Potentiel de la spectrométrie infrarouge comme outil de caractérisation des indicateurs microbiologiques des sols

Pierre VALLANTIN

Abréviations :

ACP : analyse en composante principale
AFES : l'Association Française d'étude du Sol
B-glu : β-glucosidase
CEC : capacité d'échange cationique
Cmic : carbone de la biomasse microbienne
LOO : leave-one-out
MIR : moyen-infrarouge
Nag : N-acétyl-β-D-glucosaminidase
Nmic : azote de la biomasse microbienne
PCR : régressions en composantes principales
PIR : proche-infrarouge
PLFA : acides gras dérivés des phospholipides
PLSR : régressions par les moindres carrés partiels
Pmic : phosphore de la biomasse microbienne
RMSE : erreur quadratique moyenne
RMSEcv : erreur quadratique moyenne de la validation croisée
RMSEP : erreur quadratique moyenne de prédiction
RPD : ratio de performance à la déviation
RPIQ : ratio de performance à l'écart interquartile
SECV : erreur standard de validation croisée
Smic : soufre de la biomasse microbienne
SPIR : spectroscopie proche infrarouge
VIS-PIR : visible-proche-infrarouge

Table des matières

1 INTRODUCTION : CONTEXTE ET ENJEUX

1.1 Le sol

1.1.1 Le sol : un milieu complexe et hétérogène

Le sol est défini selon l'Association Française d'étude du Sol (AFES) comme « un volume qui s'étend depuis la surface de la Terre jusqu'à une profondeur marquée par l'apparition d'une roche dure ou meuble, peu altérée ou peu marquée par la pédogenèse. L'épaisseur du sol peut varier de quelques centimètres à quelques dizaines de mètres, ou plus. Il constitue, localement, une partie de la couverture pédologique qui s'étend à l'ensemble de la surface de la Terre. Il comporte le plus souvent plusieurs horizons correspondant à une organisation des constituants organiques et/ou minéraux (la terre). Cette organisation est le résultat de la pédogenèse et de l'altération du matériau parental. Il est le lieu d'une intense activité biologique (racines, faune et microorganismes). »

Le sol est une matrice complexe (Gobat et al. 2010) constituée de particules minérales et organiques, ainsi que d'organismes vivants organisés de tel sorte qu'il constitue un milieu triphasique composé :

- D'une phase solide constituée majoritairement de minéraux de taille et nature minéralogique variable, ainsi que de constituants organiques dans lesquels on retrouve les organismes vivants du sol et les matières organiques mortes.
- D'une phase liquide qui correspond à la solution du sol, lieu de réalisation de nombreuses transformations chimiques.
- D'une phase gazeuse contenue dans la partie non occupée par la solution du sol de l'espace poral résultant de l'organisation des particules et des agrégats de la phase solide.

L'arrangement des divers constituants des sols à différentes échelles est à l'origine de leur structure. On distingue plusieurs niveaux d'organisation allant des plus fins (assemblage des particules élémentaires) jusqu'aux plus élevés comme les horizons (volumes considérés comme suffisamment homogènes). Cette organisation joue un rôle prépondérant dans les processus d'aération, l'alimentation en eau et le transport de matières modulant ainsi les échanges avec les autres compartiments du globe terrestre, et les conditions de développement des organismes vivants dans les sols (Baize et al. 2013).

1.1.2 Les services éco-systémiques rendus par les sols

Du fait de leur position à l'interface de la lithosphère, de l'hydrosphère, de l'atmosphère et de la biosphère, et de ses processus à l'origine de nombreux échanges avec ces derniers, les sols sont au cœur des grands enjeux du XXIe siècle tels que la sécurité alimentaire, le changement climatique, la préservation de la biodiversité et de la ressource en eau. En effet il assure de nombreuses fonctions écologiques et non écologiques à l'origine de services écosystémiques, qui se définissent comme des bénéfices délivrés par les écosystèmes aux sociétés humaines (Millennium Ecosystem Assessment 2005). En s'inspirant des travaux menés dans le cadre de l'évaluation des écosystèmes pour le millénaire, les services rendus par les sols peuvent être classés en quatre catégories :

- Les services de support ;
- Les services d'approvisionnements ;
- Les services de régulations ;
- Les services culturels.

Les sols constituent le support physique pour les activités humaines et un habitat pour les organismes qu'il héberge constituant ainsi un support pour la biodiversité. Il fournit des nutriments nécessaires à la production de biomasse et régule l'eau quantitativement et qualitativement. Il stock du carbone et influe sur la température du globe en jouant sur la composition de l'atmosphère contribuant ainsi à la régulation du climat (Cousin et al., s. d.).

1.1.3 Les menaces qui pèsent sur les sols et sa biodiversité

Malgré l'importance des sols et la multitude de services qu'ils fournissent aux sociétés humaines, les sols sont sous la menace croissante de l'activité de ces dernières. Dans une communication datant de 2002 [COM(2002)179] la Commission européenne faisait état de la dégradation des sols en Europe et y a présenté les huit principales menaces : l'érosion, le tassement, l'imperméabilisation, la contamination, la salinisation, les inondations et glissements de terrain, la diminution des teneurs en matières organiques et la perte de biodiversité.

Le besoin de protection de la biodiversité tellurique et de l'intégration de la composante biologique dans l'évaluation de la qualité biologique des sols a fortement progressé au cours des dernières décennies (Bonilla-Bedoya et al. 2023). Le sol est l'un des plus grands réservoirs de biodiversité. La biomasse vivante peut représenter de 0 à 15% de la masse totale du carbone organique total du sol et plus de 40% des organismes vivants des écosystèmes terrestres seraient hébergés dans le sol (Calvet 2023). Il inclut de nombreux organismes vivants de nature et de taille très variées. Ils peuvent être classés selon leur taille, ce qui amène à considérer quatre catégories (microorganismes et microfaune <0,1

mm, mésofaune [0,1;2 mm], macrofaune [2;20 mm] et mégafaune >20 mm) ou bien être classés selon trois catégories leur rôle au sein du sol et de leur participation aux fonctions du sol :

- Les chimistes : décomposent et transforment les matières organiques en éléments assimilables par les plantes
- Les régulateurs biologiques : contrôle l'activité des décomposeurs et organismes pathogènes
- Les ingénieurs : entretiennent la structure du sol.

Au vu de l'immensité de cette biodiversité et de son rôle dans le fonctionnement des sols, il est urgent de mieux connaitre cette composante biologique afin de mieux la préserver mais aussi d'en tirer une meilleure valorisation notamment dans un contexte de transition agroécologique nécessitant une gestion plus efficiente des intrants en s'appuyant sur les processus écologiques du sol.

1.2 Les communautés microbiennes du sol :

Au sein de la grande biodiversité des sols, les microorganismes sont parmi les plus abondants et les plus diversifiés au niveau taxonomique et fonctionnel, plusieurs milliards de bactéries pouvant être contenues dans un gramme de sol (Torsvik et Øvreås 2002) et le réseau mycélien formé par les champignons peut atteindre jusqu'à 200 m. Au sein de ce compartiment microbien, les organismes peuvent être séparés en deux grandes catégories : les procaryotes et les eucaryotes :

- Les procaryotes sont des microorganismes ne possédant pas de noyau cellulaire et ni d'autres organites. Ils comprennent les bactéries et les archées.
- Les eucaryotes quant à eux présentent un noyau cellulaire et d'autre organites remplissant une fonction spécifique. On retrouve les champignons, les levures, les algues et les protozoaires.

Bien qu'ils ne représentent qu'une biomasse de 1% à 4% de la biomasse de carbone organique du sol (Calvet 2023), la biomasse microbienne est responsable de nombreuses fonctions dans le sol constituant ainsi un acteur majeur du fonctionnement des sols.

1.2.1 Les méthodes de mesure et de caractérisation des communautés microbiennes

Les microorganismes jouent un rôle clé dans les services de support en étant impliqué dans les différents cycles biogéochimiques C, N, P et S, dans la structuration du sol ou bien dans leur contribution à la production végétale (formation de symbiose et amélioration de la résilience des plantes). Ils sont également impliqués dans les services de régulation du fait de leur contribution à la régulation des pathogènes et des gaz à effet de serre ainsi que dans la dégradation des composés xénobiotiques. Dans un contexte d'adoption de pratiques culturales durables, il est important de comprendre et d'optimiser le rôle des communautés microbiennes du sol si l'on veut prédire voire gérer les processus biologiques sous-jacents. Il est ainsi nécessaire de mieux comprendre l'implication des microorganismes dans les processus biologiques du sol à la base de services écosystémiques et les effets bénéfiques ou délétères de certaines pratiques agricoles sur ces communautés microbiennes et il est alors important de disposer d'outils pour caractériser ces dernières. L'accès aux différentes dimensions de la composante microbiologique des sols en termes de composition, d'abondance, de diversité, de fonctionnalité et d'activité est longtemps resté difficile à cause de la complexité à accéder à ses microorganismes dans la matrice sol marqué par une forte hétérogénéité mais également la difficile compréhension d'une information constituée de nombreuses espèces différentes pour un gramme de sol (Maron et al. 2011). Toutefois, les 30 dernières années, ont montré des avancées considérables dans les outils permettant la caractérisation des communautés microbiennes avec notamment le développement de méthodes biochimiques et de nouvelles techniques en biologie moléculaire permettant une meilleure connaissance de leur diversité. Initialement les outils de microbiologie pasteurienne reposant sur la culture, l'observation et le dénombrement des

microorganismes étaient utilisés. Cependant ces techniques ne permettent pas de prendre en considération l'ensemble des communautés microbiennes étant donné qu'environ 90% de ces dernières ne sont pas cultivables. Face à cette limitation, d'autre techniques reposant sur le dosage de constituants biochimiques spécifiques aux microorganismes et celles basés sur l'extraction et la caractérisation de l'ADN ont été développés permettant ainsi de se passer de la mise en culture et d'étudier le compartiment microbien directement à partir d'échantillons environnementaux.

1.2.1.1 Mesure de l'abondance des communautés microbiennes

La biomasse microbienne correspond à la partie vivante de la matière organique du sol et permet de rendre compte de l'abondance des microorganismes du sol. Elle peut être estimée de différentes manières : biomasse microbienne totale, ou représentée en termes de carbone de la biomasse microbienne (MBC, Cmic), d'azote de la biomasse microbienne (MBN, Nmic), de phosphore de la biomasse microbienne (MBP, Pmic), ou soufre de la biomasse microbienne (Smic). Elle peut également être caractérisée par les composés lipidiques cellulaires (acides gras dérivés des phospholipides - PLFA), les alcools polycycliques cellulaires (ergostérol) et l'ADN ou l'ARN. Les méthodes de mesures comprennent :

- fumigation-extraction (FE),
- acides gras dérivés des phospholipides (PLFA),
- ergostérol,
- approche moléculaire : extraction d'ADN (biomasse microbienne moléculaire), ARN18S, ARN16S.

Chacun de ces indicateurs présente ses propres avantages et limites en termes de représentativité, de coût et de durée d'analyse. L'extraction et la quantification de l'ADN total constituent une méthode robuste pour estimer la biomasse microbienne globale, car l'ADN étant présent dans toutes les cellules vivantes, son abondance reflète directement l'importance de la communauté microbienne du sol. Bien qu'il ait été montré que l'ADN extrait du sol ne représente pas parfaitement l'ensemble de la communauté microbienne, pour plusieurs raisons — présence d'ADN extracellulaire, lyse cellulaire biaisée, pertes liées à la dégradation enzymatique, adsorption sur les colloïdes et pertes lors de la purification — cette approche demeure pertinente (Bakken et Frostegård 2006). Elle présente l'avantage de s'affranchir des contraintes de la mise en culture et permet d'obtenir une estimation plus complète de l'abondance des communautés microbiennes. Dans le cadre de cette étude, l'ADN total est utilisé comme indicateur de la biomasse microbienne en complément de la mesure des activités de cette dernière.

1.2.1.2 Mesure de l'activité des communautés microbiennes

La mesure des activités des communautés microbiennes peut se faire au travers de l'étude des activités enzymatiques et de la respiration.

La respiration microbienne est un processus biochimique qui rend compte de l'activité des microorganismes au travers de la décomposition de la matière organique du sol. L'activité est suivie au travers de la mesure de l'O_2 consommé ou du CO_2 produit afin de déterminer l'intensité de la respiration. Deux types de respiration peuvent être mesurés sur un échantillon de sol, la respiration basale et la respiration induite.

Les enzymes du sol sont fondamentales dans le fonctionnement des sols car elles sont impliquées dans les étapes clés des cycles biogéochimiques (C, N, P, S), contribuant ainsi au cycle des nutriments, mais aussi à la dégradation de certains polluants. Il existe de nombreuses enzymes dont certaines sont impliquées spécifiquement à l'un des cycles biogéochimiques. Les quantités d'enzymes dans le sol varient selon la composition, la teneur en matière organique et de l'activité microbienne. Les enzymes peuvent être constitutives, lorsqu'elles sont systématiquement présentes dans les cellules puis libérées, indépendamment de la présence du substrat, ou bien inductibles lorsqu'elles ne sont pas toujours présentes dans le sol, mais rapidement produites et sécrétées suite à l'ajout d'un substrat. Les enzymes sont classées en fonction de la nature et du type de réaction qu'elles catalysent. On retrouve ainsi six

grands types d'enzymes (oxydoréductases, transférases, hydrolases, lyases, isomérases et ligases). Par leur capacité à réagir rapidement à des changements de l'environnement du sol suite à des changements de pratiques ou d'usages, les enzymes ont été proposés comme indicateurs pertinents de la dynamique de la qualité des sols (Dick et al. 1996).
Les méthodes de mesure des activités enzymatiques peuvent varier en fonction de la nature du substrat utilisé, des conditions opératoires (notamment le pH), des durées d'incubation et des méthodes de détection (colorimétrie, fluorimétrie ou radiomarquage). Cette variabilité méthodologique, ainsi que la grande variabilité naturelle des activités enzymatiques en fonction de la saison ou des conditions édaphiques rendent difficile l'interprétation des mesures et leur utilisation comme indicateur de qualité universel. Des efforts de recherche restent à mener pour développer des référentiels d'interprétation spécifiques à chaque site pour une utilisation plus précise des activités enzymatiques comme paramètre de qualité des sols.
Dans cette étude on s'intéresse plus spécifiquement à la N-acétyl-β-D-glucosaminidase (Nag) et à la β-glucosidase. La Nag, appelée communément chitinase, entre en jeu dans la dégradation et l'hydrolyse des chaines de chitine élément structural fondamental de la paroi cellulaire de nombreux champignons. La β-glucosidase est une enzyme qu'on retrouve dans de nombreux milieux, responsable de la dégradation de la cellulose, notamment en glucose. Elle est ainsi à la base d'une source majeure d'énergie carbonée pour les bactéries. Elle contribue également la stabilisation de la matière organique dans les sols.

Appréhender l'impact des pratiques agricole sur les communautés microbiennes et les fonctions qu'ils remplissent dans le sol afin de piloter les pratiques et d'adopter des pratiques permettant leur préservation nécessite des outils et des références (Chemidlin Prévost-Bouré et al. 2018). Cependant le déploiement de cette évaluation de l'état microbiologique des sols et son suivi régulier ne pourra se faire à grande échelle que si les méthodes de caractérisation deviennent plus rapides, moins laborieuses et abordables.

1.3 La spectroscopie proche infrarouge des sols

La spectroscopie proche infrarouge (SPIR) est une technologie ayant largement été utilisé dans le domaine de l'agroalimentaire (Bertrand et Dufour 2006) et qui semble prometteuse pour la caractérisation des propriétés du sol. La spectroscopie de réflectance diffuse dans le domaine du proche-infrarouge (PIR) et du moyen-infrarouge (MIR) a l'avantage d'être plus rapide et moins couteuse que les analyses de laboratoires conventionnelles. De plus elle ne nécessite pas l'utilisation de produits chimiques nocifs, elle est non destructive, demande peu de préparation et permet d'obtenir une estimation de plusieurs propriétés de sol à partir d'un seul spectre, si des modèles de prédiction ont été construits au préalable (Rossel et al. 2008).
Elle repose sur l'étude de l'interaction entre un rayonnement électromagnétique du domaine du PIR et de la matière, en l'occurrence les échantillons de sols soumis à l'analyse. La région de l'infrarouge est particulièrement intéressante du fait que l'on obtienne des spectres qui reflètent la présence des liaisons chimiques des composés organiques et inorganiques des sols. Plus précisément la SPIR repose sur l'étude de l'absorption des radiations à des fréquences spécifiques correspondants aux vibrations des liaisons des molécules que l'on trouve dans les constituants des sols. Les fréquences moléculaires fondamentales se situent dans la région du MIR, qui s'étend de 2500 à 25 000 nm (4000–250 $cm_{\mathbf{0}}^{1}$). Les harmoniques et bandes de combinaisons se retrouvent à la fois dans le MIR et dans la région du PIR, allant de 700 à 2500 nm. Les fréquences d'absorptions observées dépendent du nombre et de la masse des atomes, ainsi que des forces de liaisons entre eux. Les spectres du sol dans le domaine du PIR sont complexes du fait que de nombreuses fréquences vibratoires sont observées, résultantes de la présence du nombre important de composants retrouvés dans les sols. Il est donc difficile d'attribuer spécifiquement les pics, d'autant plus qu'ils sont très nombreux du fait de la modification des fréquences vibratoires liées à l'environnement moléculaire et aux importantes interactions. Il en résulte ainsi un grand nombre de pics, larges se chevauchants (Janik et al. 1998). Les bandes d'absorption observées

dans la région du proche infrarouge sont dues aux harmoniques des groupements OH, SO_4 et CO_3, ainsi que des combinaisons des bandes fondamentales de H_2O et CO_2 (Genot et al. 2014).

1.3.1 Prétraitements et étalonnages des données spectrales

Les spectres obtenus sont caractérisés par de larges bandes qui se chevauchent et rendent nécessaires l'utilisation de méthodes de prétraitements et de modèles de prédictions (Dwivedi 2017).
Les prétraitements sont des manipulations mathématiques des données spectrales appliquées avant la construction des modèles de prédiction. Elles permettent d'améliorer la qualité des données en réduisant le bruit et les variations non pertinentes. Différentes techniques sont disponibles comme la normalisation, la correction de la ligne de base, le lissage ou les dérivées des spectres.
Du fait de la complexité des spectres obtenus et de leur manque de spécificité, des méthodes d'étalonnages multivariées sont nécessaires pour extraire l'informations qu'ils contiennent. Les méthodes les plus courantes sont basées sur des régressions linéaires. Les régressions linéaires multiples pas à pas (SMLR), les régressions en composantes principales (PCR) et les régressions par les moindres carrés partiels (PLSR) sont les principales techniques utilisées. La PCR et la PLSR sont deux techniques particulièrement utiles dans le cas des données spectroscopiques du fait du grand nombre de variables prédictives et leur forte colinéarité (Dwivedi 2017). La PLSR est plus généralement utilisée car elle relie les variables de réponse et les variables prédictives de manière que les modèles expliquent plus de variance dans la réponse avec moins de composants. Les modèles sont plus interprétables et l'algorithme est plus rapide à calculer (Rossel et al. 2008). D'autres méthodes non linéaires peuvent également être utilisées.
Les performances des modèles prédictifs construits doivent être validés de manière à déterminer les dimensions des modèles et d'appréhender la fiabilité des modèles. La validation croisée est ainsi utilisée pour la sélection du nombre de variables latentes à utiliser dans la construction des modèles. Pour étudier les performances de prédiction des modèles, un ensemble d'échantillons de test représentatifs et indépendant du jeu d'étalonnage et de validation est normalement utilisé. Les performances prédictives sont généralement évaluées au travers du coefficient de détermination (R^2), de l'erreur quadratique moyenne (RMSE) et du RPD (ratio de performance à la déviation) qui est le ratio de l'écart type observé au RMSE

1.3.2 Considérations pour le développement d'étalonnages

Lors du développement des étalonnages la sélection des échantillons servant à la calibration et au test indépendant du modèle est importante. Les échantillons d'étalonnage doivent couvrir la variabilité (spectrale et de la mesure de référence) de l'ensemble de la population sur laquelle on souhaite faire des prédictions. Les échantillons servant au test du modèle doivent être indépendants du jeu qui a servi à étalonner le modèle de manière à éviter une évaluation optimiste du modèle (Soriano-Disla et al. 2014). Lorsque le jeu de test est sélectionné de manière aléatoire dans un ensemble d'échantillons comportant des réplicats, il y a un risque de surestimer les performances prédictives du modèle (Brown et al. 2005). De grandes bibliothèques spectrales de l'ordre de 10^3 à 10^5 échantillons ont été construits afin de tenter de couvrir une grande variabilité de types de sols. Cependant il arrive parfois que de nouveaux échantillons correspondent à des valeurs aberrantes vis-à-vis des spectres utilisés pour l'étalonnage des modèles. Cela se produit lorsque le jeu d'étalonnage est trop restreint ou la bibliothèque spectrale trop petite (Soriano-Disla et al. 2014). Une alternative consiste à construire des modèles à moins grande échelle qui représente une zone homogène en termes de géologie, de texture, de type de sol, d'amplitude des valeurs de références ou de leur combinaisons (Reeves 2010).

Lorsque des bibliothèques spectrales sont combinées à partir d'un ensemble de sols provenant de différents ensembles de données, il faut s'assurer que pour un même attribut d'intérêt les méthodes analytiques utilisées soient identiques, ce qui dans le cas contraire peut fortement diminuer les capacités prédictives de la SPIR (Cécillon et al. 2009).

1.4 Spectroscopie infrarouge des sols et propriétés microbiologiques

La littérature montre que la SPIR permet la prédiction avec précision de paramètres chimiques et physiques tels que les concentrations en carbone et en azote (Barthès et al. 2006; Brunet et al. 2007), la texture, le pH ou la capacité d'échange cationique (Stenberg et al. 2010). Cependant les performances de prédiction des propriétés microbiologiques des sols de cette technique est beaucoup moins renseigné. Pourtant le besoin d'outils rapide et peu couteux pour la caractérisation d'indicateurs microbiologiques reste un enjeu pour le suivi de la qualité des sols et de la cartographie à l'échelle d'un paysage qui demandent de nombreux échantillons.

La biomasse microbienne du sol ne représentant généralement pas plus de 5 % de la matière organique totale du sol, il est peu probable de détecter un signal spectral ou un motif directement associé dans les régions VNIR ou MIR (Soriano-Disla et al. 2014). De plus si des caractéristiques sont détectables, il y a de grandes chances qu'elles soient masquées par d'autres pics au vu de leur faible concentration. Plusieurs études ont néanmoins montré que les propriétés microbiologiques du sol pouvaient être estimés par spectroscopie PIR ou MIR avec une précision variable en fonction des indicateurs considérés relatifs à la biomasse microbienne ou à leur activité. Le succès des prédictions de variables microbiologiques semble être attribué à la covariation entre les paramètres microbiens et les quantités de carbone et d'azote total ou organique (Stenberg et al. 2010).

La plupart des études recensées portent sur la prédiction de la biomasse microbienne en termes de carbone, azote ou phosphore microbien (Soriano-Disla et al. 2014). Les prédictions basées sur la quantification des PLFA ou de l'ergostérol sont retrouvés dans quelques études (Terhoeven-Urselmans et al. 2008; Rinnan et Rinnan 2007; Zornoza et al. 2008) tandis que celles portant sur des méthodes de biologie moléculaire sont quasi inexistantes. En effet, hormis la prédiction du nombre de copies du gène de l'ARNr 16S par spectroscopie MIR dans la publication de Rasche (Rasche et al. 2013) aucune autre étude n'a été trouvé pour la prédiction d'indicateurs microbiologiques basés sur des techniques de biologie moléculaire. En termes d'activités, la respiration microbienne a été étudiée au travers de la respiration induite par substrat ou la respiration basale, ainsi que diverses activités enzymatiques. Les différences dans les méthodes de référence, ajoutées à la grande variabilité en taille et en homogénéité des échantillons de sols, rendent difficile la comparaison des résultats entre les études.

1.5 Objectifs

L'objectif général de nos travaux est d'apporter un nouvel éclairage sur la pertinence de la SPIR dans l'évaluation des propriétés microbiologique du sol. Cet objectif se décline en plusieurs sous-objectifs :

- Identifier les meilleures combinaisons de pré-traitement des échantillons et de choix de méthode d'étalonnage pour optimiser les prédictions par SPIR.
- Identifier l'effet de l'hétérogénéité de la population d'étalonnage sur la précision des prédictions.

2 MATERIELS ET METHODES

2.1 Echantillons de sols étudiés

2.1.1 Population d'échantillons

Des échantillons de surface (0-15 cm de profondeur) issus du projet AlterAgro mené dans l'ancienne région Haute Normandie dans le département de l'Eure (27) ont été étudiés. Les échantillons proviennent de sols de type Luvisols. Ce sont des sols limoneux épais. Les matériaux limoneux, d'origine éolienne reposent sur des matériaux d'altération de type argiles à silex. Ces sols sont non carbonatés et de texture limoneuse, limono-argileuse ou limono-sablo-argileuse. Quinze parcelles

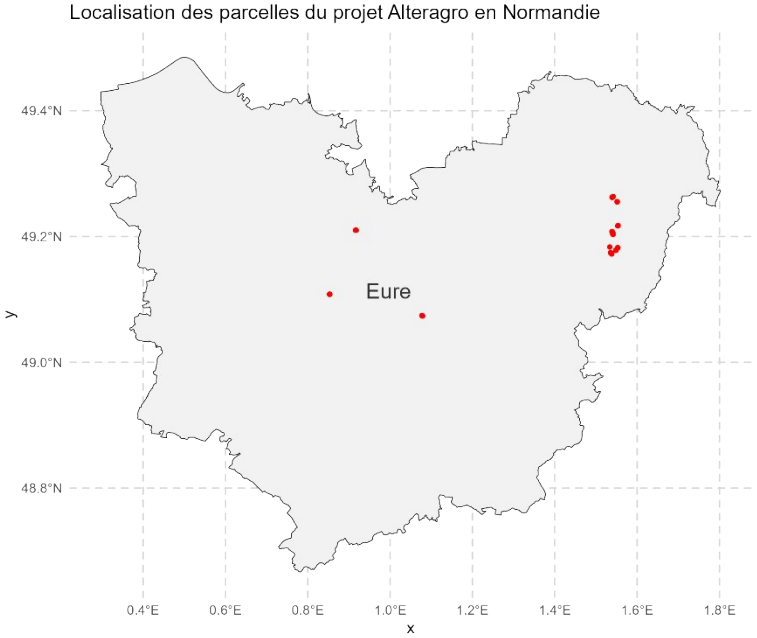
provenant de six communes différentes ont fait l'objet de prélèvement : Amfreville (AC), Combon (CO), Guitry (GU), Richeville (RI), Sacquenville (SA), Tourny (TO). Cinq systèmes de gestion contrastés ont été choisis : un système conventionnel (Conv) avec des apports chimiques réglementaires et un travail du sol occasionnel ; un système intégré avec travail du sol et apports chimiques réduits (Int_2) ; un système intégré avec travail du sol occasionnel et apports chimiques réduits (Int_1) ; un système biologique (Bio) avec travail du sol mais sans apport chimique ; et un système de prairie permanente (PP). Les parcelles conduites en grande culture étaient recouvertes de culture de blé l'année du prélèvement.

Tableau 1. Description des sites de l'étude.

Mode de production	Commune	Date de prélèvement	Parcelle	Nombre d'échantillons	Coordonnées géographiques
Agriculture conventionnelle	Tourny	12/03/2012	TO-GC-A	5	N 49° 10,454' E 1° 32,151'
	Tourny	12/03/2012	TO-GC-B	5	N 49° 10,351' E 1° 32,284'
	Tourny	12/03/2012	TO-GC-C	5	N 49° 10,924' E 1° 33,162'
Intégrée 1	Guitry	12/03/2012	GU-GC-A	5	N 49° 12,460' E 1° 32,333'
	Guitry	12/03/2012	GU-GC-B	5	N 49° 12,355' E 1° 32,435'
	Guitry	12/03/2012	GU-GC-C	5	N 49° 12,207' E 1° 32,484'
Intégrée 2	Richeville	12/03/2012	RI-GC-A	5	N 49° 15,742' E 1° 32,357'
	Richeville	12/03/2012	RI-GC-B	5	N 49° 15,810' E 1° 32,475'
	Richeville	12/03/2012	RI-GC-C	5	N 49° 15,310' E 1° 33,080'
Biologique	Combon	12/03/2012	CO-GC	5	N 49° 06,491' E 0° 51,185'
	Sacquenville	12/03/2012	SA-GC-A	5	N 49° 04,467' E 1° 04,639'
	Sacquenville	12/03/2012	SA-GC-B	5	N 49° 04,469' E 1° 04,654'
Prairie permanente	Guitry	12/03/2012	GU-PP	5	N 49° 13,023' E 1° 33,188'
	Tourny	12/03/2012	TO-PP	5	N 49° 10,683' E 1° 32,878'
	Amfreville	12/03/2012	AC-PP	5	N 49° 12,600' E 0° 55,054'

Les échantillons ont été collectés en mars 2012. Cinq échantillons par parcelle ont été prélevés amenant à soixante-quinze échantillons au total. La localisation et les coordonnées GPS des parcelles sélectionnées sont présentées dans le Tableau 1 et la Figure 1.

Figure 1. Localisation des parcelles.



2.1.2 Prélèvement et préparation des échantillons

Une première stratégie d'échantillonnage a consisté à matérialiser un carré de 30m² au sein des parcelles. A chaque coin du carré et en son centre (soit cinq points d'échantillonnage), cinq échantillons ont été prélevés à la tarière sur 15 cm de profondeur à environ 1m autour du point. Une deuxième stratégie a consisté à réaliser un transect de cinq points d'échantillonnage pour lesquels cinq échantillons ont été prélevés.

Les cinq échantillons ont été mélangés pour obtenir un échantillon moyen à chaque point puis transportés au laboratoire, tamisés à 2 mm et stockés au réfrigérateur (4°C) en attente des analyses microbiologiques et spectrométriques.

2.2 Analyses conventionnelles des propriétés du sol (i.e. méthodes de références)

2.2.1 Propriétés physico-chimiques

La texture du sol a été déterminée à l'aide d'un granulomètre laser Malvern Mastersizer (Malvern Instruments, Malvern, Royaume-Uni). Le carbone organique du sol a été mesuré à l'aide d'un analyseur TOC Shimadzu SSM-5000A/TOC-VCSH Carbone (Shimadzu, Kyoto, Japon). L'azote total du sol a été déterminé selon la méthode Kjeldahl (Kjeldahl 1883). Le pH du sol a été déterminé selon la norme NF ISO 10390. La capacité d'échange cationique (CEC) a été mesurée selon la méthode de Metson (Metson 1957). Le phosphore a été déterminé selon la norme NF ISO 11263 (Dyer NF X 31-160 - Joret-Hébert NF X 31-161).

2.2.2 Propriétés biologiques

2.2.2.1 **Activités enzymatiques**

Deux enzymes ont été mesurées, à savoir la N-acétyl-glucosaminidase (Nag, E.C. 3.2.1.30) et la β-glucosidase (β-glu, E.C. 3.2.1.21). Ces enzymes ont été analysées par spectrophotométrie (Tabatabai et Bremner, 1969). Les substrats utilisés sont listés dans le **Tableau 2** (SigmaeAldrich Co. Ldt, crystalline form).

Pour la mesure des activités de la N-acétyl-glucosaminidase et de la β-glucosidase la procédure est similaire, seuls les substrats diffèrent. Pour chaque échantillon, 1 g de sol frais tamisé à 2 mm a été introduit dans un tube Falcon de 50 ml avec 4 ml de tampon (tris hydroxymethyle aminomethane 0.09 M, acide maléique 0.09 M, acide citrique 0.06 M, and acide borique 0.75 M) et 1 ml de substrat (voir Tableau 2). Chaque Falcon a été vortexé pendant 2 secondes, bouché, puis placé sur un agitateur rotatif (Innova 4430 Incubator Shaker, New Brunswick Scientific, New Jersey, USA) et incubé pendant deux heures à 25 °C.

Après incubation, 1 ml de $CaCl_6$ 0,5 M et 4 ml de NaOH 0,5 M ont été ajoutés à chaque tube échantillon afin d'arrêter la réaction responsable du changement de couleur de l'échantillon. Les tubes ont ensuite été vortexés pendant 2 secondes, puis centrifugés pendant 3 minutes à 9000 rpm (centrifugeuse Eppendorf 5810 R, Hambourg, Allemagne) à température ambiante. Après centrifugation, des aliquotes de 4 ml de surnageant clair ont été prélevées, et l'absorbance a été mesurée à l'aide d'un spectrophotomètre à 410 nm (50 Scan UV-Visible, Varian, Australie). Des échantillons et des contrôles de substrat ont été analysés pendant la période d'incubation afin de vérifier le changement de couleur. Les contrôles ont été réalisés en mélangeant 4 ml de tampon avec soit 1 g de sol, soit 1 ml de solution de substrat.

Tableau 2. Enzymes et substrats pour la détermination des activités enzymatiques par spectrophotométrie.

Enzymes	Substrats
N-acetyl glucosaminidase (Nag)	4-pNP-N-acetyl glucosaminide
β-glucosidase (β-glu)	4-pNP-b-D-glucopyranoside

L'ensemble des résultats des activités enzymatiques sont exprimés en nmol de substrat hydrolysé par minute par gramme de sol sec.

2.2.2.2 **Extraction et dosage de l'ADN total du sol**

L'ADN total du sol est extrait à partir de 500 mg de sol frais selon le protocole du « FastDNA SPIN kit for soil » couplé au système FastPrep-24 bead-beating (Bio 101, Inc., Ca, EU). Cette technique est basée sur le couplage entre une lyse mécanique et une lyse chimique *in situ* des microorganismes. Les acides nucléiques extraits sont ensuite conservés à -20°C.

L'ADN extrait est dosé par fluorimétrie. L'ADN est couplé à un fluorophore spécifique, le Hoechst 33258 (*Fluorescent DNA quantitation Kit* commercialisé par BIORAD), qui se fixe au niveau du petit sillon de l'ADN double brin et dont les longueurs d'ondes d'excitation/émission sont de 355 et 460 nm. Une gamme étalon réalisée par dilution d'une solution d'ADN de concentration connue permet d'estimer la concentration des échantillons. Les résultats sont exprimés en µg d'ADN par g de sol sec.

2.3 Acquisition des spectres

Les spectres proche infrarouge des échantillons de sol ont été acquis à l'aide d'un spectromètre Nicolet iS10 Thermo Scientific, équipé du module « Smart Integrating Sphere Near-IR Nicolet iZ10 ». Les spectres ont été obtenus sur sol tamisé à 2mm séché à l'air pendant 7 jours puis à l'étuve à 40°C pendant 12 heures.

Les spectres de réflectance ont été acquis de 10 000 à 4000 cm^{-1}, avec un intervalle de 8 cm^{-1} sur des échantillons de sol d'environ 25 g placés sur une coupelle en quartz. Pour chaque échantillon l'acquisition de 32 spectres pendant la rotation de la coupelle a été réalisée et l'ensemble de ces spectres sont moyennés par l'appareil. Chaque spectre de réflectance a été converti en absorbance (log [1/reflectance]) puis en longueur d'onde de 1001 à 2500 nm avec un pas de 1 nm. L'analyse a portée sur le domaine PIR 1100-2500 nm. Afin de réduire les données spectrales, les spectres ont été condensés en ne conservant qu'un point spectral sur quatre points adjacents, ce qui donne 351 points par spectre.

2.4 Analyses statistiques

Les différences statistiques entre les modes d'usage du sol (prairie permanente vs grande culture) ainsi qu'entre les systèmes de gestion ont été testées pour les paramètres physico-chimiques et biologiques par des comparaisons de moyennes deux à deux ou des comparaisons multiples en utilisant les tests de Wilcoxon ou Kruskal-Wallis. Des comparaisons multiples ont été réalisées à l'aide de la procédure de Tukey avec un niveau de significativité de $P<0.05$. Les corrélations entre variables ont été testées à l'aide de la procédure de corrélation de rang de Spearman. Les hypothèses de normalité ont été testées à partir du test de Shapiro Wilk. Les calculs ont été réalisé à l'aide du logiciel R.

2.5 Développement de modèles de prédiction

Les données spectrales ont été analysées avec le logiciel R. Différents pré-traitements mathématiques usuels ont été appliqués aux spectres afin d'améliorer le signal et réduire d'éventuels déformations de ligne de base ou l'effet de la taille des particules, ainsi que les effets additifs et multiplicatifs : lissage ou calcul de la première dérivée à l'aide du filtre de Savitzky–Golay appliqué sur un segment de neuf points de mesure (noté SG1 pour première dérivée lissée avec Savitzky–Golay et SG0 pour spectre lissé avec Savitzky–Golay, sans dérivée), combiné éventuellement à un centrage-réduction (SNV, standard normal variate), à un abattement du spectre sur la ligne de base (D, detrend, pouvant être combiné à SNV ou sans autre transformation (None) (Barthès et al. 2010).

Une analyse non supervisée des données spectrales a été réalisée au travers de deux techniques de classification, la classification ascendante hiérarchique (CAH) et une technique de partitionnement grâce à l'algorithme des K-means (ou des centres mobiles), afin de réaliser une exploration préliminaire des données.

Une analyse en composante principale (ACP) a été réalisée sur les spectres d'absorbance prétraités (ou non) afin de calculer dans l'espace des composantes principales la distance de Mahalonis H moyenne entre les échantillons et le centre de la population. Les échantillons avec H > 3 ont été considérés comme déviants (« outliers ») spectraux et éliminés de la suite de l'analyse.
L'étalonnage des spectres sur les variables de référence (carbone organique, activités enzymatiques, adn total a été réalisé par régression modifiée des moindres carrés partiels (PLS pour partial least squares). La PLS est une méthode de régression multivariée qui réduit les données spectrales (composées de plusieurs centaines d'absorbances fortement corrélées) en quelques variables latentes (composantes PLS) orthogonales entres elles, qui sont-elles mêmes des combinaisons linéaires des absorbances et dont la covariance avec la valeur de référence étudiée est maximisée. Le nombre de variables latentes à retenir pour construire un modèle est déterminé par validation croisée. Etant donné que la taille de la population de sol est restreinte, les modèles ont été construits et évalués par validation croisée. Compte tenu de la présence de réplicas au sein de la population de sols, une méthode de validation croisée leave-one-out (LOO) et une autre de validation croisée k-fold ont été réalisées. Pour la validation croisée leave-one-out un échantillon est écarté de la population et un modèle est construit à partir des autres échantillons. L'échantillon mis de côté sert à valider le modèle. Chaque échantillon est écarté à tour de rôle. L'opération est donc opérée autant de fois qu'il y a d'échantillons. Pour la validation croisée k-fold chaque fold correspond à une parcelle différente, autrement dit les échantillons d'une même parcelle se retrouvent dans le même fold afin d'éviter qu'ils servent à la fois à la construction et au test du modèle de façon à éviter des conclusions sur-optimistes des performances du modèle. Le nombre de fold correspond donc au nombre de parcelles de la population de sols. Les résidus de l'ensemble des prédictions ont été regroupés de manière à calculer l'erreur standard de validation croisée (SECV) et le coefficient de détermination (R^2cv). Le nombre de variables latentes retenus pour la construction des modèles correspond à celui où la SECV cesse de diminuer lorsqu'on ajoute une variable latente dans le modèle. La performance des validations croisées a été évaluées au travers de la SECV (que l'on souhaite le plus faible possible), du R^2cv (que l'on souhaite le plus grand possible) et du RPIQ qui est le ratio entre l'écart interquartile de la population et SECV (que l'on souhaite le plus élevé possible : une erreur faible pour une large distribution des données). On considère comme « excellent » un modèle de prédiction quand RPIQ > 4.05, « bon » quand 3,37 < RPIQ < 4,05, « approximatif » quand 2,7 < RPIQ < 3,37 et mauvais quand RPIQ < 2,7. Sur la base des R^2 on considère un modèle comme « excellent » quand R^2 > 0.90, « bon » quand 0.81 < R^2 < 0.90, « approximatif » quand 0.66 < R^2 < 0.80 et « mauvais » quand R^2 < 0.66 Saeys et al. 2005).

3 RESULTATS ET DISCUSSION

3.1 Propriétés chimiques du sol

Les propriétés physico-chimiques des échantillons de sol sont montrées dans les tableaux 3 et 4. La concentration en carbone organique, en azote total, en phosphore ainsi que la capacité d'échange cationique (CEC) sont significativement supérieurs pour les sols de prairies par rapport aux sols de grandes cultures. La concentration en carbone organique varie de 14,64 à 36.44 g.kg^{-1} pour les sols prairies contre une variation de 8.48 à 11.9 g.kg^{-1} pour les sols de grandes cultures tous modes de gestions confondus (Tableau 3). Le pH est significativement plus basique pour les échantillons provenant de parcelles conduites en prairies par rapport à celles en grandes cultures. De plus les échantillons des sols de prairies ont une teneur en sable plus importante.

Tableau 3. Caractéristiques physico-chimiques par occupation du sol.

Occupation du sol	N		C. org (g.kg^{-1})	Azote total (g.kg^{-1})	pH eau	CEC (cmol.kg^{-1})	P (mg.kg^{-1})	Argile (%)	Limon (%)	Sable (%)
GC	60	min-max	8,48 - 11,9	0,66 - 1,34	5,94 - 8,16	6,14 - 14,1	59,77 - 443,59	16,76 - 21,41	68,74 - 70,925	7,7 - 13,46
		moyenne ± ET	10,07 ± 0,82 **a**	0,97 ± 0,14 **a**	7,45 ± 0,52 **a**	10,83 ± 1,54 **a**	265,67 ± 79,96 **a**	19,19 ± 1,51	70,88 ± 1,45	9,93 ± 1,67
Prairie	15	min-max	14,64 - 36,44	0,61 - 3,27	4,65 - 6,97	5,34 - 19	46,36 - 960,82	15,83 - 20,38	60,95 - 66,39	15,32 - 23,22
		moyenne ± ET	27,82 ± 5,76 **b**	2,40 ± 0,74 **b**	6,02 ± 0,83 **b**	13,76 ± 5,16 **b**	367,31 ± 306,92 **b**	18,17 ± 1,93	63,06 ± 2,47	18,77 ± 3,42

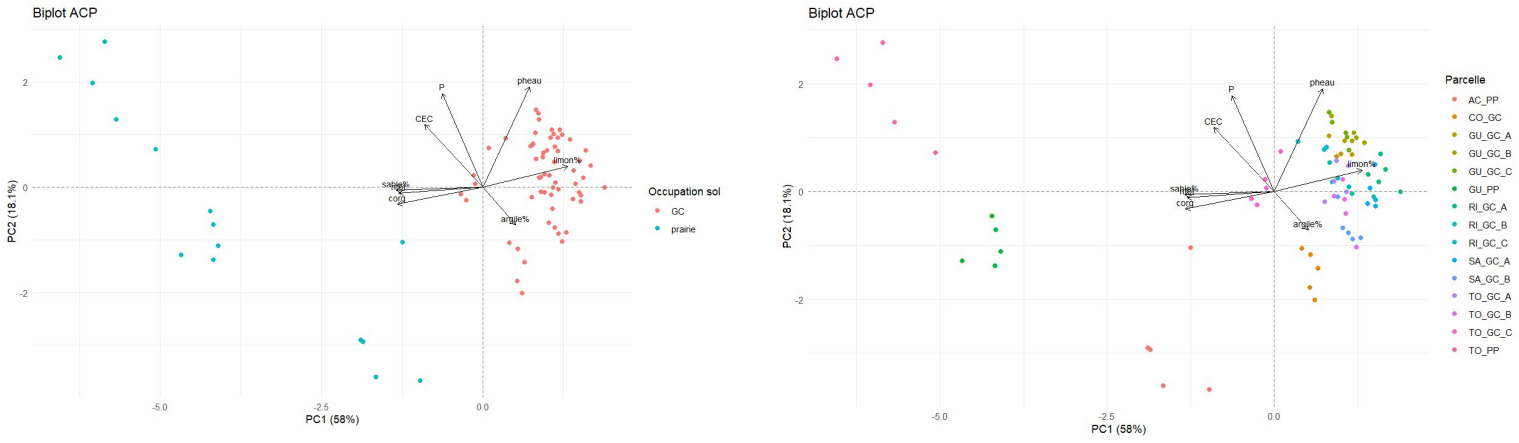
Lorsqu'on prend en considération le système de gestion du sol aucune différence significative n'est observable entre les échantillons provenant de sols de grandes cultures pour la concentration en azote total (Tableau 4). En revanche il existe une différence significative entre certaines modalités de système de gestion du sol pour le carbone organique, le pH, la CEC et la concentration en phosphore. Les échantillons de sols issus de parcelles en agriculture biologique ont une concentration en carbone organique statistiquement inférieure à celle de ceux conduits en intégré 1. De plus les échantillons de sols issus de parcelles en agriculture biologique et conventionnelle présentent des pH significativement plus faibles que ceux des parcelles conduites en intégré 1, eux-mêmes inférieurs à ceux des parcelles conduites en intégré 2. Pour la CEC seuls les échantillons de la modalité intégré 1 sont significativement supérieurs à ceux de la modalité biologique. Pour la concentration en phosphore les échantillons de sols issus de parcelles en agriculture biologique et conventionnelle présentent des concentrations significativement inférieures à celles de parcelles en intégré 1 et 2.

Tableau 4. Caractéristiques physico-chimiques par système de gestion du sol.

Occupation du sol	Système de gestion du sol	N		C. org (g.kg^{-1})		Azote total (g.kg^{-1})		pH eau		CEC (cmol.kg^{-1})		P (mg.kg^{-1})		Argile (%)	Limon (%)	Sable (%)
Prairie	PP	15	min-max	14,64 - 36,44		0,61 - 3,27		4,65 - 6,97		5,34 - 19,00		46,36 - 960,82		15,83 - 20,38	60,95 - 66,39	15,32 - 23,22
			moyenne ± ET	27,82 ± 5,76	c	2,40 ± 0,74	b	6,02 ± 0,83	d	13,76 ± 5,16	ab	367,31 ± 306,92	ab	18,17 ± 1,93	63,06 ± 2,47	18,77 ± 3,42
GC	Int_1	15	min-max	8,90 - 11,22		0,86 - 1,34		7,72 - 8,10		10,20 - 12,80		235,94 - 433,04		16,76 - 18,79	71,62 - 73,46	9,30 - 9,79
			moyenne ± ET	10,25 ± 0,65	b	1,01 ± 0,14	a	7,93 ± 0,12	b	11,51 ± 0,66	b	305,59 ± 53,95	b	18,09 ± 0,97	72,33 ± 0,84	9,58 ± 0,21
	Int_2	15	min-max	8,48 - 11,86		0,71 - 1,17		7,30 - 7,96		6,14 - 12,80		223,54 - 443,59		19,46 - 20,42	69,40 - 72,71	7,83 - 10,87
			moyenne ± ET	10,07 ± 0,77	ab	0,95 ± 0,14	a	7,70 ± 0,20	c	10,27 ± 1,68	ab	320,16 ± 66,57	b	19,87 ± 0,42	70,71 ± 1,49	9,42 ± 1,29
	Conv	15	min-max	8,66 - 11,90		0,73 - 1,04		6,71 - 8,07		9,09 - 14,10		122,71 - 315		17,20 - 21,04	68,74 - 72,05	7,70 - 13,46
			moyenne ± ET	10,40 ± 1,07	ab	1,04 - 0,14	a	7,17 ± 0,36	a	11,66 ± 1,56	ab	237,83 ± 47, 85	a	18,68 ± 1,75	70,68 ± 1,46	10,64 ± 2,44
	Bio	15	min-max	8,54 - 10,14		0,66 - 1,11		5,94 - 8,16		8,07 - 12,3		59,77 - 299,57		17,96 - 21,41	69,24 - 70,60	8,43 - 12,44
			moyenne ± ET	9,56 ± 0,52	a	0,90 ± 0,13	a	7,00 ± 0,58	a	9,89 ± 1,34	a	199,08 ± 82,71	a	20,11 ± 1,59	69,81 ± 0,60	10,07 ± 1,78

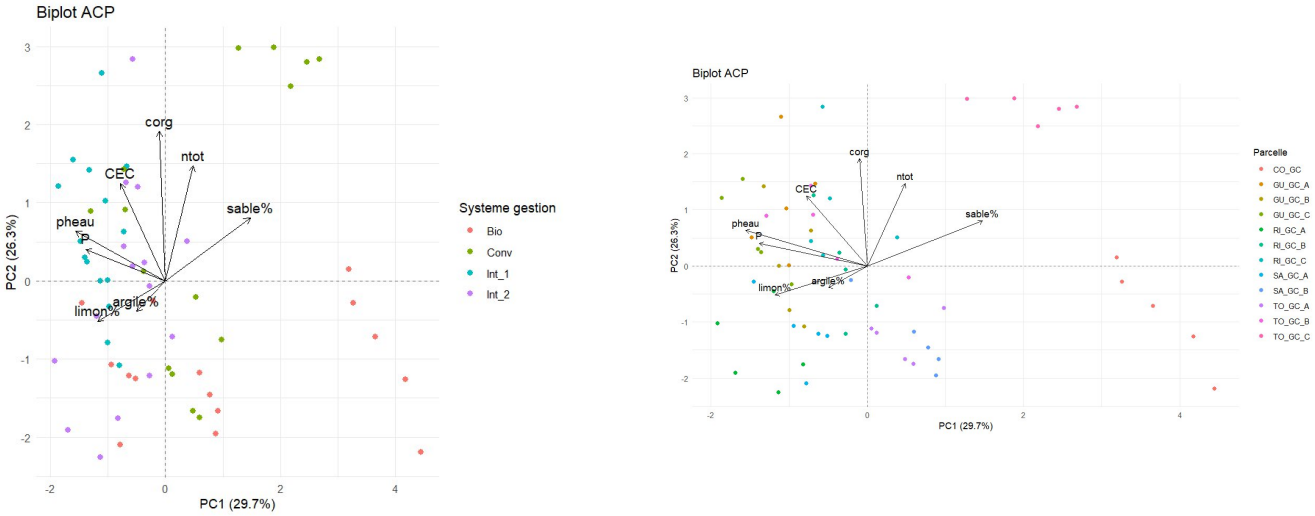
Une analyse en composante principale (ACP) effectuée sur les variables physico-chimiques des échantillons pour l'ensemble des échantillons (prairies et grandes cultures) est montrée figure 2. Les deux premières composantes principales expliquent respectivement 58 % et 18,1 % de la variabilité totale. L'ACP met en évidence une distinction nette entre les échantillons de grandes cultures, positionnés majoritairement au centre et à droite de l'axe 1, et ceux issus de prairies, localisés plutôt à gauche. Les échantillons de prairies se caractérisent par de plus fortes teneurs en carbone organique, en azote total, en sable et de plus faibles pH. L'axe 1 ne permet pas de distinguer clairement les différents modes de gestion au sein des échantillons de grandes cultures. L'axe 2 semble différencier les différents modes de gestion notamment au travers de la concentration en phosphore et du pH. On observe également que les échantillons issus d'une même parcelle se regroupent étroitement, suggérant une homogénéité relative à l'échelle parcellaire.

Figure 2. Analyses en composante principale des variables physico-chimiques pour toutes occupations du sol confondues. Coloration selon l'occupation du sol (à gauche) ou selon la parcelle (à droite).



L’ACP réalisée uniquement sur les échantillons provenant des parcelles menées en grandes cultures est montrée figure 3. Les deux premières composantes principales expliquent respectivement 29,7 % et 26,3 % de la variabilité totale. L’ACP ne révèle pas de séparation marquée entre les différents systèmes de gestion, en raison de la variabilité observée entre les parcelles au sein d’un même système. Les échantillons d’une même parcelle sont moins regroupés que lors de l’ACP précédente et peuvent parfois se mélanger, en raison d’une variabilité notable de certains échantillons au sein de quelques parcelles.

Figure 3. Analyses en composante principale des variables physico-chimiques pour les parcelles de grandes cultures. Coloration selon le système de gestion (à gauche) ou selon la parcelle (à droite).



3.2 Effets du mode de gestion des sols sur les propriétés biologiques du sol

Les propriétés biologiques des échantillons de sol sont montrées dans les tableaux 5 et 6. Les activités de la N-acetylglucosamidase (Nag) et de la β-glucosidase (β-glu) sont significativement supérieures pour les parcelles conduites en prairies par rapport à celles en grandes cultures. Les échantillons provenant de grandes cultures ont des activités pour la Nag qui varient de 0,55 à 2,86 nmol de substrat hydrolysé par minute par gramme de sol sec, tandis qu’elles varient de 3,06 à 8,79 nmol de substrat hydrolysé par minute par gramme de sol sec pour les prairies. En ce qui concerne la β-glu, elles varient de 3,55 à 9,41 nmol de substrat hydrolysé par minute par gramme de sol sec pour les grandes cultures et de 7,2 à 13,42 nmol de substrat hydrolysé par minute par gramme de sol sec pour les prairies.

Tableau 5. Caractéristiques biologiques par occupation du sol.

Occupation du sol		N-acetylglucosaminidase (nmol substrat hydrolysé.min^{-1}.g^{-1} sol sec)		β-glucosidase (nmol substrat hydrolysé.min^{-1}.g^{-1} sol sec)		ADN total (µg.g^{-1} sol sec)	
GC	nb. éch.	48		48		60	
	min-max	0,55 - 2,86		3,55 - 9,41		10,54 - 22,09	
	moyenne ± ET	1,26 ± 0,44	**a**	6,47 ± 1,42	**a**	15,76 ± 2,67	**a**
Prairie	nb. éch.	12		12		15	
	min-max	3,06 - 8,79		7,2 - 13,42		22,54 - 43,6	
	moyenne ± ET	5,34 ± 1,46	**b**	10,56 ± 1,78	**b**	29,99 ± 5,24	**b**

Lorsqu'on prend en considération le système de gestion du sol aucune différence significative n'est observable entre les échantillons provenant de sols de grandes cultures pour l'activité de la Nag. En revanche les activités de la β-glu diffèrent entre les parcelles menées en agriculture biologiques et celles menées en conventionnelle et en intégré 2. En effet les activités sont significativement supérieures pour les modalités agriculture conventionnelle et intégré 2 par rapport à la modalité agriculture biologique. La modalité intégré 1 n'est quant à elle pas statistiquement différente des autres modalités de gestion des échantillons de grandes cultures.
La biomasse microbienne totale obtenue par la quantification de l'ADN total montre une variation allant de 22,54 à 43,6 µg par gramme de sol secs pour les échantillons provenant des sols de prairies et une variation allant de 10,54 à 22.09 µg par gramme de sol secs pour les échantillons provenant des sols de grandes cultures. Les concentrations en ADN total sont statistiquement supérieures pour les sols de prairies. Aucune différence significative n'est observée entre les différentes modalités de gestion pour les échantillons de sols de grandes cultures.

Tableau 6. Caractéristiques biologiques par système de gestion du sol.

Occupation du sol	Système de gestion du sol		N-acetylglucosaminidase (nmol substrat hydrolysé.min^{-1}.g^{-1} sol sec)		β-glucosidase (nmol substrat hydrolysé.min^{-1}.g^{-1} sol sec)		ADN total (µg.g^{-1} sol sec)	
Prairie	PP	nb. éch.	12		12		15	
		min-max	3,06 - 8,79		7,2 - 13,42		22,54 - 43,6	
		moyenne ± ET	5,34 ± 1,46	**a**	10,56 ± 1,78	**c**	29,99 ± 5,24	**a**
	Int_1	nb. éch.	12		12		15	
		min-max	0,57 - 1,91		4,18 - 8,03		11,22 - 22,09	
		moyenne ± ET	1,01 ± 0,44	**b**	6,12 ± 1,24	**ab**	15,98 ± 3,22	**b**
	Int_2	nb. éch.	12		12		15	
		min-max	0,84 - 2,18		6,34 - 9,41		10,54 - 18,27	
		moyenne ± ET	1,24 ± 0,35	**b**	7,41 ± 0,95	**b**	14,40 ± 2,34	**b**
GC	Conv	nb. éch.	12		12		15	
		min-max	0,55 - 1,59		5,46 - 7,97		12,6 - 21,49	
		moyenne ± ET	1,21 ± 0,35	**b**	7,12 ± 0,77	**b**	17,18 ± 2,81	**b**
	Bio	nb. éch.	12		12		15	
		min-max	1,18 - 2,86		3,55 - 7,91		13,61 - 18,99	
		moyenne ± ET	1,59 ± 0,44	**b**	5,21 ± 1,53	**a**	15,48 ± 1,41	**b**

Lorsque l'ensemble des échantillons (prairies et grandes cultures) est pris en compte, des corrélations significatives apparaissent entre le carbone organique et les activités enzymatiques (Nag et β-glu), ainsi qu'avec l'ADN total (Figure 15). En revanche, lorsque les prairies sont exclues de l'analyse, ces corrélations disparaissent, à l'exception de celle reliant le carbone organique à l'activité de la β-glu (Figure 16). Cela suggère que la relation observée entre le carbone organique et les variables microbiologiques est principalement portée par les différences marquées entre prairies et grandes cultures. Une fois cette variabilité inter-occupation des sols retirée, la relation devient moins claire au sein des grandes cultures seules, ce qui est expliqué par une gamme plus restreinte des teneurs en carbone organique et une moindre variabilité des activités microbiennes pour cette couverture de sol.
Les résultats mettent en évidence un effet fort de l'occupation du sol sur les propriétés biologiques. Ces différences peuvent être attribuées à une plus grande quantité de carbone organique en prairies mais aussi possiblement à une moindre perturbation mécanique des sols, favorisant ainsi le développement de communautés microbiennes plus abondantes et actives. Au sein des grandes cultures, les contrastes liés au système de gestion apparaissent plus nuancés L'activité de la Nag ne varie pas entre modalités,

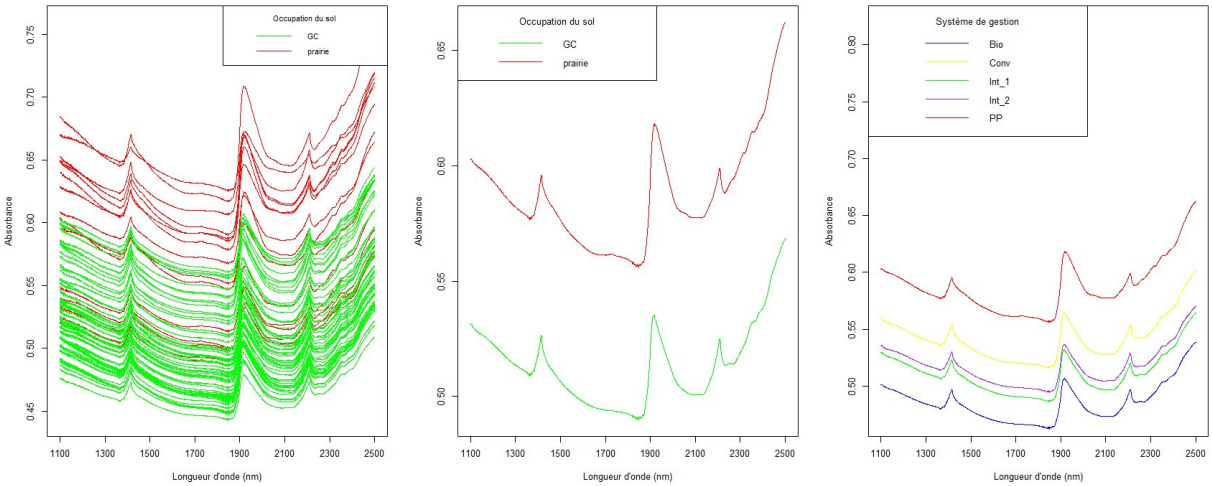
tandis que celle de la β-glu est plus élevée en conventionnel et en intégré 2 qu'en agriculture biologique. Cette observation pourrait refléter des différences dans la disponibilité en substrats carbonés ou dans les pratiques culturales associées (travail du sol, fertilisation, rotations), mais l'absence de différence significative pour la biomasse microbienne totale suggère que ces effets restent limités à certaines fonctions enzymatiques spécifiques.

3.3 Analyse exploratoire des spectres de sols

Les spectres PIR de tous les échantillons ainsi que les spectres moyens en fonction de l'occupation de sol ou du système de gestion sont présentés sur la Figure 4. Les spectres de sol présentent une apparence similaire, la principale différence observée étant un décalage vertical (vers le haut ou vers le bas) des courbes, qui correspond généralement à une variation de l'intensité globale de la réflectance plutôt qu'à une différence spectrale liée à la composition chimique. Les sols de prairies présentent des valeurs d'absorbance plus élevées, traduisant une absorption plus importante de la lumière, vraisemblablement associée à leur teneur plus élevée en matière organique. De même, au sein des sols de grandes cultures, les échantillons des parcelles en agriculture conventionnelle présentent l'absorbance la plus élevée, tandis que ceux des parcelles biologiques montrent la plus faible, en accord avec leur teneur en carbone organique.

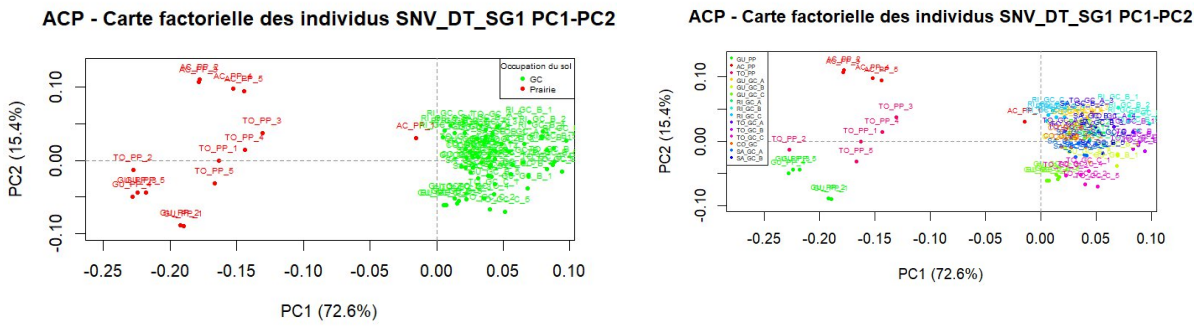
Les spectres PIR résultent des harmoniques et des combinaisons de vibrations fondamentales dans le MIR. Tous les spectres PIR des échantillons de sol montrent des pics d'absorbances élevées autour de 1400, 1900 et 2200 nm. La bande à 1400 nm est souvent associée avec les liaisons O-H, et les liaisons C-H des groupements aliphatiques, tandis que la bande d'absorbance autour de 1900 nm est reliée aux liaisons N-H des amides et O-H. Dans la bande autour de 2200 nm on retrouve les liaisons O-H des groupes phénols, les liaisons N-H des groupes amides et amines ainsi que les liaisons C-H des groupements aliphatiques (Reeves 2010). Des changements dans l'intensité ou la position des bandes d'absorbance des spectres PIR reflète des différences dans la composition du sol. Cependant du fait des nombreux chevauchements entre bandes, l'assignation de certaines bandes à des groupes chimiques particuliers reste difficile rendant nécessaire le recours à des méthodes chimiométriques pour extraire l'information concernant les constituants du sol contenue dans les spectres.

Figure 4. Spectres PIR de l'ensemble des sols (à gauche), spectres moyens selon l'occupation des sols (au centre) et selon le système de gestion (à droite).



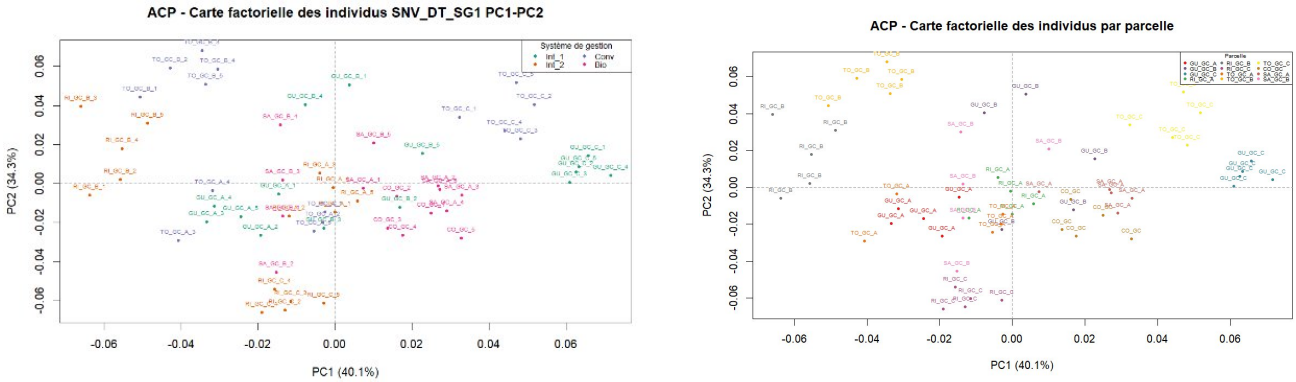
Une analyse en composante principale a d'abord été réalisée sur les spectres PIR de l'ensemble de la population de sol (prairies et grandes cultures) pour explorer la structure des données et d'éventuels clusters. Comme on peut le voir dans la Figure 5, dans l'espace définie par les deux premières composantes principales les échantillons de prairies et ceux de grandes cultures forment clairement deux groupes séparés. Cette séparation nette indique que le type d'usage du sol constitue un facteur majeur de variation dans les spectres PIR. Autrement dit, les sols de prairies et de grandes cultures présentent des signatures spectrales distinctes, reflétant probablement des différences de composition chimique, de matière organique ou de texture. L'ACP montre également que les échantillons provenant d'une même parcelle sont proches les uns des autres sur le plan spectral, ce qui reflète une homogénéité locale des propriétés du sol. En revanche, aucun regroupement clair n'apparaît en fonction du système de gestion, suggérant que ce facteur a une influence moindre sur les variations spectrales que le type d'usage du sol (prairie vs grande culture). Ces observations amènent à penser que la construction de modèles séparés (prairie vs grande culture) serait plus pertinente. En effet, bien qu'un modèle global permette d'exploiter toute la variabilité et d'obtenir un modèle potentiellement plus robuste et plus généralisable, des modèles séparés peuvent être plus spécifiques et donner de meilleures performances sur leur type de sol. Brunet et al. (2007) ont d'ailleurs démontrés que les prédictions PIR des propriétés des sols sont plus précises sur des populations d'échantillons homogènes que sur les populations hétérogènes. A noter toutefois que la construction de modèles séparés réduis la taille du jeu d'apprentissage qui peut être limitant lorsque peu de données sont disponibles et engendre une plus forte homogénéité qui ne doit pas être excessive, car une certaine variabilité doit subsister pour construire un étalonnage (Van Groenigen et al. 2003). Le fait que les répétitions d'une même parcelle soient spectralement proches souligne la dépendance spatiale des échantillons et la nécessité de rester vigilant lors de la validation des modèles. En particulier, il est important d'éviter que des échantillons issus de la même parcelle se retrouvent à la fois dans les jeux d'entraînement et de test, afin de ne pas surestimer la performance des modèles. Une bonne pratique consiste donc à faire une validation croisée groupée par parcelle plutôt qu'une validation croisée leave-one-out (LOO) qui consiste à évaluer la performance d'un modèle en retirant un échantillon à la fois du jeu de données pour servir de jeu de validation, tandis que le modèle est entraîné sur les N−1 échantillons restants. Bien qu'étant adaptée aux petits jeux de données (Cécillon et al. 2008), la validation croisée LOO peut conduire à une surestimation des performances, car certaines répétitions similaires se retrouvent à la fois dans l'entraînement et la validation.

Figure 5. Analyse en composante principale sur les spectres PIR pour toutes occupations du sol confondues. Le pré-traitement SNV+DT+SG1 a été appliqué.



Lorsque l'ACP est effectuée uniquement sur les spectres PIR des échantillons de grandes cultures, aucune structuration des données en fonction du système de gestion des sols n'est apparente (Figure 6). Seule une proximité des échantillons issus d'une même parcelle est observable. Le fait qu'aucune structuration liée au système de gestion ne soit visible indique que, dans les grandes cultures, le type de gestion (labour, non-labour, fertilisation, etc.) n'a pas d'effet majeur sur la variabilité spectrale détectée par le PIR.

Figure 6. Analyse en composante principale sur les spectres PIR pour les parcelles de grandes cultures. Le pré-traitement SNV+DT+SG1 a été appliqué.



3.4 Prédictions par spectrométrie proche infrarouge

Le Tableau 7 présente les meilleures prédictions des différentes variables testées (carbone organique, Nag, β-glu et ADN total) pour l'ensemble des échantillons tandis que le Tableau 8 présente celles des échantillons de grandes cultures uniquement. Les résultats détaillés pour ces variables sont présentés à l'Annexe 1.

3.4.1 Prédictions sur l'ensemble du jeu de données (prairies et grandes cultures)

Le coefficient de détermination de calibration (R^2) correspond à la part de la variance des données mesurées expliquée par le modèle lorsqu'il est appliqué aux échantillons qui ont servi à son ajustement. Il reflète donc la qualité de l'ajustement du modèle aux données d'entraînement. Le coefficient de détermination de validation croisée (R^2cv) est, quant à lui, calculé à partir des prédictions obtenues sur des sous-ensembles de données temporairement exclus de la calibration lors des itérations de la validation croisée. Il mesure ainsi la capacité du modèle à prédire correctement de nouvelles données. En pratique, R^2 est généralement plus élevé que R^2cv , car le modèle est construit pour coller au mieux

aux données de calibration, ce qui peut conduire à un surajustement, tandis que la validation croisée donne une estimation plus réaliste de la performance prédictive. Pour l'ensemble du jeu de données comprenant tous les échantillons, on observe bien une diminution du R^2cv par rapport au R^2 pour l'ensemble des variables testées (Tableau 7).

Le meilleur coefficient de détermination de validation croisée (R^2cv) est obtenu par le carbone organique suivi de l'activité de la Nag, de la quantification de l'ADN total et enfin de l'activité de la β-glu que ce soit pour la validation croisée LOO ou la validation croisée k-fold (Tableau 7). Ce résultat concorde avec ce que l'on retrouve dans littérature où on observe généralement une meilleure performance des prédictions du carbone organique ou de la matière organique par rapport aux variables biologiques (Soriano-Disla et al. 2014). Les R^2cv diminuent lorsqu'on passe de la validation croisée LOO à la validation croisée k-fold pour le carbone organique et l'activité de la β-glu (respectivement, 0.94 vs 0.89 et 0.71 vs 0.52) tandis que pour l'activité de la Nag et la quantification de l'ADN total il reste relativement stable (respectivement, 0.88 vs 0.87 et 0.80 vs 0.81).

Bien que le coefficient de détermination en validation croisée (R^2cv) soit un indicateur central de la capacité prédictive d'un modèle, il ne doit pas être utilisé seul pour juger de sa valeur. En pratique, il est indispensable de le compléter par d'autres métriques de performance, telles que le ratio of performance to deviation (RPD) ou le ratio of performance to interquartile distance (RPIQ). Le RPD, défini comme le rapport entre l'écart-type des valeurs de référence et l'erreur de prédiction (RMSE), permet d'apprécier la précision relative du modèle par rapport à la variabilité des données. Toutefois, cet indicateur présente des limites : il peut être artificiellement élevé en présence de données fortement dispersées ou de structures en clusters, et il repose sur l'hypothèse implicite d'une distribution normale des données, ce qui n'est pas toujours vérifié en pédologie (Bellon-Maurel et al. 2010). Le RPIQ, basé sur l'intervalle interquartile, est plus robuste dans ces situations car il n'est pas influencé par la distribution globale ou par la présence de valeurs extrêmes. Par ailleurs, dans les jeux de données comportant des réplicats d'une même parcelle, la validation croisée de type leave-one-out (LOO) est inadaptée, car elle peut conduire à inclure des échantillons très proches dans le jeu d'entraînement et dans le test, donnant ainsi une estimation trop optimiste des performances (Brown et al. 2005). Ainsi, une évaluation rigoureuse d'un modèle doit s'appuyer conjointement sur plusieurs indicateurs (R^2cv, RMSEcv, RPDcv, RPIQcv,) et utiliser une stratégie de validation croisée adaptée à la structure des données afin de fournir une vision plus fiable et nuancée de ses performances réelles.

Ainsi sur la seule base du R^2cv les modèles de prédictions sont considérés dans le cas de la validation croisée LOO comme excellent pour le carbone organique, bon pour la Nag, approximatif pour la β-glu et l'ADN total. Dans le cas de la validation croisée k-fold ils sont considérés comme bon pour le carbone organique, la Nag et l'ADN total et mauvais pour la β-glu. La validation croisée LOO optimiste ainsi que les R^2cv laissent présager d'apparentes bonnes prédictions alors que pour le carbone organique la RMSEcv relative importante (24 %) et un RPIQ très faible (0,68) indiquent que les prédictions restent peu précises et limitées pour une utilisation quantitative. De même, pour la Nag, le R^2cv (0,87) reste élevé mais la RMSEcv très élevée (47 %) et un RPIQ faible (1,17) soulignent la faible fiabilité des valeurs prédites. Les modèles pour β-glu et l'ADN total présentent des performances un peu meilleures en termes de RMSEcv et de RPIQ (R^2cv de 0,52 et 0,81, RMSEcv 20 % et 17 %, RPIQ 1,54 et 2,08), indiquant une capacité prédictive modérée, mais encore insuffisante pour une estimation précise des valeurs absolues.

Tableau 7. Résumé des résultats des calibrations et validations croisées à partir de spectres PIR pour l'ensemble des échantillons.

Variable	Méthode de validation	Nb étalons	Outliers	Prétraitements	VLs	RMSE	RMSEcv	RMSEcv%	R^2	R^2cv	RPDcv	RPIQcv
Corg	Loo cv	75	0	SNV+SG1	6	1,14	1,81	17,59	0,98	0,94	4,21	0,94
	k-fold cv	75	0	SNV+SG1	3	1,7	2,48	24,1	0,95	0,89	3,06	0,68
NAG	Loo cv	60	0	SNV+SG0	3	0,55	0,62	44,6	0,9	0,88	2,88	1,23
	k-fold cv	60	0	SNV+SG0	3	0,55	0,65	46,76	0,9	0,87	2,76	1,17
β-Glu	Loo cv	60	0	DT+SG0	10	0,63	1,18	16,6	0,92	0,71	1,89	2
	k-fold cv	60	0	DT+SG0	12	0,56	1,53	21,51	0,93	0,52	1,45	1,54
ADN total	Loo cv	75	0	SNV+SG1	3	2,63	2,92	18,02	0,84	0,8	2,27	2,05
	k-fold cv	75	0	SNV+SG1	4	2,6	2,87	17,71	0,84	0,81	2,3	2,08

La représentation graphique des valeurs mesurées versus prédites (Figures 7 à 10) révèle également l'influence de la structure des données : on observe clairement la présence de clusters liés à l'usage du sol (prairies vs grandes cultures) pour les modèles prédisant le carbone organique et la Nag avec de faibles valeurs pour les valeurs provenant des échantillons de grandes cultures et de fortes valeurs pour ceux provenant des prairies. Cette structure en clusters peut biaiser certains indicateurs tels que le RPD, et explique en partie pourquoi certains modèles présentent un R^2cv élevé mais une RMSEcv ou un RPIQ insuffisants. Ces observations soulignent l'importance de combiner plusieurs métriques (R^2, RMSE%, RPD, RPIQ) pour évaluer la qualité des modèles et de rester prudent quant à leur utilisation pour des prédictions quantitatives fines, notamment dans des jeux de données hétérogènes ou structurés en clusters. L'ensemble de ces observations rejoignent les résultats de Ludwig et al. (2017), qui ont obtenus de bonnes performances en validation croisées pour des activités enzymatiques mais des performances insuffisantes lors de la validation avec un jeu indépendant.

Figure 7. Comparaisons entre les prédictions par PIR et les déterminations par mesure conventionnelles pour le Corg pour l'ensemble des échantillons.

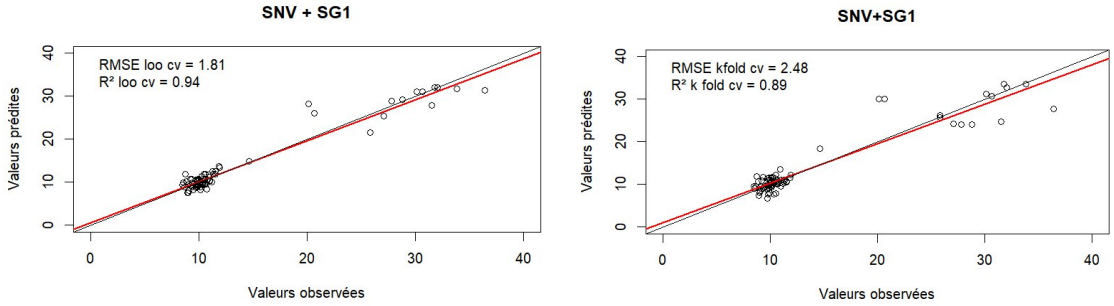


Figure 8. Comparaisons entre les prédictions par PIR et les déterminations par mesure conventionnelles de la Nag pour l'ensemble des échantillons.

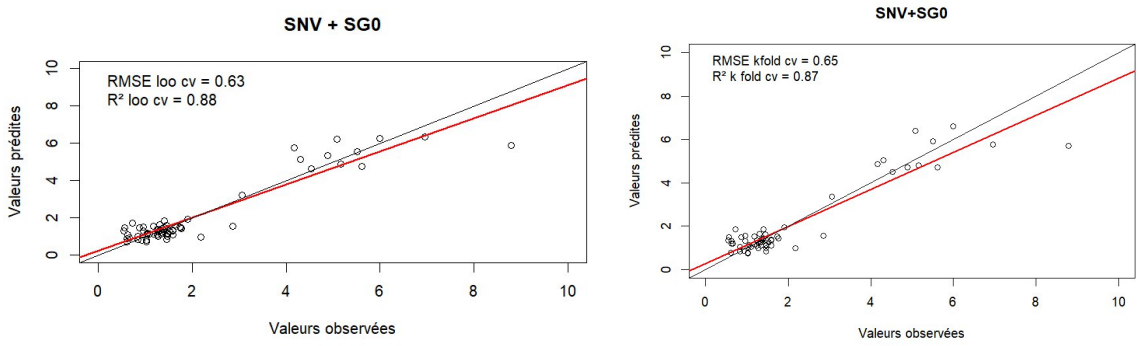


Figure 9. Comparaisons entre les prédictions par PIR et les déterminations par mesure conventionnelles de la β-glu pour l'ensemble des échantillons.

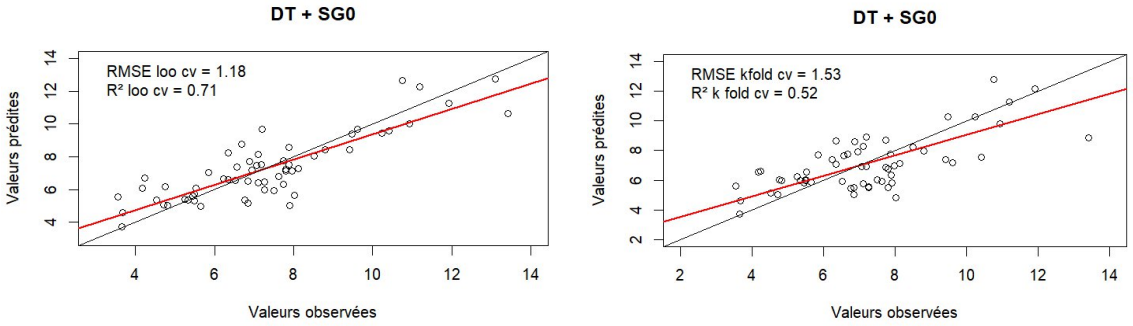
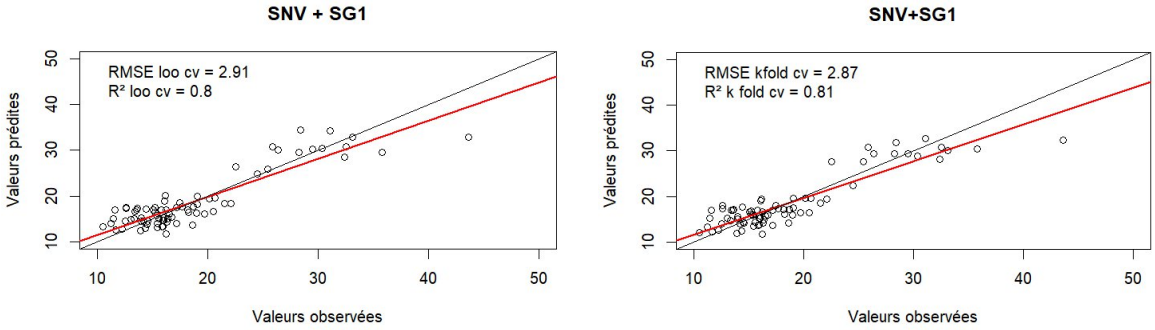


Figure 10. Comparaisons entre les prédictions par PIR et les déterminations par mesure conventionnelles de l'Adn total pour l'ensemble des échantillons.



3.4.2 Prédictions sur la population d'échantillons de grandes cultures

Pour le jeu de données comprenant uniquement les échantillons de grandes cultures, on remarque une diminution des R^2cv par rapport au jeu de données avec l'ensemble des usages de sol (grandes cultures et prairies) (Tableau 8). L'ensemble des modèles sont considérés comme mauvais lors de la validation croisée k-fold avec un R^2cv maximum de 0.5 obtenu pour le carbone organique (Figures 11 à 14). La construction de modèles plus spécifiques à l'occupation de sol à permis de réduire les RMSEcv pour l'ensemble des variables testées, cependant les RPIQ restent tous inférieurs à 2,7 ce qui indique une capacité prédictive insuffisante pour être utilisée.

Tableau 8. Résumé des résultats des calibrations et validations croisées à partir de spectres PIR pour les échantillons de grandes cultures.

Variable	Méthode de validation	Nb étalons	Outliers	Prétraitements	VLs	RMSE	RMSEcv	RMSEcv%	R^2	R^2cv	RPDcv	RPIQcv
Corg	Loo cv	60	0	DT+SG0	12	0,15	0,42	4,2	0,97	0,74	1,98	2,66
	k-fold cv	60	0	DT+SG0	12	0,15	0,58	5,79	0,97	0,5	1,43	1,91
NAG	Loo cv	48	0	SNV+DT+SG0	7	0,25	0,38	30,23	0,66	0,23	1,15	1,33
	k-fold cv	48	0	SNV+DT+SG0	6	0,28	0,42	32,56	0,6	0,07	1,05	1,21
β-Glu	Loo cv	48	0	DT+SG0	5	0,85	1,02	15	0,64	0,47	1,39	2,12
	k-fold cv	48	0	DT+SG0	3	0,99	1,28	18,82	0,5	0,17	1,11	1,69
ADN total	Loo cv	60	0	SNV+SG1	2	2,07	2,24	14,16	0,39	0,28	1,19	1,47
	k-fold cv	60	0	SNV+SG1	1	2,1	2,24	14,16	0,37	0,29	1,19	1,47

La dégradation des performances lorsque l'on ne considère que les échantillons de grandes cultures peut s'expliquer par plusieurs facteurs liés à la taille réduite et à la distribution des données. Tout d'abord, le nombre plus faible d'échantillons limite la capacité du modèle à apprendre correctement la relation entre spectres et propriétés du sol, ce qui augmente le risque de surapprentissage et réduit la robustesse des prédictions. Ensuite, l'homogénéité plus forte des sols au sein des grandes cultures diminue la variance globale, ce qui rend plus difficile pour le modèle de capturer des tendances significatives.

Figure 11. Comparaisons entre les prédictions par PIR et les déterminations par mesure conventionnelles du Corg pour les échantillons de grandes cultures.

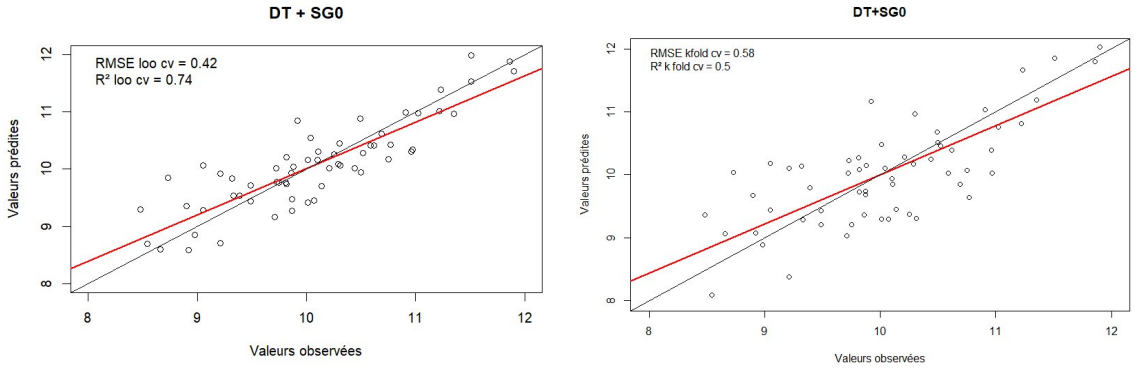


Figure 12. Comparaisons entre les prédictions par PIR et les déterminations par mesure conventionnelles de la Nag pour les échantillons de grandes cultures.

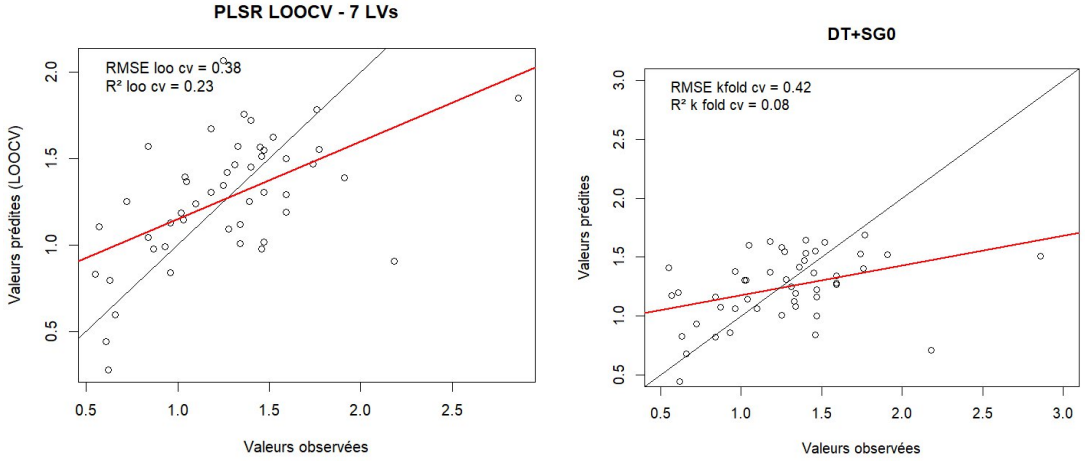


Figure 13. Comparaisons entre les prédictions par PIR et les déterminations par mesure conventionnelles de la β-glu pour les échantillons de grandes cultures.

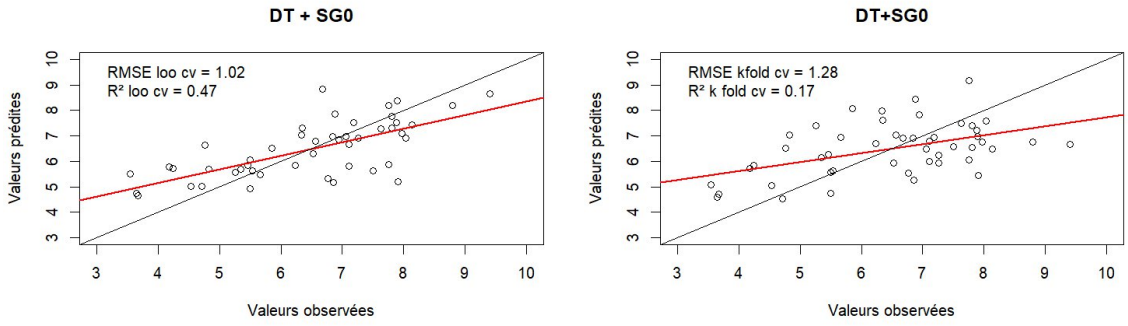
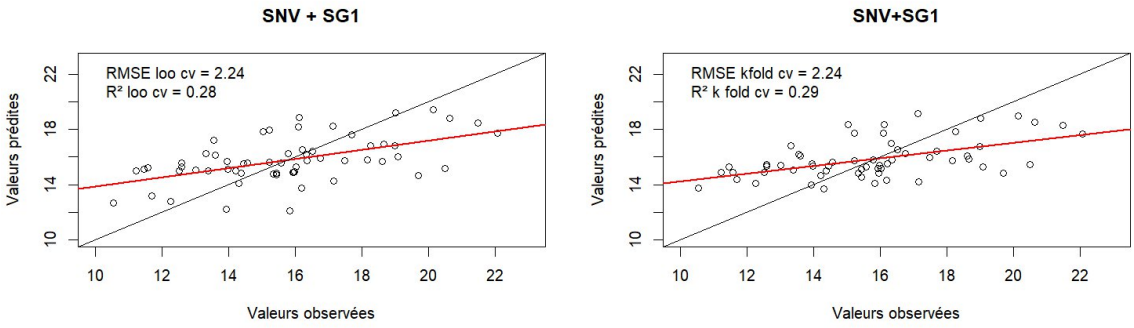


Figure 14. Comparaisons entre les prédictions par PIR et les déterminations par mesure conventionnelles de l'Adn total pour les échantillons de grandes cultures.



Plusieurs études se sont intéressées à la capacité de la spectroscopie infrarouge (PIR ou MIR) à prédire l'activité de différentes activités enzymatiques dans les sols sous différentes couvertures végétales (agricoles, forêts). Les prédictions de la phosphatase (Reeves et al. 2000; Zornoza et al. 2008), phosphatase alcaline (Mimmo et al. 2002), phosphatase acide (Chodak 2011; Mimmo et al. 2002), β-Glucosidase (Dick et al. 2013; Zornoza et al. 2008), de la β-Glucosaminidase (Dick et al. 2013), Uréase (Chodak 2011; Reeves et al. 2000; Zornoza et al. 2008), arylsulfatase (Mimmo et al. 2002; Reeves et al. 2000) et de la déshydrogénase (Chodak 2011; Reeves et al. 2000) ont été testées.

Zornoza et al. (2008) sur des populations assez hétérogènes de sols méditerranéens rapportent d'excellentes performances de prédictions dans le PIR pour la phosphatase acide (R^2cv = 0.90, RPDcv = 3.66) et la β-Glucosidase (R^2cv = 0.93, RPDcv = 3.66), ainsi que de bonnes performances pour l'uréase (R^2cv = 0.8, RPDcv = 2.26). Cependant ces résultats sont à prendre avec précautions du fait que des pseudo-réplicats étaient présents dans le jeu de données étudiées, que les modèles ont été validés par validation croisée optimiste LOO et que le RPD a été utilisé comme indicateur de performance sans toutefois préciser la distribution des variables.

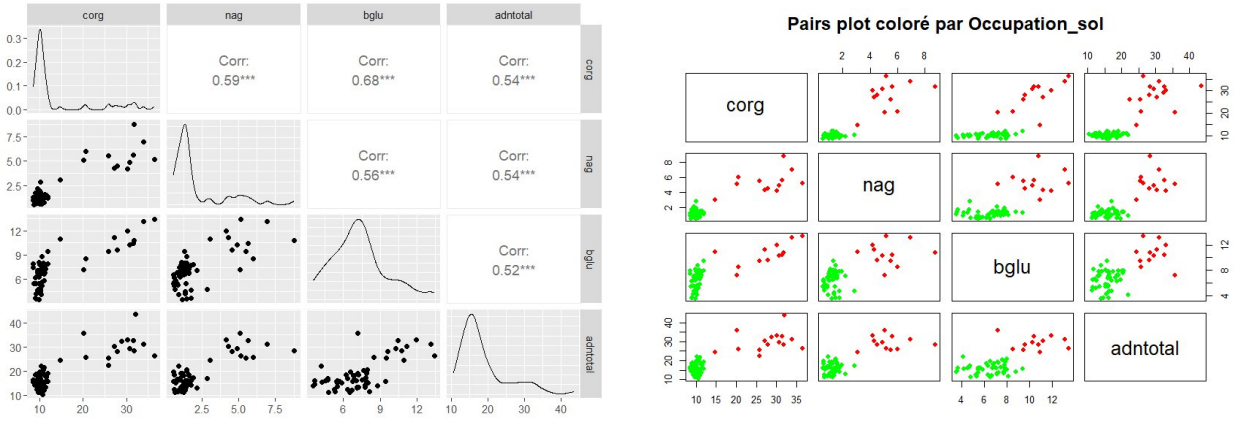
Mimmo et al. (2002) rapportent de moins bons résultats dans le MIR sur une population provenant d'une même parcelle pour l'arylsulfatase (R^2cv = 0.59), la phosphatase acide (R^2cv = 0.59) et même une calibration non significative pour la phosphatase alcaline malgré que la performance de prédiction du carbone total soit bonne pour ce même jeu de données. Les moins bonnes performances seraient attribuées à une diversité en termes de minéralogie du sol malgré que les échantillons soient proches dans l'espace ou aux faibles valeurs des mesures d'activités enzymatiques de référence. Reeves et al. (2000), en utilisant une population de sols agricoles échantillonnés à cinq profondeurs différentes, provenant de deux parcelles expérimentales situées à deux sites différents du Maryland et ayant subis les mêmes traitements (modalités de travail du sol et d'apport d'engrais différentes) sur chaque site ont tentés de prédire diverses activités enzymatiques par SPIR. La phosphatase (R^2cv = 0.77) et l'uréase (R^2cv = 0.67) ont été prédites de manière acceptable tandis que l'arylsulfatase (R^2cv = 0.61) et la déshydrogénase (R^2cv = 0.62) ont eu des résultats plutôt médiocres, malgré que les prédictions du carbone total et de l'azote étaient excellentes. L'auteur conclu qu'une prédiction par SPIR est envisageable seulement si une précision accrue n'est pas nécessaire. Cette conclusion concorde avec les résultats obtenus par Dick et al. (2013), qui sur un ensemble diversifié de sols de l'Ohio, aux États-Unis, ont obtenus des R^2 acceptables pour la β-Glucosidase et la β-Glucosaminidase (R^2cv = 0.82), mais des RMSEP (root mean square error of prediction) élevés, représentant près de la moitié de la valeur moyenne des activités mesurées sur ces échantillons. Ces performances laissent présager la possibilité de suivre des tendances globales, mais pas à donner la valeur exacte pour chaque échantillon.

3.4.3 Déterminants des prédictions par SPIR

Les déterminants des relations entre spectres PIR et variables biologiques demeure discuté. Les enzymes sont des molécules organiques comportant des groupes fonctionnels directement détectables par spectroscopie infrarouge, donc sont susceptibles d'influer directement la réponse spectrale des sols. Cependant certains auteurs comme Rinnan et Rinnan (2007) estiment que les variables

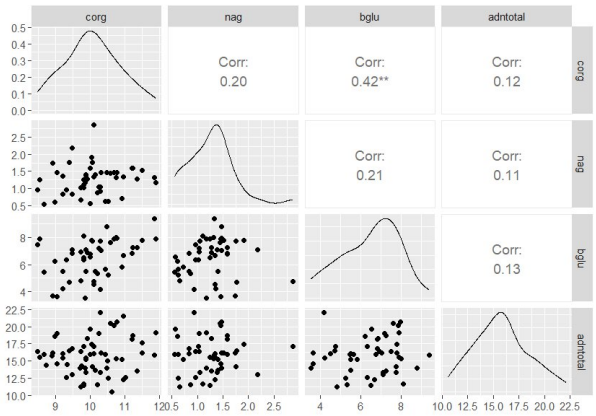
microbiologiques du sol sont présentes en trop faibles concentrations dans la matrice du sol pour qu'on puisse capter un signal direct en spectroscopie. L'obtention de bonnes prédictions serait plutôt dû à de bonnes corrélations avec la matière organique qui elle est bien détectée par spectroscopie. Dans l'étude de Chodak (2011), des prédictions bonnes à approximatives ont été rapportées pour trois enzymes et ces dernières étaient fortement corrélées avec la teneur en carbone organique. De plus les coefficients de corrélation entre les propriétés microbiennes et les longueurs d'onde particulières dans les régions visible et PIR étaient très similaires à ceux du carbone organique.
De manière assez contradictoire, dans l'étude de Zornoza et al. (2008), bien que rapportant de bonnes performances de prédictions des activités enzymatiques, ces derniers n'ont pas observé de fortes similitudes entre les coefficients de régression des propriétés biochimiques et ceux du carbone organique du sol. De plus les longueurs d'onde d'importance, contribuant à la capacité des modèles spectraux à prédire les enzymes diffèrent en partie de celles qui sont importantes dans les modèles prédisant le carbone organique des sols.
Dans notre étude des corrélations significatives ont été obtenues entre les variables biologiques et le carbone organique. Bien que significatives, ces corrélations sont modérées à relativement fortes car elles varient de 0,54 à 0,68 (Figure 15). Cependant ces coefficients de corrélations sont artificiellement bons du fait de données présentant une structure en clusters bien séparés (par exemple prairies vs grandes cultures). Dans ce cas, la différence moyenne entre les deux groupes suffit à générer une relation apparente entre les valeurs mesurées et prédites, même si, à l'intérieur de chaque cluster, le modèle est incapable de capturer la variabilité fine. Autrement dit, le modèle explique surtout la séparation entre groupes plutôt que les différences réelles entre individus au sein d'un groupe. Cela conduit à un R^2 élevé mais trompeur, car il reflète la structure globale des données plutôt qu'une véritable capacité prédictive généralisable.

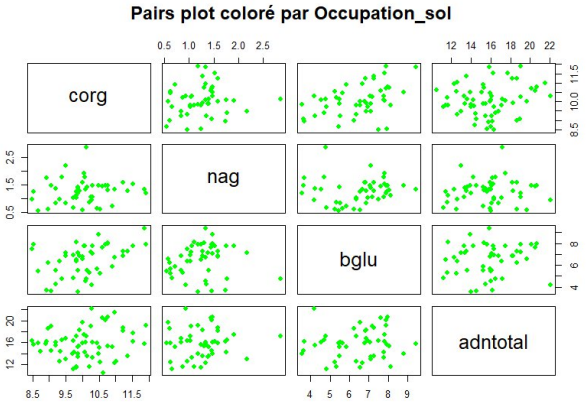
Figure 15. Matrice des nuages de points et des coefficients de Spearman entre les propriétés biologiques du sol et le Corg pour l'ensemble des échantillons.



On observe ainsi une chute marquée des coefficients de corrélation lorsque seuls les échantillons issus de grandes cultures sont pris en compte (Figure 16). Cela confirme que les bonnes corrélations initialement observées étaient en partie liées à la présence de clusters, et donc artificielles.

Figure 16. Matrice des nuages de points et des coefficients de Spearman entre les propriétés biologiques du sol et le Corg pour les échantillons de grandes cultures.





Les performances limitées des modèles obtenues dans notre étude peuvent s'expliquer par plusieurs facteurs. Tout d'abord, les variables microbiologiques ciblées sont des propriétés dynamiques, fortement influencées par des conditions environnementales fluctuantes (température et humidité), et ne présentent pas de relation directe avec les signaux spectroscopiques. Le spectre PIR reflète principalement la composition chimique et structurale de la matière organique et des minéraux du sol, ce qui n'apporte qu'une information indirecte et partielle sur l'abondance ou l'activité microbienne. De plus, les spectres ont été acquis sur des sols séchés et tamisés, alors que les analyses enzymatiques et d'ADN sont réalisées sur des sols frais et humides, dans des conditions particulières de pH et de température pour les enzymes. Ce décalage méthodologique entre l'état des échantillons pour la spectroscopie et celui requis pour les mesures biologiques contribue à expliquer la faiblesse des relations observées. Enfin, la taille réduite et l'hétérogénéité des jeux de données, ainsi que la présence de réplicats proches ou de structures en clusters (prairies vs grandes cultures), limitent la robustesse statistique des calibrations et peuvent générer des performances artificiellement gonflées dans certains cas (validation croisée LOO). L'ensemble de ces éléments explique que, malgré quelques coefficients de détermination élevés (pour les modèles de calibrations et de validation croisée LOO), les indicateurs de précision (RPIQcv) restent insuffisants pour qualifier les modèles de robustes et performants.

4 CONCLUSION

Cette étude a été réalisée sur un ensemble d'échantillons de sols de surface pour différentes occupations de sols et montre que l'utilisation d'une validation croisée LOO plutôt que k-fold dans des modèles de prédictions par SPIR amène à des résultats artificiellement optimistes. De plus la structure des données, avec la présence de clusters liés au type d'usage du sol (prairies vs grandes cultures), contribue artificiellement à certaines performances apparentes, comme en témoigne la chute des corrélations lorsque l'on considère uniquement les échantillons de grandes cultures. Cela est d'autant plus problématique pour des propriétés biologiques dont les bandes d'absorptions spécifiques sont encore mal connues et dont les prédictions sont soupçonnées d'être obtenues par corrélation avec le carbone organique. La construction de modèle plus spécifique en termes d'occupation de sol semble améliorer l'erreur de prédiction, mais le manque d'échantillons et la trop forte homogénéité (entrainant une faible variation des données) n'a pas permis de mettre en évidence une assez bonne capacité des modèles à être utilisée. De nouveaux jeux de données avec une même occupation de sol mais une plus large distribution des valeurs permettrait d'améliorer la performance des modèles.

Enfin, la méthodologie expérimentale — acquisition des spectres sur sols tamisés et séchés, tandis que les analyses biologiques sont réalisées sur sols humides— peut limiter la corrélation directe entre signal spectroscopique et variables microbiologiques et nécessiterai d'être investigué.

5 REFERENCES CITEES

Baize, Denis, Odile Duval, et Guy Richard. 2013. *Les sols et leurs structures Observations à différentes échelles*.

Bakken, Lars R., et Åsa Frostegård. 2006. « Nucleic Acid Extraction from Soil ». In *Nucleic Acids and Proteins in Soil*, édité par Paolo Nannipieri et Kornelia Smalla, vol. 8. Soil Biology. Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-29449-X_3.

Barthès, Bernard G., Didier Brunet, Alain Brauman, et al. 2010. « Determination of Potential Denitrification in a Range of Tropical Topsoils Using near Infrared Reflectance Spectroscopy (NIRS) ». *Applied Soil Ecology* 46 (1): 81-89. https://doi.org/10.1016/j.apsoil.2010.06.009.

Barthès, Bernard G., Didier Brunet, Henri Ferrer, Jean-Luc Chotte, et Christian Feller. 2006. « Determination of Total Carbon and Nitrogen Content in a Range of Tropical Soils Using near Infrared Spectroscopy: Influence of Replication and Sample Grinding and Drying ». *Journal of Near Infrared Spectroscopy* 14 (5): 341-48. https://doi.org/10.1255/jnirs.686.

Bellon-Maurel, Véronique, Elvira Fernandez-Ahumada, Bernard Palagos, Jean-Michel Roger, et Alex McBratney. 2010. « Critical Review of Chemometric Indicators Commonly Used for Assessing the Quality of the Prediction of Soil Attributes by NIR Spectroscopy ». *TrAC Trends in Analytical Chemistry* 29 (9): 1073-81. https://doi.org/10.1016/j.trac.2010.05.006.

Bertrand, Dominique, et Éric Dufour. 2006. *La Spectroscopie Infrarouge et ses Applications Analytiques*.

Bonilla-Bedoya, Santiago, Kevin Valencia, Miguel Ángel Herrera, Magdalena López-Ulloa, David A. Donoso, et José Eduardo Macedo Pezzopane. 2023. « Mapping 50 years of contribution to the development of soil quality biological indicators ». *Ecological Indicators* 148 (avril): 110091. https://doi.org/10.1016/j.ecolind.2023.110091.

Brown, David J., Ross S. Bricklemyer, et Perry R. Miller. 2005. « Validation Requirements for Diffuse Reflectance Soil Characterization Models with a Case Study of VNIR Soil C Prediction in Montana ». *Geoderma* 129 (3-4): 251-67. https://doi.org/10.1016/j.geoderma.2005.01.001.

Brunet, Didier, Bernard G. Barthès, Jean-Luc Chotte, et Christian Feller. 2007. « Determination of Carbon and Nitrogen Contents in Alfisols, Oxisols and Ultisols from Africa and Brazil Using NIRS Analysis: Effects of Sample Grinding and Set Heterogeneity ». *Geoderma* 139 (1-2): 106-17. https://doi.org/10.1016/j.geoderma.2007.01.007.

Calvet, Raoul. 2023. *Le sol - Raoul Calvet - Librairie Eyrolles*. https://www.eyrolles.com/Sciences/Livre/le-sol-9782855578439/.

Cécillon, Lauric, Bernard Barthès, Cécile Gomez, et al. 2009. « Assessment and monitoring of soil quality using near infrared reflectance spectroscopy (NIRS) ». *European Journal of Soil Science* 60 (5): 770-84. https://doi.org/10.1111/j.1365-2389.2009.01178.x.

Cécillon, Lauric, Nathalie Cassagne, Sonia Czarnes, Raphaël Gros, et Jean-Jacques Brun. 2008. « Variable Selection in near Infrared Spectra for the Biological Characterization of Soil and Earthworm Casts ». *Soil Biology and Biochemistry* 40 (7): 1975-79. https://doi.org/10.1016/j.soilbio.2008.03.016.

Chang, Cheng-Wen, David Laird, Maurice Mausbach, et Charles Hurburgh. 2001. « Near-Infrared Reflectance Spectroscopy–Principal Components Regression Analyses of Soil Properties ». *Soil Science Society of America Journal* 65 (mars): 480-90. https://doi.org/10.2136/sssaj2001.652480x.

Chemidlin Prévost-Bouré, N., M. Cannavacciuolo, E. D'Oiron-Verame, et al. 2018. « Appréhender l'impact des pratiques agricoles sur l'état biologique et le fonctionnement du sol. Quelles recommandations et pistes de R en matière de pilotage biologique des sols ? » *Innovations Agronomiques* 69: 39-46. https://doi.org/10.15454/OW3CJS.

Chodak, Marcin. 2011. « Near-infrared Spectroscopy for Rapid Estimation of Microbial Properties in Reclaimed Mine Soils ». *Journal of Plant Nutrition and Soil Science* 174 (5): 702-9. https://doi.org/10.1002/jpln.201000430.

Cousin, Isabelle, Maylis Desrousseaux, Denis Angers, et al. s. d. *Préserver la qualité des sols : vers un référentiel d'indicateurs. Rapport d'étude*. INRAE. https://doi.org/10.17180/QNPX-X742.

Dick, Richard P., Donald P. Breakwell, et Ronald F. Turco. 1996. « Soil Enzyme Activities and Biodiversity Measurements as Integrative Microbiological Indicators ». In *SSSA Special Publications*, édité par John W. Doran et Alice J. Jones. Soil Science Society of America. https://doi.org/10.2136/sssaspecpub49.c15.

Dick, Warren A., Basanthi Thavamani, Shannon Conley, Robert Blaisdell, et Aditi Sengupta. 2013. « Prediction of β-glucosidase and β-glucosaminidase activities, soil organic C, and amino sugar N in a diverse population of soils using near infrared reflectance spectroscopy ». *Soil Biology and Biochemistry*, Special Issue: Interactions of Soil Minerals with Organic Components and Microorganisms VII and Enzymes in the Environment IV, vol. 56 (janvier): 99-104. https://doi.org/10.1016/j.soilbio.2012.04.003.

Dunn, Brian, Graeme Batten, H. Beecher, et S. Ciavarella. 2002. « The potential of near-infrared reflectance spectroscopy for soil analysis — a case study from the Riverine Plain of south-eastern Australia ». *Australian Journal of Experimental Agriculture* 42 (juillet): 607-14. https://doi.org/10.1071/EA01172.

Dwivedi, Ravi Shankar. 2017. « Spectral Reflectance of Soils ». In *Remote Sensing of Soils*, par Dwivedi Ravi Shankar. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-53740-4_6.

Genot, Valérie, Laurent Bock, Pierre Dardenne, et Gilles Colinet. 2014. « L'intérêt de la spectroscopie proche infrarouge en analyse de terre (synthèse bibliographique) ». *Biotechnol. Agron. Soc. Environ.* 18 (2): 247-61.

Gobat, Jean-Michel, Michel Aragno, et Willy Matthey. 2010. *Le sol vivant: bases de pédologie, biologie des sols*. EPFL Press.

Janik, L. J., R.H. Merry, et J. O. Skjemstad. 1998. « Can Mid Infrared Diffuse Reflectance Analysis Replace Soil Extractions? » *Australian Journal of Experimental Agriculture 38* 38: 681-96. https://doi.org/10.1016/S0960-9822(97)70976-X.

Kjeldahl, J. 1883. « New Method for the Determination of Nitrogen in Organic Matter. » *Journal of Analytical Chemistry* 22: 366-82.

Ludwig, Bernard, Svendja Vormstein, Jana Niebuhr, Stefanie Heinze, Bernd Marschner, et Michael Vohland. 2017. « Estimation Accuracies of near Infrared Spectroscopy for General Soil Properties and Enzyme Activities for Two Forest Sites along Three Transects ». *Geoderma* 288 (février): 37-46. https://doi.org/10.1016/j.geoderma.2016.10.022.

Maron, Pierre-Alain, Christophe Mougel, et Lionel Ranjard. 2011. « Soil Microbial Diversity: Methodological Strategy, Spatial Overview and Functional Interest ». *Comptes Rendus. Biologies* 334 (5-6): 403-11. https://doi.org/10.1016/j.crvi.2010.12.003.

Metson, A. J. 1957. « Methods of Chemical Analysis for Soil Survey Samples ». *Agronomy Journal* 49 (4): 208. https://doi.org/10.2134/agronj1957.00021962004900040024x.

Millennium Ecosystem Assessment, éd. 2005. *Ecosystems and Human Well-Being: Synthesis*. The Millennium Ecosystem Assessment Series. Island Press.

Mimmo, Tanja, J. B. Reeves, G. W. McCarty, et G. Galletti. 2002. « DETERMINATION OF BIOLOGICAL MEASURES BY MID-INFRARED DIFFUSE REFLECTANCE SPECTROSCOPY IN SOILS WITHIN A LANDSCAPE »: *Soil Science* 167 (4): 281-87. https://doi.org/10.1097/00010694-200204000-00005.

Rasche, Frank, Sven Marhan, Doreen Berner, Daniel Keil, Ellen Kandeler, et Georg Cadisch. 2013. « midDRIFTS-Based Partial Least Square Regression Analysis Allows Predicting Microbial Biomass, Enzyme Activities and 16S rRNA Gene Abundance in Soils of Temperate Grasslands ». *Soil Biology and Biochemistry* 57 (février): 504-12. https://doi.org/10.1016/j.soilbio.2012.09.030.

Reeves, James B. 2010. « Near- versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done? » *Geoderma*, Diffuse reflectance spectroscopy in soil science and land resource assessment, vol. 158 (1): 3-14. https://doi.org/10.1016/j.geoderma.2009.04.005.

Reeves, J.B., G.W. McCarty, et J.J. Meisinger. 2000. « Near Infrared Reflectance Spectroscopy for the Determination of Biological Activity in Agricultural Soils ». *Journal of Near Infrared Spectroscopy* 8 (3): 161-70. https://doi.org/10.1255/jnirs.275.

Rinnan, Riikka, et Åsmund Rinnan. 2007. « Application of near infrared reflectance (NIR) and fluorescence spectroscopy to analysis of microbiological and chemical properties of arctic soil ». *Soil Biology and Biochemistry* 39 (7): 1664-73. https://doi.org/10.1016/j.soilbio.2007.01.022.

Rossel, R. A. Viscarra, Y. S. Jeon, I. O. A. Odeh, et A. B. McBratney. 2008. « Using a Legacy Soil Sample to Develop a Mid-IR Spectral Library ». *Soil Research* 46 (1): 1. https://doi.org/10.1071/SR07099.

Saeys, W., A.M. Mouazen, et H. Ramon. 2005. « Potential for Onsite and Online Analysis of Pig Manure Using Visible and Near Infrared Reflectance Spectroscopy ». *Biosystems Engineering* 91 (4): 393-402. https://doi.org/10.1016/j.biosystemseng.2005.05.001.

Soriano-Disla, José M., Les J. Janik, Raphael A. Viscarra Rossel, Lynne M. Macdonald, et Michael J. McLaughlin. 2014. « The Performance of Visible, Near-, and Mid-Infrared Reflectance Spectroscopy for Prediction of Soil Physical, Chemical, and Biological Properties ». *Applied Spectroscopy Reviews* 49 (2): 139-86. https://doi.org/10.1080/05704928.2013.811081.

Stenberg, Bo, Raphael A. Viscarra Rossel, Abdul Mounem Mouazen, et Johanna Wetterlind. 2010. « Visible and Near Infrared Spectroscopy in Soil Science ». In *Advances in Agronomy*, vol. 107. Elsevier. https://doi.org/10.1016/S0065-2113(10)07005-7.

Terhoeven-Urselmans, Thomas, Harald Schmidt, Rainer Georg Joergensen, et Bernard Ludwig. 2008. « Usefulness of Near-Infrared Spectroscopy to Determine Biological and Chemical Soil Properties: Importance of Sample Pre-Treatment ». *Soil Biology and Biochemistry* 40 (5): 1178-88. https://doi.org/10.1016/j.soilbio.2007.12.011.

Torsvik, Vigdis, et Lise Øvreås. 2002. « Microbial Diversity and Function in Soil: From Genes to Ecosystems ». *Current Opinion in Microbiology* 5 (3): 240-45. https://doi.org/10.1016/s1369-5274(02)00324-7.

Van Groenigen, J W, C S Mutters, W R Horwath, et C van Kessel. 2003. *NIR and DRIFT-MIR Spectrometry of Soils for Predicting Soil and Crop Parameters in a Flooded Field*.

Zornoza, R., C. Guerrero, J. Mataix-Solera, K.M. Scow, V. Arcenegui, et J. Mataix-Beneyto. 2008. « Near Infrared Spectroscopy for Determination of Various Physical, Chemical and Biochemical Properties in Mediterranean Soils ». *Soil Biology and Biochemistry* 40 (7): 1923-30. https://doi.org/10.1016/j.soilbio.2008.04.003.

6 ANNEXES

6.1 Annexe 1. Résultats détaillés des prédictions

6.1.1 Prédictions du carbone organique

Ensemble de données	Variable	Méthode de validation	Nb étalons	Outliers	Prétraitements	VLs	RMSE	RMSEcv	R^2	R^2cv	RPDcv	RPIQcv
Total	C. org	Loo cv	75	0	None+SG0	8	1,45	2,13	0,963	0,92	3,57	0,79
			75	0	None+SG1	4	1,12	2,33	0,98	0,91	3,27	0,73
			75	0	SNV+SG0	8	1,22	1,69	0,97	0,95	4,52	1
			75	0	SNV+SG1	6	1,14	1,81	0,98	0,94	4,21	0,94
			75	0	SNV+DT+SG0	6	1,42	1,85	0,96	0,94	4,12	0,91
			75	0	SNV+DT+SG1	6	1,19	1,9	0,98	0,94	4	0,89
			75	0	DT+SG0	6	1,55	2,16	0,96	0,92	3,52	0,78
			75	0	DT+SG1	5	1,6	2,25	0,96	0,91	3,39	0,75
		k-fold cv	75	0	None+SG0	7	1,59	4,05	0,96	0,71	1,88	0,42
			75	0	None+SG1	5	1,59	4,26	0,96	0,68	1,79	0,4
			75	0	SNV+SG0	6	1,29	2,92	0,97	0,85	2,6	0,58
			75	0	SNV+SG1	3	1,7	2,48	0,95	0,89	3,06	0,68
			75	0	SNV+DT+SG0	5	1,58	3,02	0,96	0,84	2,52	0,56
			75	0	SNV+DT+SG1	4	1,62	2,68	0,95	0,87	2,84	0,63
			75	0	DT+SG0	4	1,96	3,79	0,93	0,75	2,01	0,45
			75	0	DT+SG1	4	1,78	4,27	0,94	0,68	1,78	0,39
Grandes cultures	C. org	Loo cv	60	0	None+SG0	7	0,33	0,42	0,83	0,74	1,98	2,66
			60	0	None+SG1	7	0,25	0,48	0,91	0,66	1,73	2,32
			60	0	SNV+SG0	13	0,16	0,44	0,96	0,71	1,88	2,53
			60	0	SNV+SG1	5	0,35	0,5	0,81	0,63	1,66	2,22
			60	0	SNV+DT+SG0	12	0,15	0,44	0,96	0,72	1,89	2,54
			60	0	SNV+DT+SG1	6	0,29	0,5	0,87	0,63	1,66	2,23
			60	0	DT+SG0	12	0,15	0,42	0,97	0,74	1,98	2,66
			60	0	DT+SG1	6	0,27	0,46	0,89	0,68	1,79	2,4
		k-fold cv	60	0	None+SG0	7	0,33	0,5	0,83	0,62	1,64	2,2
			60	0	None+SG1	6	0,3	0,64	0,86	0,38	1,28	1,72
			60	0	SNV+SG0	12	0,2	0,64	0,94	0,4	1,3	1,74
			60	0	SNV+SG1	5	0,35	0,73	0,81	0,21	1,14	1,52
			60	0	SNV+DT+SG0	6	0,43	0,65	0,72	0,36	1,26	1,69
			60	0	SNV+DT+SG1	5	0,35	0,74	0,81	0,19	1,12	1,5
			60	0	DT+SG0	12	0,15	0,58	0,97	0,5	1,43	1,91
			60	0	DT+SG1	12	0,09	0,56	0,99	0,53	1,46	1,96

6.1.2 Prédictions de l'activité de la Nag

Ensemble de données	Variable	Méthode de validation	Nb étalons	Outliers	Prétraitements	VLs	RMSE	RMSEcv	R^2	R^2cv	RPDcv	RPIQcv
Total	NAG	Loo cv	60	0	None+SG0	4	0,57	0,7	0,9	0,85	2,58	1,1
			60	0	None+SG1	4	0,49	0,65	0,92	0,87	2,76	1,17
			60	0	SNV+SG0	3	0,55	0,62	0,9	0,88	2,88	1,23
			60	0	SNV+SG1	3	0,57	0,65	0,9	0,87	2,76	1,17
			60	0	SNV+DT+SG0	3	0,58	0,66	0,9	0,86	2,75	1,17
			60	0	SNV+DT+SG1	3	0,57	0,66	0,9	0,86	2,74	1,16
			60	0	DT+SG0	4	0,54	0,68	0,91	0,86	2,66	1,13
			60	0	DT+SG1	4	0,48	0,64	0,93	0,87	2,81	1,19
		k-fold cv	60	0	None+SG0	4	0,57	0,77	0,9	0,81	2,34	0,99
			60	0	None+SG1	4	0,49	0,67	0,92	0,86	2,7	1,15
			60	0	SNV+SG0	3	0,55	0,65	0,9	0,87	2,76	1,17
			60	0	SNV+SG1	3	0,57	0,68	0,9	0,86	2,66	1,13
			60	0	SNV+DT+SG0	2	0,58	0,69	0,89	0,85	2,63	1,12
			60	0	SNV+DT+SG1	3	0,57	0,69	0,9	0,85	2,6	1,1
			60	0	DT+SG0	4	0,54	0,64	0,91	0,87	2,82	1,2
			60	0	DT+SG1	4	0,48	0,7	0,93	0,84	2,58	1,09
Grandes cultures	NAG	Loo cv	48	0	None+SG0	5	0,32	0,38	0,45	0,25	1,16	1,34
			48	0	None+SG1	6	0,24	0,4	0,71	0,16	1,1	1,27
			48	0	SNV+SG0	6	0,28	0,38	0,58	0,24	1,16	1,33
			48	0	SNV+SG1	5	0,24	0,39	0,7	0,21	1,14	1,31
			48	0	SNV+DT+SG0	7	0,25	0,38	0,66	0,23	1,15	1,33
			48	0	SNV+DT+SG1	5	0,24	0,38	0,71	0,23	1,15	1,33
			48	0	DT+SG0	5	0,32	0,4	0,48	0,18	1,11	1,28
			48	0	DT+SG1	6	0,24	0,4	0,71	0,15	1,09	1,26
		k-fold cv	48	0	None+SG0	5	0,32	0,45	0,45	-0,04	0,99	1,14
			48	0	None+SG1	7	0,21	0,46	0,78	-0,09	0,97	1,12
			48	0	SNV+SG0	6	0,28	0,46	0,58	-0,11	0,96	1,11
			48	0	SNV+SG1	6	0,21	0,46	0,77	-0,09	0,97	1,12
			48	0	SNV+DT+SG0	6	0,28	0,42	0,6	0,07	1,05	1,21
			48	0	SNV+DT+SG1	6	0,21	0,44	0,76	-0,03	0,99	1,15
			48	0	DT+SG0	5	0,32	0,42	0,48	0,08	1,05	1,21
			48	0	DT+SG1	7	0,19	0,47	0,8	-0,14	0,95	1,09

6.1.3 Prédictions de l'activité de la β-glucosidase

Ensemble de données	Variable	Méthode de validation	Nb étalons	Outliers	Prétraitements	VLs	RMSE	RMSEcv	R^2	R^2cv	RPDcv	RPIQcv
Total	β-Glu	Loo cv	60	0	None+SG0	6	1,04	1,28	0,78	0,66	1,73	1,83
			60	0	None+SG1	5	0,9	1,29	0,83	0,66	1,72	1,83
			60	0	SNV+SG0	11	0,69	1,25	0,9	0,68	1,77	1,88
			60	0	SNV+SG1	3	1,17	1,29	0,72	0,65	1,72	1,82
			60	0	SNV+DT+SG0	10	0,7	1,24	0,9	0,68	1,78	1,89
			60	0	SNV+DT+SG1	3	1,16	1,29	0,72	0,66	1,72	1,83
			60	0	DT+SG0	10	0,63	1,18	0,92	0,71	1,89	2
			60	0	DT+SG1	6	0,68	1,2	0,9	0,7	1,85	1,97
		k-fold cv	60	0	None+SG0	12	0,56	1,34	0,93	0,63	1,66	1,76
			60	0	None+SG1	8	0,62	1,47	0,92	0,56	1,51	1,6
			60	0	SNV+SG0	11	0,69	1,73	0,9	0,38	1,28	1,36
			60	0	SNV+SG1	3	1,17	1,48	0,72	0,55	1,5	1,59
			60	0	SNV+DT+SG0	12	0,56	1,64	0,93	0,45	1,35	1,44
			60	0	SNV+DT+SG1	4	1,15	1,48	0,73	0,55	1,5	1,59
			60	0	DT+SG0	12	0,56	1,53	0,93	0,52	1,45	1,54
			60	0	DT+SG1	6	0,68	1,54	0,9	0,51	1,44	1,53
Grandes cultures	β-Glu	Loo cv	48	0	None+SG0	4	1,03	1,15	0,46	0,33	1,24	1,88
			48	0	None+SG1	4	0,87	1,1	0,62	0,39	1,29	1,97
			48	0	SNV+SG0	6	0,84	1,12	0,64	0,37	1,27	1,93
			48	0	SNV+SG1	3	1,06	1,27	0,43	0,18	1,12	1,7
			48	0	SNV+DT+SG0	6	0,83	1,08	0,66	0,41	1,31	2
			48	0	SNV+DT+SG1	3	0,99	1,16	0,5	0,33	1,23	1,88
			48	0	DT+SG0	5	0,85	1,02	0,64	0,47	1,39	2,12
			48	0	DT+SG1	3	0,95	1,11	0,54	0,38	1,28	1,95
		k-fold cv	48	0	None+SG0	4	1,03	1,41	0,46	-0,01	1,01	1,53
			48	0	None+SG1	3	0,97	1,28	0,52	0,18	1,11	1,7
			48	0	SNV+SG0	2	0,84	1,52	0,64	-0,16	0,94	1,43
			48	0	SNV+SG1	3	1,06	1,65	0,43	-0,37	0,86	1,31
			48	0	SNV+DT+SG0	6	0,83	1,37	0,66	0,05	1,04	1,58
			48	0	SNV+DT+SG1	3	0,99	1,56	0,5	-0,23	0,91	1,39
			48	0	DT+SG0	3	0,99	1,28	0,5	0,17	1,11	1,69
			48	0	DT+SG1	3	0,95	1,32	0,54	0,12	1,07	0,64

6.1.4 Prédictions de la quantification de l'ADN total

Ensemble de données	Variable	Méthode de validation	Nb étalons	Outliers	Prétraitements	VLs	RMSE	RMSEcv	R^2	R^2cv	RPDcv	RPIQcv
Total	ADN	Loo cv	75	0	None+SG0	6	2,63	3,07	0,84	0,78	2,15	1,94
			75	0	None+SG1	4	2,69	3,12	0,83	0,77	1,12	1,91
			75	0	SNV+SG0	6	2,58	3,1	0,85	0,78	2,13	1,92
			75	0	SNV+SG1	3	2,63	2,92	0,84	0,8	2,27	2,05
			75	0	SNV+DT+SG0	5	2,63	3,07	0,84	0,78	2,15	1,94
			75	0	SNV+DT+SG1	5	2,41	3,08	0,87	0,78	2,15	1,94
			75	0	DT+SG0	5	2,64	3,03	0,84	0,79	2,18	1,97
			75	0	DT+SG1	5	2,43	3,04	0,86	0,79	2,18	1,96
		k-fold cv	75	0	None+SG0	7	2,54	3,02	0,85	0,79	2,19	1,97
			75	0	None+SG1	5	2,49	3,2	0,86	0,76	2,06	1,86
			75	0	SNV+SG0	5	2,68	3,13	0,83	0,77	2,11	1,9
			75	0	SNV+SG1	4	2,6	2,87	0,84	0,81	2,3	2,08
			75	0	SNV+DT+SG0	5	2,63	2,98	0,84	0,79	2,22	2
			75	0	SNV+DT+SG1	5	2,41	2,89	0,87	0,81	2,29	2,07
			75	0	DT+SG0	6	2,47	3,05	0,86	0,78	2,17	1,96
			75	0	DT+SG1	6	2,18	3,06	0,89	0,78	2,16	1,95
Grandes cultures	ADN	Loo cv	60	0	None+SG0	2	2,21	2,33	0,31	0,22	1,14	1,41
			60	0	None+SG1	2	2,13	2,26	0,36	0,27	1,18	1,46
			60	0	SNV+SG0	3	2,09	2,33	0,38	0,22	1,14	1,41
			60	0	SNV+SG1	2	2,07	2,24	0,39	0,28	1,19	1,47
			60	0	SNV+DT+SG0	3	2,1	2,33	0,37	0,23	1,15	1,41
			60	0	SNV+DT+SG1	1	2,1	2,24	0,37	0,29	1,19	1,47
			60	0	DT+SG0	3	2,13	2,31	0,35	0,24	1,15	1,43
			60	0	DT+SG1	2	2,12	2,26	0,36	0,27	1,18	1,46
		k-fold cv	60	0	None+SG0	2	2,21	2,3	0,31	0,25	1,16	1,43
			60	0	None+SG1	2	2,13	2,26	0,36	0,27	1,18	1,45
			60	0	SNV+SG0	3	2,09	2,34	0,38	0,22	1,14	1,4
			60	0	SNV+SG1	1	2,1	2,24	0,37	0,29	1,19	1,47
			60	0	SNV+DT+SG0	3	2,1	2,34	0,37	0,22	1,14	1,41
			60	0	SNV+DT+SG1	2	2,05	2,26	0,4	0,27	1,18	1,46
			60	0	DT+SG0	3	2,13	2,29	0,35	0,25	1,17	1,44
			60	0	DT+SG1	2	2,12	2,28	0,36	0,26	1,17	1,45