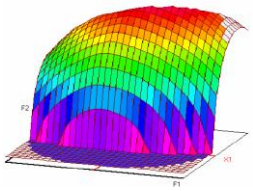
Université Sidi Mohammed Ben Abdellah
Faculté des Sciences et Techniques
www.fst-usmba.ac.ma

Année Universitaire : 2015-2016



Master Sciences et Techniques CAC Agiq
Chimiométrie et Analyse Chimique : Application à la gestion de la qualité

MEMOIRE DE FIN D'ETUDES
Pour l'Obtention du Diplôme de Master Sciences et Techniques

Développement de modèle statistique d'estimation du contenu organique et le taux d'azote total d'un sol agricole à partir de spectres d'absorbance proche infrarouge

Présenté par:

RABIE Reda

Encadré par:

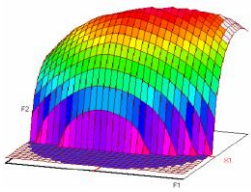
- **SAIDI Ouadi** **MAScIR**
- **SAFFAJ Taoufiq** **FST Fès**

Soutenu Le 14 Juin 2016 devant le jury composé de:

- **T. SAFFAJ**
- **E. M. EL HADRAMI**
- **A. FARAH**

Stage effectué à :MAScIR :Moroccan Foundation for Advanced Science, Innovation & Reserarch.

Université Sidi Mohammed Ben Abdellah
Faculté des Sciences et Techniques
www.fst-usmba.ac.ma



Master ST CAC Agiq

Mémoire de fin d'études pour l'obtention du Diplôme de Master Sciences et Techniques

Nom et prénom: Rabie Reda

Année Universitaire : 2015/2016

Titre: Développement de modèle statistique d'estimation du contenu organique et le taux l'azote total d'un sol agricole à partir de spectres d'absorbance proche infrarouge

Résumé

L'analyse quantitative de sol est très complexe, c est un défi de taille pour le chimiste. Généralement, pour être quantifiés, les constituants d'un sol doivent d'abord être isolés et purifiés. Toutefois, ces étapes de séparation et de purification sont souvent longues, fastidieuses et coûteuses. Le présent document traite d'une technique d'analyse statistique qui, lorsque couplée à une tchèque d'analyse chimique conventionnelle, permet d'évaluer les concentrations de tous les constituants d'un mélange complexe sans passer par les étapes de séparation et de purification. Cette analyse multivariable, appliquée ici à la technique de Spectroscopie infrarouge à transformée de Fourier (FTIR) précisément le PIR (proche infrarouge), comprend trois phases. Dans un premier temps, un modèle mathématique est construit à partir d'un ensemble d'échantillons de calibration de concentrations connues et de spectre infrarouge connus. Ce modèle doit ensuite être validé afin de permettre dans un troisième temps d'estimer les concentrations des divers constituants d'un inconnu à partir de son spectre infrarouge. La sélection d'une méthode expérimentale, l'évaluation de la faisabilité de celle-ci, l'élaboration du design expérimental, le choix d'un algorithme de calcul approprié et les procédés de validation du modèle qui en résulte font partis des étapes d'une démarche systématique permettant d'obtenir un modèle d'analyse quantitative précis et efficace

Ce rapport décrit des méthodes d'analyse innovatrice basée sur des techniques mathématiques de régression comme RLM, PCR et PLS. Ces techniques sont été appliqué avec succès à l'analyse Spectroscopique infrarouge compositionnelle de sol de région Rabat-Sallé et ce, sans traitement chimique de l'échantillon de sol. Dans les pages suivantes sont donnés une brève revue des différentes techniques mathématiques disponibles ainsi qu'un mode d'emploi pour l'application des méthodes RLM, PCR et PLS à n'importe quel type de sol de cette région.

Mots clés: sol, Matière organique, Azote totale, proche infrarouge, Chimiométrie, analyses multivariées,

Remerciements

Avant tout, je tiens à exprimer ma profonde reconnaissance à toutes les personnes qui ont contribué, par leur aide et assistance, à l'élaboration de ce travail.

Mes remerciements iront d'abord à mes tuteurs de stage à MAScIR, **SAIDI Ouadi** , **BOUZIDA Ilhame** , **Meftah KADMIRI Issam** et **YAAKOUBI Kaoutar** pour tout le temps qu'ils m'ont consacrée, et pour la qualité de leur suivi durant toute la période de mon stage.

Mes plus vifs remerciements s'adressent également, à **M.LAKSSIR Brahim** chef de service Micro&Packaging. Sa gentillesse, sa modestie et l'accueil cordial qu'il m'a réservée, m'a inspirée une grande admiration à son égard.

Je remercie grandement, mon encadrant pédagogique, **SAFFAJ Taoufiq**, pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui ont contribué à alimenter ma réflexion.

Professeur **EL Mustapha EL HADRAMI,** que je tiens à remercier pour m'avoir accueilli dans son master, je suis très reconnaissant pour la confiance qu'il m'avait accordée, sa gentillesse, sa disponibilité, et ses encouragements m'ont été indispensables.

Mes remerciements également aux membres de jury d'avoir accepté de juger ce travail.

Il est évident que je ne peux oublier de remercier ma famille, notamment mes parents, mes frères, qui ont toujours répondu «présent» et ont été d'un grand secours moral en toutes situation.

Sommaire

Liste de figure

Liste des tableaux

Les abréviations

MAScIR : Moroccan Foundation for Advanced Science, Innovation and Research

FTIR : La spectroscopie infrarouge à transformée de Fourier

PIR : proche infrarouge

nm : nanomètre

OM : matière organique

Nt : Azote totale

RLM : régression linéaire multiple

ACP : analyse en composante principale

CP : composante principale

PLS : Partial Least Squares

PRESS : Predicted Error Sum of Squares

PCR : Régression en Composante Principale

SVM : support vectore machine

RMSEP : *root mean square error of prediction* – racine carrée de l'erreur quadratique moyenne de prédiction.

RMSECV : *root mean square error of cross-validation* – racine carrée de l'erreur quadratique moyenne de validation croisée correspondant à l'écart-type de prédiction.

PRESS :Prediction Re51dual Error Sum of Squares

R^2 : coefficient de détermination

RPD : ratio entre la déviation standard du vecteur y et le RMSEP ou le RMSECV

SG : savitzky golay derivative

MSC : multiple scatter correction

SNV : standard normal variate

Introduction

L'analyse quantitative des mélanges complexes est un défi de taille pour le chimiste. Généralement, pour être quantifiés, les constituants d'un mélange doivent d'abord être isolés et purifiés. Le sol est un exemple d'un mélange relativement complexe difficile à analyser quantitativement. Il est constitué par plusieurs éléments chimiques comme (la matière organique, l'azote total, le nitrate…). Les protocoles analytiques développés jusqu'à maintenant pour en effectuer l'analyse quantitative font appel à toute une série de techniques incluant la chromatographie gazeuse (GC), la spectrophotométrie ultraviolet (UV), la chromatographie liquide à haute performance (HPLC), la chromatographie liquide à haute performance en phase inverse (RPHPLC). Ces analyses sont faites subséquemment à une extraction solide/liquide et bien que certaines des étapes puissent être automatisées, ces méthodes prennent beaucoup de temps et sont fastidieuses et coûteuses. Il en est ainsi de la plupart des mélanges complexes d'où l'intérêt de parvenir à développer une méthode précise et rapide qui permettrait d'obtenir simultanément la composition de tous les constituants d'un mélange complexe.

Il est apparu que ce problème de complexité des mélanges pouvait être contourné par l'application de méthodes statistiques à une technique d'analyse spectroscopique. Ces méthodes statistiques de calibration et de prédictions multivariables sont appliquées ici à la technique de spectroscopie infrarouge à transformée de Fourier (FTIR). En effet, l'avènement des processeurs mathématiques et des techniques d'acquisition assistée par ordinateur met à la disposition du spectroscopiste une quantité phénoménale de données analytiques. La nature de ces informations se prête admirablement aux méthodes d'analyse statistique qu'utilise l'analyse multivariable quantitative. Couplée à la grande spécificité et à la rapidité d'acquisition de la spectroscopie FTIR, l'analyse multivariable quantitative s'avère être la technique de choix pour l'analyse quantitative de l'ensemble des constituants d'un mélange et ce, sans étape de séparation ou de purification. Elle a été appliquée avec succès à l'analyse FTIR de sol.

L'analyse multivariable quantitative comprend une phase de calibration lors de laquelle l'information contenue dans un ensemble de données expérimentales (spectres infrarouge) est mise en relation avec une propriété intrinsèque connue du système que l'on veut analyser (la concentration des constituants chimique de sol par exemple). Ces informations sont utilisées pour construire un modèle mathématique. En pratique, la phase de calibration requiert la préparation d'un ensemble d'échantillon de calibration dont les concentrations sont connues. L'analyse infrarouge de ces solutions fournit une série de spectres. Un algorithme mathématique permet alors d'établir la relation existant entre les données de concentrations et les données spectrales. Par exemple, selon la loi de Beer, l'absorbance d'un constituant d'un mélange est directement proportionnelle à sa concentration dans le mélange. Le modèle résultant, s'il est juste, c'est-à-dire si la relation énoncée par l'algorithme est le reflet de la réalité et qu'il n'y a pas d'écart à cette règle, devrait permettre d'estimer les concentrations des constituants d'un mélange inconnu talque le sol à partir de son spectre infrarouge.

Les analyses quantitatives effectuées sur le sol ont permis de démontrer l'efficacité de cette technique. De façon générale, l'analyse multivariable permet d'accroître considérablement le potentiel des méthodes FTIR quantitatives car, en plus de présenter un avantage indéniable pour l'analyse de routine de sol, elle peut être appliquée à tous les types de mélanges. Couplée à la spectroscopie FTIR, l'analyse multivariable est une technique précise et rapide qui permet d'obtenir simultanément les concentrations de plusieurs constituants de la plupart des mélanges.

Chapitre 1 :Présentation MAScIR : Moroccan Foundation for Advanced Science, Innovation & Reserarch

La Fondation MAScIR, *Moroccan Foundation for Advanced Science, Innovation and Research*, est une institution publique à but non lucratif qui a pour objectif la promotion de la recherche scientifique et le développement technologique en vue d'accompagner le développement du Maroc et participer au développement d'une nouvelle économie de savoir.

Outre les moyens scientifiques de pointe dont elle dispose, la fondation MAScIR rassemble d'éminents chercheurs et ingénieurs, marocains et étrangers, œuvrant dans des domaines aussi innovants que complémentaires.

De l'environnement, aux énergies renouvelables, en passant par la santé, la recherche à MAScIR s'adapte aux besoins de la société et de l'industrie.

MAScIR rassemble près de 100 chercheurs et ingénieurs, son chiffre d'affaire a été de 93.6 millions de Dirham au cours de l'année 2014. La figure suivante présente l'évolution des chiffres clés de la fondation au cours des 5 dernières années.

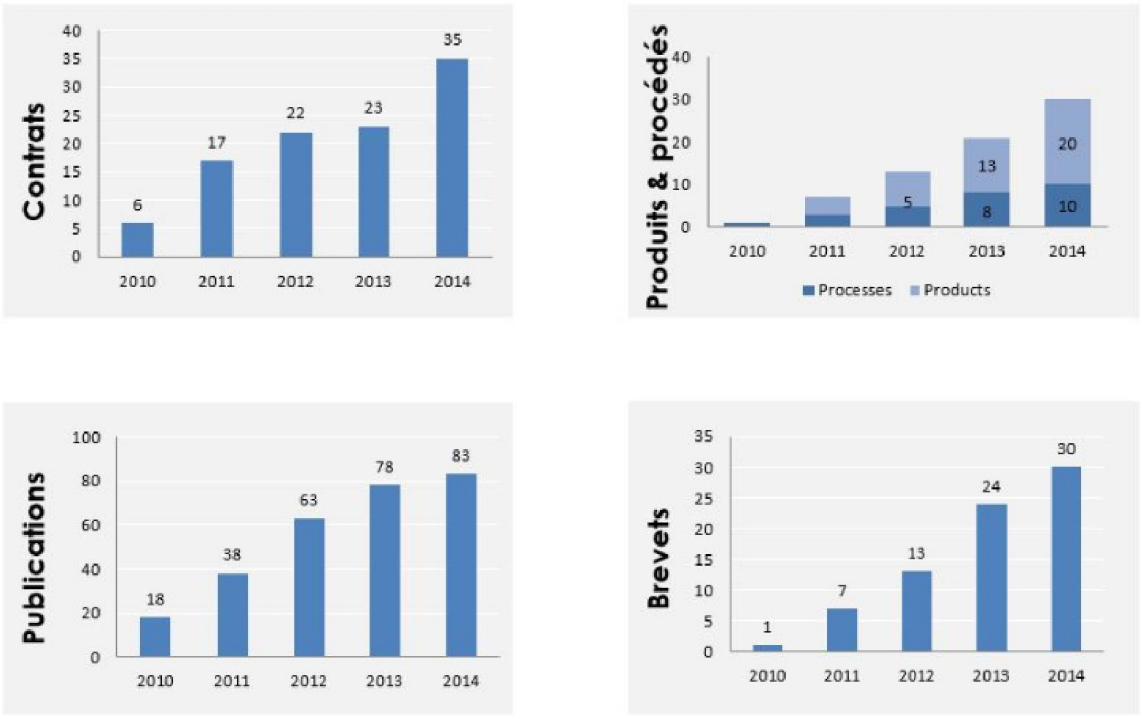


Figure 1 : Evolution des indicateurs clés de MAScIR pendant les 5 dernières année

1. Domaines d'activité de l'entreprise

Initialement fondée en 2007 par le Gouvernement Marocain en tant que fondation à but non lucratif, MAScIR compte aujourd'hui 3 pôles:

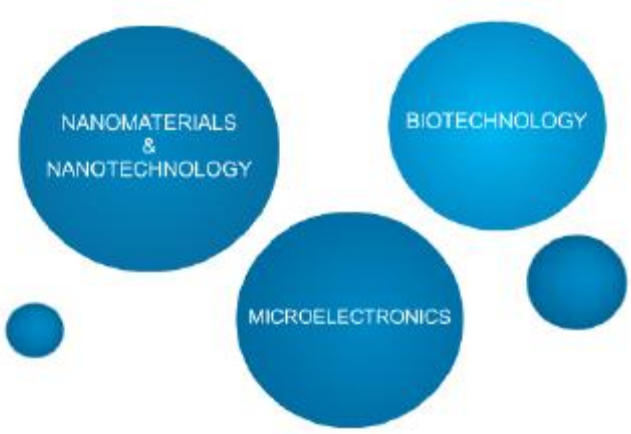


Figure 2 : Secteur d'activité de la fondation MAScIR

MAScIR MicroElectronics: ce pôle a pour objectif le développement de capteurs, la conception et l'assemblage de cartes électroniques ainsi que la définition d'algorithme et le traitement d'images et vidéos.
Le pôle Microélectronique est composé de deux centres :

- Packaging

Le centre packaging dispose de plateformes techniques chargées de produire des solutions complètes dans ses propres laboratoires à savoir la simulation et le packaging, ainsi que l'assemblage et les tests de fiabilité.

- Systèmes embarqués

Le centre systèmes embarqués répond aux besoins des industriels en matière de *reverse engineering*, conception et développement des solutions à savoir: les contrôleurs industriels, le développement de systèmes médicaux low cost, l'inspection visuelle...

MAScIR BioTechnology: Les médicaments et la santé, les agros industries ou encore l'agriculture, tels sont les domaines dans lesquels les équipes de MAScIR ont entrepris des recherches de haut niveaux.
Le pôle biotechnologie est composé de deux centres :

- Bio végétale

Ce centre est spécialisé dans la recherche biologique au niveau des végétaux. Parmi les projets en cours de valorisation les bioénergies à partir de micro-algues, les bio-fertilisants, les bio-stimulants…

- Bio médicale

Les chercheurs de ce centre sont entièrement dédiés à développer des kits de diagnostic innovant ciblant les maladies spécifiques au Maroc telles que les maladies infectieuses et les cancers.

MAScIR NanoTechnology : qui a pour mission de mener des recherches appliquées, innovantes et à la fine pointe de la technologie dans le domaine des nanomatériaux et des nanotechnologies.

2. Structure et organisation générale

Le Conseil d'Administration est investi des pouvoirs les plus étendus, pour gérer, diriger et administrer les activités scientifiques et administratives de MAScIR. Il est présidé par le Ministre de l'Industrie, du Commerce, de l'Investissement et de l'Economie Numérique, et est assisté par un Conseil Scientifique, un Comité Directeur et un Comité de Rémunération.

Le Conseil Scientifique a pour rôle d'évaluer les projets proposés par les chercheurs et de formuler des recommandations quant à la réalisation et la budgétisation de ses derniers.

Une équipe de business développement vient renforcer les équipes scientifiques en mettant à leur disposition une expertise et des outils leur permettant de répondre aux besoins des clients et des partenaires en valorisant les produits de leur recherche.

Le Comité Directeur a pour mission de coordonner les projets des centres de recherche et d'optimiser les ressources en assurant le lien entre les centres de recherche et les autres comités.

Le Comité de Rémunération dont la principale mission est d'élaborer des procédures d'évaluation du personnel et d'en assurer le suivi afin d'attirer et de retenir les chercheurs de haut calibre.

Chapitre 2 : Partie théorique

1. Le sol

Le sol est le support de la vie terrestre. Il résulte de la transformation de la couche superficielle de la roche-mère, la croûte terrestre, dégradée et enrichie en apports organiques par les processus vivants. Hors des milieux marins et aquatiques d'eau douce, il est ainsi à la fois le support et le produit du vivant, il est La partie arable homogénéisée par le labour et explorée par les racines des plantes cultivées.

Généralement il y a 4 types de sol (Sol argileux ; Sol humifère ; Sol sableux ; Sol calcaire)

a) Constituants du sol

Le sol comporte trois fractions :

- Une fraction solide (insoluble dans l'eau) qui, du fait de la double origine du sol, contient à la fois des éléments minéraux (sables, limons, argiles) et des éléments organiques (matière organique dont l'humus)
- Une fraction liquide = la solution du sol, contenant des éléments solubles dissous dans l'eau
- Une fraction gazeuse = l'atmosphère du sol, composée d'air et de gaz provenant de la vie microbienne et de la décomposition de la matière organique.

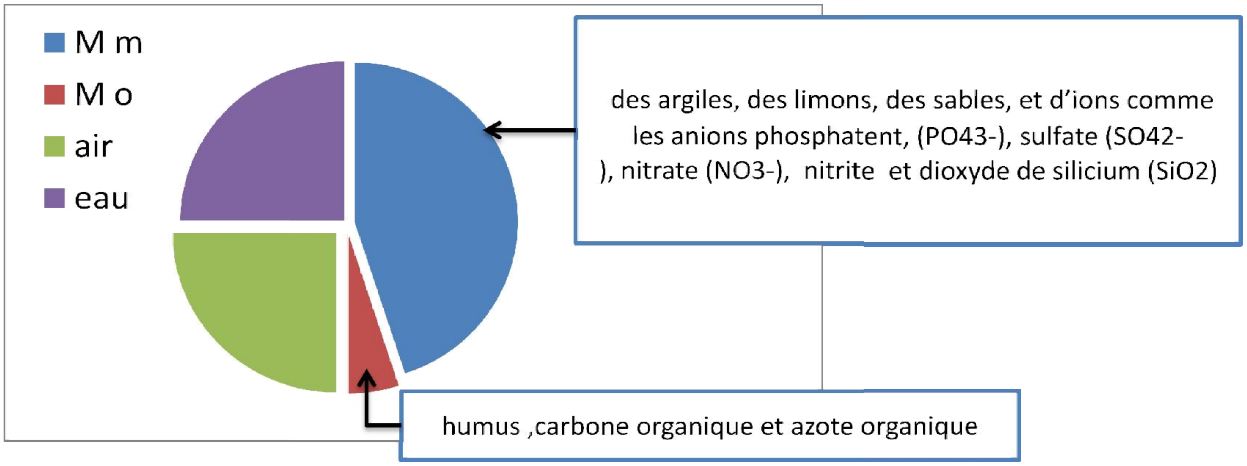


Figure 3 : le pourcentage des éléments de sol

b) Éléments fertilisants

i. La matière organique

Le terme «matières organiques du sol» regroupe l'ensemble des constituants organiques morts ou vivants, d'origine végétale, animale ou microbienne, transformés ou non, présents dans le sol. Elles représentent en général 1 à 10 % de la masse des sols.

Elles se répartissent en trois groupes

(1) les Matières Organiques Vivantes (MOV), animale, végétale, fongique et microbienne, englobent la totalité de la biomasse en activité .

(2) les débris d'origine végétale (résidus végétaux, exsudats), animale (déjections, cadavres), fongique et microbienne (cadavres, exsudats) appelés «Matières Organiques fraîches». Associés aux composés organiques intermédiaires issus de l'activité de la biomasse microbienne, appelés produits transitoires (évolution de la matière organique fraîche), elles composent les MO facilement décomposables.

(3) des composés organiques stabilisés (MO stable), les matières humiques ou humus, provenant de l'évolution des matières précédentes. La partie humus représente 70 à 90 % du total.

ii. Azote

L'azote entre, avec d'autres éléments (carbone, oxygène, hydrogène…), dans la composition des acides aminés formant les protéines. L'azote est un élément essentiel pour la con stitution des cellules et la photosynthèse (chlorophylle). C'e st le principal facteur de croissance des plantes et un facteur de qua lité qui influe sur le taux de protéines des végétaux.

2) La spectroscopie et le sol

a) Le principe de la spectroscopie

La spectroscopie infrarouge repose sur le principe que chaque groupement chimique absorbe la lumière différemment en fonction de sa longueur d'onde. Les bandes d'absorption (zone où la lumière est absorbée) permettent donc d'identifier les groupements atomiques. Beer (1729) et Lambert (1760) ont ainsi proposé d'observer l'atténuation d'un faisceau de la lumière afin de prédire la concentration d'un composé selon l'expression suivante :

$$A_\lambda = \varepsilon_\lambda \times \ell \times c \qquad (1)$$

où : A_λ :est l'Absorbance à une longueur d'onde λ donnée,

ε_λ :est le coefficient d'extinction molaire (L.mol−1·cm−1),

ℓ :la longueur du trajet optique dans l'échantillon (cm),

c :la concentration de la solution (mol.L-1).

Selon les énergies des radiations mises en jeu, la spectroscopie fournie des informations concernant

- les électrons formant les liaisons chimiques (spectroscopie U.V.- visible)
- les atomes impliqués dans les liaisons (spectroscopie infrarouge)
- les noyaux atomiques (spectroscopie de résonance magnétique nucléaire)

b) Spectroscopie infrarouge (IR)

i. Principe

La spectroscopie IR repose sur l'absorption d'énergie électromagnétique pour des radiations dont la longueur d'onde appartient au domaine de l'infrarouge (de 750 nm à 1 mm). Pour ces domaines de longueur d'onde, ce sont les molécules qui absorbent les rayonnements afin d'atteindre des états de vibration.

Les molécules subissent des mouvements de vibrations internes de deux types :

- ✓ vibrations d'élongation (lorsque la longueur d'une liaison covalentes se met à osciller autour de sa valeur moyenne)
- ✓ vibrations de déformation (lorsque l'angle défini entre deux liaisons covalentes se met à osciller autour de sa valeur moyenne)

ii. La zone de l'infrarouge

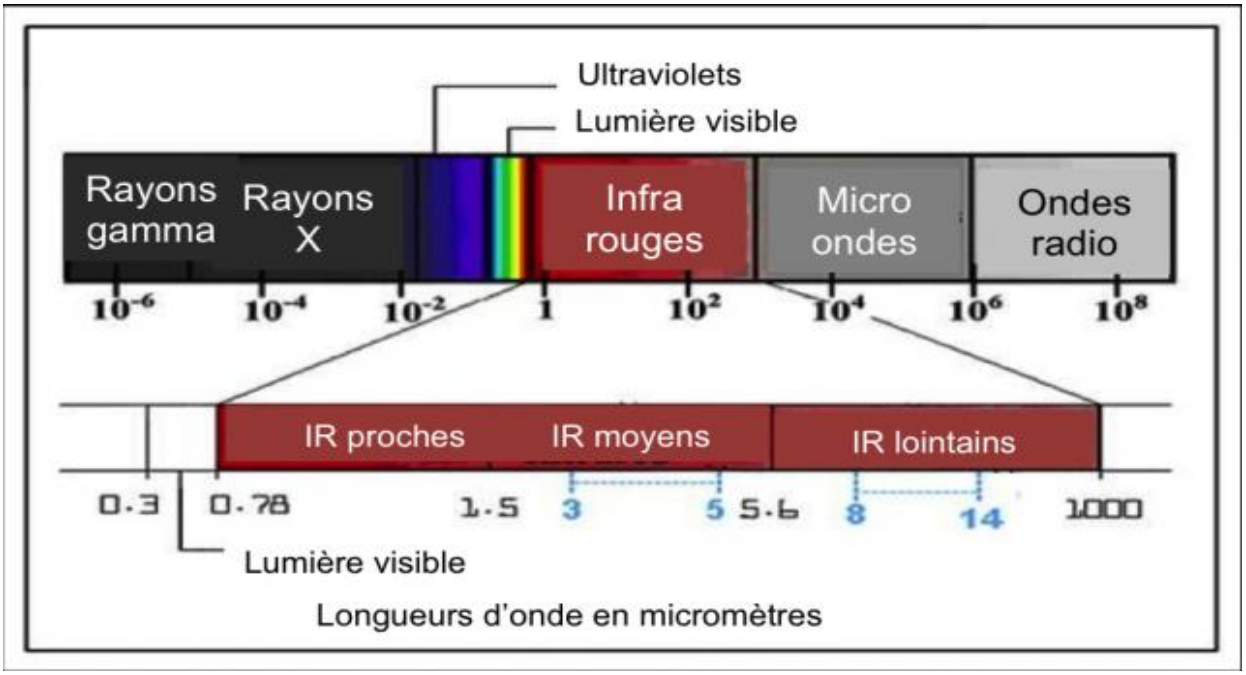


Figure 4 : Les zones des longueurs d'onde électromagnétique et la zone de proche infrarouge

Pour cet étude on se focalisé sur la zone de l'infrarouge précisément Proche infrarouge

iii. Origine des bandes d'absorption

Dans la zone du proche infrarouge, les produits absorbent beaucoup moins que dans l'infrarouge moyen, ce qui peut être dans certains cas un avantage : il sera

possible d'étudier directement en transmission des échantillons de plusieurs millimètres d'épaisseur. Ces absorptions ne sont pas dues aux vibrations fondamentales des molécules, mais à des phénomènes plus complexes : les vibrations harmoniques et les vibrations de combinaisons.

iv. L'évolution d'utilisation de PIR

La première application analytique de la SPIR en 1962 par Hart et Norris (2) utilisait un modèle linéaire basé sur cette loi avec une seule longueur d'onde.

Néanmoins, comme les pics d'absorption dans le PIR sont larges (l'absorption de la lumière se fait sur les harmoniques des fréquences de vibrations des molécules), se superposent et que de nombreuses interactions entre molécules (changement de la fréquence de vibration) modifient le signal, il s'avère difficile d'appliquer la loi de Beer-Lambert. Certains composés

peuvent présenter des bandes communes et l'identification d'une molécule nécessite donc l'analyse de la signature spectrale sur de multiples longueurs d'onde (3). Les bandes d'absorption des composés peuvent également changer en fonction de l'environnement dans lequel elles sont placées (température, charges ioniques des molécules voisines).

Par ailleurs, un autre phénomène vient très souvent s'ajouter à l'absorption : la diffusion. La structure physique de l'échantillon influence de manière importante le trajet des photons en son sein selon qu'il s'agit de fi bres, de cellules, d'une poudre, d'une suspension... Cette modification du trajet optique se caractérise par une succession de changement de direction des photons par des phénomènes de réflexion/réfraction. Ce phénomène appelé « diffusion multiple » est même, bien souvent, plus fréquent que les phénomènes d'absorption dans le milieu. Ainsi, pour des produits comme les poudres, on estime qu'il y a un phénomène d'absorption pour 100 phénomènes de diffusion. Ces phénomènes ont deux conséquences principales : premièrement, la loi de Beer-Lambert n'apparaît plus valable, deuxièmement, le spectre obtenu contient à la fois des informations de nature chimique et de nature physique car l'atténuation du faisceau n'est plus uniquement due à l'absorption.

Quoi qu'il en soit cette mixité de l'information dans un seul spectre peut présenter un intérêt dès lors que les propriétés physiques du produit doivent être intégrées dans la prise de décision finale

v. Réflexion, Transmission et Absorption

La réflexion est le processus par lequel un rayonnement électromagnétique est renvoyée soit à la limite entre les deux milieux (surface de réflexion) ou à l'intérieur d'un milieu (volume réflexion), tandis que la transmission est le passage d' un rayonnement électromagnétique à travers un milieu. Les deux processus peuvent être accompagnés par diffusion (également appelé diffusion), qui est le processus de dévier un faisceau unidirectionnel dans de nombreuses directions. Dans ce cas, nous parlons de réflexion diffuse et transmission diffuse.

La Transmission T : est définie comme la fraction d'énergie lumineuse traversant l'échantillon de part en part . Elle est égale au rapport d'intensités

$$T = I_T/I_0 \times 100 \qquad (2)$$

I_T : L'intensité transmise

I_0 : L'intensité initial

La Réflectance R : est définie comme la fraction d'énergie lumineuse réfléchie

$$T = I_R/I_0 \times 100 \qquad (3)$$

L'absorbance A : est définie comme étant le logarithme décimal de l'inverse de la transmittance ou de réflectance

$$A = \log(1/T) = \log(I_T/I_0) \qquad (4)$$

$$A = \log(1/R) = \log(I_R/I_0) \qquad (5)$$

c) Avantages et inconvénients de l'analyse dans le proche infra-rouge

Avantages

- Peu ou pas de préparation de l'échantillon (qui peut être récupéré aprés l'analyse)
- analyse en transmission sur des échantillons relativement épais
- analyse en réflexion sans contact (peu d'influence CO2 ou H2O atmosphérique)
- Analyse rapide multicomposant, en temps réel
- Coût de l'analyse peu élevé
- Cellule de mesure résistante et assez bon marché (verre ou quartz)
- Gamme importante d'appareils robustes pour l'analyse en ligne
- Possibilité d'analyse de produits toxiques ou dangereux à distance (+ de 500 m en utilisant des fibres optiques)
- Amortissement de l'investissement généralement rapide
- Méthode puissante pour l'étude de la liaison hydrogéne
- Méthode de choix pour le contrôle industriel, analyse et monitoring in situ en temps réel

Inconvénients :

- Manque de corrélation structurale (difficultés pour l' interprétation des spectres)
- Besoin de calibration pour les mélanges (analyse directe très difficile en général) utilisation de méthode de chimiométrie Phase d'étalonnage longue et delicate
- En réflexion la surface de l'échantillon doit être identique au cœur (faible pénétration du faisceau dans l'échantillon)

- Taille des particules ainsi que l'orientation modifient les spectres (même probleme qu'en IR moyen)
- Problème de transfert de calibration d'une méthode d'un appareil à l'autre
- Manque de bibliothèques de spectres diversifiées

d) Spectroscopie : Bandes Caractéristiques identifiée

Tableau 1 : les bandes caractéristique des fertilisants de sol

Bande	VIS 400-700nm	NIR 700-2500nm	MIR 2500-25000nm	LIR 25-2500µm
Eau	P401 P449 P514 P606 P660 P739	P836 P970 P1200 P1470 P1900	P2870 P3050 P4650 P6080 P15000 P25000	P55 P200
OM	500 - 590 610 - 682 550-700	780 - 890 **1550 - 1890** 1400 - 1900 P960, P1100 P1400, P1900 1200-1717	2500 - 25000 33336-3480 7231-6892 P7230 P6887	---
OC	P410, P570 P660	P1700, P2150 P2310, P2350	2500 - 25000	---
N total	P550	P940, P1150 P1200, P1300	2500 - 25000	---
Nitrates	P240	1200 - 1400	P7194	---
P	200-400	1950-2000 1500-2500	---	---
K	P220	---	---	---

Le nombre des données fournies par les appareils sont très variables : un pH mètre ou un photomètre avec un filtre donnera une valeur pour une solution donnée, tandis qu'un spectrophomètre donnera en un court instant de nombreuses données pour l'analyse d'un seul échantillon. Donc il faut appeler les techniques multivariables (Chimiométrie) qui doivent être gérées au mieux ces données afin d'utiliser toutes les données disponibles avec un temps de calcul rapide

L'efficacité de la spectroscopie proche infra-rouge dépend en partie des méthodes statistiques utilisées pour les analyses quantitatives de produit complexes comme le sol.

2. La Chimiométrie et la spectroscopie PIR

La Chimiométrie peut être définie comme l'ensemble des méthodes statistiques, graphiques ou symboliques qui permettent d'améliorer la compréhension d'information obtenues dans le domaine de la chimie. La Chimiométrie s'applique a tout les étapes de l'analyse depuis la 1er conception de l'expérience jusqu'à ce que les données soient complètement exploitées.

a) L'objectif des méthodes chimiométriques

Les méthodes chimiométriques peuvent avoir trois objectifs principaux : la description des données sous une forme synthétique. La prédiction ou la planification expérimentale

Le spectre PIR de sol contient des informations qui présentent un intérêt analytique. Cependant, l'extraction de ces informations ne s'effectue pas immédiatement et demande toujours un prétraitement mathématique, le spectre PIR est le résultat de l'interaction de la lumière avec les éléments chimiques du sol, il est constitué de centaines de variables (564 variable pour ce projet) et chaque variable représente une longueur d'onde dans la gamme (1200 nm -2100 nm).

Au coure de scannage des échantillons on tombe dans le problème de la variation incontrôlable enregistre dans le spectre, cette variation lié avec la variation de la température, la variation de la granulométrie du sol ou le problème de diffusion de la lumière. Pour tenir compte de ces variations, il est nécessaire d'accumuler des spectres de nombreux échantillons, et de traiter la collection spectrale obtenue par des méthodes chimiométriques

b) Méthodes chimiométriques

1) Les méthodes exploratoires

Correspondent à l'objectif de description des données et comprendre a la fois des méthodes élémentaires (la moyenne écart-type , variance des spectre s…) , ainsi que des méthodes sophistique (des méthodes factorielle) comme l'analyse en composante principale ou la classification

2) Les méthodes prédictives

Sont objectif set de prédire ou estimer les repenses inconnues d'un échantillons (la régression en générale) qui permet de trouver des relations mathématiques entre la transmittance correspond a chaque longueur d'onde et la concentration inconnue de l'éléments chimique dans le sol

A) Analyse en composante principale ACP

A partir de la matrice X des données de départ on construit la matrice M de corrélation

$$\mathbf{M = 1/n\ (X'X)} \qquad (6)$$

Avec X' la matrice transposé de X et X peut être préalablement transformée matrice centrées ou centrées réduites

La matrice M de corrélation est diagonalisée pour fournir :

Une matric dia gonale D des valeurs propres λ1, λ2 ,…. λ n et une matrice P des vecteurs propres

Les vecteurs colonnes de la matrice P représentent les combinaisons linéaires des variable de départ conduisant aux nouveaux axes factoriels = les Composantes Principales

Les valeurs propres représentent les variances des individus sur les composantes principales

Donc l ACP nous permette de réduire la dimension d'une matrice ,donc il permette de surmonter le problème de la forte colinéarité (la concordance) entre les variable (des problème dans l'estimation des coefficient de régression dans le cas de régression multiple), et aussi ACP permette de faire une représentation (projection) graphique des donnes sur des axes factorielle ce qui facilite l'analyse des donnes

B) Modélisation

1) Méthodes de régression linéaire

L'analyse spectroscopique quantitative est régie par la loi de Beer-Lambert. Celle-ci stipule que l'absorbance, A d'un composé dépend de son absorptivité a du parcours optique du rayonnement b et de la concentration c du composé.

$$\mathbf{A = abc} \qquad \mathbf{(7)}$$

Pour simplifier la quantification du système, les paramètres a et b sont fixés. Ceci se traduit expérimentalement par l'utilisation du même montage optique pour tous les échantillons. L'équation (7) s'exprime alors de la façon suivante

$$\mathbf{A = kc} \qquad \mathbf{(8)}$$

où k est la constante de proportionnalité.

Pour un système comportant plusieurs constituants, l'expression de la loi de Beer-Lambert doit inclure la contribution de chacun des constituants à la valeur d'absorbance et ce, à chaque longueur d'onde.

$$\begin{aligned} A_1 &= k_{11}c_1 + k_{12}c_2 + \ldots + k_{1j}c_i \\ A_2 &= k_{21}c_1 + k_{22}c_2 + \ldots + k_{2j}c_i \\ &\vdots \quad \vdots \qquad \vdots \qquad \vdots \\ A_\lambda &= k_{\lambda 1}c_1 + k_{\lambda 2}c_2 + \ldots + k_{\lambda j}c_i \end{aligned} \qquad (9)$$

A_λ est l'absorbance totale à la longueur d'onde λ , ci est la concentration du constituant j et $k_{\lambda j}$ est la constante de proportionnalité du composé j à la longueur d'onde λ.

Lorsque le nombre de longueurs d'onde à laquelle est effectuée l'analyse augmente et que le système comprend plusieurs constituants, il devient utile d'exprimer l'équation

(9) sous forme matricielle. Ainsi, pour un spectre individuel.

$$\begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_\lambda \end{bmatrix} = \begin{bmatrix} k_{11} & k_{12} & & k_{1j} \\ k_{21} & k_{22} & \ldots & k_{2j} \\ \vdots & \vdots & & \vdots \\ k_{\lambda 1} & k_{\lambda 2} & & k_{\lambda j} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_{\lambda j} \end{bmatrix} \qquad (10)$$

Où

$$A = KC \qquad (11)$$

Où chaque vecteur colonne de la matrice d'absorbance, A, correspond au spectre d'un échantillon et où chaque vecteur colonne de la matrice des concentrations, C correspond aux concentrations de chacun des constituants de ce même échantillon.

Les éléments de matrices A_λ et c_j correspondent respectivement à l'absorbance à la longueur d'onde λ et à la concentration du constituant j de l'échantillon n.

Pour opérer l'analyse univariable quantitative d'un composé, une bande spectrale appartenant à ce composé doit être isolée dans le spectre FTIR. Cette condition essentielle implique que, pour une valeur A_λ donnée, un seul terme à droite de l'équation (9) do1t être différent de zéro, i-e qu'à une longueur d'onde précise, un seul constituant doit contribuer à la valeur d'absorbance. Généralement, les spectres FTIR de mélanges présentent de larges zones de superposition à l'intérieur desquelles aucune bande typique à l'un des constituants d'intérêt ne peut être isolée. D'autres méthodes, telles la chromatographie et la spectroscopie de masse, opérant une séparation des différents constituants sont alors privilégiées. Avec l'analyse multivariable, il n'est pas nécessaire d'isoler une bande spectrale pour chacun des constituants. Cette technique procède plutôt à l'analyse matricielle des données analytiques brutes.

2) Méthode classique des moindres carrés

La méthode CLS (Classical Least Squares), aussi appelée méthode de la matrice K , consiste à résoudre l'équation (11) afin d'obtenir la matrice des constantes de proportionnalité, K

Pour résoudre K , il faut d'abord faire en sorte que les matrices d'absorbance et de concentrations soient carrées, ce qui n'est pas le cas quand le système est sur-déterminé , c'est-à-dire, quand il compte plus d'échantillons que de constituants. Pour ce faire, il suffit de multiplier A et C par la transposée de C, C'. La matrice K peut ensuite être isolée. Ainsi, l'équation (11) conduit à

$$\mathbf{AC' = KCC'}$$

$$\mathbf{AC'(CC')^{-1} = KCC'(CC')^{-1}}$$

$$\mathbf{AC'(CC')^{-1} = k} \qquad \mathbf{(12)}$$

En pratique, il n'existe pas de solution exacte pour K car chaque élément de la matrice A est entaché d'une erreur. Il faut donc en déterminer la meilleure approximation. CLS utilise la méthode des moindres carrés et minimise la somme des carrés des erreurs associées aux valeurs d'absorbance, E_A. Ces erreurs sont définies comme la différence entre les valeurs d'absorbance mesurées et celles calculées en multipliant K et C

$$\boldsymbol{E_A = \hat{K}C - A} \qquad \mathbf{(13)}$$

$\hat{K}$ étant une approximation de K

L'inverse de la matrice $\hat{K}$ permet d'estimer les concentrations ĉ de chacun des constituants d'un échantillon inconnu à partir de son spectre d'absorbance a.

$$\mathbf{\hat{c} = \hat{K}^{-1}a} \qquad \mathbf{(14)}$$

La méthode CLS requiert deux inversions matricielles. L'inverse de CC' doit être calculé pour obtenir $\hat{K}$ et l'inverse de $\hat{K}$ doit être déterminé de façon à pouvoir estimer les concentrations d'un inconnu à partir de son spectre mesuré. La dimension de $\hat{K}$ est λ x j et à moins de n'utiliser qu'une région spectrale très restreinte, c'est-à-dire, autant de longueurs d'ondes spectrales que de constituants, la matrice $\hat{K}$ ne peut être inversée directement. Il est toutefois possible de procéder à l'inversion de K en multipliant a et $\hat{K}$ par la transposée de $\hat{K}$, $\hat{K}$ ' . La matrice K peut ensuite être isolée, Ainsi, l'équation (11) conduit à

$$\mathbf{\hat{K}'a = \hat{K}'\hat{K}\hat{c}}$$

$$\mathbf{(\hat{K}'\hat{K})^{-1}\hat{K}'a = (\hat{K}'\hat{K})^{-1}\hat{K}'\hat{K}\hat{c}}$$

$$\mathbf{(\hat{K}'\hat{K})^{-1}\hat{K}'a = \hat{c}} \qquad \mathbf{(15)}$$

Cette opération permet l'utilisation du spectre entier, c'est là un des avantages de la méthode CLS. Toutefois, elle n'est applicable que dans les cas où tous les constituants du système sont connus. Ainsi, chaque impureté susceptible d'introduire des interférences doit être incluse dans l'analyse, ce qui est difficilement réalisable dans les cas de systèmes complexes talque le sol

Des difficultés supplémentaires surviennent s'il existe une déviation à la loi de Beer-Lambert, si la relation entre l'absorbance et la concentration n'est pas linéaire. Dans ce dernier cas, la courbe résultante peut, dans la région d'intérêt, être assimilée à une droite si une valeur d'ordonnée à l'origine est ajoutée à l'équation (8). Celle-ci devient alors,

$$\mathbf{A : kc + k_0} \qquad \mathbf{(16)}$$

où k est la pente de la section de courbe assimilée à une droite et k_0, son ordonnée à l'origine. L'ordonnée à l'origine peut être incluse dans le système à plusieurs constituants en ajoutant à la matrice K de l'équation (10) un vecteur colonne constitué d'éléments k_λ et la matrice C un vecteur ligne constité d'éléments unitaires

$$\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & & A_{2n} \\ \vdots & & \ddots & \vdots \\ A_{\lambda 1} & A_{\lambda 2} & \cdots & A_{\lambda n} \end{bmatrix} = \begin{bmatrix} k_{11} & k_{12} & \cdots & k_{10} \\ k_{21} & k_{22} & & k_{20} \\ \vdots & & \ddots & \vdots \\ k_{\lambda 1} & k_{\lambda 2} & \cdots & k_{\lambda 0} \end{bmatrix} \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & & c_{2n} \\ \vdots & & \ddots & \vdots \\ c_{j1} & c_{j2} & \cdots & c_{jn} \\ 1 & 1 & \cdots & 1 \end{bmatrix} \qquad \mathbf{(17)}$$

La matrice K résultante n'est pas dans ce cas, une matrice carrée et ne peut étre inversée directement. L'équation (12) devient

$$\mathbf{C = (K'K)^{-1}k'A} \qquad \mathbf{(18)}$$

Cette équation est semblable à une régression des moindres carrés traitant les éléments de la matrice K comme des variables indépendantes. La difficulté réside ici dans le fait qu'il y a plus d'inconnus que d'équations, qu'il y a plus de colonnes que de lignes. Il n'est pas possible alors de calculer l'inverse de (K'K) . La méthode classique des moindres carrés ne s'applique pas à la résolution de ce genre de système.

3) Méthode inverse des moindres carrés (ILS)

Dans les cas de déviations à la loi de Beer-Lambert, le modèle causal qu'emploie la méthode CLS n'est pas approprié et ne permet pas de résoudre l'équation (11). Le problème peut toutefois être contourné en utilisant un modèle empirique basé sur la corrélation plutôt que sur la causalité, en inversant la loi de Beer-Lambert et en exprimant la concentration comme une fonction de l'absorbance.

$$\mathbf{C = PA'} \qquad \mathbf{(20)}$$

ou

$$\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & & c_{2n} \\ \vdots & & \ddots & \vdots \\ c_{\lambda 1} & c_{\lambda 2} & \cdots & c_{\lambda n} \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{10} \\ p_{21} & p_{22} & & p_{20} \\ \vdots & & \ddots & \vdots \\ p_{\lambda 1} & p_{\lambda 2} & \cdots & p_{\lambda 0} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & & A_{2n} \\ \vdots & & \ddots & \vdots \\ A_{j1} & A_{j2} & \cdots & A_{jn} \\ 1 & 1 & & 1 \end{bmatrix} \quad (21)$$

Où la matrice C est telle que définie précédemment, où P est la matrice représentant la constante de proportionnalité entre C et A' et où A' est la matrice d'absorbance A décrite précédemment mais comportant en plus un vecteur ligne d'éléments unitaires. Chaque vecteur ligne de la matrice P est associée à un constituant du mélange et chaque élément d'un vecteur ligne, p_{jn} correspond au coefficient de régression associé au constituant j à une longueur d'onde λ. Ainsi, P comprend autant de lignes qu'il y a des constituants et autant de colonnes qu'il y a de longueurs d'ondes. La dernière colonne de P permet d'inclure les valeurs d'ordonnées à l'origine.

La méthode inverse (Inverse Least Squares), également appelée la méthode de la matrice P , consiste à résoudre l'équation (19) afin d'obtenir la matrice des constantes de proportionnalité. Pour ce faire, il faut d'abord multiplier chaque côté de l'équation par la transposée de la matrice d'absorbance avant d'isoler P.

$$\mathbf{CA'^t = PA'A'^t}$$

$$\mathbf{CA'^t (A'A'^t)^{-1} = P} \quad (22)$$

Pour que l'inverse de $(A'A'^t)$ existe, la matrice des absorbances doit avoir au moins autant de colonnes que de lignes. Puisque A' possède une ligne pour chaque longueur d'onde, une ligne d'éléments unitaires et une colonne pour chaque échantillon ceci signifie que le système doit comporter un échantillon de plus que le nombre de longueurs d'ondes analysées. Auss1, pour que l'équation (18) soit valide, il faut compter au moins autant de longueurs d'ondes que de constituants.

Comme dans le cas de la méthode CLS, la méthode ILS tient compte de l'incertitude sur les éléments de matrices. Toutefois, contrairement à la méthode classique, la méthode inverse assume que l'erreur origine exclusivement des valeurs de concentrations. Pour déterminer la meilleure approximation pour la matrice P, ILS utilise la méthode des moindres carrés et minimise la somme des carrés des erreurs associées aux valeurs de concentrations, $E_{C.}$ Ces erreurs sont définies comme la différence entre les valeurs de concentrations réelles et celles calculées en multipliant P et A'

$$\mathbf{EC = C - \hat{P}A'} \quad (25)$$

$\hat{P}$ étant une approximation de P

La matrice $\hat{P}$ peut ensuite être utilisé directement afin d'estimer les concentrations d'échantillons inconnus à partir des valeurs d'absorbance en utilisant l'équation (18)

$$\hat{c} = \hat{P}a' \qquad (26)$$

Il est facile de concevoir que les prédictions obtenues par les méthodes classique et inverse soient différentes et c'est, en général, le cas. Toutefois, ces différences sont négligeables. L'important ici n'est pas de déterminer quelle méthode est la meilleure mais de constater que pour un même système, des approches différentes peuvent être utilisées.

Contrairement à la méthode CLS, le nombre de longueurs d'onde qui peuvent être incluses dans l'analyse ILS est restreint. En conséquence, la matrice d'absorbance utilisée avec la méthode ILS est généralement plus petite que celle utilisée avec la méthode CLS. On ne peut utiliser plus de longueurs d'onde qu'il y a d'échantillons de calibration. Cette restriction est imposée par le fait que chaque nouvelle fréquence ajoute un inconnu à l'équation. En conséquence, une équation supplémentaire doit être ajoutée afin d'éviter que l'équation ne soit indéterminée.

4) La Régression linéaire multiple pas à pas

La question qui suit généralement l'approche par la régression multiple est de choisir parmi les variables X le plus petit nombre d'entre elles qui explique au mieux la variabilité de Y.

Une méthode courante est une régression itérative qui inclut d'abord dans le modèle la variable qui propose le meilleur coefficient de détermination. Ensuite, celle qui améliore le plus le coefficient de détermination et ainsi de suite.

Alternativement, toutes les variables sont entrées dans le modèle et les variables sont progressivement exclues, en fonction de celles qui contribuent le moins au modèle.

Il faut noter que la seconde variable qui entre dans le modèle n'est pas forcément celle qui présente, à elle seule, le second meilleur coefficient de détermination avec Y. Sinon, la solution serait triviale. En effet, X1 et X2 peuvent être très corrélées, voire quasi redondantes. Dans ce cas la qualité du modèle ne sera pas améliorée. C'est donc la variable qui contribue le plus à réduire la variabilité résiduelle, du modèle en voie d'élaboration qui sera sélectionnée à chaque étape

Donc on va appliquer un teste de signification des coefficients fondé sur un test de F partiel, on utilise SCR_k et SCR_{k+1} les sommes des carrées des résidus obtenus respectivement aux étapes k et k+1 , L'apport d'une nouvelle variable à l'étape k+1 se traduit par une diminution de la somme des carrées des résidus

$$\text{Fpartiel} = \frac{\text{SCR}(k) \text{ et } \text{SCR}(k+1)}{\text{SCR}(k+1)} \qquad (27)$$

5) Méthodes d'analyse factorielle

Les méthodes d'analyse factorielle permettent de pousser un peu plus les limites des méthodes classique et inverse en réduisant la dimension du système. Ces techniques de compression de données permettent de transposer les données spectrales dans un nouveau système de coordonnées. Les données spectrales sont ainsi décomposées en vecteurs de base auxquelles sont associés des coefficients représentant les intensités dans ce nouveau référentiel. Chaque spectre d'origine peut être reconstitué par une combinaison linéaire des vecteurs de base. Les méthodes d'analyse factorielle peuvent être appliquées au spectre entier ou à une sélection de régions spectrales.

a) Méthode de régression de la composante principale (PCR)

L'une des méthodes les plus utilisée pour déterminer les vecteurs de base du nouveau système de coordonnées est la méthode de régression de la composante principale, PCR (Principal Component Régression). Cette méthode procède en deux étapes. La première étape consiste à déterminer les vecteurs de base définissant le nouveau référentiel des spectres d'absorbances. Les intensités spectrales de ceux-ci sont ensuite projetées sur le nouvel espace vectoriel. Une matrice de coefficients représentant les nouvelles intensités spectrales est ainsi formée. Dans un second temps, une régression impliquant cette matrice d'intensités spectrales et les valeurs de concentration est opérée. Les coefficients de régression servent ensuite é faire l'estimation des concentrations sur des échantillons inconnus.

Mathématiquement, 1a matrice absorbance, A, est décomposée en deux matrices de plus petite dimension.

$$\mathbf{A = TB} \qquad \mathbf{(28)}$$

$$\mathbf{E_A = \check{T}\check{B} - A} \qquad \mathbf{(29)}$$

où T est une matrice n x r représentant les intensités dans le nouveau système de coordonnées et B, une matrice r x λ constituant les r vecteurs de base de ce nouveau référentiel pour lesquels les résidus spectraux, E_A sont minimisés.

Une fois la décomposition spectrale effectuée, il est possible de procéder à la régression de la matrice de concentrations, C, de dimension n x j sur la matrice transposée des coefficients spectraux en utilisant la méthode des moindres carrés inverse.

$$\mathbf{C = \check{T}\check{V}} \qquad \mathbf{(30)}$$

$$\mathbf{V = (\check{T}^t \check{T})^{-1} \check{T}^t C} \qquad \mathbf{(32)}$$

où V est la matrice r x j des coefficients de régression. Ceux-ci permettent d'établir le lien entre les intensités résultant de la décomposition spectrale et les concentrations. La somme des carrés des erreurs ou résidus, associées aux valeurs de concentrations, E_C est ici encore, minimisée de façon à déterminer la meilleure approximation pour la matrice V.

$$\mathbf{E_C = C - \check{T}\check{V}} \qquad \mathbf{(32)}$$

La matrice V peut ensuite être utilisée afin d'estimer les concentrations des constituants d'un mélange à partir du vecteur t obtenu suite à la décomposition spectrale du spectre infrarouge de l'échantillon inconnu.

$$a = \hat{t}\,\check{B} \qquad (33)$$

$$\hat{t} = a\,\check{B}^t(\check{B}\check{B}^t)^{-1} \qquad (34)$$

$$\hat{c} = \hat{t}\check{V} \qquad (35)$$

b) Partial least square PLS Régression des moindres carrés partiels

La méthode partielle des moindres carrés, PLS (Partial Least Squares), est une technique de décomposition très semblable à la méthode PCR. Toutefois, plutôt que de décomposer la matrice des données spectrales et d'ensuite procéder à la régression comme le fait PCA, PLS utilise l'information reliée aux concentrations des constituants pendant le processus de décomposition. Ceci a pour effet de pondérer les coefficients relatifs au nouveau système de coordonnées. Ainsi, PLS tire avantage de la relation de corrélation existant entre les données spectrales et les concentrations de chacun des constituants.

De la même façon que la matrice des absorbances, la matrice des concentrations peut être décomposée en un produit de coefficients et de vecteurs de base. Deux ensembles de vecteurs de base et leurs deux ensembles de coefficients respectifs sont alors générés; le premier pour les données spectrales, le second pour les données de concentrations. Les deux ensembles de coefficients sont alors régressés.

L'algorithme de régression PLS

il existe de nombreuses versions de l'algorithme de régression PLSI. Elles déifient au niveau des normalisations et des calculs intermédiaires, mais elles aboutirent toutes à la même régression. Nous allons présenter la version de SIMCA-P qui découle tout naturellement de l'algorithme NIPALS.

On cherche à réaliser une régression d'une variable a expliquer y sur des variables explicatives $x_1, \ldots, x_p$, qui peuvent être hautement corrélé entre elles. Il peut même y avoir plus de variables explicatives que d'observations.

Par ailleurs les coefficients de régression doivent être interprétable. C'est-à-dire qu'on souhaite prendre en compte le fait que le chercheur mesure la contribution de la variable x à la construction de la variable y a l'aide du coefficient de régression. Lorsqu'un coefficient de regréions a un signe opposé à celui de la corrélation entre la variable explicative correspondante et la variable y il y a difficulté d'interprétation.

L'explication mathématique de la différence entre une corrélation simple et une corrélation partielle peut satisfaire le statisticien, mais a du mal à convaincre le praticien.

On peut atteindre l'objectif d'une régression interprétable de la manière suivante. On construit tout d'abord une composante

$$\mathbf{t_1 = w_{11}x_1 + \ldots\ldots + w_{1p}x_p} \qquad \mathbf{(36)}$$

$$\frac{\mathbf{cov(xj,y)}}{\sqrt{\sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{cov}^2\mathbf{(xj,y)}}} \qquad \mathbf{(37)}$$

Lorsqu'il y a des données manquantes, les formules (36) et (37) sont modifiée. On note xij la valeur de la variable xj pour l'observation i . On pose pour chaque observation i

Lorsqu'il y a des données manquantes, les formules (36) et (37) sont modifiées. On note xij la valeur de la variable xi pour l'observation i. On pose pour chaque observation

$$\mathbf{t_{1i}} = \frac{\sum_{\{\mathbf{j:xj\ ,existe}\}} \mathbf{w''_{1J}x_{ji}}}{\sum_{\{\mathbf{j:xj\ ,existe}\}} (\mathbf{w''_{1J}})^{\mathbf{2}}} \qquad \mathbf{(38)}$$

où le coefficient w' est défini à partir de

$$\boldsymbol{w'_{1i}} = \frac{\sum_{\{j:xj\ ,existe\}} x_{1J}y_{ji}}{\sum_{\{j:xj\ ,existe\}} (y_J)^2} \qquad \mathbf{(39)}$$

par normalisation

$$\boldsymbol{w''_{1j}} = \frac{\boldsymbol{w'_{1j}}}{\sqrt{\sum_{j=1}^{p} (w'_{1j})^2}} \qquad \mathbf{(40)}$$

S'il n'y a pas de données manquantes, lues formules (38) et (39) correspondent exactement aux formules (37) et (36). Les formules (37) et (36) correspondent a une application des principes de l'algorithme NIPALS et nous pouvons en redonner la signification. Dans la formule (36) t1, représente la pente de la droite des moindres carrés, passant par l'origine, du nuage de points (w_{1j},x_{ji}).

De même dans la formule (39) w_{1j} représente la pente de la droite des moindres carrés, passant par l'origine, du nuage de points (y_{ij} x_{ij}). Il est clair que ces pentes peuvent être calculées sans problème lorqu'il y a des données manquantes.

Puis on effectue une régression simple de y sur t1,

$$\mathbf{y = c_1t_1 + y_1} \qquad \mathbf{(41)}$$

où en est le coefficient de régression et y. le vecteur des résidus. D'où une première équation de régression

$$y = c_1 w_{11} x_1 + \ldots + c_1 w_{1p}\, x_p + y_1 \qquad (42)$$

Dont les coefficients sont très faciles à interpréter pour le chercheur.

Si le pouvoir explicatif de cette région est trop faible, on cherche a construire une deuxième composante t2, combinaison linéaire des xj, non corrélée a t1 et expliquant bien le résidu y1 . Cette composante t2 est combinaison linéaire des résidus x1j des régressions des variables xj, sur la composante t1 . On obtient t2 a l'aide de la formule

$$t_2 = w_{21} x_{11} + \ldots\ldots + w_{2p} x_{1p} \qquad (43)$$

$$w_{2j} = \frac{cov(x_{1j}, y_1)}{\sqrt{\sum_{j=1}^{p} cov^2(x_{1j}, y_1)}} \qquad (44)$$

Ces formules sont adaptées comme précédemment au cas de données manquantes on effectue ensuite une régression de y sur t1 et t2

En exprimant t1 et t2 en fonction des variables xj, l'équation de régression (41) peut s'écrire en fonction de ces variables. D'où une deuxième équation de régression plus précise que la première.

Cette procédure itérative peut se poursuivre en utilisant de la même manière les résidus y2, $x_{21}.....x_{2p}$, des régressions de $y, x_1, \ldots, x_p$, sur t_1, t_2.

6) La méthode non factorielle

Support Vector Machine (SVM)

Support Vector Machine (SVM) est principalement une méthode qui exécute des tâches de classification en construisant hyperplans dans un espace multidimensionnel qui sépare les cas d'étiquettes de classe différents. SVM prend en charge les tâches de régression et de classification et peut gérer des variables continues et catégorielles multiples. Pour les variables catégorielles une variable factice est créé avec des valeurs de cas que 0 ou 1. Ainsi, une variable dépendante composée de trois niveaux, disent (A, B, C), est représenté par un ensemble de trois variables indicatrices:
A: {1 0 0}, B: {0 1 0}, C: {0 0 1}

Pour construire un hyperplan optimal, SVM utilise un algorithme d'apprentissage itérative, qui est utilisé pour minimiser une fonction d'erreur. Selon la forme de la fonction d'erreur, les modèles SVM peuvent être classés en quatre groupes distincts:

- Classification SVM type 1 (également connu sous le nom de classification C-SVM)
- Classification SVM type 2 (également connu sous le nu-classification SVM)
- Régression SVM type 1 (également connu sous le nom de régression epsilon-SVM)
- SVM régression de type 2 (également connu sous le nu-régression SVM)

Régression SVM

y = f (x) + bruit

La tâche est alors de trouver une forme fonctionnelle pour f qui peut prédire correctement les nouveaux cas que le SVM n'a pas été présenté auparavant. Ceci peut être réalisé par la formation du modèle SVM sur un ensemble d'échantillons, à savoir, l'ensemble d'apprentissage, un processus qui implique, comme le classement (voir ci-dessus), l'optimisation séquentielle d'une fonction d'erreur. En fonction de la définition de cette fonction d'erreur, deux types de modèles SVM peuvent être reconnus:

RÉGRESSION SVM TYPE 1

Pour ce type de SVM la fonction d'erreur est:

$$\frac{1}{2} w^t w + C \sum_{i=1}^{N} \xi_i + C \sum_{i=1}^{N} \xi_i' \qquad (45)$$

que nous minimisons sous réserve de:

$$w^t \phi(x_i) + b - y_i \leq \varepsilon + \xi_i'$$

$$y_i - w^t \phi(x_i) - b_i \leq \varepsilon + \xi_i \qquad (46)$$

$$\xi_i' \; \xi_i \geq 0 \quad , i=1,\ldots,N$$

RÉGRESSION SVM TYPE 2

Pour ce modèle SVM, la fonction d'erreur est donnée par:

$$\frac{1}{2} w^t w - C(\varepsilon v + \frac{1}{N} \sum_{i=1}^{N} (\xi_i + \xi_i')) \qquad (47)$$

que nous minimisons sous réserve de:

$$(w^t \phi(x_i) + b) - y_i \leq \varepsilon + \xi_i$$

$$y_i - (w^t \phi(x_i) + b_i) \leq \varepsilon + \xi_i' \qquad (48)$$

$$\xi_i \; \xi_i' \geq 0 \quad , i=1,\ldots,N$$

Il y a nombre de grains qui peuvent être utilisés dans les Support Vector Machines modèles. Ceux-ci comprennent linéaire, polynôme, fonction radiale de base (RBF) et sigmoïde:

où $\vec{w}$ le (pas nécessairement normalisé) un vecteur normal à l'hyperplan. Le paramètre $\frac{b}{\|\vec{w}\|}$ détermine le décalage de l'hyperplan à l'origine le long du vecteur normal $\vec{w}$.

Remarque : il ya des autre méthodes factoriel et non factoriel qu'on peut utilise pour faire la régression avec un très grande nombre de variable comme les réseaux de neurone, régression locale pondéré, Ridge régression. NPLS … ce sont des méthodes capables de gérer un très grand nombre de variable et des espaces multidimensionnelles.

Chapitre 3. Partie expérimentale

1) Méthode

L'analyse multivariable quantitative comporte deux phases. La première phase consiste a l'établissement d'un model de calibration. Celui-ci doit ensuite être soumis à une phase de validation de façon à déterminer l'exactitude et la précision des estimations obtenues avec le modèle lors de la phase de prédiction.

a) Phase de calibration

Les résultats statistiques générés par le logiciel (R) utilisé proviennent d'une technique de validation interne appelée méthode de validation croisée (cross-validation), cette dernière procède à la mise l'épreuve du modéle de calibration en utilisant les échantillons de calibration préparés au laboratoire comme inconnus. La méthode consiste à retirer l'un des échantillons de calibration, à calculer le modèle de calibration à partir de l'ensemble de calibration comptant un échantillon de moins et à estimer la concentration des constituants de l'échantillon retiré. L'échantillon est ensuite remis dans l'ensemble de calibration et un second échantillon est retiré. Le programme calcule de nouveau le modèle de calibration à partir du nouvel ensemble de calibration et les valeurs de concentrations pour cet échantillon. Les concentrations de tous les échantillons de calibration sont ainsi prédites jusqu'au dernier. À partir des valeurs de concentrations prédites, la somme des carrés des erreurs résiduelles de prédiction (PRESS), le coefficient de détermination (R^2) et l'erreur standard de validation croisée (SECV) peuvent être calculés pour ce constituant.

Pour sélectionner le nombre approprié de vecteurs de base, le test statistique F est appliqué aux valeurs de PRESSr calculées pour chaque modèle comportant un nombre r de vecteurs de base (r = l, 2, 3 …). Généralement, lorsque les valeurs de PRESSr sont rapportées en fonction du nombre de vecteurs de base, un minimum peut être clairement identifié. La valeur F rapport entre le PRESSr, et le PRESS minimum, est alors calculée pour chaque modèle comportant un nombre de vecteurs de base inférieur ou égal à celui du modèle de PRESS minimum, puis comparée à la valeur critique (avec $\alpha = 0{,}95$ et (m-l) degrés de liberté au numérateur et au dénominateur (m) , m étant le nombre total d'échantillons dans l'ensemble de calibration moins le nombre d'échantillons retiré lors du processus de validation croisée)

b) Phase de validation

Une fois le modèle de calibration optimisé, celui-ci doit être mis à l'épreuve avec un ensemble de validation indépendant. Il ne s'agit pas ici d'employer la méthode de validation croisée décrite précédemment, mais bien d'utiliser des échantillons de validation dont la composition se rapproche de celle des échantillons de sol qu'on souhaite analyser. Les prédictions obtenues sur ces échantillons et sur les échantillons réels de sol subissent des tests statistiques (Limit Tests) permettant d'évaluer la performance du modèle de calibration.

En premier lieu, les valeurs de coefficients spectraux calculés pour l'échantillon de validation sont comparées aux coefficients spectraux des échantillons de calibration. Ceci dans le but de vérifier si les valeurs obtenues pour l'échantillon de validation sont du même ordre de grandeur que celles extraites de la décomposition spectrale de l'ensemble de calibration. Ensuite, un test statistique F est appliqué aux résidus spectraux de l'échantillon afin de, déterminer si ceux-ci sont comparables aux résidus spectraux des échantillons de calibration. Ce test permet entre autre d'identifier les échantillons contaminés. Ensuite, les concentrations prédites sont comparées aux concentrations des échantillons de calibration de façon à s'assurer qu'elles se situent à l'intérieur des domaines de concentration du modèle pour chacun des constituants. Les valeurs limites choisies pour les tests statistiques sont de 0,95 pour le test F , finalement il faut faire des testes de comparaison entre les distances (la concentration réel et la concentration prédite par le model)…

2) Echantillonnage

On va travailler dans cette étude avec deux types d'échantillons, des échantillons naturellement collectaient dans des endroits et des profondeurs déférentes de la zone (Rabat Salé), et des autres échantillons artificiels préparer au laboratoire, on prépare des échantillons artificiels pour assurer une très bonne variation et une bonne dispersion dans l'intervalle de mesure.

a) Préparation des échantillons

Donc on va préparer un plan expérimental pour encadrer tous les cas possibles dans la nature. Le plan expérimental dont les variables et leur niveau sont.

Tableau 2: les niveaux de chaque variable

	Le sol	La tourbe	Azote
Niveau bas	1 000 g	250 g	12 g
Niveau Haut	750 g	0	0

On obtient le plan suivant (on prépare les échantillons à l'aide des mélanges motionnés dans la matrice d'expérience suivante).

Tableau 3 : matrice d'expérience pour la préparation des mélanges (le plan composite a trois facteurs)

N°Exp	SOL	MO	NO3
1	1.0000	0	0
2	0.9700	0	0.0300
3	0.5000	0.5000	0
4	0.4700	0.5000	0.0300
5	0.9850	0	0.0150
6	0.7500	0.2500	0
7	0.7200	0.2500	0.0300
8	0.4850	0.5000	0.0150
9	0.7350	0.2500	0.0150
10	0.8675	0.1250	0.0075
11	0.8525	0.1250	0.0225
12	0.6175	0.3750	0.0075
13	0.6025	0.3750	0.0225

b) Prétraitement physique

Apres la collecte et la préparation des échantillonnes on passe au prétraitement physique

Le prétraitement physique consiste à :

Séchage a l'aire sec pendant 6h 50°C pour enlever l'eau non liée, car l'eau fossé les résultats spectrales (la liaison (O-H) il donne un pic très large)

Tamisage1 : pour enlever les caïeux les feuilles, les tiges et les corps externe du sol

Broyage : pour homogénéiser le milieu

Tamisage2 : à 0,2 mm pour éviter la forte hétérogénéité des graines de sol

Compactage (pour l'analyse spectrale) : pour éviter le vide entre les grains et pour diminuer la diffusion de la lumière dans l'analyse spectrale

c) Traitement chimique des échantillons :

L'objectif de ce traitement est la détermination de concentration inconnue de l'élément dans le sol, Au cours de ce traitement on va analyser la teneur du sol en matier organique et le taux d'azote totale, pour effectuer ces analyses on va suivre les procédures analytiques standard

On va travailler avec 20 échantillons pour faire la calibration et pour la validation on va utiliser la méthode de validation croisée avec huit échantillons.

Le calcul est assuré par deux logiciels, le IBM SPSS pour faire la RLM pas a pas ascendante et le logiciel R pour calculer tous les autre type de calcul

d) Analyse spectrale

L'objectif de cette partie est l'obtention de l'empreinte digitale issue de l'interaction entre la lumière et les éléments chimiques de sol.

Avant de commencer l'analyse spectrale, il faut vérifier certains paramètres liés au spectromètre, tout d'abord il faut vérifier que la résolution de spectromètre égale 16 et le pas de variation entre les longueurs d'ondes égale 2,4 ou 2,5 nm ,De plus, il est très important de vérifier la distance entre la source lumineuse et le sol , les paramètres mentionnés doivent être constants pendant toute la durée de l'analyse et pour tous les échantillons .

Nous allons se focaliser sur la réflectance car la nature de la matrice de sol nous impose d'utiliser la quantité de lumière réfléchie ce qui nous a poussé à utiliser un détecteur de réflectance.

Le spectromètre nécessite un étalonnage après un certain nombre d'analyse (on étalonne le spectromètre lorsqu'il commence de nous donner des résultats suspects). Nous utilisons alors un plan (le Back-Ground) (figure a) pour régler le 100%.

Apres le réglage de l'instrument il faut préparer le sol (on va utiliser le sol qui est déjà prétraité physiquement), il faut bien le compacter dans une boite pétrie d'une épaisseur de 1cm, Ensuite, on va scanner l'échantillon de sol cinq fois dans des zones différentes et on va prendre la moyenne des spectres, on fait tout cela pour affaiblir le problème de la diffusion de la lumière et pour assurer le contact entre tous les éléments du sol et la lumière.

On obtient des résultats de transmittance sous forme numérique dans un fichier Excel, Enfin, il faut regrouper les résultats dans une seule matrice afin de faciliter le traitement mathématique.

3) Analyse des données, modélisation et discutions

Le schéma si dessous résume le travaille à suivre pour effectuer la modélisation

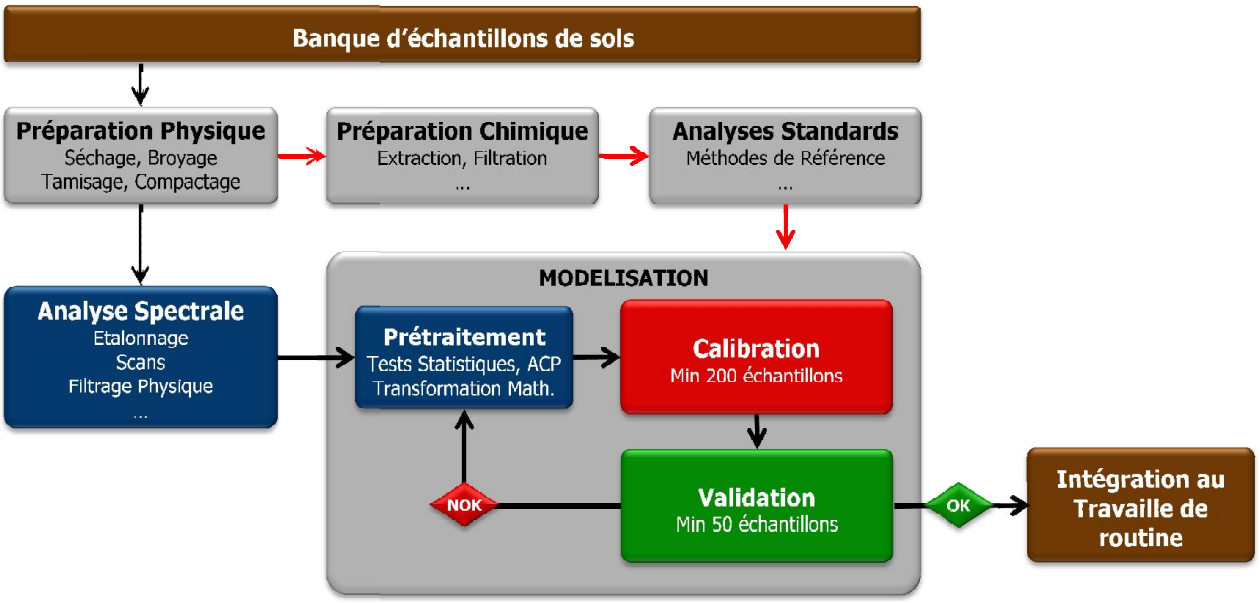


Figure 5 : les étapes chimiométrique de la modélisation

a) Traitement statistique des données

La première étape de traitement statistique commence par le calcul des paramètres de position et de dispersion. Les résultats sont rassemblés dans l'annexe A

On trouve que l'on a une très grande variation et grande dispersion des données avec une grande étendue, c'est pourquoi avoir jugé important de réaliser un test des points aberrant

Test des points aberrants

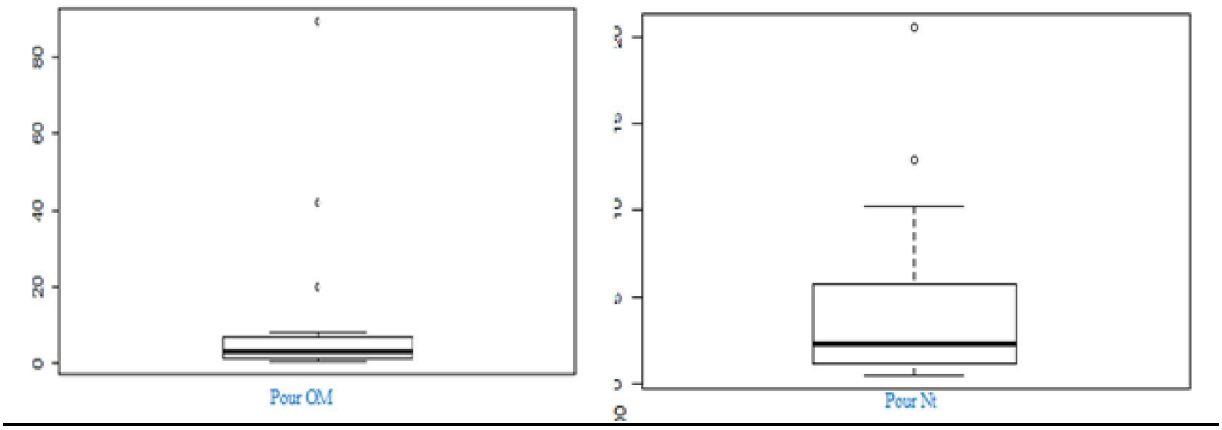


Figure 6 : box plot pour la manière organique et azote totale

L'analyse de la figure 7 montre l'existence de trois points, en effet il s'agit des échantillons synthétisés au laboratoire dont lesquels nous avions introduit des fortes quantités de la tourbe et de nitrate.

b) Acquisition des spectres et sélection de la région d'analyse

Les Figures 8 et 9 présentent respectivement les spectres des échantillons de calibration et le spectre moyen. À la Figure 8, les variations de concentration en MO et la distribution de ces concentrations sont clairement visibles dans la zone (1800-1950nm). Par contre, les changements spectraux associés aux variations de concentrations Nt constituants minoritaires sont nettement moins évidents dans la zone (1300-1500nm).

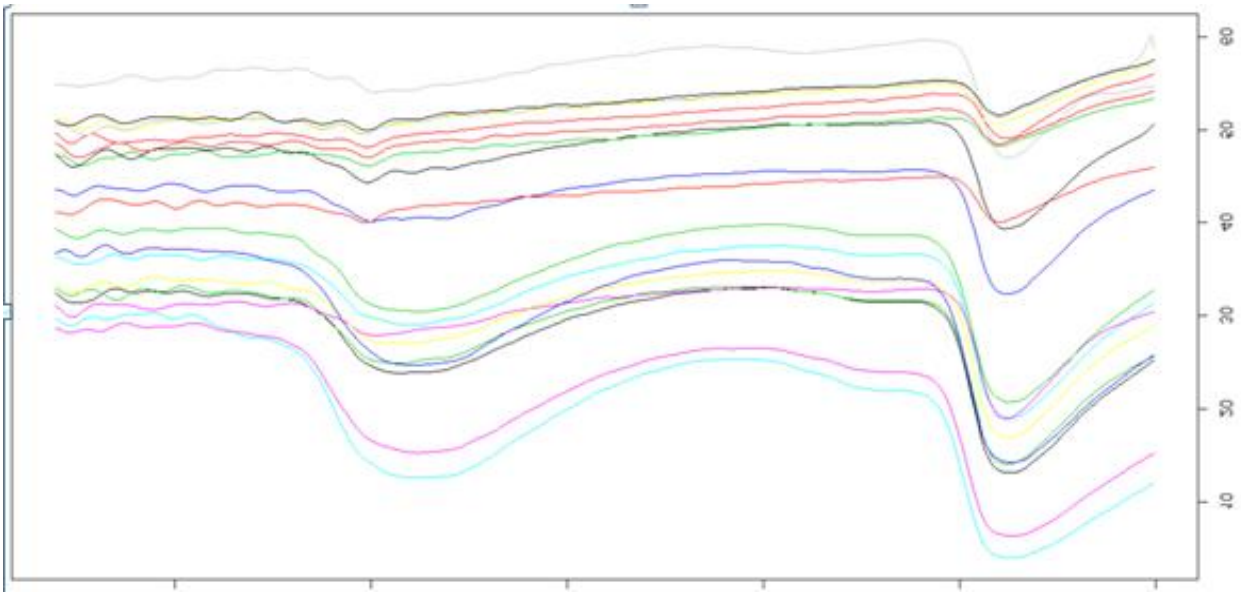


Figure 7 : représentation du spectre sans transformation mathématique

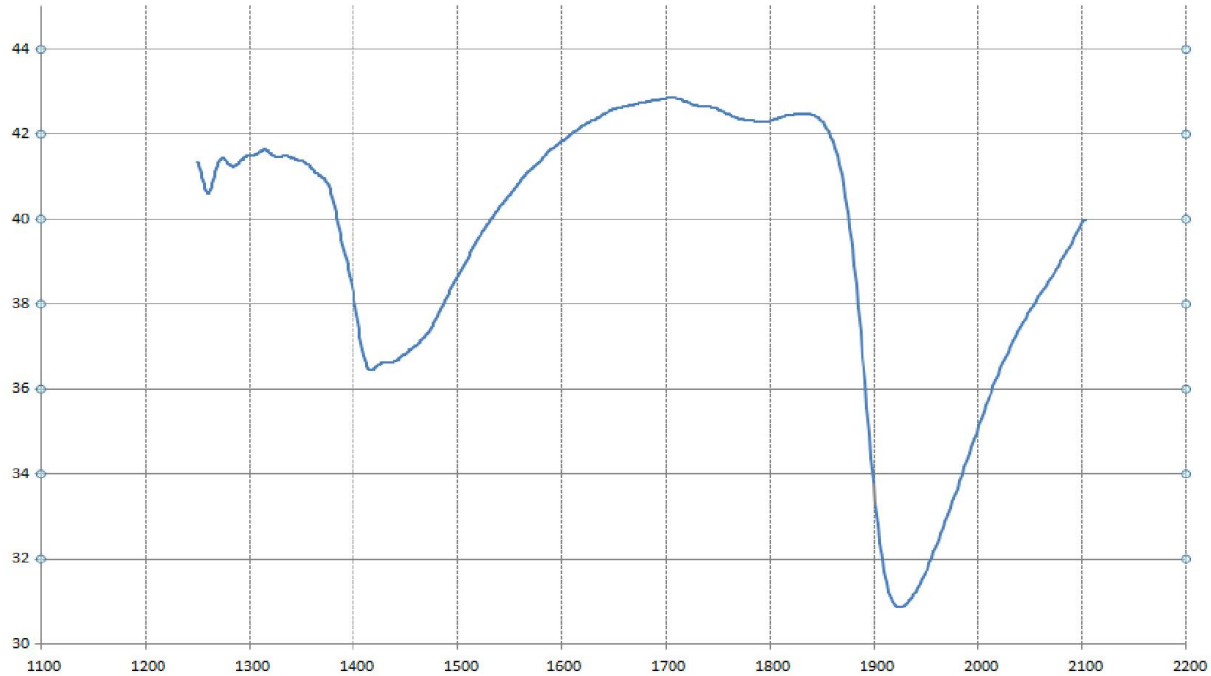


Figure 8: representation de spectre moyen

Suivant respectivement les deux figures 8 et 9, quatre zones de variation ont été générées par l'empreinte digitale. Nous avons une bande intense entre (1850 et 2100nm). c'est une bande liée à la variation de la MO et la liaison O-H, nous avons aussi une bande moyenne entre 1350 et 1600 nm, cette zone est liée à la variation de nitrate et d'azote totale, ainsi que des faibles variations de spectre dans les zones entre 1700-1800nm et entre 1250-1300nm sont des zones liée avec la MO et le carbone organique.

A) La Modélisation de la matière organique

L'analyse spectrale (PIR) toujours nous donne un nombre de variables très grandes (600 variables) que des observations ce qui provoque la singularité de la matrice X^TX et aussi le problème de redondance et la forte colinéarité entre les variables spectrales peut conduire à une situation de quasi-singularité de la matrice X^TX , donc pour régler ces problèmes il faut cibler certaines variables pour créer le modèle de régression de tel sort le coefficient de détermination doit être maximal et la significativité des variables doit être hautement significative (faire le RLM pas à pas)

1) La régression linéaire multiple pas a pas

D'abor on va commencer la modélisation par un RLM pas à pas pour sélectionner les variable qui ont un poids sur la réponse , le calcul est assuré par SPSS IBM.

On obtient les résultats suivants

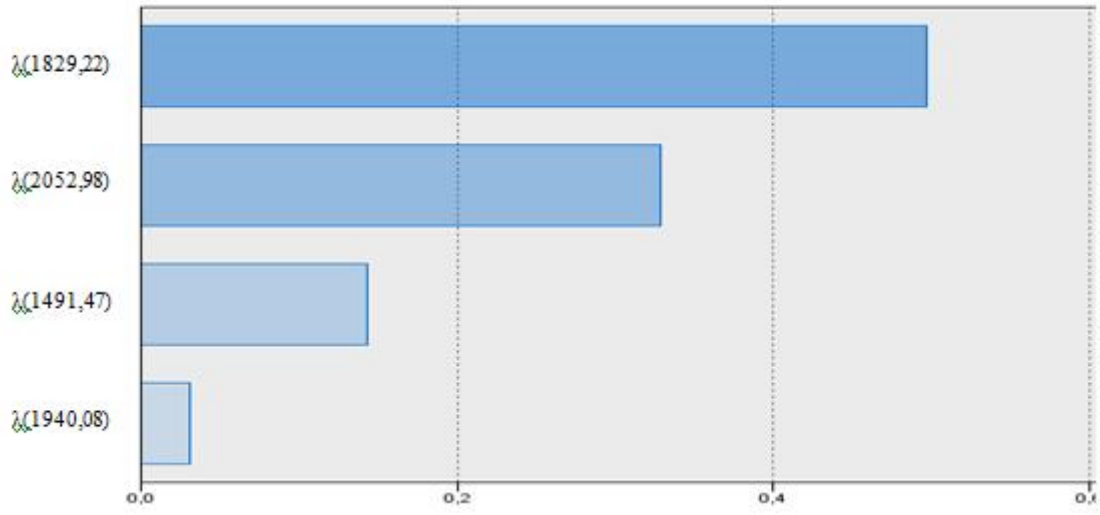


Figure 9 : le poids des variables sur la réponse après l'élimination des variables qui n ont pas de poids

Suite au calcul du test Fisher partiel, Il nous reste quatre variables qui ont un effet significatif sur la réponse. Les variable qui reste sont des longueurs d'onde inclue dans la zone liée a la variation de la matière organique

On trouve un $R^2 = 93,7$ et $R^2_{adj} = 91,9$ avec une erreur standard d'estimation = 0.75, On a un R^2 satisfaisant, c'est-à-dire on arrive à crier un modèle capable d'expliquer 93,7 des points, mais l'erreur standard reste discutable.

Pour l'étude de résidus dans la figure 11 on 'a conclu que le résidu suit une loi normal, on constate qu'aucune transformation est nécessaire car on ne trouve pas de concavité ou d'anomalie dans la distribution des résidus

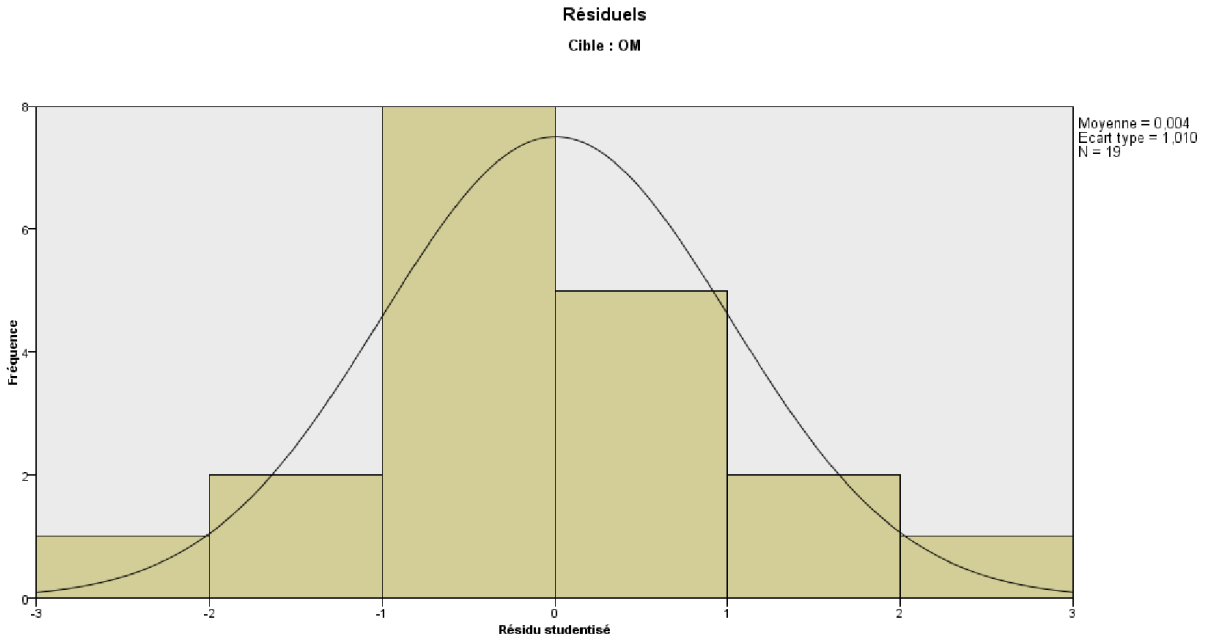


Figure 10 : distribution des résidus

Pour juger la significativité de la pente du modèle on va calcule les sommes des carrés des écarts pour faire le test d'ANOVA

Tableau 4 : ANOVA pour la RLM

Modèle	SCE	Ddl	Carré moyen	F	p-value
Régression	116,654	4	29,16	51,783	0,00
Résidus	7,885	14	0,563		
Totale	123 ;539	18			

On conclu a partir de tableau 4 que le test d'ANOVA donne un p-value supérieur a 0,05 se qi signifie que l'hypothèse nulle est rejeté donc on a l'existence d'une pente significative, et selon le test de student (annexe B) on conclu que les quatre variable ont un poids significative sur la repense

Validation avec six nouveaux échantillons

Tableau 5 : le calcule de résidu et somme care des écart

		x_Axis:Wavelength (nm)								
Ech	Y	2052.9	1940.0	1829.2	1491.4	y^	(y-ybar)²	(y-y^)²	(y^ -ybar)²	Ei
DM2	5,611	42,453	37,248	45,452	42,734	3,043	5,269	6,593	0,074	2,568
DM3	2,768	41,663	35,341	43,990	41,251	2,451	0,300	0,100	0,747	0,317
DM4	2,754	43,628	38,241	44,413	41,413	1,130	0,314	2,639	4,775	1,624
DM5	1,530	50,072	41,230	52,799	49,593	3,744	3,186	4,900	0,184	-2,214
DM6	2,227	50,708	44,505	51,560	48,292	1,920	1,184	0,094	1,948	0,307
DM7	3,526	50,612	45,329	51,866	49,060	2,361	0,044	1,356	0,910	1,165
DM8	3,560	49,798	44,557	51,108	48,114	2,248	0,060	1,722	1,139	1,312
DM9	3,641	50,084	44,985	51,910	49,374	2,895	0,106	0,557	0,177	0,746
DM10	4,219	55,537	50,025	57,228	54,014	3,143	0,817	1,159	0,030	1,076

A partir des valeurs prédites par le modèle on va calculer la divination entre la valeur vraie et la valeur prédite et on vas faire un test de Fisher (ANOVA)

Tableau 6 : ANOVA pour les échantillons de validation

Sv	SCE	Ddl	CM	F	F(5%,1;N-2)
Résiduelle	1,29	8	0,162		
Régression	9,98	1	9,982		
Total	11,28	9		61,48	5,31

A partir du tableau d'ANOVA, on 'a F_{cal} supérieur a F_{cri} ce qui signifie que l'hypothèse nulle est rejetée donc il existe un effet de modèle sur la régression et aussi on a un R^2= 88,4 % avec une distribution de résidus aléatoire comme l'illustre la figure 14

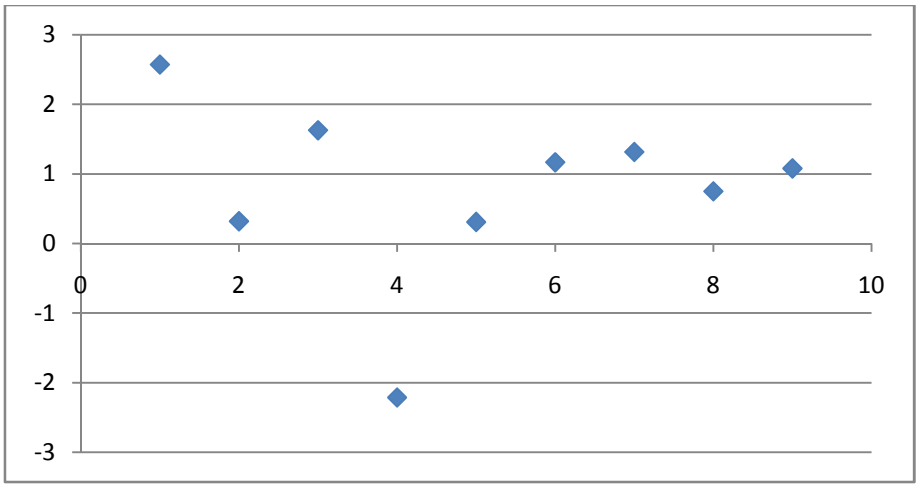


Figure 11 : la distribution des residus en fonction de la valeur prédite

Dans ce type de régression (RLM) , nous avons rejeté plusieurs variables qui comportent l'information issue de l'interaction entre la matière organique et la lumière. Ce qui implique qu'il ne faut pas prendre le risque de ne pas choisir les variables convenables.

Alors il faut trouver des méthodes qui permettent de gérer un grand nombre de variables et de résoudre le problème de redondance et la forte colinéarité qui existe entre les variables, En effet, la solution est bien évidemment les méthodes factorielles.

2) Modalisation avec les méthodes factorielles

Avon de commencer la modélisation factorielle il faut faire une étude factorielle sur les données pour déterminer le nombre des axes factoriels nécessaires pour réduire la taille de matrice et pour concentrer l'information dans une dimension plus réduite

a) Analyse en composante principale

ACP est la méthode factorielle la plus utilisée pour comprendre la variation des variables dans l'espace des individus et la variation des individus dans l'espace des variables, et aussi dans le cas où on a des données de grande taille, son principe est basé sur la réduction de dimension initiale (k) en une dimension réduite (p) avec p<<k et basé aussi sur la concentration des informations sur des axes factoriels.

A partir des valeurs propres on conclu qu'on a besoin de 6 axes factoriels pour expliquer 99,99% d'information contenus dans une matrice de 563 colonnes et on observe qu'une seule composante principale capable d'expliquer 98,75 % d inerties totales et les autres axes résume le bruit (annexe B)

A partir de cos² (annexe B) on peut dire que tous les individus sont bien expliques par la 1er composante car le coefficient de détermination est grand R^2= con² > 80 % , et à partir de la contribution on constate que la teneur on matière organique est corrélé négativement avec la première composante, (S3 et S4 sont des échantillons riches en MO et S22 et S20 sont des échantillons pauvres en MO)

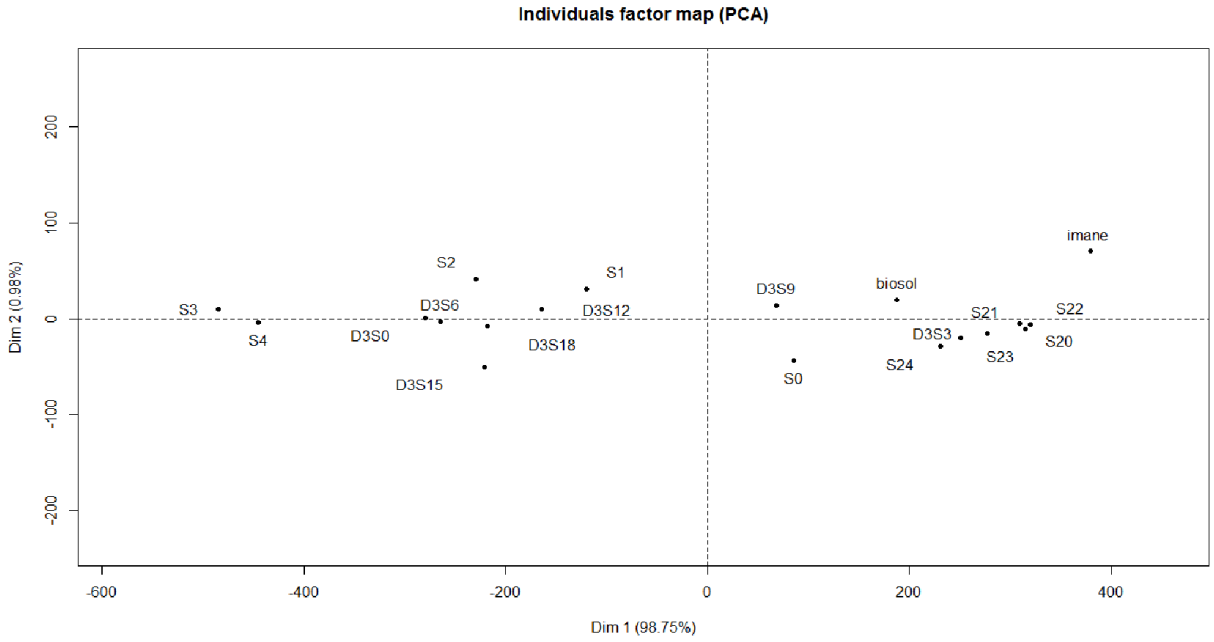


Figure 12 : projection des individus sur un plan de deux dimensions (CP1 et CP2) dans l'espace des variables

3) La régression sur la composante principale (PCR)

On va effectuer une régression sur la composant principale avec des données brutes sans transformation mathématique, et on va seulement centre notre matrice des données, on obtienne les résultats suivants :

A partir de figure 14 Projection spectre de loading on observe que le taux de bruit enregistré par le spectre augmente en fonction du nombre composante principale, on observe qu'un modèle avec trois composante principale, donne une bonne variation dans les zones liées avec la matière organique.

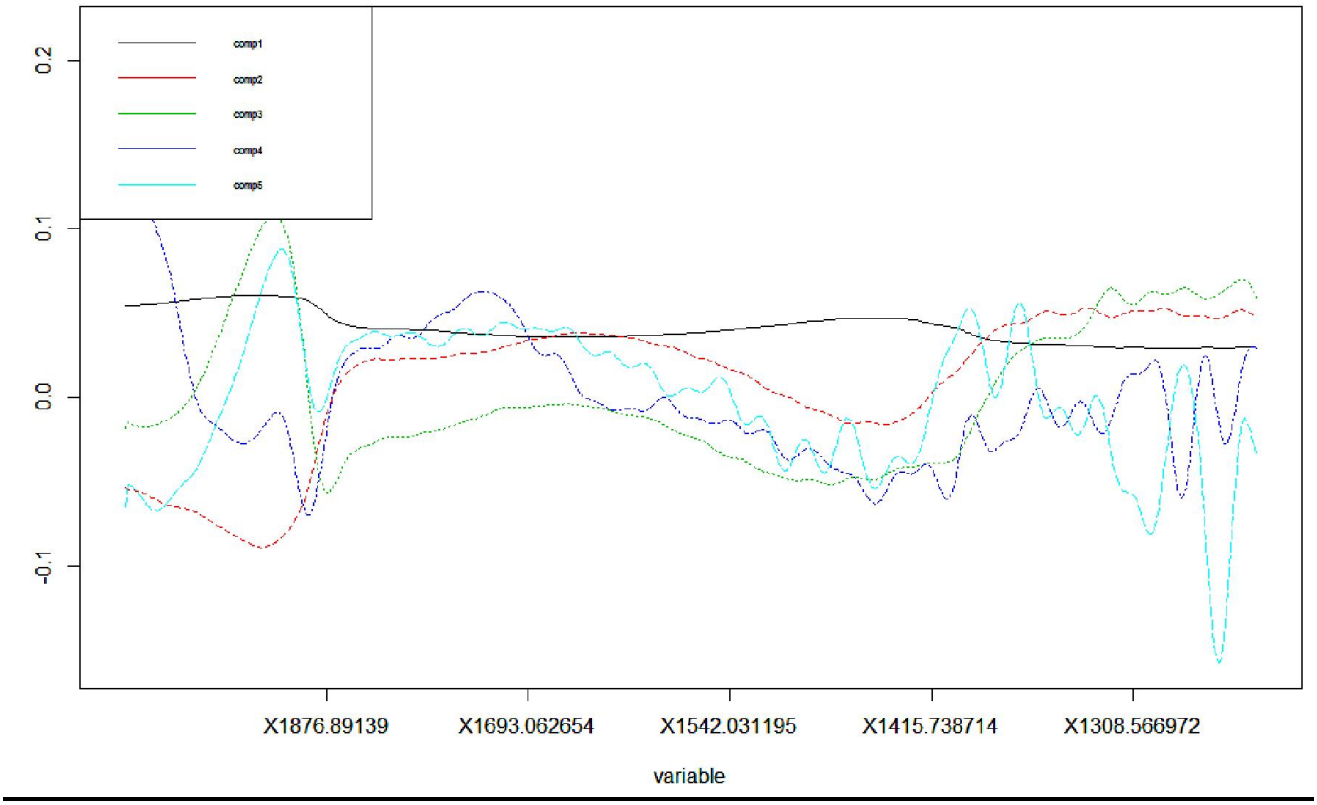


Figure 13 : projection de spectre de loading avec six CP

Dans la figure 15 de variation de R^2 et RMSEP en fonction de nombre de CP on observe que trois CP donnent un meilleur R^2 qui égale 84 ,64 % et un RMSEPc le plus petit qui égale 1%

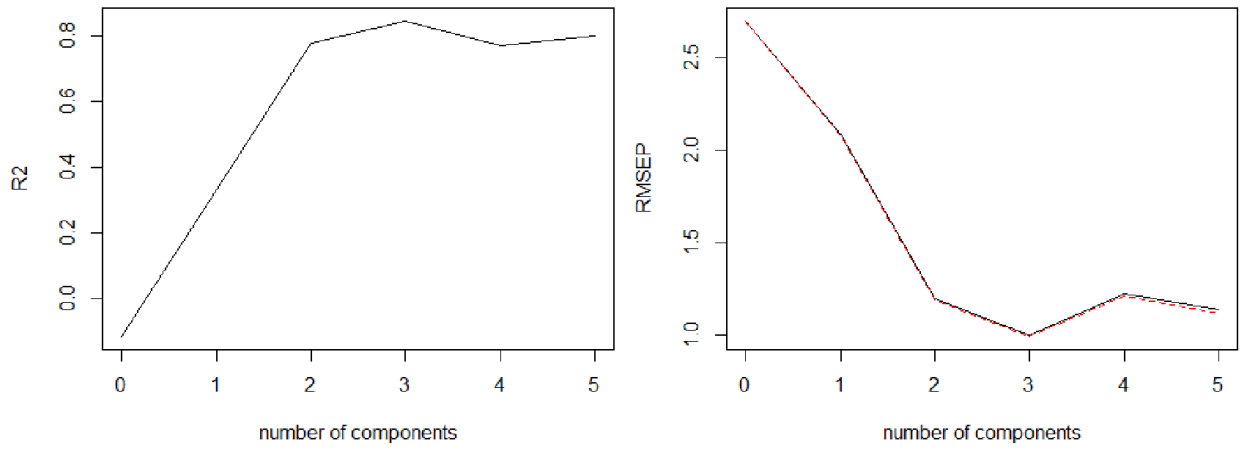


Figure 14 : variation de R^2 et RMSEP en fonction de nombre de CP

Par conséquent on va garder trois composantes principales pour crier un modèle PCR et on va calculer la RMSEPv et RDP à partir des échantillons de validation

La validation de modèle PCR

Ce pendent on remarque, avec trois CP on obtienne un RMSEPv = 0,68 qui est satisfaisante et un RPD = 1,74 inférieur à 3(selon la littérature un RPD inférieur à 3 signifié une mauvaise qualité de régression), on trouve un R^2 de validation égale 96 % donc on peut conclure qu'on a un modèle bien validé et une mauvaise qualité de régression.

Problème : L'ACP est réalisée sans garantie que les CP seront optimales pour expliquer la réponse (peut être les trois CP expliquer la variation d'un autre élément et n est pas la teneur en MO) , et aussi peut être on a un problème de déformation des distances an cours de la projection.

Donc en passe a la régression PLS

4) La régression des moindre carre partial PLS

À partir d'ACP on a conclu que 1 CP résume 98,7 d'informations. Est-ce que cette CP est suffisant pour construire un modèle PLS de grande capacité prédictive ou il faut intégrer autre CP pour améliorer le modèle ?

On a effectuer un PLS sans réduction des données avec l'utilisation de la validation croisée (leave one out) , pour valider ce modèl.

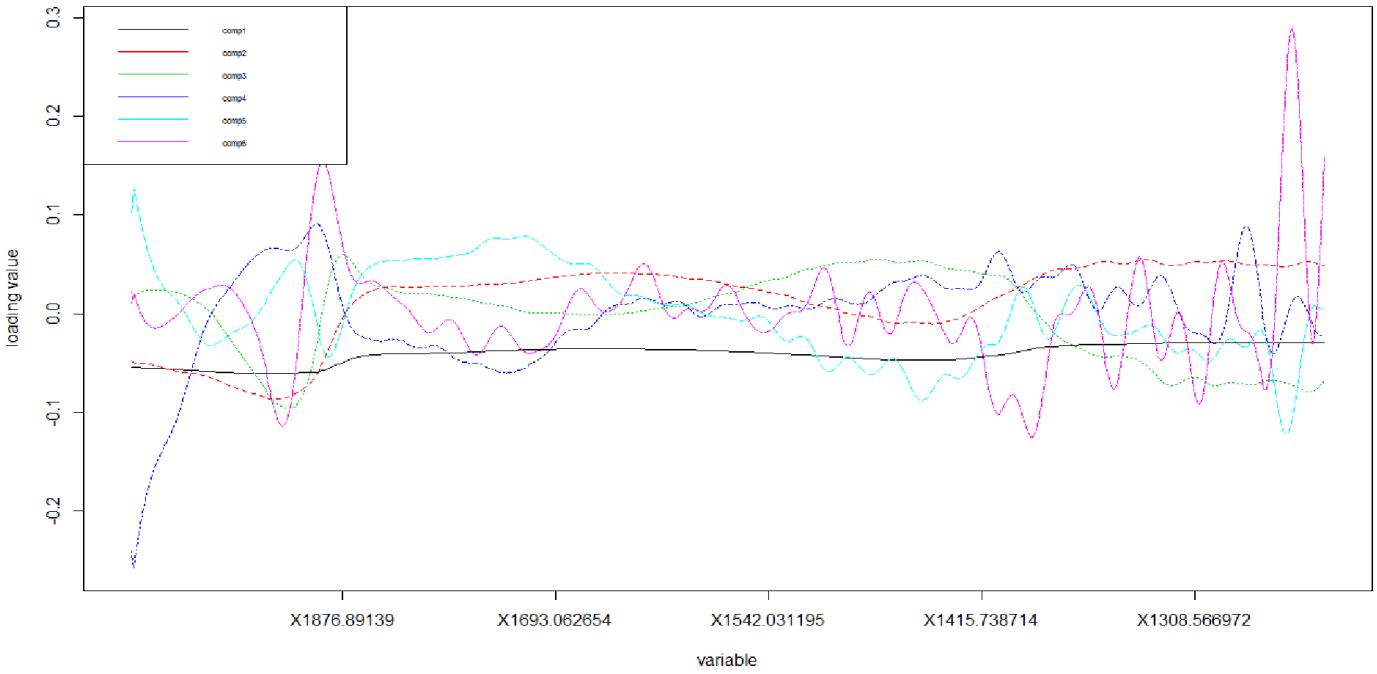


Figure 15 : projection de loading pour PLS

D'après la figure 18 On observe que plus le nombre de composante principale augmente plus le bruit enregistré augmente, et que la premier composante principale ne donne pas une forte variation de bruit de fond et une très faible variation liée avec la zone de MO (presque une ligne constante), et on a une variation remarquable entre (1800 nm et 1900 nm) et (1400 1200) lorsque nombre de CP est supérieur a 1 (les deux zone d'après la littérature sont les zone ou on a l'enregistrement de variation de la MO).

Il faut analyser la variation de R^2 et de RMSEP en fonction du nombre CP pour choisir le nombre de CP qu'il faut retenu pour construire le modèle PLS

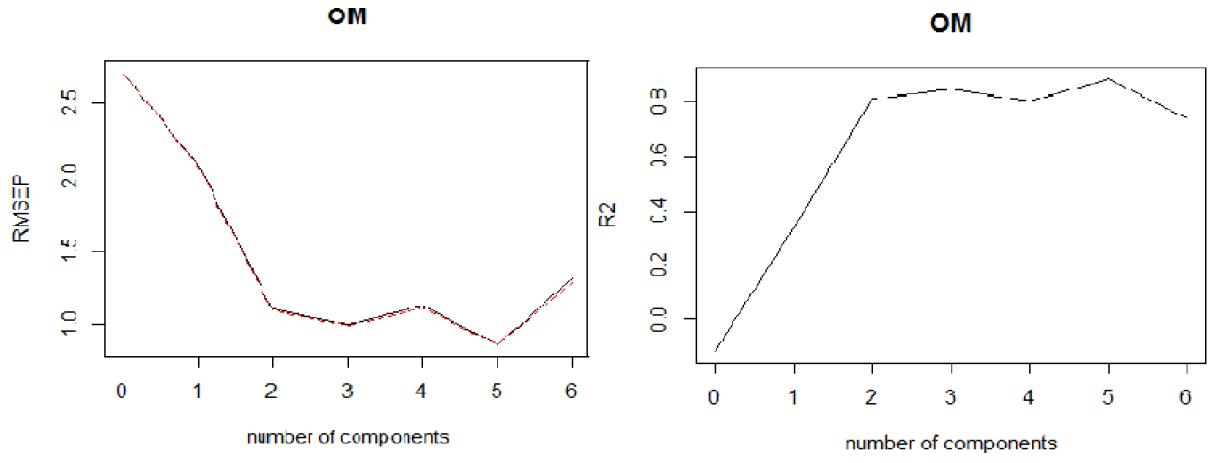


Figure 16 : la variation de R^2 et RMSEP en fonction du nombre CP pour PLS

a partir de la figure 17 la variation de R^2 et RMSEP en fonction du CP, on va choisir 5 CP pour construire le modèle PLS car 5 CP nous donne un R^2 le plus élevé 88,4 et un RMSEP le plus petite 0,8.

La validation du modèle

Pour faire la validation on va calculer les paramètres suivants

$$RMSEPc = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_{ci})^2}{n_c}} \qquad (49)$$

$$RMSEPv = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_{pi})^2}{n_p}} \qquad (50)$$

$$RPD = \frac{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}{\sum_{i=1}^{n}[(y_i - y_{pi})\sqrt{\frac{\sum_{i=1}^{n}(y_i - y_{pi})^2}{n_p}}} \qquad (51)$$

Pour 5 CP on trouve un RMSEPc de calibration égale 0,11% et un R^2 =88,5% , sont ac ceptables et très satisfaisants c'est pourquoi on prend la décision de créer un model PLS ave c 5 composante principale et on va essayer de valider ce modèle avec dix nouvelles échantill ons et on va calculer le RMSEPv et RPD.

Après le calcul de paramètres de validation on trouve un RMSEPv de validation égale à 0,4 très satisfassent avec un coefficient de détermination maximale égale à 99,8% et un RPD égale à 3.572. à cet égard on accepter le modèle PLS pour la MO comme un modèle bien validé avec une bonne qualité prédictif

N B ! : le risque qu'on a prendre pour les tests statistiques égale 5% . (α = 5%.)

Comme on a vu précédemment qu'on a besoin de 5 CP pour construire le modèle, donc on n'arrive pas a concentrer les données sur un ou deux CP pour affaiblir la bruit enregistré par le spectre.

Pour faire face à ce problème, nous avons opté pour un prétraitement mathématique des spectres afin de diminuer leur bruit de fond et aussi améliorer la qualité prédictif des modèles

5) Prétraitement mathématique et amélioration des résultats

A fin d'améliorer la qualité du spectre et pour éviter le problème de la délusion de la lumière on va faire deux typee de prétraitement mathématique, le premier consiste a utiliser la première dérivée de Savitzky-golay (spectres de figure 18) avec un degré de polynôme égale 3 créer avec 7 point , la deuxième transformation consiste a utiliser la MSC (spectres de figure 19).

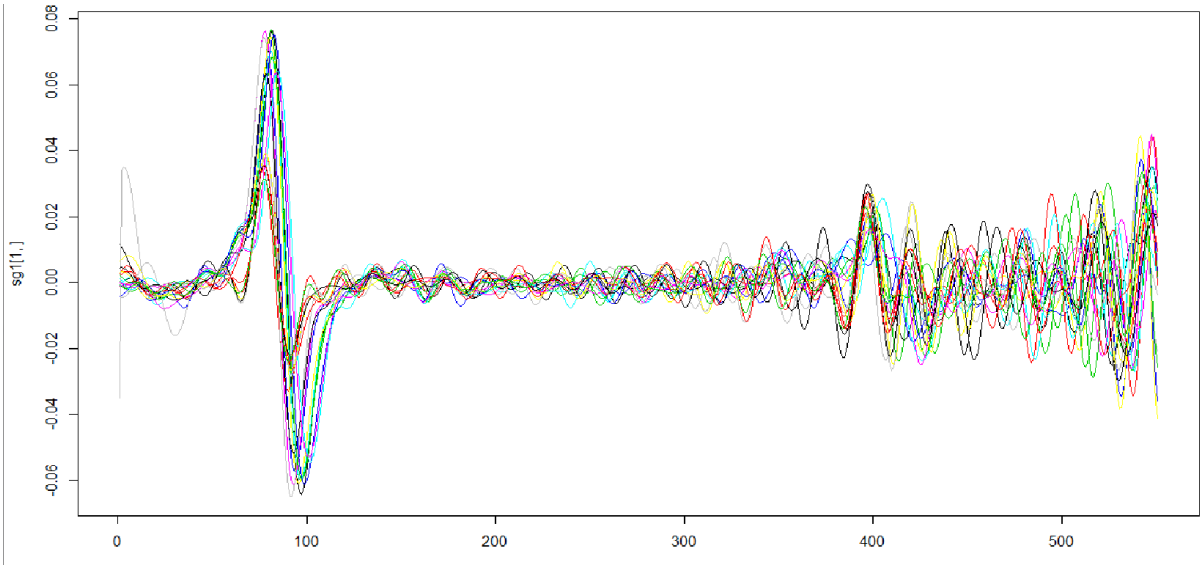


Figure 17 : le spectre après la première dérivation de Savitzky-golay

Multiplicative Scatter Correction MSC

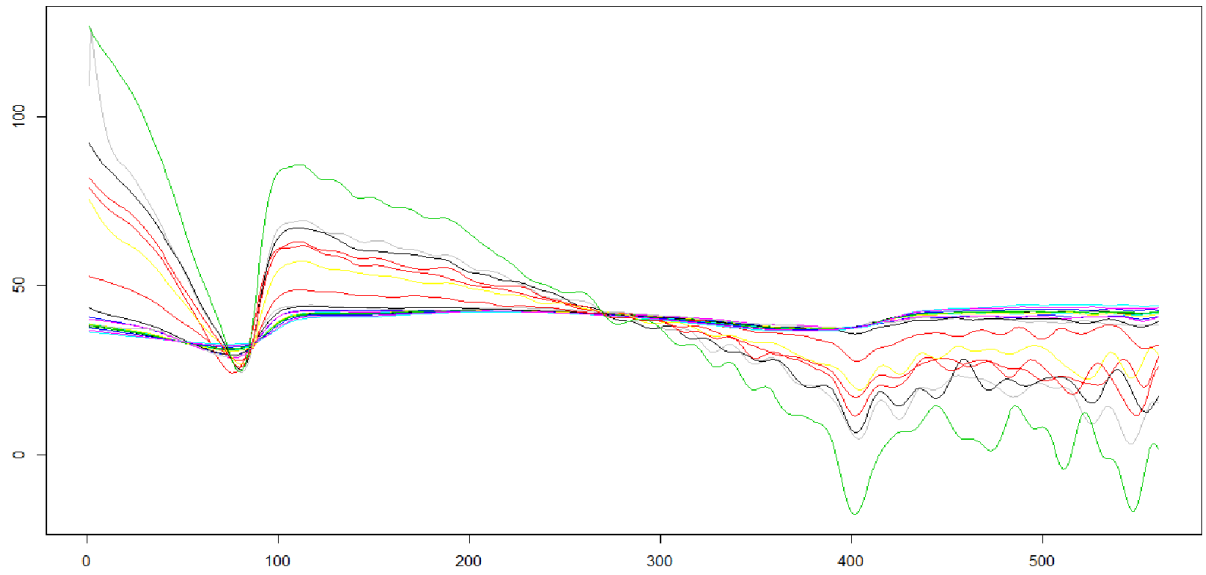


Figure 18 : le spectre après le traitement de Multiplicalive Scatter Correction MSC

Pour la MSC et standard normal variate SNV on a trouve les mêmes résultats (mêmes spectres)

6) Régression PLS Pour la première dérivé de savitzky-golay

On constat que 8 composante principale résume 96,72 de l'information initial donc la capacité de concentration des données est diminué.

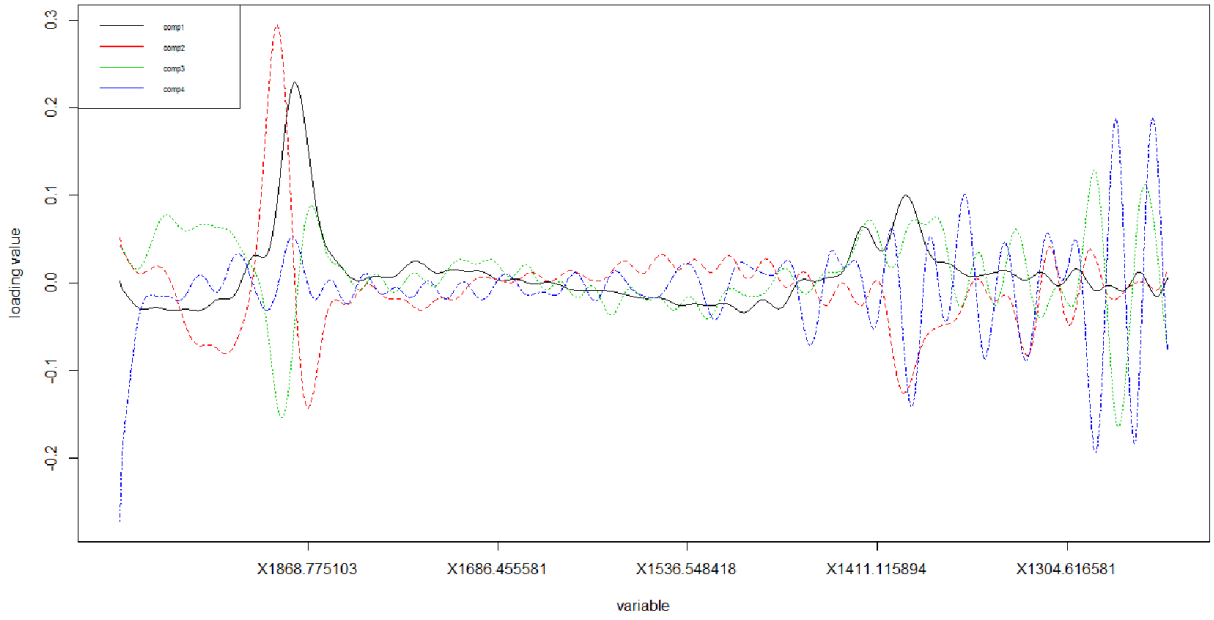


Figure 19 : projection des loading

On observe que les deux premier CP donnent des bons spectres de loading ou il apparie deux pics intense dans la zone (1900 – 1800 nm) et autour de (1500 – 1600nm)

Pour le choix de nombre de CP on va analyser la variation de R^2 et RMSEP

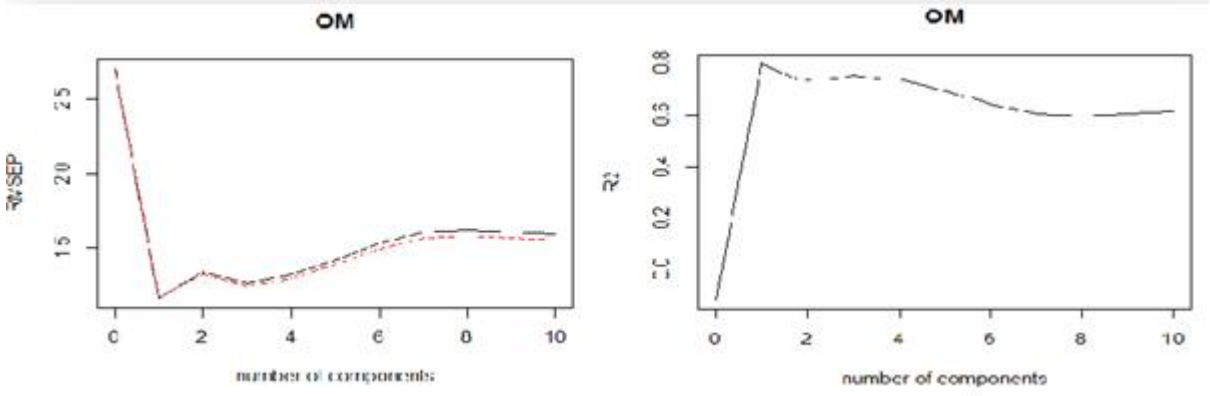


Figure 20 : variation de R^2 et RMSEP pour PLS avec la première dérivative S-G

A partir de R^2 et RMSEP on va choisie 1 CP donc on a réussi a diminuer le nombre de CP mais on n'arrive pas a améliorer le R^2 et le RMSEP, mais le modèle reste satisfaisante.

7) Régression PLS avec prétraitement mathemathique des données (MSCet SNV)

On n arrive pas a améliorer la qualité de régression avec la transformation mathématique pour la MO en trouve un R^2 maximale de prédiction égale 57% et RMESp minimale égale 1,7 %.

B) Modélisation de l'Azote totale

On va suivre la même procédure analytique de modélisation de la MO pour construire un modèle d'azote totale

1) Analyse en composante principale

On trouve qu'une seul CP capable d'expliquer 98,254 d inertie totale et sept CP expliquent 99,99 d'informations.

2) Modélisation PLS

On va faire une projection de laoding sur les trois premiers CP.

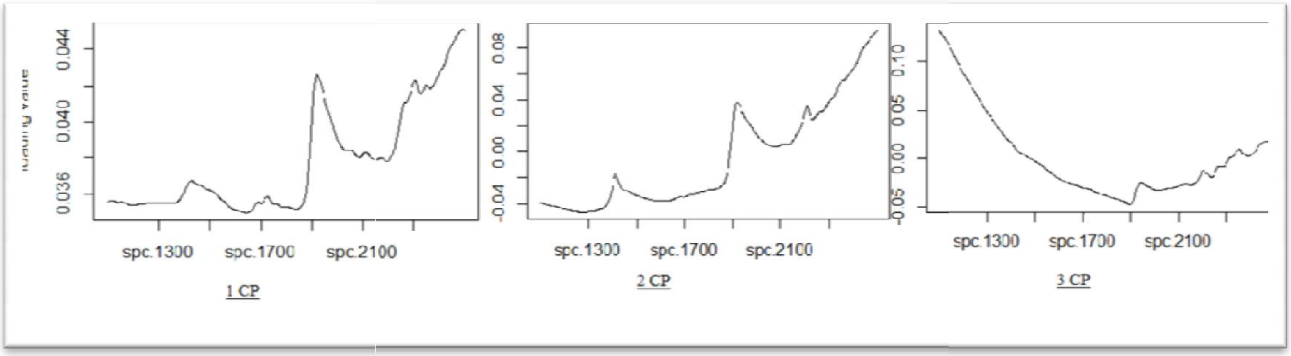


Figure 21:projection de spectre de laoding sur les trois premiers CP

on observe dans la figure 23 que dans les deux premiers CP on a des piques dans les zones correspondant à la bande spécifique d'azote totale, mais le chois de nombre de CP est liée avec la variation de R^2 et RMSEP en fonction de nombre de CP

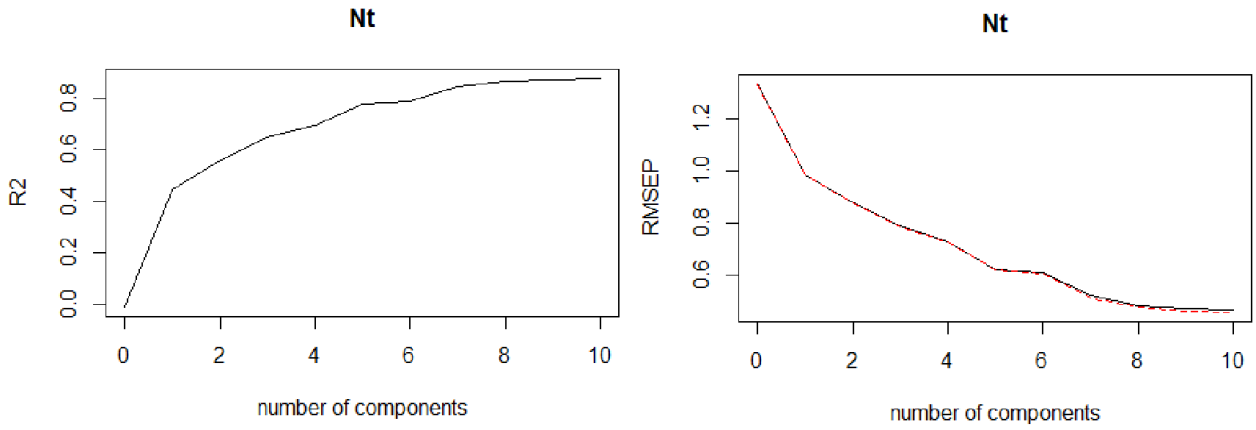


Figure 22 : variation de R^2 et RMSEP en fonction du nombre de CP pour Nt

On observe dans (la figure 25) qu'on a besoin de 9 CP pour construire le modèle PLS de bonne qualité, car la neuvième composante principale nous donne un R^2 maximale 87 ,69 % et un RMSEPc minimale 0,46

Validation du modèle

Pour la validation on utilise des données de validation, et on va calculer l'erreur quadratique de modèle et le RPD avec l'estimation de R^2

Le modèle qu'on a crée nous donne un RMSEP de validation égale 0,36 un erreur qui reste acceptable, un R^2 de validation 85,4% satisfassent, mais on a un RPD égale 1 se qui segnifie qu'on a un le modèle bien validé mais la qualité prédictive de ce modèle est faible

Est-ce qu'il y a des méthodes non factoriel ou des algorithmes qui nous permettre de faire une régression avec l'utilisation de tous les variables ?

Oui , il y a plusieurs méthodes qui génèrent un très grand nombre de variables (comme le RNN , SVM , RLP….)

3) La régression avec SVM pour Nt

Selon (Ahmed Chac 2014) , la méthode de régression SVM donne des meilleurs résultats pour l'azote totale et aussi cette algorithme capable de gérer un très grande nombre de variables

Donc on va créer un model SVM de type epsilon régression polynomiale, avec un degré de polynôme égale 3 et epsilonégale 0,01

Pour ce type de régression on va utiliser un polynôme de 3^{eme} de type epsilon avec epsilon=0,01
L'objectif de ce type de régression et de trouver un hyper plans capable d'explique la variation des points dans l'espace

On trouve que RMSEPc = 0,62 et erreur moyenne =0,47 et on a aussi R²= 65, par conséquent le modèle n'est pas valider, pour essayer d'améliorer ce type de modèle il faut effectue une transformation mathématique, selon (Ahmed Chac 2014) ils ont utilisé MSC, ils ont obtenu ($0.90 \leq R^2 \leq 0.93$ et $0.12 \leq$ RMSEP≤ 0.14)

Tableau récapitulative des résultats

Tableau 7 : Les résultats pour tous les types de régression

			calibration			validation	
élément chimique	*Prétraitement mathématique*	*model de régression*	*R²*	*RMSEP*	*RPD*	*R²*	*RMSEP*
la matière organique	sans	RLM	93,7	0,75	3,13	88,4	0,6
	sans	PCR	84,64	1	1,74	96	0,68
	sans	PLS	88,5	0,11	3,57	99,8	0,4
	SG	PLS	79,15	1,16	2,85	7,2	1,2
	SNV/ MSC	PLS	56,46	1,69	0,9	56	2
Azote totale	sans	PLS	85,4	0,36	1	85,4	0,36
	sans	SVM	65	0,62			

Parmi les types de régression que l'on a utilisés, on constate que la PLS donne de meilleurs résultats, il nous donne un coefficient de détermination supérieure 88.5pour la gamme de calibration et 99.8 pour la gamme de validation, on observe aussi que le prétraitement de données ne donne aucune amélioration, mais au contraire il baisse la qualité de régression , la RLM donne aussi une bonne qualité de régression avec R² égale 93.7, RMSEP égale 0.7 et RPD égale 3.13 , mais le problème de RLM, on a un doute sur la représentativité des variables choisies, par ce qu'on a éliminé plus de 550 variables .

Donc, notre étude montre que la régression PLS est le plus adéquat au ce type d'analyse, et la transformation mathématique ou le prétraitement mathématique n'est pas nécessaire dans ce cas même s'il y a un problème de diffusion de la lumière et la non-homogénéité des particules de sol

Conclusion

Les analyses effectuées sur le sol ont permis de démontrer l'efficacité de l'analyse multivariable quantitative. Couplée à une technique d'analyse spectroscopique (proche infrarouge), les techniques statistiques d'extraction de données peuvent être fort utiles pour la quantification d'un ou de plusieurs composés d'un mélange complexe comme le sol et ce, sans étapes de séparation ou de préparation chimique. Les résultats présentés dans le présent rapport démontrent la faisabilité de l'application de l'analyse multivariable à la détermination des concentrations des éléments fertilisent de sol.

Afin de couvrir les domaines de concentrations adéquats, deux ensembles de calibration ont été nécessaires. Une fois optimisée, les modèles de calibration obtenus avec le premier ensemble ont permis d'estimer les concentrations en MO et Nt avec moins de 2% d'erreur avec un coefficient de détermination supérieur a 95% . La performance des modèles utilisés pour la prédiction des concentrations en MO et Nt a été nettement améliorée avec un ensemble de calibration dont les domaines de concentrations sont plus larges.

Bibliographie

- Moslem Ladoni ,Hosein Ali Bahrami, Estimating soil organic carbon from soil reflectance: a review , 2009
- Eunyoung Choe et all , An alternate method for FTIR spectroscopic determination of soil Nitrate
- using derivative analysis and sample treatments ,2009
- K. Piikki ,J. Wetterlind , 3D digital soil mapping of agricultural fields by integration of multiple proximal sensor data obtained from different sensing methods, 2014
- L.Cécillon and J , -J.Brun , Near-infrared reflectance spectroscopy (NIRS) a practical tool for the assessment of soil carbon and nitrogen budge 2010
- Ahmed Chac ,Vis/NIR spectroscopic measurement of selected soil fertility parameters of Cuban agricultural Cambisols, 2014
- Miss Yogita Kulkarni , Krishna K. Warhad ,Primary nutrient determination in the cultivated Soil, 2014
- Martial Bernoux , Adrian Chappel , A global spectral library to characterize the world's soil , 2016
- Yi Peng, Xiong Xiong , Modeling soil organic carbon at regional scale by combining mmulti-spectral images with laboratory spectra
- J.B. Reeves,G. Mccarty , Mid- versus Near-infrared spectroscopy for on-site analysis of soil 2010
- K.A.Sudduth;N.Kitchen , Vis/NIR spectroscopy estimates of within-field variability in soil properties , 2010
- B.H.Kusumo;;M.Hedley , Predicting soil carbon and nitrogen concentrations and pasture root densities from proximally sensed soil spectral seflectancE , 2010
- J. B. Reeves III & J. S. Van Kessel , Near-Infrared Spectroscopic Determination of Carbon,Total Nitrogen, and Ammonium-N in Dairy Manures , 2000

Annexe A

Résultat de RLM pas a pas

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation	Variation de R-deux
1	,968[a]	,937	,919	,7504585339	,937

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	116,654	4	29,163	51,783	,000[b]
	Résidus	7,885	14	,563		
	Total	124,539	18			

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	-3,586	1,842		-1,947	,072
	VAR00022	-,689	,243	-4,213	-2,835	,013
	VAR00071	-,083	,119	-,553	-,698	,496
	VAR00125	,502	,256	2,217	1,964	,070
	VAR00339	,378	,377	1,829	1,001	,334

Annexe B

La valeur propre et la variabilité totale de l ACP pour MO

Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
Variance	76473.262	761.277	170.569	17.132	4.747	2.735
% of var.	98.755	0.983	0.220	0.022	0.006	0.004
Cumulative % of var.	98.755	99.738	99.958	99.980	99.986	99.990

La contrebutions et le cos ²Ɵ et la projection des scores sur les deux 1 er axes

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2
D3S0	279.435	-279.181	5.364	0.998	0.985	0.007	0.000
D3S3	278.387	277.894	5.315	0.996	-15.164	1.590	0.003
D3S6	264.431	-264.014	4.797	0.997	-3.469	0.083	0.000
D3S9	70.609	68.226	0.320	0.934	13.542	1.268	0.037
D3S12	164.916	-164.135	1.854	0.991	9.946	0.684	0.004
D3S15	227.395	-220.490	3.346	0.940	-50.179	17.408	0.049
D3S18	219.085	-218.397	3.283	0.994	-8.377	0.485	0.001
imane	386.331	379.621	9.918	0.966	70.117	33.990	0.033
biosol	188.717	187.436	2.418	0.986	19.245	2.561	0.010
S0	96.469	85.813	0.507	0.791	-43.573	13.126	0.204

La valeur propre et la variabilité totale de l ACP pour Nt

Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9	Dim.10	Dim.11	Dim.12	Dim.13
Variance	5.214	0.075	0.010	0.004	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
% of var.	98.254	1.422	0.195	0.071	0.035	0.012	0.004	0.002	0.002	0.001	0.001	0.001	0.000
Cumulative % of var.	98.254	99.676	99.871	99.942	99.977	99.989	99.993	99.995	99.996	99.998	99.998	99.999	