Towards a new optical system to characterize soils by Visible and Near Infrared Spectroscopy

Alexia Gobrecht

THÈSE

Pour obtenir le grade de

Docteur

Délivré par le

Centre International d'Etudes Supérieures en Sciences Agronomiques de Montpellier

Préparée au sein de l'école doctorale

Science des Procédés – Science des Aliments

Et de l'Unité Mixte de Recherche ITAP

Spécialité : Génie des Procédés

Vers une nouvelle approche optique pour la caractérisation des sols par spectrométrie visible et proche infrarouge

Présentée par

Alexia Gobrecht

Soutenue le 1er décembre 2014 devant le jury composé de

Mr Abdul MOUAZEN, Professeur - Université de Cranfield	Président du jury
Mme Ana GARRIDO-VARO, Professeur - Université de Cordoue	Rapportrice
Mr Alex McBRATNEY, Professeur - Université de Sydney	Rapporteur
Mme Véronique BELLON-MAUREL, ICPEF - Irstea	Directrice de thèse
Mr Jean-Michel ROGER, ICPEF - Irstea	Co-directeur de thèse
Mr Bernard BARTHES, IR - IRD	Invité
Mr Ryad BENDOULA, CR - Irstea	Invité

T H E S I S

In Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

of the International Center for Higher Education in Agricultural Sciences of Montpellier

Doctoral School : Sciences des Procédés Sciences des Aliments

---

Towards a new optical system to characterize soils by Visible and Near Infrared Spectroscopy

---

defended on December 1st 2014 by

Alexia Gobrecht

JURY

Reviewers :	Pr. Ana GARRIDO-VARO,	University of Cordoba (SP)
	Pr. Alex McBRATNEY,	University of Sydney (AUS)
President of the jury :	Pr. Abdul MOUAZEN,	Cranfield University (UK)
Thesis Director :	Pr. Véronique BELLON-MAUREL,	Irstea Montpellier (F)
Thesis Co-Director :	Dr. Jean-Michel ROGER,	Irstea Montpellier (F)
Invited Members :	Dr. Bernard BARTHES,	IRD Montpellier (F)
	Dr. Ryad BENDOULA,	Irstea Montpellier (F)

Irstea UMR ITAP - COMiC - Capteurs Optiques pour les Milieux Complexes - Montpellier, France

An Herrn Prof. Dr.-Ing. Heinrich "PIPA" Gobrecht (1909 - 2002)

A Thomas, Louise et Nils

Remerciements

Quelle belle aventure que cette thèse !

Et c'est avec la satisfaction (et une pointe de fierté) du travail accompli que je prends enfin le temps de coucher sur cette page, les traditionnels mots de remerciements qui préambulent ce mansucrit ...

A l'Institut Irstea tout d'abord:

Je suis très sensible à la chance que j'ai d'avoir pu, dans le cadre de mes fonctions d'Ingénieur de l'Agriculture et de l'Environnement d'Irstea, faire évoluer mon métier vers celui d'Ingénieur-Chercheur. Pour cela, je remerice l'Institut qui rend possible la formation doctorale de ses agents.

A mon "International" jury :

Professor Ana Garrido - Varo et Professor Alex McBratney, quel honneur d'avoir pu vous confier mon manuscrit pour son évaluation. Merci beaucoup d'avoir pris le temps de le lire, et surtout le temps de parcourir ces nombreux kilomètres pour venir assister à ma soutenance... Muchas gracias Ana ! Thanks a lot Alex ! Et merci également au Professeur Abdul M. Mouazen, qui a accepté de présider ce jury.

A mon Comité de thèse élargi:

Dr. Sebastien Preys, le chimiométricien, que j'ai un peu perdu en route car j'ai fait le choix de l'optique. Mais merci beaucoup d'être venu assister à mes comités de thèse.

Mister Dr. Bernard Barthès, THE soil scientist. Merci tout d'abord pour ton soutien, tout au long de cette thèse, merci aussi d'avoir assurer un certain contre-poids face à tous ces physiciens !!! J'espère vivement que d'autres collaborations suivront entre ITAP et Eco&Sols !!!

A ma garde rapprochée...

Ryad-le-chercheur, que dire, cette aventure, nous l'avons vécue ensemble, elle nous a fait grandir tous les deux et surtout elle nous ouvre des portes pour la suite : y'a plus qu'à !!! et j'ai bien hâte ... Je n'oublie pas non plus Ryad-le-coach: merci pour tout !!! tu as assuré !

A Jean-Mi, merci de remplir le cahier des charges du chercheur sénior ideal : Exigeant, Optimiste, Pédagogue, Disponible, Attentif ... et merci pour ton amitié aussi.

A Véronique, enfin, et peut-être j'ose rajouter, surtout ... Je profite de ce petit paragraphe de liberté d'expression pour te remercier tout d'abord, de m'avoir fait confiance, en 2005 lorsque j'ai candidaté au poste d'Ingénieur à l'UMR ITAP et ensuite de m'avoir offert cette belle opportunité de réaliser ma thèse en spectrométrie. Enfin, merci d'avoir dirigé mes travaux, jusqu'au bout, malgré un emploi du temps plus que contraint, je t'en suis sincèrement reconnaissante.

Merci aussi à tous les membres de l'équipe COMiC: Gilles, Christophe, Arnaud, Daniel, et surtout Nathalie, qui m'a montré qu'il etait possible de mener de front une thèse avec de petits

enfants !!! merci pour votre soutien et vos conseils ! Je n'oublie pas les jeunes, qui viennent et qui malheureusement repartent, tout en laissant une empreinte dans notre équipe : Xavier, Sylvain, Faten, Ana, Sarah, Benoit, Sylvia ... je vous souhaite de belles choses dans votre vie de "grand" !

Merci surtout, Michèle, pour tout ce que tu fais qui nous rends la vie plus simple, c'est tellement précieux !

Mes remerciements s'adressent également à tous les membres de l'UMR ITAP, aux collègues des autres instituts de la place de Montpellier venus me soutenir le jour J.

Bien entendu, je n'oublie pas les copines, Virginie qui m'a tenue la main, Emmanuelle, qui m'a prodigué les derniers conseils de respiration et toutes celles et ceux qui ont eu une pensée pour moi ce jour là.

Je remercie sincèrement Jeanne, Jean et Marc d'avoir fait cet aller-retour pour venir partager cette belle journée avec moi et Astrid pour ton soutien par la pensée, tu nous as manquée.

Maman, Papa, merci pour TOUT (i.e la logistique) TOUT (i.e le soutien moral) TOUT (i.e. le champagne) !!! Je vous aime (meme si Papa a posé une question à la fin de la soutenance !!!). Ca y'est, la lignée des Dr. Gobrecht se poursuit, OUF, l'honneur est sauf!

A Isabel, ma très chère grande soeur, merci d'être ce que tu es ...

Merci aussi à Cyril, Majdouline, Xavier et les "COUSINS", je vous expliquerai plus tard ce que je fais !!!

Louise et Nils, avec vous, le sprint final était peut-être un tout petit peu plus difficile, mais en même temps, vous m'avez apporté l'oxygène nécessaire pour franchir la ligne d'arrivée ! Quel bonheur de vous avoir, je suis tellement fière de vous !

Tomy, merci, tout simplement. C'est à ton tour maintenant, si je compte bien, ça fera Bac+21 !!! *hi hi hi* ...

Publications and communications

Papers in international peer-reviewed journals

Following articles directly result from this thesis and are referenced in the manuscript as :

Art. I Gobrecht, A., Roger, J. M., Bellon-Maurel, V. (2014). **Major issues of diffuse reflectance NIR spectroscopy in the specific context of soil carbon content estimation: A review**. Advances in Agronomy Vol. 123, 123, 145-175. DOI: 10.1016/B978-0-12-420225-2.00004-2

Art. II Bendoula, R., Gobrecht, A., Moulin, B., Roger, J. M., Bellon-Maurel, V. (2014). **Improvement of the chemical content prediction of model powder system by reducing multiple scattering using polarized light spectroscopy**. Applied Spectroscopy, Accepted.

Art. III Gobrecht, A., Bendoula, R., Roger, J.M., Bellon-Maurel, V. (2015). **Combining linear polarization spectroscopy and the Representative Layer Theory to measure Beer-Lambert's Law absorbance of highly scattering media**. Analytica Chimica Acta. Vol. 853, 486-494. DOI:10.1016/j.aca.2014.10.014

Art. IV Gobrecht, A., Bendoula, R., Roger, J.M., Bellon-Maurel, V. (2014). **Improvement of soil carbon content prediction by reducing multiscattering using polarized light spectroscopy**. Soil and Tillage Research. Submitted in October 2014.

Other article :

- Minasny, B., McBratney, A. B., Bellon-Maurel, V., Roger, J. M., Gobrecht, A., Ferrand, L., Joalland, S. (2011). **Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon**. Geoderma, 167, 118-124. DOI: 10.1016/j.geoderma.2011.09.008

Oral communications

1. McBratney, A. B., Minasny, B., Bellon-Maurel, V., Gobrecht, A., Roger, J. M., Ferrand, L., Joalland, S. (2011, May). **Removing the effect of soil moisture from NIR diffuse reflectance spectra for prediction of soil carbon**. In The 2nd Global Workshop on Proximal Soil Sensing, Montreal, Canada.
2. Gobrecht, A., Bellon-Maurel, V., (2013, April). **Near Infrared Spectroscopy, Application in Soil Science**. Soil Spectral Inference Workshop, University of Sydney. Australia.
3. Gobrecht, A., Bendoula, R., Roger, J. M., Bellon-Maurel, V. (2014, May). **A new optical method coupling light polarization and Vis-NIR spectroscopy to improve the measured absorbance signal's quality of soil samples**. In EGU General Assembly Conference Abstracts (Vol. 16, p. 5657).

This thesis is part of the research project INCA (In-field Spectroscopy for Carbon Accounting), financially supported by ADEME (Agency for the Environment and Energy Management).

Resumé en français

Une nouvelle approche optique pour améliorer la caractérisation des sols par spectrométrie visible et proche infrarouge

---

Contexte de la thèse

L'un des défis majeurs de ce XXI^{ème} siècle est le changement climatique et ses conséquences sociales, économiques et environnementales. L'attention portée au réchauffement global et à l'augmentation des concentrations en gaz à effet de serre (GES) dans l'atmosphère, principalement le dioxyde de carbone (CO_2), le méthane (CH_4), et l'oxyde nitreux (N_2O), a conduit à s'interroger sur le rôle des sols en tant que source ou puits de carbone (C). Les sols seuls constituent le plus grand réservoir de carbone organique de l'écosystème terrestre, approximativement trois fois le stock de la biomasse continentale et deux fois celui de l'atmosphère.

Le stock de carbone du sol étant fortement dépendant du mode d'usage des terres ou des pratiques culturales, une modification de ceux-ci peut conduire à des changements importants des stocks des horizons de surface (entre 0 et 30 cm de profondeur), dans le sens d'une diminution ou d'une augmentation. La question de la comptabilisation des stocks de carbone dans les sols agricoles et forestiers fait l'objet de nombreuses discussions, à la fois dans le cadre des négociations internationales sur le climat sous l'égide des Nations-Unies, mais aussi dans le cadre des marchés volontaires, en plein essor.

Dans ce contexte, il devient nécessaire de pouvoir comptabiliser précisément les stocks de carbone et leur évolution dans le temps. Les méthodes actuelles, basées sur des campagnes d'échantillonnage associées à des méthodes analytiques de laboratoires longues et couteuses, constituent un frein pour le développement de ces actions en faveur de la séquestration de carbone dans les sols.

La spectroscopie proche-infrarouge (SPIR), technique connue depuis plus de 40 ans pour mesurer la qualité et la composition des produits agricoles et alimentaires, présente un potentiel indéniable pour remplacer les campagnes de mesure couteuses. Cependant, alors qu'elle est depuis plusieurs décennies utilisée en routine dans l'industrie laitière ou céréalière, ou en ligne - en agro-alimentaire et plus récemment pour le tri des déchets-, elle reste, en ce qui concerne le sol, encore du domaine de la recherche. Si la quantification de différents constituants ou certaines fonctions (teneur pondérale en carbone organique et inorganique, en azote, capacité d'échange cationique, granulométrie) a fait l'objet de nombreuses publications, plusieurs verrous méthodologiques et technologiques doivent être levés pour en faire une méthode d'analyse de routine pour la comptabilité des crédits C.

Principes et limites de la SPIR appliquée aux sols

La loi de Beer-Lambert constitue le cadre théorique qui régit les principes analytiques de la spectroscopie proche-infrarouge. Elle établit le lien linéaire entre l'absorbance de la lumière et la concentration c d'un élément chimique constituant le milieu analysé, son coefficient d'extinction $\varepsilon(\lambda)$ et le trajet l parcouru par la lumière dans le milieu:

$$A(\lambda) = -\log \frac{I_T(\lambda)}{I_0(\lambda)} = \varepsilon(\lambda) \cdot c \cdot l$$

Cependant, cette loi ne s'applique que dans le cas de milieux translucides faiblement concentrés (donc peu absorbants). Dans le cas des sols, qui sont des milieux particulaires hétérogènes, l'interaction de la lumière avec la matière est beaucoup plus complexe. La

lumière n'est plus simplement transmise ou absorbée mais elle est également diffusée dès qu'elle rencontre une particule et que l'indice de réfraction change. Le chemin optique de la lumière est fortement dévié et rallongé. Cela impacte directement la qualité du signal d'absorbance qui n'est plus linéairement reliée à la concentration de la variable d'intérêt du fait d'effets additifs et multiplicatifs se superpose au signal (cf. figure 1).

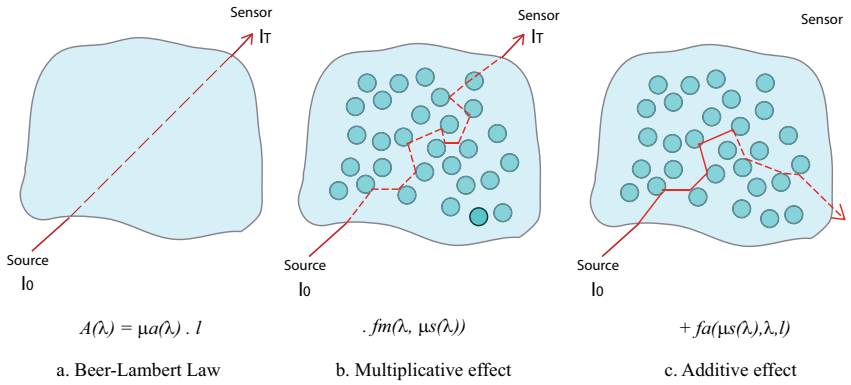

Figure 1: Représentation des effets additifs et multiplicatifs de la diffusion sur le signal d'absorbance. μ_a est le coefficient d'absorbance et μ_s est le coefficient de diffusion. λ est la longueur d'onde.

L'analyse multivariée en spectroscopie proche infrarouge consiste à trouver un modèle capable de relier les spectres d'absorbance à une variable d'intérêt, la concentration par exemple. Les modèles sont principalement construits à partir de méthodes d'analyse multivariées linéaires, du fait de la loi de Beer-Lambert. La méthode la plus couramment utilisée étant la régression PLS.

Dans le cas des sols, et plus généralement des milieux très diffusants, les modèles chimiométriques construits à partir de spectres d'absorbance dont la linéarité avec la concentration est remise en cause du fait de la diffusion, ne sont pas toujours de qualité optimale, ni robustes.

Des prétraitements mathématiques sont généralement appliqués sur les spectres pour limiter l'impact de la diffusion et rétablir, dans une certaine mesure, cette linéarité. Mais ces prétraitements ne suffisent pas toujours.

Objectifs de la thèse

Dans cette thèse, nous proposons une démarche alternative aux prétraitements mathématiques en nous focalisant sur la première étape de la méthode analytique par spectroscopie proche infrarouge: la formation du signal.

L'objectif est de mesurer un signal d'absorbance de qualité optimale, c'est à dire, le moins impacté possible par les phénomènes de diffusion de la lumière. L'hypothèse que nous posons est que la qualité du modèle de prédiction du carbone du sol est fortement liée à la qualité du signal d'absorbance à partir duquel il est construit.

Ainsi, nous avons apporté des réponses originales aux questions scientifiques suivantes:

1. Comment réduire l'effet de la diffusion sur le signal spectroscopique ?

2. Comment, à partir de ces signaux, modéliser l'absorbance chimique du milieu?

PoLiS, une méthode optique pour réduire l'impact de la diffusion sur le signal spectroscopique

Principes théoriques de la correction par polarisation

Le dispositif de mesure optique développé ici, et dénommé PoLiS, utilise les propriétés ondulatoires et les principes de polarisation de la lumière pour sélectionner la part du signal qui aura été moins diffusée par le milieu. Lorsqu'un flux lumineux incident, linéairement polarisé, interagit avec le milieu, il perd progressivement, mais assez rapidement, son état de polarisation initial. Ainsi, au moyen d'un analyseur placé devant le détecteur, il est possible de mesurer les deux composantes de ce flux : celle qui a conservé son état de polarisation initial, $I_{\parallel}(\lambda)$ et celle qui l'a perdue $I_{\perp,\Omega}(\lambda)$ (cf. figure 2).

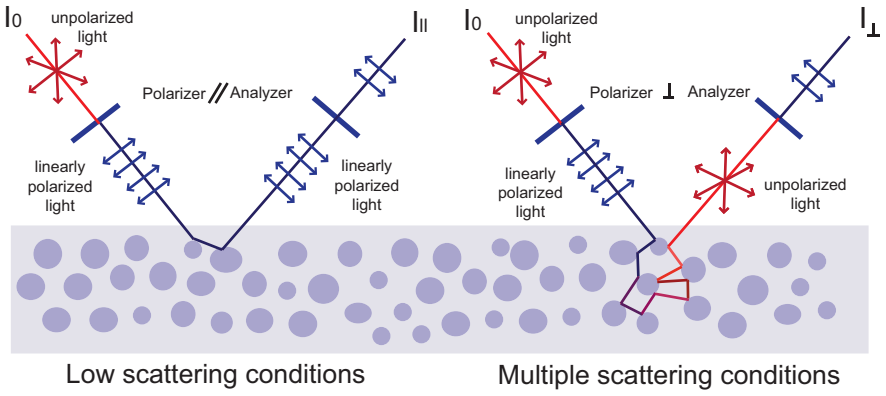

Figure 2: Principe de la mesure des deux composantes $I_{\parallel}(\lambda)$ et $I_{\perp}(\lambda)$ de la lumière réémise par le milieu au moyen d'un polariseur et d'un analyseur

Ce principe de mesure nous a permis de calculer la réflectance totale réémise par le milieu en faisant la somme de composantes parallèle et perpendiculaire de la lumière:

$$R_{BS}(\lambda) = R_{\parallel}(\lambda) + R_{\perp}(\lambda)$$

En faisant la différence de ces deux composantes, nous avons mesuré une réflectance corrigée des effets de la diffusion:

$$R_{SS}(\lambda) = R_{\parallel}(\lambda) - R_{\perp}(\lambda)$$

Principes théoriques de la modélisation de l'absorbance

Les deux types de signaux mesurés par le dispositif optique PoLiS ont été implémentés dans la fonction d'absorption et de rémission $A(R,T)$ proposée par Dahm et Dahm dans leur cadre théorique de la couche représentative (Representative Layer Theory).

$$A(R,T) = \frac{(1-R)^2 - T^2}{R} = \frac{a}{r} \cdot (2 - a - 2r)$$

Cette fonction relie la réflectance R et la transmittance T mesurées sur un échantillon, à la fraction absorbée (a) et réémise (r) d'une couche hypothétique de faible épaisseur

mais représentative de l'échantillon. Dahm et Dahm stipulent que l'absorbance calculée à partir de a, la fraction de lumière absorbée par cette couche représentative, est une bonne approximation de la vraie absorbance (selon la loi de Beer-Lambert) :

$$A = -log(1 - a)$$

Nous nous sommes placés dans ce cadre théorique pour résoudre la fonction A(R,T) en posant les hypothèses suivantes :

- La réflectance R totale de l'échantillon peut être approximée par $R_{BS}(\lambda)$, la réflectance totale mesurée avec le dispositif PoLiS;
- La fraction réémise (r) par la couche représentative théorique peut être approximée par $R_{SS}(\lambda)$, la part du signal n'ayant subi que peu de diffusion par le milieu étudié.

La résolution de cette équation nous a permis de proposer une expression de l'absorbance de milieux diffusants, fonction des mesures permises par le dispositif PoLiS, $R_{BS}(\lambda)$ et $R_{SS}(\lambda)$:

$$Abs_{Po}(\lambda) = -\ log\ \left(R_{SS}(\lambda) + \sqrt{(1 - R_{SS}(\lambda))^2 - \frac{R_{SS}(\lambda)}{R_{BS}(\lambda)}\,(1 - R_{BS}(\lambda))^2} \right)$$

Cette absorbance, obtenue par la méthode de mesure PoLiS est, en théorie, moins impactée par la diffusion et plus linéairement reliée à la concentration.

Matériel et méthodes

Instrumentation

Le dispositif PoLiS était constitué d'une source lumineuse, d'un polariseur linéaire, d'un analyseur linéaire et d'un spectromètre opérant dans la gamme spectrale 350 - 800 nm, soit le visible - très proche infrarouge (Figure 3). Des lentilles permettaient la collimation de la lumière et la collection du signal réémis.

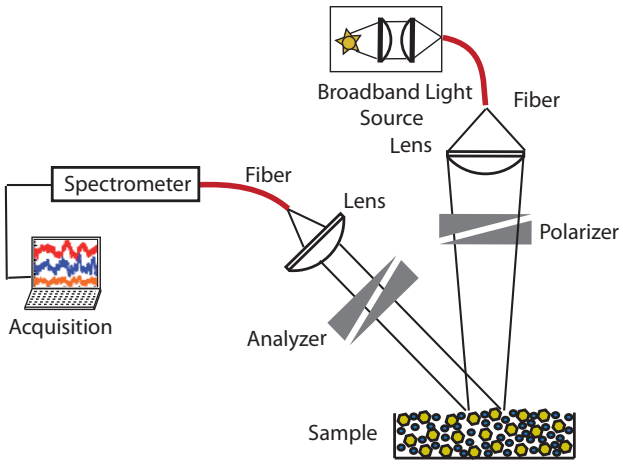

Figure 3 : schéma du dispositif optique PoLiS.

Échantillons

Trois types d'échantillons ont été mesurés par la méthode:

- Des échantillons liquides, mélangeant du lait, dont les micelles et particules de gras jouent le rôle de diffuseur, avec du colorant alimentaire, E141, l'absorbant dont on connait la concentration;
- Des échantillons poudreux, mélangeant du sable de Fontainebleau (diffuseur) avec le même colorant E141 en poudre à différentes concentrations;
- 52 échantillons de sols, provenant de la région du Vercors dont la variable d'intérêt est le carbone organique total. Chaque échantillon a été préparé selon trois tailles de particules différentes: grossiers (agrégats <5mm), tamisés à 2 mm et broyés à 0.2 mm.

Analyse multivariée

Sur les échantillons de sol, des modèles PLS de prédiction de la teneur en carbone organique ont été construits à partir du spectre de réflectance totale $R_{BS}(\lambda)$, du spectre d'Absorbance classique $\{-logR_{BS}(\lambda)\}$ et à partir des spectres d'absorbance obtenus avec la méthode PoLiS $Abs_{Po}(\lambda)$. Pour évaluer la plus-value de la méthode PoliS par rapport aux prétraitements mathématiques, les spectres ont été prétraités par SNV (Standard Normal Variate), MSC (Multiplicative Scatter Correction et OPLECm (Modified Optical Pathlength estimation and correction), qui sont classiquement appliqués pour réduire l'impact de la diffusion sur les signaux spectroscopiques.

Résultats

Diminution de l'effet de la diffusion sur les spectres

L'analyse des spectres d'absorbance obtenus avec la méthode PoLiS a montré, dans un premier temps, que la réduction de l'effet de la diffusion sur le signal se traduisait par signatures spectrales plus marquées. La figure 4 illustre ce résultat sur les poudres mélangeant du sable et un colorant.

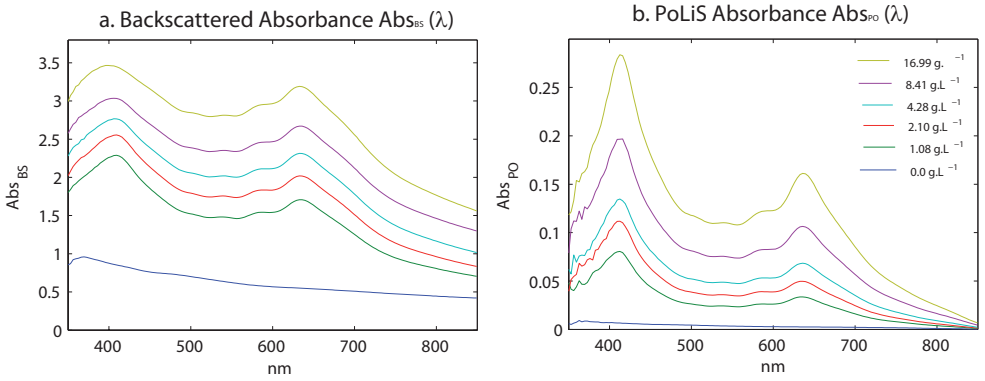

Figure 4: Absorbance totale $Abs_{BS}(\lambda)$ (a.) et absorbance PoLiS $Abs_{Po}(\lambda)$ (b.) mesurées sur des échantillons pulvérulents mixant du sable et du colorant E141 à différentes concentrations en $g.L^{-1}$.

Nous avons pu observer que la ligne de base était réduite et que les pics d'absorption étaient beaucoup plus fins et marqués.

Amélioration de la linéarité entre l'absorbance et la concentration

Les spectres d'absorbance mesurés à partir de la méthode PoLiS présentent une meilleure linéarité, à une longueur d'onde λ donnée, avec la concentration de l'absorbant. La figure 5 montre que cela est vrai quelque soit le milieu: liquide, poudreux et même pour les sols.

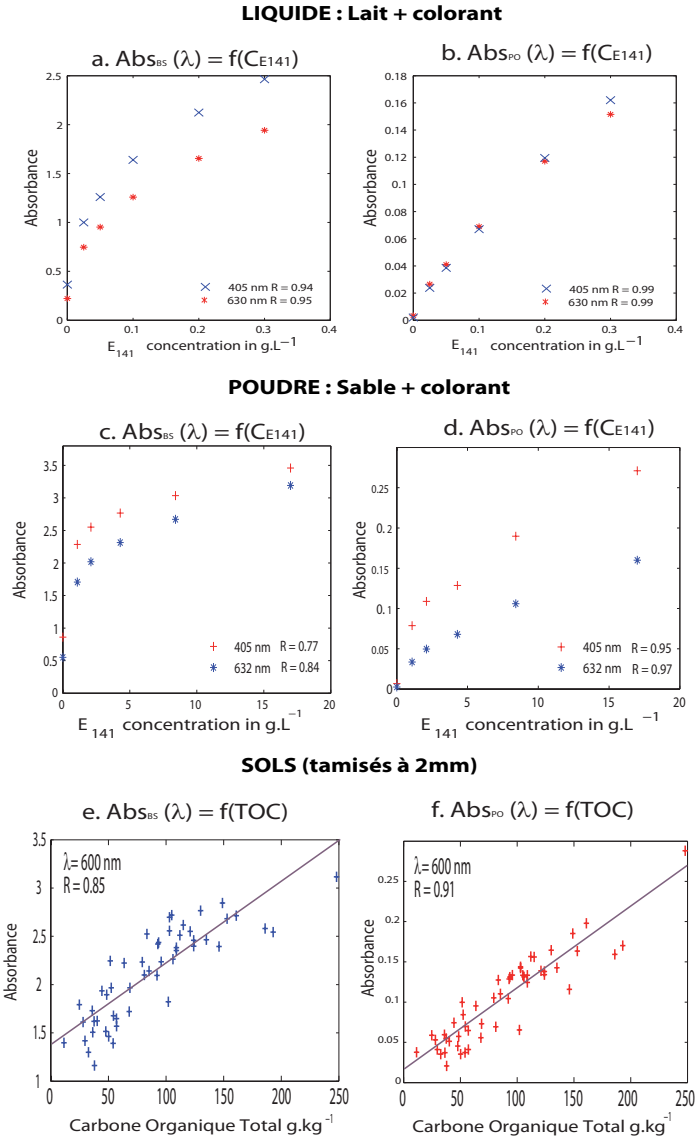

Figure 5: Absorbance totale $Abs_{BS}(\lambda)$ et absorbance PoLiS $Abs_{Po}(\lambda)$ mesurée à la longueur d'onde λ en fonction de la concentration de l'absorbant. Sur le milieu liquide (a. et b.) et sur les poudres (c. et d.), à $\lambda = 405nm$ et $\lambda = 630nm$ en fonction de la concentration en colorant E141 en $g.L^{-1}$ et sur les sols (e. et f.), à $\lambda = 600nm$, en fonction de la teneur en carbone organique en $g.kg^{-1}$. R est le coefficient de Pearson.

Les propriétés des spectres d'absorbance modélisés par la méthode PoLiS se rapprochent de celles de l'absorbance de la loi de Beer-Lambert: la quantité de lumière absorbée par le milieu est linéairement proportionnel à la concentration.

Amélioration des prédictions

Les modèles construits à partir des spectres d'absorbance PoLiS $Abs_{Po}(\lambda)$ se sont avérés être toujours de meilleure qualité que ceux construits à partir de la réflectance $R_{BS}(\lambda)$ ou de l'absorbance totale $Abs_{BS}(\lambda)$, même lorsque ces derniers ont été prétraités. Nous avons pu observer également que les prétraitements n'avaient aucun effet positif sur l'absorbance PoLiS. La figure 6 présente les indicateurs de qualité de tous les modèles construits dans cette étude, le coefficient de détermination R^2 et l'erreur standard de cross-validation (SECV). Les modèles ont été construits sur les trois types préparation des sols: grossiers, tamisés et broyés.

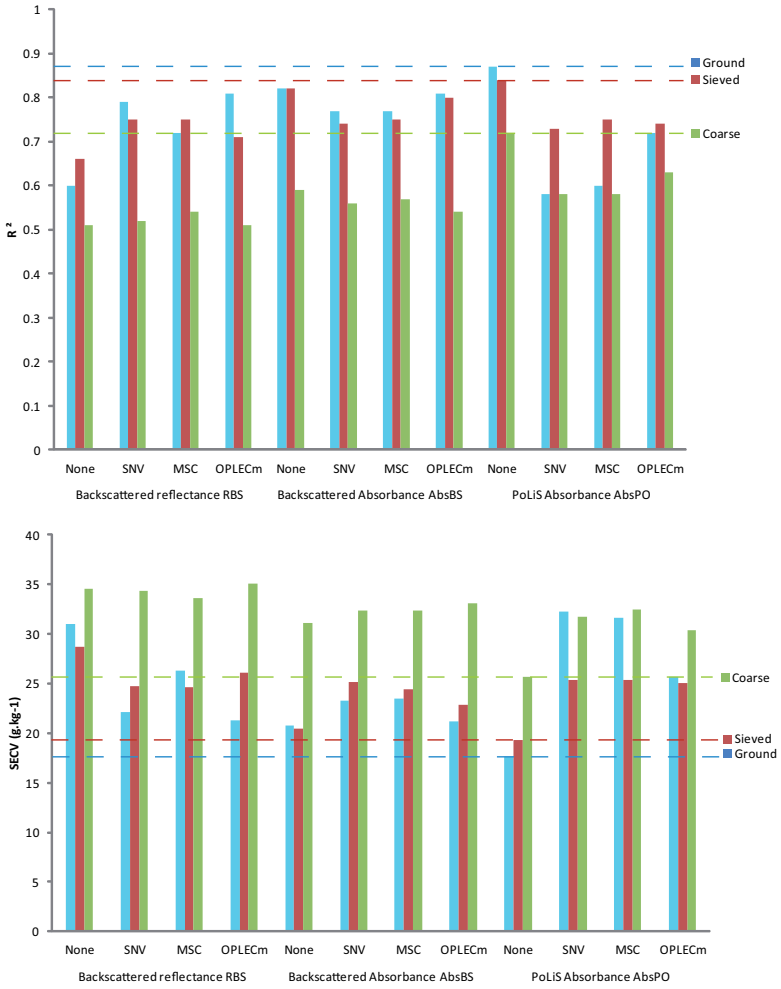

Figure 6: Qualité des prédictions réflectance totale, absorbance totale et absorbance PoLiS, Prétraitements none: aucun, SNV: standard Normal Variate, MSC: multiple scatter correction, OPLEC: optical pathlength estimation and correction, R2 cofficient de détermination, SECV erreur standard de cross validation

Ces résultats nous ont permis de confirmer que (i) l'absorbance PoLiS était de meilleure qualité et présentait un meilleur potentiel pour l'étalonnage, (ii) que ce signal atteignait une qualité optimale qui ne peut être améliorée par des prétraitements mathématiques et (iii) que ce signal est moins impacté par la taille des particules.

Conclusions

Les travaux menés dans cette thèse ont montré qu'il est possible, en mettant en oeuvre une architecture optique dédiée, d'améliorer significativement la qualité des signatures spectrales mesurées sur des milieux diffusants tels quel les sols. Il a été également confirmé que l'amélioration de la qualité des spectres, notamment par le biais d'une meilleure linéarité avec la concentration, avait un impact positif sur la qualité des modèles chimiométriques. Le potentiel de cette méthode est important, tant pour améliorer la caractérisation des sols par spectroscopie proche-infrarouge, mais pour permettre une meilleure compréhension des interactions lumière-matière se produisant dans les milieux hautement diffusants.

Contents

List of Figures

List of Tables

List of symbols and abbreviations

λ	Wavelength $[nm]$
Λ_w	Wilk's lambda criterion [–]
$\mathbf{X}$	Capital bold characters used for matrices
μ_a	Absorption coefficient $\left[cm^{-1}\right]$
μ_s	Scattering coefficient $\left[cm^{-1}\right]$
Ω	Solid collection angle of the optical setup $[sr]$
ε_{141}	E141 extinction coefficient $\left[(L \cdot g^{-1} \cdot mm^{-1}\right]$
a,r,t	Absorbed, remitted and transmitted fraction of light by the representative layer [–]
$A(R,T)$	Absorption-Remission Function [–]
Abs	Absorbance [–]
Abs_{BS}	Backscattered Absorbance computed from the backscattered reflectance [–]
Abs_{Po}	Absorbance computed with the PoLiS method [–]
Abs_{RL}	Absorbance of the representative layer [–]
c	Chemical concentration $\left[g.L^{-1}\right]$
dx	pathlength $[cm]$
I_0	Intensity of the incident light $[Watt]$
$I_\perp$	Intensity of the perpendicular component of polarized light $[Watt]$
$I_\parallel$	Intensity of the parallel component of polarized light $[Watt]$
I_{BS}	Intensity of the totally backscattered light $[Watt]$
I_{MS}	Intensity multiple scattered light $[Watt]$
I_{SS}	Intensity of low scattered light $[Watt]$
R	Reflectance [–]
R^2	Coefficient of determination [–]
R_{BS}	Backscattered reflectance [–]
R_{SS}	Low scattered reflectance [–]

T	Transmittance [–]
x	Non bold characters for column vectors
ANN	Artificial Neural Network
BLUE	Best Linear Unbiaised Estimators
BPNN	Backpropagation Neural Network
BRT	Boosted Regression Trees
CART	Classification and Regression Trees
E141	Coloring dye of copper complexes of chlorophyll
EPO	External Parameter Orthogonalization
ERT	Equation of Radiative Transfer
FOM	Figures of Merit
InGaAs	Indium Gallium Arsenide
K-NN	K-nearest Neighbor
KM	Kubelka-Munk
LS-SVM	Least Squares Support Vector Machine
LWR	Locally Weighted Regression
MARS	Multivariate Adaptative Regression Splines
MIR	Mid Infrared
MLP	Multiple layer perceptron
MLR	Multiple Linear Regression
MSC	Multiplicative Signal Correction
N.A	Numerical Aperture [–]
NAS	Net Analyte Signal
NIR	Near Infrared
OC	Organic Carbon
OLS	Ordinary Least Squares
OPLEC	Optical Path Length Estimation and Correction
OSC	Orthogonal Signal Correction
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS	Partial Least Squares Regression
PLS-NN	Partial Least Squares Neural Network

RBFN	Radial Basis Function Network
RF	Random Forest
RL	Representative Layer
RLT	Representative Layer Theory
RMSE	Root Mean Square Error
RPD	Ratio of Performance to Deviation
SECV	Standard Error of Cross Validation $\left[g.kg^{-1}\right]$
SEP	Standard Error of Prediction $\left[g.kg^{-1}\right]$
SEP	Standard Error of Prediction
SG	Savistzky-Golay
SNR	Signal to Noise Ratio
SNV	Standard Normal Variate
SOC	Soil Organic Carbon
SOM	Soil Organic Matter
SVMR	Support Vector Machine Regression
SWIR	Short Wave Infrared
TOC	Total Organic Carbon
Vis–NIRS	Visible and Near Infrared Spectroscopy

Chapter 1

General Introduction

Contents

1.1 Overview and objectives

Concerns about global warming and increasing concentrations of atmospheric greenhouse gas (CO_2, CH_4 and N_2O) have led to questions on the role of soils as a source or sink of carbon (C). Soil is the largest surface carbon pool, almost three times the quantity stored in the terrestrial biomass and twice the amount stored in the atmosphere (Eswaran *et al.*, 2000; Bernoux *et al.*, 2006). Carbon sequestration in soil is a real win-win strategy: it restores degraded soils, increases the production of biomass, purifies surface and ground waters, and reduces the rate of enrichment of atmospheric CO_2 by offsetting emissions due to fossil fuels (Lal, 2004b). From an economic perspective, as carbon sequestration in soil has become relevant to reduce the amount of greenhouse gas emissions, policymakers have made carbon trading markets emerge (Lal, 2004a; Gehl, 2007). It is therefore of utmost importance to assess soil carbon stocks and fluxes in terrestrial systems to understand the global dynamics of carbon.

Collecting soil at sufficiently high spatial and temporal resolution to meet soil C verification needs, and analyzing them using traditional laboratory-based methods, may be prohibitively expensive (Smith, 2004). Thus, new methods are required to rapidly and accurately measure soil C at field- and landscape-scales to improve field, regional and global soil C stock and flux estimates (Gehl, 2007).

Visible and Near Infrared Spectroscopy (Vis–NIRS) has become an extremely important analytical technique over the past 50 years as evidenced by the high number of different applications and products analyzed by NIRS (Williams & Norris, 2001). And, a little later than for agricultural and food products, NIRS has been naturally considered as a potential and credible substitute for the traditional analytical methods used for soil properties assessment (Reeves III, 2009; Stenberg *et al.*, 2010; Bellon-Maurel & McBratney, 2011). A new community of research targeting NIR as a rapid tool for soil analysis, either in the laboratory or in the field, has emerged among the soil science community. NIR-scientists have progressively joined the soil community because studying a complex and heterogeneous material like soils entails new NIR-related research opportunities, in

fields like instrumentation, light–matter interactions, chemometrics or sampling strategies (Bellon-Maurel, 2009). Bridging the two communities (Soil and NIR) is probably the best strategy to reach the *Grail* : a portable low-cost NIR sensor providing precise information about (many) soil properties.

Contributing to this common endeavor is the core objective of this thesis. Thus, our approach is based on the following sub-objectives:

1. Analyzing the challenges Vis-NIR spectroscopy applied to soils faces in order to identify the key factors influencing the quality of the measurement and the possible paths of improvement;

2. Designing an original optical setup dedicated to measure a spectroscopic signal of optimal quality and to model the chemical absorbance of scattering materials;

3. Testing the feasibility and assessing the added value of the proposed method to predict total organic carbon content of soils.

1.2 Outline of the thesis

Each sub-objective is addressed by a chapter of the thesis referring to a scientific publication (**Art I - IV**, listed p vii), forming the spine of this manuscript.

Chapter 2 reviews the major issues that NIR applied to soil is facing and offers a panorama of the mathematical solutions implemented by soil scientists. Light scattering is the main source of problems as it impacts directly the quality of the signal and thus, the reliability of the calibration model. Hence, developing new optical methods to increase the signal quality is a new research path that has to be invested. This chapter is the reproduction of **Art. I** published in the book series *Advances in Agronomy* in 2014. Following the conclusions drawn by the review paper, we present the scientific issues addressed in this thesis.

Chapter 3 and **chapter 4** develop an original approach to circumvent the issue of light scattering impacting a spectroscopic signal to model the absorbance of highly scattering materials.

First, **chapter 3** focuses on the design of an optical setup, based on light polarization spectroscopy. The output of the method is a reflectance signal freed from multiscattering. Reproducing **Art. II**, published in Applied Spectroscopy in 2014, the chapter presents the theory which underlies the method and which is then experimentally validated on simple model media.

Next, in **chapter 4**, the theoretical framework of the Representative Layer Theory is used to propose a model of the chemical absorbance of the sample. Again, theory and experimental validation of the approach are presented trough the reproduction of **Art. III**, published in Analytica Chimica Acta in 2014.

To close the loop, in **chapter 5**, we test the feasibility of the method, hereafter named PoLiS, to predict Total Organic Carbon content of soils. The performances of the method are compared to the classical calibration strategy, based on mathematically preprocessed spectra. **Chapter 5** is the reproduction of **Art. IV** submitted in Soil and Tillage Research in October 2014.

In **chapter 6**, which is ending this manuscript, the main contributions of this work are discussed and put into perspective with future research and development ideas that this thesis has brought to light.

Chapter 2

Major issues of NIR spectroscopy in Soil Science

Contents

Preamble

In this chapter, with the objective of being both pedagogical and practical, we review and discuss why the basic theoretical concepts underpinning NIR spectroscopy and linear chemometric modeling may be questioned in the specific context of soil: (i) light scattering due to soil particles causes departure in the assumed linear relationship between the spectrum and the carbon content and (ii) the other classical linear regression assumptions (constant residual variance, normal error distribution ...) are also put into question.

With reference to these specific issues, the different chemometric methods presented as possible solutions to perform better calibration models are discussed. We focus on classical linear methods associated with various preprocessing, local methods and finally non linear methods.

Based on the concluding remarks of this chapter, the scientific issues addressed in this research are presented.

Major Issues of Diffuse Reflectance NIR Spectroscopy in the Specific Context of Soil Carbon Content Estimation: A Review[1]

2.1 Introduction

Soil carbon sequestration is one possible way of reducing greenhouse gas emissions in the atmosphere (Lal, 2004a). However, to evaluate the real benefits offered by these methods (new agricultural practices, reforestation ...), large scale estimations of the carbon stock in the soils are necessary. Therefore, chemical analysis of a large amount of samples must be performed and this requires rapid, precise and low-cost analytical tools (Morgan *et al.*, 2009; Reeves III, 2009; Kuang *et al.*, 2012).

Near infrared spectroscopy (NIRS) entails acquiring and processing spectra on materials in the 700 nm - 2500 nm wavelength range. This technology enables rapid analysis and is optimized for chemical compound determination. Today it is widely used for the characterization of organic materials such as agricultural and food products or for petrochemicals and pharmaceuticals (Williams & Norris, 2001). For several years there has been a growing interest in NIRS among soil scientists (Bellon-Maurel, 2009), which is now commonly used to measure different physical and chemical parameters of soils, including carbon content. In this field, the potential of this technology is very high. It offers rapid cost-effective acquisition requiring a minimum sample preparation and measurement can be performed directly in the field. However, prediction model accuracy is insufficient for NIRS to replace routine laboratory analysis and/or to make in-situ measurements, whatever the type of soil (Reeves III, 2009).

Several recent review papers (Bellon-Maurel & McBratney, 2011; Stenberg *et al.*, 2010; Reeves III, 2009; Cécillon *et al.*, 2009) have detailed the latest improvements made by the community but also highlighted the research avenues that need to be pursued if NIRS is to become a reference technique for the measurement of soil carbon content.

[1] Alexia Gobrecht, Jean-Michel Roger, Véronique Bellon-Maurel, *Major Issues of Diffuse Reflectance NIR Spectroscopy in the Specific Context of Soil Carbon Content Estimation: A Review*, In: Donald L. Sparks, Editor(s), Advances in Agronomy, Academic Press, 2014, Volume 123, Pages 145-175

One of the biggest issues that needs to be addressed concerns the calibration process: how does the mathematical method or the sample selection influence the model quality?

In most cases, there is not a lot of thoughts put into the choice of the mathematical method, which is often made empirically (test and try). This is especially due to the fact that analytical devices used generally include software that allows users to quickly and easily apply most of the multivariate analysis methods, without being aware of the underlying theories. This reduces the relevance of the calibration.

It is therefore essential to return to fundamental laws governing spectrum formation and in particular to understand light/matter interaction in order to optimize calibration.

The aim of this paper is to review the basic theoretical assumptions underpinning NIRS and classical linear chemometric modeling and to confront them with the actual phenomena during light/matter interaction. In the specific context of soil carbon content measurement, there is an enormous gap between reality and theory. Our objective is not to quantify this gap but to evaluate its impact on the metrological quality of the measurement. This will be of significant pedagogical and practical use.

The first part of this paper presents the theoretical concepts supporting NIRS and linear chemometrics and introduces the assumptions that have to be fulfilled to build a linear model. Then, the question of NIRS compliance with these assumptions in soil related application to evaluate the resulting metrological quality of the prediction is addressed. Finally, the mathematical solutions that the authors have proposed to overcome these model quality issues are reviewed and discussed.

2.2 Theoretical concepts underlying multivariate calibration based on Near infrared spectra

2.2.1 Spectroscopy and Beer - Lambert's law in diffuse media

Beer's law is the cornerstone of quantitative analysis with Near Infrared Spectroscopy. The first assumption based on Beer's law in spectroscopy is that there is a relationship

between spectrometric response and the concentration of an analyte in a sample. It assumes that the ratio $(I_T(\lambda)/I_0(\lambda))$ of the transmitted intensity $I_T(\lambda)$ and the incident beam intensity $I_0(\lambda)$ is equivalent to:

$$T(\lambda) = \frac{I_T(\lambda)}{I_0(\lambda)} = 10^{-\varepsilon(\lambda)\cdot c\cdot l} \tag{2.1}$$

where $T(\lambda)$ is the transmittance at wavelength λ, $\varepsilon(\lambda)$ is the molar extinction coefficient (in $L \cdot mol^{-1} \cdot cm^{-1}$), c is the concentration (in $mol \cdot L^{-1}$), and l is the path length (in cm) (Workman & Springsteen, 1998).

Absorbance is a more standard form used in spectrometry, where the logarithm is applied to linearize the relationship between spectrophotometer response and concentration:

$$A(\lambda) = -\log \frac{I_T(\lambda)}{I_0(\lambda)} = \varepsilon(\lambda) \cdot c \cdot l \tag{2.2}$$

$\varepsilon(\lambda) \cdot c$ characterizes the absorption capacity of the analyzed sample and may be replaced by the absorption coefficient:

$$\mu_a(\lambda) = \varepsilon(\lambda) \cdot c \tag{2.3}$$

$\mu_a(\lambda)$ is the probability per length unit that has a photon of wavelength λ to be absorbed by the material with which it interacts. If the purpose of the measurement is to determine the concentration of a compound, the absorption coefficient of the material becomes the key parameter, because it is related to concentration (Dahm & Dahm, 2001).

This law is fundamental to spectroscopy but is strictly applicable only to transmission measurements on low concentrated transparent materials. If the sample is turbid, particulate or solid, another phenomenon, called scattering, occurs along with absorption. The scatter effect characterizes photon path changing phenomenon when it encounters a particle or when the refractive index changes (Ciani *et al.*, 2005). The photons may not only be absorbed or transmitted, but they can also be reflected, refracted or diffracted.

Beer–Lambert's law is also frequently applied to diffuse reflectance measurement of light scattering media, replacing $I_T(\lambda)$ by $I_R(\lambda)$, the intensity of the remitted radiation (Dahm & Dahm, 2001).

Through analogy with the absorption coefficient $\mu_a(\lambda)$, $\mu_s(\lambda)$ is the scattering probability of a photon per length unit. The analytical expression of the scattering coefficient μ_s is not straightforward, because the changes of direction of the photons depend not only on the size and shape of the particles, but also on their wavelength, the direction of the incident light and changes of refractive indices. Scattering has a direct impact on absorbance because the more photons are scattered, the more likely they are to be absorbed by the medium as the optical path-length increases (Dahm & Dahm, 2001).

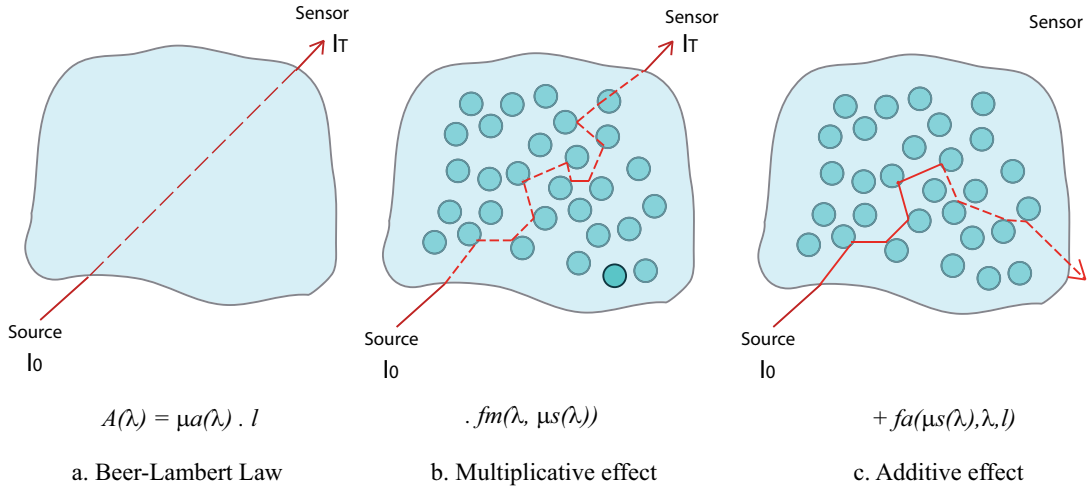

Figure 2.1: Representation of additive and multiplicative effects in diffuse material

Figure 2.1 illustrates and decomposes the effects of scattering on the absorbance signal. Figure 2.1.a shows the Beer-Lambert law when a transmittance measurement in applied on low concentrated homogeneous samples. The increase in the optical path length traveled by the photons in a scattering medium reflects a multiplicative effect on the absorbance A_λ (2.1.b). It can be characterized by a function f_m , which depends on the physical properties of the medium and the wavelength.

In addition to a multiplicative effect, light scattering occurring in the analyzed material is responsible for an additive effect on the absorbance value. Theoretical models like Beer-Lambert or Kubelka-Munk (Kubelka & Munk, 1931) assume that all the scattered

light is collected. But optical instruments are built in such a way that only a fraction $1/\alpha$ of light is detected.

$$
\begin{aligned}
I_{measured}(\lambda) &= 1/\alpha \cdot I_R(\lambda) && (2.4)\\
A_{measured}(\lambda) &= -log\ (I_{measured(\lambda)}/I_0(\lambda))) && (2.5)\\
&= log\ \alpha + log\ (I_0(\lambda)/I_R(\lambda)) && (2.6)\\
&= f_a + A(\lambda) && (2.7)
\end{aligned}
$$

The additive term $log\ \alpha = f_a$ is closely related to the scattering properties of the material and depends on the wavelength and the thickness l of the sample. Thus, $log\ \alpha = f_a(\mu_s(\lambda), \lambda, l)$, which depends on the configuration of the measuring system, is sample specific, implying inter-sample variability resulting in baseline drifts of the ideal absorption spectrum. The observed absorbance in the case of scattering samples, is no longer a linear function of concentration. Based on the above, we propose an expression of the absorbance. It integrates multiplicative and additive effects due to radiation scattering by the medium:

$$A(\lambda) = \mu_a(\lambda) \cdot l \cdot f_m(\lambda, \mu_s(\lambda)) + f_a(\mu_s(\lambda), \lambda, l) \quad (2.8)$$

with $f_m(\lambda, \mu_s(\lambda))$ the multiplicative function and $f_a(\mu_s(\lambda), \lambda, l)$ the additive function resulting in a departure from the linear relationship between the absorbance spectrum $A(\lambda)$ and the concentration c of the analyte of interest.

When a medium is complex and scatters, the useful part of the information of the signal (in our case, $\mu_a(\lambda)$, which is related to the concentration) is relatively small compared to what we can call useless information, which is due to scattering effect. Moreover, in addition to scattering, other factors, such as interactions between chemicals, may be responsible for nonlinearities (Bertran *et al.*, 1999). As the relationship between the spectrum and the concentration is not linear, it requires complex mathematical treatments to

extract useful information. This is the purpose of multivariate analysis in chemometrics.

2.2.2 NIRS and Linear Multivariate Calibration Methods

The assumptions of linear multivariate calibration methods

Multivariate calibration in near infrared spectroscopy, consists in finding a model f, that is able to relate a property y of a sample set to signal intensities or absorbance, $\mathbf{x}$, measured on these samples at several wavelengths.

Considering the theoretical Beer-Lambert law lying behind near infrared spectroscopy (section 2.2.1), it was usually assumed that the function f, with $y = f(\mathbf{x})$, was linear, which largely contributed to the development of linear multivariate calibration methods adapted to spectral data (Martens *et al.*, 2003).

When the data are centered or when it is assumed that there is no intercept, the model can be written as:

$$y = \mathbf{x}^T\mathbf{b} + e \tag{2.9}$$

With $\mathbf{b}$ the model coefficient to be estimated and the residual e, representing the deviation of the measurement y from its value $\hat{y}$ predicted by $\mathbf{x}^T\mathbf{b}$.

From a classical statistical point of view, one supposes that the model exists. The aim of the regression is to estimate the best model coefficient $\mathbf{b}$. Several methods exist, the simplest and most popular one being the Ordinary Least Squares (OLS) method. In order to use the OLS estimator, the following basic assumptions must hold (Massart *et al.*, 1998):

- H_1: Condition of linearity: the relation between x and y is linear in the parameter;
- H_2: Condition of no-multicollinearity : The regressors x_i must all be linearly independent;
- H_3: The number of observations is greater than the number of independent variables;

- H_4: Condition of homoscedasticity: The residuals e_i all have the same variance $var(e_i) = \sigma^2$;

- H_5: Condition of normality: For each individual i the residual e_i is normally distributed with mean zero, $N(0, \sigma)$. Consequently, it is assumed that for each specific x_i, the probability distribution function of y_i is also normal;

The hypothesis H_1, H_2 and H_3 have to be fulfilled in order for the OLS method to give meaningful results. The hypothesis H_4 and H_5 concern the residuals and mainly condition the quality of the estimates. These assumptions can only be tested once the regression has been performed.

If the assumptions hold, the estimated OLS parameter are the best one from the point of view of their statistical properties (convergent, non biased and of minimal variance). They are called BLUE (Best Linear Unbiaised Estimators) (Allen, 1997). Unfortunately, this is rarely the case in NIR Spectroscopy, which requires other approaches.

Chemometric approach of Multiple linear regression

The hypothesis H_1 is assumed to be met because the Beer-Lambert's law underlies the relationship between $\mathbf{x}$ and y and which is supposed to be linear but we showed in section 2.2.1 that is was false. Moreover, in spectroscopy, the spectral variable space ($\mathbf{X}$) is multidimensional, suggesting the existence among this space of a subspace where the relationship can be linear. However, the hypothesis H_2 is systematically violated: the predictors $\mathbf{x}$, composed by the absorbance measured at different wavelengths, are highly correlated with each other. Consequently, the variance of the model parameters can be very large (Bertrand & Dufour, 2006). Furthermore, assumption H_3 is not satisfied when spectral data are used as predictor variables: the number of predictor variables are generally more important than the number of available individuals. This poses a problem for computing the OLS regression coefficients $\mathbf{b}_{OLS}$ (equation 2.10) because a matrix inversion is required ($\mathbf{X'X}$) and which is not possible if H_2 et H_3 are not fulfilled.

(dimensionality problem and ill-condition of $\mathbf{X}'\mathbf{X}$) (Næs *et al.*, 2002).

$$\mathbf{b}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y \tag{2.10}$$

In chemometrics, the approach of multiple linear regression is a little bit different from that of classical statistics. In classical statistics, one assumes *a priori* that the model exist and the objective is to find the best model parameter. In chemometrics, the model is built from the available data (a calibration set ($\mathbf{X}$, y)), and then tested, if possible on an independent set or by cross validation (Geladi *et al.*, 1999). Model quality is assessed by comparing the predicted values to the measured values of the test set. If the model satisfies validation criterion, it can then be used for predicting a new sample. During the model building step, i.e. estimating the model coefficients, the pre-listed assumptions are generally not tested and this whatever the estimation method chosen.

Furthermore, to overcome the problems posed by the violation of assumptions H_2 and H_3, chemometricians have developed new methods. To reduce the number of explanatory variables and limit the risk of collinearity, the linear regression is performed in a spectral space of limited dimension. To reduce the spectral space dimension, it is possible to select a certain number of variables assuming that the excluded ones do not significantly improve the model. Stepwise Multiple Linear Regression (MLR) (Martens & Næs, 1989) or CovSel (Roger *et al.*, 2011) are some examples of variable selection methods. One limitation to this approach is that the variable selection can be arbitrary (in stepwise MLR for example) and highly dependent on the available data set (Williams & Norris, 2001).

Another way to reduce the spectral space dimension is to build new variables from linear combinations of descriptors x_i.

$$\mathbf{T} = \mathbf{XP} \tag{2.11}$$

With $\mathbf{T}$ the constructed latent variables and $\mathbf{P}$ the loadings.

The regression is performed between the variable to predict (y) and so called *la-*

tent variables (or scores). This approach is among the most used in chemometrics and the principal algorithms are Principal Component Regression (PCR) and Partial Least Squares regression (PLS), the latter being developed by Wold (Wold *et al.*, 2001).

With PLS, the regression model based on these new variables can still be written in a simple form like equation 2.12 and the OLS method is used to predict the model coefficient $\mathbf{q}_{OLS}$ because assumptions H_2 and H_3 become fulfilled: the latent variables are not correlated one with another (by construction, they are orthogonal) and they are fewer than the number of observations.

$$y = \mathbf{T}\mathbf{q}_{OLS} + e \tag{2.12}$$

With $\mathbf{T}$ containing $\mathbf{t}_k$ latent variables (or scores) and $\mathbf{q}_{OLS}$ the regression coefficients to be estimated by OLS.

$$\mathbf{q}_{OLS} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'y \tag{2.13}$$

$$\hat{y} = \mathbf{T} * (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'y = \mathbf{X}\mathbf{P}(\mathbf{P}'\mathbf{X}'\mathbf{X}\mathbf{P})^{-1}\mathbf{P}'\mathbf{X}'y = \mathbf{X}\mathbf{b}_{PLS} \tag{2.14}$$

with $\mathbf{b}_{PLS} = \mathbf{P}(\mathbf{P}'\mathbf{X}'\mathbf{X}\mathbf{P})^{-1}\mathbf{P}'\mathbf{X}'y$.

However, the PLS (or PCR) remains within the paradigm of linear regression, and is therefore subject to the same constraints of the assumptions regarding the residuals (assumptions H_4 and H_5). If the conditions are strongly violated, the lower quality of the model parameter estimates will directly impact the prediction quality (accuracy and robustness). This may be the case if the residuals are heteroscedastic (their variance is not constant) or if the distribution of e_i is not normal (Fearn, 2012).

2.3 Does diffuse reflectance spectroscopy satisfy the assumptions of the linear multivariate calibration methods when applied on soils?

2.3.1 Hypothesis H_1: Violation of the Beer-Lambert law

Soil is a mixture of mineral and organic matter with a physical structure composed of macroscopic aggregates of particles and porous spaces which may contain water or air (Ben-Dor *et al.*, 2009). Soil is therefore a highly absorbing and scattering medium. In this case, the ideal theoretical conditions that are required by the Beer-Lambert law cannot be satisfied. This affirmation is not specific to soils, indeed it is commonly applicable for all scattering media. But soil is, from this point of view, extremely challenging, especially when the samples are measured *in situ* or without prior sample preparation (drying and grinding for example). The application of near infrared spectroscopy to measure soil carbon concentrations has yet to be performed in a satisfactory manner and principally becomes a chemometric challenge. Indeed, nonlinearities introduced by the scattering effect are a real constraint for linear multivariate calibration as they impact hypothesis H_1.

2.3.2 Violation of the assumptions about residuals (H_4, H_5)

The carbon content of a given set of samples does usually not follow a normal distribution. Some authors maintain that it is positively skewed (Vistelius, 1960; Reimann & Filzmoser, 2000) with a high occurrence in low carbon concentrations and other state that the distribution is lognormal (Ahrens, 1954; Parkin *et al.*, 1988; Clark, 1999; Brejda *et al.*, 2000). According to Reimann & Filzmoser (2000), this asymmetry is frequent for many environmental variables of low values as they can not be given negative values and are thus truncated at 0. Also, the important spatial dependence of these type of variables may reflect the existence of several subpopulations, which is inconsistent with a strictly

gaussian distribution.

It is essential to know the distribution function of a variable to be able to characterize a given data set. Describing a population when the distribution is symmetric is unambiguous because the mean, median and mode coincide and all can be taken as the center. In most studies applied to soils, the statistics provided are insufficient to adequately describe the variables. Very often, only the mean and standard deviation are shown with sometimes the extreme values (min, max), although the two first values are of little interest if the data are not normally distributed, which is usually the case.

As a consequence of this asymmetric distribution of carbon content in soils, there is a high probability that residuals neither satisfy the condition of homoscedasticity (H_4) nor the condition of normality (H_5). The causes of heteroscedasticity are difficult to identify, but authors agree that the variation of the residual variance is a *by-product* of the violation of other assumptions (Osborne & Waters, 2002) like an asymmetric distribution of y. Measurement errors of x can also contribute to the residual error term. Consequently, an increase of the residual error as a function of y is often observed (Geladi *et al.*, 1999).

2.4 Impacts on the metrological quality of the prediction

In metrology, several indicators are usual to characterize the properties of a method or an instrument: reproducibility, repeatability, sensitivity, precision, accuracy or uncertainty. Zeaiter *et al.* (2006) recalls the definitions of these terms. How are these parameters, specifically signal quality and prediction model quality, affected when NIRS is applied to estimate the carbon concentration of soils?

2.4.1 Signal quality

The quality of a spectroscopic signal determines the quality of the resulting measurement. This quality can be assessed using the signal to noise ratio (SNR). Technological advances in instrumentation have reduced optical noise and improved signal quality. But in scattering materials such as soils, only a part of the signal contains relevant information related to the absorbance and therefore useful for calibration. The remaining information, which results from scattering, contributes to noise. Signal sensitivity decreases as the scattering effect increases compared to chemical absorbance ($\mu_s >> \mu_a$). In addition, scattering changes the optical path of the photon in a random manner, which in turn, impacts another metrological quality criterion, i.e measurement reproducibility.

2.4.2 Model quality

As seen above, building a linear model to directly estimate soil carbon content from a soil spectrum will probably be lacking in performance as the ideal conditions are not met: non-linearity of the spectra - concentration relationship, non-normal distribution of y, which leads to biased estimation of the model parameter and to heteroscedasticity. Fearn (2012) discusses the possible consequences of non-normally distributed data on robustness of the least square fit, the validity of significance tests and the relevance of statistics used to assess the fit such the Standard Error of Calibration (SEC).

However, as these optimal conditions are rarely met the BLUE can not be found and the overall quality of the model has to be evaluated through a validation step. To perform it, the model is tested on an independent data set in order to calculate performance indicators, of which the following are the most frequently used in soil science:

- The SEP (Standard Error of Prediction) is the root mean square average error recorded on a independent dataset (validation set). It can be broken down as follows: $SEP^2 = biais^2 + SEP_c^2$. The bias reflects systematic error, related to systematic variations of influence factors (e.g instrument, the analysis methodology ...). SEP_c (for SEP corrected for biais) is the residual variance (Davies & Fearn,

2006a).

- The RPD (Ratio of Performance to Deviation), $RPD = SD/SEP$, is a popular indice used in soil science. It standardizes the value of the SEP with respect to sample population dispersion (i.e the standard deviation).

Bellon-Maurel *et al.* (2010) offers a critical overview on the use of these indicators in the specific case of spectroscopy applied to soils analysis. In particular, RDP based on skewed data is considered irrelevant since it is calculated from the standard deviation of the dataset ($RPD = SD/SEP$), whereas SD is not a good indicator to correctly describe the dispersion of a skewed dataset. They propose to improve this indicator by introducing a more representative distribution parameter of y based on the interquartile distance $Q1-Q3$. With the same objective, Limpert *et al.* (2001) offer another alternative to characterize the log normal data with the geometric mean.

Beyond the overall performance assessment of the model, a new analytical technique should be able to predict a value for a new sample with the least uncertainty. Calculating the SEP is insufficient, because (i) it contains a part of systematic error (bias), (ii) its value does not provide any information about an individual sample, and (iii) because uncertainty will vary from one sample to another. It is necessary to compute $var(\hat{y})$, i.e. the uncertainty attached to the estimated sample i.

Several expressions of linear model uncertainty, which can be expressed as $var(\hat{y})$, can be found in the literature. Zhang & Garcia-Munoz (2009) review most of them. They are based on the theoretical error propagation framework, which identifies and assesses the contribution of all sources of uncertainty associated to the model parameters. The terms of these expressions differ depending on the simplifying assumptions made. Fernandez-Ahumada *et al.* (2012) discuss these assumptions and propose a general expression of the uncertainty based on least restrictive assumptions.

Whatever the expression, one term is common: $(\widehat{\mathbf{x}} - \widehat{\mathbf{x_c}})$, with $\widehat{x}$ the measured spectrum and $\widehat{x_c}$ the spectrum at the center of the model. This term, which reflects the distance of the spectrum $\widehat{x}$ to the center $\widehat{x_c}$ of the model is called *leverage* (Cook &

Weisberg, 1982). Prior to regression, the data are centered in relation to a fixed point. This helps to compare the scales of both dependent and independent variables. Thus, analysis of the deviation from the center and the distribution around this point becomes possible , thereby facilitating the interpretation of the regression model. The further the sample is from the center of the model, the higher the prediction uncertainty.

In most cases, the models are mean-centered. Hence, $\widehat{x_c} = \bar{x}$, which coincides with the mode, if the dataset is symmetrically distributed. Close to the mean, the predictions will have a smaller variance. This phenomenon, which is called *Dunne effect*, leads to an overall improvement of model performance when most of the samples to be predicted are close to the distribution mean (Martens & Næs, 1989).

The SEP can be written $SEP^2 = SEP_c^2 = \Sigma var(y_i)P(y_i)$. In order to minimize it, it is necessary to have the maximum of the distribution of y, $P(y_i)$, coincide with the minimal uncertainty $var(\widehat{y_i})$. If the center of the model $\widehat{x_c}$ is close to the mode, then the leverage effect will be reduced.

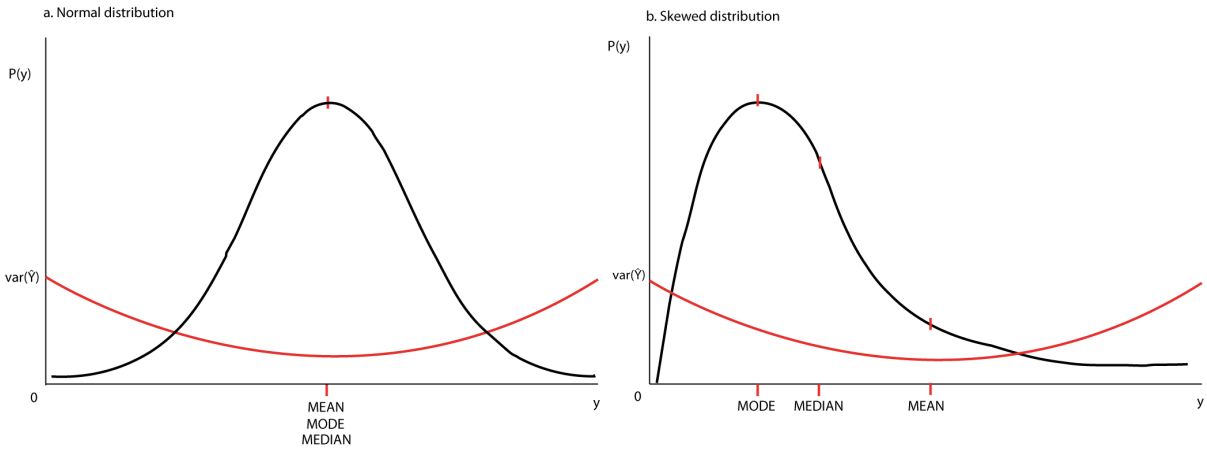

Figure 2.2: Mode, Median and Mean in (a) normally distributed and (b) positively skewed data and representation of the leverage effect on the prediction uncertainty when the model is mean centered

From that, we may state that :

- When the distribution is symmetric, it makes sense to center the model on the mean, because it corresponds to the mode;

- In the case of soil carbon content which distribution is skewed (Figure 2.2.b), the mode and the mean do not coincide. The overall quality of a mean-centered model

will be directly affected as there will be more predicted samples with high uncertainty (because of high leverage) in the lower values of y;

- The mode determination is not straightforward to achieve. For statisticians, the geometric mean ($\sqrt[n]{\prod y_i}$) is less sensitive to high values in a positively skewed dataset than the arithmetic mean ($\frac{1}{n}\Sigma y_i$). It gives, therefore, another more accurate data centrality in the case of positively skewed distributions. Thus, the geometric mean, zero (Seasholtz & Kowalski, 1992) or even the median are alternatives worth considering when choosing the center of the model, these values being closer to the mode than the mean.

2.5 Data pretreatment help to fulfill the assumptions of linear multivariate calibrations

Preprocessing methods provide mathematical transformation of the signal in order to amplify the useful (i.e chemically related) part of the signal and reduce the irrelevant (i.e the scattered) information. A first approach involves applying transformations on spectral variables (spectral preprocessing) in order to remove the scattering effect and to restore, to a certain extend, the linear relationship between the spectrum and concentration, which tends to satisfy H_1. The other approach is to restore the symmetry of the distribution of y by applying statistic transformations (Martens & Næs, 1989).

2.5.1 Spectral preprocessing

The performance of linear calibration depends on the degree of linearity between the independent variables and the predicted variables. Sources of spectral variations explained in section 2.2.1 are identified causes of nonlinearities and can be partly corrected with pretreatments. Thus, based on the proposed expression of the absorbance (equation 2.8), the spectral pretreatments are designed to reduce the impact of multiplicative and additive effects while maintaining a sufficient amount of useful information to ensure an

efficient prediction model. Ideally, the corrected absorbance spectrum should look like the absorbance coefficient $\mu_a(\lambda)$.

Geometric spectral pre-processing methods

A first group of widely used methods, effectively remove the additive and multiplicative effects due to scattering from the spectra. Among them: Multiplicative Signal Correction (MSC) (Geladi *et al.*, 1985), Extended MSC (EMSC) (Martens, 1991), Standard Normal Variate (SNV), Detrend (Barnes *et al.*, 1989) and normalization.

These methods all correct the measured spectrum by describing the multiplicative and the additive effect based on explicit functions like those introduced in equation 2.8.

$$\mathbf{x_{corr}} = \frac{\mathbf{x_{org}} - f_a}{f_m} \tag{2.15}$$

where $\mathbf{x_{corr}}$ is the corrected spectra, $\mathbf{x_{org}}$ the measured spectra, f_m and f_a the function describing the multiplicative and additive effect of scattering on the original spectra respectively.

These preprocessing methods differ in the parameter estimation of the explicit functions. For example, in the MSC method, the corrective parameter are the regressors (intercept and slope) of the ordinary least squares regression line between a reference spectrum and the spectrum to be corrected. The f_a function corresponds to the intercept and f_m to the slope. The reference spectra commonly used is the mean spectra of the calibration set. This choice is open to discussion when used with skewed population as the mean spectrum is not necessarily the most representative spectrum of the database.

SNV method has been introduced by Barnes *et al.* (1989) in order to reduce the multiplicative effect of scattering. Each spectrum is centered and reduced. Therefore, in equation 2.15, f_a corresponds to the mean value of the spectrum to be corrected and f_m to its standard deviation. With this approach, a reference spectrum in not required to estimate the corrective parameter.

With these two methods, the corrective functions are quite simple but are based on

the hypothesis that the multiplicative effect is not wavelength dependent. However, in section 2.2.1, we stated that the additive function depends on (i) the structure of the material, which can be characterized by the scattering coefficient $\mu_s(\lambda)$, (ii) the volume traveled by the photons (or the thickness of the sample) and (iii) the wavelength.

In order to get more closely the physical realities of the scattering phenomenon, some methods introduce more complexity in the corrective function of the additive effect. Detrend, often associated to SNV, removes the baseline curvature by adjusting it with polynomial function of the wavelength λ (for example of a second order). EMSC, an extended version of MSC, is comprised of a second order adjustment of the reference spectra, a quadratic function of the wavelength or also pure spectra of the chemical compounds of the studied material. There is, in theory, no mathematical limitation to an increase in the complexity of the corrective functions (Rinnan *et al.*, 2009). Thus, Thennadil & Martin (2005) substitute the wavelength dependent term in EMSC ($d_i\lambda + e_i\lambda^2$) by $d_i log(\lambda)$. Their reasoning is based on the fact that, for small-particulate media, the scattering intensity is proportional to λ^{-4} (Rayleigh approximation). By extension, they suppose that light scattering can be expressed using the form $\alpha\lambda^{\beta}$, which becomes proportional to $log(\lambda)$ when transformed in absorbance units. The introduction of a semi-empirical model based on the physics of the scattering phenomenon produced better calibration models on simulated data (Thennadil & Martin, 2005).

Derivative methods are other very popular pretreatment techniques used in chemometrics. The first derivative removes an additive constant and the second derivative removes also the slope of the baseline. In other words, these methods suppose that the additive function, f_a has a null derivative (first and second), which eliminates the non-informative parameter. Among the most popular methods, the Savistzky-Golay algorithm (SG) (Savitsky & Golay, 1964) associates the derivative with a smoothing step to reduce (i) the random noise of the measurements and (ii) its amplification by the differentiation step. The derivative methods have the advantage of accentuating the spectral resolution but it is a deep spectral transformation which runs the risks of removing part of the useful information for the calibration process.

Orthogonal projections

Another preprocessing strategy consists in identifying a sub-space of the total spectral space, which supports variability caused by the non-informative part of the signal. This subspace is then subtracted by orthogonal projection so that the resulting space is, *in theory*, independent of the scatter - or any other - effect. Differences between the various orthogonal projection methods exist and mainly concern the method used to identify the sub-pace to be orthogonalized. The Orthogonal signal correction (OSC) method, proposed by Wold *et al.* (1998), extracts and remove the useless sub-space by selecting the principal components that contain most of the spectral variability and are not correlated to y. The main drawback of this approach is that the calibration set does not necessarily contain all the variability due to the scatter effect. Moreover, the PLS algorithm is subject to the same constraint so OSC does not add any value to the calibration quality if it is based on a PLS (Verron *et al.*, 2004).

Another approach called External Parameter Orthogonalization (EPO) was developed by Roger *et al.* (2003). It consists in building the subspace containing the influence factor using a dedicated experimental design where spectra, with and without perturbation, are collected. The spectral influences are then removed from the total spectral space by orthogonal projection. The identified subspace contains, for example, the additive effect due to scattering, but also all the linear combinations of this effect. On the other hand, if y and the spectral variation due to the influence factor are correlated, EPO will remove some useful information and therefore impoverish the calibration database.

Discussion

In the NIRS publications applied to soils, *on the shelf* pretreatments like MSC or SNV are almost systematically used as they have the advantage of being widely implemented in multivariate analysis software. On the other hand, orthogonal projection methods are not so popular, although promising for materials as complex as soils. Minasny *et al.* (2011) tested the EPO method to overcome the moisture effect on the spectra. In this

case, it would seem appropriate to set up a dedicated experimental design to characterize the scattering effect on the spectra by varying the physical characteristics of the soil. Preys *et al.* (2008) also propose a method, that combines EPO and OSC and which seems particularly suited to soil related spectroscopy. In an approach similar to EPO, a spectral database is built by making the factor of influence vary and then OSC is used to select the principal components that carry a maximum spectral variability while remaining orthogonal to y. These identified directions are then used to remove the useless subspace from an existing dataset.

2.5.2 Restoration of y distribution symmetry

The spectral preprocessing methods presented above do not solve the problem arising from the skewed soil carbon content distribution which impacts H_4 and H_5. Statisticians (Webster, 2001; Kleinbaum *et al.*, 2008), recommend a variable transformation to restore its symmetry. Thus, Vasques *et al.* (2008) log transform the *Total Carbon* and Bartholomeus *et al.* (2008) apply $(SOC)^{1/4}$ to reduce the skewness index of the soil organic carbon (SOC) distribution from 2.85 to 0.91.

Mathematically, this approach makes sense and contributes to the reduction of the model prediction error by lowering leverage (the mode is closer to the model center). After back transformation (i.e. the inverse transformation), the most probable values (in the mode neighborhood) are predicted with less uncertainty, which improves the quality of the model (cf section 2.4.2).

However, these statements lead us to make the following comments:

- If the objective is model comparison, it is possible to provide model quality indexes without back transformation, as in Minasny *et al.* (2011), but if the purpose is to assess absolute prediction quality, it is mandatory to back transform the predicted result, in order not only to be able to assess the weight of the extreme values but also to retrieve the original units in the performance indexes of the model.

- One may ask the question of the relevance of these transformations in spectrometry.

The modeling process is based on the existence of a relationship dictated by physical law (Beer-Lambert's law or others) between variables. How do these variable transformation offset the non-linearities of the signal/variable relationship?

- If the objective of variable transformation is to restore its distribution symmetry in order to lessen the total uncertainty of model prediction, then precautions must be taken to avoid hazardous conclusions from these results.

To conclude, linear methods, and in particular PLS are, by far, the most used calibration methods to predict soil parameters (Viscarra Rossel *et al.*, 2006). These methods comply with the theoretical framework of the Beer-Lambert law, are easy to implement, and the model parameters can be interpreted from a spectroscopic point of view.

However, the PLS method often fails to circumvent all the difficulties exposed above because basic assumptions are not satisfied. Therefore, PLS should be associated with other mathematical techniques such as pretreatments (2.5.1), wavelets, Neural Networks (Mouazen *et al.*, 2010), which may help approaching these assumptions.

And even then, the predictions are sometimes not satisfactory or the interpretation of the model significance theoretically difficult. Thus, new calibration strategies not based on linear model may be used for soils. We will study the ones discussed in NIRS literature related to soil science.

2.6 Other calibration strategies as an alternative to linear models

2.6.1 Local strategies

Local methods, although considered as non-linear methods, take advantage of the simplicity associated with linear methods. The common principle of local methods is to select a sub-set of samples from a training set in order to build a local linear model used to predict an unknown sample. The sample selection can either be based on auxiliary

information or on proximity of the samples in terms of their spectral characteristics. The global model built on all the created local models is capable of modeling the non-linear structure of the data set (Gogé *et al.*, 2012).

Sub-set selection based on auxiliary information

The sub-set selection of spectra in order to build a local model can be based on expert knowledge or information related to the data. The type of soils, geology or geographic location (Sankey *et al.*, 2008) can be used to select proximal samples and thus reduce, to a certain degree, the spectral variability between samples (Stevens *et al.*, 2010). The concentration range of the soil carbon content can be narrower and, consequently, the distribution of y more symmetric (Janik *et al.*, 2009). Some of the basic assumptions of multiple linear regression will be better fulfilled.

Although this strategy has not been subject to a lot of testing, this approach seems interesting in the case of NIRS applied to soils as there is a large number of auxiliary information (McBratney *et al.*, 2003) that could be used to perform intelligent sampling designed to create this sub-calibration set.

Sub-set selection based on the spectral characteristics of the neighborhood samples

The selection of the sub-set based on the spectral characteristics of the samples is a more commonly used technique in soil science. These approaches originates from the K-NN (K-nearest neighbor) classification method (Kowalski & Wold, 1982). A set of spectra that are spectrally similar to the unknown sample to be predicted is selected from a larger database. The so created sub-set is used to build a calibration model specifically dedicated to the prediction of the new sample.

The three main local methods found in the literature are LWR (Locally Weighted Regression) (Cleveland & Devlin, 1988; Næs *et al.*, 1990), the LOCAL algorithm (Shenk *et al.*, 1997) and CARNAC methods (Davies & Fearn, 2006b). They differ in their approaches to select the neighboring samples. LOCAL and LWR use an euclidean dis-

tance between samples while CARNAC uses local averaging instead of local regression for prediction and performs data compression (Fast Fourier Transformation) before the distances are computed.

These methods are very attractive for complex and heterogeneous matrices such as soils. The spectral similarity between the selected samples suggests a certain homogeneity with regards to the structure (Fernández Pierna & Dardenne, 2008).

However, implementation of local methods within soil spectral databases raise some difficulties. In its basic principle, the selection of local samples within the database should be related to the spectral absorbance feature, i.e. related to the analyte of interest concentration. In a soil spectrum, the impact of scattering is greater than absorbance ($\mu_s >> \mu_a$). Therefore, the subset selection step in local methods will be mainly based on the physical properties of the soil rather than it's chemical content. As a consequence, the local model will probably not meet the expected quality. Performing the best strategy to select the local samples still remains an open question in soil science. The ideal case would be to be able to compare the samples regarding their absorbance coefficient μ_a.

Several solutions can solve, to a certain extent, this issue:

- In order to homogenize the scatter effect between the samples and therefore enhance the chemical absorbance compared to scattering, soil samples are dried, sieved at $2mm$ and sometimes grounded at a smaller particle size ($< 0.2mm$). Sample preparation before spectral measurement is a way to control the effects of the physical influence factors such as moisture, particle size, bulk density ... (Stenberg & Viscarra-Rossel, 2010). Soil sample preparation is a very common procedure. If it is suited for laboratory analysis, it is not possible if the spectral acquisition is performed in-field, on bulk samples.

- Spectral pretreatment (section 2.5.1) have, in a certain way, the same objective than the sample preparation as both aims at reducing the impact of the scatter effect against the chemical absorbance. If the method has proven its efficiency, it does not always solve the problem as the scatter effect is very complex and multidimensional.

Part of μ_s effect can still remain, even after preprocessing. Furthermore, selection of the best preprocessing method is generally based on the model performance and not in order to optimize the sample subset selection for local methods (Igne *et al.*, 2010).

- Another possibility is to perform the local sample selection within an exhaustive database, i.e. containing all the possible variability regarding the carbon concentration (which is usually expected from a dataset) but also all the possible variability due to the physical properties of the samples. If some soil scientists actively work on building an global soil spectral library (Viscarra Rossel, 2008), it is far from achieved. An alternative approach consists in spiking (i.e. completing) global datasets with some local (geographically) samples (Wetterlind & Stenberg, 2010; Brown, 2007).

Results

Despite these obstacles, local methods generally give good results (Igne *et al.* (2010) and Gogé *et al.* (2012) with LWR and Fernández Pierna & Dardenne (2008) with LOCAL). In their study, Gogé *et al.* (2012) apply the LWR method on 2500 soil samples representative of the entire french territory. An analysis of the auxiliary characteristics of these samples show a strong relationship between these samples and the geology of their respective sites. These findings merit deeper analysis but provide arguments for the use auxiliary information to improve the similar samples selection.

2.6.2 Non-linear processing methods

To overcome the non-linear relationship between the spectrum and the reference value, the use of nonlinear methods is becoming increasingly popular in soil science (Stenberg, 2010). Table 2.1 divides the references in three main categories of methods.

1. **Tree based regression**: Tree-based methods for classification and regression were introduced from a statistical perspective by Breiman (1984) (CART, *classification*

Table 2.1: A review of non-linear methods used to predict soil carbon content with NIR diffuse reflectance spectroscopy

Calibration methods[1]		Parameter	n	R^2	RMSE/SEP	Reference
	MARS	OC $g \cdot kg^{-1}$	300	$R^2_{val} = 0.8$	$RMSE = 3.1$	Shepherd & Walsh (2002)
	MARS	OC %	1104	$R^2_{cv} = 0.8$	$RMSE = 1.02$	Viscarra Rossel & Behrens (2010)
Tree	BRT	OC $g \cdot kg^{-1}$	3793	$R^2_{cv} = 0.82$	$RMSE = 9$	Brown *et al.* (2006)
based	BRT	OC $g \cdot kg^{-1}$	52	$R^2_{val} = 0.96$	$SEP = 3.8$	Sankey *et al.* (2008)
regressions	CART	OC $g \cdot kg^{-1}$	30	$R^2_{cv} = 0.48$	$RMSE = 2.64$	Ballabio (2009)
	Cubist	OC %	157	$R^2_{test} = 0.96$	$RMSE = 0.35$	Minasny & McBratney (2008)
	Treenet	OC %	257	$R^2_{test} = 0.71$	$RMSE = 0.76$	Minasny & McBratney (2008)
	RF	OC %	1104	$R^2_{cv} = 0.8$	$RMSE = 1.23$	Viscarra Rossel & Behrens (2010)
	ANN	OC %	214	$R^2_{cv} = 0.94$	$RMSE = 0.89$	Udelhoven & Schütt (2000)
Artificial	PLS-NN	OC %	256	$R^2_{test} = 0.94$	$RMSE = 0.5$	Janik *et al.* (2009)
neural	ANN	OC %	1104	$R^2_{cv} = 0.89$	$RMSE = 0.75$	Viscarra Rossel & Behrens (2010)
network	BPNN	OC %	45	$R^2_{test} = 0.94$	$RMSE = 0.54$	Mouazen *et al.* (2010)
	BPNN	OC %	157	$R^2_{test} = 0.71$	$RMSE = 0.76$	Fernández Pierna & Dardenne (2008)
	ANN	SOM $mg \cdot g^{-1}$	10	$R^2_{val} = 0.86$		Daniel *et al.* (2003)
	MLP	SOM %	60	$R^2_{test} = 0.88$	$RMSE = 0.35$	Fidêncio *et al.* (2002)
	RBFN	SOM %	60	$R^2_{test} = 0.92$	$RMSE = 0.25$	Fidêncio *et al.* (2002)
	LS-SVM	OC %	40	$R^2_{val} = 0.89$	$RMSE = 0.5$	Vohland *et al.* (2011)
Support	SVMR	OC $g \cdot kg^{-1}$	30	$R^2_{cv} = 0.61$	$RMSE = 2.35$	Ballabio (2009)
Vector	SVMR	OC $g \cdot kg^{-1}$	102	$R^2_{val} = 0.83$	$RMSE = 5.37$	Stevens *et al.* (2010)
Machine	LS-SVM	OC %	157	$R^2_{test} = 0.89$	$RMSE = 0.56$	Fernández Pierna & Dardenne (2008)
	SVMR	OC %	1104	$R^2_{cv} = 0.84$	$RMSE = 0.92$	Viscarra Rossel & Behrens (2010)
	LS-SVM	OC $g \cdot kg^{-1}$	106	$R^2_{val} = 0.6$	$SEP \approx 0.8$	Igne *et al.* (2010)

1: Abbreviations: MARS=multivariate adaptative regression splines; BRT= boosted regression trees;
CART= Classification and regression trees; RF=random forest; ANN = artificial neural network;
PLS-NN= Partial least squares neural network; BPNN = backpropagation neural network;
MLP= Multiple layer perceptron; RBFN = Radial basis function network;
LS-SVM=least squares support vector machine; SVMR=support vector machine regression;
OC=organic carbon; SOM = Soil organic matter; RMSE = root mean square error; SEP= standard error of prediction

and regression trees). Regression trees are used when the response variable is continuous, while classification trees are used for a categorical response. The fundamental idea is to split the data into subsets giving a splitting criteria. Examples of splitting criteria are given in Breiman (1984). The same procedure is applied in turn to the descendant nodes, sometimes called recursive partitioning. Usually, the trees are grown until a stopping criterion is met, for example, all nodes contain fewer than some fixed number of cases, then pruned back to prevent over-fitting (Breiman, 1984). Once a tree has been grown and possibly pruned, it will have some non-partitioned nodes called terminal nodes. Predicted values are obtained by computation of the terminal node outputs in order to have an unique value. The CART algorithm (Breiman, 1984) underpins theses methods which differ in the way of calculating the final output. The main methods found in the soil related literature are MARS and BRT (Friedman, 1991), Random Forest (Breiman, 2001), Cubist (Quinlan, 1992).

2. **Artificial Neural Network**: The design and the basic concept of Artificial Neural Networks (ANN) have been adopted from data processing in biological nervous system: a group of cells receives information, others forwards or stores it and a last group processes it before releasing it. In a modeling process, input variables of the network undergo non-linear transformations (sigmoid or logistic function), run through several layers where information is combined (often weighted) and the output value obtained will be assigned to the target parameter. ANN are very flexible functions with many parameters to be determined. They are adaptive and possess the ability to model almost any relationship (linear or not). However, in NIRS, as the input data (x) are generally the spectral variables and therefore,very numerous, the network will be large and the parametrization tricky with no experience. To overcome this issue, some authors use principal components or latent variables as inputs (Mouazen *et al.*, 2010; Janik *et al.*, 2009). But ANN remain a *black box*-type approach where the network architecture does not give clues to

understand how the calibration really works.

3. **Support Vector Machine**: SVM are kernel-based learning methods. Initially a classification method, the concept has been extended to multivariate regression (Cogdill & Dardenne, 2004). A hyperplane describing as precisely as possible the spectral data set is defined using a kernel function. The model reduces the complexity of the data by the construction of subset of support vectors. The parameters to be defined are (i) the distance between the hyperplane and the dataset and (ii) the kernel function. Computational times are huge (Vohland *et al.*, 2011; Igne *et al.*, 2010) and this method has not yet much applied on soil data.

As shown in Table 2.1, the prediction model for these applications can be of very good quality. In theory, the methods are capable to model the relationship between the spectrum and the variable of interest, even if it is nonlinear. Moreover, they presuppose nothing about the distribution of y. These are advantages in favor of using non-linear methods for the calibration of soil parameters using NIRS.

However, their use remains still marginal in soil science because the algorithms are not present in commercial NIR spectrometers and are sometimes considered as black boxes. Training and structure optimization may require long computation time.

2.6.3 Bayesian methods

To our knowledge, Bayesian methods have not yet been applied to soil data. Chen *et al.* (2007) provide a theoretical perspective on the value of using Bayesian methods in chemometrics, in comparison with more traditional methods. The Bayesian framework seems relevant when addressing a number of issues posed by NIRS applied to soil. Fearn *et al.* (2010) and Pérez-Marín *et al.* (2012) used the Bayesian framework to predict the composition of materials showing some similarities with soils, i.e. forages. Forages have got similar spectral signal behavior with soils (scattering, heterogeneity of composition and structure). In addition, these authors identify both nonlinearity problems and issues related to the non-gaussian distribution of variables to predict.

Whereas previously described methods involve regression, linear or nonlinear, between reference values and spectral data (inverse regression) Bayesian estimation methods combine prediction model distribution and a *a priori* distribution describing the population of samples to be predicted (Fearn *et al.*, 2010).

With Bayesian approaches, a prediction consists in computing the distribution probability ($p(y|x)$) of the reference measurement y knowing the spectra x. This term describes the variability of y for samples with the same spectra. A kernel function (kernel density estimate) is used to model $p(x|y)$. The Bayes theorem gives:

$$p(y|x) = p(x|y)p(y)/p(x) \tag{2.16}$$

If the distribution is normal, equation 2.16 can be solved analytically. However, in order to be able to apply this method in a more general situation, specifically with skewed $p(y)$ distribution, Fearn *et al.* (2010) use a discrete set I of $y_1, \ldots, y_I$ values for y. Thus, equation 2.16 becomes:

$$p(y_i|x) = p(x|y_i)p(y_i)/p(x) \quad \text{for } i = 1, \ldots, I \tag{2.17}$$

As an output, a discrete distribution of $p(y|x)$ is obtained. The mean of this distribution can be used to assign a value to y, but the other characteristics of this distribution (i.e the median) can also be used if necessary. This is of high interest because it becomes possible to calculate a prediction interval for $\hat{y}$, from this distribution, which informs in another way about the quality (or uncertainty) of the model (Pérez-Marín *et al.*, 2012).

Predictions of forage compounds using this method are better than with PLS and similar in terms of quality to local methods, with advantages such that (i) of not requiring any calibration set to build the model and (ii) to be able to affect, in addition to a value for y, a prediction interval (Chen *et al.*, 2007).

2.7 General conclusions

The purpose of this review paper was to offer a focus on the basic theoretical concepts supporting NIRS and the use of linear multivariate calibration in soil applications especially related to soil carbon content measurement. Compared to other studied materials, soil does present some specific features that needed to be highlighted regarding their influence on these theoretical concepts:

- soils are highly diffuse materials where the scattering effect dominates in the spectra and introduces non-linearities in the relation with the carbon content;
- soils are extremely complex in terms of chemical composition and physical structure, which present a high variability between samples, especially in-field;
- soil carbon content presents a highly skewed distribution.

Because of this, the spectral measurement conditions (including sample preparation) and the choice of calibration methods will directly impact the quality of the prediction model. We showed that on the one hand, the optical phenomena, and especially the scatter of photons and on the other hand, the data structure are limiting factors to build good and robust calibrations.

New approaches are emerging in the soil literature, mostly in chemometrics, such as nonlinear or local methods, but their added value has still to be confirmed.

To conclude, the main goal of this study was to make soil scientists fully aware of the critical point of using classical chemometric methods without completely being aware of the underlying theory. It is also an opportunity to show that specific developments are needed to adapt NIRS and chemometrics to soil applications:

- A need in better understanding the light-soil interaction in order to better express the absorbance as a function of μ_a and μ_s;
- A need in new optical acquisition methods capable of overcoming the issues of scattering, especially in the case of in-field measurements;

- A need in adapted preprocessing methods and chemometric calibration methods.

Investing, simultaneously or not, these paths of research will allow to take an important step in the metrological quality of the soil carbon content measurement by NIRS.

Scientific issues at stake

According to the analysis carried out in this chapter, the main issue faced by NIR Spectroscopy applied to soil is the negative impact of light scattering on the signal quality, which directly affects the quality of the prediction models. Although a considerable effort has been made in the empirical scattering correction techniques, they are not sufficient to solve the problem of multiple scattering completely.

The conclusions drawn from this chapter lead us to investigate this issue through the prism of signal quality. In other terms, to provide answers to the following question as the scientific heart of this thesis:

How can we measure an absorbance signal of optimal quality on highly scattering materials ?

The signal formation is the first stage of the whole analytical method which also includes the calibration step. When light interacts with matter, it picks up both physically and chemically related information about the material. Hence, improving the quality of the signal means increasing its sensitivity and selectivity to the analyte of interest. The figures of merit of the analytical method such as precision and robustness will be positively impacted. In the case of highly scattering materials such as soils, the challenge is mainly to restore the linearity between the absorbance and the chemical property of interest, which is affected by light scattering.

In this context, the objective of the thesis is to provide an original approach solving the following two scientific questions:

1. *How to optically reduce the impact of light scattering on the spectroscopic signal ?*

2. *How to model the chemical absorbance of highly scattering materials ?*

The first question, addressed in following chapter, requires to invoke optical theories to act on the quality of the spectroscopic signal. Based on the principles of light polarization, we design an original optical setup, which selects light being less impacted by multiscattering. The chapter first present the theoretical aspects underlying the proposed optical approach, which is then experimentally validated on model samples in powdered form.

The second question is addressed in chapter 4. Here it is about comprehension of the information contained in the signals measured with the new method and their link with the chemical absorbance. We combine the optimized signals with the Absorption/Remission function of Dahm's Representative Layer Theory (RLT) (Dahm & Dahm, 1999) to model the absorption which becomes, in theory, linearly proportional to concentration of constituents. This approach, named PoLiS (for Polarized Light Spectroscopy), is tested on liquid and particulate model samples in the visible range.

In chapter 5, we validate the method for the estimation of Total Organic Carbon content in soils by applying it on real soil samples and benchmark the results (i.e. the prediction accuracy) with the ones achieved using empirical preprocessing.

Chapter 3

Optical methodology for reducing scattering effects on the spectroscopic signal

---

Contents

Preamble

The main problem, in quantitative analysis of highly scattering samples using Vis-NIR spectroscopy, is that multivariate calibration models built on conventional spectroscopic measurements such as Transmittance or Reflectance are adversely affected by variations arising from non-linear multiple light scattering effects. Because these variations are not necessarily related to changes in the chemical composition, it makes the extraction of chemical information from such samples challenging.

Instead of spectral pre-processing, which is commonly used by Vis – NIR spectroscopists to deal with undesirable scattering effects, this chapter presents an optical methodology to reduce multiple scattering. A new optical setup, based on polarized light spectroscopy is specifically designed to select photons that have been only weakly scattered. When tested in Visible range (400-800 nm) on powdered samples mixing scattering and absorbing particles, the set-up provides significant improvements in the capacity to predict the absorber's concentration. This optical pretreatment allows us to retrieve linear and steady conditions for spectral analysis.

Improvement of the chemical content prediction of a model powder system by reducing multiple scattering using polarized light spectroscopy[1]

3.1 Introduction

Visible - Near infrared spectroscopy (Vis-NIRS) is a well-known technique used for measuring the chemical composition of a wide variety of media and products. Although Vis-NIRS has been quoted in articles for approximately 50 years (Hart *et al.*, 1962; Massie & Norris, 1965) with this purpose, it really took off in the late 80's in agricultural and food applications (jumping from around 10 publications per year in the late 80's to 150 publications per year in the turn of century), and then in the 90's for pharmaceutical and biomedical applications. Today, it plays a major role in these sectors, as a routine laboratory method for in-vivo or in-line monitoring system. On the one hand Vis-NIRS presents several advantages: Vis-NIR extinction coefficients are small compared to mid-infrared (MIR) ones, which allows light to penetrate deeper into objects and avoids time-consuming sample preparation; Vis-NIR light scattering makes it possible to analyze bulk samples with a retro-diffusion optical configuration, thus turning it into a non-destructive technique. In addition Vis-NIR optical components are low cost and with high Signal-to-Noise Ratio (SNR). On the other hand, VIS-NIRS has several drawbacks: the VIS-NIR spectrum is poorly resolved as it is made up of scattering effects and of wide low-intensity harmonics and combinations of MIR fundamental absorption bands. Consequently, retrieving chemical information from Vis-NIR spectra is quite painstaking and requires advanced chemometrics: it is based on calibration models to be built between VIS-NIR spectra and known concentrations of a set of calibration samples. Traditionally, linear multivariate calibration methods such as Principal Component Regression (PCR) and Partial Least Square Regression (PLS) are used in Vis-NIRS. However, scattering

[1]Ryad Bendoula, Alexia Gobrecht, Benoit Moulin, Jean-Michel Roger, Véronique Bellon-Maurel, *Improvement of the chemical content prediction of a model powder system by reducing multiple scattering using polarized light spectroscopy*, Accepted in Applied Spectroscopy, June 2014

effects are troublesome in the VIS-NIR spectra of turbid media, defined by Shi and Anderson (Shi & Anderson, 2010) as "samples exhibiting multiple scattering events". Scattering can be several orders of magnitude larger than absorption and may invalidate the use of such data processing methods, which are themselves based on the underlying assumption of a linear Beer-Lambert law relationship between absorbance spectra and chemical concentration. It is therefore necessary for VIS-NIR spectroscopists working on highly scattering media to use strategies to release Vis-NIR spectra from scattering effects. The most common strategy is spectral pre-treatment. These preprocessing step is specifically designed to reduce multiplicative and additive effects caused by variations in sample physical properties (Rinnan *et al.*, 2009; Martens, 1991) . Among them, standard normal variate (SNV) often associated to detrend, multiplicative signal correction (MSC) (Geladi *et al.*, 1985), Extended MSC (EMSC) (Martens, 1991), normalization or Optical Path Length Estimation and Correction (OPLEC) (Chen *et al.*, 2006; Jin *et al.*, 2012). However, these approaches remain questionable : they consider that scattering is nearly constant over the wavelengths, which is not the case3; they may eliminate chemical-related information, which is very small with regard to scattering effects (Martens *et al.*, 2003); they are inappropriate when light scattering varies greatly from sample to sample (Steponavicius & Thennadil, 2011).

Another option is to acquire the spectrum in a way that separates the part related to absorption from the part related to scattering. Specific experimental techniques, related to the application of light propagation theory and resolution of the Radiative Transfer Equation (Shi & Anderson, 2010) have been proposed, including adding-doubling set-ups (Steponavicius & Thennadil, 2011; Prahl, 1995; Steponavicius & Thennadil, 2009), spatially-resolved spectroscopy (Farrell *et al.*, 1992), time-resolved spectroscopy (Chauchard *et al.*, 2005; Abrahamsson *et al.*, 2005b) and frequency-resolved spectroscopy (Martens, 1991).

Although powerful, these methods have their limitations, particularly when applied on highly scattering samples. First, they may require complex and sometimes expensive optical implementations, which may not be compatible with conventional spectrometers

or with highly turbid samples (for which transmission measurement is not possible). Secondly, as they rely on the estimation of absorption and scattering coefficients achieved by model inversion, parameters describing the studied medium (sample thickness, refractive index, particle size and shape...) must be known or approximated, which may be a troublesome task as they are often unknown in complex media (Steponavicius & Thennadil, 2011; Swartling *et al.*, 2003).

Whereas separating absorption and scattering from a Vis-NIR signal is still an open research issue on highly turbid samples, the main demand from Vis-NIR spectroscopists is merely for spectra with reduced impact of scattering in order to better fit Beer-Lambert's Law conditions (Hebden *et al.*, 1997; Lu *et al.*, 2006). Light polarization subtraction is a simple technique to reduce directly the effects of multi-scattering on the measured signal (Lu *et al.*, 2006; Backman *et al.*, 1999). This approach has been based on the fact that, when light interacts with matter, a small number of scattering events do not significantly modify the polarization status of the beam whereas multiple scattering leads to depolarization (Swartling *et al.*, 2003; Abrahamsson *et al.*, 2005a). Polarization subtraction technique (Hebden *et al.*, 1997; Schmitt *et al.*, 1992; Morgan & Ridgway, 2000; Demos & Alfano, 1997) was used to select light beams that retain initial polarization and which are therefore less impacted by multiple scattering events.

Although this technique has gained interest in the field of biomedical (Lu *et al.*, 2006; Backman *et al.*, 1999; Demos & Alfano, 1996), where it is used to optically target subsurface organelles (particles suspended in water) and tissues (layered samples), it is either poorly understood or not used by NIR spectroscopists working with agricultural, food, pharmaceutical and other industrial samples. To our knowledge, polarized NIRS techniques have never been applied to routine or in-line analysis to reduce scattering effects on spectra on turbid media.

In this paper, the effectiveness of this multi-scattering correction based on the polarization subtraction is evaluated using a two-component model powder system. The objectives of this paper were to assess the effect of multi-scattering correction (i) on the performances of a calibration model and (ii) on the robustness of the prediction model

built from the corrected spectra for predicting the absorber's concentration of powder samples.

3.2 Theoretical Model : Polarization subtraction

Capital bold characters will be used for matrices, e.g. $\mathbf{X}$; non bold characters will be used for column vectors, e.g. x.

Polarization subtraction technique (Schmitt *et al.*, 1992; Morgan & Ridgway, 2000; Demos & Alfano, 1997) is based on the polarization-maintaining property of weakly scattered light. When polarized light illuminates a scattering medium, weakly scattered light will emerge in its original polarization state (Backman *et al.*, 1999; Demos & Alfano, 1996; Sokolov *et al.*, 1999; Yoo & Alfano, 1989), while multiple scattered light will emerge with random polarization. In the case of linearly polarized source, the light that is remitted in the same polarization channel as the input illumination is composed by light that has maintained its original polarization state plus a component from the randomly polarized heavily scattered light (cf equation 3.1). Light that emerges in the orthogonal polarization channel contains only randomly polarized light, approximately equal to the randomly polarized component in the original polarization state (cf equation 3.2).

$$I_{\parallel}(\lambda) = \frac{\Omega}{2\pi} \cdot I_0(\lambda) \cdot S(\lambda) + \frac{\Omega}{2\pi} \cdot I_0(\lambda) \cdot \alpha(\lambda) \cdot M(\lambda) \tag{3.1}$$

$$I_{\perp}(\lambda) = \frac{\Omega}{2\pi} \cdot I_0(\lambda) \cdot \beta(\lambda) \cdot M(\lambda) \tag{3.2}$$

Where $I_{\parallel}(\lambda)$ and $I_{\perp}(\lambda)$ are the light scattered by the media with parallel and perpendicular polarization respect to the polarization of the illumination light. $I_0(\lambda)$ is the intensity of the illumination light. Ω is the collection solid angle, residual term of the integration on the solid angle (Schmitt *et al.*, 1992; Morgan & Ridgway, 2000; Demos & Alfano, 1997), of the optical device. $S(\lambda)$ and $M(\lambda)$ are the probabilities of light undergoing single and multiple scattering respectively.

Since all light undergoes scattering :

$$S(\lambda) + M(\lambda) = 1 \tag{3.3}$$

Finally, $\alpha(\lambda)$ and $\beta(\lambda)$ are the multiple light scattered ratio by the media with the parallel and perpendicular polarization respect to the polarization of the illumination light. The sum $\alpha(\lambda)$ and $\beta(\lambda)$ must be one :

$$\alpha(\lambda) + \beta(\lambda) = 1 \tag{3.4}$$

By subtracting 3.1 to 3.2, the intensity of light undergoing single scattering ($I_{ss}(\lambda)$) is equal to :

$$I_{ss}(\lambda) = I_{\parallel}(\lambda) - I_{\perp}(\lambda) = \frac{\Omega}{2\pi} \cdot I_0(\lambda) \cdot [S(\lambda) + (\alpha(\lambda) - \beta(\lambda)) \cdot M(\lambda)] \tag{3.5}$$

In conclusion, the part of the single-scattering effect in the signal is preserved and the multi-scattering effect is highly reduced.

3.3 Materials and Methods

3.3.1 Instrumentation

In the experimental setup (figure.3.1), a halogen light source (150 W, Leica Cls) was coupled with a 940 μm core diameter optical fiber of numerical aperture (N.A) of 0.25, Sedi & ATI). The light delivered by the fiber was collimated by an aspheric lens (F220SMA-B - Thorlabs). The incident beam was a 1.5 cm diameter circular spot with 1° divergence. The incident and reflected beam were polarized through two broad-band (400 nm - 800nm) polarizers (NT52-557, Edmunds Optics). Incident light was linearly polarized and reflected light was collected in a narrow cone (1°). The output from the

analyzer was coupled inside an optical fiber ($N.A = 0.25$, Sedi & ATI) by an aspheric lens (F220SMA-B - Thorlabs). This fiber was connected to a spectrometer (MMS1, Zeiss) featuring a detection range of 400 nm - 800 nm, with 3 nm resolution. A constant angle of 70^o was maintained between the excitation and collection arms. This angle was chosen to optimize intensity of the reflected beam and to avoid specular reflection.

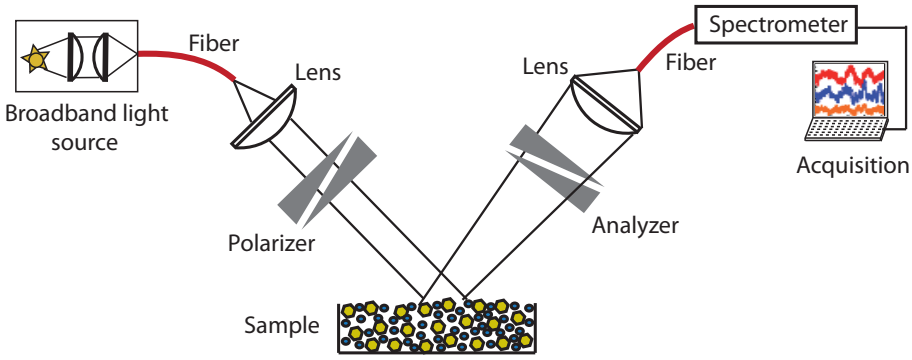

Figure 3.1: Schematic diagram of polarized light spectroscopy system.

3.3.2 Experimental design and sample preparation

Powdered samples mixing sand (Fontainebleau sand VWR International) and coloring dyes (brilliant blue FCF-E133 and chlorophyllin E141, purchased from Colorey, respectively named E133 and E141 in the text) were prepared. Two sand particle size classes were used : S_1 with a diameter less than 250 μm and S_2 with a diameter greater than 250 μm. Sand plays the role of a scattering but non absorbing matrix. One or both of the coloring dyes have been added at different densities to the sand, playing the role of absorbing substance in the mixture. Note that absorbers in powdered form also have scattering properties. Particle sizes of the coloring powders were less than 50 μm, with E133 being about three times smaller than E141.

Overall, 42 samples were prepared for spectral acquisition composing a calibration set and 12 samples were prepared, afterward and with the same procedure to create an independent test set. The range of sample's colorant densities (in $g \cdot L^{-1}$) are specified in Figure 3.2.

Each sample was directly prepared in an airtight plastic container of 100 mL by

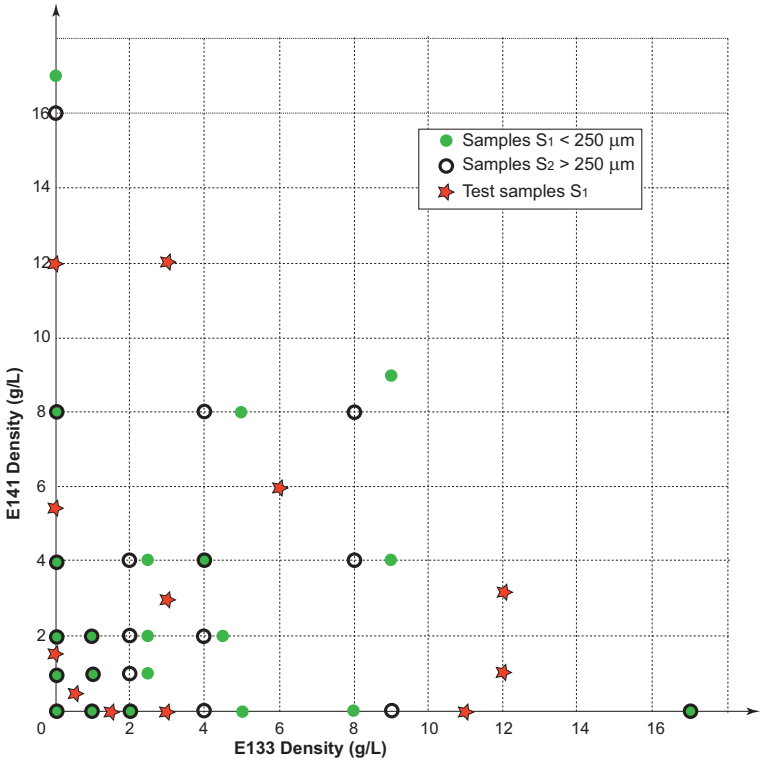

Figure 3.2: Experimental design presenting the dye densities $g \cdot L^{-1}$ of 42 samples for the calibration set and 12 samples for the independent test set.

adding the precisely weighted corresponding amount of colorant in 20 g of sand using an analytical balance (Kern 770). The maximum dye volume added was not higher than 4% of the total sand volume. Considering that the pore volume for sand is about 40%, and that the dye particles are between 5 to 10 times smaller than the sand, one can make the assumption that the dye would fill the interstices between the sand particles and therefore not increase the initial volume of sand. The density of dye (d_{dye}) in a sample was obtained from :

$$d_{dye} = \frac{m_{dye} \cdot d_{sand}}{m_{sand}} \tag{3.6}$$

with m_{dye} the added mass of dye, d_{sand} the density of sand (which differ for S_1 and S_2) and m_{sand} the mass of sand (here $20g$). The colorant density ranged from $[0 - 18\ g \cdot L^{-1}]$.

To ensure homogeneity of the mixture, the sample was agitated after preparation and again just before it was carefully transferred in an adapted 5 cm of diameter cup to get an even and horizontal surface.

3.3.3 Spectral acquisition

For each sample, light was measured with the polarized spectrometer with parallel and perpendicular respect to the polarization of the illumination light. Dark current (I_b) was recorded from all measured spectra and subtracted.

A broadband dielectric mirror (BB3-E02, Thorlabs) was used as a reference (I_0) to standardize spectra from non-uniformities of all components of the instrumentation (light source, fibers, lens, polarizer and spectrometer).

From these measurements and the equation (3.5), a raw reflectance (R_W) and a corrected reflectance (R_C), for each sample, were calculated :

$$R_W(\lambda) = \frac{(I_{\parallel}(\lambda) - I_{b\parallel}(\lambda)) + (I_{\perp}(\lambda) - I_{b\perp}(\lambda))}{(I_{0\parallel}(\lambda) - I_{b\parallel}(\lambda))} \tag{3.7}$$

$$R_C(\lambda) = \frac{(I_{\parallel}(\lambda) - I_{b\parallel}(\lambda)) - (I_{\perp}(\lambda) - I_{b\perp}(\lambda))}{(I_{0\parallel}(\lambda) - I_{b\parallel}(\lambda))} = \frac{Iss(\lambda) - (I_{b\parallel}(\lambda) + I_{b\perp}(\lambda))}{(I_{0\parallel}(\lambda) - I_{b\parallel}(\lambda))} \tag{3.8}$$

With $I_{\parallel}(\lambda)$ and $I_{\perp}(\lambda)$, the intensities of light scattered by the media with parallel and perpendicular polarization respect to the polarization of the illumination light and $I_{0\parallel}(\lambda)$ and $I_{0\perp}(\lambda)$, the intensities of light reflected by the standard mirror with parallel polarization respect to the polarization of the illumination light (as the perpendicular component emerging from the mirror is zero). $I_{b\parallel}(\lambda)$ and $I_{b\perp}(\lambda)$ are the dark current intensities recorded for each measurement.

3.3.4 Multivariate analysis

All computations and multivariate data analysis were performed with Matlab software v. R2012b (The Mathworks Inc., Natick, MA,USA).

Linear unmixing

The first step of Classical Least Square (CLS) (Geladi, 2003) was used to extract the reflectance *pure spectra* of the absorbers from the reflectance spectra of the mixtures. Also called **K**-matrix, this linear unmixing assumes that a spectrum is a linear combination of the pure component's spectra. The whole calibration set **R** (42 samples mixing the absorbers at different concentrations) and **C**, the matrix of sample components concentrations, were used to compute the linear least square estimated $\widehat{\mathbf{K}}$-matrix of the two pure active components (E133 and E141) composing **K** knowing that:

$$\widehat{\mathbf{K}} = \mathbf{R}\mathbf{C}^T(\mathbf{C}\mathbf{C}^T)^{-1} \tag{3.9}$$

Both the raw reflectance spectra ($\mathbf{R}_W$) and the corrected reflectance spectra ($\mathbf{R}_C$) were used to compute respectively $\widehat{\mathbf{K}}_w$ and $\widehat{\mathbf{K}}_c$ containing the demixed pure spectra of E133 and E141.

Calibration

Partial Least Square (PLS) (Wold *et al.*, 2001) algorithm was used to model the chemical composition of the powder mixture using $\mathbf{R}_W$ and $\mathbf{R}_C$. A general PLS model was built using the whole calibration set (42 samples) to predict the samples of the independent test set (12 samples). Secondly, to assess the robustness of the prediction models regarding sand particle size, a PLS model was built with the samples set S_2 and tested on the independent test set S_1 (figure 3.2). The number of latent variables was determined by comparing performances by leave-one-out cross-validation (Wold, 1978). Performances (R^2, Standard Error of cross-validation (SECV)) and number of latent variables of the different prediction models built with uncorrected and corrected signals of the different models were compared.

3.4 Results and discussion

3.4.1 Spectra analysis

Bulk colorant

The raw spectra ($R_W(\lambda)$) and the corrected spectra ($R_C(\lambda)$) of the pure powder colorant E133 and E141 are represented in figure 3.3 (a) and (b).

First comment is that the polarization subtraction reduces the global reflectance intensity of the measured signal (by 10 times). It is an expected result as only a small part of the signal is selected: the single-scattered one. Despite this reflectance loss, the corrected spectrum is not noisy and contains information about the sample.

Between 400 nm and 700 nm, the raw spectrum and the corrected spectrum have similar shapes. For example, the spectroscopic signature of the colorant E133 appears to be purple (as seen in powdered form), mixing a reflectance peak at 450 nm (Blue) and at 650 nm (Red). However, these peaks are more marked on the corrected spectra. For the raw spectrum, crushing peaks can be explained by a strong increase in reflectance above 750 nm. This sharp increase in reflectance, in a spectral range where the colorant does not absorb, is due to the multi-scattering. However, this effect seems to be less important for E141 than for E133. In the corrected spectra, this effect is strongly reduced (Figure 3.3 (b)).

Sand and dye mixtures

Figures 3.3 (c) and (d) show respectively the raw spectra and the corrected spectra of sand S_1 mixing coloring powder E133 at different densities. When mixed with the colorant, sand is responsible for high multi-scattering as it is not absorbing the light. This physical phenomena come of top of the chemical information contained in the spectra and masks the spectral features linked to the absorber. The shape of the raw spectra of coloring powder E133 (Figure 3.3.(a)) and the shape of the raw reflectance of the sand–dye mixture (Figure 3.3(c)) are completely different. The multiplicative effect due to

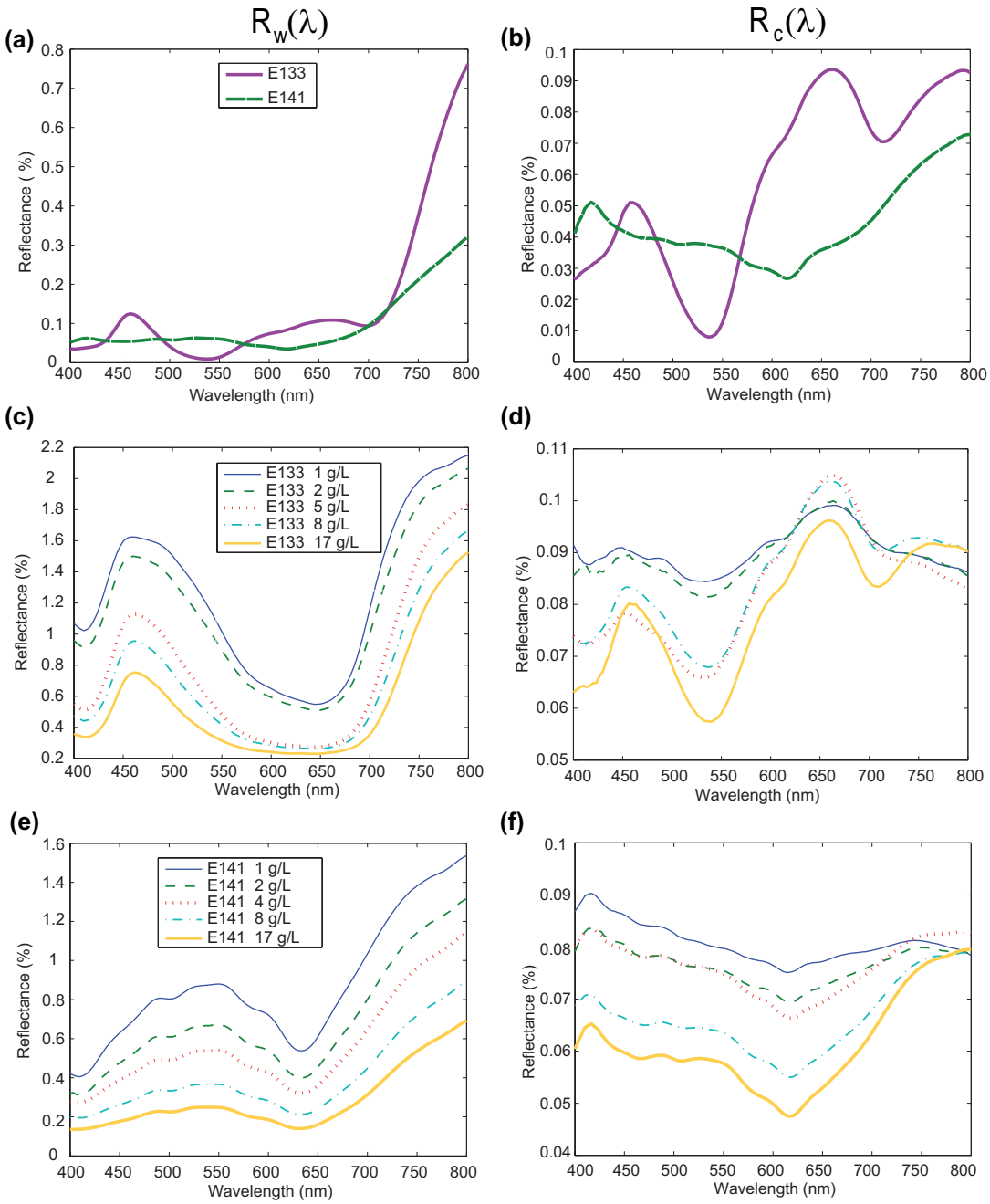

Figure 3.3: (a) Raw reflectance spectra of coloring powder E133 (b) Corrected reflectance spectra of coloring powder E133 (c) Raw reflectance spectra of sand S1 + coloring powder E133 mixed at different densities. (d) Corrected reflectance spectra of sand S1 + coloring powders E133 mixed at different densities.(e) Raw reflectance spectra of sand S1 + coloring powder E141 mixed at different densities.(f) Corrected reflectance spectra of sand S1 + coloring powders E141 mixed at different densities.

scattering is not wavelength dependent and depends on the number of scatterers in the sample. Therefore, the more dye, the more scatterers and, consequently the higher the reflectance. The order of the raw spectra is consistent with the dye's concentration, but for a physical reason. By applying the correction, the spectral features of the colorant are enhanced as it can be seen on figure 3.3 (d). The signature of the corrected spectra is similar to the spectral signal of the colorant E133 in powdered form (figure 3.3 (b)). The reflectance peaks at 450 nm and at 650 nm clearly appear, but more important, because linked to the dye's concentration, the wavelength regions where absorbance occurs (400 – 430 nm and 500 – 650 nm) are now visible. In these regions, the spectrum ordering is consistent with the dye concentration, contrary to the other areas where low absorbance occurs and reveals more complex reflectance patterns.

Figures 3.3 (e) and (f) show respectively the raw spectra and the corrected spectra of sand S_1 mixing coloring powder E141 at different densities. By comparing the raw reflectance intensities for of the sand - E133 mixtures (Figure 3.3 (c)) and sand - E141 mixtures (Figure 3.3(e)), containing the same ranges of dye's densities, it appears that the level of $R_w(\lambda)$ is two-times higher for E133 than for E141. As stated in section 3.3.2, the particle size of E141 is, at least, three-times larger than E133. This difference in particle diameter has a direct impact on the elastic scattering phenomenon occurring during light-matter interaction. First, for the same density of dye, small particles scatter more than larger particles (Backman *et al.*, 1999). Secondly, the scattering angle differ between small and larges particles: the larger the diameter, the smaller the scattering angle. Combining these two properties, the overall reflectance intensity will be higher for smaller particles, which is the case of raw reflectance of the Sand-E133 mixture.

For the corrected spectra $R_C(\lambda)$, there is no significant difference in the intensity level between sand – E133 mixtures (Figure 3.3 (d)) and sand - E141 mixtures (Figure 3.3(f)). This is coherent with the fact that the method corrects the spectra form multi-scattering, which is mainly due to the sand particles but also, and in a significant manner, to the powdered colorant.

3.4.2 Extraction of the absorber's pure spectra

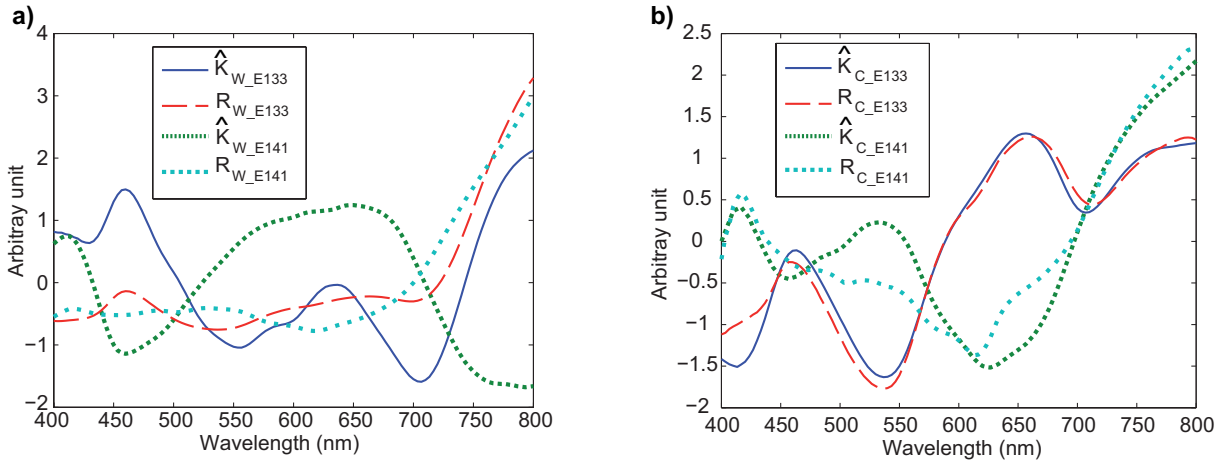

Figure 3.4: Comparison of the raw and the corrected spectra acquired on the two coloring powders ($R_{W\_E133}(\lambda)$, $R_{W\_E141}(\lambda)$ and $R_{C\_E133}(\lambda)$, $R_{C\_E141}(\lambda)$) with the demixed pure spectrum ($\widehat{K}_{W\_E133}(\lambda)$, $\widehat{K}_{W\_E141}(\lambda)$ and $\widehat{K}_{C\_E133}(\lambda)$, $\widehat{K}_{C\_E141}(\lambda)$) extracted respectively from ($\mathbf{R}_W$) and ($\mathbf{R}_C$) with Linear unmixing.

Figure 3.4 presents the demixed pure spectra $\widehat{\mathbf{K}}$ extracted from the reflectance spectra ($\mathbf{R}_W$ and $\mathbf{R}_C$). The linear unmixing (figure 3.4.a) applied to raw spectra ($\mathbf{R}_W$) provided estimated pure spectra ($\widehat{K}_{W\_E133}(\lambda)$ and $\widehat{K}_{W\_E141}(\lambda)$) that a very different from the raw spectra $R_{W\_E133}(\lambda)$ and $R_{W\_E141}(\lambda)$ measured directly on the powders. The shape of these spectra are not matching and the correlation coefficients, between $\widehat{K}_{W\_E133}(\lambda)$ and $R_{W\_E133}(\lambda)$ and between $\widehat{K}_{W\_E141}(\lambda)$ and $R_{W\_E141}(\lambda)$, are respectively equal to 0.59 and 0.74.

Linear unmixing assumes that a spectrum is a linear combination of the pure components spectra. From these results, the failure to recover the absorber's pure spectrum from the raw reflectance spectra, support the fact that interactions in the mixture are responsible of non-linearities which are directly responsible of the non-linearities in the relationship between the Absorbance and the absorbing power of the sample. It is well known (Lu *et al.*, 2006; Stockford *et al.*, 2007) that scattering and absorbance are not independent phenomenon. Scatter increases the mean free path of photons, increasing the chances of being absorbed.

On the contrary, the computation of the $\widehat{K}_{C\_E133}(\lambda)$ - matrix is successful and matches

the pure colorant corrected reflectance spectrum. Estimated pure spectra ($\widehat{K}_{C\_E133}(\lambda)$ and $\widehat{K}_{C\_E141}(\lambda)$) are very close to $R_{C\_E133}(\lambda)$ and $R_{C\_E141}(\lambda)$ measured directly on the powder and corrected. The correlation coefficients are respectively 0.98 and 0.93. As the method corrects the signal from a physical phenomenon (multi-scattering), one can affirm that the observed interactions in the raw spectra are of optical nature (and not chemical interactions). Regarding these results and this consideration, polarization subtraction correction induces, by reducing the multi-scattering effect, a better linear relationship between the light attenuation and the absorption the dyes. The recovered spectra ($\widehat{K}_{C\_E133}(\lambda)$) are corrected from the physical interactions occurring in the mixture.

3.4.3 Calibration models

General model

Table 3.1 shows the quality parameter of the prediction models of the test set absorber's densities comparing the raw and the corrected spectra.

	Cal set (n)	Test set (n)	Spectra	Predicted absorber	PLS factors	R^2	SEP_c ($\%g/l$)
General model	S1+S2 (42)	S1 (12)	$\mathbf{R}_w$	E133	8	0.75	2.52
				E141	5	0.91	1.3
			$\mathbf{R}_c$	E133	5	0.91	1.41
				E141	4	0.93	1.12

Table 3.1: Figure of merit of the calibration models

First, with the raw spectra $R_w(\lambda)$, the quality of the E133 prediction model is distinctly poorer than the quality of the E141 prediction model. The number of latent variables is much higher for E133 as well as the SEP_c. On the contrary, the figures of merit for the E141 model are good. Again, E133 and E141 behave differently. As stated before, multiscattering effect is more important when E133 is present in the mixture (with or without E141) because of a smaller particle diameter. Hence, non-linearities between absorbance and absorber's concentration are more important and the PLS model is limited in building a performant linear prediction model for E133.

When the models are built with the corrected reflectance spectra $R_c(\lambda)$, the figure of merit of the absorber's density prediction models are good and of the same level for E133 and E141. In both cases, the number of latent variables is decreasing. The improvement is less important for E141 but significant for E133. And both models have a lower standard error of prediction compared to the $R_w(\lambda)$ prediction models for the two colorants. In section 3.4.1, we stated that the proposed method mainly reduced the effect of multiscattering due to sand particles, enhancing the part of the signal related to chemical absorbance. In $R_w(\lambda)$, while this information is present, it is masked by the multi-scattering and the PLS needs more latent variable to extract this chemically related information to build a model.

Here, the different behavior of the two dyes is not obvious anymore. This agrees with the hypothesis that the correction method equalizes the mean free photon path between all the samples, regardless of the particle size and shape of the sample's constituents. For the two absorbers, the number of latent variables, which is an indicator of the complexity of the models, is still high (respectively 5 and 4 for E133 et E141 with the corrected spectra $R_c(\lambda)$) for samples mixing only two different absorbers. Theoretically, two PLS components should be sufficient. This agrees with the initial assumption that the polarization subtraction method highly reduces the multi-scattering but does not remove it completely. In addition, these results are consistent with the conclusions of section 'Extraction of the absorber's pure spectra, stating that the correction method restores, in a significant manner, the linear relationship between the spectra and the absorber's density in the powdered samples. To conclude, as the PLS models using the corrected spectra show good prediction capacities, it fulfills the assumption that even if the corrected signal intensity is highly reduced, the remaining information is of better quality in terms of signal sensitivity.

Robustness assessment

Table 3.2 presents the results of the calibration model built with samples of one particle size (S_2) and tested on samples with another particle size (S_1).

	Cal set (n)	Test set (n)	Predicted absorber	Spectra	PLS factors	R^2	SEP_c (%v/v)
Robustness assessment	S_2 (21)	S_1 (12)	E133	$\mathbf{R}_w$	5	0.69	1.61
				$\mathbf{R}_C$	4	0.94	0.71
			E141	$\mathbf{R}_W$	5	0.83	0.8
				$\mathbf{R}_C$	3	0.86	0.8

Table 3.2: Figure of merit of the calibration model built with samples of one particle size (S_2) and tested on samples with another particle size (S_1)

First, the models built with $R_W(\lambda)$ show, as previously observed, better predictions for E141 than for E133, but, in overall, lower quality than in table 3.1. This confirms a different behavior of E133 and E141, but also that the sand particle size has an effect in the quality of the predictions. A change in the physical structure of the samples usually leads to low prediction performances because of the scattering impact on the signal.

When built with the corrected spectra $R_C(\lambda)$, again, the prediction of E133 highly improves, while the gain is less significant for E141, which is also consistent with the previous conclusions. But overall, the predictions are good, confirming that the corrected spectra, composed by the single scattered part of the total reflectance signal, becomes less dependent to physical changes in the sample.

The polarization subtraction method selects by optical means only, part of information related to the powdered absorbers concentration, while discarding the unwanted effect of multi-scattering on the signal. The measured signal becomes less dependent of the particle size changes of the samples and therefore improves both quality and robustness of the prediction models.

3.5 Conclusion

This study demonstrates the effectiveness of the polarized light subtraction method, applied to a two component model powder system, which improves the performance of multivariate calibration models.

By selecting only the light which has conserved the initial polarization and therefore being less impacted by scattering events, a better linear correlation between the spectra

and the absorbers of the powder is observed. Only photons, that all have the same path length, are present. Then it's possible, by using the extracted pure spectrum from the calibration set with CLS, to have a good prediction of the absorbers concentrations in the samples.

When the corrected spectra are used to build the PLS models (more powerful than the CLS method), all the general quality parameters and the parsimony significantly improve. Although the overall signal intensity is reduced after optical correction, the remaining information in the corrected signal is sufficient and of better quality to build a good prediction model, thus meaning that the signal sensitivity increases. Inevitably there is a trade-off between making more accurate measurements and a reduction of SNR.

After the polarization correction, the measured signal becomes less dependent to physical changes (particle sizes) which also improves the robustness of the prediction models.

This "plug and play" optical method offers the potential to be easy to implement to a commercial spectrophotometry system and does not significantly increase the measurement time.

Contributions of chapter 3 and outlook

In this chapter, we showed a way of acting on the first stage of any spectroscopic analytical method: the signal formation. Therefore, we developed an original optical methodology to remove multiscattering from reflectance signals.

This method is original because it is based on light polarization principles, which are rarely implemented in NIR spectroscopy, except from some experiments conducted in the field of biomedical optics. Actually, polarization techniques are considered to present high SNR issues as the intensities of the signals are lower than for conventional spectroscopy.

So, this method overcomes these SNR issues and provides signals of improved quality. Moreover, correcting the spectrum from non-linear physical effects, the signal becomes a linear combination of the pure component spectra. This is an essential prerequisite for multivariate analysis.

To evaluate this optical approach, we have built models from the corrected spectra to predict the concentration of dyes. The models have proven better quality compared to ones built from the raw reflectance. However, according to Beer-Lambert's law, it is the absorbance which is linearly proportional to the constituents concentrations, and not the reflectance.

This raises the following questions:

- Which information has been extracted from the samples by the corrected reflectance spectra $R_C(\lambda)$?

- How are these measurements linked to the Beer-Lambert's law chemical absorbance?

In the following chapter, we will provide answers to these questions. Starting with the measurements made by the optical method (named PoLiS method) developed here, we propose a method to model the absorbance which has the same properties as the Beer-Lambert absorbance of non-scattering media.

Chapter 4

Modeling the absorbance of highly scattering materials

Contents

Preamble

The output of the PoLiS optical method, described in the previous chapter, is a reflectance signal corrected from multiscattering effects. Although this signal present a good SNR and contains relevant information related to the sample's chemical content, it only interrogates a small volume of the sample. Therefore, it is necessary to link these optical measurements to the absorbing properties of the whole sample.

We found the frame of the Representative Layer Theory (RLT) developed by Dahm & Dahm adapted to provide a link between the PoLiS measurements and the absorbing power of highly scattering samples. This chapter details the underpinning theories of this combined approach and presents the experimentation conducted to evaluate the method on scattering samples in liquid and powdered form.

To avoid the reader some redundancies between the two stand-alone chapters (chapter 3 and chapter 4), part of the introduction has been grayed. In addition, between the two chapters, the following symbols changed:

- the raw reflectance $R_W(\lambda)$ becomes the backscattered reflectance $R_{BS}(\lambda)$
- the corrected reflectance $R_C(\lambda)$ becomes the low scattered (or single scattered) reflectance $R_{SS}(\lambda)$

Combining linear polarization spectroscopy and the Representative Layer Theory to measure the Beer-Lambert's Law Absorbance of highly scattering media[1]

4.1 Introduction

Visible and Near Infrared (Vis–NIR) Spectroscopy has been widely accepted as a rapid, nondestructive analytical technique for a huge number of media and products. Today, it plays a major role in many sectors such as agricultural and food products or for petrochemicals and pharmaceuticals, as a routine laboratory, in-vivo or in-line monitoring system (Williams & Norris, 2001). The spectrometric signal is used to extract chemically related information from different materials, usually by means of chemometric modeling. This is made possible because, according to Beer–Lambert Law, absorbance is linearly related to the concentration of the chemicals composing the samples.

This ideal case occurs only in transmission measurements of low concentration in non turbid media where the derived absorbance $\{Abs = -\,log\ T\}$ is a good estimation of the Beer–Lambert law absorbance, here referred as the "absorbing power" (Dahm & Dahm, 2007). In other cases, especially when highly turbid samples are dealt with, measuring the absorbing power of samples is far from trivial. As soon as the material contains scattering centers, accounting for all the photons that have entered the sample becomes a real challenge. Some of them are absorbed, some of them reach the detector directly; some after having traveled a certain distance in the media; and, at last, some of the photons exit the sample without striking the (transmission) detector. In diffuse reflectance, the detector measures the backscattered signal R. Traditionally, a "simili-absorbance" is computed from R for an "infinitely thick" sample : $\{Abs = -\,log\ R\}$. This computed absorbance is a bad approximation of the Beer–Lambert law absorbance, because the path-length through the sample is dependent on both absorption and scatter in the

[1] Alexia Gobrecht, Ryad Bendoula, Jean-Michel Roger, Véronique Bellon-Maurel, *Combining linear polarization spectroscopy and the Representative Layer Theory to measure the Beer-Lambert's Law Absorbance of highly scattering media.* Accepted in Analytica Chimica Acta, October, 2014.

sample. This gives rise to additive and multiplicative effects, generating non-linearity in the absorbance-concentration relationship. When this phenomenon dominates the spectra formation, the chemically related absorbance can be severely overlapped by the physically related information, making the calibration step more critical.

It is therefore necessary for NIR spectroscopists working on highly scattering materials to use strategies to free NIR spectra from scattering effects. The most common strategy is spectral pre-processing, with treatments specifically dedicated to reduce multiplicative and additive effects caused by variations in sample's physical properties (Rinnan *et al.*, 2009; Martens, 1991). While they may be sufficient in some practical situations, they may not be able to integrate the whole complexity of non-linear multiple scattering effects in many situations. This may be because they consider that scattering is nearly constant over the wavelengths, which is not the case (Shi & Anderson, 2010); they may eliminate chemical-related information, which is very weak with regard to scattering effects (Martens *et al.*, 2003); they are inappropriate when sample-to-sample light scattering variations are large (Steponavicius & Thennadil, 2011). Hence, preprocessing the spectra may revert some simple variations like additive or multiplicative effects, but as scattering and absorption are not two independent phenomena (Dahm & Dahm, 2001), their effect on the spectrum can be mathematically irreversible.

Another option is to acquire the spectrum so that one can separate the signal related to absorption from the one related to scattering. Specific experimental techniques, related to the application of light propagation theory and resolution of the Equation of Radiative Transfer (ERT) (Shi & Anderson, 2010) have been proposed, including adding-doubling set-ups (Steponavicius & Thennadil, 2011; Prahl, 1995; Steponavicius & Thennadil, 2009), spatially-resolved spectroscopy (Farrell *et al.*, 1992), time-resolved spectroscopy (Abrahamsson *et al.*, 2005b; Chauchard *et al.*, 2005) and frequency-resolved spectroscopy (Torrance *et al.*, 2004). Though powerful, these methods have limitations particularly when applied to highly scattering samples. First, they may require complex and sometimes expensive optical implementation, which may not be compatible with conventional spectrometers or with opaque samples (transmission measurement may be

impossible). Secondly, as they rely on the estimation of absorption and scattering coefficients achieved by model inversion, they require knowing or approximating the parameters describing the studied medium (sample thickness, refractive index, particle size and shape), which may be critical (Steponavicius & Thennadil, 2011; Swartling *et al.*, 2003). Simpler approaches of the ERT like N-flux models (Kubelka & Munk, 1931; Thennadil, 2008; Kessler *et al.*, 2009) have been tested to separate absorption and scattering coefficients. Among them, the Kubelka-Munk theory (Kubelka & Munk, 1931) is the simplest and therefore most popular one. However, these approaches assume a continuous sample but fail when the media include spatial discontinuities such as powdered samples presenting different particles and voids (Pasikatan *et al.*, 2001; Coello *et al.*, 2008). Aware of these limitations, Dahm & Dahm (2004a) derived a more general expression of the 2-flux Kubelka-Munk equation, in the frame of plane parallel mathematics, the Representative Layer Theory (RLT) (Dahm & Dahm, 1999, 2007). The sample is, as in the K–M theory, considered as a superposition of n representative layers of thickness small enough so that there is no scatter between material in the same layer. This present the advantage that absorption and scattering occur independently in the layer and can therefore be theoretically separated. The frame of RLT has been already been used to study highly scattering materials such as milk (Bogomolov *et al.*, 2013; Dahm, 2013) or powdered mixtures (Coello *et al.*, 2008). Whereas separating absorption and scattering from Vis–NIR signal is still an open research issue on highly scattering samples, the main demand from Vis–NIR spectroscopists is *at least* to get an absorbance spectra with a reduced effect of scattering in order to better approximate Beer–Lambert conditions (Hebden *et al.*, 1997; Lu *et al.*, 2006).

In this study, we propose to use a Polarized Light Spectroscopy setup ($PoLiS$), adapted from Bendoula *et al.* (2014), to optically select photons that have undergo very few interactions with matter, i.e. photon of which paths have not been affected by multi-scattering (Bendoula *et al.*, 2014). Although light polarization has gained interest in the field of biomedical spectroscopy (Lu *et al.*, 2006; Backman *et al.*, 1999; Sokolov *et al.*, 1999) and imaging (Demos & Alfano, 1997; Arimoto, 2006), for example to optically

target subsurface organelles and tissues, it is still not used by Vis–NIR spectroscopists. We propose to combine the *PoLiS* spectral information with the *Absorption-Remission* function $A(R,T)$ defined by Dahm & Dahm (1999) in their Representative Layer Theory to compute a new absorbance spectra fulfilling Beer–Lambert law conditions. This is the aim of this paper, which first introduces the theoretical aspects underpinning this approach and second, studies experimentally its validity for scattering samples in liquid and powdered form in the 350 nm to 850 nm range corresponding to the Visible and Very-Short-Near-Infrared range (Vis-VSNIR).

4.2 Theory

4.2.1 Polarization subtraction spectroscopy

Light emitted by a source with an intensity $I_0(\lambda)$ is an electromagnetic wave vibrating in all the planes randomly, when *unpolarized*. By means of a linear polarizer, it is possible to select the light's electric field oscillation plane, either parallel or perpendicular to the plane defined by the direction of the incident and the reflected beam. After reflection, an analyzer placed before the detector makes it possible to measure the two components $I_{\parallel,\Omega}(\lambda)$ and $I_{\perp,\Omega}(\lambda)$ of the backscattered light intensity $I_{BS,\Omega}(\lambda)$, where Ω is the solid collection angle of the optical setup:

$$I_{BS,\Omega}(\lambda) = I_{\parallel,\Omega}(\lambda) + I_{\perp,\Omega}(\lambda) \tag{4.1}$$

$I_{\parallel,\Omega}(\lambda)$ is the intensity of light measured with the analyzer orientated in parallel to the polarizer. $I_{\perp,\Omega}(\lambda)$ is the light collected with the analyzer oriented perpendicularly to the polarizer.

When linearly polarized incident light penetrates a scattering medium, the remitted signal looses its initial polarization state because of the multiple scattering (including reflection) events. This is a gradual process and photons that have undergone a few scattering events maintain their initial polarization status (Stockford *et al.*, 2007; Sokolov

et al., 1999; Backman *et al.*, 1999; Demos & Alfano, 1997).

Let $I_{MS,\Omega}(\lambda)$ be the multiscattered part and $I_{SS,\Omega}(\lambda)$ the low scattered part of back-scattered light intensity $I_{BS,\Omega}(\lambda)$ such as :

$$I_{BS,\Omega}(\lambda) = I_{MS,\Omega}(\lambda) + I_{SS,\Omega}(\lambda) \tag{4.2}$$

Multiscattered light is isotropically unpolarized and half of its intensity passes through the analyzer when oriented parallel to the polarizer and the other half when oriented perpendicular. Therefore,

$$I_{\perp,\Omega}(\lambda) = \frac{1}{2} I_{MS,\Omega}(\lambda) \tag{4.3}$$

$$I_{\parallel,\Omega}(\lambda) = I_{SS,\Omega}(\lambda) + \frac{1}{2} I_{MS,\Omega}(\lambda) \tag{4.4}$$

From these relations, it is possible to select the intensity of light undergoing very few scattering events ($I_{SS,\Omega}(\lambda)$) :

$$I_{SS,\Omega}(\lambda) = I_{\parallel,\Omega}(\lambda) - I_{\perp,\Omega}(\lambda) \tag{4.5}$$

Polarization subtraction technique enables us to select the light conserving the initial polarization and therefore being less impacted by multiscattering events (Bendoula *et al.*, 2014; Stockford *et al.*, 2007; Hebden *et al.*, 1997; Schmitt *et al.*, 1992).

4.2.2 Absorbance of scattering samples

According to the *Glossary of Terms used in Vibrational Spectroscopy* compiled by John Bertie (Bertie, 2006), *Absorbance* (here abbreviated *Abs*) expressed by the Beer–Lambert law, in the case of non turbid liquids, is the product of the extinction coefficient (ε) (also called the Beer–Lambert absorption coefficient), the absorber's concentration

(c) and path length of light through the sample (dx) :

$$Abs(\lambda) = \varepsilon(\lambda)\, c\, dx \tag{4.6}$$

In transparent liquids with no scatterers, light is either absorbed (with a probability A, also called absorptance) or transmitted (T or transmittance) through the sample so that $A(\lambda) + T(\lambda) = 1$. Therefore, the absorbance value, by Beer–Lambert Law, can be expressed by the absorbance function :

$$Abs(\lambda) = -\ log\ T(\lambda) = -\ log\ (1 - A(\lambda)) \tag{4.7}$$

In the case of a scattering media, light is composed of three fractions, the absorbed (A), the transmitted (T) and the remitted (R or reflectance) one, with $A + T + R = 1$.

Hence, the absorption function can be similarly used to calculate the absorbance Abs from the measurements of R and T :

$$Abs(\lambda) = -\ log\ (R(\lambda) + T(\lambda)) = -\ log\ (1 - A(\lambda)) \tag{4.8}$$

This absorbance value is however a bad approximation of the Beer–Lambert law absorbance because in scattering samples, absorbance and scattering are not independent phenomena (Dahm, 2013). One consequence is that the relationship between absorbance and concentration is not linear anymore.

4.2.3 Absorbance of a representative layer of the sample

Based on their Representative Layer Theory, Dahm & Dahm (2007) propose an estimate for the Beer–Lambert law absorbance, which is corrected from scattering. In this theoretical approach, a sample is considered as a superposition of n plane parallel layers being representative of the sample, i.e. layers that are considered to have the same average chemical and physical properties. This layer, named Representative Layer (RL) is thin enough so that absorption, transmission and remission occur independently: a pho-

ton interacting with the representative layer can either be absorbed (with a probability a), transmitted (with a probability t) or remitted (with a probability r). The Representative Layer absorbance value computed for these fractions (a, r, t) can be considered as freed from scattering. Hence, the Beer–Lambert law absorbance can be approximated by the absorbance of the representative layer Abs_{RL} :

$$Abs_{RL}(\lambda) = -log\ (1 - a(\lambda)) \tag{4.9}$$

with a the absorbed fraction of light of a representative layer. This absorbance Abs_{RL} is linearly related to the sample's extinction coefficient, the analyte's concentration and sample thickness as a conventional absorbance value is, for non scattering samples (Dahm, 2013).

In addition, the Absorption-Remission $A(R,T)$ function (Dahm & Dahm, 1999) relates the fractions of light absorbed, remitted and transmitted by a representative layer to the spectroscopic measurements (R and T) made on the whole sample. This function is constant for any number of layers making up the sample:

$$A(R,T) = \frac{(1-R)^2 - T^2}{R} = \frac{a}{r} \cdot (2 - a - 2r) \tag{4.10}$$

In order to resolve equation 4.10 to compute the absorbed fraction of light in the representative layer a, the total reflectance R, total transmittance T and remitted fraction of the representative layer r have to be measured or approximated. This is the purpose of the next section.

4.2.4 Estimation of the PoLiS absorbance combining polarized light spectroscopy and the Representative Layer Theory

For optically thick samples, the transmitted fraction of light is null, $T = 0$. The total reflectance R can be approached by the remitted fraction of light $R_{BS,\Omega}$ computed from:

$$R_{BS,\Omega}(\lambda) = \frac{I_{BS,\Omega}(\lambda)}{I_{0,\Omega}(\lambda)} \tag{4.11}$$

with $I_{0,\Omega}(\lambda)$ the intensity if the light source.

According to Dahm, r, the remitted fraction of a representative layer of sample is, in theory, independent of any multi-scattering event. As $I_{SS,\Omega}(\lambda)$ is the intensity of light not being influenced by multiscattering, the assumption can be made that it can be used to approximate the remitted fraction of light r, by computing the related low scattered reflectance $R_{SS,\Omega}(\lambda)$:

$$r(\lambda) \approx R_{SS,\Omega}(\lambda) = \frac{I_{SS,\Omega}(\lambda)}{I_{0,\Omega}(\lambda)} \tag{4.12}$$

Considering this, equation 4.10 becomes :

$$A(R_{BS,\Omega}(\lambda), 0) = \frac{(1 - R_{BS,\Omega}(\lambda))^2}{R_{BS,\Omega}(\lambda)} = \frac{a(\lambda)}{R_{SS,\Omega}(\lambda)} \cdot (2 - a(\lambda) - 2\, R_{SS,\Omega}(\lambda)) \tag{4.13}$$

Therefore, by resolving equation 4.13, the absorbed fraction a of a representative layer can be expressed as:

$$a(\lambda) = 1 - R_{SS,\Omega}(\lambda) - \sqrt{(1 - R_{SS,\Omega}(\lambda))^2 - \frac{R_{SS,\Omega}(\lambda)}{R_{BS,\Omega}(\lambda)} (1 - R_{BS,\Omega}(\lambda))^2} \tag{4.14}$$

With $R_{SS,\Omega}(\lambda)$ and $R_{BS,\Omega}(\lambda)$ measured with $(PoLiS)$, it is possible to compute the absorbance spectrum $Abs_{Po}(\lambda)$ from equation 4.9 presenting the same properties, regarding

the Beer–Lambert Law, as the Absorbance of a representative layer, $Abs_{RL}(\lambda)$:

$$Abs_{Po}(\lambda) = -\ log\ \left(R_{SS,\Omega}(\lambda) + \sqrt{(1 - R_{SS,\Omega}(\lambda))^2 - \frac{R_{SS,\Omega}(\lambda)}{R_{BS,\Omega}(\lambda)}\,(1 - R_{BS,\Omega}(\lambda))^2} \right) \tag{4.15}$$

4.3 Material and Methods

We propose to couple a Polarized Light Spectroscopy setup ($PoLiS$) with the Representative Layer Theory (RLT) to compute a new absorbance value Abs_{Po} of highly scattering media. Experiments have been conducted on two types of scattering samples: liquid and powdered samples.

4.3.1 Samples preparation

Liquid samples

Liquid samples were composed of half-fat milk mixed with 6 different concentrations of chlorophyllin E141, a common food colouring (Colorey) : 0, 0.025, 0.050, 0.10 and 0.20 and 0.30 $g \cdot L^{-1}$.

Aliquots of 75 mL of each sample were conditioned in a beaker so that the height of liquid was about 3 cm, though optically thick.

Powdered samples

A series of 6 powdered samples was prepared mixing sand of 250 μm mean particle size (Fontainebleau sand, VWR International) with the same dye E141 in powdered form, at different concentrations. Each sample was directly prepared in a 100 mL airtight plastic container by adding the precisely weighted corresponding amount of dye in 20 g of sand using an analytical balance (Kern 770, Kern Gmbh). Considering that the dye, presenting a particle size of less than 50 μm, would fill the interstices between the sand particles and therefore not increase the total volume, the concentration of the colorant

c_{E141} in a sample was obtained from m_{E141} the added mass of dye, d_{sand} the bulk density of sand and m_{sand} the mass of sand:

$$c_{E141} = \frac{m_{E141}\ d_{sand}}{m_{sand}} \tag{4.16}$$

The colorant concentrations ranged from $[0 - 18\ g \cdot L^{-1}]$.

To ensure homogeneity of the mixture, the sample was thoroughly mixed after preparation and again just before it was carefully transferred in a layer 5 cm – diameter cup. The powder was then leveled in order to get an even and horizontal surface.

4.3.2 Instrumentation

Jasco spectrophotometer

On the liquid samples, total diffuse reflectance (R) and transmittance (T) have been measured using a double beam spectrophotometer (V670, Jasco) equipped with a 60 mm diameter integrating sphere (ISN-723, Jasco). The Jasco presented a linear photometric range of $[-2\ \ldots + 4\ Abs]$. Spectral data was collected in the wavelength region 350-850 nm at 1 nm interval. For each sample, a 1 mm path length quartz cuvette (100-QS, Hellma) was used. The baseline was measured with a white reference (Spectralon® , Labsphere) to ensure a simultaneous baseline correction during the reflectance and transmission measurements.

PoLiS setup

The *PoLiS* optical setup (figure 4.1) is composed of a halogen light source (150 W, Leica Cls) coupled with a 940 μm core diameter optical fiber of numerical aperture (N.A) of 0.25 (Sedi & ATI) . The light delivered by the fiber was collimated by an aspheric lens (F220SMA-B - Thorlabs). The incident beam was a 1.5 cm diameter circular spot with 1° divergence. The incident and reflected beam were polarized through two broad-band (400 nm - 800nm) polarizers (NT52-557, Edmunds Optics). Incident light was linearly polarized and reflected light was collected in a narrow cone (1°). The output from the

analyzer was coupled to an optical fiber ($N.A = 0.25$, Sedi & ATI) by an aspheric lens (F220SMA-B - Thorlabs). This fiber was connected to a spectrometer (MMS1, Zeiss). Spectral data were collected in the 350 – 850 nm wavelength range at 3 nm intervals, resulting in measurements at 151 discrete wavelengths per spectrum. The illumination arm was placed at the zenith so that the beam of light hit the sample perpendicularly. The collection arm was oriented at 45^o zenith angle in order to avoid specular reflections. The irradiated surface was about 1.8 cm^2.

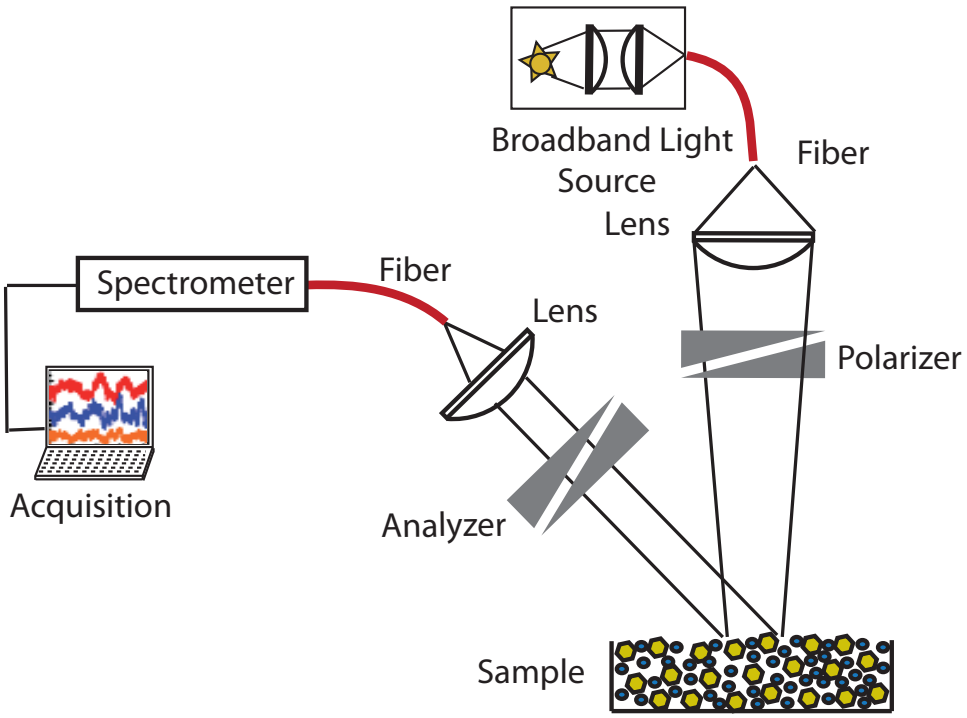

Figure 4.1: Schematic diagram of polarized light spectroscopy system (PoLiS).

4.3.3 Spectral acquisitions and computation of the absorbance

Dahm's representative layer absorbance Abs_{RL} on liquid samples

A sample of colorant E141 dissolved in distilled water (at 0.156 $g.L^{-1}$) has been measured with the Jasco V670 in transmission to computed the extinction coefficient ε_{E141} from equation 4.6.

From the diffuse reflectance (R) and transmittance (T) Jasco measurements, the three fractions of light (transmitted, reflected and absorbed) are known for samples of

1 mm thickness. Following Dahm & Dahm's procedure to compute an absorbance of a representative layer as presented in Dahm & Dahm (2013), the following set of Benford equations (4.17 – 4.19) (Benford, 1946), have been repeatedly applied to compute the three fractions (A, R and T) for a thinner and thinner layer :

$$R_{d/2} = \frac{R_d}{1+T_d} \tag{4.17}$$

$$T_{d/2} = \left[T_d\left(1-R_d^2\right)\right]^{0.5} \tag{4.18}$$

$$A_{d/2} = 1-R_{d/2}-T_{d/2} \tag{4.19}$$

with d the thickness of the sample and $d/2$ half of this thickness. And R_d, T_d and A_d the reflected, transmitted and absorbed fractions of light of a sample of thickness d.

From these iteratively computed fractions, the absorbance for each layer of thickness d/n has been computed using equation 4.8, with n the number of *representative layers* composing the sample :

$$Abs_{RL,d/n} = -log(1-A_{d/n}) \tag{4.20}$$

According to Dahm & Dahm (2013), when a minimal thickness is reached, absorbance is directly proportional to sample thickness: if the thickness is doubled, so does the absorbance, in accordance with Beer–Lambert's Law. Therefore, iterations have been stopped for the condition :

$$\frac{Abs_{RL,d/n}}{Abs_{RL,d/2n}} \approx 2 \tag{4.21}$$

PoLiS spectral acquisitions

For each type of sample (powdered and liquid), remitted light intensity was measured with the *PoLiS* setup with the analyzer set parallel ($I_{\parallel,\Omega}(\lambda)$) and perpendicular ($I_{\perp,\Omega}(\lambda)$) with respect to the polarization of the illumination light. Dark current ($I_b(\lambda)$) (i.e.

current without light) was recorded for all measured spectra and subtracted.

A diffuse reflectance white standard (Spectralon® SRS-99-010, Labsphere) was used to collect a reference spectrum ($I_{0,\Omega}(\lambda)$) to standardize spectra from non-uniformities of all components of the instrumentation (light source, fibers, lens, polarizer and spectrometer).

From these measurements and the equations (4.1) and (4.5), the backscattering reflectance ($R_{BS,\Omega}(\lambda)$) and the low scattered reflectance ($R_{SS,\Omega}(\lambda)$), has been calculated for each sample :

$$R_{BS,\Omega}(\lambda) = \frac{\left[I_{\parallel,\Omega}(\lambda) - I_{b\parallel}(\lambda)\right] + \left[I_{\perp,\Omega}(\lambda) - I_{b\perp}(\lambda)\right]}{\left[I_{0,\Omega}(\lambda) - I_{b0}(\lambda)\right]} \tag{4.22}$$

$$R_{SS,\Omega}(\lambda) = \frac{\left[I_{\parallel,\Omega}(\lambda) - I_{b\parallel}(\lambda)\right] - \left[I_{\perp,\Omega}(\lambda) - I_{b\perp}(\lambda)\right]}{\left[I_{0,\Omega}(\lambda) - I_{b0}(\lambda)\right]} \tag{4.23}$$

From the measurements performed with the Jasco on the liquid samples and the PoLiS setup on both type of samples, different absorbance spectrum have been computed and compared :

- $Abs_{RL}(\lambda)$, the absorbance of the representative layer of liquid samples computed from its absorbed fraction of light ($A_{d/n}$), obtained from the Jasco measurements (c.f. equation 4.20). ;

- $Abs_{BS}(\lambda)$, the absorbance computed from the total backscattered reflectance signal $R_{BS,\Omega}(\lambda)$ measured with PoLiS. $Abs_{BS}(\lambda) = -\,log\, R_{BS,\Omega}(\lambda)$;

- $Abs_{Po}(\lambda)$, the PoLiS absorbance computed from $I_{\parallel,\Omega}(\lambda)$ and $I_{\perp,\Omega}(\lambda)$, measured with PoLiS and implemented in equations 4.22 and 4.23, to retrieve a, the absorbed fraction of a representative layer of the samples using equation 4.15.

4.4 Results and discussion

4.4.1 E141 extinction coefficient $\varepsilon_{141}(\lambda)$

The transmission measurement performed with the Jasco laboratory spectrometer of a sample mixing E141 coloring dye in low concentration with distilled water, allowed us to compute the E141 extinction coefficient $\varepsilon_{141}(\lambda)$ from equation 4.6, with the assumption that there is no scattering in a low concentrated sample. Figure 4.2 shows the spectral signature of the extinction coefficient $\varepsilon_{141}(\lambda)$ over the wavelength range 250 – 800 nm. The dye shows two absorbance peaks at 405 nm and 630 nm.

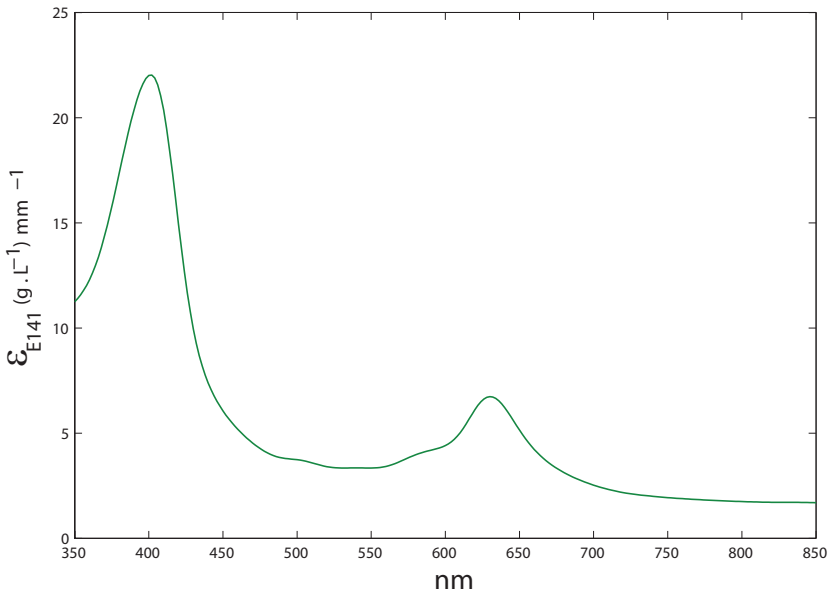

Figure 4.2: Extinction coefficient $\varepsilon_{141}(\lambda)$ of E141 dye obtained from the collimated transmittance measured with the Jasco on a sample having low concentration

4.4.2 Liquid samples

Comparison of the different absorbance signals

Figure 4.3 shows the three different absorbance signals computed according to the different types of measurements made on the milk + dye samples: $Abs_{RL}(\lambda)$, $Abs_{BS}(\lambda)$ and $Abs_{Po}(\lambda)$. For $Abs_{RL}(\lambda)$, the number of representative layers composing the sample has been set at $n = 256$ for which the condition of equation 4.21 is fulfilled.

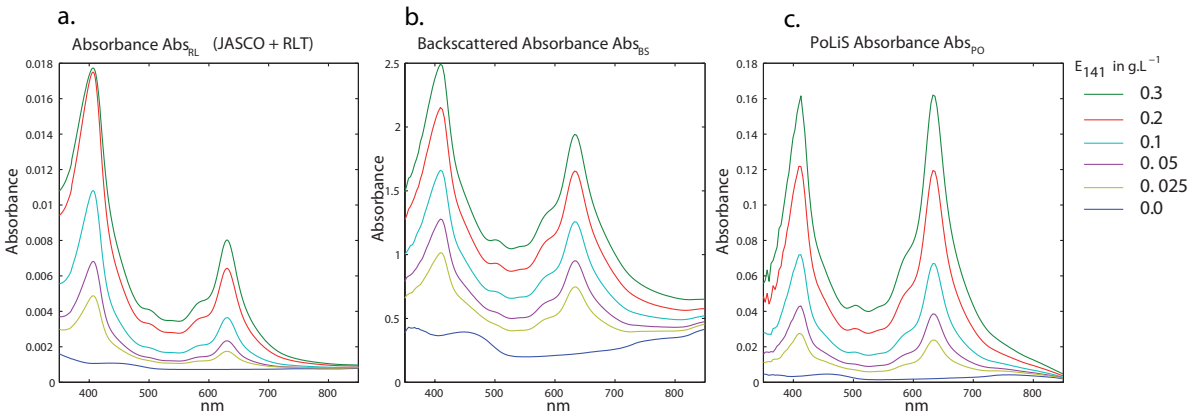

Figure 4.3: Absorbance spectra of milk + E141 sample; a. $Abs_{RL}(\lambda)$ computed from the Jasco measurements combined with the Representative Layer Theory (RLT); b. $Abs_{BS}(\lambda)$ computed from the backscattered reflectance measured with the PoLiS setup and c. $Abs_{Po}(\lambda)$ computed from the backscattered and low scattered reflectance measured with PoLiS and combined with the RLT.

The absorbance level of the three different absorbance signals computed can not be compared in their absolute value since the absorbance intensity depends on the path length traveled by the photons in the sample. For $Abs_{RL}(\lambda)$ this path length corresponds to the thickness of the representative layer. Figure 4.3 a. shows the absorbance for a RL of $1000/256 = 3.9\ \mu m$ since the samples have been measured in a 1 mm cuvette. This value is in accordance with the average size (3 - 4 μm) of fat globules in milk, which are responsible of scattering (Cabassi *et al.*, 2013).

$Abs_{RL}(\lambda)$ (figure 4.3 a.) show similar spectral features to those of $\varepsilon_{141}(\lambda)$ (c.f. figure 4.2), with two narrow peaks at 405 nm and 630 nm. By measuring both the transmittance and the reflectance with the Jasco spectrometer, the goal is to collect all the photons interacting with the sample within the integration sphere, including those scattered by the fat globules present in milk (Cattaneo *et al.*, 2009). However, a baseline is present, even when there is no colorant added to the milk. This baseline can be explained by a non negligible loss of photons (in transmission for example), which leads to overestimate the absorbance over the whole studied wavelength range. In addition we observe that for the high concentrated sample ($c_{E141} = 0.3\ g.L^{-1}$), at 405 nm, the absorbance value is lower than expected. The Absorbance value of the whole sample (1 mm thick) can be

retrieved from $Abs_{RL}(\lambda)$:

$$Abs_j(\lambda) = n \cdot Abs_{RL}(\lambda) \tag{4.24}$$

with n the number of layers composing the sample. Hence, for $n = 256$, the Absorbance $Abs_j = 4.5$ Abs at 405 nm which is just outside the upper limit of the photometric range of the Jasco (section 4.3.2), which can explain the ceiling reached by this sample. But inside the linear photometric range of the Jasco, $Abs_{RL}(\lambda)$ can be considered as a reliable reference measurement of the absorbing power of the liquid samples studied.

For the absorbance computed from the PoLiS measurements ($Abs_{BS}(\lambda)$ and $Abs_{Po}(\lambda)$), the path length is not known, although it is supposed to approximate the mean particle size for $Abs_{Po}(\lambda)$ (Dahm, 2013). However, a visual analysis can be carried out on the shapes of these different spectra to compare with the Jasco absorbance (figure 4.3).

The absorbance $Abs_{BS}(\lambda)$ computed from only the backscattered diffuse reflectance $R_{BS,\Omega}(\lambda)$ (figure 4.3 b.) shows the same absorbance peaks at 405 nm and 630 nm. The overall shape is also similar to the shape of E141 colorant without milk. It is known that diffuse reflectance measurements do provide relatively coherent information about the studied material. However, compared to the Jasco measurements, here considered as the laboratory reference measurement, some qualitative differences can be observed: an important base line, larger peaks and an overall intensity dynamic (i.e. the difference between absorbing and non absorbing zones) which is reduced compared to the one seen for the Jasco measurements. This is typical of scattering which increases the light path length, especially in the non or low absorbing wavelength ranges. The longer the path in the medium, the higher the probability of the photon to be absorbed. This results in larger peaks in the highly absorbing ranges and an increase of the absorbance level in the low absorbing ranges.

The PoLiS absorbance $Abs_{Po}(\lambda)$ spectrum (figure 4.3 c.) also presents two distinct peaks at 405 nm and 630 nm. The baseline is highly reduced as the absorbance for raw milk is close to zero. Compared to $Abs_{BS}(\lambda)$, $Abs_{Po}(\lambda)$ shows narrower absorbance

features. The absorbance intensity at 405 nm is lower than expected for all concentrations (comparing the the higher peak with the Jasco). This can be a direct consequence of the computation of $Abs_{Po}(\lambda)$. One hypothesis is that at 405 nm, where the absorbance is high, the $R_{SS}(\lambda)$ component is less scattered and therefore, underestimated compared to 630 nm. At 405 nm, the light is rapidly absorbed by the absorbing liquid before reaching any scattering center. At 630 nm, as the absorbing power is lower, the light is more able to reach a scattering center and can be remitted. However, this hypothesis would need further investigations to be confirmed.

The visual inspection of the three different types of spectra shows that with the absorbance obtained with the PoLiS optical setup combined to the Absorption – Remission function of the representative layer theory, it is possible to retrieve a signal less impacted by multiscattering than the one of mere reflectance and therefore better revealing its chemically related information.

Does the linearity with the concentration improves ?

In Beer–Lambert law theory, absorbance is linearly related to the concentration of the absorber, the optical path traveled by the photons and the extinction coefficient (c.f. equation 4.6). The latter is the same for all the samples as milk is not absorbing in this wavelength range and only one absorber (E141) has been added. For $Abs_{Po}(\lambda)$, the optical path length can also be considered as constant for all the samples, as the PoLiS setup selects the photons that have been weakly scattered, by the scatterers contained in the superficial layer. Therefore, concentration in E141 (C_{E141}) is the only changing parameter and should linearly affect the absorbance. Figure 4.4 shows the degree of linearity between the absorbance value at 405 nm and 630 nm and the dye concentrations for all the absorbance computed in this experiment: the Jasco absorbance $Abs_{RL}(\lambda)$, the backscattered absorbance $Abs_{BS}(\lambda)$ and finally the PoLiS absorbance $Abs_{Po}(\lambda)$.

At 630 nm, the absorbance $Abs_{RL}(\lambda)$ of the representative layer computed from the Jasco measurement is linearly related with E_{141} concentration, with a Pearson's coefficient of more than 0.99 (figure 4.4 a.). However, at 405 nm, all the samples are lined up besides

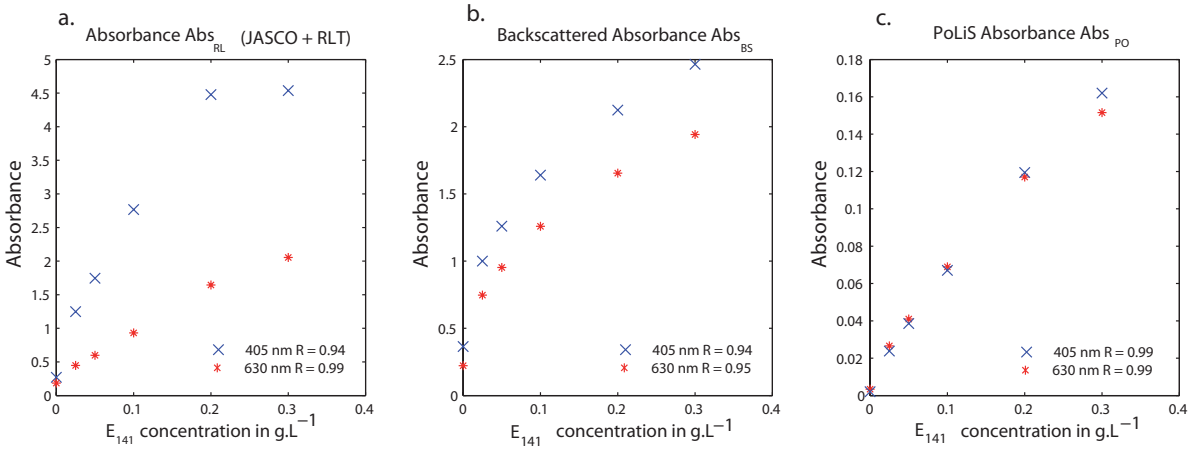

Figure 4.4: Absorbance at 405 nm and 630 nm of $Abs_{RL}(\lambda)$ computed from the Jasco measurements (a.), $Abs_{BS}(\lambda)$ (b.) and $Abs_{Po}(\lambda)$ (c.) computed form the PoLiS measurements vs. the concentration of E141 in $g.L^{-1}$

the highly concentrated sample which absorbance value is clearly reaching a limit as already observed in figure 4.3 (a.).

The backscattered absorbance $Abs_{BS}(630)$ show a less linear relation with the concentration, with a Pearson's coefficient of 0.94. Although the shape of the backscattered absorbance spectra $Abs_{BS}(\lambda)$ appeared very similar to the extinction coefficient $\varepsilon_{E141}(\lambda)$, the multiscattering is responsible of a certain degree of non-linearity as shown in figure 4.4 (b.).

The Pearson's coefficient between the PoLiS absorbance $Abs_{Po}(630)$ and the concentration is higher than 0.99, comparable to the performances of the Jasco measurements. Figure 4.4 (c.) shows the good alignment of the points for both wavelengths ($\lambda = 405nm$ and $\lambda = 630nm$) and they nearly pass through zero. The improvement of the linearity between $Abs_{Po}(\lambda)$ and the concentration of the dye is due to the fact that the optical setup architecture of PoLiS allowed us to select only the weakly scattered photons which have traveled a short distance in the matrix. From a spectroscopic point of view, this photon path is similar to a transmission measurement with no (or very few) scattering, hence approaching the ideal conditions of Beer–Lambert's law.

4.4.3 Powdered samples

Powdered samples used in our experiment are highly scattering and absorbing samples. Hence it is not possible to perform the transmission measurement necessary to compute a reference absorbance value on a representative layer, as it has be done on liquid samples with the Jasco. The backscattered absorbance $Abs_{BS}(\lambda) = -logR_{BS}(\lambda)$ signal is usually the one used in multivariate analysis, even knowing that it can be strongly affected by multiscattering effects and therefore far from the Beer–Lambert law conditions. Here, the performance of the PoLiS method on powdered samples is assessed by comparing the backscattered absorbance $Abs_{BS}(\lambda)$ and the PoLiS absorbance $Abs_{Po}(\lambda)$ signals (figure 4.5), both computed from PoLiS measurements $I_{\parallel}(\lambda)$ and $I_{\perp}(\lambda)$, respectively implemented in equations 4.22 and 4.23 .

For both absorbance spectra, the characteristic spectral features of E141 are present, with peaks at 405 nm and 630 nm. For the raw absorbance $Abs_{BS}(\lambda)$ spectra (fig. 4.5 a.), the peaks are large, which is characteristic of the multiscattering occurring in the samples. On the contrary, the shape for the PoLiS absorbance $Abs_{Po}(\lambda)$ spectra (fig. 4.5 c.) are very close to the spectral signature of the colorant E141 characterized by $\varepsilon_{E141}(\lambda)$ (figure 4.2). At 405 nm and 630 nm, the peaks are narrow. More, the baseline is highly reduced – though not completely removed. On the contrary to what we observed on the liquid samples, the absorbing peak at 405 nm is higher than at 630 nm and more consistent with $\varepsilon_{E141}(\lambda)$. This confirms the assumption made in section 4.4.2 that part of the photons are absorbed in the continuous phase, and therefore the low scattering component $R_{SS}(\lambda)$ in highly absorbing regions (around 405 nm) is slightly underestimated, and consequently so does the absorbance. In particulate samples, absorbance occurs within the particles as they are surrounded by voids filled with air, which do not absorb. The single scattered component of the remitted light do reach the detector, even in the region where absorbance is high.

To validate that the PoLiS absorbance $Abs_{Po}(\lambda)$ is a better approximation of the Beer–Lambert absorbance than the backscattered absorbance $Abs_{BS}(\lambda)$, the degrees of

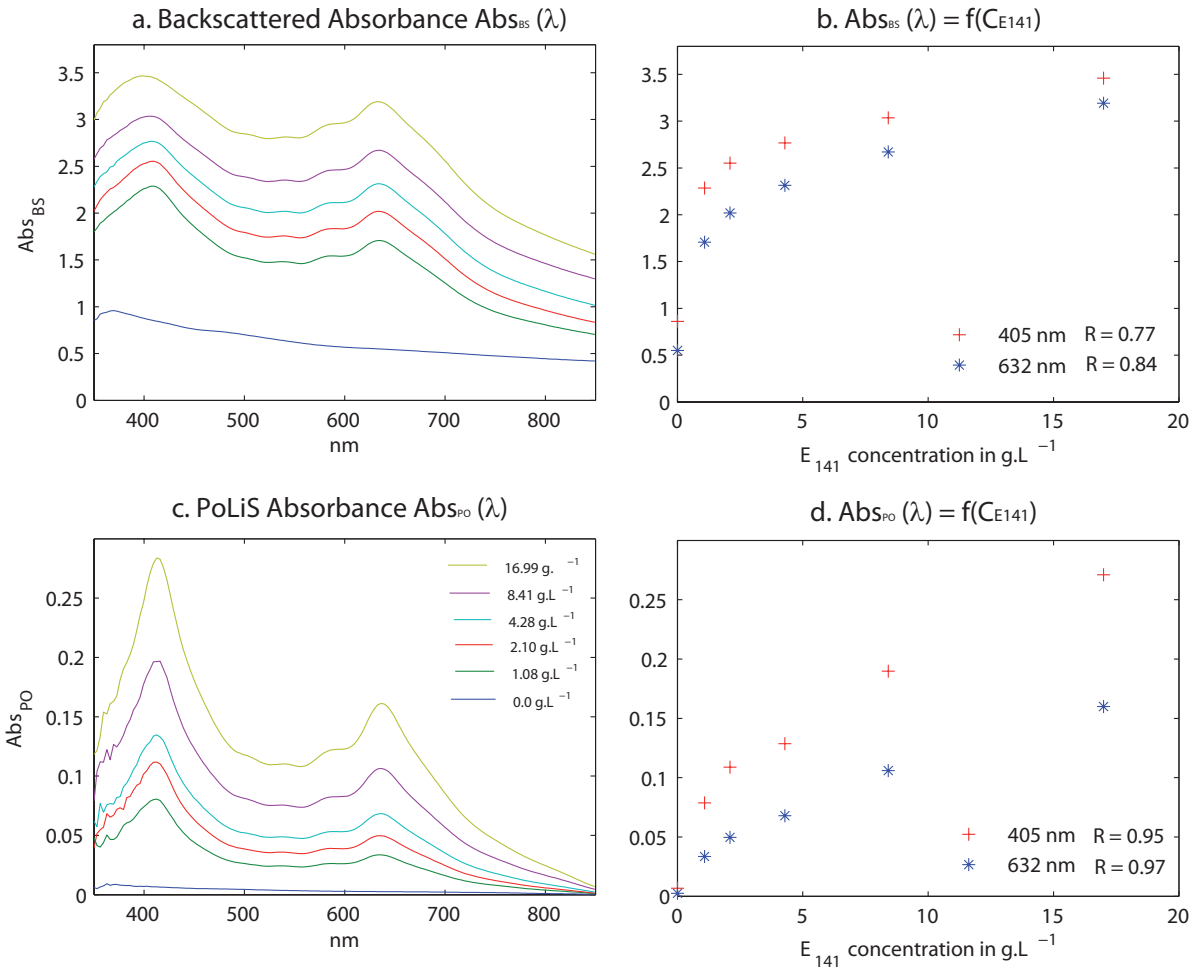

Figure 4.5: Backscattered absorbance spectra $Abs_{BS}(\lambda)$(a.) and PoLiS absorbance $Abs_{Po}(\lambda)$ (c.) of powder samples mixing sand with E141 at different concentrations. Relationship between E141 concentration and the absorbance level at 405 nm and 630 nm for $Abs_{BS}(\lambda)$ (b.) and $Abs_{Po}(\lambda)$ (d.)

linearity of the relationship between the dye concentration (C_{E141}) and the two computed absorbance $Abs_{BS}(\lambda)$ and $Abs_{Po}(\lambda)$ have been compared at $\lambda = 405nm$ and $\lambda = 630nm$ (respectively Figure 4.5 b. and d.). At 405 nm and 630 nm, $Abs_{Po}(\lambda)$ show a better linear correlation with concentration than Abs_{BS} as highlighted by the Pearson's coefficient. It is known that a mixture of particles of different sizes (here 250 μm for the sand and less than 50 μm for the dye) may produce a non linear behavior when absorbance is measured in reflectance mode, as it is here for Abs_{BS} (Dahm & Dahm, 2007). This can explain the curve in figure 4.5 (b.). The fact that PoLiS absorbance is linear with the concentration shows that the effect of differences in particle size is negligible. We find ourselves in conditions close to those of representative layer conditions where the effect of particle size has no foundation (Dahm & Dahm, 2007). This helps to validate our hypothesis of using the polarization components, $R_{\parallel}(\lambda)$ and $R_{\perp}(\lambda)$ to solve the Absorption/Remission function A(R,T).

By selecting the weakly scattered photons, the PoLiS method homogenizes the photon path–lengths and lessens the multiplicative and additive effects, leading to a more accurate signal. The computed absorbance value $Abs_{Po}(\lambda)$ combining the PoLiS spectral measurements with the representative layer theory, approximates, also for powdered samples, an absorbance which is more linearly related to the extinction coefficient (ε), the dye's concentration (c) and the path length (dx).

4.5 Conclusions

In turbid or particulate samples, scattering is strongly and negatively impacting the spectroscopic signature when it comes to extract chemically relevant information, i.e. the absorbance. Preprocessing methods are used to reduce the scattering effect the spectra, but they sometimes fail because the effect of multiscattering is deep and complex and can not always be revert mathematically. This is why it is crucial to measure a signal as free as possible from scattering effects prior calibration.

In this work, we propose a method which can be used as an alternative to select,

by optical means only, the spectral information related to chemical absorbance of highly scattering samples, while discarding the unwanted effect of multi-scattering on the signal. Combining the PoLiS signal with the Representative Layer Theory (RLT) we computed an absorbance spectra being a reasonable approximation of the Beer–Lambert absorbance of a representative layer of the sample.

Applied on liquid and powdered samples, the results confirmed the following assumptions:

- The backscattered reflectance $R_{BS}(\lambda)$ measured with PoLiS on infinitely thick samples is a good approximation of total reflectance of samples of infinite optical thickness;

- The low scattered reflectance $R_{SS}(\lambda)$, measuring the light being weakly scattered, is a satisfying approximation of the remitted fraction r of a representative layer;

- The combination the PoLiS measurements with the Representative Layer Theory allows us to compute a good estimation of an absorbance spectrum, $Abs_{Po}(\lambda)$, being freed from scattering effects and peaks of which are linearly related to the concentration of the absorber.

The PoLiS optical setup allows to perform reflectance measurements, on optically thick samples, which presents a real advantage comparing to methods like Kubleka-Munk, which need differently prepared samples (different thickness for example) or different optical setups (Transmission and Reflectance).

PoLiS method presents a high potential to increase the diffuse reflectance signal quality on highly scattering media, in solid form (soils, waste, pharmaceutical tablets) or liquid form (algae, sludge). Building calibration models using high quality signals will increase their overall quality and robustness.

Contributions of chapter 4 and outlook

This chapter presents an original approach to meet the challenge of modeling the absorbance of highly scattering materials. This approach combines optimized optical measurements and the theoretical concept of the Representative Layer.

The theoretical framework of the Representative Layer proved to be useful to establish our approach. The RLT models a sample as a series of identical layers and is based on discontinuous theories which are more appropriate to describe and to understand the optical properties of highly scattering and absorbing samples (Dahm & Dahm, 2004b). Here, the signals measured by PoLiS found their counterpart in the expression of Absorption/Remission function which could then be solved to compute a, the absorbed fraction of light by the representative layer. From this absorptance, the discontinuous mathematics allow us to derive the absorption of the whole sample, provided that it is homogenous.

At this point, only simple mixtures (in liquid and powdered form) were considered. The next chapter intends to validate the results of chapter 4 in a more complex application, which is the objective of this thesis: prediction of total organic carbon content (TOC) in soils.

The main interrogations are:

- Is the PoLiS method adapted to highly complex materials such as soils? In other terms, does the polarized light have the same behavior after interacting with soils?

- Are the measured signals of sufficient intensity and quality to be processed?

- Does the PoLiS method improve the linearity between the absorbance and a more

complex variable of interest such as TOC ?

- Is the TOC calibration model of better quality when built with the PoLiS Absorbance ? And does the PoLis method present an added value compared to preprocessing methods?

Chapter 5

Application of the PoLiS method to predict soil carbon content

---

Contents

Preamble

Soil are highly scattering and absorbing samples. Scattering is due to particles, which are large compared to the wavelength and which can either exclusively scatter (quartz for example) or both scatter and absorb (organic-mineral complex). To add complexity, the analyte of interest, here the carbon content, is itself a complex chemical parameter to be studied in soils. In this chapter, we test the PoLiS method (described in chapter 3 and 4) on a set of real soil samples to predict Total Organic Content. First, the ambition is to validate the whole PoLiS process in a more complex situation and also to confirm the assumption that the quality of the calibration models directly depend on the quality of the spectroscopic signals. Therefore, we benchmarked the models built with the PoLiS spectra against models built from classically preprocessed spectra.

Improvement of soil carbon content prediction by reducing multiple scattering using polarized light spectroscopy[1]

5.1 Introduction

Although Visible and Near Infrared Spectroscopy (Vis–NIRS) is becoming a very popular analytical technology in soil science, it is still steps away from being used as a routine analytical tool, both in field and laboratory. One of the reasons is that calibration models lack of robustness as soon as influence factors, which are numerous in soils, interfere. One of the main issues is that soils are highly scattering materials. As a direct consequence, the measurement conditions are far from the ideal conditions stated by Beer-Lambert's law where the absorbance should be linearly related to the chemical concentration (Gobrecht *et al.*, 2014a). Light scattering depends on the physical structure of the soil samples and directly contributes to the shape of the measured spectrum by hiding (or overlapping) the chemically related information. The absorbance at wavelength λ is not linear with concentration and there is a real contradiction in building calibration models based on linear multivariate methods such as the commonly-used Partial Least Squares Regression (PLS). Overcoming this signal quality issue is of great interest because the accuracy of the prediction is directly related to the quality of the measured signal (MacDougall & Crummett, 1980).

The most common strategy to reduce scattering effects is spectral pretreatment. This preprocessing step is specifically designed to reduce multiplicative and additive effects caused by variations of physical properties (Rinnan *et al.*, 2009; Martens, 1991). Among them, standard normal variate (SNV) often associated to detrend (Barnes *et al.*, 1989), multiplicative signal correction (MSC) (Geladi *et al.*, 1985), Extended MSC (EMSC) (Martens, 1991), normalization or Optical Path Length Estimation and Correction (OPLEC) (Chen *et al.*, 2006; Jin *et al.*, 2012). However, these approaches remain questionable: they

[1] Alexia Gobrecht, Ryad Bendoula, Jean-Michel Roger, Véronique Bellon-Maurel, *Improvement of soil carbon content prediction by reducing multiple scattering using polarized light spectroscopy*, submitted in Soil and Tillage Research, October 2014.

consider that scattering is nearly constant allover the wavelengths, which is not the case (Shi & Anderson, 2010); they may eliminate chemical-related information, which is very small with regard to scattering effects (Martens *et al.*, 2003); they are inappropriate when light scattering varies greatly from sample to sample (Steponavicius & Thennadil, 2011). As a consequence, the model may sometimes fail when applied on a new set of samples.

Another option is to acquire the spectrum in a way that separates the part related to chemical absorption from the part related to scattering. Specific experimental techniques, based on the application of the light propagation theory or resolution of the Radiative Transfer Equation (Shi & Anderson, 2010) have been proposed: adding-doubling set-ups (Steponavicius & Thennadil, 2011; Prahl, 1995; Steponavicius & Thennadil, 2009), spatially-resolved spectroscopy (Farrell *et al.*, 1992), time-resolved spectroscopy (Chauchard *et al.*, 2005; Abrahamsson *et al.*, 2005b) and frequency-resolved spectroscopy (Martens, 1991). Although powerful, these methods have their limitations, particularly when applied on highly scattering samples. First, they may require complex and sometimes expensive optical implementations, which may not be compatible with conventional spectrometers or with highly scattering samples (for which transmission measurement is not possible). Secondly, as they rely on the estimation of absorption and scattering coefficients achieved by model inversion, parameters describing the studied medium (sample thickness, refractive index, particle size and shape...) must be known or approximated, which may be a troublesome task as they are often unknown in complex media (Steponavicius & Thennadil, 2011; Swartling *et al.*, 2003).

Bendoula *et al.* (2014) proposed to combine light polarization and VIS-NIR reflectance spectra acquisitions. The Polarized Light Spectroscopy (PoLiS) method is an original technique to reduce directly the effects of multi-scattering on the measured signal by using the wave theory of light (Lu *et al.*, 2006; Backman *et al.*, 1999). When linearly polarized light interacts with a scattering material, the backscattered light progressively looses its initial polarization and oscillates randomly in all the planes. Using the principle of polarization subtraction, Bendoula *et al.* (2014) measured a reflectance spectra that has been less impacted by multiscattering. In Gobrecht *et al.* (2014b), the signals measured

with the PoLiS method have been processed in the frame of Dahm's Representative Layer Theory (Dahm & Dahm, 2007) to propose a model of the absorbing power. The method has been successfully tested on model particulate samples (sand + dye) showing that the newly computed absorbance signal is more linearly related to the concentration of dye in the sample.

The aim of this study is to test the PoLiS method on real soil samples to predict Total Organic Carbon (TOC) content in order to:

- validate that PoLiS absorbance measured on soil samples is more linearly related to TOC ;
- evaluate the benefit of using the PoLiS absorbance in TOC calibration models ;
- compare this "optical" preprocessing method to commonly used mathematical preprocessing methods.

5.2 Material and Methods

5.2.1 Instrumentation

The PoLiS optical setup, schematized in figure 5.1, was composed of a halogen light source (150 W, Leica Cls) coupled with a 940 μm core diameter optical fiber (Numerical Aperture $N.A = 0.25$, Sedi & ATI). The light delivered by the fiber was collimated by an aspheric lens (F220SMA-B - Thorlabs). The incident beam was a 1.5 cm diameter circular spot with 1° divergence. The incident and reflected beam were polarized through two broad-band (400 nm–800 nm) polarizers (NT52-557, Edmunds Optics). Incident light was linearly polarized and reflected light was collected in a narrow cone (1°). The output from the analyzer was coupled to an optical fiber ($N.A = 0.25$, Sedi & ATI) by an aspheric lens (F220SMA-B - Thorlabs). This fiber was connected to a spectrometer (MMS1, Zeiss). Spectral data were collected in the 400 – 800 nm wavelength range at 3 nm intervals, resulting in measurements at 121 discrete wavelengths per spectrum. A

constant angle of 70^o was maintained between the incident and reflecting arms. This angle was chosen to optimize intensity of the reflected beam.

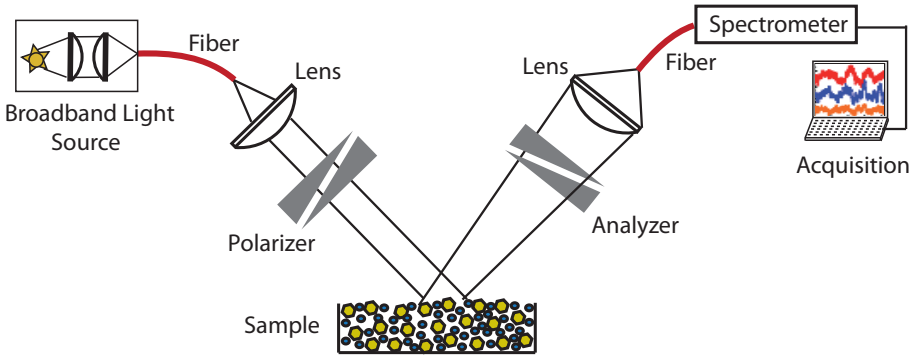

Figure 5.1: Schematic diagram of polarized light spectroscopy system (PoLiS).

5.2.2 Soil samples

The 52 studied soil samples, provided by Irstea EMGR research unit are a subset of a soil sample collection used in a previous research work published in Saenger *et al.* (2013). The samples were collected in the Vercors High Plateau Natural Reserve (VHPNR) a protected mountainous area in the French calcareous Prealps (44°97N - 5°42E). Soils of the VHPNR developed on Urgonien limestones and are generally neutral or basic. They comprise humiferous and very shallow Cambisols, Leptosols, Umbirsols and Anthroposols (FAO/IUSS/ISRIC 200). Detailed information on vegetation and soil types of the study area are provided in Saenger *et al.* (2013). The samples were collected from the Topsoil (0-10 cm) from the A horizon (Organo-mineral layer). The litter layer, when present, was removed prior to sampling.

After collection, soil samples have been air dried and stored at 4°C until chemical and spectral analysis. Total Organic Carbon was measured by dry combustion after decarbonation according to NF ISO 10694, using a N/C-Analyzer (Thermo Scientific, FLASH 2000 NC Analyzer, France) (Table 5.1).

Each sample has been prepared to get different particle sizes, namely:

- The *Coarse* form obtained by hand crushing the air-dried soil to get aggregates

smaller than 5 mm. This preparation conducted to a large variety of particle and aggregate sizes within and between samples, depending on the type of soil;

- The *Sieved* form at 2 mm, which is the classical soil preparation prior to spectral acquisition;

- The finely *Ground* form at 0.25 mm.

Each sample was carefully transferred in an adapted 5–cm diameter petri dish and moved in circles to get an even and horizontal surface before spectral analysis.

n	Mean	SD	Min	Q1	Q2	Q3	Max	Skewness
52	88.6	48.04	11.4	50.20	88.75	115.0	248.0	0.86

Table 5.1: Total Organic Carbon ($g.kg^{-1}$) descriptive statistics for the whole dataset. Q1, Q2 and Q3 correspond respectively to the first quartile, the median and the upper quartile. SD: standard deviation

5.2.3 PoLiS spectral acquisition

Each sample was illuminated with linearly polarized light and the remitted light intensity was measured with the *PoLiS* setup with the analyzer set respectively parallel, $I_{\parallel}(\lambda)$, and perpendicular, $I_{\perp}(\lambda)$, with respect to the polarization of the illumination light (Figure 5.2). Dark current, $I_b(\lambda)$, i.e. signal without light, was systematically recorded for all measured spectra with the same optical configuration and subtracted to each measurement.

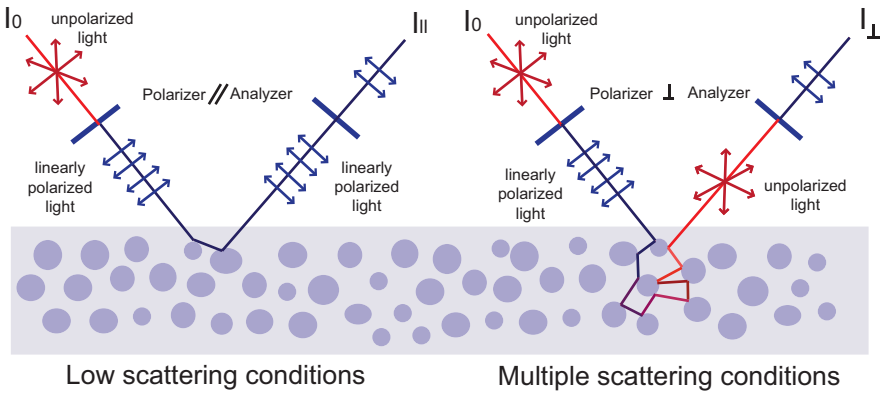

Figure 5.2: Principle of the measurement of the two components $I_{\parallel}(\lambda)$ and $I_{\perp}(\lambda)$ of the totally backscattered light by means of linear light polarization

A diffuse reflectance gray standard (Spectralon® SRS-60, Labsphere) was used to collect a reference spectrum, $I_0(\lambda)$, to standardize spectra from non-uniformities of all components of the instrumentation (light source, fibers, lens, polarizer and spectrometer).

From these measurements, the backscattered reflectance, $R_{BS}(\lambda)$, and the weakly scattered reflectance, $R_{SS}(\lambda)$, have been computed for each sample according to Bendoula *et al.* (2014):

$$R_{BS}(\lambda) = \frac{\left[I_{\parallel}(\lambda) - I_{b\parallel}(\lambda)\right] + \left[I_{\perp}(\lambda) - I_{b\perp}(\lambda)\right]}{\left[I_0(\lambda) - I_{b0}(\lambda)\right]} \tag{5.1}$$

$$R_{SS}(\lambda) = \frac{\left[I_{\parallel}(\lambda) - I_{b\parallel}(\lambda)\right] - \left[I_{\perp}(\lambda) - I_{b\perp}(\lambda)\right]}{\left[I_0(\lambda) - I_{b0}(\lambda)\right]} \tag{5.2}$$

5.2.4 PoLiS absorbance Abs_{PO}

As proposed in Gobrecht *et al.* (2014b), the PoLiS absorbance $Abs_{PO}(\lambda)$ has been computed from the backscattered reflectance $R_{BS}(\lambda)$ and low scattered reflectance $R_{SS}(\lambda)$ as :

$$Abs_{PO}(\lambda) = -\,log\,\left(R_{SS}(\lambda) + \sqrt{(1 - R_{SS}(\lambda))^2 - \frac{R_{SS}(\lambda)}{R_{BS}(\lambda)}\,(1 - R_{BS}(\lambda))^2}\right) \tag{5.3}$$

For comparison, the backscattered absorbance $Abs_{BS}(\lambda)$ has also been computed from the total backscattered reflectance signal $R_{BS}(\lambda)$ measured with PoLiS.

$$Abs_{BS}(\lambda) = -\,log\,R_{BS}(\lambda) \tag{5.4}$$

5.2.5 Multivariate Analysis

Principal Component Analysis

An exploratory analysis of the backscattered absorbance spectra $Abs_{BS}(\lambda)$ and the PoLiS absorbance spectra $Abs_{PO}(\lambda)$, has been carried out using Principal Component

Analysis (PCA). In order to evaluate the impact of the soil preparation on the spectra, the spectral data have been centered in two different ways:

- Mean centered, meaning that the mean spectrum of the entire data set (global mean) is removed from all samples (Coarse, sieved and ground) to analyze the global variance of the dataset;

- Centered according to the location: the mean of the three spectra (one for each particle size preparation) measured for each sample collected at one location is subtracted. This centering allows us to examine the variance within samples having the same TOC content but presenting different physical structure.

The Wilk's lambda criterion (Λ_w) has been applied on the scores of the PCA. Wilk's Lambda is the ratio of the between class variance to the total variance (Roger *et al.*, 2005). Λ_w ranges from 0 to 1. For a value close to one, the classes are well separated and a value close to zero indicates that the classes are confounded.

Calibration with Partial Least Squares Regression

Calibration models have been built using PLS (Wold, 1978), considered as the *benchmark* chemometric technique used for quantitative analysis of diffuse reflectance spectra. The different types of signals computed, $R_{BS}(\lambda)$, $Abs_{BS}(\lambda)$ and $Abs_{PO}(\lambda)$, were compared on the basis of the performances of leave-one-out cross-validation models built on the each particle size sample set to predict soil Total Organic Content (TOC).

Preprocessing methods such as Standard Normal Variate (SNV), Multiplicative Scatter Correction (MSC) and modified Optical Pathlength Estimation and Correction (OPLECm) have been applied to the different spectra.

Finally, the best models obtained for each particle size class have been applied on the other particle size sets.

The performances of the cross-validation models and validation models have been assessed through the number of latent variables used in the models, the coefficient of

determination R^2 and the Standard Error of Cross-Validation (SECV) and Standard error of prediction (SEP) corrected from the bias (Bellon-Maurel *et al.*, 2010).

All the computations have been performed with Matlab software (Matlab R2012b, Mathworks).

5.3 Results and discussion

5.3.1 Spectral analysis

The different mean–per–quartile spectra measured for samples having different particle sizes are plotted in figure 5.3. In the studied wavelength range (400 nm - 800 nm), soil spectra do not show characteristic spectral features and appear "flat". Hence, the differences in the intensity level are related to the brightness of the samples. The reflectance intensity is consistent with the total organic carbon content (TOC) of the studied samples, showing that the darker the sample, the higher the TOC content. This observation concerns all particle sizes classes.

The wavelength range of the PoLiS setup is limited to 400 nm - 800 nm by the range of the polarizer used. This range is not the optimal Vis-NIR region for soil carbon calibration but Viscarra Rossel *et al.* (2008), for example, suggest that the visible portion of the spectrum contains more information on the absorbance characteristics of soil organic carbon than the shortwave NIR (700 – 1100 nm) content. In regard of the objectives of this study, this range is sufficient.

Sample preparation, i.e. particle size, has an impact in the intensity level of the backscattered reflectance $R_{BS}(\lambda)$. As commonly seen in NIR diffuse reflection (Pasikatan *et al.*, 2001), the smaller the particles, the higher the reflectance. As a consequence, the absorbance computed as $Abs_{BS}(\lambda) = -log\ R_{BS}(\lambda)$ shows a lower level for ground samples. Therefore, the differences in the intensity levels are due to the combined effect of the physical structure and the brightness of the soil samples.

For the PoLiS absorbance $Abs_{PO}(\lambda)$, the intensity is about ten times smaller than for

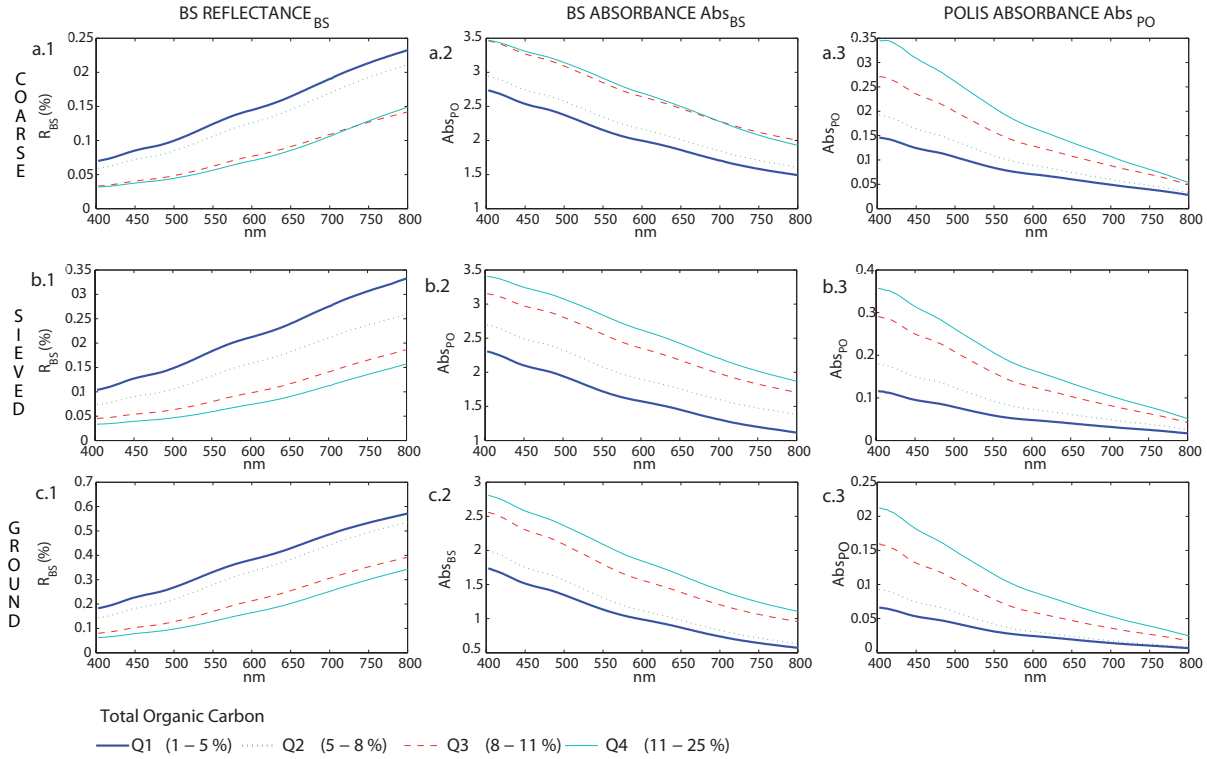

Figure 5.3: Mean reflectance $R_{BS}(\lambda)$, backscattered absorbance $Abs_{BS}(\lambda)$, PoLiS absorbance $Abs_{PO}(\lambda)$ per quartile of TOC concentration for the three different particle sizes (a.) coarse $<$ 5mm, (b.) sieved $<$ 2 mm and (c.) ground $<$ 0.25 mm

backscattered absorbance $Abs_{BS}(\lambda)$. This is partly due to the fact that the PoLiS optical set up selects only a small part of the signal (the single scattered one). The shape is also slightly different, with a small shoulder at 600 nm.

For coarse samples, the absorbance spectra $Abs_{BS}(\lambda)$ of the highly concentrated samples (quartiles Q3 and Q4) are not clearly separated, meaning that the variance due to particle size differences, and therefore scattering, dominates the chemically related information in the spectra. On the contrary, the PoLiS absorbance spectra $Abs_{PO}(\lambda)$ for quartiles Q3 and Q4 are clearly separated. This indicates that part of the spectral information due to the physical structure has been removed. Chemically related information, characterized by the brightness, becomes more visible.

Figure 5.4 shows the score plots of the PCA performed on differently centered spectral datasets ($Abs_{BS}(\lambda)$ and $Abs_{PO}(\lambda)$) according to section 5.2.5.

On the mean-centered data and for both $Abs_{BS}(\lambda)$ and $Abs_{PO}(\lambda)$, PC1 explains more than 98 % of the variability. Because of a multiplicative effect, the spectra appear to be

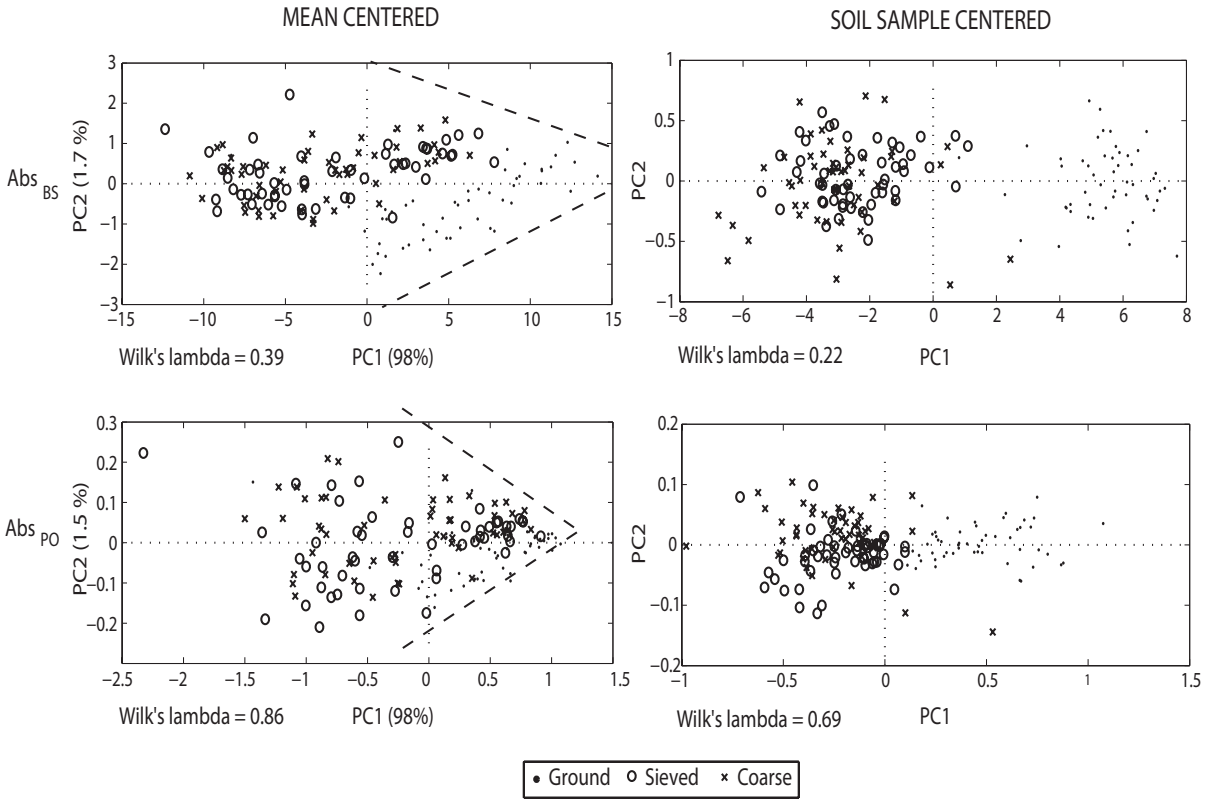

Figure 5.4: Scores plots of the two principal components of the Principal Component Analysis performed on the absorbance spectra $Abs_{BS}(\lambda)$ (first line) and $Abs_{PO}(\lambda)$ (second line) for different data centering (mean centering and centering per sample location) methods.

organized in a conic pattern represented by the triangle in figure 5.4. The responsible influence factors is acombination of soil brightness (related to TOC) and soil physical structure, i.e. the particle size.

The score plot converges both to a minima close to zero and spreads on the opposite side. This conic pattern represented on figure 5.4, is characteristic of a multiplicative effect caused by variables of influence.

For Abs_{BS}, finely ground samples are clearly separated from the two other particle size classes. For $Abs_{PO}(\lambda)$, this separation is less obvious. The summit of the cone contains the darker samples of different particle size classes. The multiplicative effect is due to TOC content as scattering is supposedly lessen.

The score plot of the data centered per sample location confirms the previous observation: for $Abs_{BS}(\lambda)$, the ground soils are clearly separated from the two other classes (sieved and coarse) as for $Abs_{PO}(\lambda)$, the classes appear more confounded. The values of the Wilk's lambda, computed on the scores of the PCA confirm these statements. When particle size classes are separated, the Wilk's lambda is lower.

The PoLiS method corrects, to a certain extent, the effect of scattering on the signal, leading to an absorbance less sensitive to the physical structure of the samples.

5.3.2 Linearity between Absorbance and TOC Concentration

The assumption that, by correcting the signal from part of the multiscattering effect, the PoLiS absorbance $Abs_{PO}(\lambda)$ is more linearly related to TOC content can be assessed through the Pearson's correlation coefficient between the absorbance and the TOC content. The correlograms presented in figure 5.5 show the correlation between the two absorbance signals ($Abs_{BS}(\lambda)$ and $Abs_{PO}(\lambda)$) and TOC as a function of the wavelength and for each sample preparation.

For coarse and sieved samples, the Pearson's coefficient R between the absorbance and the TOC concentration is always higher for $Abs_{PO}(\lambda)$ than for $Abs_{BS}(\lambda)$, over all the wavelength range. For ground samples, the two correlogram are similar, although slightly better for $Abs_{PO}(\lambda)$ between 400 and 600 nm. It is coherent with the general acceptance

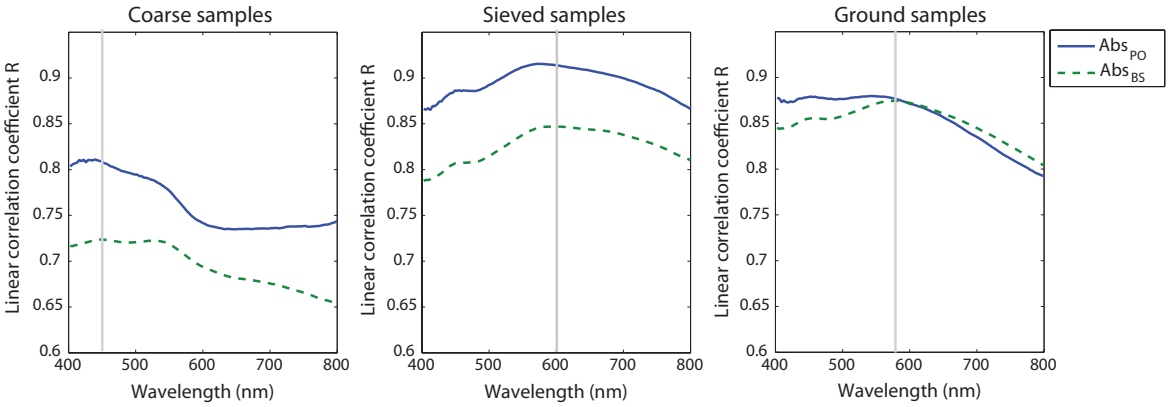

Figure 5.5: Correlogram between Absorbance and TOC for the wavelength range 400 - 800 nm. Vertical line indicates the wavelength at which the correlation coefficient for $Abs_{BS}(\lambda)$ is the highest.

that preparing the samples (sieving or grinding) has a direct impact on the signal quality and consequently on the quality of the calibration models (Morgan *et al.*, 2009; Bellon-Maurel *et al.*, 2010). Here, PoLiS method leads to an additional improvement of the correlation between the Absorbance signal and TOC.

Another way to visualize this observation is to plot TOC versus the absorbance value at the optimal wavelength of $Abs_{BS}(\lambda)$, respectively 450 nm for the coarse samples, 600 nm for the sieved samples and 570 nm for the ground samples (figure 5.6).

The degree of linearity between $Abs_{PO}(\lambda)$ and TOC is improved for coarse and sieved samples, but this effect is lessen for ground samples, for which the linear correlation coefficient for $Abs_{BS}(\lambda)$ and $Abs_{PO}(\lambda)$ are very similar and high ($>$0.87).

To conclude, this analysis shows that $Abs_{PO}(\lambda)$ is more linearly related to the TOC concentration (Figure 5.6) and additionally that the particle size has less impact on its spectral signature (Figure 5.4). Therefore, calibration conditions are more appropriate for $Abs_{PO}(\lambda)$ than for $Abs_{BS}(\lambda)$ to use linear methods like PLS in order to predict TOC in soils.

5.3.3 Model analysis

Quality of the calibration models

Figure 5.7 shows the quality of the models calibrated on the spectra obtained with the different methods : the backscattered reflectance spectra ($R_{BS}(\lambda)$), the backscattered

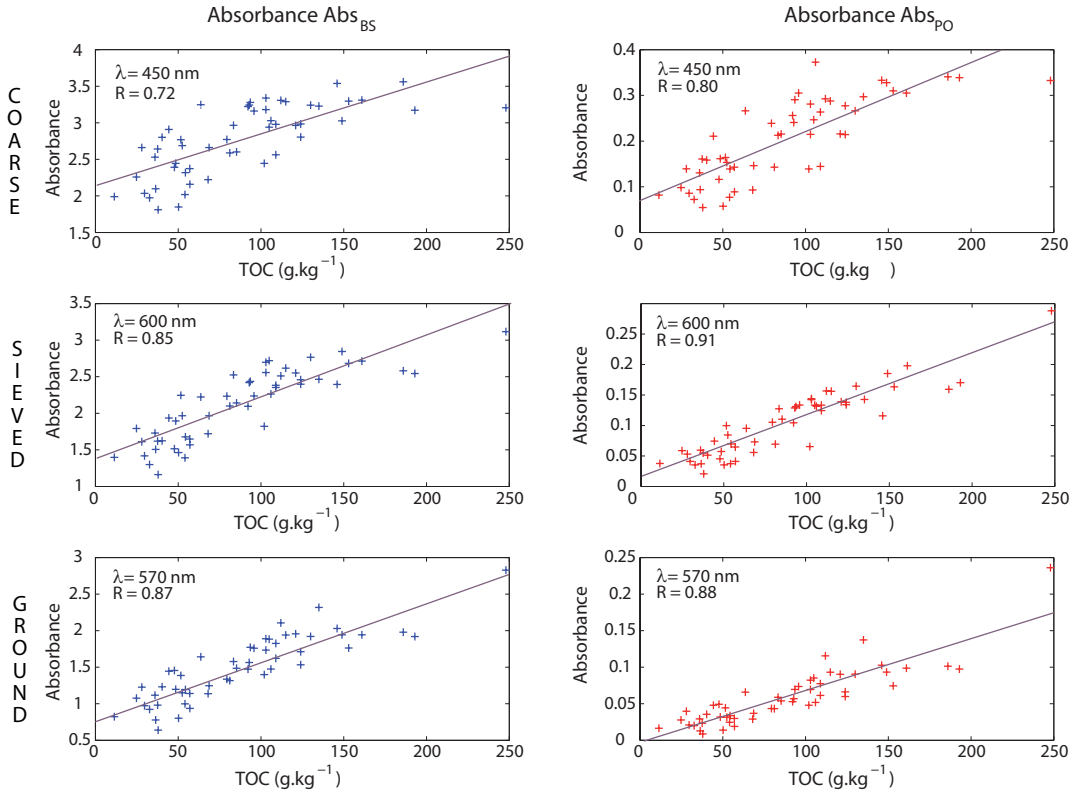

Figure 5.6: Plot of the backscattered absorbance $Abs_{BS}(\lambda)$ and the PoLiS absorbance $Abs_{PO}(\lambda)$ at wavelength λ vs the TOC concentration (in $g \cdot kg^{-1}$) for the three different particle sizes: coarse < 5 mm, sieved < 2 mm and ground < 0.25 mm) with linear fitting. R is the Pearson's coefficient.

absorbance spectra ($Abs_{BS}(\lambda)$) and the PoLiS absorbance spectra ($Abs_{PO}(\lambda)$), with no preprocessing, for each category of particle size.

First, the prediction models built with the backscattered reflectance $R_{BS}(\lambda)$ are not satisfying. They show a characteristic "banana" shaped regression curve, typical of non-linearity. However, ground and sieved samples produce better predictions than coarse samples. The latter present a high structural variability which affects the spectra. The scattering effect dominates in the spectral information but in a different manner for all the samples. This confirms the discussion of the previous section: sieving or grinding soils improves the PLS models.

The log–transformation of the backscattered reflectance $R_{BS}(\lambda)$ into backscattered absorbance, $\{Abs_{BS}(\lambda) = -log\ R_{BS}(\lambda)\}$, improves the quality of the models. Theoretically, the linear relation is between absorbance and concentration and not between reflectance and the concentration. In our case, the *log* also plays the role of a mathematical preprocessing method as it transforms multiplicative effects (due to scattering) into addi-

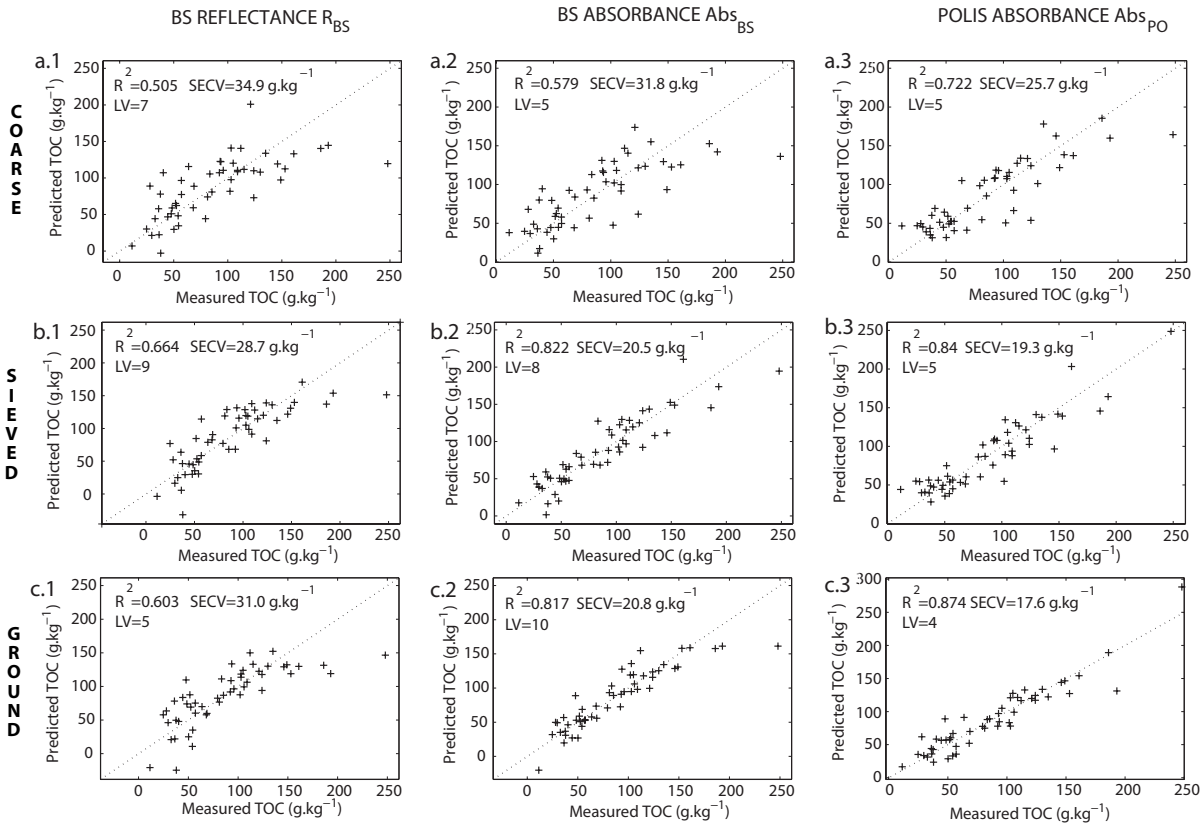

Figure 5.7: Predicted vs measured total organic carbon content from leave-one-out cross validation models calibrated with backscattered reflectance spectra (R_{BS}), backscattered absorbance ($Abs_{BS}(\lambda)$) and PoLiS Absorbance ($Abs_{PO}(\lambda)$) for the three different particle sizes: (a.) coarse $<$ 5mm , (b.) sieved $<$ 2 mm and (c.) finely ground $<$ 0.25 mm) . R^2 : coefficient of determination; SECV: standard error of cross validation; LV: number of latent variables

tive effects (Hadoux *et al.*, 2014). The PLS algorithm is capable to discard this additive effect in the regression process. R^2 and SECV are improved but need a high number of latent variables to build the models (10 for the ground samples and 8 for the sieved samples). According to the principle of parsimony, there is a risk that models will lack in robustness (Bellon-Maurel & McBratney, 2011; Seasholtz & Kowalski, 1993).

The models built with $Abs_{PO}(\lambda)$ outperform all the other models built with $R_{BS}(\lambda)$ and $Abs_{BS}(\lambda)$, whatever the particle size. R^2 and SECV are improved and, in addition, the number of latent variables decreases. However, soil sample preparation still impacts the results. PoLiS method also takes benefit from sample preparation (ground or sieved). For coarse samples, predictions are not so good, although improved compared to the predictions of the models built with the backscattered absorbance Abs_{BS}.

Comparison of optical and mathematical spectral preprocessing

The PoLiS method can be considered as an "optical preprocessing" method: prior to the calibration step, the different components of the total spectra are selected in order to compute an absorbance spectrum. The main objective of this optical preprocessing step is to enhance the quality of the signal by reducing the effect of multiscattering. We compared the calibration results using the PoLiS method with three mathematical preprocessing methods (SNV, MSC and modified OPLEC) usually applied on spectra to reduce the multiplicative and additive effects due to scattering.

Figure 5.8 present the R^2 and the SECV values for each models built.

The TOC prediction models built with the PoLiS absorbance spectra $Abs_{PO}(\lambda)$ always show better figures of merit than for the models built with $R_{BS}(\lambda)$ and $Abs_{BS}(\lambda)$, even when they are preprocessed.

The backscattered reflectance spectra $R_{BS}(\lambda)$ are highly impacted by light scattering. Hence, the preprocessing methods improve the performances of the prediction models, in particular for the sieved and ground samples. SNV and MSC have almost the same behavior on these spectral data, which is often stressed out by authors (Fearn *et al.*, 2009). Modified OPLEC gives good results and seems to be a promising preprocessing method as

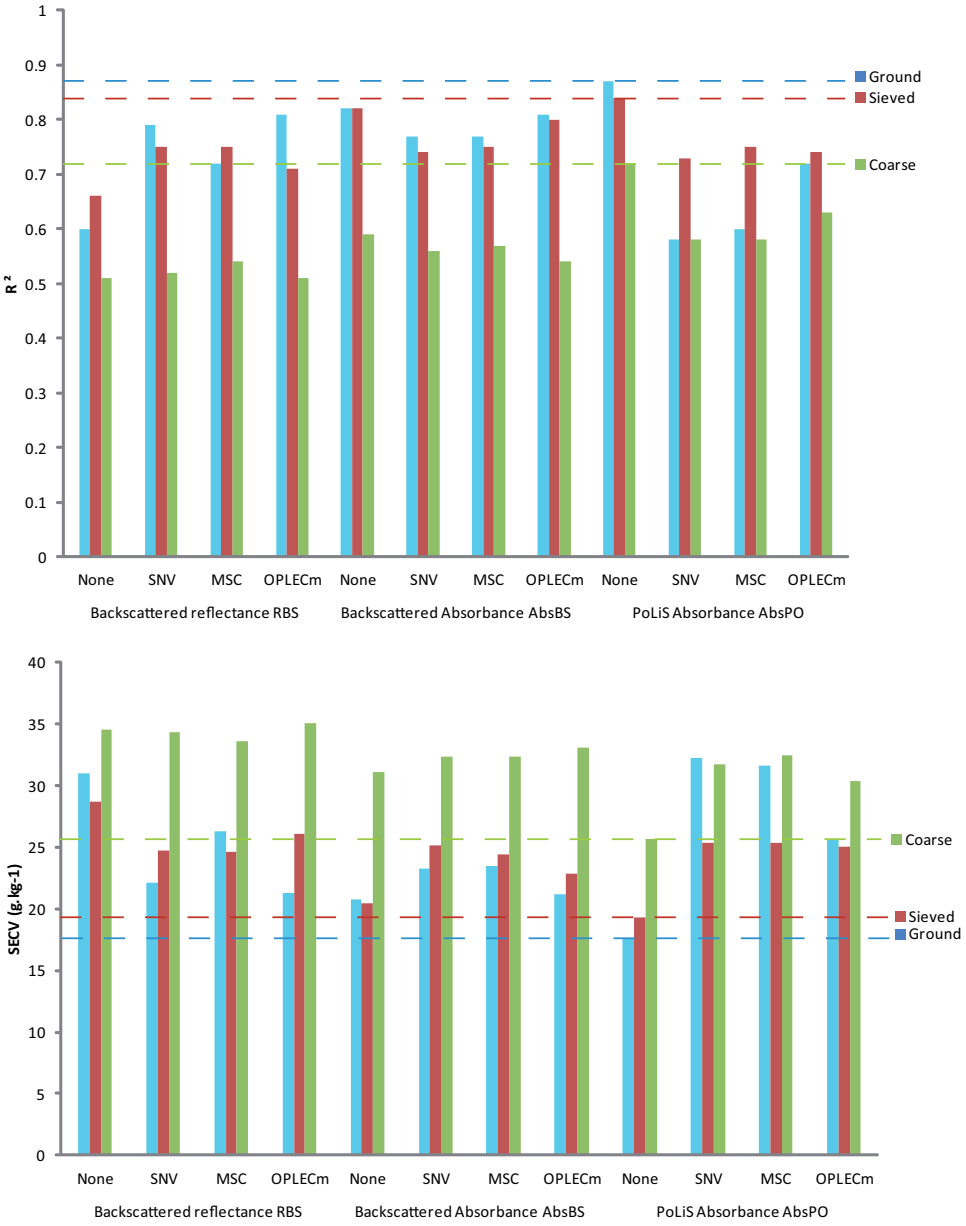

Figure 5.8: Comparison of the determination coefficient R^2 and the Standard Error of cross validation (SECV) of the prediction models built on the three types of samples. Dotted lines correspond to the performances of the models built with $Abs_{Po}(\lambda)$.

it specifically removes the multiplicative effect. For coarse samples however, none of the preprocessing methods applied do significantly increase the quality parameters. These samples present a high sample–to–sample heterogeneity and as a consequence, different levels of light – matter interactions, which are more difficult to capture and correct by the different preprocessing method. Preprocessing the backscattered absorbance spectra Abs_{BS} does not significantly changes the quality of the models, although the number of latent variables decreases from 10 to 7.

For $Abs_{PO}(\lambda)$, none of the preprocessing methods have a positive impact on the figures of merit compared to the raw absorbance spectra. On the contrary, preprocessing the PoLiS absorbance $Abs_{PO}(\lambda)$ highly degrades the quality of the models. It is known that mathematical preprocessing methods suppresses part of the spectral information, sometimes not exclusively due to physical influence but which can also be related to chemical information.

As a conclusion, the PoLiS method produces an optimal absorbance signal, which does not need to be preprocessed prior calibration as the models built from $Abs_{PO}(\lambda)$ always outperform the other models, for all the particle sizes.

Behaviour of the PoLiS method regarding particle size

The main assumption made for the PoliS method is that it reduces the multiscattering effect on the absorbance spectra. Yet, multiscattering is dependent of the particle size of the sample. In section 5.3.1, the PCA analysis on the data concluded that $Abs_{PO}(\lambda)$ is less impacted by the preparation of the samples than $Abs_{BS}(\lambda)$, although, the ground samples still behave differently. Table 5.2 show the quality parameter (R^2, bias and Standard Error of Prediction corrected from the bias (SEP_c) and slope) of the models built on a particle size class and applied to another particle size class.

First, each time finely ground samples (< 0.25 mm) are involved, either in the calibration set or in the test set, PoLiS method do not produce better predictions. R^2 is lower with $Abs_{PO}(\lambda)$ than with $Abs_{BS}(\lambda)$ and the SEP_c, the bias and the slope are

Particle size of the Calibration set	Particle size of the Test set	Signal	L.V.	R^2	SEP_c	Bias	Slope
Coarse	Sieved	$Abs_{BS}(\lambda)$	5	0.64	29	−6.5	0.74
		$Abs_{PO}(\lambda)$	**5**	**0.76**	**24**	**−5.7**	**0.86**
	Ground	$Abs_{BS}(\lambda)$	5	0.67	28	−44	0.70
		$Abs_{PO}(\lambda)$	5	0.62	31	−33	0.50
Sieved	Coarse	$Abs_{BS}(\lambda)$	8	0.53	37	24.5	0.78
		$Abs_{PO}(\lambda)$	**5**	**0.67**	**28**	**6.0**	**0.72**
	Ground	$Abs_{BS}(\lambda)$	8	0.75	24	−20	0.72
		$Abs_{PO}(\lambda)$	5	0.70	28	−34	0.54
Ground	Coarse	$Abs_{BS}(\lambda)$	10	0.45	44	12	0.8
		$Abs_{PO}(\lambda)$	4	0.50	52	23	1.1
	Sieved	$Abs_{BS}(\lambda)$	10	0.70	27	11	0.84
		$Abs_{PO}(\lambda)$	4	0.69	43	31	1.28

Table 5.2: Performance of the models built with $Abs_{BS}(\lambda)$ and $Abs_{PO}(\lambda)$ on one particle size sample set and tested on another particle size sample set. L.V. is the number of latent variables used for the calibration model, R^2 is the coefficient of determination, SEP_c is standard error of prediction corrected form the bias in $g.kg^{-1}$.

worse. We previously observed that for ground samples, $Abs_{BS}(\lambda)$ and $Abs_{PO}(\lambda)$ show a very similar correlogram, meaning that both absorbance signals show a relative linearity with TOC. Here, the PoLiS method seems to reach its limits when the particle size of the particulate samples are very small. Grinding finely the samples affects the way light travels in the samples and probably also the depolarization process. As a consequence, the backscattered reflectance $R_{BS}(\lambda)$ and the low scattered reflectance $R_{SS}(\lambda)$ used to compute the PoliS absorbance $Abs_{PO}(\lambda)$ (equation 5.3) are not completely reliable.

When particle sizes are higher that 2 mm, i.e. sieved or coarse, the models built with $Abs_{PO}(\lambda)$ always produce better results than with $Abs_{BS}(\lambda)$, as shown in figure 5.9.

Although the PoLiS calibration model built on coarse samples was the less performant in cross-validation (see figure 5.7), the prediction are not degraded when it is applied on the sieved samples. Moreover, the bias, which is a good indicator of robustness, remains small. On the other way, when the model built on sieved samples is applied on coarse samples, the figures of merit are not as good as in cross validation, but still, the results are much better with $Abs_{PO}(\lambda)$ than with $Abs_{BS}(\lambda)$. And again, the bias is very small for $Abs_{PO}(\lambda)$ compared to the high bias value for $Abs_{BS}(\lambda)$.

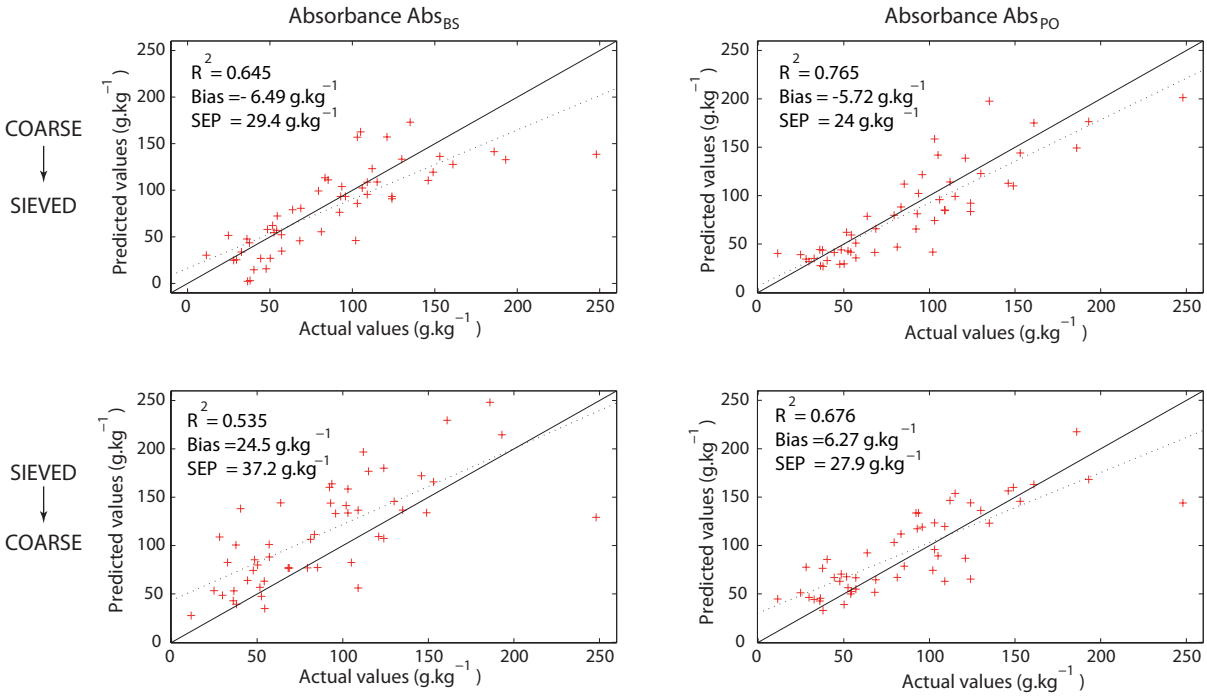

Figure 5.9: Predicted vs measured total organic carbon content. Models were calibrated with the backscattered absorbance ($Abs_{BS}(\lambda)$) and the PoLiS Absorbance ($Abs_{PO}(\lambda)$) on one particle size class and tested on another particle size class. (upperline: coarse $<$ 5 mm on sieved $<$ 2 mm and lower line: sieved $<$ 2 mm on coarse $<$5 mm). R^2 : coefficient of determination, SEP_c: standard error of Prediction corrected from the bias in $g.kg^{-1}$.

These results show that PoLiS is a promising measurement technique in the perspective of reducing the sample preparation as it is less sensitive to changes of the physical structure of the samples and well adapted to low processed samples.

5.4 Conclusions

For the first time, the issue of light scattering in Vis-NIR spectroscopy applied to soils has been studied from an optical point of view. In this study, PoLiS, an original optical setup based on light polarization spectroscopy, has been used to select backscattered light being less impacted by multiscattering effects due to particles composing soil samples. The absorbance signal computed from the PoLiS measurements has been compared to the absorbance traditionally computed by taking the log of the backscattered reflectance.

The aim of this study was to verify the assumptions underpinning the PoLiS method. We can make following statements and concluding remarks :

- On soil samples, the method produced spectral signatures of good quality, with no noise, despite the low intensity in the PoLiS wavelength range;

- Removing part of the multiscattering improved the degree of linearity between the PoLiS absorbance and the TOC, over all the wavelength range (400 - 800 nm) for coarse and sieved samples.

- TOC prediction models build with the PoLiS absorbance always outperformed the models built with the backscattered absorbance, even when mathematically preprocessed. This is an important result confirming that a signal of better quality improves the quality of the prediction models.

- The PoLiS absorbance is less impacted by a change of particle size of the samples but an effect is still visible, particularly for ground samples. As a consequence, the predictive potential of the PoLiS absorbance when only the physical structure of the sample changes is higher than the backscattered absorbance, when the particle size is > 2 mm. For finely ground samples, PoLiS seems to reach it limits.

This study confirms the high potential of the PoLiS method for the spectral analysis of soil properties. Solving the technical limits which would make the PoLiS method work beyond 800 nm, would allow to take an important step in the metrological quality of the soil carbon content measurement by NIRS.

Contributions of chapter 5 and outlook

In this chapter we tested the PoLiS method, which combines an optical setup based on light polarization spectroscopy and the Representative Layer Theory to model the absorbance signal of soils. This absorbance signal tends to be more linearly related to the concentration of organic carbon, which is an important pre-requisite to perform linear multivariate modeling.

In a second step, we showed that the method leads to calibration model which perform appreciably better than models based on preprocessed reflectance spectra.

The results of this preliminary study on soils should be confirmed on a larger soil database. In addition, the wavelength range of the actual version of the PoliS method is not the most relevant for the study of chemical soil properties. Therefore, technical improvements are needed to confirm the high potential of the PoLiS method to characterize soils.

Chapter 6

Contributions and Perspectives

Contents

6.1 Introduction

In this thesis we aim at developing an optical method based on light polarization spectroscopy to measure the absorbance of highly scattering materials. The operational problem that initiated this work was to measure soil carbon content with Vis-NIR spectroscopy and while the classical methods faced some limitations mainly due to light scattering. We proposed therefore an optical architecture capable of reducing the effect of multiscattering on the spectra, posing the assumption that the calibration models built with these spectra would be more precise and robust.

While the goal of this undertaking was very focused on a particular application, it opened new alleys of research. This final chapter synthesizes the major findings and results of the current thesis. It summaries the assumptions, capabilities and constraints of the PoLiS method.

The scientific perspectives, that have emerged during this work, are also presented, as a testimony of the bright future of light polarization spectroscopy serving multivariate analysis of complex materials.

6.2 Summary of the main contributions of the work

6.2.1 A pedagogical review : back to basics !

The first contribution of this work is a pedagogical review mainly addressed to soil scientists. We focused on the causal link between the theoretical concepts underpinning NIR and linear chemometric modeling and the question why such a promising technique, NIR, is still not largely widespread in soil analysis.

The review highlights that light scattering is an important source of limitations: it negatively impacts the NIR spectrum, which itself is not a very selective signal. As a consequence, extracting the relevant information, being usually the chemical absorbance, becomes a much bigger challenge than for non scattering materials. Indeed, the useful information is overlapped, both linearly and non linearly, by useless, and even sometimes

harmful spectral information (section 2.2).

To overcome these limitations, the main efforts have been concentrated on the development or adaptation of chemometric methods. The strategy is to either restore the linearity between signal and concentration, by preprocessing the spectra for example (section 2.5.1) or to circumvent the problem by using local and non-linear approaches (section 2.6). If the latter present a certain potential, linear approaches such as PLS remain, by far, the number one calibration method in NIR analysis. PLS is simple to implement (sometimes even already implemented in spectral analysis software), rapid and simple to interpret. However, as it is a linear method, it is also the most impacted one, in case of high level of scattering.

The conclusions drawn at the end of the review insist on the fact that overcoming the issue of signal quality should improve the performances of NIR spectroscopy as an analytical tool for soil analysis.

6.2.2 PoLiS: an original optical setup to reduce the scattering effect

Optical methods aiming at separating the absorbing coefficient from the scattering coefficients in NIR spectra already exist (section 3.1) but we found out that they are not adapted to highly scattering and absorbing materials such as soil samples. Their common principle is to solve (directly or by model inversion) a system of equations with two unknown parameters, i.e. the scattering and the absorption coefficients. Hence, it is necessary to collect at least two different type of spectral information about the studied material. The most common "set" of measurements is the Transmission and the Reflectance performed on the same sample (in the Inverse Adding-Doubling methods or more simple 2-Flux methods such as the Kubelka-Munk model). An alternative is to measure a reflectance on a optically infinite sample and a reflectance on a optically finite sample. In methods such as spatially or time resolved spectroscopy additional spectral data are acquired as a function of space or time.

In soil samples, the distance traveled by the photons is very short before they are absorbed, that it is neither possible to measure a transmittance or a reflectance on an optically thin sample (like in Kessler *et al.* (2009)) nor to have different spectral signatures with SRS. In highly scattering and absorbing samples, on which transmission measurements are not possible to perform, the optical analysis must rely on reflectance measurements.

This conducted us to find alternative ways to measure this "set" of different spectral information : we used light polarization properties. Based on the theoretical principles of polarization subtraction we designed an optical architecture aiming at decomposing a remitted signal in two complementary components: a multiscattered reflectance and a low scattered reflectance (section 3.2). This optical setup is fully adapted to highly scattering materials as the measurements are performed only in reflectance on optically thick samples.

We first tested the PoLiS setup on powdered model samples mixing sand and two coloring dyes. We observed that when corrected from multiscattering, the reflectance signal becomes a linear combination of the pure components spectra. On the contrary, the classical reflectance spectrum tends to be a non-linear mixture of the two colorant spectra (section 3.4.1). This preliminary result showed the potential of the PoLiS method to correct the spectrum of physical interactions.

In addition, whatever the type of sample, powder or liquid form (section 4.4), the spectra (multiscattered and low scattered component) showed a good signal to noise ratio in the studied wavelength range (350 nm to 800 nm).

Based on the principles of light polarization, the PoLiS method outputs two different types of signals: a classical backscattered reflectance and a corrected reflectance, which proved to be less impacted by multiscattering.

6.2.3 A model of the absorbance of highly scattering materials

According to Beer-Lambert law, it is the absorbance that is linearly related to concentration. Here, the objective is to provide a better approximation of the true absorbance of

scattering materials than the almost exclusively used expression $\{-log\ R_{BS}(\lambda)\}$, which is inherently non linear with concentration. The main reason is that applying Beer-Lambert Law to reflectance measurements is based one wrong assumptions: (i) the path-length of light is constant and (ii) the scattering coefficient for the sample is independent of absorption (Dahm & Dahm, 2001).

In this thesis, we used the frame of the Representative Layer Theory proposed by Dahm & Dahm because they explicitly raise the question of an equivalent to the Beer-Lambert Law for scattering materials (Dahm & Dahm, 2007, p. 34).The RLT allows for a layer to contain particle types of multiple materials and diameters, as well as voids between the particles so long as each layer is identical in its composition with relation to the volume and surface area ratio between particle types (Dahm & Dahm, 2007). Because of the initial assumptions about a sample, this technique is particularly applicable to the optical characterization of powdered samples, which may be contain multiple chromophores .

We put forward the hypothesis that the "set" of PoLiS reflectance measurements can be implemented in the Absorption – Remission function to model the absorbing power of a scattering sample (section 4.2.4).

We validate these assumptions experimentally for liquid and powdered samples by confirming that the PoLiS absorbance showed a better linear relation with the absorber concentration (section 4.4.2) than the classical backscattered absorbance $\{-log\ R_{BS}(\lambda)\}$.

6.2.4 Application on soils

The samples studied to validate experimentally the PoLiS method are simple samples, mixing a scattering but non absorbing matrix (milk or sand) with a unique absorber (a coloring dye). Applying the PoLiS method on real soil samples intent to confirm that the method could be applied on more complex samples (soil is a sort of ideal complex sample) to predict more complex variables of interest (e.g. Total Organic Carbon).

First, from a practical point of view, the PoLiS optical setup is fully adapted to the measurement of air dried and sieved soil samples. The collected spectra showed a good

S/N ratio in the 350 - 800 nm range. Next, the linearity between the PoLiS absorbance Abs_{PO} with TOC is improved compared to Abs_{BS} (section 5.3.2).

More important, we confirmed that building a calibration model with PLSR using the PoLiS absorbance to predict the TOC content outperforms the model built with Abs_{BS}, even when mathematical pretreatments were applied to it (section 5.3.3). Here, in our experiment, we found out that preprocessing the PoLiS absorbance degraded the model. This leads us to believe that PoLiS is an optical preprocessing method that discards only the useless information from the absorbance spectra which reaches an optimal quality regarding linear multivariate analysis.

6.3 Technical limits and areas of improvements

The application of the Polis method on soil did highlight some limits, which are presented, and discussed. As these limits are mainly of technical order, we propose some technical improvements.

6.3.1 Limits of the actual optical setup

The wavelength range of the PoLiS setup, i.e. 350 - 800 nm, tend not to be a limiting factor to study coloring dyes, which absorb in the visible range. To study soil chemical properties, however, this restricted range is a clear limitation, although for carbon, there is clearly a link between soil color and soil carbon content.

The main reason that we can not measure in the near infrared range is related to the detector of the spectrometer. The quality of the signal depends on the responsivity of the detector. In the Vis-VNIR range (350 nm – 1100 nm), the spectrometer includes a silicon detector, which present the advantage of having a high responsivity. Over 1000 nm, (SWIR - NIR), the spectrometer is generally composed of an InGaAs (Indium Gallium Arsenide) detector, which show a lower responsivity. So if the signal is to low in intensity, the noise will be relatively high. The PoLiS measurements, as they result from the difference between the two signals $R_{\parallel}$ and $R_{\perp}$, are too noisy to be used.

6.3.2 Areas of improvements

To overcome these limits, some technical adaptation of the PoLiS method should be tested:

- To augment the signal intensity , one can use a more powerful light source. In the PoliS optical setup, the light source used is an halogen lamp (150 W, Leica Cls). A lot of power is lost by collimating the beam. Using a supercontinuum source (laser), which is already collimated, will concentrate the available energy on a small surface of the sample. As a consequence, the remitted intensity will increase and therefore also the selected low scattered component. Another lever would be to rethink the architecture of the collecting part of the device, with the objective to increase the quantity of photons reaching the detector. The optical components must be chosen so as to maintain the intensity at its maximum. The right lenses have to be chosen and the use of optical fibers have to be limited, as they attenuate light.

- To build the whole spectrometer integrating a source, optical components, a monochromator (wavelength range) and a detector. Each component can be adapted to optimize the signal quality. This is a necessary stage to define the technical specifications of a fully optimized sensor.

6.4 Scientific perspectives

6.4.1 Increasing knowledge about the studied material

The PoLiS method is combining various theoretical fields such as light polarization principles, the Beer–Lambert physical law and the Representative Layer Theory. This coupling allows us to study light–matter interactions at two different, but complementary levels: the macroscopic and the microscopic one.

- From a macroscopic point a view, the light is considered as a corpuscular element and different properties of the material can be extracted from the quantity of photons reaching (or not) the detector. The frame of the Representative Layer Theory is very promising to understand how light travels in the material and how it is absorbed. But the added-value is the combination of the RLT with the PoliS method. Indeed, the optical setup can implement different polarization status that are different of linear one : the the elliptic or circular ones. The wave will interact differently with the material. For example, circular polarized light penetrated deeper in the material before it looses its polarization status (Voit *et al.*, 2012). The reflectance signals measured could be accordingly interpret and provide new knowledge about the material, a better understanding of the light–matter interaction and the mechanism of light absorption..

- From a microscopic point a view: The fraction of light that is reflected by a surface can be computed with the Fresnel equations. This fraction is a function of the complex refractive index $\{n - ik\}$ of the material and the state of polarization of the incident beam; k is known as the absorption index and it is related to the absorbing power of the material (Wendlandt & Hecht, 1966). The PoLiS method allows us to measure all the polarization states of light. Polarized light with its electric field along the plane of incidence is denoted p-polarized, while light electric field of which is normal to the plane of incidence is called s-polarized. When these wave interact with the material, their reflectance $R_s(\lambda)$ and $R_p(\lambda)$ have a different

expression. From them, it may be possible to calculate analytically, by using model inversion, the complex refractive index $\{n - ik\}$ of the medium. These values are of high interest for a several complex materials. Among them soils, for which published information on refractive indices is very scarce.

6.4.2 Assessing the signal quality prior calibration

All through this research, we were confronted to the question of assessing the quality of the signals produced by each method. This was particularly the case when we had to optimize the architecture of the optical set up. Here, we consider that a signal is of good quality, if it is sufficiently selective and sensitive to predict the variable of interest. In other terms, if it contains sufficiently information to be captured by the model.

Usually, to assess the impact of a signal, we assess the quality of the model: the model is built and figures of merit (FOM) of the prediction model are compared (Dardenne *et al.*, 2000). Among them, the correlation coefficient R^2, the standard error of prediction (SEP) and the bias. However, this procedure needs many available samples, each with a known reference value to build, validate and test the model. In addition, the FOM assess the whole analytical process (i.e. comprising the measurement and the calibration) and it is difficult to know which of the measurement stage or the model calibration stage has the higher impact on the prediction uncertainty.

Several FOM exist, dedicated to signal comparison. They mainly come from the frame of the Net Analyte Signal (NAS), a concept introduced by Lorber *et al.* (1997). The NAS is the part of the measured signal that a calibration model relates to the property of interest (e.g. analyte concentration) (Boelens *et al.*, 2004). The remaining part contains the contribution from other components. Several figures of merit are computed from the NAS, such as the selectivity, the sensibility, the signal to noise ratio and limit of detection (Olivieri *et al.*, 2006). In simple mixtures, where the pure spectra are known, the real NAS can be computed. However, if the samples are more complex or if the pure spectra are not available (which is, in NIR spectroscopy, the usual case), NAS has to be estimated. Several methods exist to estimate the NAS, depending on the available

information about the analyte of interest and the interferent. The main method is to estimate the NAS from the b coefficient of a PLS model (Faber, 1998).

If the purpose of the figure of merit is to assess the signal quality, the there is no real added value in computing FOM from an estimated NAS using the model coefficient in comparison to computing the traditional FOM of model quality from the same database: R^2, RMSEP, bias. In addition, figures of merit are computed for each sample.

Therefore, knowing about the signal quality before (and independently of) the calibration step is of practical interest when different optical setups have to be benchmarked. As far as we know, there is no such a quality parameter, which could assess the prediction capacity of a spectra without building a model.

General conclusion

The aim of the present thesis was to provide an optical methodology to measure, with Vis-NIR spectroscopy, an absorbance signal of optimized quality to characterize soils. Two main scientific questions have driven this work:

1. *How to optically reduce the impact of light scattering on the spectroscopic signal ?*
2. *How to model the chemical absorbance of highly scattering materials ?*

The first step was to design an optical setup, named PoLiS, dedicated to remove scattering from reflectance signals measured on highly absorbing and scattering materials. Using the wave theory of light, this approach was based on the fact that, when linearly polarized light interacts with a scattering medium, the remitted signal looses its initial polarization state because of the multiple scattering events. By light polarization subtraction, it was possible to select light beams that were less impacted by multiple scattering events.

The second step was to link the corrected signal measured with PoLiS to the chemical absorbance of the material. The Representative Layer Theory provided a theoretical frame to model, from the PoLiS measurements, the absorbed fraction of a hypothetical representative layer of the sample. From this absorbed fraction, an absorbance signal, less impacted by scattering could be computed and used for multivariate analysis.

The assumptions underlying our approach combining the PoLiS measurements and the RLT have been successfully verified on model samples, mixing powdered or liquid scattering matrices with coloring dyes, in the Vis-VNIR (350 - 800 nm) range: the absorbance signal retrieved its linearity with the absorbers concentration.

The feasibility of the method to be applied on soil samples has been tested to predict total organic carbon content. Again, the linearity between the PoLiS absorbance and the concentration of TOC improved compared to the classical absorbance $\{-log\, R(\lambda)\}$. But more importantly, the PLS models built from the PoLiS absorbance outperformed the models built from the classical absorbance, this, even when the signals were mathematically preprocessed to reduce scattering. The standard errors of cross validation decreased from 20.8 $g.kg^{-1}$ to 17.6 $g.kg^{-1}$ and the coefficient of determination R^2 improved from 0.82 to 0.87 on ground samples, although the wavelength range was not the optimal range for soil carbon analysis.

This work confirmed that by optical means, it is possible to significantly improve the quality of a spectroscopic signal. It also confirmed that, when the absorbance signal is more linearly related the analyte of interest concentration, the linear model is improved. These findings allow us to see the great potential of this method, both for the characterization of soils and more generally, for all materials presenting the common characteristic of being complex from the physical structure and chemical composition point of view.

Bibliography

Abrahamsson, C., Löwgren, A., Strömdahl, B., Svensson, T., Andersson-Engels, S., Johansson, J., & Folestad, S. 2005a. Scatter Correction of Transmission Near-Infrared Spectra by Photon Migration Data: Quantitative Analysis of Solids. *Applied Spectroscopy*, **59**(11), 1381–1387.

Abrahamsson, C., Johansson, J., Andersson-Engels, S., Svanberg, S., & Folestad, S. 2005b. Time-Resolved NIR Spectroscopy for Quantitative Analysis of Intact Pharmaceutical Tablets. *Analytical Chemistry*, **77**(4), 1055–1059.

Ahrens, L. H. 1954. The lognormal distribution of the elements (A fundamental law of geochemistry and its subsidiary). *Geochimica et Cosmochimica Acta*, **5**(2), 49–73.

Allen, M. P. 1997. Assumptions of ordinary least-squares estimation Understanding Regression Analysis. *Pages 181–185 of: Understanding Regression Analysis*. Springer US.

Arimoto, H. 2006. Multispectral Polarization Imaging for Observing Blood Oxygen Saturation in Skin Tissue. *Applied Spectroscopy*, **60**(4), 459–464.

Backman, V., Gurjar, R., Badizadegan, K., Itzkan, I., Dasari, R. R., Perelman, L. T., & Feld, M. 1999. Polarized light scattering spectroscopy for quantitative measurement of epithelial cellular structures in situ. *Selected Topics in Quantum Electronics, IEEE Journal of*, **5**(4), 1019–1026.

Ballabio, C. 2009. Spatial prediction of soil properties in temperate mountain regions using support vector regression. *Geoderma*, **151**(3-4), 338–350.

Barnes, R. J., Dhanoa, M. S., & Lister, Susan J. 1989. Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. *Applied Spectroscopy*, **43**(5), 772–777.

Bartholomeus, H. M., Schaepman, M. E., Kooistra, L., Stevens, A., Hoogmoed, W. B., & Spaargaren, O. S. P. 2008. Spectral reflectance based indices for soil organic carbon quantification. *Geoderma*, **145**(1-2), 28–36.

Bellon-Maurel, V. 2009. NIR and Soil Science : A teenage love-story. *Pedometron*, **28**.

Bellon-Maurel, V., & McBratney, A.B. 2011. Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils - Critical review and research perspectives. *Soil Biology and Biochemistry*, **43**(7), 1398–1410.

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.M., & McBratney, A.B. 2010. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *TrAC Trends in Analytical Chemistry*, **29**(9), 1073–1081.

Ben-Dor, E., Chabrillat, S., Demattê, J. A. M., Taylor, G. R., Hill, J., Whiting, M. L., & Sommer, S. 2009. Using Imaging Spectroscopy to study soil properties. *Remote Sensing of Environment*, **113, S1**(0), S38–S55.

Bendoula, R, Gobrecht, A., Moulin, B., Roger, J.M., & Bellon-Maurel, V. 2014. Improvement of the chemical content prediction of a model powder system by reducing multiple scattering using polarized light spectroscopy. *Applied Spectroscopy*, **(in press)**(xx).

Benford, F. 1946. Radiation in a Diffusing Medium. *Journal of the Optical Society of America A*, **36**(9), 524–554.

Bernoux, M., Cerri, C. C., Cerri, C. E. P., Neto, M. S., Metay, A., Perrin, A.S., Scopel, E., Razafimbelo, T., Blavet, D., Piccolo, M. de C, *et al.* 2006. Cropping systems, carbon

sequestration and erosion in Brazil, a review. *Agronomy for Sustainable Development*, **26**(1), 1–8.

Bertie, J. E. 2006. *Glossary of Terms used in Vibrational Spectroscopy.* John Wiley & Sons, Ltd. Chap. 1, page 49.

Bertran, E., Blanco, M., Maspoch, S., Ortiz, M. C., Sánchez, M. S., & Sarabia, L. A. 1999. Handling intrinsic non-linearity in near-infrared reflectance spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, **49**(2), 215–224.

Bertrand, D., & Dufour, E. 2006. *La spectroscopie infrarouge et ses applications analytiques* . 2 edn. Lavoisier.

Boelens, H.F.M., Kok, W.T., de Noord, O.E., & Smilde, A. K. 2004. Performance optimization of spectroscopic process analyzers. *Analytical chemistry*, **76**(9), 2656–2663.

Bogomolov, A., Melenteva, A., & Dahm, D. J. 2013. Technical note: Fat globule size effect on visible and shortwave near infrared spectra of milk. *Journal of Near Infrared Spectroscopy*, **21**(5), 435–440.

Breiman, L. 1984. *Classification and regression trees.* Belmont, CA: Wadsworth International Group.

Breiman, L. 2001. Random forests. *Machine Learning*, **45**(1), 5–32.

Brejda, J. J., Moorman, T. B., Smith, J. L., Karlen, D.L., Allan, D. L., & Dao, T. H. 2000. Distribution and Variability of Surface Soil Properties at a Regional Scale. *Soil Science Society of America Journal*, **64**(3), 974–982.

Brown, D. J. 2007. Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. *Geoderma*, **140**(4), 444–453.

Brown, D. J., Shepherd, K. D., Walsh, M.G., Dewayne Mays, M., & Reinsch, T. G. 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma*, **132**(3-4), 273–290.

Cabassi, G., Profaizer, M., Marinoni, L., Rizzi, N., & Cattaneo, T. 2013. Estimation of fat globule size distribution in milk using an inverse light scattering model in the near infrared region. *Journal of Near Infrared Spectroscopy*, **21**(5), 359–373.

Cattaneo, T., Cabassi, G., Profaizer, M., & R., Giangiacomo. 2009. Contribution of light scattering to near infrared absorption in milk. *Journal of Near Infrared Spectroscopy*, **17**(6), 337–343.

Cécillon, L., Barthès, B. G., Gomez, C., Ertlen, D., Genot, V., Hedde, M., Stevens, A., & Brun, J. J. 2009. Assessment and monitoring of soil quality using near-infrared reflectance spectroscopy (NIRS). *European Journal of Soil Science*, **60**(5), 770–784.

Chauchard, F., Roger, J.M., Bellon-Maurel, V., Abrahamsson, C., Andersson-Engels, S., & Svanberg, S. 2005. MADSTRESS: A linear approach for evaluating scattering and absorption coefficients of samples measured using time-resolved spectroscopy in reflection. *Applied Spectroscopy*, **59**(10), 1229–1235.

Chen, H., Bakshi, B. R., & Goel, P. K. 2007. Toward Bayesian chemometrics. A tutorial on some recent advances. *Analytica Chimica Acta*, **602**(1), 1–16.

Chen, Z.-P., Morris, J., & Martin, E. 2006. Extracting Chemical Information from Spectral Data with Multiplicative Light Scattering Effects by Optical Path-Length Estimation and Correction. *Analytical Chemistry*, **78**(22), 7674–7681.

Ciani, A., Goss, K. U., & Schwarzenbach, R. P. 2005. Light penetration in soil and particulate minerals. *European Journal of Soil Science*, **56**(5), 561–574.

Clark, R.N. 1999. *Spectroscopy of rocks and minerals and principles of spectroscopy.* Chichester, UK: John Wiley & Sons.

Cleveland, W. S., & Devlin, S.J. 1988. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, **83**(403), 596–610.

Coello, J., Maspoch, S., *et al.* 2008. Application of representative layer theory to near-infrared reflectance spectra of powdered samples. *Applied Spectroscopy*, **62**(12), 1363–1369.

Cogdill, R. P., & Dardenne, P. 2004. Least-squares support vector machines for chemometrics: an introduction and evaluation. *Journal of Near Infrared Spectroscopy*, **12**(2), 8.

Cook, R.D, & Weisberg, S. 1982. *Residuals and Influence in Regression*. New York, NY: John Wiley & Sons.

Dahm, D. J. 2013. Review: Explaining some light scattering properties of milk using representative layer theory. *Journal of Near Infrared Spectroscopy*, **21**(5), 323–339.

Dahm, D.J., & Dahm, K.D. 1999. Representative Layer Theory for Diffuse Reflectance. *Applied Spectroscopy*, **53**(6), 647–654.

Dahm, D.J., & Dahm, K.D. 2001. The physics of near-infrared scattering. *Chap. 1, pages 1–17 of:* Williams, P., & Norris, K.H. (eds), *Near-Infrared Technology in the Agricultural and Food Industries*. Amer Assn of Cereal Chemists, St. Paul, Minnesota,, USA.

Dahm, D.J., & Dahm, K.D. 2004a. Short communication: Illustration of failure of continuum models of diffuse reflectance. *Journal of near infrared spectroscopy*, **11**(6), 479–485.

Dahm, D.J., & Dahm, K.D. 2007. *Interpreting Diffuse Reflectance and Transmittance: A Theoretical Introduction to Absorption Spectroscopy of Scattering Materials*. 1 edn. NIR Publications.

Dahm, Kevin D, & Dahm, Donald J. 2004b. Relation of representative layer theory to other theories of diffuse reflection. *Journal of near infrared spectroscopy*, **12**(3), 189–198.

Dahm, Kevin D, & Dahm, Donald J. 2013. Separating the effects of scatter and absorption using the representative layer. *Journal of Near Infrared Spectroscopy*, **21**(5), 351–357.

Daniel, K. W., Tripathi, N. K., & Honda, K. 2003. Artificial neural network analysis of laboratory and in situ spectra for the estimation of macronutrients in soils of Lop Buri (Thailand). *Soil Research*, **41**(1), 47–59.

Dardenne, Pierre, Sinnaeve, George, & Baeten, Vincent. 2000. Multivariate calibration and chemometrics for near infrared spectroscopy: which method? *Journal of Near Infrared Spectroscopy*, **8**(4), 229–237.

Davies, A. M. C., & Fearn, T. 2006a. Back to basics: calibration statistics. *Spectroscopy Europe*, **18**(2), 31–32.

Davies, A. M. C., & Fearn, T. 2006b. Quantitative analysis via near infrared databases : comparison analysis using restructured near infrared and constituent data-deux (CARNAC-D). *Journal of Near Infrared Spectroscopy*, **14**(6), 9.

Demos, S. G., & Alfano, R. R. 1996. Temporal gating in highly scattering media by the degree of optical polarization. *Optics Letters*, **21**(2), 161–163.

Demos, S. G., & Alfano, R. R. 1997. Optical polarization imaging. *Applied Optics*, **36**(1), 150–155.

Eswaran, H., Reich, P.F., Kimble, J.M., Beinroth, F.H., Padmanabhan, E., & Moncharoen, P. 2000. Global carbon stocks. *Global climate change and pedogenic carbonates*, 15–25.

Faber, Nicolaas (Klaas) M. 1998. Efficient computation of net analyte signal vector in inverse multivariate calibration models. *Analytical chemistry*, **70**(23), 5108–5110.

Farrell, T.J., Patterson, M.S., & Wilson, B. 1992. A diffusion theory model of spatially resolved, steady-state diffuse reflectance for the noninvasive determination of tissue optical properties in vivo. *Medical Physics*, **19**(4), 879–888.

Fearn, T. 2012. Do my data need to be normally distributed? *NIR News*, **23**(1), 20–21.

Fearn, T., Riccioli, C., Garrido-Varo, A., & Guerrero-Ginel, J. E. 2009. On the geometry of SNV and MSC. *Chemometrics and Intelligent Laboratory Systems*, **96**(1), 22–26.

Fearn, T., Perez-Marin, D., Garrido-Varo, A., & Guerrero-Ginel, J. E. 2010. Inverse, classical, empirical and nonparametric calibrations in a Bayesian framework. *Journal of Near Infrared Spectroscopy*, **18**(1), 27–38.

Fernandez-Ahumada, E., Roger, J.M., & Palagos, B. 2012. A new formulation to estimate the variance of model prediction. Application to near infrared spectroscopy calibration. *Analytica Chimica Acta*, **721**(0), 28 – 34.

Fernández Pierna, J. A., & Dardenne, P. 2008. Soil parameter quantification by NIRS as a Chemometric challenge at [']Chimiométrie 2006'. *Chemometrics and Intelligent Laboratory Systems*, **91**(1), 94–98.

Fidêncio, P.H., Poppi, R. J., & de Andrade, J. C. 2002. Determination of organic matter in soils using radial basis function networks and near infrared spectroscopy. *Analytica Chimica Acta*, **453**(1), 125–134.

Friedman, J.H. 1991. Multivariate adaptive regression splines. *The annals of statistics*, 1–67.

Gehl, R.and Rice, C. 2007. Emerging technologies for "in situ" measurement of soil carbon. *Climatic Change*, **80**(1), 43–54.

Geladi, P. 2003. Chemometrics in spectroscopy. Part 1. Classical chemometrics. *Spectrochimica Acta Part B: Atomic Spectroscopy*, **58**(5), 767 – 782.

Geladi, P., MacDougall, D., & Martens, H. 1985. Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat. *Applied Spectroscopy*, **39**(3), 491–500.

Geladi, P., Hadjiiski, L., & Hopke, P. 1999. Multiple regression for environmental data: nonlinearities and prediction bias. *Chemometrics and Intelligent Laboratory Systems*, **47**(2), 165–173.

Gobrecht, A., Roger, J.J., & Bellon-Maurel, V. 2014a. Chapter Four - Major Issues of Diffuse Reflectance NIR Spectroscopy in the Specific Context of Soil Carbon Content Estimation: A Review. *Pages 145–175 of:* Sparks, Donald L. (ed), *Advances in Agronomy 123*, vol. 123. Academic Press.

Gobrecht, A., Bendoula, R., Roger, J.M., & Bellon-Maurel, V. 2014b. Combining linear polarization spectroscopy and the Representative Layer Theory to measure the Beer-Lambert Law Absorbance of highly scattering materials. *Analytica Chimica Acta*, –.

Gogé, F., Joffre, R., Jolivet, C., Ross, I., & Ranjard, L. 2012. Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. *Chemometrics and Intelligent Laboratory Systems*, **110**(1), 168–176.

Hadoux, Xavier, Gorretta, Nathalie, Roger, Jean-Michel, Bendoula, Ryad, & Rabatel, Gilles. 2014. Comparison of the efficacy of spectral pre-treatments for wheat and weed discrimination in outdoor conditions. *Computers and Electronics in Agriculture*, **108**, 242–249.

Hart, J.R., Norris, K.H., & Golumbic, C. 1962. Determination of the Moisture Content of Seeds by Near-Infrared Spectrophotometry of Their Methanol Extracts. *Cereal Chem*, **39**, 94–99.

Hebden, J. C., Arridge, S. R., & Delpy, D.T. 1997. Optical imaging in medicine: I. Experimental techniques. *Physics in Medicine and Biology*, **42**(5), 825.

Igne, B., Reeves III, J.B., McCarty, G. W., Hively, D.E.L., & Hurburgh Jr, C. R. 2010. Evaluation of spectral pretreatments, partial least squares, least squares support vector

machines and locally weighted regression for quantitative spectroscopic analysis of soils. *Journal of Near Infrared Spectroscopy*, **18**(3), 167–176.

Janik, L. J., Forrester, S. T., & Rawson, A. 2009. The prediction of soil chemical and physical properties from mid-infrared spectroscopy and combined partial least-squares regression and neural networks (PLS-NN) analysis. *Chemometrics and Intelligent Laboratory Systems*, **97**(2), 179–188.

Jin, J.W., Chen, Z.P., Li, L.M, Steponavicius, R., Thennadil, S. N., Yang, J., & Yu, R.Q. 2012. Quantitative Spectroscopic Analysis of Heterogeneous Mixtures: The Correction of Multiplicative Effects Caused by Variations in Physical Properties of Samples. *Analytical Chemistry*, **84**(1), 320–326.

Kessler, W., Oelkrug, D., & Kessler, R. 2009. Using scattering and absorption spectra as MCR-hard model constraints for diffuse reflectance measurements of tablets. *Analytica Chimica Acta*, **642**(1 - 2), 127 – 134.

Kleinbaum, D. G., Kupper, L.L., Nizam, A., & Muller, K.E. 2008. *Applied Regression Analysis and Other Multivariable Methods.* Belmont, CA.: Thomson Brooks/Cole.

Kowalski, Bruce R., & Wold, Svante. 1982. 31 Pattern recognition in chemistry. *Pages 673–697 of:* Krishnaiah, P. R., & Kanal, L. N. (eds), *Handbook of Statistics*, vol. Volume 2. Amsterdam (NL): Elsevier.

Kuang, B., Mahmood, H.S., Quraishi, M.Z., Hoogmoed, W.B., Mouazen, A.M., & van Henten, E.J. 2012. Chapter four - Sensing Soil Properties in the Laboratory, In Situ, and On-Line: A Review. *Pages 155 – 223 of:* Sparks, Donald L. (ed), *Advances in Agronomy 114.* Advances in Agronomy, vol. 114. Burlington: Academic Press.

Kubelka, P., & Munk, F. 1931. Ein Beitrag zur Optik der Farbanstriche. *Zeitschrift. für technische Physik*, **12**, 593–601.

Lal, R. 2004a. Soil Carbon Sequestration Impacts on Global Climate Change and Food Security. *Science*, **304**(5677), 1623–1627.

Lal, R. 2004b. Soil carbon sequestration to mitigate climate change. *Geoderma*, **123**, 1 – 22.

Limpert, E., Stahel, W. A., & Abbt, M. 2001. Log-normal Distributions across the Sciences: Keys and Clues. *BioScience*, **51**(5), 341–352.

Lorber, Avraham, Faber, Klaas, & Kowalski, Bruce R. 1997. Net analyte signal calculation in multivariate calibration. *Analytical Chemistry*, **69**(8), 1620–1626.

Lu, B., Morgan, S. P., Crowe, J. A., & Stockford, I. M. 2006. Comparison of Methods for Reducing the Effects of Scattering in Spectrophotometry. *Applied Spectroscopy*, **60**(10), 1157–1166.

MacDougall, D., & Crummett, W. B. 1980. Guidelines for data acquisition and data quality evaluation in environmental chemistry. *Analytical Chemistry*, **52**(14), 2242–2249.

Martens, H.and Stark, E. 1991. Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis*, **9**(8), 625–635.

Martens, H., & Næs, T. 1989. *Multivariate Calibration.* New York, NY: John Wiley & Sons.

Martens, H., Nielsen, J. P., & Engelsen, S. B. 2003. Light Scattering and Light Absorbance Separated by Extended Multiplicative Signal Correction. Application to Near-Infrared Transmission Analysis of Powder Mixtures. *Analytical Chemistry*, **75**(3), 394–404.

Massart, D.L., Vandeginste, B. G. M., Buydens, L. M. C., De Jong, J., Lewi, P.J., & Smeyers-Verbeke, J. 1998. Chapter 8 Straight line regression and calibration. *Pages 171–230 of: Handbook of Chemometrics and Qualimetrics: Part A*, vol. Part A. Amsterdam: Elsevier.

Massie, D. R., & Norris, K. H. 1965. Spectral reflectance and transmittance properties of grain in the visible and near infrared. *Transactions of the ASABE*, **8 (4)**, 598–600.

McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. 2003. On digital soil mapping. *Geoderma*, **117**(1-2), 3–52.

Minasny, B., & McBratney, A. B. 2008. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, **94**(1), 72–79.

Minasny, B., McBratney, A. B., Bellon-Maurel, V., Roger, J.-M., Gobrecht, A., Ferrand, L., & Joalland, S. 2011. Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon. *Geoderma*, **167-168**(0), 118–124.

Morgan, C.L. S., Waiser, T. H., Brown, D. J., & Hallmark, C. T. 2009. Simulated in situ characterization of soil organic and inorganic carbon with visible near-infrared diffuse reflectance spectroscopy. *Geoderma*, **151**(3-4), 249–256.

Morgan, S.P., & Ridgway, M.E. 2000. Polarization properties of light backscattered from a two layer scattering medium. *Optics Express*, **7**(12), 395–402.

Mouazen, A. M., Kuang, B., De Baerdemaeker, J., & Ramon, H. 2010. Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. *Geoderma*, **158**(1-2), 23–31.

Næs, T., Isaksson, T., & Kowalski, B. 1990. Locally weighted regression and scatter correction for near-infrared reflectance data. *Analytical Chemistry*, **62**(7), 664–673.

Næs, T., Isaksson, T., Fearn, T., & Davies, T. 2002. *A User-Friendly Guide to Multivariate Calibration and Classification*. Chichester (UK): IMP publication.

Olivieri, A.C., Faber, N.M., Ferré, J., Boqué, R., Kalivas, J.H., & Mark, H. 2006. Uncertainty estimation and figures of merit for multivariate calibration (IUPAC Technical Report). *Pure and Applied Chemistry*, **78**(3), 633–661.

Osborne, J., & Waters, E. 2002. Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation*, **8**(2), 1–9.

Parkin, T. B., Meisinger, J. J., Chester, S. T., Starr, J. L., & Robinson, J. A. 1988. Evaluation of statistical estimation methods for lognormally distributed variables. *Soil Science Society of America Journal*, **52**(2), 323–329.

Pasikatan, M. C., Steele, J. L, Spillman, C. K, & Haque, E. 2001. Near infrared reflectance spectroscopy for online particle size analysis of powders and ground materials. *Journal of Near Infrared Spectroscopy*, **9**(3), 153–164.

Prahl, S. A. 1995. The adding-doubling method. *Pages 101–129 of: Optical-thermal response of laser-irradiated tissue*. Springer.

Preys, S., Roger, J. M., & Boulet, J. C. 2008. Robust calibration using orthogonal projection and experimental design. Application to the correction of the light scattering effect on turbid NIR spectra. *Chemometrics and Intelligent Laboratory Systems*, **91**(1), 28–33.

Pérez-Marín, D., Fearn, T., Guerrero, J. E., & Garrido-Varo, A. 2012. Improving NIRS predictions of ingredient composition in compound feedingstuffs using Bayesian non-parametric calibrations. *Chemometrics and Intelligent Laboratory Systems*, **110**(1), 108–112.

Quinlan, J.R. 1992. Learning with continuous classes. *Pages 343–348. of: AI92, 5th Australian Joint Conference on Artificial Intelligence*. Hobart, Tasmania: World Scientific.

Reeves III, J. B. 2009. Near- versus mid-infrared diffuse reflectance spectroscopy for soil

analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done? *Geoderma*, **158**(1-2), 3–14.

Reimann, C., & Filzmoser, P. 2000. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. *Environmental Geology*, **39**(9), 1001–1014.

Rinnan, A., van den Berg, F., & Engelsen, S. B. 2009. Review of the most common preprocessing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, **28**(10), 1201–1222.

Roger, J. M., Palagos, B., Bertrand, D., & Fernandez-Ahumada, E. 2011. CovSel: Variable selection for highly multivariate and multi-response calibration: Application to IR spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, **106**(2), 216–223.

Roger, J.M., Chauchard, F., & Bellon-Maurel, V. 2003. EPO-PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits. *Chemometrics and Intelligent Laboratory Systems*, **66**(2), 191–204.

Roger, J.M., Palagos, B, Guillaume, S., & Bellon-Maurel, V. 2005. Discriminating from highly multivariate data by focal eigen function discriminant analysis; application to NIR spectra. *Chemometrics and Intelligent Laboratory Systems*, **79**(1), 31–41.

Saenger, A., Cécillon, L., Sebag, D., & Brun, J.J. 2013. Soil organic carbon quantity, chemistry and thermal stability in a mountainous landscape: A Rock-Eval pyrolysis survey. *Organic Geochemistry*, **54**(0), 101 – 114.

Sankey, J. B., Brown, D.J., Bernard, M.L., & Lawrence, R.L. 2008. Comparing local vs. global visible and near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay, organic C and inorganic C. *Geoderma*, **148**(2), 149–158.

Savitsky, A., & Golay, M.J.E. 1964. Smoothing and differentiation of data by simplified least-squares procedures. *Analytica Chimica Acta*, **36**, 1627 – 1639.

Schmitt, J.M., Gandjbakhche, A.H., & Bonner, R.F. 1992. Use of polarized light to discriminate short-path photons in a multiply scattering medium. *Applied Optics*, **31**(30), 6535–6546.

Seasholtz, M. B., & Kowalski, B.R. 1992. The effect of mean centering on prediction in multivariate calibration. *Journal of Chemometrics*, **6**(2), 103–111.

Seasholtz, M. B., & Kowalski, B.R. 1993. The parsimony principle applied to multivariate calibration. *Analytica Chimica Acta*, **277**(2), 165–177.

Shenk, J. S., Westerhaus, M. O., & Berzaghi, P. 1997. Investigation of a LOCAL calibration procedure for near infrared instruments. *Journal of Near Infrared Spectroscopy*, **5**(4), 223–232.

Shepherd, K. D., & Walsh, M.G. 2002. Development of Reflectance Spectral Libraries for Characterization of Soil Properties. *Soil Science Society of America Journal*, **66**(3), 988–998.

Shi, Z., & Anderson, C. A. 2010. Pharmaceutical applications of separation of absorption and scattering in near-infrared spectroscopy (NIRS). *Journal of Pharmaceutical Sciences*, **99**(12), 4766–4783.

Smith, P. 2004. Monitoring and verification of soil carbon changes under Article 3.4 of the Kyoto Protocol. *Soil Use and Management*, **20**(2), 264–270.

Sokolov, K., Drezek, R., Gossage, K., & Richards-Kortum, R. 1999. Reflectance spectroscopy with polarized light: isit sensitive to cellular and nuclear morphology. *Optics Express*, **5**(13), 302–317.

Stenberg, B. 2010. Effects of soil sample pretreatments and standardised rewetting as interacted with sand classes on Vis-NIR predictions of clay and soil organic carbon. *Geoderma*, **158**(1-2), 15–22.

Stenberg, B., & Viscarra-Rossel, R.A. 2010. Diffuse Reflectance Spectroscopy for High-Resolution Soil Sensing. *Pages 29–47 of:* Viscarra Rossel, Raphael A., McBratney, Alex B., & Minasny, Budiman (eds), *Proximal Soil Sensing*. Progress in Soil Science, vol. 1. Springer Netherlands.

Stenberg, Bo, Rossel, Raphael A. Viscarra, Mouazen, Abdul Mounem, & Wetterlind, Johanna. 2010. Chapter Five - Visible and Near Infrared Spectroscopy in Soil Science. *Pages 163 – 215 of:* Sparks, Donald L. (ed), *Advances in Agronomy 107*. Advances in Agronomy, vol. 107. Burlington: Academic Press.

Steponavicius, R., & Thennadil, S. N. 2009. Extraction of Chemical Information of Suspensions Using Radiative Transfer Theory to Remove Multiple Scattering Effects: Application to a Model Two-Component System. *Analytical Chemistry*, **81**(18), 7713–7723.

Steponavicius, R., & Thennadil, S. N. 2011. Extraction of Chemical Information of Suspensions Using Radiative Transfer Theory To Remove Multiple Scattering Effects: Application to a Model Multicomponent System. *Analytical Chemistry*, **83**(6), 1931–1937.

Stevens, A., Udelhoven, T., Denis, A., Tychon, B., Lioy, R., Hoffmann, L., & van Wesemael, B. 2010. Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. *Geoderma*, **158**(1-2), 32–45.

Stockford, I. M., Lu, B., Crowe, J. A., Morgan, S. P., & Morris, D. E. 2007. Reduction of Error in Spectrophotometry of Scattering Media Using Polarization Techniques. *Applied Spectroscopy*, **61**(12), 1379–1389.

Swartling, J., Dam, J.S., & Andersson-Engels, S. 2003. Comparison of spatially and temporally resolved diffuse-reflectance measurement systems for determination of biomedical optical properties. *Applied Optics*, **42**(22), 4612–4620.

Thennadil, S. N., & Martin, E. B. 2005. Empirical preprocessing methods and their

impact on NIR calibrations: a simulation study. *Journal of Chemometrics*, **19**(2), 77–89.

Thennadil, S.N. 2008. Relationship between the Kubelka-Munk scattering and radiative transfer coefficients. *Journal of the Optical Society of America A*, **25**(7), 1480–1485.

Torrance, S.E.a, Sun, Z.b, & Sevick-Muraca, E.M.a. 2004. Impact of excipient particle size on measurement of active pharmaceutical ingredient absorbance in mixtures using frequency domain photon migration. *Journal of Pharmaceutical Sciences*, **93**(7), 1879–1889.

Udelhoven, T., & Schütt, B. 2000. Capability of feed-forward neural networks for a chemical evaluation of sediments with diffuse reflectance spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, **51**(1), 9–22.

Vasques, G. M., Grunwald, S., & Sickman, J. O. 2008. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. *Geoderma*, **146**(1-2), 14–25.

Verron, T., Sabatier, R., & Joffre, R. 2004. Some theoretical properties of the O-PLS method. *Journal of Chemometrics*, **18**(2), 62–68.

Viscarra Rossel, R. A. 2008. The Soil Spectroscopy Group and the development of a global soil spectral library. *NIR News*, **20**(4), 14–15.

Viscarra Rossel, R. A., & Behrens, T. 2010. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*, **158**(1-2), 46–54.

Viscarra Rossel, R. A., Walvoort, D. J. J., McBratney, A. B., Janik, L. J., & Skjemstad, J. O. 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*, **131**(1-2), 59–75.

Viscarra Rossel, R.A., Fouad, Y., & Walter, C. 2008. Using a digital camera to measure soil organic carbon and iron contents. *Biosystems Engineering*, **100**(2), 149–159.

Vistelius, A.B. 1960. The Skew Frequency Distributions and the Fundamental Law of the Geochemical Processes. *The Journal of Geology*, **68**(1), 1–22.

Vohland, M., Besold, J., Hill, J., & Fründ, H.C. 2011. Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. *Geoderma*, **166**(1), 198–205.

Voit, F., Hohmann, A., Schäfer, J., & Kienle, A. 2012. Multiple scattering of polarized light: comparison of Maxwell theory and radiative transfer theory. *Journal of Biomedical Optics*, **17**(4), 045003–1–045003–8.

Webster, R. 2001. Statistics to support soil research and their presentation. *European Journal of Soil Science*, **52**(2), 331–340.

Wendlandt, W. W., & Hecht, H. G. 1966. *Reflectance spectroscopy*. Interscience New York.

Wetterlind, J., & Stenberg, B. 2010. Near-infrared spectroscopy for within-field soil characterization: small local calibrations compared with national libraries spiked with local samples. *European Journal of Soil Science*, **61**(6), 823–843.

Williams, P., & Norris, K. 2001. *Near-Infrared Technology in the Agricultural and Food Industries (2nd Ed.)*. Amer Assn of Cereal Chemists, St. Paul, Minnesota.

Wold, S. 1978. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, **20**, 397–405.

Wold, S., Antti, H., Lindgren, F., & Öhman, J. 1998. Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, **44**(1-2), 175–185.

Wold, S., Sjöström, M., & Eriksson, L. 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, **58**(2), 109–130.

Workman, J., & Springsteen, A. 1998. *Applied Spectroscopy: A Compact Reference for Practitioners*. Academic Press, Boston, USA.

Yoo, K.M., & Alfano, R.R. 1989. Time resolved depolarization of multiple backscattered light from random media. *Physics Letters A*, **142**(8-9), 531–536.

Zeaiter, M., Roger, J. M., & Bellon-Maurel, V. 2006. Dynamic orthogonal projection. A new method to maintain the on-line robustness of multivariate calibrations. Application to NIR-based monitoring of wine fermentations. *Chemometrics and Intelligent Laboratory Systems*, **80**(2), 227–235.

Zhang, L., & Garcia-Munoz, S. 2009. A comparison of different methods to estimate prediction uncertainty using Partial Least Squares (PLS): A practitioner's perspective. *Chemometrics and Intelligent Laboratory Systems*, **97**(2), 152–158.

RESUME

Avec l'objectif de réduire de la quantité de gaz à effets de serre dans l'atmosphère, les pouvoirs publics encouragent les pratiques ayant vocation à séquestrer le carbone dans les sols (reforestation, changement de pratiques agricoles). Pour en évaluer les réels bénéfices, des outils analytiques rapides, précis et peu coûteux sont nécessaires pour pouvoir comptabiliser précisément les stocks de carbone et leur évolution dans le temps. La Spectroscopie proche infrarouge est une technologie analytique adaptée à ce cahier des charges mais qui relève encore du domaine de la recherche en science du sol.

Cette thèse s'est focalisée sur la première étape de cette méthode analytique: la formation du signal. Les sols étant des milieux très complexes, en termes de composition chimique et de structure physique, le signal spectroscopique est négativement impacté par les phénomènes de diffusion. Les conditions de la loi de Beer-Lambert n'étant plus remplies, les modèles chimiométriques pour prédire la teneur en carbone des sols sont moins précis et robustes. Nous proposons un système optique de mesure spectrale original et adapté aux milieux très diffusants, qui se base sur le principe de polarisation de la lumière. Il permet de sélectionner les photons ayant été moins impactés par le phénomène de diffusion. Ce signal est utilisé pour calculer un signal d'absorbance étant une bonne approximation de l'absorbance de Beer-Lambert.

Ce dispositif, appelé PoLiS, a été validé expérimentalement sur des milieux modèles liquides et particulaires. La méthode PoLiS a été testée sur des échantillons de sols pour prédire leur teneur en carbone organique. En comparaison avec les méthodes classiques d'étalonnages, les modèles de prédiction présentent de meilleurs résultats avec la méthode développée dans cette thèse.

Mots clés : Spectrométrie Visible et Proche Infrarouge - Polarisation de la lumière - Diffusion - Sols - Carbone -

ABSTRACT

With the goal of reducing the amount of greenhouse gases in the atmosphere, policy makers encourage practices intended to sequester carbon in soils (reforestation, changes in farming practices). New methods are required to rapidly and accurately measure soil C at field- and landscape-scales. Near infrared spectroscopy (NIRS) is an analytical technology adapted to these specifications but remains experimental research in soil science.

This thesis has focused on the first step of this analytical method: signal formation. The soils are very complex materials, in terms of chemical composition and physical structure. Hence, the spectroscopic signal is negatively impacted by light scattering. Consequently, the conditions of the Beer-Lambert are no longer fulfilled, and the chemometric models to predict the carbon content of soils are less accurate and robust. We develop an original optical method based on light polarization spectroscopy to measure the absorbance of highly scattering materials. By selecting photons being less scattered, we compute a new absorbance signal which is a good approximation of the Beer-Lambert absorbance.

This method, called Polis, was experimentally validated on model materials in liquid and powdered form. Applied on soils to predict Total Organic Content, the model built with the PoLiS absorbance outperform the models built with the classical absorbance computed from the diffuse reflectance signal.

Keywords : Visible and Near Infrared Spectroscopy - Light Polarization - Scattering - Soil - Carbon -

This thesis is part of the research project INCA (In-field Spectroscopy for Carbon Accounting), financially supported by ADEME (Agency for the Environment and Energy Management).

Alexia GOBRECHT
Irstea - UMR ITAP - COMiC - Capteurs Optiques pour les Milieux Complexes
351 rue Jean François Breton - 34196 Montpellier Cedex 5 (France).