons adopté une méthode basée sur un étalonnage et mettant en œuvre des réseaux de neurones artificiels. La phase d'étalonnage a nécessité le prélèvement de sols réels sur différents sites géologiques caractérisés par des matrices différentes. Tous les échantillons ont fait l'objet de mesures ICP-AES en laboratoire afin de pouvoir construire l'étalonnage. Elles ont été réalisées par le BRGM, partenaire du projet. Les échantillons préparés sous forme de pastilles ont été analysés par LIBS directement sur site. Les analyses ont été conduites en collaboration entre le LOMA, IVEA, le BRGM et l'ADEME avec un système LIBS commercialisé par la société IVEA. Tout ce travail a permis de rassembler les données LIBS et les données ICP-AES pour une série de sols réels lors de différentes campagnes. A partir de la base de données du NIST, nous avons démontré que non seulement les raies spectrales de l'analyte devaient être sélectionnées mais aussi des raies relatives à la matrice du sol. Une erreur relative de prédiction inférieure à 20% a été obtenue pour des éléments mineurs et pour des éléments majeurs à partir de mesures sur 3 sites. Dans le cas particulier du plomb caractérisé par une très large gamme de concentrations, plusieurs modèles ANN quantitatifs ont été nécessaires pour atteindre des performances acceptables. Le choix du modèle le plus adapté repose sur une reconnaissance de la matrice qui peut être plutôt silicatée, plutôt carbonatée ou plutôt riche en minerais. Il peut aussi reposer sur la détermination d'une valeur seuil de concentration.

Cependant, le choix des données d'entrée de l'ANN reste l'étape la plus délicate. C'est pourquoi, au-delà de la sélection des quelques données d'entrée de façon experte, nous avons démontré qu'une compression des données par PLS était également satisfaisante dans le cas d'éléments majeurs comme le calcium. En revanche, dans le cas d'un élément mineur comme le cuivre la PLS a été appliquée d'abord à l'ensemble du spectre LIBS puis à des bandes spectrales plus étroites et l'on a pu observer que pour les faibles concentrations il était avantageux de réduire la bande spectrale avant de lancer l'analyse. Ces observations laissent sans réponse tranchée les questions relatives au choix des meilleures données d'entrée et ouvrent de nouvelles voies d'investigation pour améliorer encore les analyses LIBS de milieux complexes comme les sols.

Chapitre 3. **Chimiométrie appliquée à la spectroscopie térahertz**

Dans ce chapitre, nous allons présenter nos travaux visant à introduire les outils de chimiométrie dans le champ disciplinaire de la spectroscopie térahertz. Notons que la chimiométrie est très peu utilisée dans ce domaine et il y a donc un gisement potentiel de résultats à mettre en exergue dans ce domaine. La stratégie que nous avons suivie est la suivante. Nous avons préparé en laboratoire des mélanges ternaires sous forme de pastilles avec du polyéthylène comme liant et nous avons enregistré les spectres THz en transmission afin de disposer des spectres d'absorbance de ces échantillons. Notre objectif était ici d'exploiter différents outils de chimiométrie pour en démontrer les avantages dans le cadre du traitement des données THz.

3.1 **La spectroscopie térahertz**

La spectroscopie térahertz est une spectroscopie très en vogue de nos jours notamment grâce à la capacité du rayonnement térahertz à détecter des matériaux explosifs ou des drogues, y compris à travers des emballages en papier ou polyéthylène ou encore à travers des parois en béton. Le domaine térahertz, également appelé infrarouge lointain, est l'intervalle spectral situé entre 100 GHz et 10 THz, dont les longueurs d'ondes sont comprises entre 30 µm et 3 mm. Le domaine térahertz est donc situé entre l'infrarouge et les micro-ondes comme indiqué sur la Figure 3-1 et bénéfice donc des avantages de ces deux domaines [114] La spectroscopie dans le domaine micro-onde s'appuie sur les propriétés de rotations moléculaires. Une molécule polaire par exemple est caractérisée par un dipôle qui tourne à une fréquence bien déterminée, et donc elle absorbera préférentiellement les photons à cette fréquence. Mais notons que certains effets ne sont observables qu'à des fréquences associées au domaine térahertz. C'est le cas par exemple pour détecter les hydrures (le gaz toxique H_2S) ou encore le dichlorométhane qui présente une fréquence caractéristique au-delà de 2,5 THz [115].

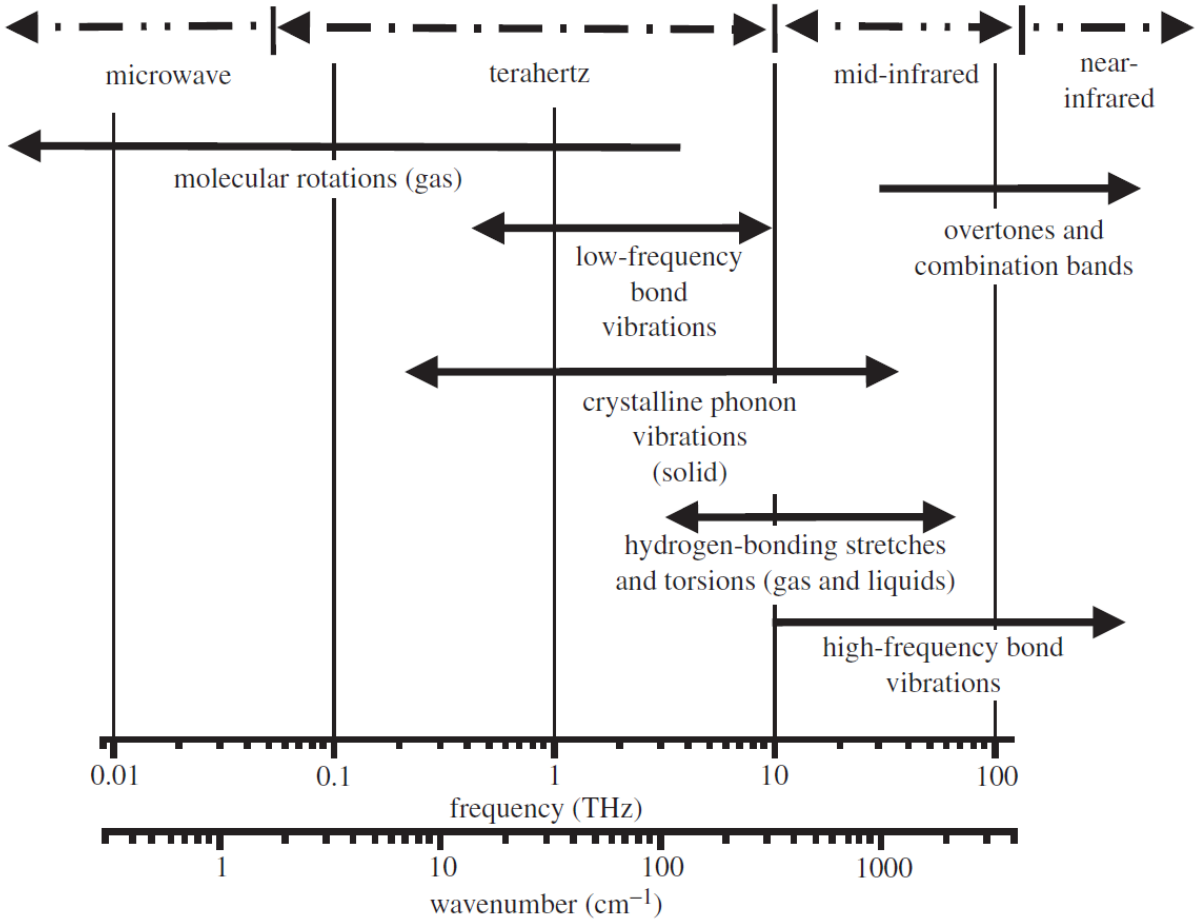


Figure 3-1 Spectre électromagnétique centré sur le domaine térahertz. Extrait de [115]

Par ailleurs, la plupart des vibrations des liaisons moléculaires présentent des fréquences caractéristiques dans le domaine infrarouge. Il y a différents types de vibrations comme le mode d'élongation associé à la modification de la longueur de la liaison moléculaire ou encore le mode de déformation associé à modification de l'angle entre deux liaisons, chacun étant associé à une fréquence caractéristique. Les modes de vibration qui interviennent dans le domaine THz sont relatifs à des liaisons faibles (liaison hydrogène) ou encore à un mouvement de torsion, c'est-à-dire au déplacement d'ensemble d'une partie de la molécule par rapport au reste. Ces « vibrations THz » sont caractéristiques des grosses molécules comme les molécules polycycliques et les molécules biologiques. Elles permettent par ailleurs d'accéder à des informations sur la structure cristalline et de discriminer des molécules chirales, ce qui n'est pas possible par spectroscopie infrarouge classique [114, 115]. Nous donnons ici une très brève description des techniques expérimentales mais le lecteur intéressé par la spectroscopie térahertz est invité à consulter l'excellent ouvrage intitulé « Optoélectronique térahertz » [114].

La Figure 3-2 représente le schéma typique d'un dispositif de spectroscopie térahertz dans le domaine temporel. Une lame semi-réfléchissante (BS) sépare le faisceau laser initial en deux faisceaux qui vont respectivement déclencher l'émission THz au niveau du dispositif émetteur et la prise de mesure au niveau =du dispositif détecteur. Une ligne à retard optique corrige la

différence entre les deux chemins optiques. L'émetteur et le détecteur sont constitués de photocommutateurs à semi-conducteurs [116]. Une impulsion THz est générée par interaction avec une impulsion laser ultracourte [114] puis est mesurée par échantillonnage en fonction du temps grâce au déplacement de la ligne à retard. Le temps d'acquisition est optimisé pour améliorer le rapport signal sur bruit, sachant que le signal mesuré est en fait le résultat d'une convolution entre le signal réel et la réponse du détecteur, elle-même étant le résultat de la convolution entre la réponse impulsionnelle du photocommutateur utilisé en réception et l'impulsion laser utilisée pour déclencher l'acquisition. En ajoutant un échantillon dans le trajet de l'impulsion THz entre l'émetteur et le détecteur et en déterminant le profil temporel de l'impulsion électromagnétique ultra-brève, on peut déterminer les caractéristiques de transmission de l'échantillon.

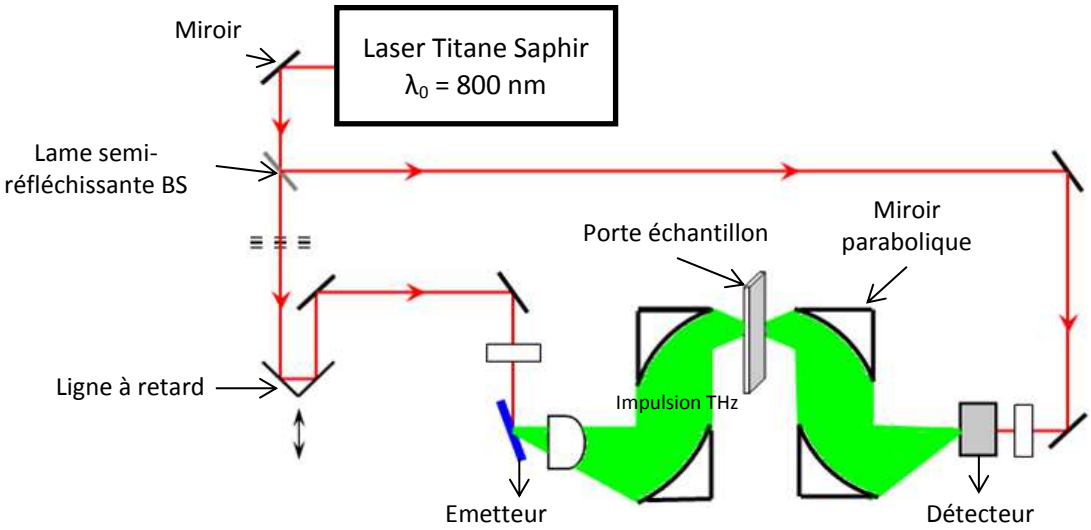


Figure 3-2 Schéma de principe d'un montage de spectroscopie térahertz dans le domaine temporel.

L'impulsion THz est ainsi mesurée une première fois en l'absence d'échantillon afin de fournir le signal de référence S_{ref} (t) – en pointillé sur la Figure 3-3 – puis en présence de l'échantillon S_{ech} (t) – en trait plein sur la Figure 3-3 – afin de pouvoir remonter aux propriétés physiques de l'échantillon. On note que des échos peuvent être présents sur le signal enregistré en présence de l'échantillon à cause des réflexions aux interfaces air-échantillon.

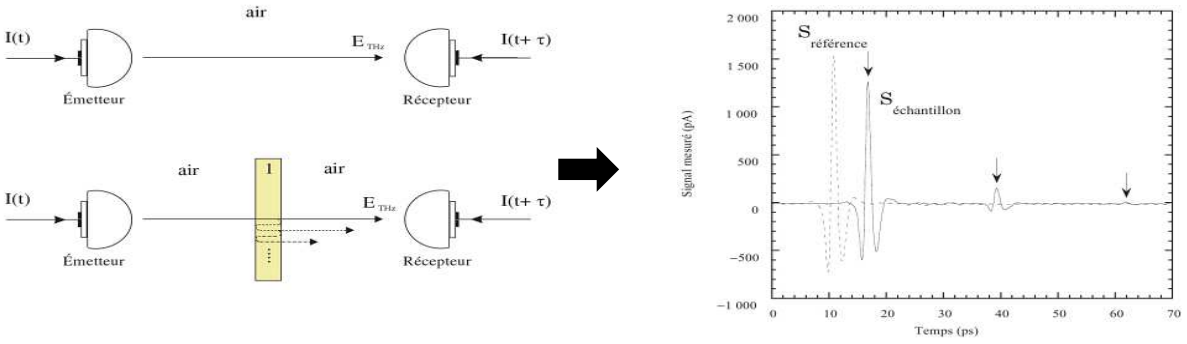


Figure 3-3 Schéma de principe d'une mesure de spectroscopie THz dans le domaine temporel – Extrait de [114]

Dans le cadre de nos travaux, chaque signal enregistré dans le domaine temporel puis nettoyé du bruit à l'aide d'un filtre Blackman-Harris 3 est ensuite traité par transformée de Fourier si bien qu'on obtient deux spectres, le spectre de référence et le spectre observé après passage à travers l'échantillon. La fonction de transfert complexe de l'échantillon est obtenue en effectuant le rapport des deux spectres :

$$T_{exp}(\omega) = \frac{S_{ech}(\omega)}{S_{ref}(\omega)} \quad \textbf{(3-1)}$$

On en déduit le spectre d'absorbance :

$$A = \log(T_{exp}(w)) = \log\left(\frac{S_{ech}(W)}{S_{ref}(W)}\right) \quad \textbf{(3-2)}$$

Alors qu'en LIBS, le spectre est donné en fonction des longueurs d'onde λ (nm) et qu'en spectroscopie infrarouge il est donné en fonction du nombre d'onde $\tilde{\nu}$ (cm^{-1}), on notera qu'en spectroscopie térahertz, on donne généralement le spectre en fonction de la fréquence ν (THz).

3.1.1 Les précautions expérimentales

Le polyéthylène (PE) est traditionnellement utilisé comme liant en spectroscopie infrarouge et en spectroscopie térahertz car cela permet d'améliorer le rapport signal-sur-bruit. Il faudra cependant normaliser les spectres de tous les échantillons en fonction de celui du PE [117]. Par ailleurs, la molécule d'eau présente de forts mouvements rotationnels dans la région THz si bien que l'humidité agit fortement sur l'absorption des échantillons. Sachant que les échantillons sont généralement préparés sous forme de pastilles à l'air libre, la surface de chaque pastille peut contenir une faible quantité d'eau, différente d'une pastille à l'autre, susceptible d'affecter les mesures de spectroscopie THz. Le porte échantillon est ainsi généralement accompagné d'un dispositif de pompage sous air sec ou sous azote sec pour minimiser ces contributions [118]. De plus, une étude sur l'effet du séchage en fonction de temps a été réalisée [117] et a révélé que la transmission de la théophylline diminuait de 9,52 (u.a.) après 10 minutes à 6,97 (u.a.) après 14 heures. Notons aussi que l'effet du séchage n'est pas le même tout le long du spectre THz. Par exemple pour un spectre de microcristalline de cellulose (MCC) mélangé avec le PE, le spectre a été divisé en trois régions : la région I (0 cm^{-1} <ω< 80 cm^{-1}) s'est révélée insensible à l'humidité, tandis que pour la région II (80 cm^{-1}< ω < 200 cm^{-1}), la transmission augmente avec le temps de séchage, et pour la région III (200 cm^{-1} < ω < 400 cm^{-1}) la transmission diminue en augmentant le temps de séchage. De plus, pour les deux région I et II, les auteurs ont remarqué que les spectres devenaient plus lisses après une heure de séchage [117]. En nous appuyant sur ces résultats, nous avons adopté un temps de séchage identique pour tous les échantillons dans le but de réduire le biais lié à l'influence de l'humidité.

3.2 Application de la chimiométrie aux spectres térahertz

Nous avons rédigé en 2013 une revue des applications récentes de la chimiométrie pour la spectroscopie térahertz dans un article intitulé «Review in terahertz spectral analysis» [47]. Plus tôt, en 2011, Shen [119] avait publié un article citant les différentes études quantitatives en pharmaceutique basées sur l'utilisation de la PLS en précisant cependant que les techniques chimiométriques n'avaient encore jamais été exploitées dans le cas des échantillons solides réels. Le premier usage de la chimiométrie pour des données THz remonte à 2003. Il s'agissait d'une application pharmaceutique ayant pour finalité de quantifier le paracétamol et l'aspirine dans un mélange de cellulose et de lactose [115]. Il a été démontré que la composition chimique affectait beaucoup les spectres THz, et tout particulièrement les formes cristallines comme celles du lactose et de la théophylline conduisent à des fortes raies dans le spectre THz, reliées au mouvement vibratoire de la structure moléculaire et à ses interactions avec le milieu solide. Par opposition, les milieux amorphes ainsi que ceux contenant des molécules présentant des poids moléculaires élevés comme l'amidon conduisent à de très faibles pics dans le spectre THz qui rendent d'autant plus pertinent le recours aux techniques d'analyse multivariées [117]. L'analyse quantitative par PLS est d'ores et déjà fréquemment utilisée en spectroscopie THz. A titre d'exemple, Hua et al. [120] ont détecté différents pesticides dans le riz et ont réussi à mener une analyse quantitative par PLS sur des spectres d'absorption THz dans la bande 0,5-1,6 THz avec une erreur relative moyenne inférieure à 5% pour un mélange d'imidaclopride dans le polyéthylène ainsi que dans le riz.

En analyse multivariée, il est usuel d'appliquer un prétraitement aux données expérimentales. Plusieurs types de prétraitements ont été testés sur les données THz comme le lissage et la normalisation, le calcul de la dérivée première ou de la dérivée seconde, ou encore l'application d'un filtre de type MSC, Savitzky Golay, ou SNV [10, 121]. Pour illustrer les avantages du prétraitement, notons que pour un mélange de 4 produits, à savoir la théophylline, le lactose et le MCC liés par de l'amidon, il a été rapporté une valeur de RMSEP (w/w) =0.074 (7.4 %) en mettant en œuvre une analyse PLS après un prétraitement de dérivée seconde [122]. Il n'y a pas de règle absolue pour choisir le prétraitement et il est nécessaire de les tester un à un.

3.3 Préparation des échantillons

Pour notre étude, nous avons sélectionné trois produits chimiques très différents : i) l'acide citrique qui se trouve dans le citron, ce produit est de la marque «Sigma-Aldrich» et de pureté ≥ 99.0 %. ii) le D (-) Fructose qui existe dans les fruits, ce produit est de la marque «Sigma Life Science » Lot ♯ SLBB6798V, la pureté est ≥ 99% et iii) le α-Lactose monohydrate qui se trouve dans le lait, ce produit est aussi de la marque «Sigma Life Science » Lot ♯ SLBD56375V, la pureté est ≥ 99%. La matrice utilisée est du polyéthylène « Aldrich Chemistry » Lot : MKBC9115V, de taille de particules (53-75) µm. Ces quatre produits se trouvent sous forme de poudre et leurs compositions chimiques sont décrites dans le Tableau 3-1.

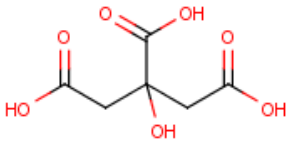
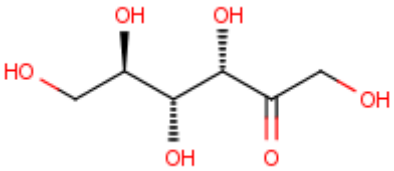
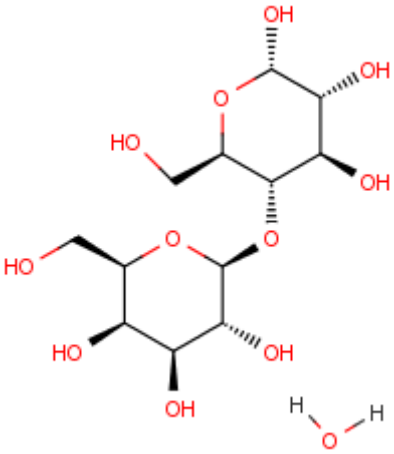
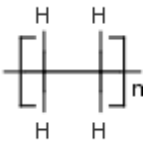
Composant	Formule	Structure
Acide Citrique	$C_6H_8O_7$	
D-(-) Fructose	$C_6H_{12}O_6$	
α-Lactose monohydrate	$C_{12}H_{22}O_{11} \cdot H_2O$	
Polyéthylène	$-(CH_2-CH_2)_n-$	

Tableau 3-1 Formules chimiques et structures des 3 produits (Acide Citrique, D-(-) Fructose, α-Lactose monohydrate) et de la matrice de polyéthylène.

L'objectif est de déterminer la quantité de chaque produit dans les différents échantillons par spectroscopie térahertz. Les produits, initialement des poudres, sont liés à l'aide de polyéthylène (PE) et préparés sous forme de pastilles. L'intérêt d'utiliser le polyéthylène comme liant est qu'il est transparent dans le domaine des fréquences térahertz [120]. Nous avons d'abord préparé 39 échantillons. Les échantillons A1 à A19 constituent la première série et les échantillons B1 à B20 la deuxième série. Tous les échantillons contiennent 80% en masse de PE comme liant. Sur les 20% restant, l'échantillon A1 contient 100% de Fructose, l'échantillon A6 contient 100% de Lactose et l'échantillon A11 contient 100% d'acide citrique comme on peut le vérifier sur la **Figure 3-4.** Notons que certains échantillons tels que A2 ou A8 par exemple représentent des mélanges binaires alors que d'autres comme ceux de la série B représentent des mélanges ternaires. Le diagramme ternaire présenté sur la **Figure 3-4** permet de connaître directement les concentrations relatives de chaque produit du mélange. Par exemple, l'échantillon B10 contient 10% en masse d'acide citrique, 50% en masse de fructose, 40% en masse de lactose. Chaque échantillon est préparé selon le protocole décrit ci-après qui consiste à préparer 1200 mg de mélange et donné ici dans le cas de l'échantillon B7 choisi à titre d'exemple.

1- pesage d'une masse de 240 mg de polyéthylène (PE) en poudre qui sera ensuite placée dans un mortier.
2- pesage d'une masse de 72 mg en fructose en poudre qui sera ajoutée au PE dans le mortier
3- broyage et homogénéisation de ce mélange à l'aide d'un pilon
4- pesage d'une masse 240 mg de polyéthylène (PE) en poudre qui sera ajoutée au mélange dans le mortier puis homogénéisée avec un pilon.
5- pesage d'une masse 72 mg de lactose en poudre qui sera ajoutée au mélange dans le mortier puis homogénéisée avec un pilon.
6- pesage d'une masse 240 mg de polyéthylène (PE) en poudre qui sera ajoutée au mélange dans le mortier puis homogénéisée avec un pilon.
7- pesage d'une masse 96 mg d'acide citrique en poudre qui sera ajoutée au mélange dans le mortier puis homogénéisée avec un pilon.
8- pesage d'une masse de 240 mg de polyéthylène (PE) en poudre qui sera ajoutée au mélange dans le mortier puis homogénéisée avec un pilon.
9- De ce mélange de masse totale 1200 mg et qui contient 80% en masse de PE et 20% en masse du mélange ternaire, on pèse deux masses de 400 mg et on fabrique deux pastilles à l'aide d'une presse manuelle sur laquelle on applique une pression de 8 tonnes pendant 1 minute.

Ce même protocole est appliqué pour la préparation des autres échantillons simplement en changeant les masses de fructose, lactose et acide citrique. Chaque échantillon est donc préparé sous forme de deux réplicas, à savoir deux pastilles identiques de 400 mg, afin de disposer de données statistiques.

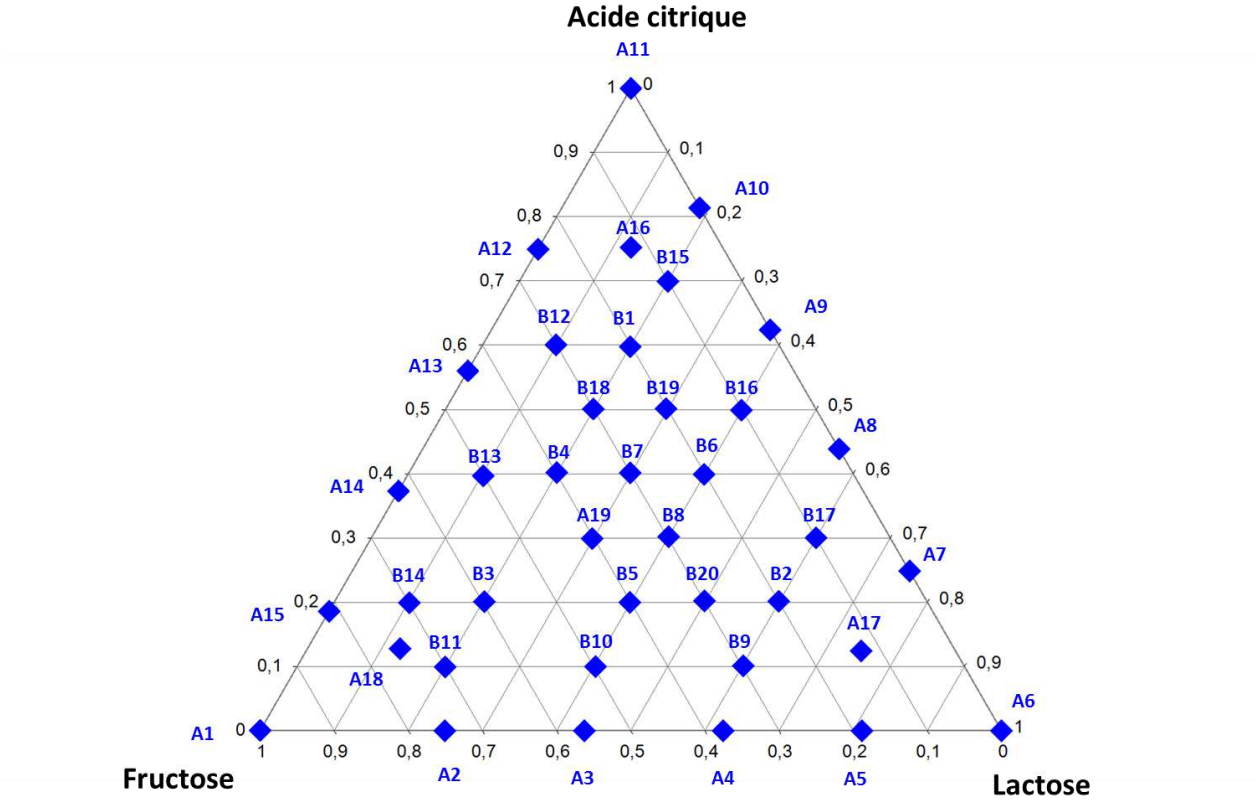


Figure 3-4 Diagramme ternaire présentant les concentrations relatives en Fructose, Lactose et acide citrique des échantillons préparés pour l'analyse en spectroscopie THz.

3.4 **Spectres d'absorbance**

Les spectres d'absorbance THz étudiés ici se situent dans la gamme spectrale 0-3 THz. Nous avons en effet observé qu'après 3 THz, les spectres n'étaient pas répétables et ne permettaient donc pas de faire des déductions fiables. Le polyéthylène (PE) est parfaitement transparent dans cette gamme spectrale comme le montre la courbe en bleu sur la Figure 3-5 et il est donc idéal comme liant. Notons que cette courbe a été obtenue pour une pastille de 400 mg de PE pur qui sert d'échantillon de référence. L'absorbance caractéristique du Fructose, obtenue grâce à l'échantillon A1, est décrite par la courbe en rouge sur la Figure 3-5. De même, celle du lactose, obtenue pour l'échantillon A6, est donnée en vert et celle de l'acide citrique, obtenue pour l'échantillon A11, est donnée en violet. On remarque que l'absorbance augmente globalement avec la valeur de la fréquence. De plus, nous avons vérifié la validité de la loi de Beer-Lambert [123], à savoir l'augmentation de l'absorbance avec la concentration de l'élément absorbant. Enfin, nous avons observé l'effet additif des bandes d'absorption lorsqu'on mélange les différents produits.

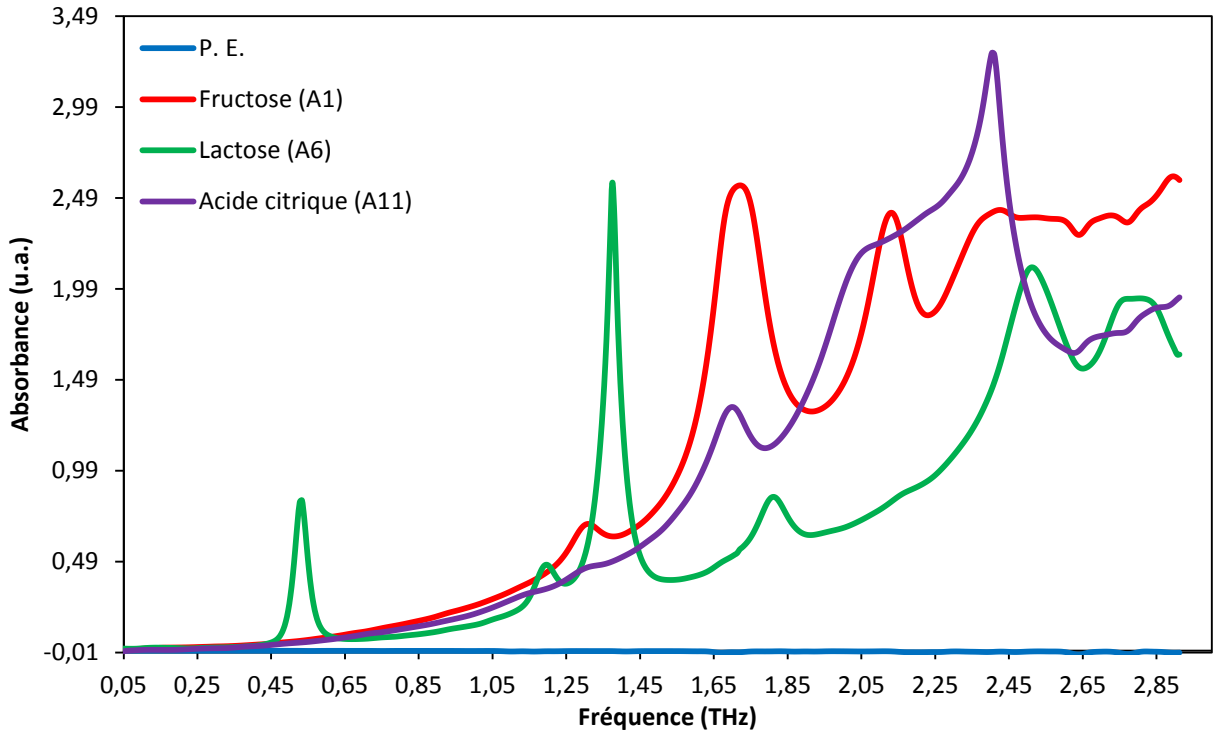


Figure 3-5 Spectres d'absorbances du polyéthylène (PE) en bleu, du fructose en rouge, du lactose en vert et de l'acide citrique en violet sur la bande 0,05-2,90 THz.

D'après la Figure 3-5, notons que le fructose est caractérisé par 3 pics à 1,3-1,71-2,13 THz. Le lactose est quant à lui caractérisé par un pic très intense à 1,37 THz, et un pic bien isolé dans les basses fréquences à 0,53 THz, dans une gamme spectrale où les autres produits présentent une très faible absorption. Le lactose présente en plus deux pics plus faibles à 1,19 THz et 1,81 THz. Enfin, l'acide citrique est caractérisé principalement par trois pics aux fréquences à 1,29-1,7- 2,4 THz. On remarque que le fructose et l'acide citrique présentent tous les deux un pic d'absorption à 1,7 THz avec cependant une valeur d'absorption qui est deux fois supérieure pour le fructose que pour l'acide citrique. Cela laisse présager de possibles interférences qui pourraient dégrader les performances d'une analyse quantitative.

3.5 Analyse par ACP

3.5.1 Description des données

Un premier calcul ACP est effectué à l'aide du logiciel SIMCA-P+ sur les données brutes. Dans la bande spectrale analysée, on dispose de 454 variables. Le nombre d'observations est égal à 390 (39 échantillons ; 10 spectres par échantillon). Le résultat d'ACP est donné dans le Tableau 3-2.

A	R2X	R2X(cum)
1	0,73	0,73
2	0,114	0,845
3	0,0609	0,905
4	0,0373	0,943
5	0,0158	0,959
6	0,0103	0,969
7	0,0074	0,976
8	0,00572	0,982
9	0,00458	0,987

Tableau 3-2 Résultat d'ACP pour N=390 observations, et K=454 variables

On constate que les deux premières composantes permettent d'expliquer plus de 84 % de la variance des données. La Figure 3-6 présente les scores dans le plan t(1)/t(2) avec les mélanges binaires en rouge et les mélanges ternaires en bleu. On remarque une légère séparation entre les échantillons binaires et les échantillons ternaires suivant l'axe 1, qui pourrait être en partie expliquée par l'absorption globale des échantillons sur la gamme de fréquences considérée. Pour vérifier cette hypothèse, nous avons reporté sur la Figure 3-7 les spectres de 5 échantillons répartis le long de l'axe t(1). On remarque que plus l'absorbance est élevée, plus l'échantillon est placé vers la droite sur l'axe t(1). Pour aller plus loin, nous avons tracé sur la Figure 3-8, les courbes des loadings p(1) et p(2) en fonction de la fréquence. On remarque que les valeurs de p(1) sont positives pour la plupart des fréquences et on observe aussi que les valeurs de p(2) associées aux fréquences caractéristiques du lactose permettent de conclure que l'axe 2 de l'ACP permet essentiellement de séparer les échantillons selon leur taux de lactose. Cette analyse ACP sur des données brutes s'avère finalement être insuffisante.

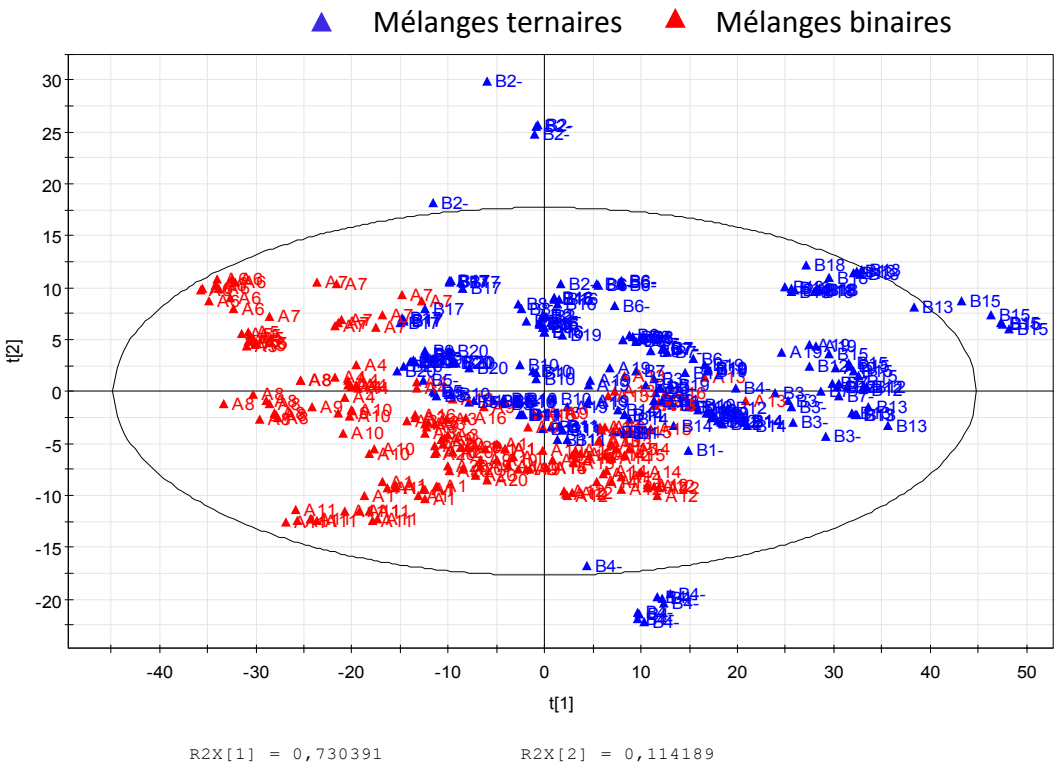


Figure 3-6 Présentation des scores dans le plan t1/t2. Les mélanges binaires sont en rouge et les mélanges ternaires en bleu

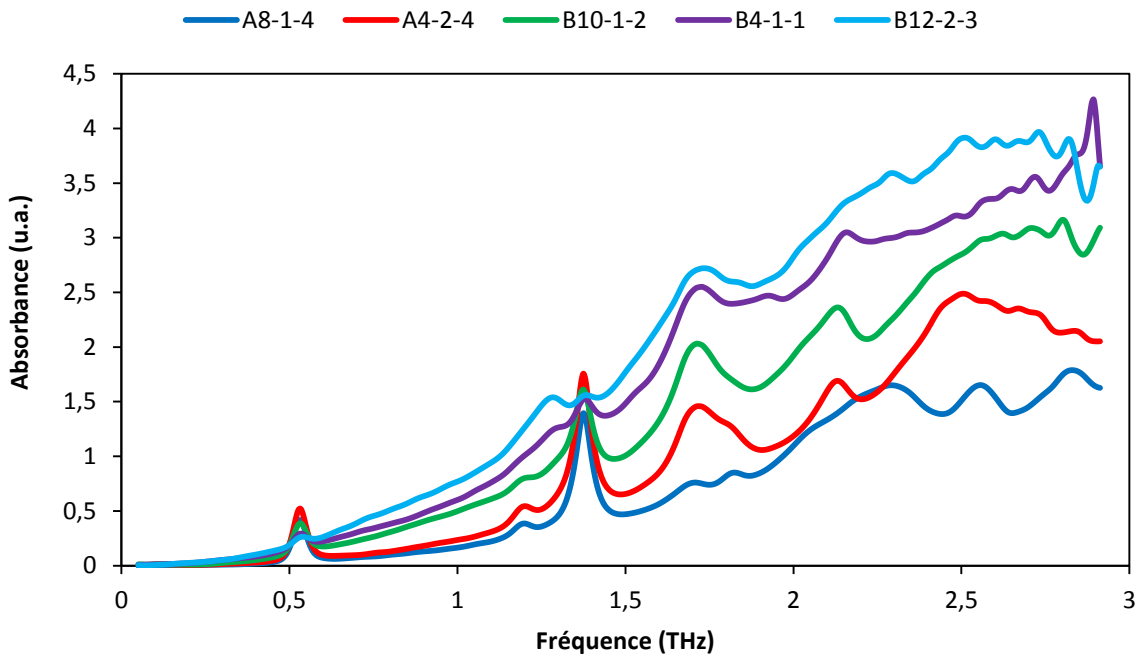


Figure 3-7 Spectres d'absorbance de 5 échantillons répartis suivant l'axe t1 de la Figure 3-6

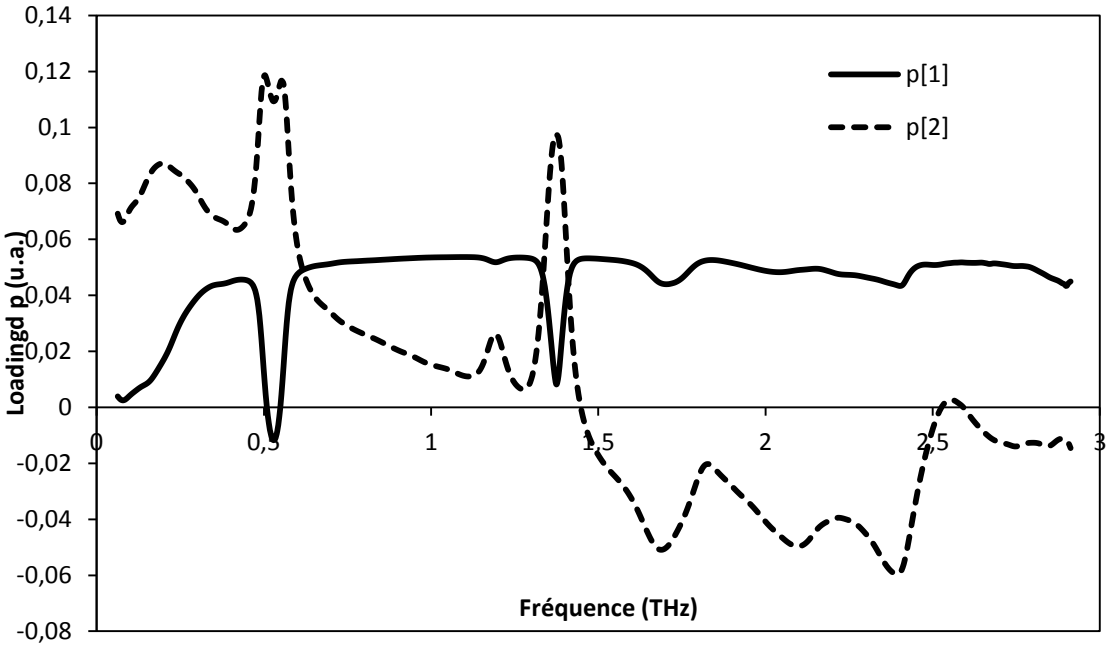


Figure 3-8 Courbes des deux loadings p(1) et p(2) en fonction de la fréquence (THz)

En considérant à présent des données centrées (après soustraction de la valeur moyenne pour chaque variable), l'ACP a permis de séparer les échantillons A11 et A1 correspondant respectivement à l'acide citrique pur et au fructose pur comme le montre la projection ACP dans le plan t(1)/t(2) donnée sur la Figure 3-9. En revanche, on ne parvient toujours pas à distinguer les 3 pôles du diagramme ternaire. Enfin, on constate que le spectre de la pastille B15-2 associée à l'échantillon B15 est considéré par ce modèle ACP comme une donnée aberrante par rapport aux autres.

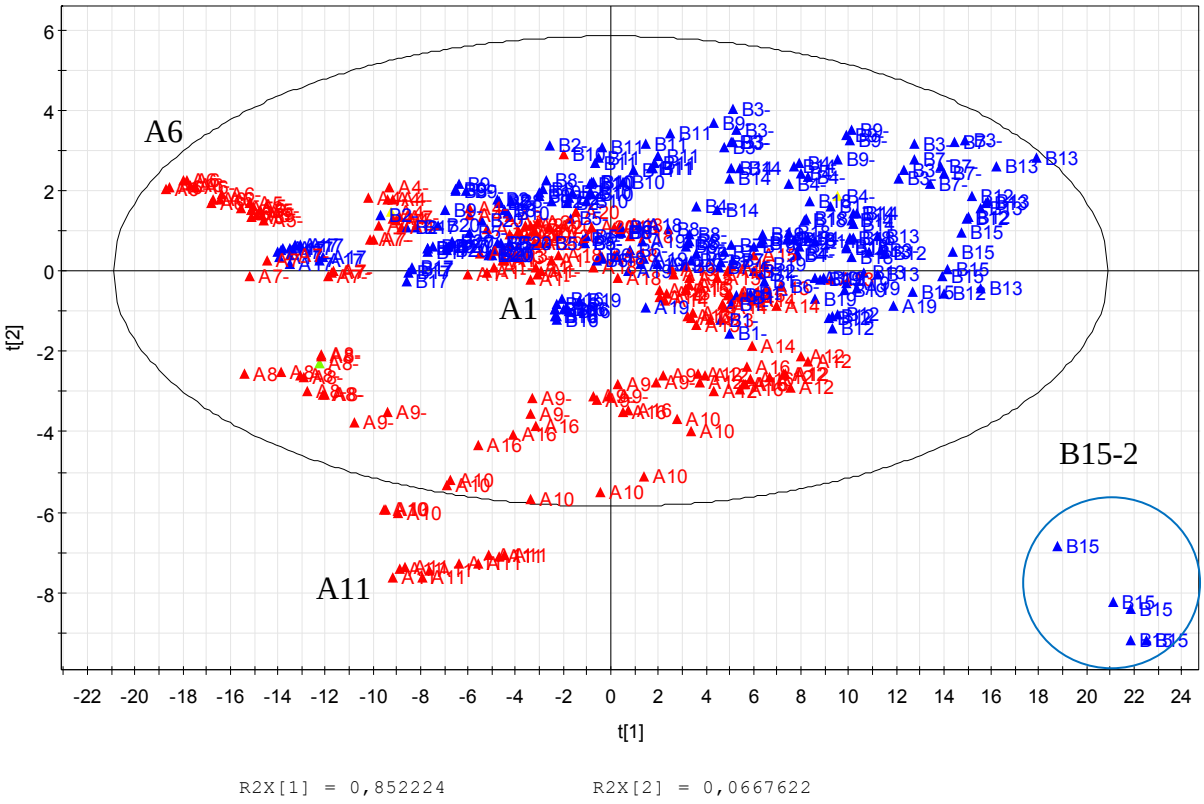


Figure 3-9 Présentation des scores dans le plan t(1)/t(2) pour des données centrées. A1, A6 et A11 représentent les trois corps purs tandis que les données relatives à B15-2 sont considérées aberrantes.

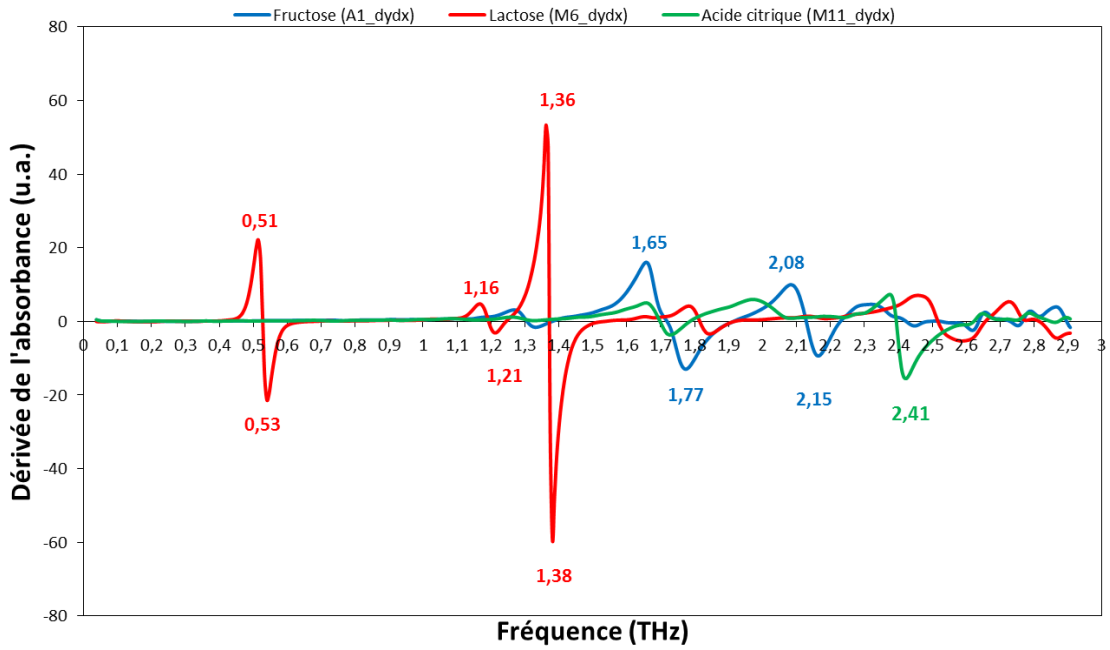


Figure 3-10 Dérivée première des spectres d'absorption du fructose (échantillon A1, bleu), du lactose (échantillon A6, rouge) et de l'acide citrique (échantillon A11, vert)

Pour aller plus loin dans le traitement ACP des spectres d'absorbance THz, on procède au calcul de la dérivée première en guise de prétraitement. Ainsi, le front montant d'une raie

d'absorption devient un pic positif et le front descendant un pic négatif comme on peut le voir sur la Figure 3-10. On remarque ainsi que la raie la plus significative pour l'acide citrique est une raie négative à 2,41 THz. Les données spectrales ainsi dérivées sont ensuite centrées avant de lancer le calcul d'ACP. Notons qu'on a retiré ici les données aberrantes de la pastille 2 de l'échantillon B15 ainsi que celles de la pastille 2 de l'échantillon B9 pour les mêmes raisons. Finalement, le nombre de variables est de 457 (0.0439 THz - 2.90619 THz) et le nombre d'observations est de 380. Le Tableau 3-3 décrit les résultats obtenus avec ce nouveau modèle ACP. Cette fois, les deux premières composantes permettent d'expliquer 60% de la variance des données.

A	R2X	R2X(cum)	Valeurs propres
1	0,506	0,506	192
2	0,104	0,61	39,4
3	0,0791	0,689	30
4	0,066	0,755	25,1
5	0,0542	0,809	20,6
6	0,037	0,846	14,1
7	0,0241	0,87	9,15
8	0,018	0,888	6,85

Tableau 3-3 Résultats d'ACP pour des données dérivées puis centrées en fonction du nombre A de composantes.

La Figure 3-11 donne la projection des scores dans le plan t(1)/t(2). Le résultat obtenu est tout à fait remarquable. En effet, on retrouve clairement les trois pôles du diagramme ternaire associés aux échantillons A1-fructose, A6-lactose et A11-acide citrique. De plus, les mélanges binaires (rouge) apparaissent bien sur les bords du triangle alors que les mélanges ternaires (bleu) apparaissent à l'intérieur. En regardant de plus près, on remarque enfin que l'ordre des échantillons est conforme à celui du diagramme ternaire. Ce résultat est très intéressant car il permet de conclure que l'ACP est un outil adapté au traitement des spectres d'absorption THz, non seulement pour apporter des réponses qualitatives mais aussi pour fournir une information semi-quantitative tout à fait robuste. Rappelons ici que l'ACP est une méthode non supervisée et que même sans apprentissage, il a été possible de retrouver le diagramme ternaire et donc de fournir des informations semi-quantitatives sur chaque échantillon.

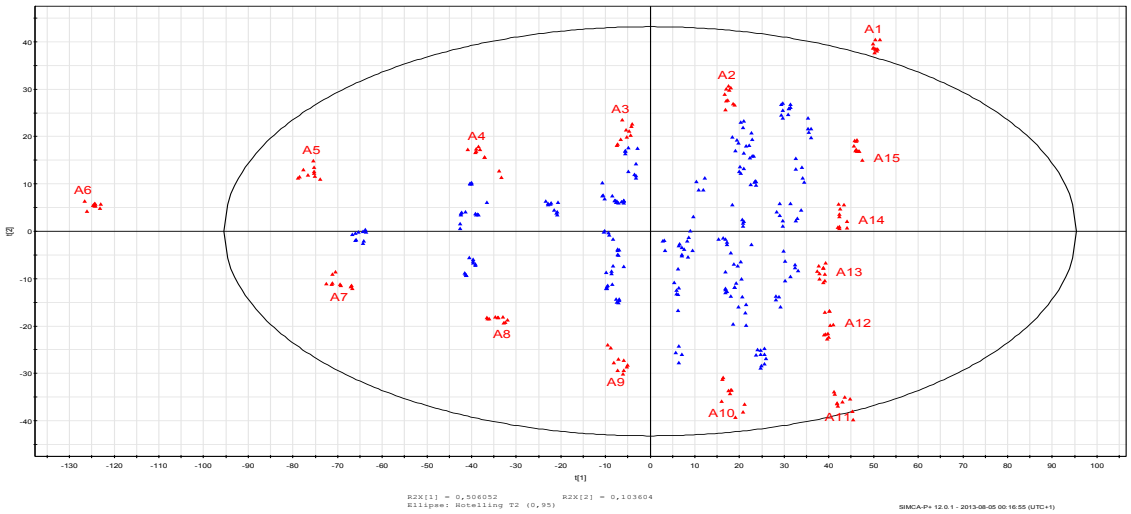


Figure 3-11 Présentation des scores des données dérivées puis centrées dans le plan t(1)/t(2). Bleu : mélanges ternaires, rouge : mélanges binaires et produits purs.

Pour mieux comprendre, reportons-nous à la Figure 3-12 qui présente les loadings p(1) en bleu et p(2) en rouge en fonction de la fréquence. On constate que p(1) traduit une anti-corrélation avec le lactose, ce qui signifie que sur la Figure 3-11 des scores, plus on va vers la droite selon l'axe 1 et plus la valeur en lactose diminue. Ceci permet d'expliquer que l'échantillon A6 associé au lactose pur se trouve complètement à gauche de la figure. Ainsi, l'axe 1 semble être exclusivement corrélé à la concentration en lactose. En poursuivant l'analyse, on remarque sur la Figure 3-12 que p(2) est corrélé au fructose et anti-corrélé à l'acide citrique. Ceci permet d'expliquer parfaitement que sur la Figure 3-11, on retrouve l'échantillon A1 (fructose pur) le plus en haut et l'échantillon A11 (acide citrique pur) le plus en bas.

Sur la base de ces observations pour les trois produits purs, il n'est finalement pas très surprenant de parvenir à reconstituer le diagramme ternaire pour les mélanges.

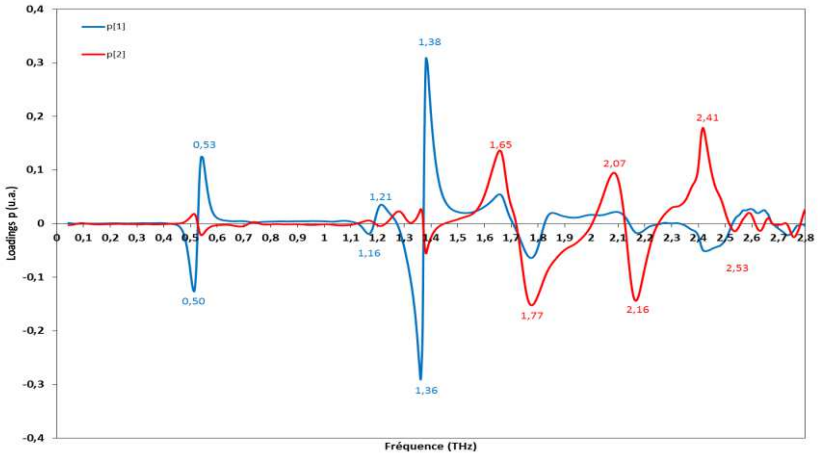


Figure 3-12 Présentation des courbes des loadings p[1] et p[2] des données dérivées puis centrées.

3.5.2 Prédiction semi quantitative en ACP

L'ACP permet de décrire un jeu de données comme nous venons de l'illustrer juste avant. Mais le résultat remarquable obtenu sur la Figure 3-11 nous a donné l'idée d'aller un cran plus loin et d'utiliser l'ACP pour prédire des concentrations. Cela revient à placer un échantillon inconnu dans le diagramme ternaire ou encore à projeter un échantillon inconnu dans le plan t(1)/t(2) d'un modèle ACP qui a été construit sans cet échantillon. Nous avons ainsi éliminé de un à six échantillons afin de tester progressivement la capacité de l'ACP à prédire les concentrations. Les échantillons retirés du modèle ACP et analysés a posteriori sont indiqués dans le Tableau 3-4. Dans un premier temps, seul l'échantillon B16 a été retiré puis prédit par un modèle ACP construit à partir de 370 observations. Ensuite, on a recommencé avec les échantillons B16 et B3. Dans ce cas, le modèle ACP a été construit à partir de 360 observations. Et ainsi de suite jusqu'à 6 échantillons retirés et un modèle ACP bâti à l'aide de 320 observations. Les colonnes du Tableau 3-4 décrivent respectivement : le nombre N de spectres inclus dans le modèle ACP, les échantillons éliminés du modèle et ensuite prédits a posteriori, le coefficient de corrélation R^2(t1/t'1) entre les valeurs de scores du modèle ACP initial (tous les échantillons) et celles du nouveau modèle (après élimination de certains échantillons) pour la composante 1, le coefficient de corrélation R^2(t2/t'2) semblable au précédent mais pour la composante 2. Notons que le calcul des coefficients de corrélation est le calcul le plus rapide pour vérifier que le nouveau modèle projette bien les échantillons dans le plan t(1)/t(2) de façon quasi similaire au modèle initial complet. On calcule aussi la valeur moyenne quadratique de l'erreur entre les anciens et les nouveaux scores pour les échantillons d'apprentissage (communs aux deux modèles) notée RMSE t1C pour la composante 1 et RMSE t2C pour la composante 2. On calcule enfin la valeur moyenne quadratique de l'erreur entre les anciens et les nouveaux scores pour les échantillons prédits (10 pour la première ligne, 20 pour la deuxième, etc…) notée RMSE t1P pour la composante 1 et RMSE t2P pour la composante 2.

N	Ech. de prédiction	R^2 (t1/t'1)	R^2 (t2/t'2)	RMSE t1C	RMSE t2 C	RMSE t1P	RMSE t2P
370	B16	1	1	0,23	1,03	0,10	1,10
360	B16-B3	1	1	0,37	1,03	0,40	1,10
350	B16-B3-B5	1	1	5,80	4,15	7,80	5,39
340	B16-B3-B5-B1	1	1	0,63	1,25	0,57	1,28
330	B16-B3-B5-B1-B2	1	1	0,61	1,35	0,55	1,26
320	B16-B3-B5-B1-B2-B4	1	1	0,16	1,51	0,13	1,37

Tableau 3-4 Analyse des valeurs des scores des deux premières composantes (1 et 2) des nouveaux modèles en fonction du modèle initial

Le Tableau 3-4 montre que jusqu'à l'élimination de 6 échantillons du modèle ACP initial, on obtient de faibles valeurs de RMSE, sachant que ces valeurs expriment l'écart entre les nouveaux scores et les anciens. En effet, rappelons que pour le modèle initial, les valeurs de t1 couvrent l'intervalle allant de -127 à + 52 ce qui correspond à une étendue de 179. De même, les valeurs de t2 couvrent un intervalle allant de -40 à +40 soit une étendue de 80. Les écarts reportés dans le tableau 4-3 sont donc extrêmement faibles si on raisonne en valeurs relatives. Par conséquent, une analyse semi-quantitative est tout à fait envisageable grâce à cette

approche. Pour cela, il suffit d'introduire les valeurs des concentrations dans le diagramme donnant les scores dans le plan t(1)/t(2). A titre d'exemple, prenons le cas selon lequel 5 échantillons ont été éliminés du modèle ACP (N=330). Les scores du modèle correspondant sont présentés sur la Figure 3-13. On y retrouve clairement les trois pôles associés aux trois produits purs qui forment un triangle non pas équilatéral comme présenté pour le diagramme ternaire de référence mais plutôt isocèle avec le côté fructose-acide citrique plus court que les deux autres. En reportant les valeurs des concentrations en particulier sur les côtés du triangle, c'est-à-dire pour les mélanges binaires, il est possible dans un second temps de prédire les concentrations. Notons cependant que les échelles ne sont pas identiques sur les trois côtés du triangle. Prenons l'exemple de l'échantillon B17 représenté en cyan sur la Figure 3-13. L'intersection de la droite parallèle au segment fructose-acide citrique et passant par la moyenne des points B17 (pointillé rouge) avec le segment lactose-fructose donne un pourcentage en lactose voisin de 60 %. De même l'intersection de la droite parallèle au segment fructose-lactose et passant par la moyenne des points B17 (pointillé vert) avec le segment lactose-acide citrique donne un pourcentage en acide citrique voisin de 30 %. Enfin, l'intersection de la droite parallèle au segment acide citrique-lactose et passant par la moyenne des points B17 (pointillé bleu) avec le segment fructose-acide citrique donne un pourcentage en fructose voisin de 10 %. On constate que les concentrations obtenues pour l'échantillon B17 (en moyenne) par cette analyse ACP sont remarquablement proches des concentrations utilisées lors de la préparation. De façon générale, l'étalement selon l'axe 2 étant plus petit que selon l'axe 1, les erreurs sur les valeurs prédites pour le fructose et pour l'acide citrique sont potentiellement plus grandes que celles du lactose et peuvent atteindre 10%.

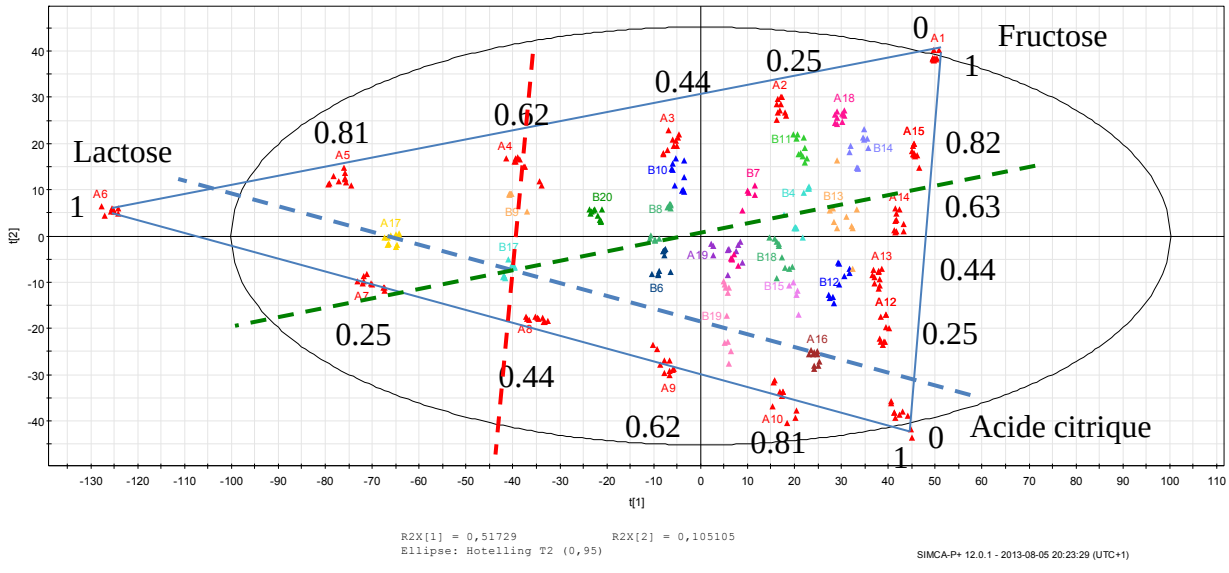


Figure 3-13 Présentation des scores du modèle ACP à N=330 après élimination de 5 échantillons

Notons cependant que l'échantillon B17 analysé précédemment avait été utilisé pour bâtir le modèle ACP. L'étape finale consiste donc à prédire les concentrations d'échantillons qui n'ont pas été inclus dans le modèle ACP. Dans le cas présent il s'agit des échantillons B1, B2, B3, B5 et B16. Après avoir projeté ces 5 échantillons dans les plan t(1)/t(2), il suffit

d'appliquer la même technique que pour l'échantillon B17 et on obtient les résultats présentés dans le Tableau 3-5.

	Lactose (%)		acide citrique (%)		Fructose (%)	
	référence	ACP	référence	ACP	référence	ACP
B1	20,0	20	59,8	60-70	20,2	10-20
B2	59,9	60	20,1	20	20,0	20
B3	20,2	20	20,1	10-20	59,7	60-70
B5	39,9	40	19,9	20-30	40,1	40-50
B16	40,0	40	49,9	40-50	10,1	10-20

Tableau 3-5 Concentrations relatives (%) de référence (pesage) et après traitement ACP des spectres THz.

On conclut de cette étude que l'ACP permet de réaliser une analyse semi-quantitative tout à fait acceptable dans le cadre de ces mélanges ternaires. On s'attend donc naturellement à ce que la méthode de régression en composantes principales (PCR) fournisse de bons résultats dans le cadre d'une analyse quantitative. Cependant, dans le but de pouvoir donner un sens physique aux composantes principales en lien avec la concentration de l'analyte, la méthode de régression aux moindres carrés partiels (PLS) sera privilégiée pour les analyses quantitatives.

Rappelons enfin que nous avons obtenu par le traitement ACP des erreurs relatives de prédiction inférieures à 10% sachant que le mélange ne constitue lui-même que 20% en masse de la pastille de 400 mg, soit 80 mg. Cela revient donc finalement à une erreur relative de 8 mg pour 400 mg que l'on peut encore écrire 2mg/100mg.

3.6 **Analyse quantitative**

Après l'analyse semi-quantitative réalisée par ACP, intéressons-nous à présent à une véritable analyse quantitative d'abord par PLS puis par ANN. De manière générale, il s'agit ici de bâtir un modèle quantitatif à l'aide d'un jeu d'échantillons dédiés à l'étalonnage puis de l'exploiter avec des échantillons connus afin de valider le modèle dans le but de vérifier que l'on n'a pas affaire à une situation de sur-apprentissage. De façon plus précise, il est nécessaire de séparer les échantillons en trois lots : lot de calibration, lot de validation et lot de test. Contrairement à l'analyse des données LIBS, nous avons pris en compte ici 10 spectres pour chaque échantillon plutôt qu'un seul spectre moyen. Ce choix a été motivé par le fait que les échantillons sont considérés ici homogènes étant donnée la dimension importante du faisceau térahertz d'analyse. Ainsi, chaque spectre individuel est supposé représentatif de l'échantillon. Afin de ne pas focaliser l'étude sur des artéfacts incompris, les échantillons qui révèlent une différence importante entre les spectres des deux réplicats sont exclus de l'étude. Il s'agit dans ce cas des échantillons B4, B15 et B 9. Par ailleurs, afin de pouvoir utiliser la méthode de validation externe, nous avons répartis les échantillons en trois lots complétement différents. Les 36 échantillons restants sont ainsi répartis de la façon suivante :

- Lot d'apprentissage : A1-A2-A4-A5-A6-A7-A9-A10-A11-A12-A14-A15-A16-A17-A18-B2-B7-B19-B20 à savoir 190 spectres.

- Lot validation : A3-A8-A13-A19-B1-B3-B8-B10-B12-B14-B16, ce qui correspond à 110 spectres.
- Lot de test : B5-B6-B11-B13-B17-B18, ce qui correspond à 60 spectres.

3.6.1 Analyse par PLS

L'analyse qualitative par ACP a été validée et l'ajout d'informations supervisées a permis de proposer une analyse semi-quantitative. En réalité, cette approche qui consiste à utiliser les composantes principales de l'ACP pour ensuite prédire des concentrations n'est rien d'autre que la méthode de régression en composante principale (PCR). L'inconvénient de cette méthode est qu'elle élimine les variables qui présentent les variances les plus faibles et qui peuvent être cependant corrélées à la concentration de l'analyte (Y). Pour éviter cela, on préfèrera à la PCR la technique de régression aux moindres carrés partiels ou PLS. Celle-ci privilégie les variables directement corrélées avec la concentration de l'analyte comme cela a été décrit dans le Chapitre 1 de ce mémoire. Nous avons testé différents types de prétraitement et pour chacun, nous avons choisi le nombre A de composantes principales par validation croisée externe. Ctr désigne des données centrées tandis que Ctr-1st-deriv désigne un prétraitement par dérivée première puis les nouvelles données obtenues centrées. Les résultats sont présentés dans le Tableau 3-6.

prétraitement	**modèle**	**analyte**	**bande (THz)**	**A**	**RMSE (%)**			**Q^2**			**R^2**		
					C	**V**	**T**	**C**	**V**	**T**	**C**	**V**	**T**
Ctr	PLS-2	F	0,04-2,9	7	0,5	1,6	1,0	1,00	0,96	0,93	1,00	0,97	0,93
Ctr	PLS-2	AC	0,04-2,9	7	0,6	1,8	1,4	1,00	0,95	0,74	0,99	0,95	0,39
Ctr	PLS-2	L	0,04-2,9	7	0,3	0,5	0,6	1,00	0,99	0,97	0,99	0,99	0,45
Ctr	PLS-1	F	0,04-2,9	6	0,5	1,6	1,0	1,00	0,96	0,94	1,00	0,97	0,94
Ctr	PLS-1	AC	0,04-2,9	6 1	0,6	1,7	1,2	1,00	0,96	0,80	1,00	0,96	0,41
Ctr	PLS-1	L	0,04-2,9	0	0,1	0,3	0,3	1,00	1,00	0,99	0,99	1,00	0,52
Ctr-1st-deriv	PLS-1	F	0,04-2,9	3	0,5	1,1	0,8	1,00	0,98	0,96	1,00	0,98	0,97
Ctr-1st-deriv	PLS-1	F	0,04-2,4	2	1,3	1,0	1,4	0,98	0,98	0,87	0,98	0,99	0,85
Ctr-1st-deriv	PLS-1	F*	0,04-2,6	3	0,6	0,9	0,7	0,99	0,99	0,96	0,99	0,99	0,97
Ctr-1st-deriv	PLS-1	L*	0,04-2,9	7	0,1	0,2	0,1	1,00	1,00	1,00	1,00	1,00	0,99
Ctr-1st-deriv	PLS-1	AC*	0,04-2,9	7	0,5	0,9	0,8	1,00	0,99	0,91	1,00	0,99	0,79
Ctr-1st-deriv	PLS-1	AC	0,04-2,6	8	0,4	0,9	0,7	1,00	0,99	0,93	1,00	0,99	0,88
Ctr-1st-deriv	PLS-2	F	0,04-2,6	4	0,7	0,9	0,7	0,99	0,99	0,96	0,99	0,99	0,94
Ctr-1st-deriv	PLS-2	AC	0,04-2,6	4	0,7	1,0	0,8	0,99	0,98	0,90	0,99	0,99	0,76
Ctr-1st-deriv	PLS-2	L	0,04-2,6	4	0,2	0,3	0,2	1,00	1,00	0,99	1,00	1,00	0,98

Tableau 3-6 Performances de différents modèles PLS pour deux types de prétraitement (Ctr pour des données centrées et Ctr-1st-deriv pour des données dérivées une fois puis centrées) et pour différentes bandes spectrales. Dans chaque cas, le nombre A de composantes principales est indiqué. Les analytes sont F : fructose, L : lactose et AC : acide citrique. C, V et T indiquent les lots de calibration, de validation et de test. (*) indique le meilleur modèle obtenu pour chaque analyte.

PLS-1 désigne l'algorithme de PLS qui permet de prédire la concentration d'un seul analyte tandis que PLS-2 désigne celui qui permet de prédire simultanément les concentrations de

plusieurs analytes. Pour chaque modèle, on précise quel est l'analyte : F pour fructose, L pour lactose et AC pour acide citrique. Les valeurs de RMSE (%) doivent être comparées aux valeurs des concentrations qui se trouvent dans la gamme allant de 0 à 20%. Le meilleur modèle est celui qui permet de minimiser les valeurs de RMSE pour les différents lots d'échantillons. On constate que PLS-1 donne de meilleurs résultats que PLS-2 et que la dérivée première permet d'améliorer les résultats par comparaison aux données centrées. Pour le lactose et l'acide citrique, les meilleurs modèles (marqués par une étoile dans le Tableau 3-6) sont obtenus pour 7 composantes dans la bande spectrale 0,04-2,9 THz. Pour le fructose, le meilleur modèle est obtenu pour 3 composantes dans la bande spectrale 0,04-2,6 THz. Dans ce cas, RMSE(V)=0,95% sur une échelle des concentrations allant jusqu'à 20% pour une pastille de 400 mg de masse totale. Ainsi 20% correspondent à une masse de 80 mg hors liant. Et finalement, 0,95% correspondent à une masse de 0,95%*80 = 0,76 mg. Finalement la quantité 0,76 mg pour 400 mg peut encore être écrite 0,19 mg / 100 mg. Pour le lactose, RMSE(V) = 0,25 (%) tandis que 20% correspondent à 80 mg hors liant. Donc finalement 0,25% correspondent à une masse de 0,25%*80 = 0,2 mg. Ceci est obtenu pour une pastille de 400 mg donc la valeur finale est de 0,05 mg/100 mg. De même pour l'acide citrique, RMSE(V) = 0,89 (%) ce qui correspond à 0,178 mg / 100 mg.

Il ne reste plus qu'à vérifier la signification statistique des différents modèles en appliquant la procédure de Y-randomisation décrite au chapitre 1. Pour cela, nous avons permuté 25 fois la variable Y de sortie, à savoir la concentration en fructose, en lactose ou en acide citrique selon les cas. Les résultats des calculs de R^2 à partir du lot d'apprentissage et de Q^2 à partir du lot d'apprentissage par validation croisée interne sont reportés dans le Tableau 3-7. Ils montrent que les modèles optimisés ont une réelle signification statistique.

M-PLS1	R^2	R^2_Yrandom	Q^2	Q^2_ Yrandom
Fructose	0,98	0,48	0,98	0,43
Lactose	1,00	0,05	1,00	0,62
A. citrique	0,99	0,77	0,98	0,60

Tableau 3-7 Comparaison des coefficients R^2 et Q^2 entre les meilleurs modèles PLS (notés par (*) dans le Tableau 3-6) et ceux résultant de la procédure de Y-randomisation, le tout pour le lot d'apprentissage.

La Figure 3-14 décrit les capacités des meilleurs modèles PLS pour le fructose, l'acide citrique et le lactose à prédire les concentrations des échantillons des lots de validation et de test. On constate que les prédictions pour le lactose sont excellentes pour toutes les valeurs de concentrations. Pour le fructose, certains échantillons comme B16 et A19 sont moins bien prédits que les autres mais l'ensemble des échantillons est toutefois très correctement prédit par le traitement PLS. Pour l'acide citrique, les prédictions sont également satisfaisantes excepté pour l'échantillon A19 pour lequel la concentration est surestimée. Par ailleurs, on constate que l'échantillon B13 présente des fluctuations importantes de la valeur prédite de concentration en acide citrique. Il sera intéressant de comprendre les raisons de ces fluctuations par une étude ultérieure plus approfondie. Une des raisons possible réside dans la préparation de l'échantillon et en particulier le broyage des poudres et leur mélange de façon homogène.

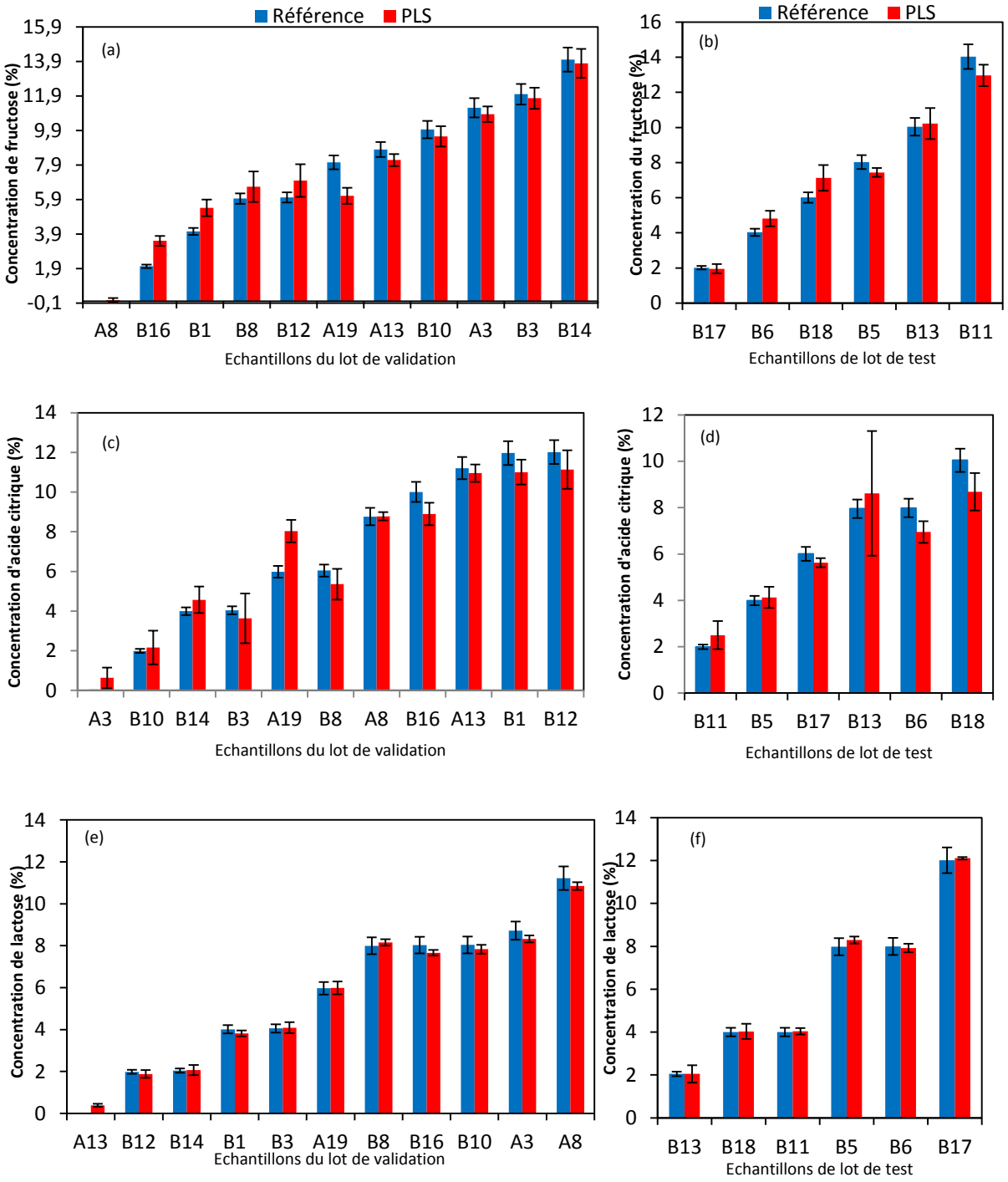


Figure 3-14 Comparaison entre les concentrations de référence en bleu (résultant du pesage) et les concentrations prédites par PLS en rouge. Les barres d'erreur indiquées sur les résultats PLS représentent l'écart-type des 10 spectres analysés pour chaque échantillon tandis que celles indiquées sur les résultats de référence correspondent à une erreur relative de 5% prise de façon arbitraire. Les résultats pour le fructose sont donnés en (a) pour le lot de validation et en (b) pour le lot de test. De même (c) et (d) concernent l'acide citrique et (e) et (f) le lactose.

On peut conclure que l'analyse quantitative des données THz par PLS est tout à fait satisfaisante. Il ne semble donc pas y avoir dans ce cas précis de problème de non-linéarité susceptible de dégrader les performances de la PLS.

3.6.2 Analyse par ANN

Même si la PLS a donné des résultats satisfaisants dans le cadre de l'analyse de ces mélanges ternaires, nous avons décidé d'évaluer l'analyse par ANN qui est potentiellement transposable à un plus grand nombre de cas d'analyse d'échantillons complexes. On applique la même répartition des échantillons en trois lots que celle présentée précédemment. Mais il n'est pas question d'injecter dans l'ANN toutes les données de la gamme spectrale 0-3 THz, à savoir 457 points, ce qui conduirait à coup sûr à un sur-apprentissage ou encore à un problème de sous-dimensionnalité si le nombre d'échantillons est très inférieur à celui des variables dans le spectre. Toutefois, la sélection des données d'entrée de l'ANN n'est pas facile car le spectre THz ne révèle pas des raies bien précises comme en LIBS mais au contraire une évolution de l'absorption sur une gamme spectrale extrêmement étendue. L'idée première est donc de compresser les données spectrales THz par un traitement ACP de façon à ne garder que quelques variables non corrélées entre elles. Nous avons adopté en guise de prétraitement la dérivée première puis les données centrées car nous avons démontré que cela apportait un réel avantage dans le traitement par ACP, tout comme dans le cas de la PLS présentée plus haut. Les 190 spectres du lot d'apprentissage ont ainsi servi à calculer les vecteurs propres (loadings), ce qui permet dans un second temps de calculer les scores pour les échantillons des lots de validation et de test.

3.6.2.1 Résultats pour le Fructose

L'analyse par PLS avait donné le meilleur résultat pour la bande 0,04-2,6 THz. On a donc considéré cette bande comme point de départ de notre calcul, ce qui correspond à 408 variables. Le Tableau 3-8 donne les résultats de l'ACP pour chaque composante principale jusqu'à A=12.

A	R^2	R^2(cum)	Valeurs propres	Q^2	Limit	Q^2(cum)
1	0,725	0,725	138	0,722	0,0077	0,722
2	0,14	0,864	26,5	0,501	0,00773	0,861
3	0,0568	0,921	10,8	0,395	0,00776	0,916
4	0,0231	0,944	4,39	0,236	0,0078	0,936
5	0,0186	0,963	3,52	0,318	0,00783	0,956
6	0,00879	0,971	1,67	0,221	0,00787	0,966
7	0,00648	0,978	1,23	0,211	0,0079	0,973
8	0,00426	0,982	0,809	0,167	0,00794	0,978
9	0,00266	0,985	0,505	0,104	0,00797	0,98
10	0,00251	0,987	0,476	0,15	0,00801	0,983
11	0,00153	0,989	0,291	0,0868	0,00805	0,984
12	0,00131	0,99	0,25	0,0716	0,00809	0,986

Tableau 3-8 Les performances des 12 premières composantes principales du modèle ACP (la valeur Q^2 est calculée par une validation croisée interne LOO). N=190 observations, K=408 variables

Pour chaque composante A, l'ACP nous permet de calculer les valeurs des loadings (p) pour chaque variable k variant entre 1 et K=408. Les entrées du modèle ANN seront les valeurs des scores (t) pour chaque échantillon. Une fois le modèle construit, les échantillons inconnus seront projetés dans le nouvel espace des composantes principales via un calcul des scores. Les scores ACP peuvent naturellement prendre des valeurs négatives ou positives. Or, l'apprentissage de l'ANN doit être effectué avec des valeurs positives seulement. On est donc amené à ajouter un offset à l'ensemble des scores ACP pour n'avoir que des données positives. On normalise enfin à 1 en divisant toutes les valeurs par la valeur maximale. Le nombre optimum d'entrée de l'ANN n'est pas connu a priori et il faut donc tester différents cas afin de faire le meilleur choix en fonction des performances de prédiction de l'ANN. Notons ici que les composantes sont classées par ordre décroissant d'importance, ce qui signifie que la première composante est la plus représentative de la variance du jeu de données, la seconde un peu moins et ainsi de suite. Les entrées de l'ANN ayant été présentées, rappelons que dans le cas présent, il n'y a qu'une seule sortie à l'ANN qui donne la concentration prédite en fructose. Pour chaque nombre d'entrées, c'est-à-dire de composantes principales, on cherche à construire le meilleur modèle ANN en optimisant le nombre de neurones dans la couche cachée, la vitesse d'apprentissage, le terme de mémoire et le nombre d'itérations. Le modèle est construit à l'aide du lot de calibration puis testé à l'aide du lot de validation pour détecter un éventuel problème de sur-apprentissage. Les paramètres correspondant au meilleur modèle pour le fructose sont donnés dans le Tableau 3-9 et les performances associées dans le Tableau 3-10.

A	Neurones cachés	Vitesse d'apprentissage	Terme de mémoire	Nombre d'itérations
5	2	0.05	0.2	12000

Tableau 3-9 Paramètres d'apprentissage du modèle ANN optimum pour l'analyse du fructose. Les 5 premières composantes principales de l'ACP ont permis de fournir les 5 données d'entrée.

	RMSE (%)	Q^2	R^2
Lot d'apprentissage (C)	0,7	0,99	0,99
Lot de validation (V)	0,7	0,99	0,99
Lot de test (T)	0,5	0,98	0,98

Tableau 3-10 performances obtenues à l'aide du modèle ANN dont les paramètres sont donnés dans le Tableau 3-9.

En comparant ces résultats avec les résultats de PLS donnés dans le Tableau 3-6, on constate que les performances de l'ANN sont légèrement supérieures. En effet, RMSE(V) passe de 0,95 pour la PLS à 0,70 pour l'ANN et RMSE(T) de 0,75 à 0,49. On remarque par ailleurs que la prédiction par ANN est meilleure pour la majorité des échantillons comme la montre la Figure 3-15 sur laquelle les données de référence (pesage) sont en bleu, les résultats ANN en rouge et les résultats PLS en vert. Les résultats ANN sont systématiquement plus proches des données de référence pour toutes les concentrations, excepté lorsque la valeur de concentration est nulle. Les barres d'erreur pour les données de référence et les résultats PLS s'appuient sur la même définition que pour la Figure 3-14. En ce qui concerne les données ANN, les barres d'erreur représentent l'écart-type mesuré sur les 10 mesures faites pour un même échantillon.

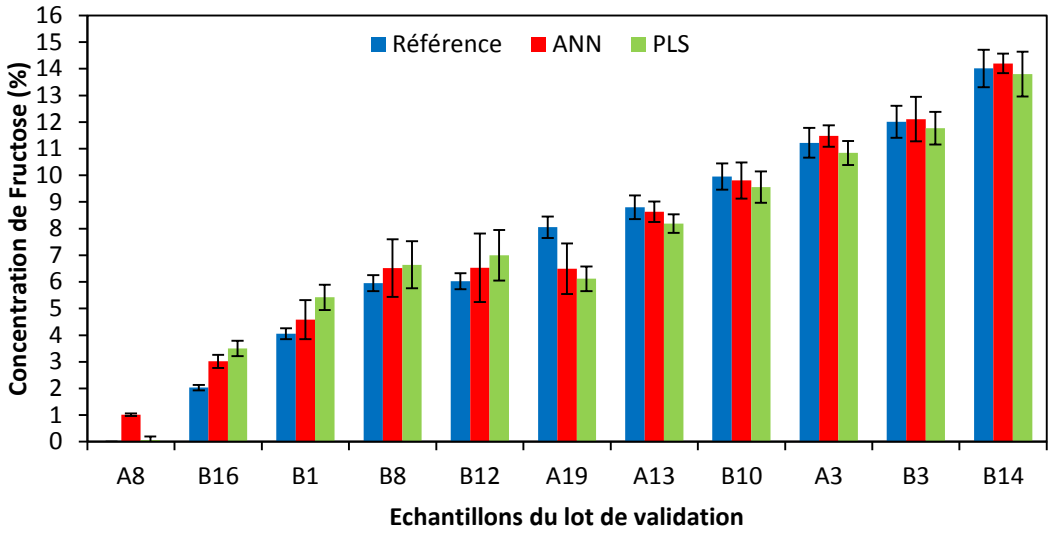


Figure 3-15 Comparaison des valeurs de concentration du fructose pour le lot de validation. En bleu : valeurs de référence (pesage), en rouge : résultat ANN, en vert : résultat PLS. Les barres d'erreur pour PLS et ANN représentent l'écart-type du résultat des 10 spectres d'un même échantillon. Pour les données de référence, il s'agit d'une valeur relative de 5% choisie arbitrairement.

Finalement, outre le mauvais résultat obtenu par ANN pour l'échantillon A8 qui ne contient pas de fructose, on peut vérifier grâce au lot de test (Figure 3-16) que l'ANN est globalement plus performant que la PLS pour l'analyse quantitative du fructose dans le cas présent d'un mélange ternaire analysé par spectroscopie THz.

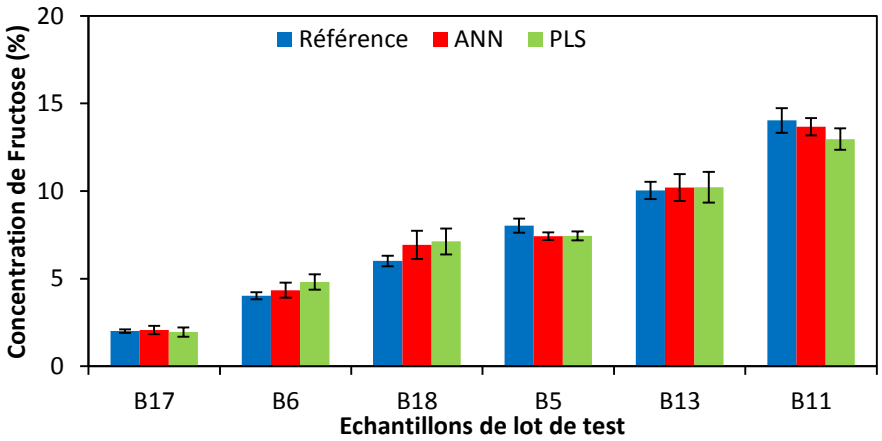


Figure 3-16 Comparaison des valeurs de concentration du fructose pour le lot de test. En bleu : valeurs de référence (pesage), en rouge : résultat ANN, en vert : résultat PLS. Les barres d'erreur pour PLS et ANN représentent l'écart-type du résultat des 10 spectres d'un même échantillon. Pour les données de référence, il s'agit d'une valeur relative de 5% choisie arbitrairement.

3.6.2.2 Résultats pour le Lactose

Nous avions obtenu un excellent résultat par PLS avec cependant un nombre de composantes principales assez élevé, A=7 par validation croisée interne, ce qui peut laisser craindre un sur-apprentissage. On a conservé ici la bande spectrale qui avait été retenue pour la PLS, à savoir

0,04-2,9 THz, ce qui correspond à K=457 variables injectées dans le calcul d'ACP. Les résultats sont présentés dans le Tableau 3-11 pour les 11 premières composantes principales.

A	R2X	R2X(cum)	Valeurs propres	Q2	Limit	Q2(cum)
1	0,716	0,716	136	0,713	0,00744	0,713
2	0,138	0,854	26,2	0,48	0,00747	0,851
3	0,0562	0,91	10,7	0,362	0,0075	0,905
4	0,024	0,934	4,57	0,22	0,00753	0,926
5	0,0188	0,953	3,58	0,27	0,00757	0,946
6	0,00925	0,962	1,76	0,175	0,0076	0,955
7	0,00824	0,97	1,57	0,206	0,00764	0,964
8	0,00491	0,975	0,933	0,128	0,00767	0,969
9	0,00316	0,978	0,601	0,106	0,0077	0,972
10	0,00273	0,981	0,519	0,102	0,00774	0,975
11	0,00188	0,983	0,357	0,0637	0,00778	0,977

Tableau 3-11 Les performances des 11 premières composantes principales du modèle. N=190 observations, K=457 variables.

Nous avions déjà constaté que par calcul ACP, la première composante était fortement corrélée à la concentration du lactose. Il n'est donc pas surprenant de voir dans le Tableau 3-11 que la composante 1 permet d'expliquer près de 72% de la variance du jeu de données. Le score n°1 sera donc une excellente donnée d'entrée pour l'ANN dans le but de quantifier le lactose. Reste cependant à évaluer combien de données d'entrée il est souhaitable d'introduire dans l'ANN. Pour cela, nous avons réalisé une étude systématique en faisant varier le nombre de données d'entrée et les paramètres de l'ANN. Les valeurs de RMSE(V) sont présentées sur la Figure 3-17.

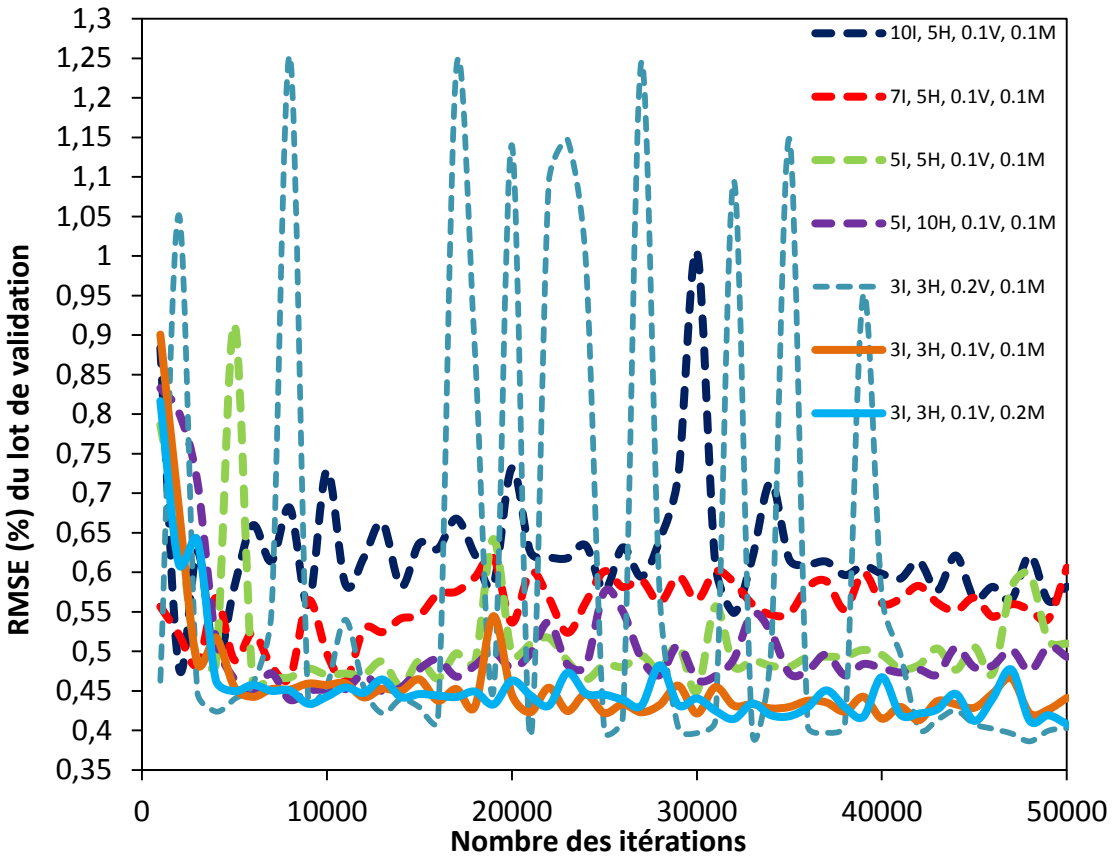


Figure 3-17 RMSE pour le lot de validation en fonction de nombre d'itérations. I : nombre d'entrées, H : nombre de neurones dans la couche cachée, V : vitesse d'apprentissage, M : terme de mémoire

En introduisant les 10 premières composantes de l'ACP en entrée de l'ANN, on obtient RMSE(V)_min= 0,48 pour 2000 itérations seulement, ce qui montre l'existence d'un sur-apprentissage rapide (pointillé gras bleu foncé). En réduisant à 7 entrées (les 7 premières composantes principales de l'ACP), RMSE(V)_min = 0,45 pour 11 000 itérations (pointillé gras rouge). Pour 5 entrées dans l'ANN (les 5 premières composantes principales de l'ACP), nous avons un résultat pour 5 neurones dans la couche cachée (pointillé gras vert) et un autre pour 10 neurones dans la couche cachée (pointillé gras violet). Dans les deux cas, RMSE(V)_min = 0,43 est légèrement plus faible et le sur-apprentissage est plus lent en fonction du nombre d'itérations par rapport aux modèles ANN à 10 ou à 7 entrées. Pour 3 entrées dans l'ANN (les 3 premières composantes principales de l'ACP) on remarque pour une vitesse d'apprentissage égale à 0,2 (pointillé gras bleu) le résultat oscille très fortement en fonction du nombre d'itérations. On préfèrera donc une vitesse d'apprentissage plus faible, que l'on fixe à 0,1. Dans ce cas, pour un terme de mémoire égal à 0,1 (trait plein gras orange) ou égal à 0,2 (trait plein gras bleu), on obtient RMSE(V)_min entre 0,4 et 0,45 ce qui montre que le terme de mémoire n'a pas une grande influence sur le résultat dans ce cas précis. On décide donc de fixer la valeur du terme de mémoire à 0,1 de même que la vitesse d'apprentissage, sachant que le nombre de neurones dans la couche cachée a été optimisé quant à lui à la valeur 3. Il ne reste plus qu'à faire varier le nombre d'itérations. En passant de 50 000 à 850 000 itérations, on remarque que la valeur de RMSE(V)_min reste stable autour 0,38 alors que dans le même temps RMSE(C)_min est passée de 0,28 à 0,19. On améliore

donc la prédiction des échantillons du lot d'apprentissage mais sans pour autant améliorer la prédiction des concentrations pour les échantillons du lot de validation. On en déduit que la valeur de 0,38 est la valeur minimale que peut atteindre RMSE(V) dans tous les cas.

Finalement, le Tableau 3-12 récapitule les performances obtenues pour deux nombres d'itérations (54000 et 388000) dans le cas de 3 entrées dans l'ANN et la Figure 3-18 présente une comparaison entre les valeurs des concentrations de lactose prédites par ANN (pour 54000 puis pour 388000 itérations), les valeurs prédites par PLS en prenant en compte le spectre complet et enfin les valeurs de référence obtenues par pesage, le tout pour le lot de validation. On constate sur la Figure 3-18 que le fait d'augmenter le nombre d'itérations de 54000 à 388000 n'améliore pas significativement les résultats. Ceci se vérifie aussi pour les échantillons du lot de test dont les résultats sont montrés sur la Figure 3-19.

	RMSE (%)		Q^2		R^2	
Nombre d'itérations	54 000	388 000	54 000	388 000	54 000	388 000
Lot d'apprentissage (C)	0,2	0,1	1,00	1,00	1,00	1,00
Lot de validation (V)	0,3	0,3	1,00	1,00	1,00	1,00
Lot de test (T)	0,3	0,3	0,99	0,99	0,99	1,00

Tableau 3-12 Performances de deux modèles ANN pour 54 000 et 388 000 itérations et pour 3 entrés (les trois premières composantes de l'ACP), nombre de neurones dans la couche cachée =3, vitesse d'apprentissage =0,1 ; terme de mémoire= 0,1.

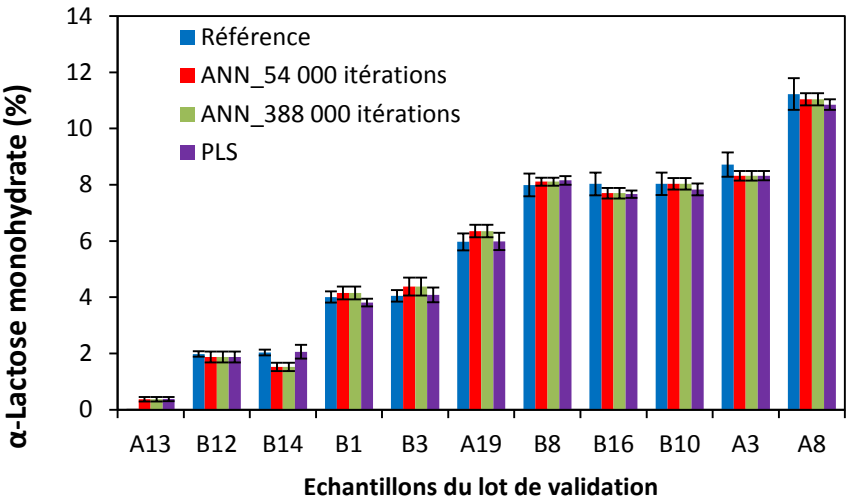


Figure 3-18 Concentrations en lactose des échantillons du lot de validation : Bleu : valeurs de référence (pesage), rouge : ANN pour 54000 itérations, vert : ANN pour 388000 itérations (voir le texte pour le détail sur les paramètres) et violet : PLS. Les barres d'erreur pour ANN et PLS sont données par l'écart-type sur les 10 spectres d'un même échantillon. Celles des données de référence correspondent à une erreur relative fixée arbitrairement à 5%.

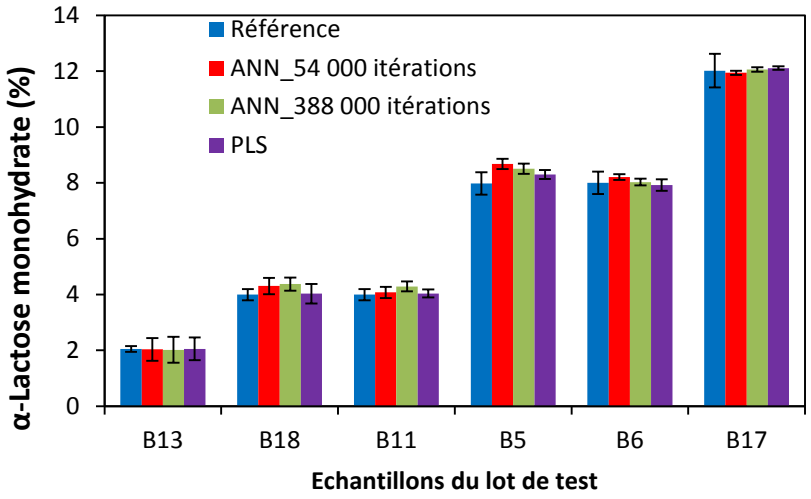


Figure 3-19 Concentrations en lactose des échantillons du lot de test : Bleu : valeurs de référence (pesage), rouge : ANN pour 54000 itérations, vert : ANN pour 388000 itérations (voir le texte pour le détail sur les paramètres) et violet : PLS. Les barres d'erreur pour ANN et PLS sont données par l'écart-type sur les 10 spectres d'un même échantillon. Celles des données de référence correspondent à une erreur relative fixée arbitrairement à 5%.

On conclut que pour l'analyse quantitative du lactose, l'ANN présente des performances tout à fait comparables à celles de la PLS. Cela est certainement dû au caractère linéaire des variations d'absorbance THz en fonction de la concentration en lactose dans les conditions expérimentales qui ont été étudiées.

3.6.2.3 Résultats pour l'acide citrique

Le choix de nombre d'entrées dans l'ANN est crucial pour atteindre des performances optimales. La Figure 3-20 décrit l'évolution de RMSE (moyenne sur 5 calculs) en fonction du nombre d'itérations pour 3 puis pour 5 données d'entrée (premières composantes principales du calcul ACP) et pour les trois lots d'échantillons : calibration, validation et test. Les barres d'erreur ne sont pas représentées afin de ne pas surcharger le graphique. Cependant, nous avons vérifié qu'elles étaient suffisamment étroites pour ne pas risquer de modifier l'interprétation des résultats. On remarque que dans le cas de 5 entrées, l'optimum est obtenu avant 10000 itérations, laissant craindre ensuite un risque de sur-apprentissage. En revanche, ce risque n'apparaît pas dans le cas 3 données d'entrée, du moins jusqu'à 50000 itérations.

Pour chacun des deux modèles précédents, c'est-à-dire avec 3 ou 5 données d'entrée, nous avons recherché le nombre d'itérations donnant les performances optimales. Nous avons trouvé 5000 itérations dans le cas de 5 données d'entrée et 36000 itérations pour 3 données d'entrée. Les meilleurs résultats obtenus dans ces conditions pour RMSE dans le cas des trois lots d'échantillons sont présentés dans le Tableau 3-13. On remarque que ces deux configurations sont très satisfaisantes pour l'analyse quantitative avec un léger avantage pour le modèle à 3 entrées et 36000 itérations.

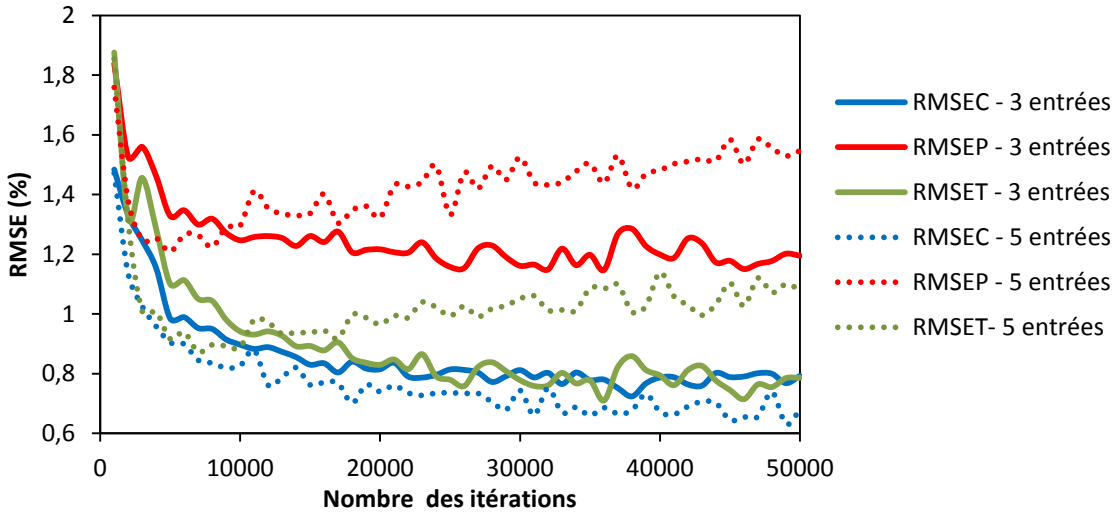


Figure 3-20 RMSE (%) en fonction du nombre d'itérations pour deux modèles ANN, l'un avec 3 entrées (traits pleins) et l'autre avec 5 entrées (traits pointillés). Bleu : calibration (190 spectres), rouge : validation (110 spectres) et vert : test (60 spectres).

	RMSE (%)		Q^2		R^2	
Nombre d'entrées	3	5	3	5	3	5
Lot d'apprentissage	0,5	0,7	1,00	0,99	1,00	0,99
Lot de validation	0,9	1,0	0,99	0,98	0,99	0,99
Lot de test	0,5	0,6	0,97	0,95	0,99	0,99

Tableau 3-13 Performances de deux modèles ANN à deux différents nombres d'entrées (3 et 5), premières composantes du calcul ACP.

La Figure 3-21 montre une comparaison des concentrations d'acide citrique de référence (par pesage) en bleu, par ANN (3 entrées, 36000 itérations) en rouge et par PLS (appliquée au spectre complet) en vert, pour le lot de validation. On constate que les valeurs fournies par l'ANN sont plus proches des valeurs de référence que celles fournies par la PLS pour des concentrations comprises entre 2 et 10%. En revanche, l'ANN a du mal à quantifier correctement l'échantillon A3 qui ne contient pas d'acide citrique. De plus, l'ANN a tendance à sous-estimer les concentrations supérieures à 10%.

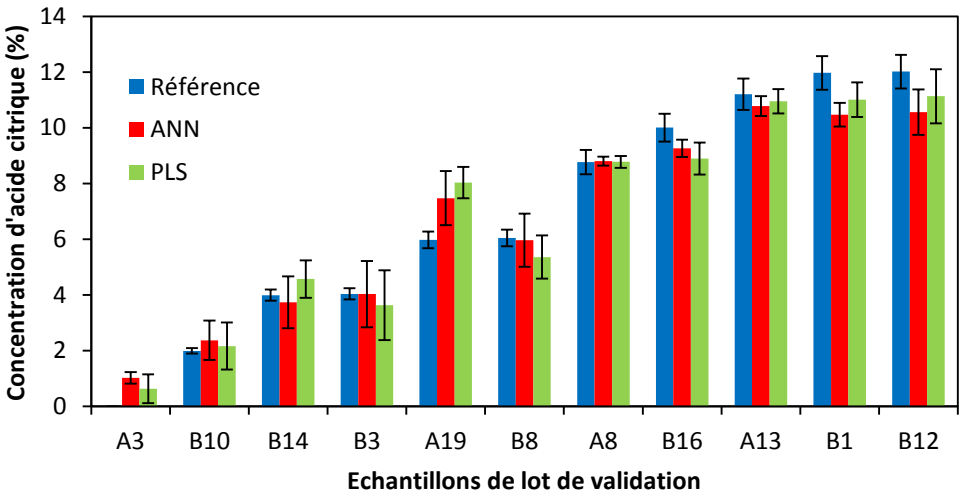


Figure 3-21 Concentrations en acide citrique des échantillons du lot de validation : Bleu : valeurs de référence (pesage), rouge : ANN, vert : PLS. Les barres d'erreur pour ANN et PLS sont données par l'écart-type sur les 10 spectres d'un même échantillon. Celles des données de référence correspondent à une erreur relative fixée arbitrairement à 5%.

Pour compléter cette analyse, la Figure 3-22 donne les résultats de RMSE obtenus pour les échantillons du lot de test. On remarque que la prédiction par ANN est toujours plus proche des valeurs de référence que celle effectuée par PLS, ce qui confirme simplement le résultat précédent établi dans la gamme allant de 2 à 10%. Pour l'échantillon B18, on retrouve cependant une valeur sous-estimée par ANN comme on l'avait observé pour le lot de validation. Par ailleurs les variations importantes observées pour l'échantillon B13 en PLS ne sont pas observées lors du traitement par ANN.

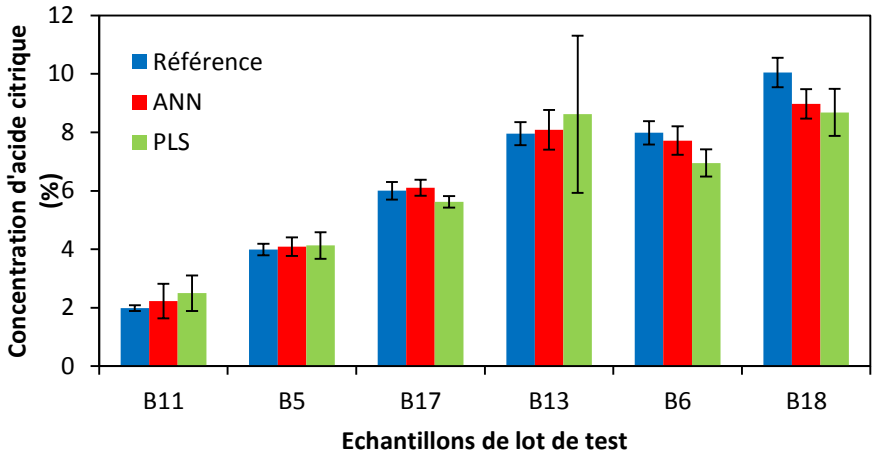


Figure 3-22 Concentrations en acide citrique des échantillons du lot de test : Bleu : valeurs de référence (pesage), rouge : ANN, vert : PLS. Les barres d'erreur pour ANN et PLS sont données par l'écart-type sur les 10 spectres d'un même échantillon. Celles des données de référence correspondent à une erreur relative fixée arbitrairement à 5%.

3.6.3 Analyse quantitative multiéléments à 3 sorties par ANN

Dans cette étude, nous avons utilisé un réseau de neurones avec 3 sorties afin de prédire simultanément les concentrations de lactose, fructose et acide citrique. Après une étude préliminaire, nous avons conclu que la bande spectrale optimale était comprise entre 0 et 2,6 THz pour le calcul d'ACP. Les trois premières composantes principales de l'ACP sont utilisées pour former les données d'entrée de l'ANN. Le Tableau 3-14 présente les résultats obtenus pour les trois lots d'échantillons, à savoir 190 spectres en calibration, 110 en validation et 60 en test.

		Lot d'apprentissage	Lot de validation	Lot de test
Fructose	RMSE (%)	0,6	1,0	0,7
	Q^2	0,99	0,99	0,97
	R^2	1,00	0,99	0,98
Lactose	RMSE (%)	0,6	0,6	0,4
	Q^2	1,00	0,99	0,99
	R^2	1,00	0,99	0,99
Acide citrique	RMSE (%)	0,9	1,1	0,9
	Q^2	0,99	0,98	0,88
	R^2	0,99	0,98	0,98

Tableau 3-14 Résultats obtenus avec un réseau de neurones à 3 sorties, pour 3 entrées (3 premières composantes principales), 3 neurones dans la couche cachée, vitesse d'apprentissage =0.05, terme de mémoire = 0.1 et nombre des itérations =18000.

Les résultats obtenus pour les RMSE sont remarquablement faibles, ce qui montre que l'ANN est tout à fait performant pour évaluer les 3 concentrations simultanément. Cependant, nous avons pu vérifier que les prédictions réalisées à l'aide d'un réseau de neurones à une seule sortie sont toujours meilleures que celles issues d'un réseau à trois sorties. Ce résultat pourrait être expliqué par le fait que la méthode de rétro-propagation de l'erreur qui permet à l'algorithme d'ANN de converger s'appuie sur deux calculs d'erreur très différents. Pour un ANN à une seule sortie, l'erreur est simplement la différence entre la valeur prédite et la valeur de référence. Pour un ANN à trois sorties, on calcule la moyenne des erreurs pour les trois analytes. Ce dernier calcul conduit donc à trouver un compromis qui vise à minimiser trois paramètres au lieu d'un seul.

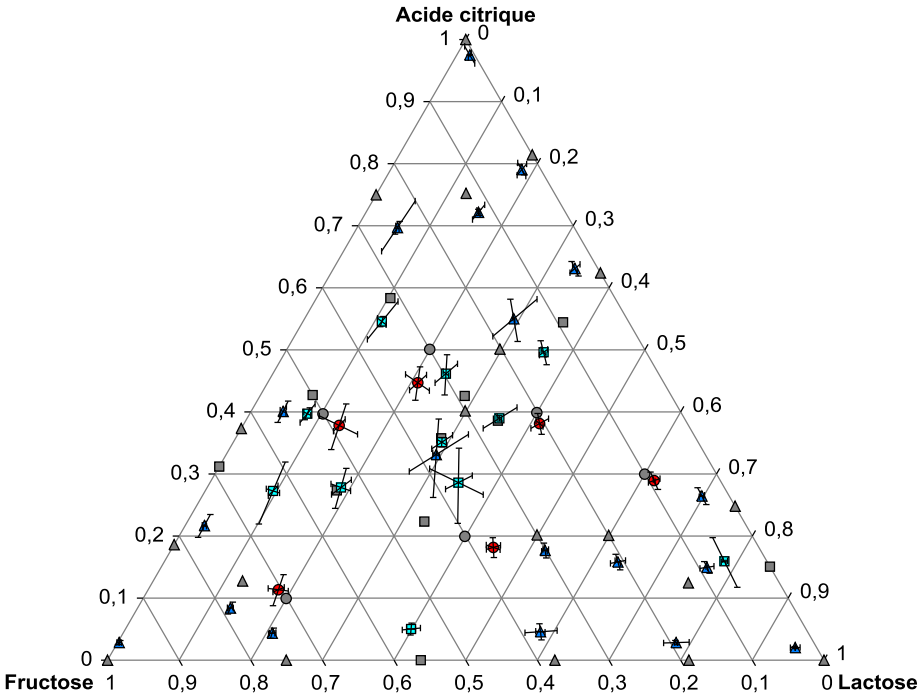


Figure 3-23 Diagramme ternaire calculé par ANN à partir des données THz. Les triangles bleus présentent les échantillons du lot d'apprentissage, les carrés bleu clair présentent les échantillons du lot de validation, les disques rouges présentent les échantillons du lot de test. Les barres présentent les écart-types de prédiction des 10 spectres par échantillon. La couleur grise présente le pourcentage de référence (pesage) de l'échantillon.

Et puisque le réseau de neurones fournit directement 3 valeurs en sortie, il est possible de placer chaque point dans un diagramme ternaire comparable à celui qui a été établi au départ de cette étude. Les trois pôles sont alors le fructose, le lactose et l'acide citrique. Comme l'ANN, à travers la fonction d'activation qui est la fonction sigmoïde, fournit toujours une réponse entre 0 et 1, il n'y a pas besoin d'un calcul supplémentaire pour placer les échantillons dans le diagramme ternaire. Les résultats obtenus sont présentés sur la Figure 3-23. On peut en déduire que les valeurs prédites sont relativement éloignées des valeurs de référence, ce qui dégrade la capacité de prédiction du modèle ANN à trois sorties. En comparant ces résultats à ceux obtenus précédemment dans le cas de modèles ANN à une seule sortie, on conclue que la prédiction des concentrations par ANN est plus précise lorsqu'on utilise un modèle à une seule sortie.

3.7 **Conclusion**

Nous avons pu appliquer plusieurs méthodes de chimiométrie jusqu'ici utilisée en spectroscopie LIBS (cf. chap. 2) aux données d'absorbance THz. L'ACP a permis de décrire correctement la composition de la matrice des échantillons une fois les spectres dérivés et les

données centrées. En plus d'une simple description qualitative des données, l'ACP a aussi permis d'obtenir des résultats semi-quantitatifs. Enfin, l'ACP a été efficacement exploitée pour compresser les données spectrales dans le but de fabriquer des données d'entrée du réseau de neurones qui contiennent un maximum d'information. La PLS et l'ANN (tout deux à simple sortie) ont permis de quantifier la concentration en lactose, puis en fructose et enfin en acide citrique des échantillons avec une erreur relative de prédiction souvent inférieure à 5%. Rappelons cependant que le spectre du lactose qui se comporte de façon très linéaire avec la concentration est correctement analysé par PLS tandis que l'ANN apporte un avantage léger dans l'analyse de l'acide citrique qui présente des interférences spectrales avec le fructose.

Finalement, dans le cadre du mélange ternaire étudié ici, l'avantage de l'ANN n'est pas démontré et l'analyse PLS est tout à fait satisfaisante. Le fait que l'analyse par ANN donne des résultats comparables à ceux obtenus par PLS mais pas meilleurs confirme le fait que la relation entre l'absorbance THz et les concentrations des trois analytes est multi-linéaire et que les effets non-linéaires sont négligeables. On préfèrera dans ce cas la PLS à l'ANN étant donnée la complexité d'optimisation de ce dernier. Cependant, la méthodologie mise en œuvre pour le calcul ANN pourrait être transposée à un autre type de mélanges ternaires voire plus complexes et dans ce cas, l'ANN pourrait certainement apporter une amélioration pour l'analyse quantitative des données THz.

Conclusion

Nous avons présenté dans ce mémoire de thèse deux exemples de spectroscopie – LIBS et THz – pour lesquels la chimiométrie apporte une valeur ajoutée incontestable à l'analyse des données. Rappelons que ce travail de thèse a fait l'objet de deux contrats de recherche. Le premier est un contrat signé avec l'ADEME dans le cadre d'un projet à finalité de transfert industriel intitulé CALIPSO et dédié à l'analyse LIBS sur site de sols pollués en métaux lourds (2011-2012). Un consortium entre le BRGM, spécialiste en géosciences, IVEA entreprise privée commercialisant des systèmes LIBS de terrain et le LOMA pour l'analyse des données a ainsi été constitué pour une durée de deux ans. Le travail engagé a ensuite été complété par un contrat de recherche d'un an (2013) entre le LOMA et l'entreprise IVEA. Dans ce contexte, la partie LIBS représente de loin la partie la plus importante du travail effectué.

Une banque de données très intéressante sur les sols a pu être créée à partir de plusieurs sites géographiques. D'autre part, un algorithme de réseau de neurones artificiel (ANN) à trois couches a été écrit de A à Z et transféré à l'entreprise IVEA dans le but d'une utilisation en routine sur site. Pour cela, il a été implanté au cœur du logiciel Analibs qui permet de piloter l'instrument LIBS puis d'enregistrer et de traiter les données. Nous avons démontré qu'une analyse univariée était totalement inappropriée dans le cas de sols prélevés sur le terrain. En revanche, l'analyse des données par ANN a donné des résultats très satisfaisants pour des mesures sur site, à savoir moins de 20% d'écart relatif de prédiction, aussi bien pour des éléments mineurs que pour des majeurs. Dans ce cadre, nous avons démontré qu'il était nécessaire d'injecter en entrée de l'ANN des données autres que celles directement reliées à l'analyte afin de prendre en compte les effets de matrice. De plus, nous avons mis à jour avec l'exemple du dosage du plomb la difficulté de l'ANN à prédire des concentrations sur une très large gamme de valeurs de concentrations. Cela nous a conduit à proposer deux modèles ANN, l'un pour les faibles concentrations et l'autre pour les fortes. La conséquence était alors qu'il fallait être capable de décider pour un échantillon inconnu si on devait lui appliquer un modèle ANN ou un autre. Nous avons alors développé un ANN spécialement dédié à cette prise de décision, à savoir un ANN qui fournit une valeur de sortie binaire permettant de trier entre deux classes.

Nous avons choisi de faire progresser notre travail en analyse LIBS par paliers de difficulté croissante. Ainsi, nous avons dans un premier temps traité les données provenant d'un seul site (SLM) avant de considérer la question beaucoup plus compliquée d'une analyse globale de plusieurs sites. Dans ce dernier cas, nous avons mis en œuvre un ANN avec trois sorties pour une analyse semi-quantitative. Plus précisément, la première valeur est le résultat d'un calcul des concentrations cumulées de Al-Si correspondant au pôle silicaté, bien connu en géosciences. De même, la deuxième valeur concerne le groupe Ca-Mg associé aux sols calcaires et enfin la dernière valeur prend en compte différents métaux tels que Pb-Zn-Ba dans le but de décrire le pôle minerai. Grâce à cette approche, on dispose d'une présentation des échantillons de sols dans un diagramme ternaire et il est par conséquent facile de déterminer si un sol inconnu est plutôt associé à une matrice de type alumino-silicate, calcaire ou minerai. Pour des performances analytiques acceptables, il est important de classer les échantillons avant l'analyse quantitative. On peut ensuite appliquer le modèle quantitatif approprié à la classe déterminée et donc, par cette approche globale, on peut quantifier un échantillon inconnu. Ce travail a permis de démontrer que des analyses LIBS quantitatives sur site peuvent être réalisées en routine grâce à l'utilisation de réseaux de neurones artificiels.

A l'issue de ce travail, nous pouvons conclure que le point qui reste à améliorer est le choix des données d'entrée de l'ANN. Rappelons ici que l'on doit injecter un petit nombre de données en entrée de l'ANN. Il faut donc sélectionner quelques données parmi toutes celles qui constituent le spectre LIBS. Nous avons mené la plupart de nos études à partir d'un choix préalable de quelques raies atomiques motivé par les informations trouvées sur la base de données du NIST. Cependant, dans le but de mettre au point une méthode permettant de réduire les effets dus à la sélection des raies spectrales, nous avons étudié la possibilité de compresser les données d'entrée par l'intermédiaire d'un calcul de PLS. Les résultats obtenus ont été partiellement satisfaisants mais la généralisation n'est pas démontrée et finalement, la question du choix optimal des données d'entrée reste ouverte. Notons que la sélection des données d'entrée à partir de la base de données du NIST a permis de donner une interprétation physique aux résultats des calculs et ainsi le modèle n'est pas une « boîte noire ». Enfin, même si les performances globales des ANNs sont satisfaisantes, on peut envisager en guise de perspectives de ce travail de tester la méthode SVM (support vector machine) non seulement pour les analyses quantitatives mais aussi pour de la classification. L'avantage principal de cette technique est son aptitude à traiter un petit nombre de données, contrairement à l'ANN. Ceci fait de la SVM la technique idéale tant que le nombre de données est relativement petit en comparaison du nombre de poids ajustables dans l'ANN.

Notons enfin que le logiciel développé dans le cadre de ce travail a été validé à l'aide d'un logiciel de calcul scientifique commercial. Ceci a permis de détecter les erreurs de programmation et les problèmes de convergence de l'algorithme. Finalement, le module ANN implanté dans le logiciel Analibs s'est avéré tout à fait satisfaisant de par sa stabilité et sa capacité à fournir des résultats conformes à ceux obtenus avec un logiciel commercial.

L'analyse des données LIBS par ANN a servi de base de travail pour l'étude ultérieure des données THz qui a été présentée dans le chapitre 3. Par extension, la méthodologie qui a été mise au point peut potentiellement être transposée à n'importe quel domaine de spectroscopie. En ce qui concerne la spectroscopie THz, les données à traiter ne sont plus des raies atomiques mais au contraire de bandes spectrales très peu contrastées. Afin de mener une première étude THz, nous avons fait le choix de préparer une série d'échantillons composés de mélanges ternaires lactose-fructose-acide citrique liés avec du polyéthylène et préparés sous forme de pastilles. En traitant les données THz par ACP, les deux premières composantes principales ont permis de mener avec succès une analyse semi-quantitative. De plus, deux analyses quantitatives, l'une par PLS et l'autre par ANN ont donné chacune de très bons résultats. L'une des perspectives de ce travail consiste à analyser par spectroscopie THz des spectres d'échantillons réels tels que ceux de produits pharmaceutiques sachant que l'un des avantages de la spectroscopie THz est sa capacité à discriminer les forme chirales qui jouent un rôle spécifique dans les principes actifs.

Tout au long de ce travail, nous avons adopté la méthodologie en usage en chimiométrie qui permet d'évaluer les performances d'un modèle et de valider sa pertinence statistique. Plus précisément, nous avons systématiquement réparti les données en trois lots : un lot de calibration pour construire le modèle, un lot de validation pour déterminer le meilleur modèle et éviter les risques de sur-apprentissage et enfin un lot de test visant à évaluer les performances du modèle retenu a posteriori. De même, la méthode connue sous le nom de Y-randomization a été systématiquement appliquée afin de vérifier que les résultats obtenus avaient une vraie signification statistique.

Plus généralement, le travail de chimiométrie appliquée à la spectroscopie exposé dans ce mémoire de thèse peut être potentiellement transposé à bien d'autres types de spectroscopies. De plus, la chimiométrie ouvre la voie très prometteuse de la fusion des données qui consiste à analyser simultanément des données spectrales provenant de différentes techniques telles que LIBS, THz, Raman, NIR ou autres. C'est dans cette direction que s'orientent désormais les travaux du groupe de recherche dans lequel j'ai effectué ce travail de thèse.

Bibliographie

[1] R. Kowalski Bruce, Chemometrics: Theory and Application, Preface, in, American Chemical Society, 1977.
[2] S. Wold, Chemometrics; what do we mean with it, and what do we want from it?, Chemometrics and Intelligent Laboratory Systems, 30 (1995) 109-115.
[3] C.-C. Wang, C.-C. Lu, Y.-L. Chen, H.-L. Cheng, S.-M. Wu, Chemometric Optimization of Cation-Selective Exhaustive Injection Sweeping Micellar Electrokinetic Chromatography for Quantification of Ractopamine in Porcine Meat, Journal of Agricultural and Food Chemistry, 61 (2013) 5914-5920.
[4] S. Ruiz-Castelar, A. Checa, R. Gargallo, J. Jaumot, Combination of chromatographic and chemometric methods to study the interactions between DNA strands, Analytica chimica acta, 722 (2012) 34-42.
[5] G. Hanrahan, F.A. Gomez, Chemometric Methods in Capillary Electrophoresis, Wiley, 2009.
[6] E. Gerbino, P. Mobili, E.E. Tymczyszyn, C. Frausto-Reyes, C. Araujo-Andrade, A. Gomez-Zavaglia, Use of Raman spectroscopy and chemometrics for the quantification of metal ions attached to Lactobacillus kefir, Journal of applied microbiology, 112 (2012) 363-371.
[7] B. Pavoni, N. Rado, R. Piazza, S. Frignani, FT-IR Spectroscopy and Chemometrics as a Useful Approach for Determining Chemical-Physical Properties of Gasoline, by Minimizing Analytical Times and Sample Handling, Annali di Chimica, 94 (2004) 521-532.
[8] Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, N. Jent, A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies, Journal of Pharmaceutical and Biomedical Analysis, 44 (2007) 683-700.
[9] L.J. Janik, D. Cozzolino, R. Dambergs, W. Cynkar, M. Gishen, The prediction of total anthocyanin concentration in red-grape homogenates using visible-near-infrared spectroscopy and artificial neural networks, Analytica chimica acta, 594 (2007) 107-118.
[10] H. Wu, E.J. Heilweil, A.S. Hussain, M.A. Khan, Process analytical technology (PAT): Quantification approaches in terahertz spectroscopy for pharmaceutical application, Journal of Pharmaceutical Sciences, 97 (2008) 970-984.
[11] J.L. Gottfried, D.A. Cremers, L.J. Radziemski, Chemometric Analysis in LIBS, in: Handbook of Laser-Induced Breakdown Spectroscopy, John Wiley & Sons Ltd, 2013, pp. 223-255.
[12] G. Downey, Food and food ingredient authentication by mid-infrared spectroscopy and chemometrics, TrAC Trends in Analytical Chemistry, 17 (1998) 418-424.

[13] F. Marini, Chemometrics in Food Chemistry, 1st Edition, in: M. Federico (Ed.), Elsevier, 2013.
[14] G.W. Johnson, R. Ehrlich, State of the Art Report on Multivariate Chemometric Methods in Environmental Forensics, Environmental Forensics, 3 (2002) 59-79.
[15] J. Mocák, Chemometrics in Medicine and Pharmacy, Nova Biotechnologica et Chimica, 11 (2012).
[16] J. Chen, K.-C. Liu, On-line batch process monitoring using dynamic PCA and dynamic PLS models, Chemical Engineering Science, 57 (2002) 63-75.
[17] X. Cetó, A.M. O'Mahony, I.A. Samek, J.R. Windmiller, M. del Valle, J. Wang, Rapid Field Identification of Subjects Involved in Firearm-Related Crimes Based on Electroanalysis Coupled with Advanced Chemometric Data Treatment, Analytical Chemistry, 84 (2012) 10306-10314.
[18] C. Muehlethaler, G. Massonnet, P. Esseiva, The application of chemometrics on Infrared and Raman spectra as a tool for the forensic analysis of paints, Forensic Science International, 209 (2011) 173-182.
[19] E. Manzano, J. García-Atero, A. Dominguez-Vidal, M.J. Ayora-Cañada, L.F. Capitán-Vallvey, N. Navas, Discrimination of aged mixtures of lipidic paint binders by Raman spectroscopy and chemometrics, Journal of Raman Spectroscopy, 43 (2012) 781-786.
[20] J. Moros, J. Serrano, C. Sanchez, J. Macias, J.J. Laserna, New chemometrics in laser-induced breakdown spectroscopy for recognizing explosive residues, Journal of Analytical Atomic Spectrometry, 27 (2012) 2111-2122.
[21] M.J. Adams, R.S.o. Chemistry, Chemometrics In Analytical Spectroscopy, Royal Society of Chemistry, 2004.
[22] R. Noll, Laser-Induced Breakdown Spectroscopy: Fundamentals and Applications, Springer.
[23] A.W. Andrzej, V. Palleschi, I. Schechter, Laser-induced Breakdown Spectroscopy (Libs): Fundamentals and Applications, 2006.
[24] O. Forni, S. Maurice, O. Gasnault, R.C. Wiens, A.s. Cousin, S.M. Clegg, J.-B. Sirven, J.r.m. Lasue, Independent component analysis classification of laser induced breakdown spectroscopy spectra, Spectrochimica Acta Part B: Atomic Spectroscopy, (2013).
[25] A.K. Myakalwar, S. Sreedhar, I. Barman, N.C. Dingari, S. Venugopal Rao, P. Prem Kiran, S.P. Tewari, G. Manoj Kumar, Laser-induced breakdown spectroscopy-based investigation and classification of pharmaceutical tablets using multivariate chemometric analysis, Talanta, 87 (2011) 53-59.
[26] S. Balakrishnama, A. Ganapathiraju, Linear discriminant analysis -a brief tutorial.
[27] R.A. Multari, D.A. Cremers, J.M. Dupre, J.E. Gustafson, The Use of Laser-Induced Breakdown Spectroscopy for Distinguishing Between Bacterial Pathogen Species and Strains, Appl. Spectrosc., 64 (2010) 750-759.
[28] Q. Godoi, F.O. Leme, L.C. Trevizan, E.R. Pereira Filho, I.A. Rufini, D. Santos Jr, F.J. Krug, Laser-induced breakdown spectroscopy and chemometrics for classification of toys relying on toxic elements, Spectrochimica Acta Part B: Atomic Spectroscopy, 66 (2011) 138-143.
[29] R. Pan, S. Zhao, J. Shen, Terahertz spectra applications in identification of illicit drugs using support vector machines, Procedia Engineering, 7 (2010) 15-21.
[30] N.C. Dingari, I. Barman, A.K. Myakalwar, S.P. Tewari, M. Kumar Gundawar, Incorporation of Support Vector Machines in the LIBS Toolbox for Sensitive and Robust Classification Amidst Unexpected Sample and System Variability, Analytical Chemistry, 84 (2012) 2686-2694.
[31] Y. Pao, Adaptive pattern recognition and neural networks, (1989).

[32] P. Yaroshchyk, D.L. Death, S.J. Spencer, Comparison of principal components regression, partial least squares regression, multi-block partial least squares regression, and serial partial least squares regression algorithms for the analysis of Fe in iron ore using LIBS, Journal of Analytical Atomic Spectrometry, 27 (2012) 92-98.
[33] Y. Ma, Q. Wang, L. Li, PLS model investigation of thiabendazole based on THz spectrum, Journal of Quantitative Spectroscopy and Radiative Transfer, 117 (2012) 7-14.
[34] Y. Hua, H. Zhang, H. Zhou, Quantitative determination of cyfluthrin in n-hexane by terahertz time-domain spectroscopy with chemometrics methods, Instrumentation and Measurement, IEEE Transactions on, 59 (2010) 1414-1423.
[35] J. El Haddad, M. Villot-Kadri, A. Ismaël, G. Gallou, K. Michel, D. Bruyère, V. Laperche, L. Canioni, B. Bousquet, Artificial neural network for on-site quantitative analysis of soils using laser induced breakdown spectroscopy, Spectrochimica Acta Part B: Atomic Spectroscopy, 79-80 (2013) 51-57.
[36] H. Wold, Estimation of principal components and related models by iterative least squares, Journal of Multivariate Analysis, (1966).
[37] H. Martens, T. Naes, Multivariate Calibration, Wiley, 1991.
[38] B. Bousquet, J.B. Sirven, L. Canioni, Towards quantitative laser-induced breakdown spectroscopy analysis of soil samples, Spectrochimica Acta Part B: Atomic Spectroscopy, 62 (2007) 1582-1589.
[39] A.C. Samuels, F.C. DeLucia Jr, K.L. McNesby, A.W. Miziolek, Laser-Induced Breakdown Spectroscopy of Bacterial Spores, Molds, Pollens, and Protein: Initial Studies of Discrimination Potential, Appl. Opt., 42 (2003) 6205-6209.
[40] S.M. Clegg, E. Sklute, M.D. Dyar, J.E. Barefield, R.C. Wiens, Multivariate analysis of remote laser-induced breakdown spectroscopy spectra using partial least squares, principal component analysis, and related techniques, Spectrochimica Acta Part B: Atomic Spectroscopy, 64 (2009) 79-88.
[41] S. Jantzi, J.R. Almirall, Characterization and forensic analysis of soil samples using laser-induced breakdown spectroscopy (LIBS), Analytical and Bioanalytical Chemistry, 400 (2011) 3341-3351.
[42] S. Nakajima, H. Hoshina, M. Yamashita, C. Otani, N. Miyoshi, Terahertz imaging diagnostics of cancer tissues with a chemometrics technique, Applied Physics Letters, 90 (2007) 041102-041102-041103.
[43] Y. Shen, P.F. Taday, M.C. Kemp, Terahertz spectroscopy of explosive materials, (2004) 82-89.
[44] S.-J. Choi, K.-J. Lee, J. Yoh, Quantitative laser-induced breakdown spectroscopy of standard reference materials of various categories, Applied Physics B, (2013) 1-10.
[45] M.D. King, W.D. Buchanan, T.M. Korter, Identification and Quantification of Polymorphism in the Pharmaceutical Compound Diclofenac Acid by Terahertz Spectroscopy and Solid-State Density Functional Theory, Analytical Chemistry, 83 (2011) 3786-3792.
[46] M.D. King, T.M. Korter, Effect of Waters of Crystallization on Terahertz Spectra: Anhydrous Oxalic Acid and Its Dihydrate, The Journal of Physical Chemistry A, 114 (2010) 7127-7138.
[47] J. El Haddad, B. Bousquet, L. Canioni, P. Mounaix, Terahertz spectral analysis, TrAC Trends in Analytical Chemistry, (2013).
[48] C.B. Stipe, B.D. Hensley, J.L. Boersema, S.G. Buckley, Laser-Induced Breakdown Spectroscopy of Steel: A Comparison of Univariate and Multivariate Calibration Methods, Appl. Spectrosc., 64 (2010) 154-160.
[49] J. Amador-Hernandez, L.E. Garcia-Ayuso, J.M. Fernandez-Romero, M.D. Luque de Castro, Partial least squares regression for problem solving in precious metal analysis by laser

induced breakdown spectrometry, Journal of Analytical Atomic Spectrometry, 15 (2000) 587-593.
[50] I.S. Helland, On the structure of partial least squares regression, Communications in Statistics - Simulation and Computation, 17 (1988) 581-607.
[51] M.Z. Martin, M.A. Mayes, K.R. Heal, D.J. Brice, S.D. Wullschleger, Investigation of laser-induced breakdown spectroscopy and multivariate analysis for differentiating inorganic and organic C in a variety of soils, Spectrochimica Acta Part B: Atomic Spectroscopy, (2013).
[52] Z. Wang, T.-B. Yuan, S.-L. Lui, Z.-Y. Hou, X.-W. Li, Z. Li, W.-D. Ni, Major elements analysis in bituminous coals under different ambient gases by laser-induced breakdown spectroscopy with PLS modeling, Frontiers of Physics, 7 (2012) 708-713.
[53] K. Ishibashi, T. Arai, K. Wada, M. Kobayashi, S. Ohno, H. Senshu, N. Namiki, T. Matsui, S. Kameda, Y. Cho, Analysis Method for Minerals with Laser-Induced Breakdown Spectroscopy (LIBS) for In-Situ Lunar Mineral Measurement, in: Lunar and Planetary Institute Science Conference Abstracts, 2012, pp. 1786.
[54] R.C. Wiens, S. Maurice, J. Lasue, O. Forni, R.B. Anderson, S. Clegg, S. Bender, D. Blaney, B.L. Barraclough, A. Cousin, L. Deflores, D. Delapp, M.D. Dyar, C. Fabre, O. Gasnault, N. Lanza, J. Mazoyer, N. Melikechi, P.Y. Meslin, H. Newsom, A. Ollila, R. Perez, R.L. Tokar, D. Vaniman, Pre-flight calibration and initial data processing for the ChemCam laser-induced breakdown spectroscopy instrument on the Mars Science Laboratory rover, Spectrochimica Acta Part B: Atomic Spectroscopy, 82 (2013) 1-27.
[55] P.-Y. Meslin, A. Cousin, G. Berger, O. Forni, O. Gasnault, J. Lasue, N. Mangold, S. Schröder, S. Maurice, R. Wiens, ChemCam Analysis of Soil Diversity along Bradbury-Glenelg Traverse, in: EGU General Assembly Conference Abstracts, 2013, pp. 11711.
[56] J.B. Sirven, B. Bousquet, L. Canioni, L. Sarger, S. Tellier, M. Potin-Gautier, I.L. Hecho, Qualitative and quantitative investigation of chromium-polluted soils by laser-induced breakdown spectroscopy combined with neural networks analysis, Analytical and Bioanalytical Chemistry, 385 (2006) 256-262.
[57] Y. LeCun, Efficient learning and second-order methods. A tutorial at NIPS 93, Denver, (1993).
[58] S.O. Haykin, Neural Networks and Learning Machines: International Version, Third Edition ed., Upper Saddle River [etc.] Pearson Education, 2009.
[59] R. Asadi, N. Mustapha, N. Sulaiman, N. Shiri, New Supervised Multi Layer Feed Forward Neural Network Model to Accelerate Classification with High Accuracy, European Journal of Scientific Research, 33 (2009) 163-178.
[60] K. Keeni, K. Nakayama, H. Shimodaira, A training scheme for pattern classification using multi-layer feed-forward neural networks, in: Computational Intelligence and Multimedia Applications, 1999. ICCIMA '99. Proceedings. Third International Conference on, 1999, pp. 307-311.
[61] T.-C. Chang, R.-J. Chao, Application of back-propagation networks in debris flow prediction, Engineering Geology, 85 (2006) 270-280.
[62] M. Boueri, Laser-induced plasma on polymeric materials and applications for the discrimination and identification of plastics, in: Ecole doctorale de Physique et d'Astrophysique, Université Claude Bernard, Lyon 1, Lyon, 2010.
[63] J.-B. Sirven, Détection de métaux lourds dans les sols par spectroscopie démission sur plasma induit par laser (LIBS), in: École de sciences physiques et de l'ingénieur, Université Bordeaux 1, Bordeaux France, 2006.
[64] D. Diego-Vallejo, D. Ashkenasi, A. Lemke, H.-J. Eichler, Selective ablation of Copper-Indium-Diselenide (CIS) solar cells monitored by laser-induced breakdown spectroscopy and classification methods, Spectrochimica Acta Part B: Atomic Spectroscopy, (2013).

[65] A. Tropsha, P. Gramatica, V.K. Gombar, The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, QSAR & Combinatorial Science, 22 (2003) 69-77.
[66] L. Eriksson, J. Jaworska, A.P. Worth, M.T. Cronin, R.M. McDowell, P. Gramatica, Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs, Environ Health Perspect, 111 (2003) 1361-1375.
[67] A. Golbraikh, A. Tropsha, Beware of q2!, J Mol Graph Model, 20 (2002) 269-276.
[68] W.T.Y. Mohamed, Improved LIBS limit of detection of Be, Mg, Si, Mn, Fe and Cu in aluminum alloy samples using a portable Echelle spectrometer with ICCD camera, Optics & Laser Technology, 40 (2008) 30-38.
[69] E. Cerrai, R. Trucco, The matrix effect in laser sampled spectrochemical analysis, Journal Name: Energ. Nucl. (Milan), 15: 581-7(Sept. 1968).; Other Information: Orig. Receipt Date: 31-DEC-68, (1968) Medium: X.
[70] K.W. Marich, P.W. Carr, W.J. Treytl, D. Glick, Effect of matrix material on laser-induced elemental spectral emission, Analytical Chemistry, 42 (1970) 1775-1779.
[71] W.T.Y. Mohamed, A. Askar, Study of the matrix effect on the plasma characterization of heavy elements in soil sediments using LIBS with a portable echelle spectrometer. , Progress in Physics. Progress in Physics., 1 (2007) 46-52.
[72] B.J. Taylor, M.A. Darrah, C.D. Moats, Verification and validation of neural networks: a sampling of research in progress, (2003) 8-16.
[73] A.G. GLAROS, R.B. KLINE, Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model, Journal of Clinical Psychology, 44 (1988) 1013–1023.
[74] J.-B. Sirven, B. Salle, P. Mauchien, J.-L. Lacour, S. Maurice, G. Manhes, Feasibility study of rock identification at the surface of Mars by remote laser-induced breakdown spectroscopy and three chemometric methods, Journal of Analytical Atomic Spectrometry, 22 (2007) 1471-1480.
[75] H.A. Martens, P. Dardenne, Validation and verification of regression in small data sets, Chemometrics and Intelligent Laboratory Systems, 44 (1998) 99-121.
[76] C. Rucker, G. Rucker, M. Meringer, y-Randomization and its variants in QSPR/QSAR, J Chem Inf Model, 47 (2007) 2345-2357.
[77] H. Kubinyi, Comparative Molecular Field Analysis (CoMFA), in: Handbook of Chemoinformatics, Wiley-VCH Verlag GmbH, 2008, pp. 1555-1574.
[78] A. Shrivastava, V.B. Gupta, Methods for the determination of limit of detection and limit of quantitation of the analytical methods, Chronicles of Young Scientists, 2 (2011).
[79] M. Ostra, C. Ubide, M. Vidal, J. Zuriarrain, Detection limit estimator for multivariate calibration by an extension of the IUPAC recommendations for univariate methods, Analyst, 133 (2008) 532-539.
[80] J.B. Sirven, B. Bousquet, L. Canioni, L. Sarger, Laser-Induced Breakdown Spectroscopy of Composite Samples: Comparison of Advanced Chemometrics Methods, Analytical Chemistry, 78 (2006) 1462-1469.
[81] J. Debras-Guédon, N. Liodec, De l'utilisation du faisceau d'un amplificateur a ondes lumineuses par émission induite de rayonnement (laser à rubis), comme source énergétique pour l'excitation des spectres d'émission des éléments, CR. Acad. Sci. , 257 (1963).
[82] A.S. Eppler, D.A. Cremers, D.D. Hickmott, M.J. Ferris, A.C. Koskelo, Matrix Effects in the Detection of Pb and Ba in Soils Using Laser-Induced Breakdown Spectroscopy, Appl. Spectrosc., 50 (1996) 1175-1181.
[83] L.J. Radziemski, T.R. Loree, D.A. Cremers, N.M. Hoffman, Time-resolved laser-induced breakdown spectrometry of aerosols, Analytical Chemistry, 55 (1983) 1246-1252.

[84] D.A. Cremers, L.J. Radziemski, Handbook of Laser-Induced Breakdown Spectroscopy, Wiley, 2006.
[85] D.A. Cremers, F.-Y. Yueh, J.P. Singh, H. Zhang, Laser-Induced Breakdown Spectroscopy, Elemental Analysis, in: Encyclopedia of Analytical Chemistry, John Wiley & Sons, Ltd, 2006.
[86] R. Noll, V. Sturm, Ü. Aydin, D. Eilers, C. Gehlen, M. Höhne, A. Lamott, J. Makowe, J. Vrenegor, Laser-induced breakdown spectroscopy—From research to industry, new frontiers for process control, Spectrochimica Acta Part B: Atomic Spectroscopy, 63 (2008) 1159-1166.
[87] R. Noll, Evaporation and Plasma Generation, in: Laser-Induced Breakdown Spectroscopy, Springer Berlin Heidelberg, 2012, pp. 75-82.
[88] E. Tognoni, V. Palleschi, M. Corsi, G. Cristoforetti, N. Omenetto, I. Gornushkin, B.W. Smith, J.D. Winefordner, From sample to signal in laser-induced breakdown spectroscopy: a complex route to quantitative analysis. Laser-Induced Breakdown Spectroscopy (LIBS), Cambridge University Press, 2006.
[89] A. Ciucci, M. Corsi, V. Palleschi, S. Rastelli, A. Salvetti, E. Tognoni, New Procedure for Quantitative Elemental Analysis by Laser-Induced Plasma Spectroscopy, Appl. Spectrosc., 53 (1999) 960-964.
[90] P. Higueras, R. Oyarzun, J.M. Iraizoz, S. Lorenzo, J.M. Esbrí, A. Martínez-Coronado, Low-cost geochemical surveys for environmental studies in developing countries: Testing a field portable XRF instrument under quasi-realistic conditions, Journal of Geochemical Exploration, 113 (2012) 3-12.
[91] R. Jenkins, Quantitative X-Ray Spectrometry, Second Edition, Taylor & Francis, 1995.
[92] D.J. Kalnicky, R. Singhvi, Field portable XRF analysis of environmental samples, Journal of Hazardous Materials, 83 (2001) 93-122.
[93] C.-M. Wu, H.-T. Tsai, K.-H. Yang, J.-C. Wen, How Reliable is X-Ray Fluorescence (XRF) Measurement for Different Metals in Soil Contamination?, Environmental Forensics, 13 (2012) 110-121.
[94] R. Jeannot, B. Lemière, S. Chiron, Guide méthodologique pour l'analyse des sols pollués, in, Documents du BRGM 298, Orléans, France, 2001.
[95] P.T. Michel Hoenig, Préparation d'échantillons de l'environnement pour analyse minérale, Ed. Techniques Ingénieur.
[96] B.J. Alloway, Environmental Pollution: Heavy Metals in Soils: Trace Metals and Metalloids in Soils and Their Bioavailability, Springer London, Limited, 2013.
[97] O.T. Butler, W.R.L. Cairns, J.M. Cook, C.M. Davidson, Atomic spectrometry update. Environmental analysis, Journal of Analytical Atomic Spectrometry, 27 (2007) 187-221.
[98] D.A. Skoog, F.J. Holler, T.A. Nieman, Principes d'analyse instrumentale, De Boeck, 2003.
[99] M.A. Gondal, T. Hussain, Z. Ahmed, A.H. Bakry, Detection of contaminants in ore samples using laser-induced breakdown spectroscopy, Journal of Environmental Science and Health, Part A, 42 (2007) 879-887.
[100] X. Hou, H.L. Peters, Z. Yang, K.A. Wagner, J.D. Batchelor, M.M. Daniel, B.T. Jones, Determination of Trace Metals in Drinking Water Using Solid-Phase Extraction Disks and X-ray Fluorescence Spectrometry, Appl. Spectrosc., 57 (2003) 338-342.
[101] P. Yaroshchyk, R.J.S. Morrison, D. Body, B.L. Chadwick, Quantitative determination of wear metals in engine oils using laser-induced breakdown spectroscopy: A comparison between liquid jets and static liquids, Spectrochimica Acta Part B: Atomic Spectroscopy, 60 (2005) 986-992.
[102] W.Q. Lei, J. El Haddad, V. Motto-Ros, N. Gilon-Delepine, A. Stankova, Q.L. Ma, X.S. Bai, L.J. Zheng, H.P. Zeng, J. Yu, Comparative measurements of mineral elements in milk

powders with laser-induced breakdown spectroscopy and inductively coupled plasma atomic emission spectroscopy, Anal Bioanal Chem, 400 (2011) 3303-3313.
[103] Inductively Coupled Plasma Atomic Emission Spectrometer (ICP-AES), in: C.A.I. Laboratory (Ed.), Concordia College Moordead, Minnesota 2012.
[104] B. Bousquet, G. Travaillé, A. Ismaël, L. Canioni, K. Michel-Le Pierrès, E. Brasseur, S. Roy, I. le Hecho, M. Larregieu, S. Tellier, M. Potin-Gautier, T. Boriachon, P. Wazen, A. Diard, S. Belbèze, Development of a mobile system based on laser-induced breakdown spectroscopy and dedicated to in situ analysis of polluted soils, Spectrochimica Acta Part B: Atomic Spectroscopy, 63 (2008) 1085-1090.
[105] L. Arroyo, T. Trejos, P.R. Gardinali, J.R. Almirall, Optimization and validation of a Laser Ablation Inductively Coupled Plasma Mass Spectrometry method for the routine analysis of soils and sediments, Spectrochimica Acta Part B: Atomic Spectroscopy, 64 (2009) 16-25.
[106] D. Santos Jr, L.C. Nunes, L.C. Trevizan, Q. Godoi, F.O. Leme, J.W.B. Braga, F.J. Krug, Evaluation of laser induced breakdown spectroscopy for cadmium determination in soils, Spectrochimica Acta Part B: Atomic Spectroscopy, 64 (2009) 1073-1078.
[107] M. Anna P.M, Review: Applications of single-shot laser-induced breakdown spectroscopy, Spectrochimica Acta Part B: Atomic Spectroscopy, 65 (2010) 185-191.
[108] V. Juvé, R. Portelli, M. Boueri, M. Baudelet, J. Yu, Space-resolved analysis of trace elements in fresh vegetables using ultraviolet nanosecond laser-induced breakdown spectroscopy, Spectrochimica Acta Part B: Atomic Spectroscopy, 63 (2008) 1047-1053.
[109] D. DÍAz, D.W. Hahn, A. Molina, Evaluation of Laser-Induced Breakdown Spectroscopy (LIBS) as a Measurement Technique for Evaluation of Total Elemental Concentration in Soils, Appl. Spectrosc., 66 (2012) 99-106.
[110] M.E. Essington, G.V. Melnichenko, M.A. Stewart, R.A. Hull, Soil Metals Analysis Using Laser-induced Breakdown Spectroscopy (libs), Soil Sci. Soc. Am. J., 73 (2009) 1469-1478.
[111] A. Koujelev, M. Sabsabi, V. Motto-Ros, S. Laville, S.L. Lui, Laser-induced breakdown spectroscopy with artificial neural network processing for material identification, Planetary and Space Science, 58 (2009) 682-690.
[112] S.-L. Lui, A. Koujelev, Accurate identification of geological samples using artificial neural network processing of laser-induced breakdown spectroscopy data, Journal of Analytical Atomic Spectrometry, 26 (2011) 2419-2427.
[113] P. Mukhono, K. Angeyo, A. Dehaye, M. Massop, K. Kaduki, Laser induced break down spectro-analysis an d characterization of environmental matrices utilizing multivariate chemo metrics, (2012).
[114] L. Chusseau, J. Demaison, J.L. Coutaz, Optoélectronique térahertz, EDP Sciences, 2008.
[115] P.F. Taday, Applications of terahertz spectroscopy to pharmaceutical sciences, Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 362 (2004) 351-363; discussion 363-354.
[116] D. Dragoman, M. Dragoman, Terahertz fields and applications, Progress in Quantum Electronics, 28 (2004) 1-66.
[117] H. Wu, E.J. Heilweil, A.S. Hussain, M.A. Khan, Process analytical technology (PAT): Effects of instrumental and compositional variables on terahertz spectral data quality to characterize pharmaceutical materials and tablets, International Journal of Pharmaceutics, 343 (2007) 148-158.
[118] J.A. Zeitler, P.F. Taday, D.A. Newnham, M. Pepper, K.C. Gordon, T. Rades, Terahertz pulsed spectroscopy and imaging in the pharmaceutical setting--a review, The Journal of pharmacy and pharmacology, 59 (2007) 209-223.

[119] Y.C. Shen, Terahertz pulsed spectroscopy and imaging for pharmaceutical applications: a review, Int J Pharm, 417 (2011) 48-60.
[120] Y. Hua, H. Zhang, Qualitative and Quantitative Detection of Pesticides With Terahertz Time-Domain Spectroscopy, Microwave Theory and Techniques, IEEE Transactions on, 58 (2010) 2064-2070.
[121] M. Otsuka, J. Nishizawa, J. Shibata, M. Ito, Quantitative evaluation of mefenamic acid polymorphs by terahertz-chemometrics, J Pharm Sci, 99 (2010) 4048-4053.
[122] R. Palermo, R.P. Cogdill, S.M. Short, J.K. Drennen Iii, P.F. Taday, Density mapping and chemical component calibration development of four-component compacts via terahertz pulsed imaging, Journal of Pharmaceutical and Biomedical Analysis, 46 (2008) 36-44.
[123] K. Lien Nguyen, T. Friscic, G.M. Day, L.F. Gladden, W. Jones, Terahertz time-domain spectroscopy and the quantitative monitoring of mechanochemical cocrystal formation, Nat Mater, 6 (2007) 206-209.