RAPPORT DE STAGE

présenté par : Moubarak ISMAIL HOSKY

AOÛT 2024

TRANSFERT D'ÉTALONNAGE ENTRE DIFFÉRENTS SPECTROMÈTRES EN RÉFLEXION DIFFUSE DANS LE VISIBLE ET PROCHE INFRAROUGE APPLIQUÉ AU SOLS

ENCADRANT(E)S : Aurélie CAMBOU
Jean Michel ROGER
Bernard BARTHÈS

M2 Ingénierie Statistique

Remerciements

Je tiens à exprimer ma profonde gratitude pour l'opportunité que l'Institut de Recherche pour le Développement (IRD) m'as offerte pour la réalisation de mon stage de fin d'étude au sein de L'UMR ECO&SOL. Cette expérience a été extrêmement enrichissante sur le plan professionnel comme personnel, et a marqué une étape importante dans mon parcours académique.

Je tiens tout particulièrement à remercier **Aurélie Cambou**, Sa disponibilité, ses remarques constructives, son humanité et son écoute attentive à mes propositions ont été précieuses pour améliorer ce travail. Son professionnalisme et son leadership ont été des clés de voûte pour l'aboutissement de ce stage. Grâce à elle, j'ai pu découvrir le monde de la recherche d'une manière globale et concrète, ce qui a largement contribué à mon développement professionnel et ma vision dans ce noble domaine .

Je remercie également **Jean-Michel Roger** et **Bernard Barthes** pour leur soutien et leurs conseils avisés tout au long de cette expérience.

Je souhaite également adresser mes sincères remerciements à **Anne Philippe** et **Aymeric Stamm**, mes référents universitaires, pour leur encadrement et leur soutien pendant toute la durée du stage.

Un grand merci à **Gilles Chaix** pour sa présence et ses conseils éclairés, qui ont permis d'ouvrir de nouvelles perspectives d'amélioration. Enfin, je tiens à exprimer ma reconnaissance à **Alexandre Eymard** pour nos échanges constructifs et son soutien indéfectible tout au long de ce stage.

Moubarak Ismail Hosky

Sigles et Abréviations

— **ACP** : Analyse en composante principale

— **BDD** : Base de données

— **BF** Base de données spectrale acquise avec le spectromètre basé à Ouagadougou au Burkina Faso

— **CAL** : Calibration

— **CBP** : Correction biais pente

— **COS** : Carbone Organique du Sol

— **IR** : Infrarouge

— **MOS** : Matière Organique du Sol

— **MPL** Base de données spectrale acquise avec le spectromètre basé à Montpellier en France

— **PDS** : Piecewise Direct Standardisation (Standardisation directe par segment)

— **PLSR** : Partial Least Square Regression

— $\mathbf{R^2}$: Coefficient de détermination

— **RPD** : Prediction to Deviation Ratio (Rapport de Prédiction)

— **RPIQ** : Rapport de la performance à l'interquartile

— **RMSE** : Root Mean Square Error (Erreur Quadratique Moyenne)

— **RMSEC** : Root Mean Square Error of calibration (Erreur Quadratique Moyenne de la calibration)

— **RMSECV** : Root Mean Square Error of Cross-Validation (Erreur Quadratique Moyenne de la validation croisée)

— **RMSEP** : Root Mean Square Error pf prediction (Erreur Quadratique Moyenne de prédiction)

— **SEL** : Standard Error of Laboratory

— **SEP** : Squared Error of Prediction (Erreur Standard de Prédiction)

— **SN** Base de données spectrale acquise avec le spectromètre basé à Dakar au Sénégal

— **STD** : Standard (jeu de transfert)

— **VAL** (jeu de Validation du Burkina Faso)

— **VAL-BRE** (jeu de Validation du Brésil)

— **VAL-BF** (jeu de Validation de la base de données spectrale **BF**)

— **VAL-MPL** (jeu de Validation de la base de données spectrale **MPL**)

— **VAL-SN** (jeu de Validation de la base de données spectrale **SN**)

— **VPIR** : Visible-Proche Infrarouge

Avertissement ! avant la lecture de ce rapport : Le terme "échantillon" utilisé dans ce stage ne désigne pas un échantillon au sens propre en statistique, qui est une sous-population. Il fait plutôt référence à une **observation** ou à un individu statistique."

Table des matières

1 Introduction

L'impact des activités anthropiques sur le réchauffement climatique, à travers l'émission de gaz à effet de serre (GES ; par exemple, le dioxyde de carbone, CO2), est aujourd'hui incontestable. D'après le 6ème rapport du groupe d'experts intergouvernemental sur l'évolution du climat (GIEC), la température globale mesurée pendant la période 2011-2020 a augmenté d'environ 1,1°C par rapport à celle de la période préindustrielle (1850-1900 ; GIEC, 2023). Une solution qui avait été proposée lors de la COP21 en 2015 pour limiter l'émission de GES, était de séquestrer du carbone organique dans les sols (COS). En effet, le sol, qui constitue la couche superficielle de la croûte terrestre, d'une épaisseur allant de 10 cm jusqu'à 10-20 m, représente le plus gros réservoir continental de carbone organique. Or, en fonction du type de couvert ou du mode d'occupation, il peut se comporter comme un puits ou une source de CO2 dans l'atmosphère (Jacobson et al., 1993 ; Batjes, 1996 ; Houghton, 2003 ; Derrien et al., 2016).

Le COS est un élément majeur de la matière organique du sol (MOS, qui représente environ 1-5% des constituants du sol) (Gerzabek et al,2005). Or la MOS joue un rôle clé dans la fertilité physique (agrégation, rétention en eau, aération, etc.), chimique (fourniture en nutriments, pouvoir tampon, etc.) et biologique (activité et diversité microbiennes, animales et végétales ; Lal, 2014 ; Zhao et al,2015) des sols. La séquestration du COS représente donc un enjeu crucial, d'autant qu'il s'agit d'un indicateur clé de la qualité des sols (Reeves, 1997).

Dans cette optique, de nombreuses techniques agronomiques ont été mises en place pour favoriser la séquestration du COS comme les forêts en jachères, ou la mise en place de systèmes agroforestiers (Lal, 2004, Lorenz and Lal, 2014) etc.

L'étude du COS dans le temps et dans l'espace constitue donc un enjeu crucial. Cependant, les techniques conventionnelles permettant de quantifier le COS sont longues, coûteuses et destructives. La spectroscopie en réflexion diffuse dans le visible et proche infrarouge (VPIR) est utilisée comme complément ou alternative aux méthodes conventionnelles en science des sols depuis une trentaine d'années, justement du fait que cette méthode est rapide, avec un faible coût unitaire et non destructive (Barthès et al., 2023).

Le principe de la spectroscopie VPIR se base sur le phénomène quantique suivant : un rayonnement dans le visible et proche infrarouge (350-2500 nm) est appliqué à un échantillon. Ce rayonnement induit des vibrations des liaisons moléculaires, qui absorbent l'énergie lumineuse à des longueurs d'ondes spécifiques, en fonction de la force des liaisons, des types d'atomes concernés, et de l'environnement chimique (Siesler, 2007). Ainsi, la lumière VPIR réfléchie par un échantillon solide

est différente de la lumière incidente en ce qui concernent les régions spectrales absorbées par les liaisons chimiques de cet échantillon ; ce rayonnement réfléchi contient donc de l'information sur la composition chimique de cet échantillon.

La spectroscopie VPIR utilise des étalonnages (ou calibrations) reliant la mesure analytique d'une propriété de l'échantillon (par exemple, la teneur en COS) à son spectre VPIR. Cet étalonnage requiert des échantillons caractérisés spectralement et conventionnellement, afin de construire par régression multivariée une relation statistique entre propriété considérée et information portée par le spectre VPIR. L'étalonnage peut ensuite être appliqué sur de nouveaux échantillons de même nature pour prédire la propriété uniquement d'après leur spectre, rapidement (quelques secondes pour « scanner » un échantillon) et à faible coût unitaire (ni consommables ni déchets). Plusieurs approches mathématiques et statistiques peuvent être testées en analyse de données spectrales (chimiométrie) pour optimiser l'estimation du paramètre étudié. Parmi ces méthodes, figurent les approches non supervisées de visualisation spectrale (ACP) et des approches supervisées permettant l'étalonnage du modèle (par exemple la régression des moindres carrés partiels, ou partial least squares régression, PLSR) ...

La spectroscopie VPIR a montré une bonne capacité à prédire la teneur en COS (gC.kg-1 sol ; Viscarra Rossel et Webster, 2012 ; Clairotte et al., 2016 ; Barthès et al., 2023) ; elle offre donc des réelles opportunités pour quantifier le COS à haut débit et à faible coût.

La spectroscopie VPIR n'échappe pas au phénomène de production galopante des données dans notre ère du numérique et de la technologie. Depuis les premières applications de cette méthode en science du sol, de nombreuses bases de données (BDD) ont été construites dans différents laboratoires de différentes régions du monde. Il existe d'ailleurs beaucoup d'appareils de mesure (spectromètres) de marques différentes (par exemple, Foss, Bruker, ASD, Shimadzu), avec des technologies et des prix différents.

Dans un contexte où les instituts de recherche publique incitent à l'ouverture des données (Open Data), il est nécessaire d'envisager la mise à disposition et l'interopérabilité de BDDs spectrales obtenues avec différents appareils de mesure, à différentes périodes et dans différents laboratoires. Cette ouverture des BDDs permettrait notamment d'élargir les capacités de prédiction de la teneur en COS dans une diversité de contextes. Les perturbations liées au contexte d'acquisition (par exemple, différentes méthodes de préparation des échantillons), au vieillissement des instruments et surtout à la diversité des instruments limitent l'interopérabilité des BDD spectrales. Les erreurs engendrées par ces variations sont des grandeurs d'influence car les résultats de prédictions dépendent de la qualité des données. Il est donc nécessaire de corriger ces erreurs par le biais de corrections

mathématiques appelées transferts d'étalonnage.
Au sein de l'UMR Eco&Sols (Ecologie fonctionnelle & biogéochimie des sols & agro-écosystèmes), lieu d'accueil de ce stage, il y a trois spectromètres VPIR présents dans différentes implantations : un spectromètre présent à Montpellier (France), un autre appareil est implémenté à Ouagadougou (Burkina Faso) et autre spectromètre présent à Dakar (Sénégal).La variabilité de ces appareils et leurs contextes d'acquisition pose une réelle problématique à l'homogénéisation et partage des BDDs spectrales au sein de cette unité et avec ses partenaires scientifiques.
Face à ces enjeux, ce stage a été proposé par les UMR Eco&Sols et ITAP (Technologies et Méthodes pour les Agricultures de demain) afin de tester la capacité des méthodes de transfert d'étalonnage entre spectromètres VPIR pour réduire au maximum les erreurs liées à ces différentes perturbations.
L'objectif de ce stage était de tester différentes méthodes de transfert d'étalonnage pour optimiser l'interopérabilité des BDDs spectrales obtenues avec trois spectromètres de gamme VPIR de marque ASD, sur des échantillons de sols préparés au laboratoire (séchés et tamisés < 2 mm) afin de prédire la teneur en COS.
Ce stage a permis de renforcer les travaux menés en 2023 dans le cadre d'un stage M2 supervisé par Aurélie Cambou (stage de Vova Martirosyan) portant sur le transfert d'étalonnage entre deux appareils proche infrarouge (1100-2500 nm) de marques différentes (FOSS et ASD). De plus, il a permis de lever des verrous scientifiques étant donné qu'à notre connaissance, il n'existe pas à ce jour de travaux publiés sur le transfert d'étalonnage entre plusieurs spectromètres VPIR pour prédire les propriétés organiques des sols.

2 Matériels et Méthodes

2.1 Composition du jeu d'échantillons de sols

Le jeu d'échantillons a été choisi avant mon stage (stage M2 de Vova Martirosyan en 2023) de sorte d'être représentatif d'une partie de la base des données (BDD) spectrales de l'UMR Eco&Sols (les échantillons de sols ayant été stockés depuis les années 1990 et étant encore disponibles dans la pédothèque Bernard G. Barthès). Ce jeu d'échantillons comprenait au total 178 échantillons provenant de cinq pays d'Afrique de l'ouest (Côte d'Ivoire, Mali, Burkina Faso, Bénin, Sénégal), de deux pays d'Afrique centrale (Cameroun, Congo), d'un pays d'Amérique du Sud (Brésil), et de Madagascar (Figure 1). Au sein de ces pays, plusieurs sites ont parfois été étudiés dans différentes régions amenant à un total de 22 sites d'étude.

FIGURE 1 – Localisation géographique des pays d'étude

2.1.1 Jeux d'échantillons d'étalonnage et de transfert

Le jeu dit d'apprentissage (99 échantillons) était composé d'un jeu d'échantillons d'étalonnage (ou jeu de calibration, CAL ; permettant de construire un modèle de prédiction d'une variable d'intérêt) et d'un jeu d'échantillons de transfert (ou jeu standard, STD ; permettant de construire un modèle de transfert d'étalonnage). La séparation entre les jeux CAL et STD avait été réalisée lors d'une étude précédant mon stage (stage M2 Vova Martirosyan, 2023). Pour cela, 40 échantillons ont d'abord été sélectionnés parmi les 99 échantillons initiaux, de sorte d'être représentatifs des différents sites d'étude, des profondeurs de sol (couches de 0-5 cm à 60-80 cm), des types de sol et textures du sol (e.g., Ferralsol argileux, Lixisol sableux) et des différents types d'usage (e.g., forêt, culture, savane) rencontrés dans le jeu d'apprentissage. Dans un second temps, afin de réduire ce jeu à environ 30 échantillons, une analyse en composantes principales (ACP) a été effectuée à partir des données spectrales des 99 échantillons. Pour cela, les 40 échantillons précédemment sélectionnés ont été distingués par leur couleur dans le plan des deux premières composantes principales (CP) et lorsque deux d'entre eux étaient très proches visuellement dans ce plan, l'un a été déplacé dans le jeu CAL car jugé redondant pour le transfert ; ainsi neuf échantillons ont été déplacés dans le jeu CAL. Le jeu STD comprenait finalement 31 échantillons et le jeu CAL 68 échantillons. Les jeux CAL et STD provenaient de 21 et 20 sites, respectivement, représentant tous les pays de cette étude. Au sein de ces deux jeux, le climat entre les sites d'étude est variable : la température annuelle moyenne est comprise entre 16 et 29°C et et les précipitations annuelles moyennes entre 600 et 2000 mm. La gamme des textures est également large allant de sableuse (e.g., échantillons provenant du nord du Burkina Faso, ou encore des plateaux Bakété au Congo) à argileuse (e.g. région du Paraná au Brésil ou dans la vallée du Niari au Congo).

2.1.2 Jeu d'échantillons de validation du Burkina Faso (VAL)

Le jeu d'échantillons de validation externe (jeu VAL ; permettant de valider le modèle de prédiction de la variable d'intérêt) est issu du nord du Burkina Faso (près de Torokoro). Il a été choisi avant le début de mon stage afin d'être géographiquement indépendant des sites présents dans les jeux CAL et STD, tout en étant bien représenté par le jeu CAL. Concernant le climat, le site d'étude de VAL est caractérisé par une température annuelle moyenne de 27°C et des précipitations annuelles moyennes de 1100 mm. Le type de sol est un Haplic Lixisol (Dengiz et al., 2018). La texture du sol était sableuse et les principaux types d'usages rencontrés étaient des cultures, forêts et savanes arborées. Ce jeu VAL était constitué de 38 échantillons prélevés à 0-10 cm de profondeur. Parmi ces 38 échantillons, seuls 30 ont été retenus pour la validation ; les huit échantillons restants ont été utilisés pour enrichir le jeu CAL, à travers une technique dite de **spiking** (voir section **2.5**). Le second jeu de validation était situé dans l'Etat de São Paulo, au Brésil (**VAL-BRE** ; n = 41) ; ce dernier ne sera cependant pas abordé dans la suite de ce document du fait d'un manque de temps pour l'étudier.

2.2 Analyse conventionnelle de la teneur en carbone organique des sols

Chacun des 137 échantillons de sol (issus des jeux **CAL**, **STD** et **VAL** ; **VAL-BRE** n'étant pas étudié) était caractérisé par deux sous-échantillons déjà préparés et stockés : le premier sous-échantillon avait été tamisé à 2 mm, le second avait été tamisé à 2 mm puis broyé finement à 0,2 mm. Les sous-échantillons broyés finement ont été séchés à 40°C pendant plus de 24 h afin d'éliminer toute trace d'humidité. Puis, environ 25 mg de chaque échantillon broyé ont été pesés à l'aide d'une balance de précision afin d'analyser leur teneur en carbone total ($gC.kg^{-1}$) selon le principe de la "méthode Dumas" qui implique l'oxydation complète et instantanée de l'échantillon par combustion sèche à l'aide d'un analyseur élémentaire (Pansu et Gautheyrou, 2006) . En absence supposée de carbonates, le carbone total analysé était considéré comme organique.

2.3 Acquisition des spectres

Trois spectromètres de la même marque ASD et avec la même gamme spectrale (VNIR ; 350-2500 nm avec un pas d'un nm) ont été considérés dans cette étude (Figure 2) :

- LabSpec 2500 (nommé LabSpec-MPL dans la suite du document) : ce modèle, présent dans l'implantation d'Eco&Sols à Montpellier, est une version plus ancienne de la gamme LabSpec, conçu pour être utilisé de manière portable ;

- LabSpec 4 (nommé LabSpec-SN dans la suite du document) : ce modèle, présent dans l'implantation partenaire d'Eco&Sols à Dakar (Sénégal), est plus récent que le LabSpec 2500 et est également conçu pour être utilisé de manière portable ;
- LabSpec 4 Bench (nommé LabSpec-BF dans la suite du document) : ce modèle, présent dans l'implantation partenaire d'Eco&Sols à Ouagadougou (Burkina Faso), est spécifiquement conçu pour des applications de laboratoire.

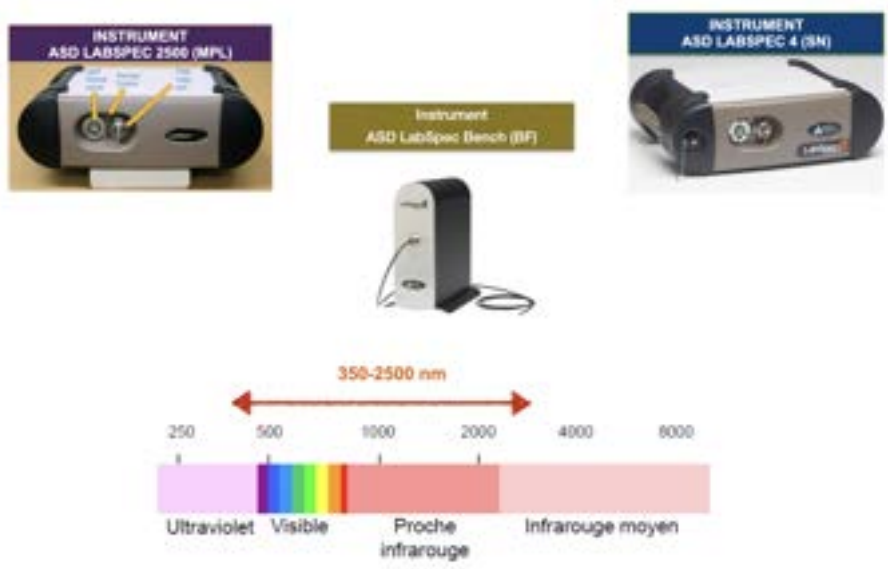


FIGURE 2 – Présentation des trois spectromètres visible et proche infrarouge considérés dans cette étude

Les acquisitions spectrales ont été réalisées (i) dans le laboratoire de Montpellier avec les spectromètres LabSpec-MPL (acquisitions en 2023) et LabSpec-BF (acquisitions au cours de ce stage), (ii) dans le laboratoire de Dakar avec le spectromètre LabSpec-SN (acquisitions en 2023) ; chaque campagne d'acquisitions ayant été réalisée par un opérateur différent. Pour chacun des 137 échantillons, les acquisitions spectrales ont été effectuées à la fois sur le sous-échantillon seulement tamisé et sur le sous-échantillon finement broyé, après un séchage préalable à 40°C pendant 24h. L'acquisition spectrale des sols broyés n'était pas le sujet de cette étude, mais s'inscrit dans un projet à moyen terme de l'équipe d'Eco&Sols ; les spectres des sols broyés ne seront donc pas traités dans la suite du manuscrit. Chaque campagne d'acquisitions spectrales a été réalisée par contact direct entre la sonde et l'échantillon. Un blanc a été réalisé en début d'acquisition à l'aide d'un disque blanc d'absorbance nulle, le Spectralon (poudre de polytétrafluoroéthylène compressée qui

est un réflecteur Lambertien presque parfait). La mesure de sa réflectance a été renouvelée toutes les 10 acquisitions. Concernant les acquisitions spectrales réalisées à Montpellier (LabSpec-MPL et LabSpec-BF), chaque spectre était la résultante de 50 scans réalisés automatiquement par le spectromètre puis moyennés. Pour chaque échantillon tamisé, deux spectres ont été acquis, puis, après vérification visuelle de leur similitude, les deux spectres ont été moyennés. A Dakar (LabSpec-SN), le nombre de scans réalisés automatiquement par le spectromètre puis moyennés ne nous a pas encore été communiqué par l'opérateur. Pour chaque échantillon tamisé, trois spectres ont été acquis, puis, après vérification visuelle de leur similitude, ils ont été moyennés. Les données spectrales, obtenues en réflectance (R ; sans unité) ont été converties en absorbance (A ; sans unité) selon la formule suivante :
$A = \log(\frac{1}{R})$. Trois BDDs spectrales (incluant chacune les jeux CAL, STD et VAL ; n = 137) ont donc été obtenues à l'issue de ces acquisitions spectrales : la BDD MPL acquise avec le LabSpec-MPL, la BDD SN acquise avec le LabSpec-SN et la BDD BF acquise avec le LabSpec-BF.

2.4 Prétraitements spectraux

Les prétraitements spectraux sont essentiels pour améliorer la qualité des données spectrales et rendre les analyses plus fiables. Ils permettent de réduire les effets indésirables causés par les variations instrumentales, les interférences environnementales, et les variations granulométriques des échantillons. Ainsi, les prétraitements spectraux visent à minimiser le bruit, à corriger les dérives de base, à normaliser les spectres, et de ce fait à améliorer la résolution des signaux d'intérêt. Dans le cadre des modèles prédictifs, la qualité d'un modèle dépend de la qualité des données. En chimiométrie, les prétraitements spectraux sont effectués systématiquement.
Dans cette étude, plusieurs techniques ont été appliquées parallèlement sur les trois BDDs spectrales afin d'améliorer la qualité du signal :

* L'algorithme de Savitzky-Golay est utilisé pour lisser des données (Savitzky & Golay, 1964), notamment dans le traitement des spectres, afin de réduire le bruit sans altérer les caractéristiques spectrales importantes, comme les pics. Chaque point (i.e., absorbance à une longueur d'onde donnée) du spectre est lissé en ajustant un polynôme d'un certain ordre sur une fenêtre englobant ce point. Dans notre étude, le degré du polynôme était 2 et les fenêtres testées étaient 11 et 31.Le polynôme était donc de la forme :

$$y(x) = a_0 + a_1 x + a_2 x^2$$

avec a_0, a_1 et a_2 les coefficients de ce polynôme. Ces coefficients sont déterminés par une régression linéaire minimisant l'erreur quadratique sur les points de la fenêtre, ce qui produit une matrice de convolution spécifique. La valeur lissée est obtenue en appliquant cette matrice de convolution aux points de la fenêtre. Une matrice de convolution est un outil mathématique utilisé principalement dans les traitements de signal et d'image pour appliquer des filtres. Elle fonctionne en multipliant les valeurs des pixels ou des points de données d'une région locale (fenêtre) avec des poids spécifiques (les éléments de la matrice de convolution) et en additionnant les résultats pour obtenir une nouvelle valeur. Ce lissage peut être accompagné d'une dérivée. Dans notre étude, le lissage de Savitzky-Golay (polynôme de degré 2 calculé sur une fenêtre de 11 et 31) a été testé sans dérivée, avec dérivée 1ère et dérivée 2nde.

* La correction de ligne de base (Detrend) a également été testée. Cette méthode utilise une régression polynomiale pour ajuster une fonction $T(x)$ au spectre et soustraire cette fonction du spectre original :

$$\sum_{i=1}^{n}(y_i - T(x_i))^2 \tag{1}$$

où y_i est la valeur d'absorbance mesurée au point x_i, et $T(x_i)$ est la valeur du polynôme (d'ordre 1 ou 2) ajustée à ce point. Dans notre étude, la fonction Detrend a été basée sur un polynôme d'ordre 0 (centrage), 1 ou 2.

* La normalisation dite Standard Normal Variate (SNV) a aussi été testée dans cette étude. La normalisation est utilisée pour rendre les spectres comparables entre eux (Barnes et al, 1989). Elle est calculée ainsi pour chaque échantillon :

$$y'_i = \frac{y_i - \bar{y}}{s} \tag{2}$$

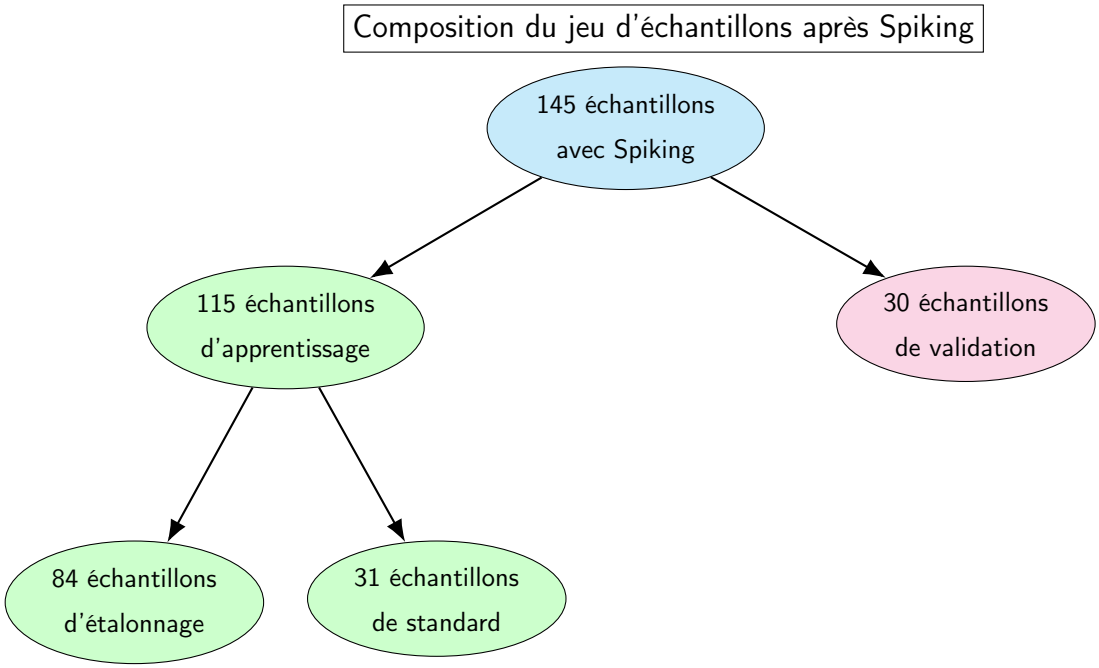
avec y'_i, l'absorbance corrigée à la longueur d'onde i ; $\bar{y}$, la moyenne des absorbances pour l'ensemble des longueurs d'onde ; yi, l'absorbance mesurée à la longueur d'onde i ; et s l'écart-type des absorbances pour l'ensemble des longueurs d'onde.

Ces prétraitements spectraux (lissage de Savitzky-Golay avec ou sans dérivée 1ère ou 2nde, SNV, Detrend) ont été utilisés seuls ou combinés (par exemple, SNV suivi d'un Detrend d'ordre 1), amenant à un total de 20 prétraitements spectraux testés dans cette étude.

2.5 Spiking et surpondération

Le spiking est une technique utilisée en traitement de données, en particulier dans le contexte de la spectroscopie infrarouge et de l'analyse chimique, pour améliorer la précision des modèles de prédiction. Cette méthode consiste à ajouter intentionnellement de petites quantités d'une substance connue (appelée "spike") à un échantillon (Kuang et Mouazen, 2013) . Le but est d'améliorer la robustesse du modèle en renforçant certaines caractéristiques ou en simulant des conditions réelles plus complexes. Le spiking peut aider à corriger des biais dans un modèle de prédiction, notamment lorsqu'un jeu cible (sur lequel le modèle est appliqué) est insuffisamment représenté dans la construction de ce modèle.

Dans cette étude, le spiking a été réalisé via l'algorithme de Kennard-Stone (Kennard et Stone, 1969). Cet algorithme permet de séparer des lots d'échantillons en fonction de leur représentativité spectrale. Son principe est le suivant : une ACP est construite à partir des spectres d'un jeu d'échantillons donné en utilisant un certain nombre de CPs (ici, 10). A partir des scores des échantillons (ou individus) sur cette ACP, l'algorithme identifie d'abord les deux échantillons les plus éloignés (dans le cas de cette étude, la distance considérée était euclidienne) et les injecte dans un nouveau jeu dit "représentatif spectralement". Puis, il identifie l'échantillon qui est le plus éloigné des deux échantillons préalablement sélectionnés et l'ajoute dans le jeu représentatif spectralement. Ce processus se poursuit jusqu'à l'obtention du nombre souhaité d'échantillons représentatifs, assurant ainsi une couverture étendue et représentative de l'ensemble de données. Dans cette étude, l'algorithme de Kennard-Stone a été appliqué parallèlement sur les spectres VAL (n = 38) acquis avec les trois spectromètres (LabSpec-BF, LabSpec-MPL et LabSpec-SN), afin d'identifier huit échantillons en commun, représentatifs spectralement, qui ont été utilisés pour le spiking. Le nouveau jeu VAL comprenait alors 30 échantillons. La méthode dite de surpondération (extra-weighting) proposée par Guerrero et al. (2014) a enfin été appliquée à ces huit échantillons avant leur ajout dans le jeu CAL. Plus précisément, ils ont été artificiellement dupliqués afin d'augmenter leur poids dans le jeu CAL, amenant à 16 échantillons ajoutés dans le jeu CAL (n = 84). La figure 3 résume les jeux d'échantillons obtenus à l'issue de cette étape de spiking et surpondération.



2.6 Présentation des bases de données spectrales

Chaque BDD spectrale obtenue avec les trois spectromètres (MPL, SN et BF) est représentée par une matrice $\mathbf{X_{n \times p}}$ avec :

— **n** : le nombre d'échantillons de sol (un échantillon correspondant à une observation ou un individu statistique dans notre cas) ;

— **p** : le nombre de variables prédictives (ou caractéristiques) ; chaque variable prédictive représente une valeur d'absorbance (sans unité) à une longueur d'onde donnée variant de 350 à 2500 (nm).

Chaque BDD spectrale était constituée de $\mathbf{n} = 145$ échantillons (en incluant les duplicats des huit échantillons de spiking) et $\mathbf{p} = 2150$ absorbances.

Finalement, plusieurs jeux d'échantillons sont issus de ce travail :

◇ Les jeux CAL (n = 84) issus des BDDs spectrales MPL, SN et BF seront notés CAL-MPL, CAL-SN et CAL-BF, respectivement, dans la suite de ce document ;

◇ Les jeux VAL (n = 30) issus des BDDs spectrales MPL, SN et BF seront notés VAL-MPL, VAL-SN et VAL-BF, respectivement, dans la suite de ce document ;

◇ Les jeux STD (n = 31) issus des BDDs spectrales MPL, SN et BF seront notés STD-MPL, STD-SN et STD-BF, respectivement, dans la suite de ce document.

2.7 Modélisation

Tout le travail présenté dans cette section a été réalisé avec le logiciel R.

2.7.1 Régression des moindres carrés partiels (PLSR)

Le principal objectif de ce stage était de prédire la teneur en COS avec précision tout en minimisant les perturbations dues au changement de spectromètre, grâce au transfert d'étalonnage. Dans cette étude, les BDDs spectrales étaient caractérisées par un plus grand nombre de variables (absorbances) que d'observations (échantillons ; section 2.3). Il s'agit donc d'un exemple de statistiques en grande dimension, ce phénomène étant appelé malédiction de la dimensionnalité (Curse of dimentionnality ; Lavergne et Patilea, 2008). Dans cette étude, le modèle de régression des moindres carrés partiels (partial least squares regression ; PLSR) a été choisi car il est robuste face à la malédiction de la dimensionnalité et permet de supprimer l'impact lié à la multicolinéarité des variables (Wold,1975).

La PLSR est une méthode de régression linéaire qui vise à construire des variables latentes orthogonales entre elles, tout en maximisant la covariance entre les variables prédictives X (absorbances) et la (ou les) variable(s) réponse(s) y (teneur en COS).

Mathématiquement, la PLSR est construite ainsi :

$$\boxed{\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\mathbf{W}_{p\times p}(\mathbf{W}'\boldsymbol{\beta})_{p\times 1} + \mathbf{e}_{n\times 1} = \mathbf{T}_{n\times p}\boldsymbol{\gamma}_{p\times 1} + \mathbf{e}_{n\times 1}} \tag{3}$$

$$\boxed{\mathbf{T}_{n\times p} = \mathbf{X}_{n\times p}\mathbf{W}_{p\times p}} \tag{4}$$

$\mathbf{y}_{n\times 1}$: Vecteur des variables réponses

$\mathbf{X}_{n\times p}$: Matrice des variables prédictives

$\mathbf{W}_{p\times p}$: Matrice des poids

$\mathbf{W}'_{p\times p}$: Transposée de la matrice des poids

$\boldsymbol{\beta}_{p\times 1}$: Vecteur des coefficients de régression

$\mathbf{e}_{n\times 1}$: Vecteur des résidus

$\mathbf{T}_{n\times p}$: Matrice des nouvelles variables (variables latentes de la PLSR)

$\boldsymbol{\gamma}_{p\times 1}$: Vecteur des coefficients de régression pour les nouvelles variables

Étapes de la PLSR

1. A partir d'une base d'étalonnage (jeu CAL), pour laquelle la variable réponse (y) et les variables prédictives (X) sont connues, les poids $\mathbf{W}$ qui maximisent la covariance (ou la corrélation si les variables X et y sont centrées-réduites) sont calculés entre la variable réponse $\mathbf{y}$ et $\mathbf{T} = \mathbf{X}_{n\times p}\mathbf{W}_{p\times p}$.
2. La procédure Moindres Carrés Ordinaires (MCO) est alors utilisée pour produire $\boldsymbol{\gamma}_{p\times 1}$.
3. Le calcul des $\boldsymbol{\gamma}_{p\times 1}$ permet de retrouver l'équation **4** (modèle complet qui exprime le vecteur y de chaque échantillon en fonction de la matrice X initiale de cet échantillon).

Dans cette étude, deux étapes de modélisation ont été suivies systématiquement :

- **Phase d'étalonnage du modèle** : cette première étape permet de paramétrer et de résoudre l'équation **4** à partir du jeu CAL. Par un processus de validation croisée, les paramètres déterminés à l'issue de cette étape sont : le nombre de variables latentes de la PLSR (NLV) et le prétraitement spectral optimal (20 prétraitements testés). Dans cette étude, l'indicateur de performance utilisé pour choisir NLV et le prétraitement a été l'erreur quadratique moyenne de validation croisée (root mean square error of cross-validation ; RMSECV).
- **Phase de la validation externe du modèle** : Cette phase consiste à valider le modèle avec les prétraitement et **NLV** optimaux choisis lors de la phase d'étalonnage. Lors de la validation, le modèle est testé sur des échantillons qui n'ont pas été utilisés pour le construire. Dans cette étude, les indicateurs qui ont été utilisés pour juger de la qualité de la validation sont ci-dessous.

Soient les notations suivantes :

— y_i : les valeurs observées ;

— $\hat{y}_i$: les valeurs prédites par le modèle ;

— n : le nombre total d'observations ;

— $\bar{y}$: la moyenne des valeurs observées, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$;

— s_y : l'écart-type des valeurs observées ;

— $IQR(y)$: l'écart interquartile des valeurs observées.

Les métriques de validation sont définies comme suit :

- **L'erreur quadratique moyenne de prédiction (RMSEP)** :

$$\text{RMSEP} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

- **Le biais de la validation** :

$$\text{Biais} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{6}$$

- **Le coefficient de détermination pour la validation externe (R^2)** :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

- **Le ratio de l'écart-type du jeu de validation par la RMSEP (RPD)** :

$$\text{RPD} = \frac{s_y}{\text{RMSEP}} \tag{8}$$

- **Le ratio de l'écart interquartile du jeu de validation par la RMSEP (RPIQ)** :

$$\text{RPIQ} = \frac{IQR(y)}{\text{RMSEP}} \tag{9}$$

Dans cette étude, deux approches de PLSR (globale et localement pondérée) ont été testées en suivant ces deux étapes (étalonnage, validation) et en utilisant les mêmes indicateurs de performance.

2.7.2 PLSR global

La PLSR globale est la plus couramment employée en raison de sa simplicité et de son efficacité. Elle est construite en attribuant un poids identique à tous les échantillons du jeu CAL. L'utilisation égale de tous les échantillons de la PLSR globale aboutit à un modèle qui reflète de manière fidèle la diversité des données disponibles, améliorant ainsi sa capacité à être généralisée à de nouveaux échantillons, dans le cas où ces derniers sont bien représentés par cette diversité.
Néanmoins, cette méthode n'est pas capable de capter l'information locale de certains groupes d'échantillons.
C'est pourquoi la PLSR localement pondérée a également été testée et comparée à la PLSR globale pour prédire la teneur en COS.

2.7.3 PLSR localement pondéré

La PLSR localement pondérée (lwplsr) permet de modéliser des relations complexes entre les variables X et y en tenant compte des variations locales dans les données. La méthode a été développée pour améliorer la précision des prédictions en pondérant les échantillons du jeu CAL en fonction de leur similarité avec un échantillon cible (chaque échantillon VAL dans notre cas).
Dans les étalonnages multivariés, la lwplsr est une méthode de prédiction efficace lorsque l'hétérogénéité des données génère des relations non linéaires (courbures et regroupements) entre la variable réponse et les variables prédictives (Metz et al., 2021).
Pour comprendre la lwplsr, il est pertinent de connaître la définition de la localité et de la pondération.

Définitions clés de la PLSR localement pondérée

Localité :
- Considérer un point cible x_0 dans l'espace spectral des prédicteurs.
- Définir une métrique de distance (euclidienne, Mahalanobis, corrélation) pour mesurer la proximité des autres points x_i par rapport à x_0 dans cet espace.

Poids de localité :
-Calculer les poids w_i pour chaque observation i en fonction de sa distance à x_0.

La méthode lwplsr fonctionne comme suit :
soient X (matrice des absorbances) et Y (teneur en COS)

◇ **Réduction de dimension via une PLSR globale préliminaire** :

- Construire une PLSR globale à partir du jeu CAL avec un NLV donné (dans notre cas, 10 variables latentes ont été utilisées) .
- Les variables X peuvent donc être projetées dans ce nouvel espace pour obtenir X_{proj}. Cette étape permet de réduire la dimension pour simplifier le calcul de la distance, elle permet aussi d'éliminer le bruit en se concentrant sur les directions les plus importantes pour la modélisation.

- **Calcul de la distance de Mahalanobis** :
  - Dans l'espace construit précédemment, la distance entre chaque échantillon x_i et l'échantillon cible x_0 peut être calculée par :

$$d_i(x_0) = \sqrt{(X_{\text{proj},i} - X_{\text{proj},0})^T \Sigma^{-1} (X_{\text{proj},i} - X_{\text{proj},0})} \tag{10}$$

    où Σ est la matrice de covariance des variables prédictives.

- **Calcul des poids à partir d'un vecteur de distances à l'aide d'une fonction exponentielle décroissante** :
  - Les poids préliminaires sont calculés par :

$$w_i = \exp\left(-\frac{d_i}{h \times \text{mad}(d_i)}\right) \tag{11}$$

    où h est un facteur d'échelle (un scalaire $h > 0$), d_i est la distance de Mahalanobis pour l'échantillon x_i, et $\text{mad}(d_i)$ est la déviation absolue médiane des distances d_i.
  - Les poids correspondant aux distances supérieures à $\text{median}(d) + \text{cri} \times \text{mad}(d)$, où cri est un scalaire $\text{cri} > 0$, sont mis à zéro pour éliminer les valeurs aberrantes.
  - Enfin, les poids sont "normalisés" entre 0 et 1 par :

$$w_i = \frac{w_i}{\max(w)} \tag{12}$$

◇ **Sélection des k plus proches voisins** :

- Les échantillons du jeu CAL sont triés selon leur distance $d_i(x_0)$ et les k plus proches voisins de x0 sont alors sélectionnés. Dans cette étude, ce paramètre a été fixé à 84, i.e.,

tous les échantillons du jeu CAL ont systématiquement été utilisés et pondérés pour la construction du modèle lwplsr.

◇ **Pondération des données** :

- Les matrices X et Y sont alors pondérées selon les poids calculés, avant d'être utilisées pour construire l'équation Eq. XX :

$$X_w = W^{1/2}X \quad \text{et} \quad Y_w = W^{1/2}Y \tag{13}$$

où W est une matrice diagonale contenant les poids $w_i(x_0)$ sur la diagonale, X_w et Y_w les nouvelles matrices X et Y pondérées.

◇ **Régression PLS locale** :

- Une PLSR est enfin appliquée sur les matrices pondérées X_w et Y_w pour ajuster un modèle local.

2.7.4 Optimisation des modèles PLSR et lwplsr

La sélection des NLV et prétraitement optimaux est une étape cruciale de la phase d'étalonnage (présentée en section 2.7.1) de la PLSR globale ou de la lwplsr pour garantir une performance optimale. Néanmoins, il n'existe pas à ce jour de méthodologie claire et systématique pour le choix du NLV. Cette étude a mis en œuvre deux stratégies distinctes pour déterminer le NLV optimal, testé entre 1 et 15, chacune offrant des avantages spécifiques en termes de robustesse et de fiabilité.

Quelque soit la stratégie, l'optimisation du modèle s'est appuyée sur une validation croisée par site (**one-site-out** cross-validation ; 21 sites dans le jeu CAL). Pour chaque prétraitement spectral et NLV testé, deux étapes ont été suivies :
- étalonnage du modèle sur $\mathbf{k-1}$ sites ;
- validation de ce modèle sur le kième site, n'ayant pas servi à l'étalonnage. Cette étape a été répétée jusqu'à ce que les 21 sites du jeu CAL aient permis de valider une fois le modèle. A l'issue de cette validation croisée, la RMSECV a été calculée pour chaque prétraitement et NLV considéré. En traçant la RMSECV en fonction du NLV par prétraitement, des points d'inflexion locaux, correspondant à des minima locaux de la RMSECV, ont été identifiés. Ces points d'inflexion sont particulièrement importants car ils peuvent indiquer qu'un ajout de variables latentes supplémentaires n'apporte qu'une amélioration faible de la performance, voire qu'il peut introduire du surajustement. Les NLV associés à ces points d'inflexion ont donc été sélectionnés pour chaque prétraitement, repré-

sentant un compromis judicieux entre biais et variance. Cette méthode sera appelée "méthode de double inflexion" dans la suite de ce document.

La deuxième approche testée repose sur l'utilisation du Bootstrap, qui vise à stabiliser la sélection du NLV en introduisant une forme de répétabilité statistique ; elle permet ainsi d'identifier une tendance globale qui assure une convergence stable du modèle. Le Bootstrap repose sur l'échantillonnage avec remplacement, permettant de capturer la variabilité inhérente aux données. Pour cette étude, 100 itérations de Bootstrap ont été testées. À chaque itération, une validation croisée par site a été réalisée et la RMSECV a été calculée. La moyenne des RMSECVs issues du Bootstrap a ensuite été calculée (RMSECV_Boot) et tracée pour chaque NLV. Cette méthode offre une vision plus complète et robuste de la performance attendue du modèle en réduisant l'impact des fluctuations dues à la variabilité des données.

Pour une question de temps, c'est la méthode de la double inflexion qui a été finalement choisie pour valider les modèles PLSR de ce manuscrit. Le prétraitement qui a été finalement considéré comme optimal était celui qui amenait à la RMSECV la plus faible pour ces deux inflexions.

2.8 Méthodes de correction

2.8.1 Motivation

Les grandeurs d'influence telles que (i) le changement d'appareil (spectromètre), (ii) l'origine des échantillons (site d'étude pour les sols), (iii) les variations entre les méthodes de préparation des échantillons (par exemple, échantillons de sol tamisés ou broyés) sont susceptibles d'engendrer des perturbations spectrales (par exemple, génération de lignes de base, bruit aléatoire, décalage des absorbances en fonction des longueurs d'ondes). Ces perturbations constituent un frein à la reproductibilité du modèle. Cette étude s'est particulièrement intéressée à l'effet du changement de spectromètre sur les modèles. Afin d'atténuer l'impact du changement de spectromètre, des corrections mathématiques (ou transferts d'étalonnage) peuvent être appliquées entre une BDD spectrale acquise avec un spectromètre source (BDD source qui sert à étalonner le modèle de prédiction de la variable Y) et une BDD spectrale acquise avec un spectromètre cible (BDD cible qui sert à valider le modèle déjà étalonné). Pour cette étude, trois spectromètres ont été utilisés (section **2.3)**, et la stratégie adoptée a été de les considérer deux à deux. Dans cette optique, trois différents scénarios ont été testés :

* Le spectromètre **LabSpec-SN** (générant les jeux CAL-SN, VAL-SN et STD-SN) a été consi-

déré comme l'appareil source et le spectromètre **LabSpec-BF** (générant les jeux CAL-BF, VAL-BF et STD-BF) a été considéré comme l'appareil cible ;

* Le spectromètre **LabSpec-MPL** (générant les jeux CAL-MPL, VAL-MPL et STD-MPL) a été considéré comme l'appareil source et le spectromètre **LabSpec-SN** a été considéré comme l'appareil cible ;
* Le spectromètre **LabSpec-MPL** a été considéré comme l'appareil source et le spectromètre **LabSpec-BF** a été considéré comme l'appareil cible.

Dans chaque scénario, quatre situations ont été étudiées pour prédire la variable Y à partir des données spectrales. Un premier modèle de prédiction (PLSR globale et lwplsr) a été construit sur le jeu CAL de la BDD source (CAL-Source) et validé sur le jeu VAL de la BDD source (VAL-Source ; modèle source sur source). Un second modèle de prédiction (PLSR globale et lwplsr) a été construit sur le jeu CAL de la BDD cible (CAL-Cible) et validé sur le jeu VAL de la BDD cible (VAL-Cible ; modèle cible sur cible). Un troisième modèle de prédiction (PLSR globale et lwplsr) a été construit sur le jeu CAL-Source et validé sur le jeu VAL-Cible sans correction (modèle source sur cible sans transfert). Un troisième modèle de prédiction (PLSR globale et lwplsr) a été construit sur le jeu CAL-Source et validé sur le jeu VAL-Cible avec correction (modèle source sur cible avec transfert). Pour cette dernière situation, le même prétraitement spectral que celui sélectionné lors du modèle source sur source a été utilisé ; de plus, plusieurs méthodes de correction ont été testées et sont décrites ci-dessous.

2.8.2 Update & Double Update

La méthode Update ("mise à jour" en français) consiste à enrichir le jeu CAL-Source avec le jeu STD de la BDD cible (STD-Cible). Un modèle PLSR ou lwplsr est alors reconstruit à partir de cette nouvelle base d'étalonnage (choix d'un nouveau NLV optimal) qui est ensuite appliqué sur VAL-Cible (NI et al., 2010) voir Figure 3 .

FIGURE 3 – La méthode Update et Double Update

La méthode Double Update est similaire à la méthode Update mais au lieu d'ajouter uniquement le jeu STD-Cible au jeu CAL-Source, on y ajoute également le jeu STD-Source avant la construction du modèle d'étalonnage (PLSR ou lwplsr).

2.8.3 External Parameter Othogonalisation (EPO)

Cette méthode de robustification consiste à identifier l'espace des perturbations lié aux paramètres externes (dans le cas de cette étude, un changement d'appareil). Une fois cet espace identifié, un processus d'orthogonalisation est effectué pour supprimer cet espace des données spectrales (Roger et al., 2003).

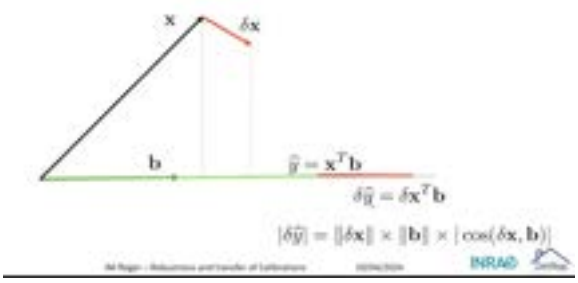


FIGURE 4 – représentation graphique de la méthode

La figure 4 constitue le point de départ qui permet d'expliquer de manière géométrique cette méthode. Ce graphique est en deux dimensions (hypothèse simplificatrice). Trois vecteurs sont définis sur ce graphique : $\mathbf{X}$ représente le vecteur des spectres, $\mathbf{b}$ le vecteur des coefficients de régression et $\delta_{\mathbf{X}}$ représente le vecteur des perturbations dues aux paramètres externes (i.e., changement de spectromètre). L'objectif principal est de minimiser $\delta_{\mathbf{X}}$ (déviations que l'on cherche à corriger),

ce qui revient à minimiser sa projection orthogonale, $\delta_{\hat{y}}$, sur l'axe de $\mathbf{b}$. La valeur absolue de $\delta_{\hat{y}}$ s'exprime ainsi :

$$|\delta_{\hat{y}}| = \|\delta_X\| \cdot \|\mathbf{b}\| \cdot |\cos(\delta_X, \mathbf{b})|$$

La valeur absolue $\delta_{\hat{y}}$ peut donc être minimisée en réduisant la valeur absolue de $\delta_{\mathbf{X}}$, mais aussi celle de $\mathbf{b}$ ou de $\cos(\delta_{\mathbf{X}}, \mathbf{b})$. L'objectif de l'orthogonalisation est notamment de minimiser$|\cos(\delta_{\mathbf{X}}, \mathbf{b})|$.

Les étapes de la méthode EPO

- Une collection de spectres pour laquelle l'effet de la perturbation considérée est connue est utilisée pour construire une matrice $\mathbf{D}$.
- Une matrice $\mathbf{P}$ de l'espace vectoriel engendré par les variations dues à la perturbation est estimée par construction d'une SVD (Singular Value Decomposition) sur $\mathbf{D}$.
  - La SVD décompose $\mathbf{D}$ en $\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, où $\mathbf{U}$ et $\mathbf{V}$ sont des matrices orthogonales et $\mathbf{\Sigma}$ est une matrice diagonale contenant les valeurs singulières. Le nombre de CPs utilisées pour construire cette SVD est un paramètre à régler lors de l'étalonnage de l'EPO.
  - Les colonnes de $\mathbf{V}$ correspondant aux plus grandes valeurs singulières (racine carrée des valeurs propres) forment la matrice $\mathbf{P}$.
- $\mathbf{X}$ est alors projeté orthogonalement à $\mathbf{P}$ pour donner $\tilde{\mathbf{X}}$:

$$\tilde{\mathbf{X}} = \mathbf{X}(\mathbf{I} - \mathbf{P}(\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T)$$

Dans cette étude, l'algorithme de clustering Kmeans (package prospectr ; Stevens et Ramirez-Lopez, 2022) a été appliqué sur le jeu STD pour le diviser en deux sous-ensembles (**STDA** contenant 16 échantillons et **STDB** contenant 15 échantillons). Le sous-ensemble **STDA** a été utilisé pour construire l'EPO pour un nombre de CPs (NCP) allant de 0 à 10 ; le sous-ensemble **STDB** a été utilisé pour choisir NCPs optimal de l'EPO.
La matrice $\mathbf{D}$ a donc été construite à partir des spectres sources et cibles de **STDA** :

$$\mathbf{D} = \mathbf{X}_{\mathbf{STDA_{cible}}} - \mathbf{X}_{\mathbf{STDA_{source}}} \tag{14}$$

avec :

$\mathbf{X}_{\mathbf{STDA_{cible}}}$: matrice des spectres cibles du jeu STDA.

$\mathbf{X}_{\mathbf{STDA_{source}}}$: Matrice des spectres sources du jeu STDA.

Puis la matrice P a été construite à partir de D, pour chaque NCP. Le jeu CAL de la BDD source a ensuite été corrigé par P pour chaque NCP et un modèle d'étalonnage PLSR (globale et lwplsr) a été construit à partir du jeu CAL corrigé (choix du NLV optimal pour chaque NCP). Chaque modèle a ensuite été appliqué aux BDDs sources et cibles du jeu **STDB** (STDB-Source et STDB-Cible, respectivement), puis validé dans chaque cas avec la variable réponse connue pour **STDB**. Le NCP optimal était celui pour lequel le biais entre les prédictions obtenues pour STDB-Source et STDB-Cible était minimal. La correction EPO a ensuite été appliquée avec le NCP optimal au jeu CAL-Source, puis un modèle PLSR (globale et lwplsr) a de nouveau été construit (choix du NLV optimal) puis validé sur VAL-Cible.

2.8.4 Piecewise Direct Standardisation (PDS)

La standardisation optique consiste à transformer les spectres obtenus avec l'appareil source pour qu'ils ressemblent aux spectres obtenus avec l'appareil cible via une fonction mathématique.

$$\tilde{X_s} = f(X_s)$$

Avec :
X_s : Matrice des spectres sources
$\tilde{X_s}$: Matrice des spectres sources transformés pour qu'ils ressemblent à ce qu'ils auraient été s'ils avaient été acquis avec le spectromètre cible.
La méthode la plus utilisée est la Piecewise Direct Standardisation (PDS) $\longrightarrow$ suppose que la fonction f est linéaire (Wang et al., 1991).

Les étapes clés du processus de la PDS au sein du jeu STD (permettant d'étalonner la méthode de correction) sont présentées ci-dessous :

- Les matrices des spectres sources (Xs) et cibles (Xc) du jeu STD sont centrées, i.e., l'absorbance de chaque échantillon à chaque longueur d'onde est soustraite par la moyenne des absorbances de tous les échantillons à cette même longueur d'onde :

$$\mathbf{X}_s = \mathbf{S} - \mu_s \tag{15}$$

$$\mathbf{X}_c = \mathbf{C} - \mu_c \tag{16}$$

où :

— **S** est la matrice des spectres sources du jeu STD.

— **C** est la matrice des spectres cibles du même jeu STD.

— μ_S est la matrice des absorbances moyennes de S calculées par longueur d'onde.

— μ_C est le matrice des absorbances moyennes de C calculées par longueur d'onde.

- Une fenêtre de largeur w (généralement un nombre impair pour assurer la symétrie et la centralité) est ensuite définie autour de chaque longueur d'onde (dans notre étude, plusieurs w ont été testées : 3, 5, 7, 9, 11).

$$w = 2l + 1 \tag{17}$$

où l est la demi-largeur de la fenêtre.

- Une décomposition en valeurs singulières (singular value decomposition ; SVD) est ensuite construite à partir des spectres sources centrés du jeu STD sur la fenêtre w.

$$\mathbf{X}_s^{(i)} = \mathbf{UDU}^T \tag{18}$$

où :

— $\mathbf{X}_s^{(i)}$ est la matrice des spectres sources centrés pour chaque longueur d'onde i.

— **U**, **D**, $\mathbf{U}^T$ sont les matrices issues de la décomposition SVD calculées sur la fenêtre w : U est la matrice des CPs (ou vecteurs propres), D est la matrice diagonale des valeurs singulières et U^T la transposée de U.

- Puis, une régression linéaire est construite de sorte à exprimer l'absorbance centrée Xc des spectres cibles du jeu STD pour chaque longueur d'onde i à partir de la matrice U calculée précédemment .Cette étape permet de calculer les coefficients de régression $\mathbf{b}_i$ et ϵ_i :

$$\mathbf{X}_c^{(i)} = \mathbf{U}_{(i)}\mathbf{b}_i + \epsilon_i \tag{19}$$

où :

— $\mathbf{X}_c^{(i)}$ est l'absorbance centrée à la longueur d'onde i des spectres du jeu cible STD ;

— $\mathbf{U}_{(i)}$ est la matrice des composantes principales calculée à i à partir des spectres centrés du jeu source STD ;

— $\mathbf{b}_i$ est le coefficient de régression calculé pour la longueur d'onde i ;

— ϵ_i est le terme d'erreur obtenu pour cette même longueur d'onde i.

- Enfin, l'ordonnée à l'origine (offset ; qui permet d'ajuster la différence moyenne entre les spectres sources et cibles) est calculée. En d'autres termes, l'offset permet de corriger des décalages systématiques entre les instruments.

$$\mathbf{o} = \mu_C - \mu_S \mathbf{B} \tag{20}$$

où :

— **B** est la matrice des coefficients de régression (bi) calculée pour toutes les longueur d'onde i lors de l'étape précédente.

- Une fois ces étapes terminées, la fonction PDS est construite pour la fenêtre w et peut être appliquée sur des spectres sources :

$$\mathbf{S}_{\text{new}} = \mathbf{S}.\mathbf{B} + \mathbf{o} \tag{21}$$

où :
$\mathbf{S}_{\text{new}}$ est un nouveau spectre source corrigé par application de la fonction PDS précédemment construite ; S est ce même spectre source (acquis avec le spectromètre source) avant correction PDS. Snew est alors considéré comme comparable aux mesures obtenues avec l'appareil cible. Pour cette étude, les choix de **NLV** et de la fenêtre **w** ont été effectués via la procédure suivante : la courbe de **RMSECV** en fonction de **NLV** a été tracée pour chaque w et le **NLV** et **w** qui minimisaient la **RMSECV** ont été choisis.

2.8.5 Correction Biais-Pente (CBP)

La Correction Biais-Pente (CBP) est une autre méthode de transfert d'étalonnage, dite a posteriori, car elle utilise un algorithme de correction basé sur les prédictions (Wang et al., 1986).
Elle vise à corriger les différences systématiques (biais) et les divergences de pente entre deux BDDs (dans notre cas, entre la BDD source et la BDD cible). Les prédictions sont obtenues par un modèle construit à partir des spectres sources et appliqué sur les spectres cibles.

Le principe de cette méthode est le suivant :

- A partir du jeu STD, une relation linéaire entre la variable réponse mesurée conventionnellement, y, et cette même variable prédite par VPIRS, $\hat{y}$, est d'abord construite :

$$y = a \cdot \hat{y} + b$$

- Les valeurs de y et $\hat{y}$ étant connues pour ce jeu STD, la pente a et le biais b peuvent donc être estimés. Pour cela, une régression linéaire univariée est appliquée. L'objectif est de minimiser la somme des carrés des résidus (i.e., la différence entre les valeurs mesurées et les valeurs prédites ; SCR) donnée par :

$$SCR = \sum_{i=1}^{n} (y_i - (a \cdot \hat{y}_i + b))^2$$

La méthode des moindres carrés ordinaires est ensuite appliquée pour calculer a et b :

$$a = \frac{\sum(y_i - \overline{y})(\hat{y}_i - \overline{\hat{y}})}{\sum(\hat{y}_i - \overline{\hat{y}})^2}$$

Où :

— y_i est la valeur mesurée de la variable réponse pour un échantillon (ou observation) i ;
— $\hat{y}_i$ est la valeur prédite de la variable réponse pour ce même échantillon i ;
— $\overline{y}$ et $\overline{\hat{y}}$ sont les valeurs mesurées et prédites moyennes pour le jeu d'échantillons STD, respectivement.

Une fois a calculé, le biais b peut être défini par :

$$b = \overline{y} - a \cdot \overline{\hat{y}}$$

- La valeur prédite $\hat{y}_{new}$ d'un nouvel échantillon peut alors être corrigée par la CBP comme suit :

$$\hat{y}_{cor} = a \cdot \hat{y}_{new} + b$$

avec y^cor, la valeur corrigée par CBP de y^new pour cet échantillon.
Aussi faut-il noter que pour cette méthode de transfert, la phase d'étalonnage du modèle PLSR (globale ou lwplsr) à partir de la BDD source ne doit pas être réitérée étant donné que cette méthode corrige les prédictions, après application du modèle. Le **NLV** n'a donc pas été modifié par rapport à celui du modèle source sur source.

3 Résultats et discussion

3.1 Analyse exploratoire des données

3.1.1 Analyse descriptive de la teneur en COS mesurée conventionnellement

L'analyse statistique de la teneur en carbone organique du sol (COS) est essentielle pour comprendre la variabilité et les caractéristiques des sols dans différents contextes agronomiques et environnementaux. Cette étude se concentre sur trois jeux d'échantillons distincts : CAL, VAL, et STD. Chaque jeu d'échantillons a été analysé pour fournir des statistiques descriptives clés de la teneur en COS, telles que la valeur minimale (Min), la valeur moyenne (Mean), la variance (Var), l'écart-type (SD), la valeur maximale (Max) et l'étendue (la différence entre les valeurs maximale et minimale). Ces mesures permettent d'évaluer la distribution et la dispersion des teneurs en COS et d'identifier ainsi les tendances au sein de chaque jeu d'échantillons.

Jeu	Min	Mean	Var	SD	Max	Étendue
CAL	2.19	11.59	81.27	9.01	49.80	47.61
VAL	4.39	7.89	12.39	3.52	22.50	18.10
STD	1.94	12.31	83.15	9.12	37.27	35.33

TABLE 1 – Statistiques descriptives de la teneur en COS par jeu d'échantillons

D'après le Tableau 1 :

- Les jeux CAL et STD montrent une distribution similaire de leurs teneurs en COS, avec, dans chaque cas, une étendue et un SD importants, ce qui indique une hétérogénéité marquée dans ces jeux d'échantillons.
- VAL BF présente une variation plus modérée des valeurs de COS, avec une étendue et un SD plus faibles, suggérant une distribution plus homogène des valeurs de COS.

La **Figure 5** illustre la distribution de la teneur en COS par jeu de d'échantillons. Elle montre notamment que la distribution de la teneur en COS du jeu CAL couvre celles des jeux VAL et STD. De plus, elle met en évidence une asymétrie positive pour chaque jeu, avec plusieurs échantillons atypiques éloignés des autres échantillons, pour les valeurs de COS les plus hautes.

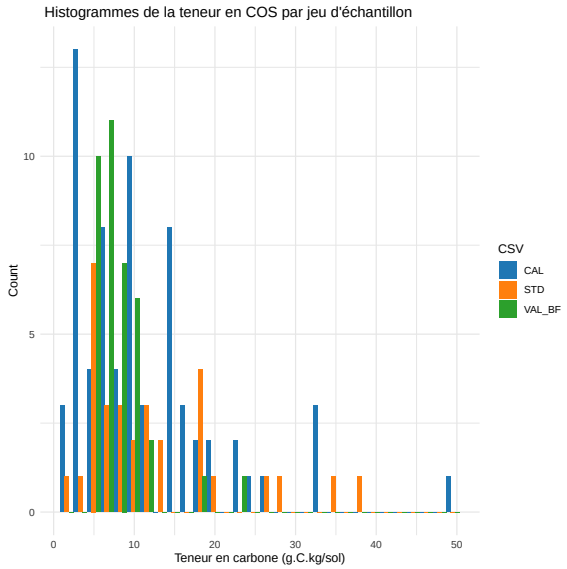


FIGURE 5 – La teneur en COS par jeu d'échantillons

La **Figure 6** illustre la distribution de la teneur en COS pour les 22 sites considérés dans cette étude. La représentation sous forme de boîtes à moustaches (Boxplots) permet de visualiser la

médiane, les quartiles et les valeurs extrêmes de teneurs en COS pour chaque site, offrant une vue d'ensemble comparative des différentes distributions par site. Une forte hétérogénéité inter-site est remarquable ; elle peut être attribuée à des facteurs environnementaux tels que le climat, le type de sol, la végétation et l'historique d'utilisation des terres. En revanche, la teneur en COS du jeu VAL (noté Torokoro dans la Figure 6) est incluse dans la distribution de celle des autres sites et est donc supposément bien représentée par la plupart des autres sites.

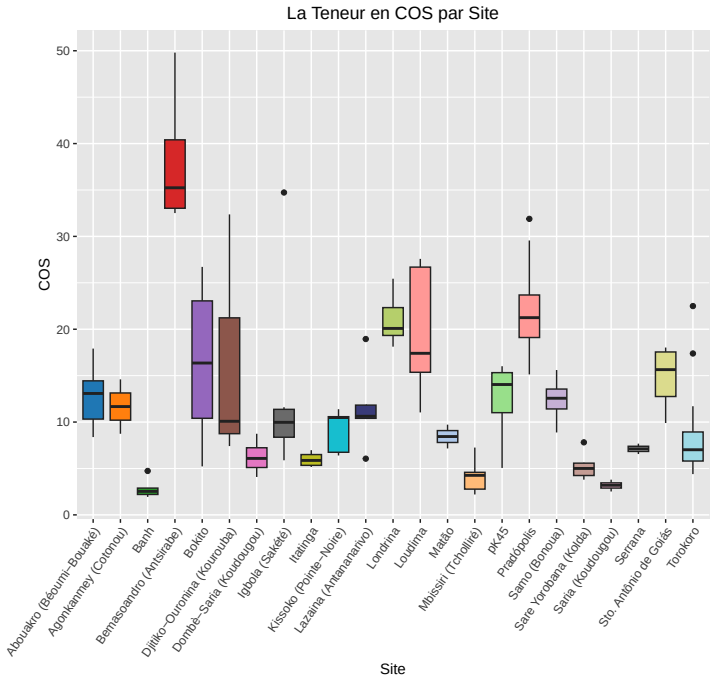


FIGURE 6 – Comparaison de la teneur en COS par site

3.1.2 Analyse descriptive des spectres

L'analyse des spectres est une étape cruciale dans l'étude de la teneur en COS à travers des techniques de spectroscopie infrarouge. Les spectres, qui représentent l'intensité de la réflexion (ou absorption) lumineuse à différentes longueurs d'onde, fournissent des informations détaillées sur la composition chimique et physique des échantillons de sol.

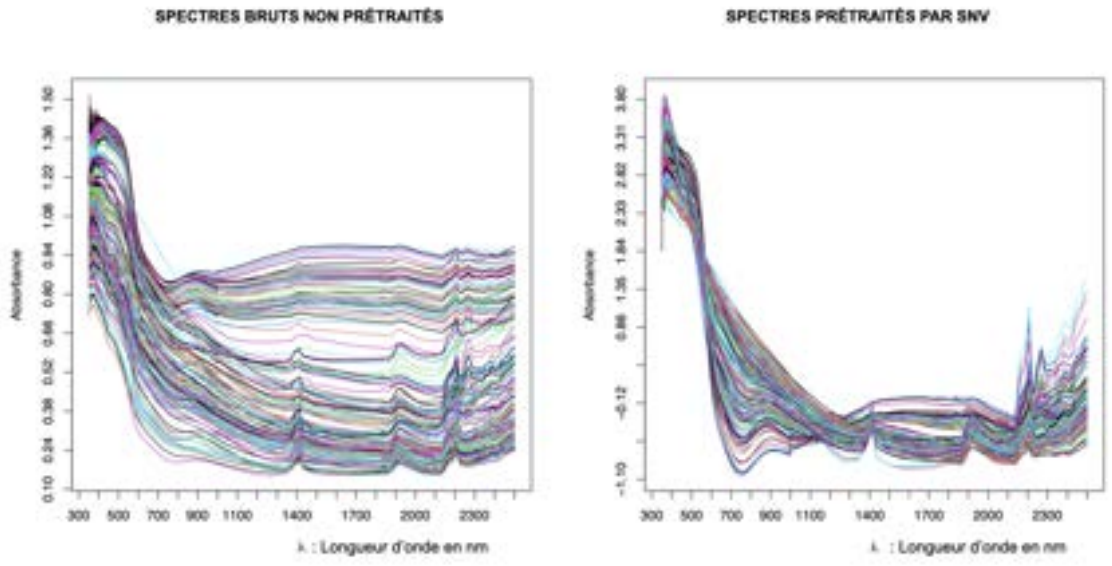


FIGURE 7 – Présentation des spectres **bruts** VS **prétraités**

La Figure 7 montre des spectres d'absorbance bruts (non prétraités ; graphe de gauche) et des spectres d'absorbance prétraités par SNV (graphe de droite). Le prétraitement SNV a permis de diminuer les fortes divergences qui étaient visibles entre les spectres bruts, ces divergences étant principalement expliquées par des effets optiques liés à des grandeurs d'influence autres que la teneur en COS (granulométrie et agencement des particules dans l'échantillon, température, humidité, etc.). La Figure 7 montre ainsi le fort effet que peut avoir un prétraitement sur les spectres d'un jeu de données.

La **Figure 8** représente le plan défini par les deux premières CP d'une ACP construite à partir des données spectrales prétraitées par SNV du jeu CAL-MPL et sur lequel les spectres SNV du jeu VAL-MPL on été projetés. La variance expliquée par ces deux premières CPs était de 90,3%. Bien que les échantillons du jeu VAL-MPL n'aient pas influencé la création des axes de l'ACP, leur position dans cet espace a permis d'observer leur comportement par rapport aux échantillons CAL-MPL. Sur ce plan, le jeu VAL-MPL est bien représenté spectralement par le jeu CAL-MPL. Le même résultat a été obtenu pour les jeux BF et SN, justifiant ainsi l'applicabilité des modèles prédictifs

construits à partir des jeux CAL (CAL-MPL, CAL-SN, CAL-BF) sur les jeux VAL (VAL-MPL, VAL-SN, VAL-BF).

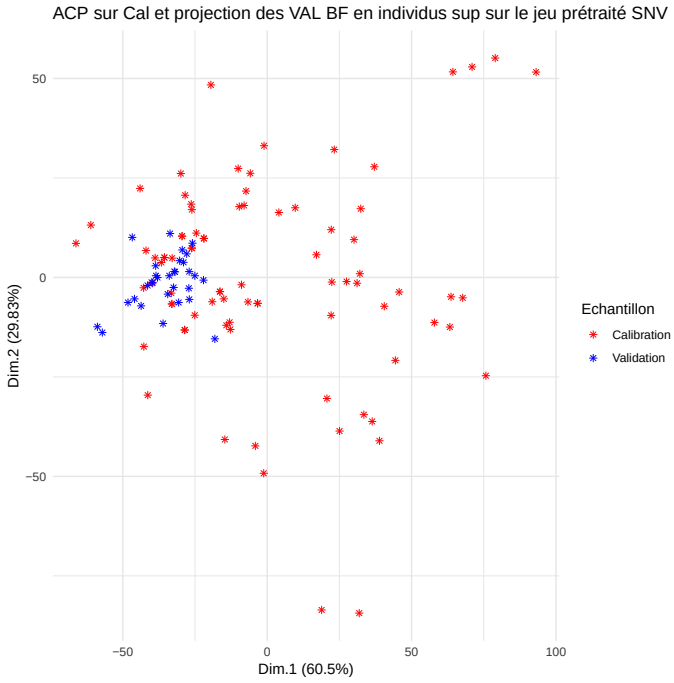


FIGURE 8 – Représentation du jeu de calibration et de validation

3.2 Présentation des modèles sans transfert d'étalonnage

Les modèles sans transfert d'étalonnage concernent les modèles (i) **sans changement d'appareil** et (ii) **avec changement d'appareil**.

Les modèles sans changement d'appareil correspondent aux modèles à la fois étalonnés et validés sur le jeu de données obtenu avec le **même** appareil (i.e., en utilisant respectivement CAL-MPL et VAL-MPL ; CAL-SN et VAL-SN ; CAL-BF et VAL-BF). Pour ces modèles, les deux approches de PLSR (globale et lwplsr) ont été testées afin d'en comparer les performances.

Les modèles avec changement d'appareil correspondent aux modèles étalonnés avec l'appareil source et validés sur l'appareil cible (MPL sur SN, MPL sur BF et SN sur BF).

3.2.1 Modèles sans changement d'appareil

Une validation croisée par site a d'abord été construite à partir de chaque BDD spectrale (CAL-MPL, CAL-SN et CAL-BF), pour la **lwplsr** et la PLSR **globale**. Cette étape de la phase d'étalonnage a permis de choisir pour chaque modèle, le **NLV** optimal et le prétraitement optimal en considérant RMSECV. La Figure 9 montre la RMSECV en fonction de NLV pour le prétraitement **SNV** ; ce dernier ayant amené à la RMSECV la plus faible parmi les 20 prétraitements spectraux testés (données non présentées).

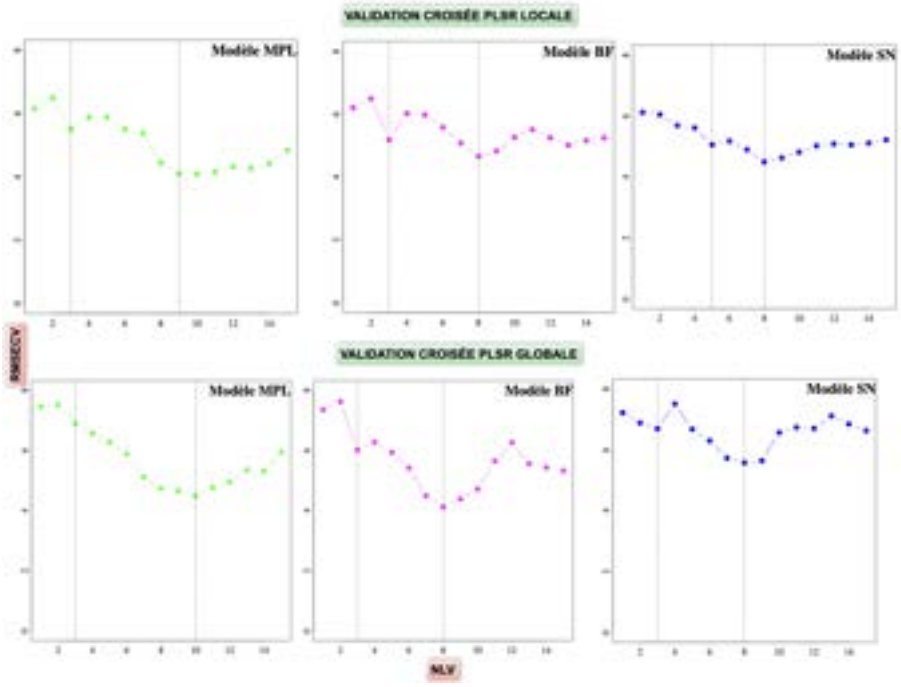


FIGURE 9 – RMSECV vs. NLV issus de la lwplsr (graphes du haut) et de la PLSR globale (graphes du bas)

En chimiométrie, il n'existe à ce jour aucune méthode clairement définie pour choisir NLV. Dans ce stage, une importante réflexion a été portée sur ce choix et la stratégie qui a été finalement adoptée a été de sélectionner (i) NLV amenant au RMSECV le plus faible mais aussi (ii) NLV > 0 pour lequel un changement de pente dans les valeurs de RMSECV étaient visible (ce changement de pente étant souvent un minimum local ; Figure 9). Ces NLVs ont ensuite été utilisés pour la construction du modèle de prédiction PLSR (lwplsr et globale). Dans la suite de ce manuscrit, les résultats présentés sont ceux obtenus avec le NLV amenant à un changement de pente (ou minimum

local) dans les valeurs de RMSECV. Ainsi, pour les approches lwplsr et PLSR globale sans transfert, le NLV choisi était de 3 et 3 pour MPL, de 5 et 3 pour SN et de 3 et 3 pour BF, respectivement. Le prétraitement optimal était systématiquement SNV.

La Figure 10 présente les résultats de validation externe obtenus par **LWPLSR** (graphes du haut) et **PLSR globale** (graphes du bas) sans changement d'appareil, en utilisant les paramètres (NLV et prétraitement) définis précédemment.

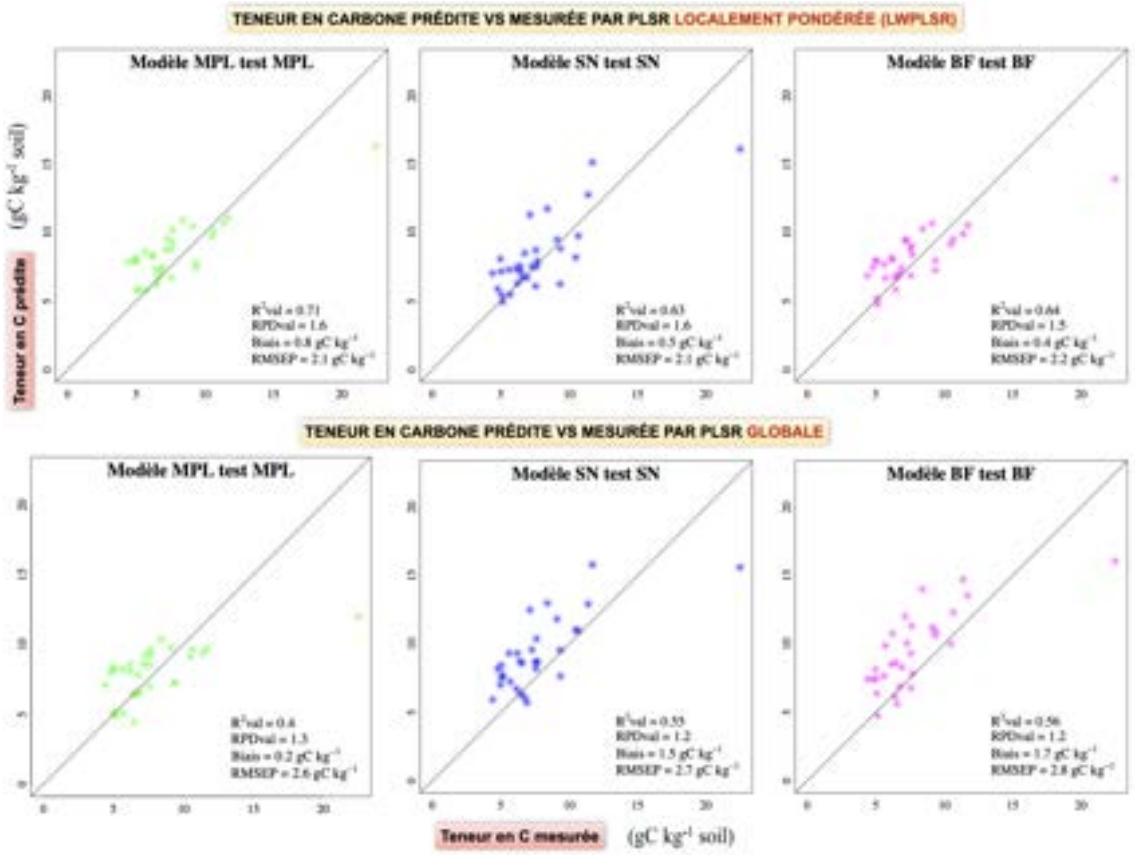


FIGURE 10 – Graphes des teneurs en COS (gC.kg-1) mesurées vs. prédites par la lwplsr (graphes du haut) et la PLSR globale (graphes du bas) en validation externe, sans changement d'appareil

Considérant la RMSEP, les résultats obtenus en validation externe pour chaque BDD spectrale étaient presque identiques. Elles étaient comprises entre **2.1** (pour MPL et SN) et **2.2** gC kg^{-1} (pour BF) avec l'approche locale et entre **2.6** (pour MPL) et **2.8** gC kg^{-1} (pour BF) avec l'ap-

proche globale. Dans cette étude, l'approche locale était donc plus robuste que l'approche globale. De plus, pour cette dernière, un biais important était notable dans les résultats de validation pour les BDD SN et BF (entre 1.5 et 1.7 gC kg-1) qui n'était pas notable pour la lwplsr appliquée à ces mêmes BDDs (biais entre 0.4 et 0.5 gC kg-1). Ceci peut être expliqué par un effet "site" (le jeu VAL était géographiquement indépendant du jeu CAL dans cette étude), ce qui a été mieux corrigé par le spiking lorsque le modèle était la lwplsr que lorsque le modèle était la PLSR globale. Les bénéfices plus importants du spiking en approche locale qu'en approche globale, lorsque le nombre d'échantillons de spiking était faible, a déjà été rapporté dans la littérature (Barthès et al., 2020 ; Cambou et al., 2024 ; Gogé et al., 2014).

La lwplsr ayant montré les meilleurs résultats de validation sans changement d'appareil, seule cette approche est présentée dans la suite de ce manuscrit (les résultats obtenus avec l'approche globale sont quant à eux présentés en annexes).

3.2.2 Modèles avec changement d'appareil

La **Figure 11** présente les résultats de validation externe des modèles avec changement de spectromètre mais sans transfert d'étalonnage. Les trois cas de figure étudiés étaient (i) un modèle construit à partir de CAL-SN et validé sur VAL-BF (Modèle SN test BF), (ii) un modèle construit à partir de CAL-MPL et validé sur VAL-SN (Modèle MPL test SN), (iii) un modèle construit à partir de CAL-MPL et validé sur VAL-BF (Modèle MPL test BF).

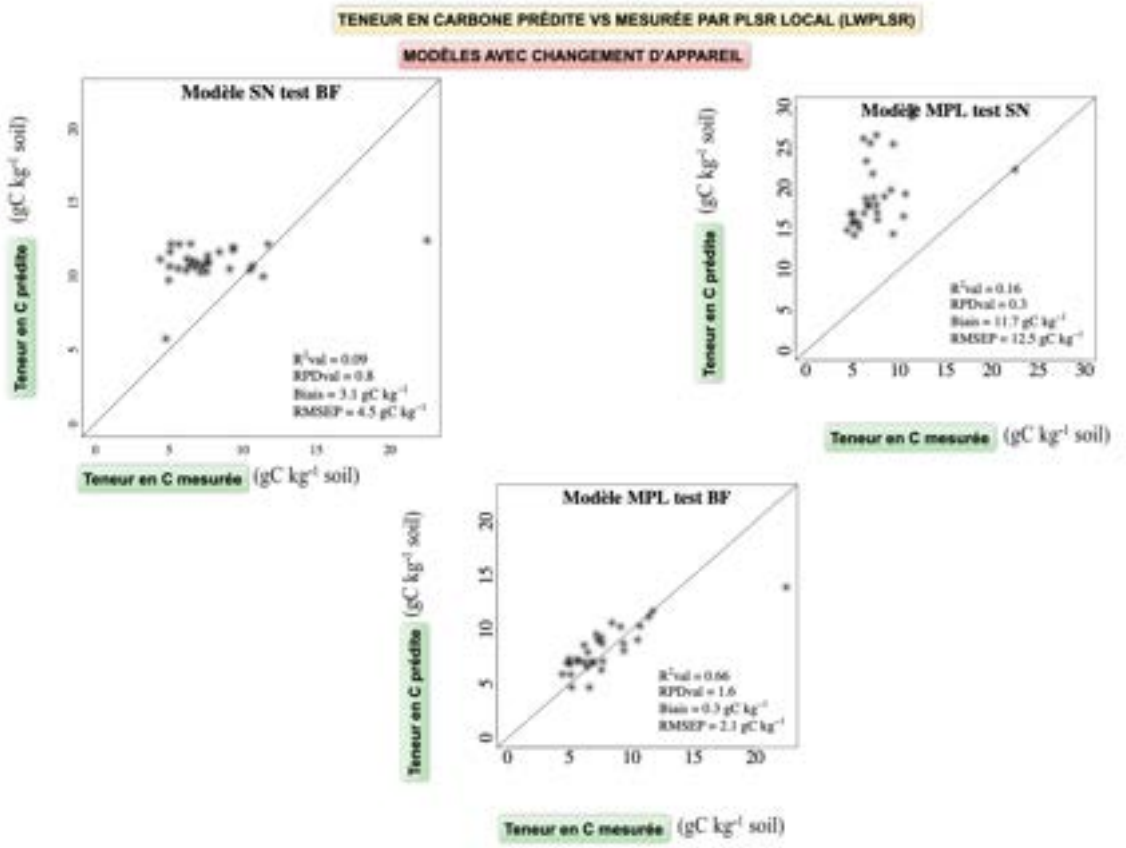


FIGURE 11 – Graphes de la teneur en COS (gC.kg-1) mesurée vs. prédite par lwplsr en validation externe, avec changement d'appareil

Que ce soit pour le "Modèle SN test BF" ou pour le "Modèle MPL test SN", le changement d'appareil a eu un impact important sur les résultats de validation. Pour le "Modèle SN test BF", la RMSEP et le biais ont atteint 4.5 gC.kg-1 et 3.1 gC.kg-1, respectivement (alors que RMSEP était de 2.1 gC.kg-1 et le biais de 0.5 gC.kg-1, lorsque le jeu VAL était VAL-SN). Pour le "Modèle MPL test SN", la RMSEP et le biais ont atteint 12.5 gC.kg-1 et 11.7 gC.kg-1, respectivement (alors que RMSEP était de 2.1 gC.kg-1 et le biais de 0.8 gC.kg-1, lorsque le jeu VAL était VAL-SN). En revanche, pour le "Modèle MPL test BF", le changement d'appareil a eu peu d'impact sur les résultats de validation (la RMSEP a stagné à 2.1 gC.kg-1 et le biais a même légèrement diminué par rapport à celui d'une validation sur VAL-MPL passant de 0.8 à 0.3 gC.kg-1). Ces résultats montrent que pour le "Modèle SN test BF" et le "Modèle MPL test SN", un transfert d'étalonnage est nécessaire. De plus, la mauvaise prédiction des teneurs en COS du jeu VAL par ces deux modèles

n'est pas due à un effet "site" entre le jeu CAL et le jeu VAL mais bien au changement de conditions d'acquisitions spectrales (différents appareils, opérateurs, et conditions au laboratoire), étant donné que sans changement d'appareil, les résultats de validation avec la lwplsr étaient corrects. Enfin, il semblerait que la BDD SN soit particulièrement différente des deux autres BDDs car seuls les modèles impliquant SN ont amené à des prédictions très imprécises avec changement d'appareil. Entre SN et les deux autres BDDs, une perturbation systématique doit donc être corrigée ; ceci étant probablement dû à des conditions d'interaction entre l'échantillon et l'appareil très spécifiques lors de la construction de cette BDD par rapport aux deux autres.

3.3 Présentation des modèles avec transfert d'étalonnage

La suite du manuscrit ne présentera que les résultats de transfert d'étalonnage entre l'appareil **SN (source)** et l'appareil **BF (cible)** obtenus avec un modèle lwplsr. Les résultats obtenus avec transfert d'étalonnage seront systématiquement comparés aux résultats obtenus sans transfert (**Figures 10 et 11**).

3.3.1 Update % Double Update

La **Figure 12** présente les résultats de (i) validation croisée (graphes en haut) et (ii) validation externe (graphes en bas) obtenus avec la méthode de correction Update (graphes à gauche) et Double Update (graphes à droite), respectivement. En validation croisée, le RMSECV minimal était obtenu avec un NLV de 8 et 9 pour les méthodes Update et Double Update, respectivement ; et un minimum local a été observé avec NLV égal à 3 et 2, respectivement. Le NLV correspondant à ce minimum local a été retenu dans chaque cas pour la modélisation de la teneur en COS.

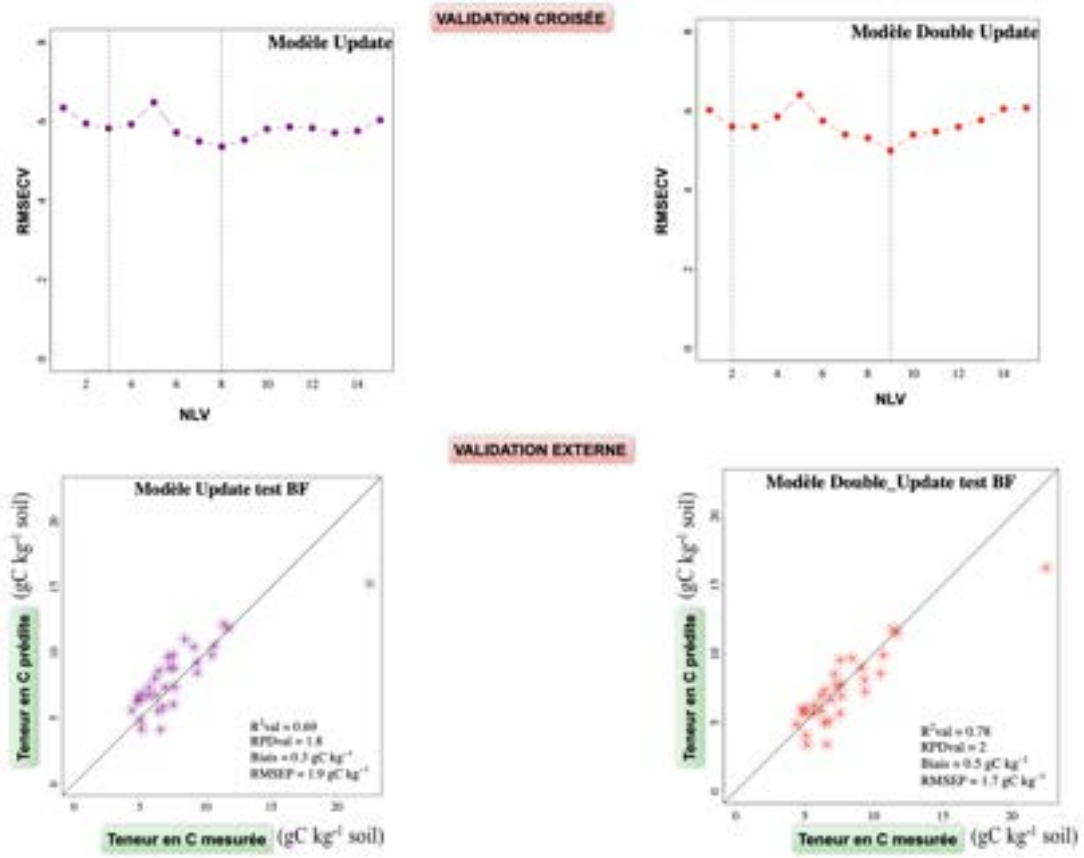


FIGURE 12 – Résultats Update & Double Update

Les résultats de validation externe montrent que ces deux méthodes corrigent les perturbations causées par le changement d'appareil car elles diminuent la RMSEP d'environ 58% et de 62% pour les méthodes Update et Double Update, respectivement, par rapport au Modèle SN test BF, sans transfert. On observe aussi une réduction du biais qui est passé de 3.1 gC.kg-1 à 0.3-0.5 gC.kg-1, témoignant ainsi d'une bonne qualité d'ajustement. La RMSEP est même plus faible que celle obtenue avec le modèle SN appliqué sur VAL-BF (RMSEP = 1.7-1.9 gC.kg-1 vs. 2.1 gC/kg-1). Cette amélioration par rapport à un modèle sans changement d'appareil peut être expliquée par le fait qu'ajouter une proportion non négligeable d'échantillons supplémentaires au jeu CAL-Source (environ un tiers d'échantillons ajoutés pour Update et environ 50% d'échantillons ajoutés pour Double Update), initialement de taille assez faible, a permis d'enrichir l'information spectrale liée à la teneur en COS dans le jeu CAL. Ainsi, le modèle a pu mieux apprendre lors de la phase de calibration.

3.3.2 External Parameter Othogonalisation (EPO)

La **Figure 13** présente les résultats de **RMSECV** en fonction de NLV pour chaque nombre de composante principale (NCP) testée pour le paramétrage de la méthode EPO. A partir de ce graphique, la NLV optimale pour chaque NCP a pu être choisie.

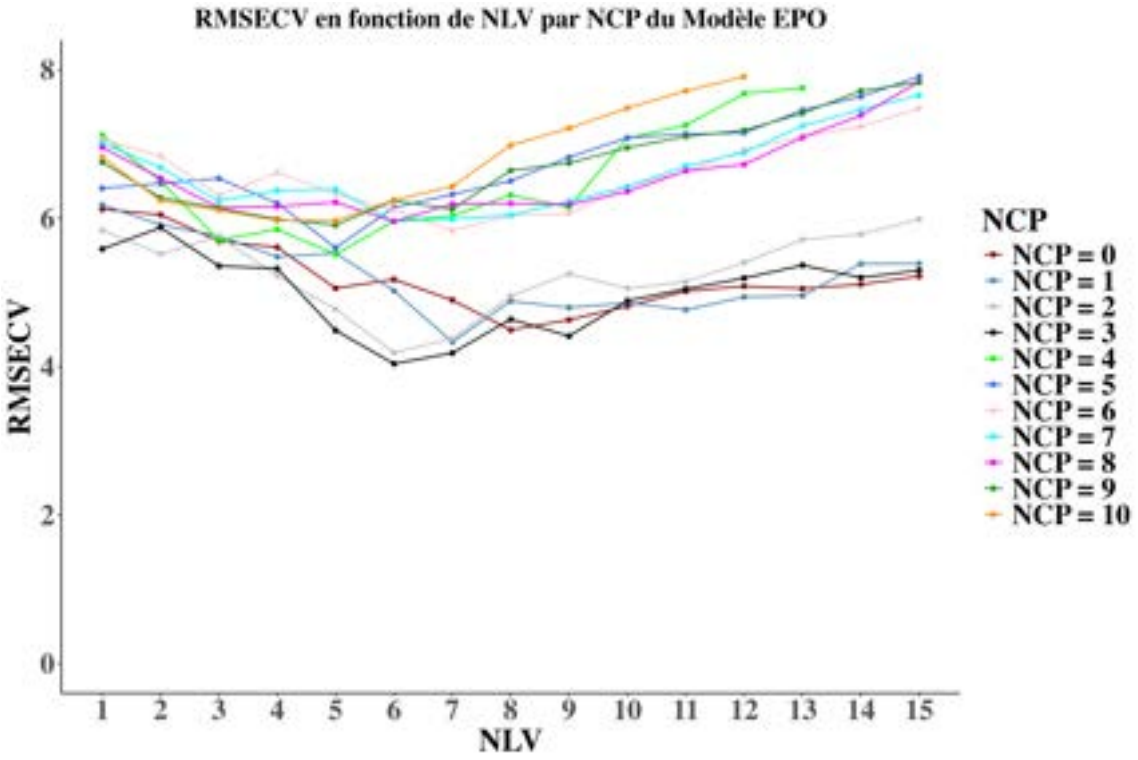


FIGURE 13 – Résultats de validation croisée obtenus après correction par la méthode EPO

La **Figure 14** présente pour chaque NCP testé, l'écart quadratique moyen entre (i) les valeurs prédites de la seconde partie du jeu STD (15 échantillons qui n'ont pas servi à construire l'EPO ; voir section 2.8.3) de la BDD SN et (ii) les valeurs prédites de ce même jeu de la BDD BF, après correction du modèle d'étalonnage par EPO. Cet écart sera appelé **RMSEP sur STD** . La RMSEP sur STD a ainsi été utilisée pour choisir le NCP optimal de l'EPO (i.e., celui qui minimisait **RMSEP sur STD**) égal à **8**. Le NLV optimal correspondant à ce NCP a été fixé à 6 (Figure 13).

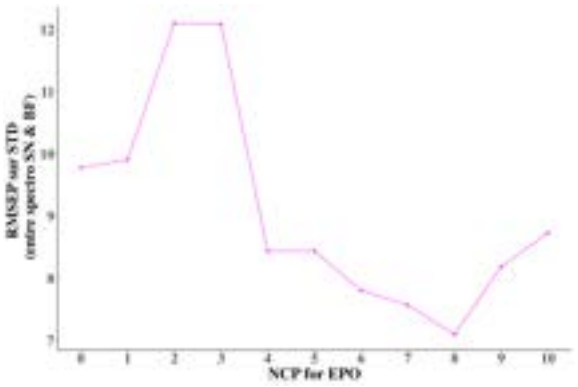


FIGURE 14 – Résultats de méthode EPO

La **Figure 15** présente les résultats de validation externe après application du modèle lwplsr corrigé par EPO sur le jeu VAL-BF.

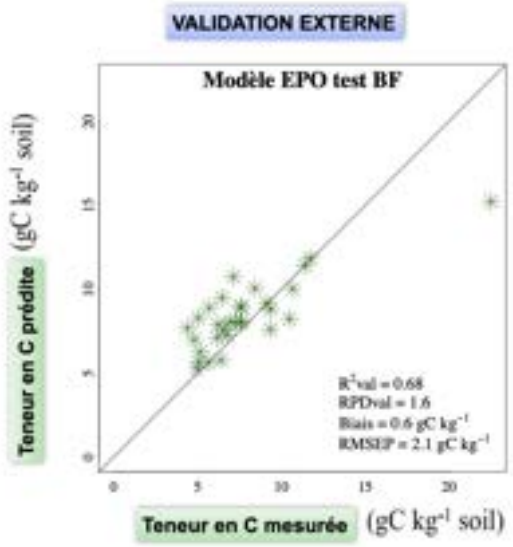


FIGURE 15 – Graphes de la teneur en COS (gC.kg-1) mesurée vs. prédite par lwplsr en validation externe, après correction par EPO

La méthode EPO a permis de corriger la RMSEP et le biais liés au changement d'appareil entre SN et BF : la RMSEP est passée de 4.5 à 2.1 gC.kg-1 (identique à la RMSEP du modèle SN sur VAL-SN) et le biais est passé de 3.1 à 0.6 gC.kg-1 (le biais du modèle SN sur VAL-SN était de 0.5 gC.kg-1). La méthode EPO a donc permis de retrouver une précision de prédiction similaire à celle

du modèle SN sans changement d'appareil. Elle est en revanche moins efficace que les méthodes Update et Double Update, l'avantage de ces dernières étant l'enrichissement du jeu CAL avec de nouveaux échantillons.

3.3.3 Piecewise Direct Standardisation (PDS)

La **Figure 16** illustre d'une part les résultats de validation croisée (RMSECV vs. NLV ; graphe à gauche) du modèle lwplsr obtenus pour chaque fenêtre w testée avec la méthode PDS. Dans le cas de la méthode PDS, la validation croisée a permis de choisir le NLV optimal, fixé à **3** (amenant au RMSECV le plus faible) et pour ce NLV, le w optimal qui a été fixé à 3 également (peu d'effet de w pour NLV = 3). D'autre part, la Figure 16 présente les résultats de validation externe de la prédiction de la teneur en COS après correction PDS en utilisant NLV = 3 et w = 3 (graphe à droite).

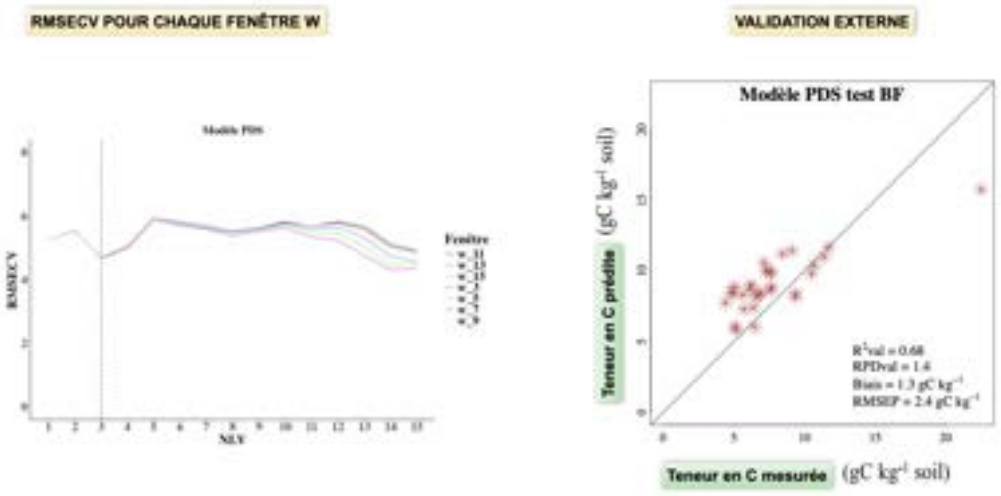


FIGURE 16 – Résultats de la méthode PDS

D'après les résultats de validation externe, la méthode PDS a également permis de corriger les perturbations liées au changement d'appareil, avec une bonne qualité d'ajustement (RMSEP de 2.3 gC.kg-1 vs. 4.5 gC.kg-1 sans transfert d'étalonnage ; biais de 1.3 gC.kg-1 vs. 3.1 gC.kg-1 sans transfert d'étalonnage). Néanmoins, la correction PDS est moins efficace que les méthodes Update, Double Update et EPO car elle a réduit l'erreur d'environ 48% (cette réduction était de 53%-62%

pour les méthodes Update, Double Update et EPO) et le biais est resté presque trois fois supérieur à celui obtenu avec le modèle SN sur VAL-SN.

3.3.4 Correction Biais-Pente (CBP)

Cette méthode est mathématiquement simple à mettre en oeuvre et corrige le biais et la pente des prédictions de la variable d'intérêt. Pour cette méthode dite "a posteriori", l'étalonnage du modèle n'est pas modifié par rapport à un modèle source sur source, donc le même NLV que celui du modèle SN sur SN a été utilisé (i.e., NLV = 5). La CBP a également permis d'améliorer les prédictions du modèle source sur cible étant donné qu'elle a amené à une réduction de la RMSEP de 40% (la RMSEP est passée de 4.5 gC kg^{-1} sans correction à 2.7 gC kg^{-1} après correction CBP). Elle a aussi permis de réduire de biais de **81**% ; le biais étant passé de **3.1** gC kg^{-1} sans correction à **0.6** gC kg^{-1} après correction CBP. Cette méthode réduit toujours efficacement le biais mais dans cette étude, elle n'a pas permis de réduire autant la RMSEP que les méthodes précédentes.

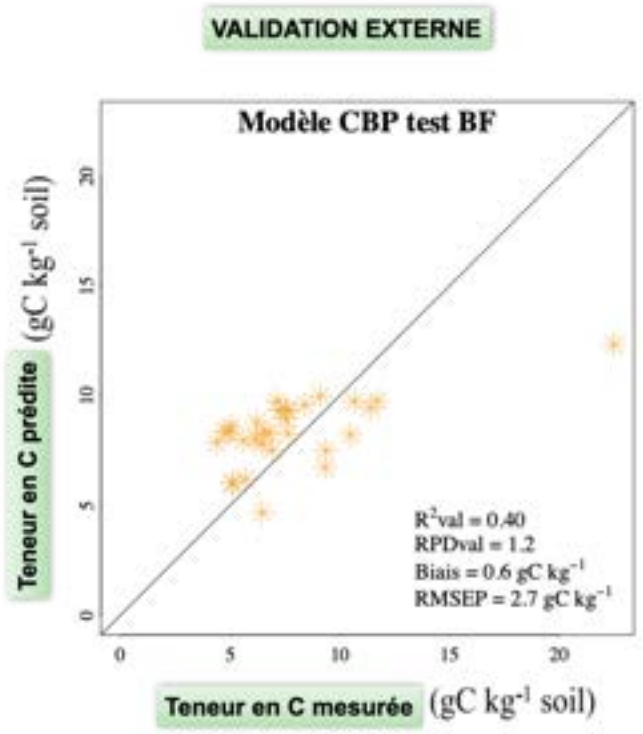


FIGURE 17 – Graphes de la teneur en COS (gC.kg-1) mesurée vs. prédite par lwplsr en validation externe, après correction par CBP

3.4 Bilan des résultats

Le Tableau 2 dresse le bilan des résultats de validation externe pour prédire la teneur en COS avec les différents modèles lwplsr testés (modèle SN sur SN, modèle BF sur BF, modèle SN sur BF sans correction, puis modèle SN sur BF avec les différentes méthodes de correction testées).
La précision des modèles SN sur SN et BF sur BF est similaire (RMSEP de 2.1-2.2 gC.kg-1 ; R^2_{val} de 0.63-0.64 ; biais de 0.4-0.5 gC.kg-1). Cependant, lorsque le modèle est calibré sur SN puis validé sur VAL-BF (SN sur BF) sans correction, les performances ont considérablement chuté avec une R^2_{val} de 0.09, un biais de 3.1 gC kg^{-1}, un RMSEP de 4.5 gC kg^{-1}. Ce résultat a donc souligné la nécessité d'appliquer une méthode de transfert d'étalonnage pour corriger le bruit lié au changement de conditions d'acquisitions spectrales entre SN et BF.
Chaque méthode de transfert testée a permis d'améliorer les résultats de validation par rapport au modèle SN sur BF sans transfert. C'est la méthode Double Update qui a permis d'améliorer le plus efficacement les prédictions, à tel point que celles-ci étaient mêmes plus précises que celles des modèles SN sur SN et BF sur BF (R^2_{val} = 0.78, biais = -0.5 gC kg^{-1}, RMSEP = 1.7 gC kg^{-1}). La méthode "Update" a également permis d'atteindre une précision de validation externe supérieure à celle des modèles SN sur SN et BF sur BF (R^2_{val} = 0.69, biais = 0.3 gC kg^{-1}, RMSEP = 1.9 gC kg^{-1}). L'ajout d'échantillons supplémentaires (31 voire 31 x 2 échantillons) dans un jeu CAL initialement de petite taille (n = 67) a très certainement permis d'ajouter de l'information sur la teneur en COS dans la BDD d'étalonnage. Ainsi, ces deux méthodes n'ont pas seulement servi à corriger les perturbations liées au changement de spectromètre mais aussi à alimenter la BDD d'étalonnage. La méthode EPO a permis de revenir à une précision proche de celle des modèles SN sur SN et BF sur BF (R^2_{val} = 0.68, biais = 0.6 gC kg^{-1}, RMSEP = 2.1 gC kg^{-1}). En revanche, les méthodes PDS et CBP, bien qu'elles aient permis de nettement améliorer la qualité du modèle SN sur BF, n'ont pas permis de revenir à une précision initiale SN sur SN ou BF sur BF : avec PDS, R^2_{val} = 0.68, biais = 1.3 gC kg^{-1}, RMSEP = 2.4 gC kg^{-1} ; avec CBP, R^2_{val} = 0.40, biais = 0.6 gC kg^{-1}, RMSEP = 2.7 gC kg^{-1}). Cette moindre efficacité des méthodes les plus simples mathématiquement (correction des spectres sources pour qu'ils ressemblent à ce qu'ils auraient été s'ils avaient été acquis avec le spectromètre cible ; ou encore, correction des prédictions par simple ajustement du biais et de la pente) montre que dans ce cas d'étude, la perturbation liée au changement de contexte des acquisitions spectrales (changement d'appareil, changement d'opérateur, et entre SN et BF, changement de laboratoire), n'est pas une simple perturbation systématique des spectres. Elle correspond plutôt à une perturbation complexe de la ligne de base

pour laquelle les méthodes PDS et CBP ne suffisent pas. A notre connaissance, il n'existe pas à ce jour de référence bibliographique ayant testé ces quatre méthodes pour corriger des erreurs liées à un changement de contexte d'acquisitions spectrales dans le VPIR. En bilan, malgré ces différences entre méthodes, les quatre méthodes de correction qui ont été testées, en particulier "Double Update", ont considérablement amélioré les performances des modèles en réduisant les biais et les erreurs de prédiction, démontrant leur efficacité dans l'harmonisation des données spectrales entre différents appareils et conditions d'acquisitions.

Métriques / **Modèles**	**Nval**	$R^2_{\mathbf{val}}$	**Biais**	**RMSEP**	**RPDval**
SN sur SN	30	0.63	0.5	2.1	1.6
BF sur BF	30	0.64	0.4	2.2	1.5
SN sur BF	30	0.09	3.1	4.5	0.8
Update	30	0.69	0.3	1.9	1.8
Double Update	30	**0.78**	-0.5	**1.7**	**2**
EPO	30	0.68	0.6	2.1	1.6
PDS	30	0.68	1.3	2.4	1.4
CBP	30	0.40	0.6	2.7	1.2

TABLE 2 – Table de Bilan des résultats

4 Perspectives

Ce stage a permis de tester cinq méthodes de transfert d'étalonnage : Update, Double Update, EPO, CBP et PDS. Ces méthodes ont toutes prouvé leur capacité à corriger des perturbations liées à un changement d'appareil. Néanmoins, les méthodes les plus efficaces dans ce cas d'étude étaient différentes de celles trouvées dans l'étude de Vova Martirosyan (stagiaire M2 d'Aurélie Cambou en 2023). Dans son cas, le jeu d'échantillon était le même mais la gamme spectrale étudiée était uniquement le proche infrarouge (1100-2500 nm) et un seul opérateur avait acquis les BDDs spectrales dans un même laboratoire (Montpellier). De plus, il avait appliqué une PLSR globale et non une lwplsr. Dans son étude, CBP et PDS donnaient les meilleurs résultats de correction, alors que Double Update et EPO amenaient aux pires résultats. Des études approfondies doivent donc être menées sur de nouveaux jeux de données pour mieux comprendre les effets de ces différentes méthodes, dans différents contextes. Dans cette étude, les différents cas testés lors de ces travaux

ont montré que ce changement d'appareil ne générait pas systématiquement des perturbations. Par exemple, pour le cas MPL source et BF cible, le modèle de prédiction de la teneur en COS n'a pas été détérioré malgré le changement de spectromètre (Tableau 2 ; Figure 11 ; Annexes). En revanche, dans les deux autres cas (SN source/BF cible et MPL source/SN cible), le changement d'appareil a amené à une détérioration forte du modèle de prédiction de la teneur en COS. L'implication de la BDD SN dans ces deux derniers cas pose question, notamment sur le mode d'acquisition des données spectrales. Par exemple, les conditions de conservation des échantillons avant les acquisitions spectrales ont-elles été similaires avec celles adoptées pour les BDDs MPL et BF ? L'effet "opérateur" a également pu jouer un rôle important dans ces divergences, puisque chaque BDD a été acquise par un opérateur différent. Etant donné que pour les spectromètres de marque ASD, les acquisitions spectrales sont réalisées par contact directe de la sonde avec l'échantillon, il est possible que l'angle de la sonde ou le poids de la main appliqué sur la sonde ait été différent, ce qui a pu avoir un effet sur les spectres. De plus, le spectromètre BF est installé au Burkina Faso mais les acquisitions spectrales ont eu lieu à **Montpellier**, dans le même laboratoire que celui des acquisitions spectrales de la BDD MPL. L'effet "conditions de laboratoire" (e.g., température, humidité, propreté du Spectralon utilisé comme référence) peut donc avoir induit une différence entre les acquisitions des BDDs MPL et BF vs. les acquisitions de la BDD SN.

La partie dédiée à l'optimisation des modèles (choix du NLV optimal) a pris une place importante dans mon travail de stage bien qu'initialement ce n'était pas l'objectif principal du stage. En effet, le choix du NLV a eu un très fort effet sur les résultats de validation externe, et il était donc difficile de ne pas étudier cette problématique plus précisément. Ainsi, plusieurs stratégies ont été testées pour choisir le NLV optimal. Ces stratégies ont été : (i) une validation croisée basée sur une séparation des blocs par site d'étude, (ii) une validation croisée basée sur un choix aléatoire de deux blocs répété 30 fois, (iii) un Bootstrap (100 répétitions avec remise) intégré à une validation croisée (voir Annexes). Pour chacune de ces méthodes, la RMSECV a été calculée pour chaque NLV. Pour la méthode de validation croisée par site, des minima locaux étaient souvent visibles lorsque NLV était compris entre 3 et 5, particulièrement pour l'approche locale. La sélection du NLV correspondant à ces minima locaux amenait à des résultats de validation externe parfois deux fois meilleurs que lorsque c'était le NLV correspondant au RMSECV minimal qui était choisi. Cette tendance était beaucoup plus visible avec l'approche locale qu'avec l'approche globale.

Ainsi, ce stage a mis en avant que les méthodes usuelles de sélection de NLV ne sont pas toujours adéquates et qu'il est nécessaire d'étudier plus précisément l'effet du NLV sur les résultats de validation externe dans différents cas d'étude, afin d'aller vers une méthode de sélection plus générique

et plus représentative. Les méthodes qui ont été testées pendant ce stage ne sont pas exhaustives. Il pourrait être intéressant d'aller vers d'autres méthodes de sélection de NLV telles que des méthodes multi-critères, donc basées sur une gamme plus large d'indicateurs que la simple observation de la RMSECV (les critères sont encore à définir), le Bootstrap à l'intérieur de la validation croisée, ou encore des méthodes basées sur la vraisemblance (BIC, AIC, CP de Mallow, etc). Ce travail de stage a donc contribué à ouvrir la réflexion sur l'optimisation des modèles PLSR.
L'optimisation de la lwplsr demeure un point important à étudier dans la suite de ces travaux. Cette approche est pertinente dans le cas où l'on observe une certaine hétérogénéité des données. Malgré sa performance, la lwplsr requiert une régularisation fine des hyperparamètres tels que le nombre de voisins à considérer, la forme de la fonction de poids, le NLV nécessaire pour calculer la dissimilarité entre les observations. Pour ce stage, ces hyperparamètres ont été fixés à l'avance suite aux travaux de stage de Vova Martirosyan et l'expertise de mes encadrants de stage. Pour la suite de ces travaux, ces paramètres pourraient varier en utilisant des approches utilisées en Machine Learning telles que la "recherche par grille", l'optimisation bayésienne ou d'autres approches d'optimisation mathématique.
Malgré les bons résultats produits par la lwplsr, il pourrait être intéressant de tester d'autres méthodes de prédiction, notamment les méthodes Machine Learning et Deep Learning qui nécessitent une quantité bien plus importante de données d'entraînement. L'utilisation de ces méthodes peut néanmoins se heurter à la complexité d'acquisition des données pédologiques (campagnes d'échantillonnage sur le terrain, analyses des données au laboratoire, nécessité parfois de stocker les échantillons de sol). Ces données sont souvent coûteuses à obtenir, rendant la construction de larges BDDs spectrales longue au sein d'une même équipe de recherche. Il pourrait être pertinent d'utiliser des techniques de data augmentation en Deep Learning pour challenger ce déficit. Enfin, les méthodes de Transfer Learning, notamment "Domaine Adaptation" sont à tester pour offrir une alternative aux méthodes classiques de transfert d'étalonnage qui ont été appliquées pendant ce stage.

5 Conclusion

Les cinq méthodes de transfert d'étalonnage testées dans ce travail de stage ont été efficaces pour corriger les perturbations liées au changement de contexte d'acquisitions spectrales. Entre les BDDs SN et BF, la méthode Double Update, puis Update suivies par EPO ont été plus efficaces en termes de qualité d'ajustement et de précision par rapport aux méthodes PDS et CBP, en utilisant

une approche locale. La qualité de correction de ces méthodes constitue ainsi une avancée vers la levée de la limite de l'interopérabilité des BDDs obtenues avec différents spectromètres. Cependant, les méthodes de transfert d'étalonnage les plus **optimales** dans ce stage étaient à l'opposé de celles trouvées lors du stage de Vova Martirosyan (2023) à partir du même jeu d'échantillons mais en PLSR globale et avec d'autres spectromètres. Ceci suggère la nécessité d'aller plus loin sur l'étude de ces méthodes avec de nouveaux jeux de données, différents spectromètres et différentes conditions d'acquisitions.

Dans ce stage, la lwplsr a prouvé sa performance de prédiction de la teneur en COS ; elle était notamment plus précise que l'approche globale. Néanmoins, pour ces deux approches, le choix du NLV optimal s'est avéré crucial et fastidieux. Ce stage a permis une ouverture de la réflexion vers d'autres méthodes de sélection des NLVs, notamment le Bootstrap qui semble offrir une alternative robuste basée sur le ré-échantillonage. Néanmoins, il est encore nécessaire de creuser cette problématique pour de nouveaux jeux de données, afin d'aller vers une meilleure généricité et représentativité des méthodes de sélection du NLV.

6 Annexes

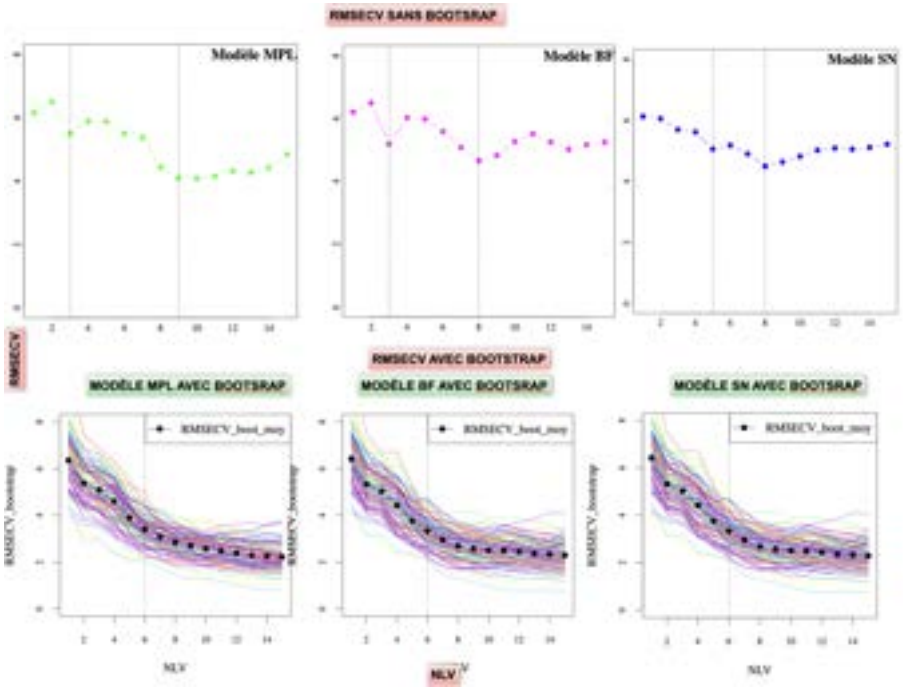
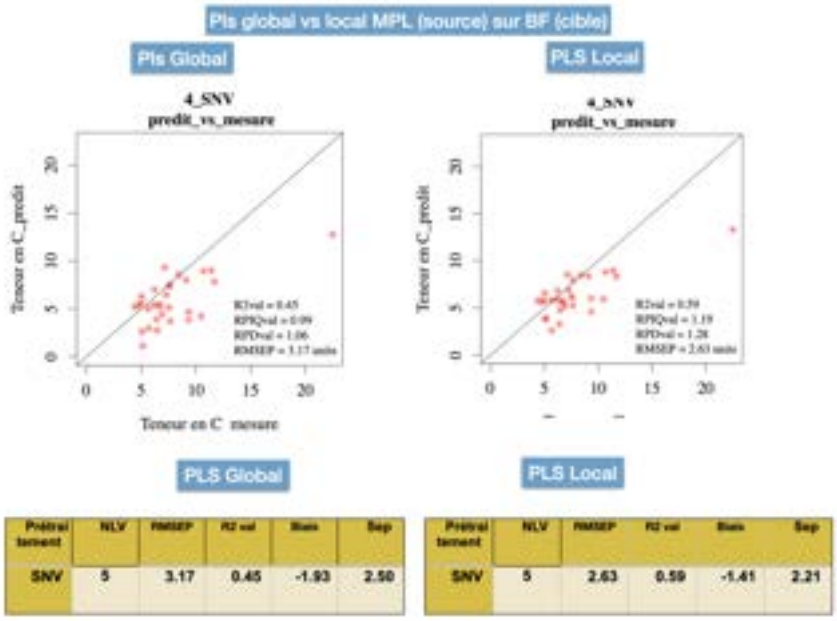


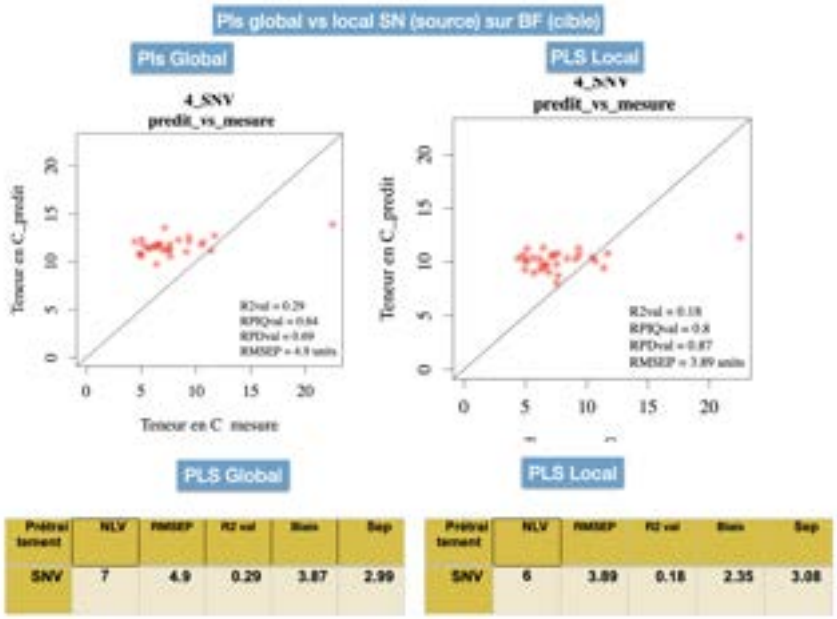
FIGURE 18 – RMSECV lwplsr



Prétraitement	NLV	RMSEP	R2 val	Biais	Sep
SNV	5	3.17	0.45	-1.93	2.50

Prétraitement	NLV	RMSEP	R2 val	Biais	Sep
SNV	5	2.63	0.59	-1.41	2.21

FIGURE 19 – RMSECV lwplsr



Prétraitement	NLV	RMSEP	R2 val	Biais	Sep
SNV	7	4.9	0.29	3.87	2.99

Prétraitement	NLV	RMSEP	R2 val	Biais	Sep
SNV	6	3.89	0.18	2.35	3.08

FIGURE 20 – RMSECV lwplsr

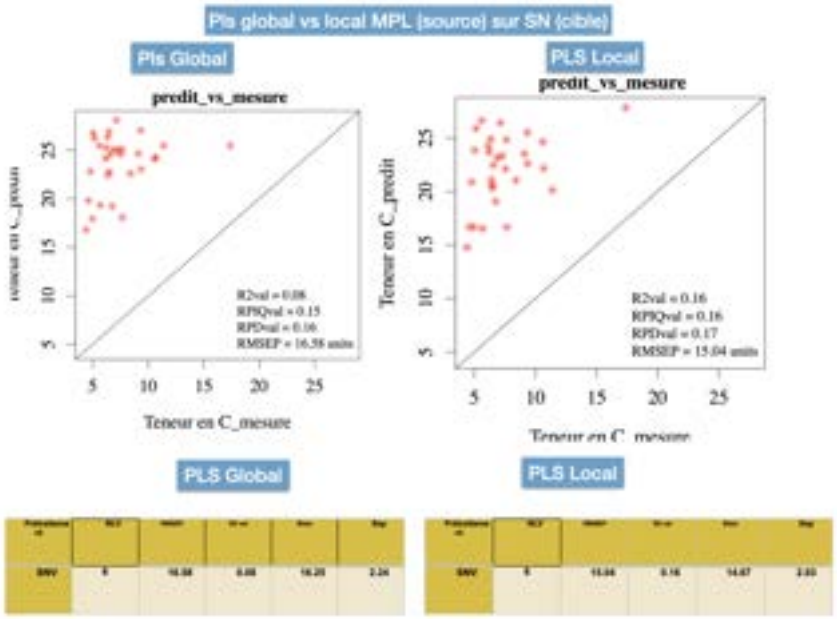
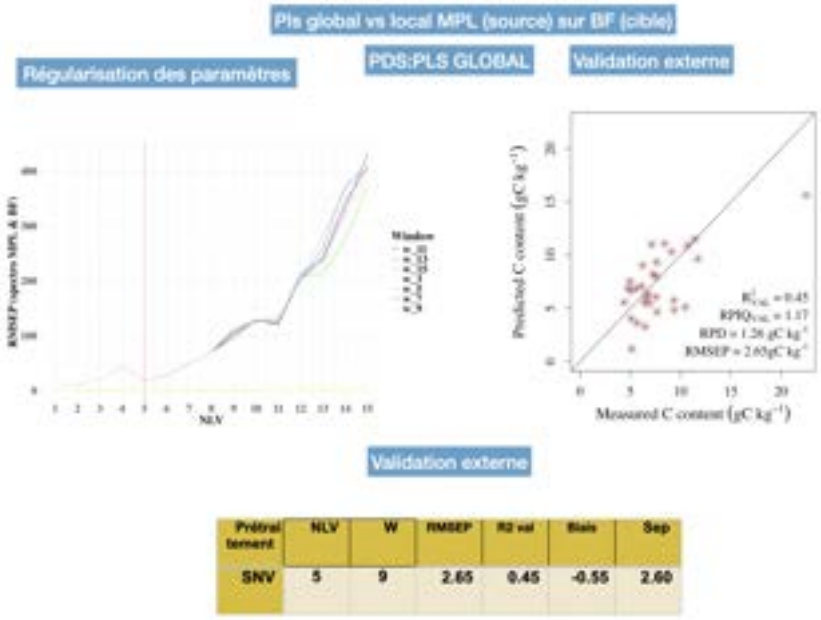
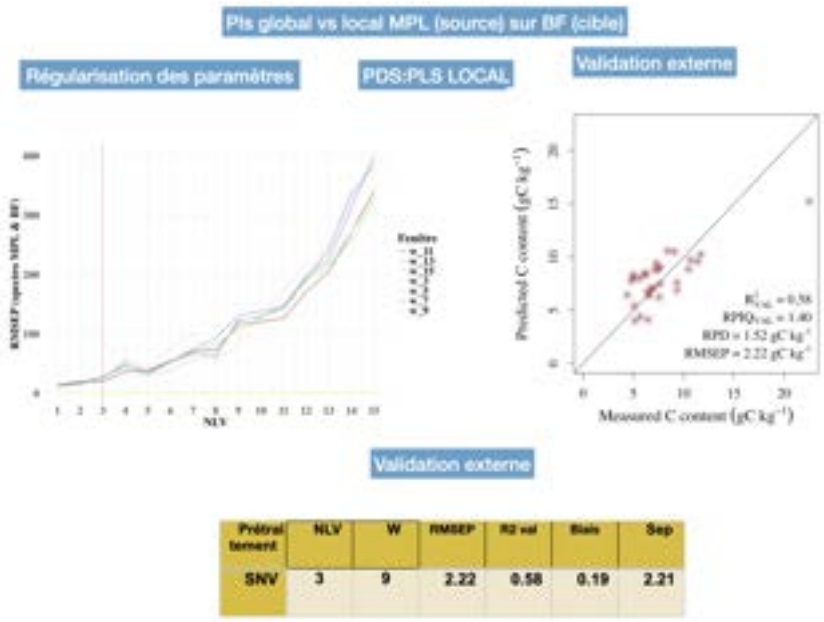


FIGURE 21 – RMSECV lwplsr



Prétraitement	NLV	W	RMSEP	R2 val	Biais	Sep
SNV	5	9	2.65	0.45	-0.55	2.60

FIGURE 22 – RMSECV lwplsr



Prétraitement	NLV	W	RMSEP	R2 val	Biais	Sep
SNV	3	9	2.22	0.58	0.19	2.21

FIGURE 23 – RMSECV lwplsr

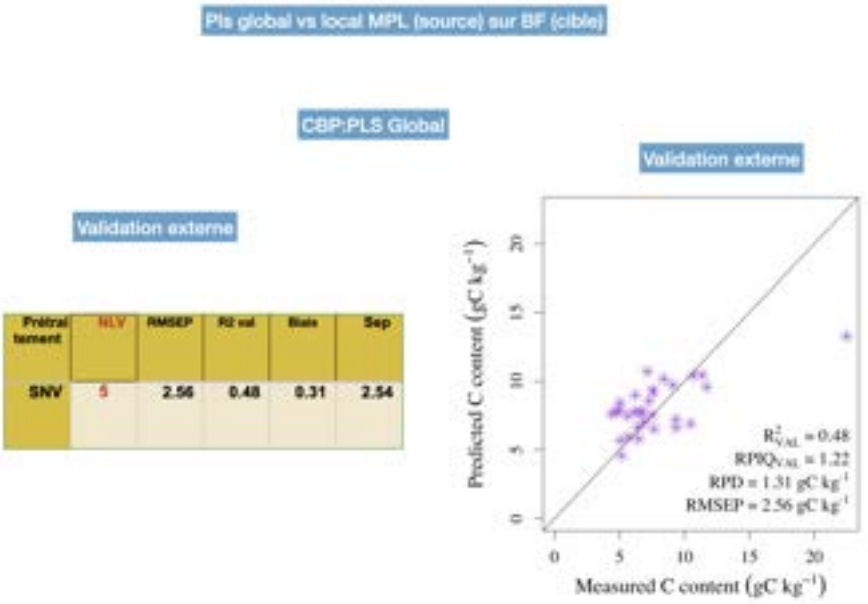


FIGURE 24 – RMSECV lwplsr

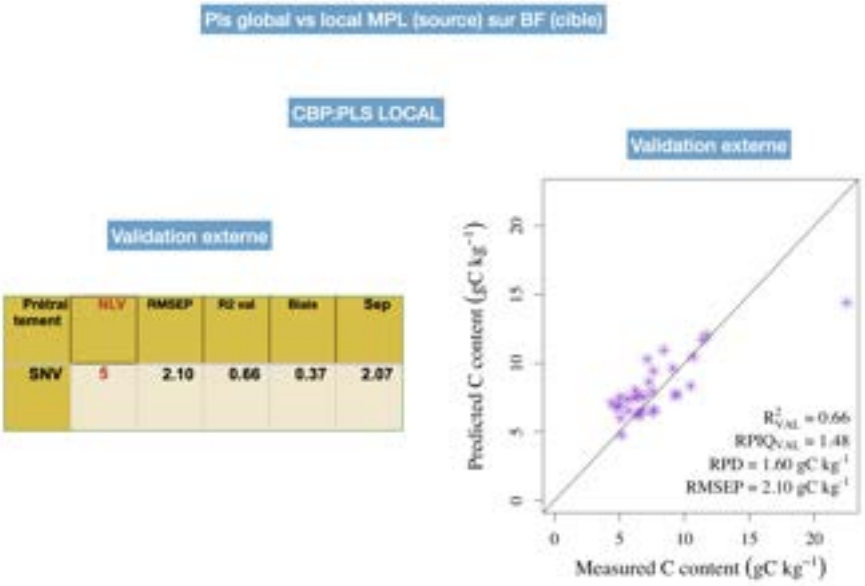


FIGURE 25 – RMSECV lwplsr

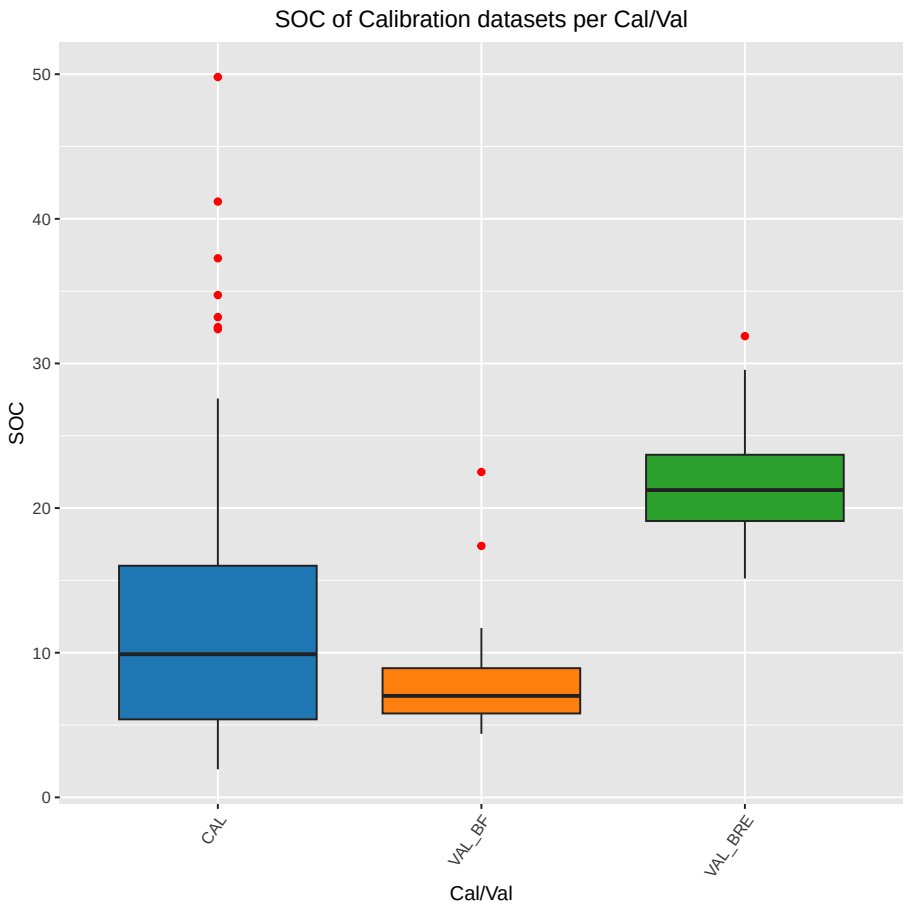


FIGURE 26 – COS par cal et projection sur val

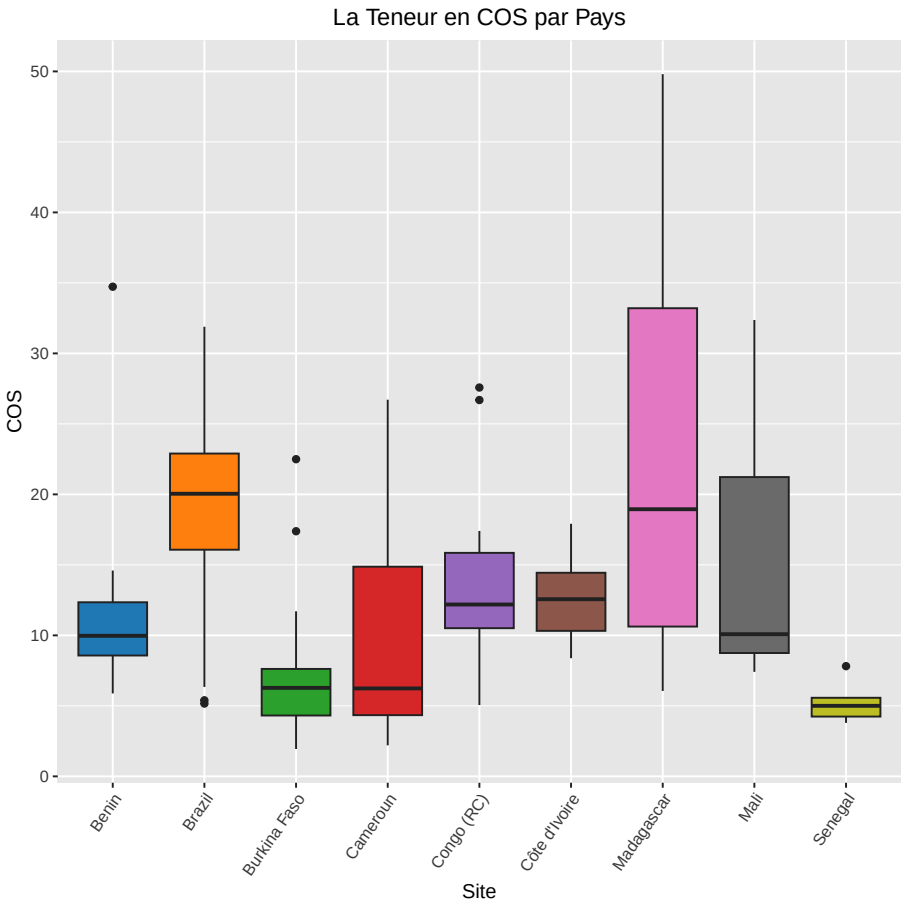


FIGURE 27 – COS par PAYS

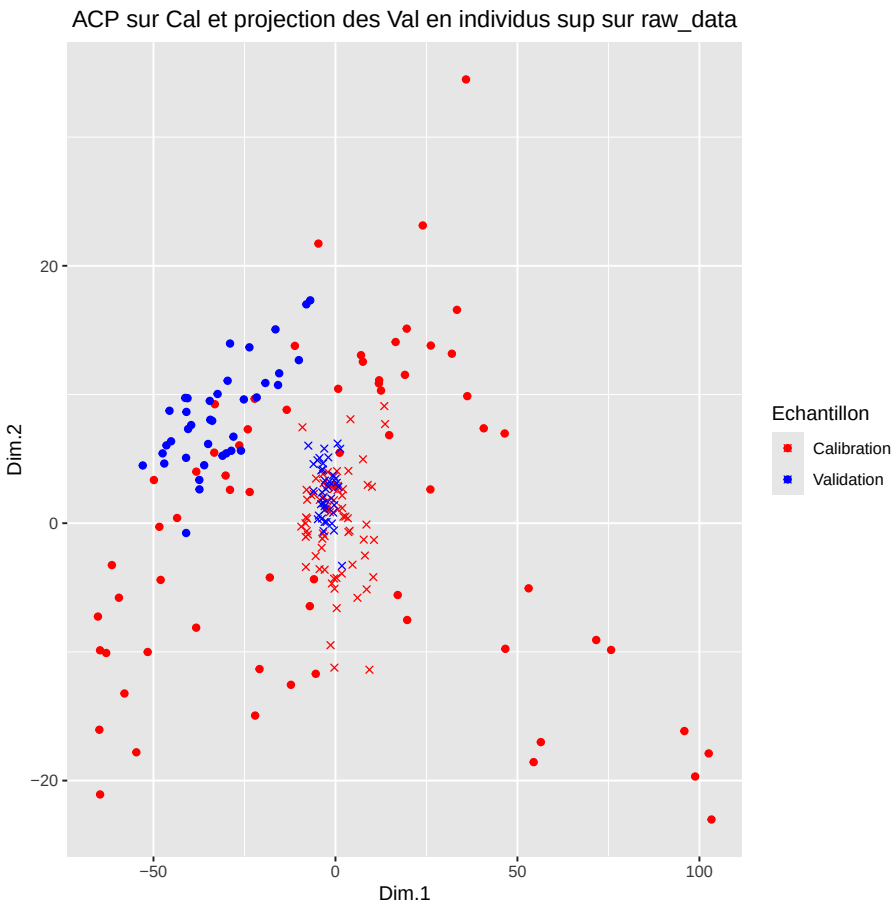


FIGURE 28 – ACP sur cal et prjeition des val en individus sup

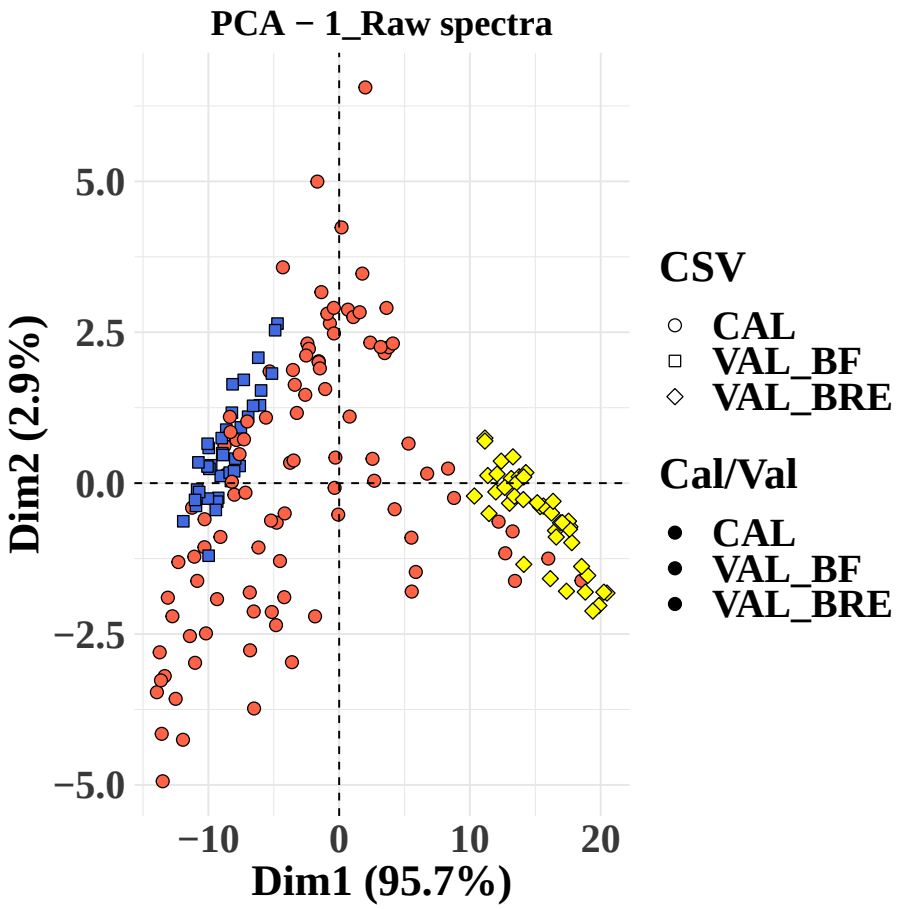


FIGURE 29 – Spectres

7 Bibliographie

Références

[1] Global change and carbon cycle : The position of soils and agriculture - universites montpellier.

[2] lwplsr : KNN-LWPLSR in mlesnoff/rchemo : Dimension reduction, regression and discrimination for chemometrics.

[3] Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra - r. j. barnes, m. s. dhanoa, susan j. lister, 1989.

[4] Standardisation and calibration transfer for near infrared instruments : A review - tom fearn, 2001.

[5] Thermal degradation of poly(3-hydroxybutyrate) and poly(3-hydroxybutyrate-co-3-hydroxyhexanoate) in nitrogen and oxygen studied by thermogravimetric–fourier transform infrared spectroscopy - christian vogel, shigeaki morita, harumi sato, isao noda, yukihiro ozaki, heinz w. siesler, 2007.

[6] S. Amat-Tosello, N. Dupuy, and J. Kister. Contribution of external parameter orthogonalisation for calibration transfer in short waves—near infrared spectroscopy application to gasoline quality. 642(1) :6–11.

[7] Erik Andries, John H. Kalivas, and Anit Gurung. Sample and feature augmentation strategies for calibration updating. 33(1) :e3080. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/cem.3080.

[8] Emmanuel A. Badewa, Chun C. Yeung, Joann K. Whalen, and Maren Oelbermann. Compost and biosolids increase long-term soil organic carbon stocks. 103(3) :483–492. Publisher : NRC Research Press.

[9] R. J. Barnes, M. S. Dhanoa, and Susan J. Lister. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. 43(5) :772–777. Publisher : SAGE Publications Ltd STM.

[10] B. Barthès, E. Kouakoua, G. H. Sala, C. Hartmann, and B. Nyeté. Effet à court terme de la mise en culture sur le statut organique et l'agrégation d'un sol ferrallitique argileux du congo. 76(4) :493–499. Publisher : NRC Research Press.

[11] Marion Brandolini-Bunlon, Benoit Jaillais, Jean-Michel Roger, and Matthieu Lesnoff. rchemo : Dimension reduction, regression and discrimination for chemometrics.

[12] Aurélie Cambou, Tiphaine Chevallier, Bernard G. Barthès, Delphine Derrien, Patrice Cannavo, Adeline Bouchard, Victor Allory, Christophe Schwartz, and Laure Vidal-Beaudet. The impact of urbanization on soil organic carbon stocks and particle size and density fractions. 23(2) :792–803.

[13] Aurélie Cambou, Issiakou A. Houssoukpèvi, Tiphaine Chevallier, Patricia Moulin, Nancy M. Rakotondrazafy, Eltson E. Fonkeng, Jean-Michel Harmand, Hervé N. S. Aholoukpè, Guillaume L. Amadji, Fritz O. Tabi, Lydie Chapuis-Lardy, and Bernard G. Barthès. Quantification of soil organic carbon in particle size fractions using a near-infrared spectral library in west africa. 443 :116818.

[14] Cheng-Wen Chang, David A. Laird, Maurice J. Mausbach, and Charles R. Hurburgh. Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. 65(2) :480–490. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.2136/sssaj2001.652480x.

[15] Yuan-yuan Chen and Zhi-bin Wang. Cross components calibration transfer of NIR spectroscopy model through PCA and weighted ELM-based TrAdaBoost algorithm. 192 :103824.

[16] Michaël Clairotte, Clovis Grinand, Ernest Kouakoua, Aurélie Thébault, Nicolas P. A. Saby, Martial Bernoux, and Bernard G. Barthès. National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy. 276 :41–52.

[17] John B. Cooper, Christopher M. Larkin, and Mohamed F. Abdelkader. Calibration transfer of near-IR partial least squares property models of fuels using virtual standards. 25(9) :496–505. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/cem.1395.

[18] Minerva J. Dorantes, Bryan A. Fuentes, and David M. Miller. Calibration set optimization and library transfer for soil carbon estimation using soil spectroscopy—a review. 86(4) :879–903. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/saj2.20435.

[19] Tom Fearn. Standardisation and calibration transfer for near infrared instruments : A review. 9(4) :229–244. Publisher : SAGE Publications Ltd STM.

[20] Valeria Fonseca Diaz, Jean-Michel Roger, and Wouter Saeys. Unsupervised dynamic orthogonal projection. an efficient approach to calibration transfer without standard samples. 1225 :340154.

[21] K. Fujisaki, A.-S. Perrin, M. Boussafir, S. Gogo, M. Sarrazin, and M. Brossard. Decomposition kinetics and organic geochemistry of woody debris in a ferralsol in a humid tropical climate. 66(5) :876–885. Num Pages : 10 Place : Hoboken Publisher : Wiley Web of Science ID : WOS :000361187000006.

[22] Yasas Gamagedara, Nuwan K. Wijewardane, Gary Feng, Cathy Seybold, Michael Williams, Mary Love Tagert, and Vitor S. Martins. Can we use a mid-infrared fine-ground soil spectral library to predict non-fine-ground spectra ? 443 :116799.

[23] Yufeng Ge, Cristine L. S. Morgan, Sabine Grunwald, David J. Brown, and Deoyani V. Sarkhot. Comparison of soil reflectance spectra and calibration models obtained using multiple spectrometers. 161(3) :202–211.

[24] M. H. Gerzabek, F. Strebl, M. Tulipan, and S. Schwarz. Quantification of organic carbon pools for austria's agricultural soils using a soil information system. 85 :491–498. Publisher : NRC Research Press.

[25] C. Grinand, B. G. Barthès, D. Brunet, E. Kouakoua, D. Arrouays, C. Jolivet, G. Caria, and M. Bernoux. Prediction of soil organic and inorganic carbon contents at a national scale (france) using mid-infrared reflectance spectroscopy (MIRS). 63(2) :141–151. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2389.2012.01429.x.

[26] C. Guerrero, B. Stenberg, J. Wetterlind, R. A. Viscarra Rossel, F. T. Maestre, A. M. Mouazen, R. Zornoza, J. D. Ruiz-Sinoga, and B. Kuang. Assessment of soil organic carbon at local scale with spiked NIR calibrations : effects of selection and extra-weighting on the spiking subset. 65(2) :248–263. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/ejss.12129.

[27] John Kort, Jim Richardson, Raju Soolanayakanahally, and William Schroeder. Innovations in temperate agroforestry : the 13th north american agroforestry conference. 88(4) :563–567.

[28] Boyan Kuang and Abdul Mounem Mouazen. Effect of spiking strategy and ratio on calibration of on-line visible and near infrared soil sensor for measurement in european farms. 128 :125–136.

[29] Pascal Lavergne and Valentin Patilea. Breaking the curse of dimensionality in nonparametric testing. 143(1) :103–122.

[30] Matthieu Lesnoff. mlesnoff/rchemo. original-date : 2021-04-06T09 :36 :30Z.

[31] Maxime Metz, Florent Abdelghafour, Jean-Michel Roger, and Matthieu Lesnoff. A novel robust PLS regression method inspired from boosting principles : RoBoost-PLSR. 1179 :338823.

[32] Muhammad Abdul Munnaf and Abdul Mounem Mouazen. Removal of external influences from on-line vis-NIR spectra for predicting soil organic carbon using machine learning. 211 :106015.

[33] Muhammad Abdul Munnaf and Abdul Mounem Mouazen. Removal of external influences from on-line vis-NIR spectra for predicting soil organic carbon using machine learning. 211 :106015.

[34] Wartini Ng, Leigh Ann Winowiecki, Valentine Karari, Elvis Weullow, Dickens Alubaka Ateku, Tor-Gunnar Vågen, Zampela Pittaki, and Budiman Minasny. Exploring mid-infrared spectral transfer functions for the prediction of multiple soil properties using a global dataset. 88(4) :1234–1247. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/saj2.20697.

[35] Wangdong Ni, Steven D. Brown, and Ruilin Man. Data fusion in multivariate calibration transfer. 661(2) :133–142.

[36] Wangdong Ni, Steven D. Brown, and Ruilin Man. Data fusion in multivariate calibration transfer. 661(2) :133–142.

[37] Wangdong Ni, Steven D. Brown, and Ruilin Man. Stacked PLS for calibration transfer without standards. 25(3) :130–137. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/cem.1369.

[38] J. Padarian, B. Minasny, and A. B. McBratney. Transfer learning to localise a continental soil vis-NIR calibration model. 340 :279–288.

[39] Zampela Pittaki-Chrysodonta, Alfred E. Hartemink, Jonathan Sanderman, Yufeng Ge, and Jingyi Huang. Evaluating three calibration transfer methods for predictions of soil properties using mid-infrared spectroscopy. 85(3) :501–519. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/saj2.20225.

[40] Jean-Michel Roger and Jean-Claude Boulet. A review of orthogonal projections for calibration. 32(9) :e3045. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/cem.3045.

[41] Jean-Michel Roger, Fabien Chauchard, and Véronique Bellon-Maurel. EPO–PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits. 66(2) :191–204.

[42] José L. Safanelli, Jonathan Sanderman, Dellena Bloom, Katherine Todd-Brown, Leandro L. Parente, Tomislav Hengl, Sean Adam, Franck Albinet, Eyal Ben-Dor, Claudia M. Boot, James H. Bridson, Sabine Chabrillat, Leonardo Deiss, José A. M. Demattê, M. Scott Demyan, Gerd Dercon, Sebastian Doetterl, Fenny van Egmond, Rich Ferguson, Loretta G. Garrett, Michelle L. Haddix, Stephan M. Haefele, Maria Heiling, Javier Hernandez-Allica, Jingyi Huang, Julie D. Jastrow, Konstantinos Karyotis, Megan B. Machmuller, Malefetsane Khesuoe, Andrew Margenot, Roser Matamala, Jessica R. Miesel, Abdul M. Mouazen, Penelope Nagel, Sunita Patel, Muhammad Qaswar, Selebalo Ramakhanna, Christian Resch, Jean Robertson, Pierre Roudier, Marmar Sabetizade, Itamar Shabtai, Faisal Sherif, Nishant Sinha, Johan Six, Laura Summerauer, Cathy L. Thomas, Arsenio Toloza, Beata Tomczyk-Wójtowicz, Nikolaos L. Tsakiridis,

Bas van Wesemael, Finnleigh Woodings, George C. Zalidis, and Wiktor R. Żelazny. An interlaboratory comparison of mid-infrared spectra acquisition : Instruments and procedures matter. 440 :116724.

[43] Abraham. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. 36(8) :1627–1639. Publisher : American Chemical Society.

[44] Parviz Shahbazikhah and John H. Kalivas. A consensus modeling approach to update a spectroscopic calibration. 120 :142–153.

[45] Peng Shan, Yuhui Zhao, Qiaoyun Wang, Shuyu Wang, Yao Ying, and Silong Peng. A nonlinear calibration transfer method based on joint kernel subspace. 210 :104247.

[46] Zefang Shen, Leonardo Ramirez-Lopez, Thorsten Behrens, Lei Cui, Mingxi Zhang, Lewis Walden, Johanna Wetterlind, Zhou Shi, Kenneth A Sudduth, Philipp Baumann, Yongze Song, Kevin Catambay, and Raphael A. Viscarra Rossel. Deep transfer learning of global spectra for local soil carbon monitoring. 188 :190–200.

[47] Heinz W. Siesler. Vibrational spectroscopy of polymers. 16(8) :519–541. Publisher : Taylor & Francis _eprint : https ://doi.org/10.1080/1023666X.2011.620234.

[48] Erik Tengstrand, Lars Erik Solberg, Katinka Dankel, Tiril Aurora Lintvedt, Nils Kristian Afseth, and Jens Petter Wold. Calibration transfer between different spectrometers by wavelength correspondence. 132 :103667.

[49] Yan-bin Wang, Hong-fu Yuan, and Wan-zhen Lu. [a new calibration transfer method based on target factor analysis]. 25(3) :398–401.

[50] Nuwan K. Wijewardane, Yufeng Ge, Jonathan Sanderman, and Richard Ferguson. Fine grinding is needed to maintain the high accuracy of mid-infrared diffuse reflectance spectroscopy for soil property estimation. 85(2) :263–272. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/saj2.20194.

[51] Herman Wold. Soft modelling by latent variables : The non-linear iterative partial least squares (NIPALS) approach. 12 :117–142.

[52] Zhuopin Xu, Shuang Fan, Jing Liu, Binmei Liu, Liangzhi Tao, Jin Wu, Shupeng Hu, Liping Zhao, Qi Wang, and Yuejin Wu. A calibration transfer optimized single kernel near-infrared spectroscopic method. 220 :117098.

[53] Bin Yang, Lijun Yao, and Tao Pan. Near-infrared spectroscopy combined with partial least squares discriminant analysis applied to identification of liquor brands. 9(2) :181–189. Number : 2 Publisher : Scientific Research Publishing.

[54] Xien Yin Yap, Kim Seng Chia, and Nur Aisyah Syafinaz Suarin. Adaptive artificial neural network in near infrared spectroscopy for standard-free calibration transfer. 230 :104674.

[55] Feiyu Zhang, Wanchao Chen, Ruoqiu Zhang, Boyang Ding, Heming Yao, Jiong Ge, Lei Ju, Wuye Yang, and Yiping Du. Sampling error profile analysis for calibration transfer in multivariate calibration. 171 :234–240.

[56] Feiyu Zhang, Ruoqiu Zhang, Jiong Ge, Wanchao Chen, Wuye Yang, and Yiping Du. Calibration transfer based on the weight matrix (CTWM) of PLS for near infrared (NIR) spectral analysis. 10(18) :2169–2179. Publisher : The Royal Society of Chemistry.

[57] Ya-Nan Zhao, Yue-Qiang Zhang, Hong-Xia Du, Yue-Hong Wang, La-Mei Zhang, and Xiao-Jun Shi. Carbon sequestration and soil microbes in purple paddy soil as affected by long-term fertilization. 97(3) :464–476. Publisher : Taylor & Francis _eprint : https ://doi.org/10.1080/02772248.2015.1050200.

[58] Yong Zheng, Zhengkun Hu, Xu Pan, Xiaoyun Chen, Delphine Derrien, Feng Hu, Manqiang Liu, and Stephan Hättenschwiler. Carbon and nitrogen transfer from litter to soil is higher in slow than rapid decomposing plant litter : A synthesis of stable isotope studies. 156 :108196.