Mathématiques appliquées et traitement du signal pour l'évaluation de la dégradation de la biomasse lignocellulosique

Abbas Rammal

UNIVERSITE DE REIMS CHAMPAGNE-ARDENNE

ECOLE DOCTORALE SCIENCES TECHNOLOGIE SANTE (547)

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE DE REIMS CHAMPAGNE-ARDENNE

Discipline : Automatique, signal, productique, robotique

Spécialité : Mathématique appliquées et traitement de signal

présentée et soutenue publiquement par :

Abbas Rammal

Le 25 Janvier 2016

Titre : Mathématiques appliquées et traitement du signal pour l'évaluation de la dégradation de la biomasse lignocellulosique

Jury

M. Douglas Rutledge, Professeur des Universités, AgroParisTech	Rapporteur
M. Jérôme Mars, Professeur des Universités, Grenoble-INP	Rapporteur
M. Guillaume Gellé, Professeur des Universités, URCA	Examinateur
M. Sebastian Miron, Maître de Conférences, Université de Lorraine	Examinateur
M. Valeriu Vrabie, Maître de Conférences HDR, URCA	Directeur de Thèse
M. Eric Perrin, Maître de Conférences, URCA	Co-Directeur de Thèse

Table de matière

Liste de figures

Liste de figures de l'annexe

Liste des tableaux

Liste des tableaux pour l'annexe

Introduction générale

Les gaz à effet de serre (GES) sont des composants qui absorbent le rayonnement infrarouge émis par la surface terrestre. L'augmentation de leur concentration dans l'atmosphère terrestre est l'un des facteurs à l'origine du réchauffement climatique. Les principaux gaz à effet de serre qui existent naturellement dans l'atmosphère sont : la vapeur d'eau, le dioxyde de carbone, le méthane, le protoxyde d'azote, l'ozone. Cet effet de serre est naturel et existe depuis l'origine de notre planète. Cependant, les activités humaines, par leurs rejets, en amplifient la puissance, soit en inoculant des GES artificiels jusqu'alors inconnus dans l'atmosphère, soit en augmentant de façon considérable les concentrations naturelles.
La demande croissante en énergie dépend fortement des combustibles fossiles tels que le charbon, le pétrole brut et le gaz naturel. Les énergies fossiles sont très émettrices de gaz à effet de serre lors de leur combustion. Parmi les combustibles fossiles, le charbon représente le combustible fossile le plus important au niveau des émissions de dioxyde de carbone dans l'atmosphère, avec l'émission d'environ 2,5 tonnes de dioxyde de carbone (CO_2) par tonne de charbon. Ce CO_2, une fois dans l'air, absorbe les rayonnements ultra-violets. C'est un des principaux GES responsables du réchauffement climatique.Face à la raréfaction des ressources fossiles et au défi climatique, la production de carbone renouvelable représente un enjeu considérable. Il est donc nécessaire de développer des procédés de production plus respectueux de l'environnement qui utilisent des ressources renouvelables afin de réduire les émissions de GES.

La biomasse constitue une source alternative d'énergie à travers sa transformation dans des bioraffineries. Bien raisonnée, son utilisation pourrait permettre de réduire les émissions de GES. La biomasse, et en particulier la biomasse végétale, représente la première ressource renouvelable à l'échelle du globe devant l'hydraulique, le solaire et l'éolien [Bou08, UGS12, ZZL12]. Les principales sources de biomasse végétale sont variées : agriculture, forêt, industries agro-alimentaires, industries papetières, transformation du bois, etc. La biomasse végétale constitue une des principales alternatives à l'utilisation du pétrole pour de très nombreuses applications dans les domaines de l'énergie et des bioproduits industriels [ELB14]. D'autre part, l'incorporation de la biomasse végétale dans les sols agricoles est aussi un moyen de maintenir la teneur en matière organique du sol, de renforcer l'activité biologique, d'améliorer les propriétés physiques des sols et d'augmenter la disponibilité des nutriments. Il s'agit d'une problématique environnementale : maintenir la fertilité des sols pour, entre autres, produire des biomasses végétales.

La matière lignocellulosique est le constituant principal de la biomasse végétale. Elle est la source de carbone renouvelable la plus abondante de la planète. La matière lignocellulosique est constituée essentiellement de trois fractions polymériques qui sont la cellulose, l'hémicellulose et la lignine [OD06] et qui influencent les cinétiques de décomposition de la biomasse végétale. Par exemple, les flux de carbone et d'azote vers l'hydrosphère et l'atmosphère et leur cycle dans les sols agricoles sont ainsi étroitement liés à la nature des matières organiques restituées au sol. Il est donc essentiel de pouvoir analyser, à partir de leurs caractéristiques initiales, le comportement des matières lignocellulosiques au cours de la dégradation. Les fractions polymériques de la matière lignocellulosique peuvent être quantifiées par plusieurs méthodes chimiques telles que la méthode de Van Soest ou d'autres méthodes dégradatives qui renseignent plus spécifiquement ces composés pariétaux [VRL91]. Néanmoins il s'agit de méthodes destructives, le plus souvent couteuses en temps d'analyse et parfois critiquables par manque de spécificité des composés ciblés.

Différentes méthodes analytiques de type spectroscopique permettent également de caractériser la biomasse lignocellulosique, notamment les spectroscopies de résonnance magnétique nucléaire, Raman, Infrarouge (IR), UV-Visible, ou de fluorescence. La spectroscopie IR à transformée de Fourier (FTIR) est une technique d'analyse fondée sur le principe d'absorption de l'énergie des rayonnements infrarouges. Cette technique d'identification et de quantification est classiquement utilisée pour la caractérisation de la biomasse végétale en raison de sa rapidité de mise en œuvre [STW05, NCB92]. Elle permet d'identifier des groupements moléculaires et d'obtenir de nombreuses informations sur leur conformation et leurs éventuelles interactions [Koe75, Khe09]. La gamme moyen IR (Mid-IR en anglais, ou MIR) est particulièrement utilisée en raison de sa sensibilité à la fois à des constituants organiques et inorganiques et a trouvé un intérêt croissant pour le développement de bio-marqueurs liés aux caractéristiques intrinsèques de plantes et de leur mode de dégradation. La gamme proche IR (Near-IR en anglais ou NIR) a été largement utilisée pour des analyses quantitatives et qualitatives pour la détermination des caractéristiques telles que les lipides, les protéines, les glucides, etc. Du fait de la rapidité d'analyse et de son caractère non destructif, la FTIR est une méthode de choix pour comparer et classer un grand nombre d'échantillons et également pour évaluer l'évolution de systèmes dans le temps tels que les processus de dégradation et de transformation des biomasses végétales dans différents contextes industriels et environnementaux. Néanmoins, un des enjeux importants est d'établir, en utilisant des informations spectrales, les bases d'un modèle fonctionnel de déstructuration de tissus végétaux dont les applications sont nombreuses : prédiction de la dégradation de végétaux dans des problématiques industrielles et environnementales.

Dans ce contexte, l'objectif de ce travail de thèse est de développer des outils mathématiques et des algorithmes innovants permettant d'identifier des marqueurs spectroscopiques discriminants de résidus lignocellulosiques en fonction de leur niveau de dégradation. En raison de leur complexité et leur nature complémentaire, un défi important est de mettre en place des outils mathématiques qui permettent de combiner les informations spectrales MIR et NIR. Ces outils doivent nous permettre d'extraire des informations plus complètes l'évolution des matrices organiques telles que les lignocelluloses. Le second verrou consiste à prendre en compte les différents stades de la cinétique de biodégradation et pas uniquement au point initial et final (majorité des travaux effectués jusqu'à maintenant). Cette démarche devrait permettre les définir les bases des modèles suffisamment génériques de traitement de spectres en mode cinétique. Notre étude porte sur différentes typologies de lignocelluloses en lien avec les domaines d'application retenus : des racines de maïs, dont les processus de biodégradation dans le sol [MBB11, Amin12] et la biomasse aériennes de miscanthus ou de peuplier, espèces candidates pour la production de bioéthanol par voie biochimique [HDL08, CSJ04].

Ce mémoire de thèse est organisé de la façon suivante :

Le premier chapitre présente des généralités sur la biomasse lignocellulosique, les processus de biodégradations et les techniques permettant de caractériser cette dégradation, dont la spectroscopie IR. Ensuite, nous décrivons les principes de la spectroscopie IR, le fonctionnement d'un spectromètre IR, les techniques d'acquisition de spectres FTIR et les méthodes de préparation des échantillons de biomasse lignocellulosique. Après avoir introduit les spécificités de l'analyse dans les gammes MIR et NIR et leur application pour l'étude de la biodégradation de la biomasse lignocellulosique, nous présentons également les méthodes de combinaison des informations spectrales MIR et NIR.

Le deuxième chapitre est consacré au choix des prétraitements de spectres IR et des gammes spectrales les mieux adaptées pour l'étude de la biomasse lignocellulosique. Pour cela, nous proposons d'utiliser uniquement l'information spectrale MIR et NIR à travers des algorithmes de classification non supervisés qui permettent de réaliser ce choix en optimisant la classification de la biomasse lignocellulosique. Après avoir exposé les bases théoriques des principales méthodes de prétraitement existantes et leur application aux spectres MIR et NIR, nous introduisons le principe de fonctionnement des méthodes de classifications non supervisées, notamment la méthode de classification Fuzzy C-Means (FCM). L'un des paramètres les plus importants étant la métrique entre deux spectres, nous présentons ensuite des extensions FCM basées sur la distance de Mahalanobis telles que les algorithmes de Gustafson-Kessel, Gath-Geva, FCM-M, FCM-CM et FCM-SM. Pour notre application, les spectres pouvant se retrouver regroupés dans des classes de formes non sphériques, nous proposons une nouvelle extension de l'algorithme FCM basée sur une distance Euclidienne pondérée par un facteur de covariance. Une autre spécificité de notre application étant le nombre limité d'échantillons, nous proposons d'utiliser la méthode de ré-échantillonnage « Bootstrap » qui permet d'améliorer la précision de classification. L'application de ces outils mathématiques sur des spectres enregistrés sur les deux typologies de lignocelluloses considérées (maïs, miscanthus) nous a permis de choisir les prétraitements les mieux adaptés et les gammes spectrales les plus pertinentes et de montrer que l'information spectrale est pertinente par rapport à l'information chimique classiquement utilisée dans ce type d'application (concentrations des différents constituants de la biomasse).

Dans le troisième chapitre, nous nous intéressons à développer des méthodes permettant de sélectionner des bandes spectrales IR discriminantes de résidus lignocellulosiques en fonction de leur niveau de dégradation en utilisant uniquement les informations spectrales. Après avoir analysé différentes méthodes de sélection de variables, nous proposons dans un premier temps une méthode basée sur un algorithme génétique en choisissant les paramètres les plus adaptés à notre application. Cette méthode permet de sélectionner des bandes discriminantes dans les deux gammes spectrales considérées, MIR et NIR, mais également des couples des bandes spectrales MIR - NIR qui permettent d'extraire des informations plus complètes et mieux discriminer la biomasse lignocellulosique au cours du processus de dégradation. Ensuite, nous développerons un nouvel algorithme basé sur une approche d'optimisation qui requiert moins des paramètres et qui permet d'extraire également les poids des bandes spectrales. Les résultats obtenus sur des spectres IR de différentes biomasses lignocellulosiques montrent les capacités de ces nouvelles méthodes à sélectionner des bandes spectrales discriminantes par rapport au processus de dégradation.

Le quatrième chapitre est consacré à développer des modèles capables de prédire des activités biologiques reflétant l'état de dégradation de la biomasse à un instant donné. Pour cela, en plus de l'information spectrale, nous considérons également l'information chimique. Après avoir introduit le principe des méthodes de régression et plus particulièrement les méthodes de modélisation telle que la régression des moindres carrés partiels (Partial Least Square en anglais, PLS), y compris des algorithmes permettant d'améliorer la calibration en sélectionnant des bandes spectacles discriminantes comme la projection des variables d'importances (Variable Importance in Projection ; VIP), nous étudions l'association de la PLS avec une méthode basée sur l'algorithme génétique. Cette approche permet d'améliorer la calibration en sélectionnant de bandes discriminantes. Nous présentons ensuite une extension de cette approche en combinant les informations spectrales MIR et NIR. Cette nouvelle approche permet une calibration optimale en s'appuyant sur des couples des bandes spectrales MIR – NIR, ce qui améliore la prédiction de l'état de dégradation de la biomasse lignocellulosique.

Chapitre 1 : Biomasse lignocellulosique et méthodes de caractérisation

I.1. Introduction

La première partie de ce chapitre représente un état de l'art consacré à des généralités sur la biomasse lignocellulosique, les processus de biodégradations et les techniques spectroscopiques permettant de caractériser cette dégradation. Nous décrivons les différentes ressources de biomasses végétales et le rôle de la matière organique des sols. Puis nous décrivons les atouts de la bioconversion des ressources lignocellulosiques pour la production de biocarburant. Nous présentons les facteurs déterminant la décomposition de la matière organique dans le sol ainsi que les prétraitements de la biomasse végétale. Ensuite, nous introduisons l'importance des structures des biomasses lignocellulosiques sur leur processus de dégradation.

La deuxième partie est consacrée à la caractérisation de la biomasse lignocellulosique par chimie humide et par méthodes analytiques. Nous décrivons les techniques spectroscopiques pour caractériser la dégradation des lignocelluloses telles que la spectroscopie Raman, la spectroscopie par Résonnance Magnétique Nucléaire, la spectrophotométrie U.V-Visible, la spectroscopie de fluorescence et la spectroscopie infrarouge.

La partie suivante expose les fondements théoriques de la spectroscopie infrarouge, les caractéristiques des données spectrales IR et les types de spectromètres IR. Nous introduisons les spécificités de l'analyse dans les gammes spectrales moyen infrarouge (MIR) et proche infrarouge (NIR), puis nous présentons quelques applications des techniques de spectroscopie MIR et NIR pour l'étude de la biodégradation de la biomasse lignocellulosique. De par la nature complémentaire de ces deux techniques, la combinaison des spectres MIR et NIR pourrait permettre d'améliorer les analyses. De ce fait, nous nous intéressons aux techniques de combinaison de spectres MIR et NIR et nous présentons quelques applications qui ont utilisé ces méthodes, en particulier celles qui portent sur l'étude de la biomasse lignocellulosique. Ensuite, nous présentons les techniques de spectroscopie à Transformée de Fourier (FT-IR), en particulier le principe et les modes de fonctionnement. Nous discutons des avantages et des inconvénients de la technique FT-IR, puis nous abordons les techniques d'acquisition de spectres FT-IR et plus particulièrement les procédés par transmission et réflexion, comme la réflexion totale atténuée.

La dernière partie de ce chapitre est consacrée aux descriptions des matériels et des données de biomasses lignocellulosiques. Pour cela, nous détaillons les méthodes de préparation des échantillons de biomasse lignocellulosique. Ensuite nous décrivons les différents paramètres d'acquisition des spectres FT-IR et leurs effets.

I.2. Biomasses végétales

I.2.1. Différentes ressources de biomasses végétales

Les résidus végétaux issus de plantes annuelles telles les grandes cultures ou de plantes pérennes tels les taillis à courte rotation contiennent de la lignocellulose, c'est-à-dire, des assemblages complexes de lignine, de cellulose et d'hémicellulose que l'on peut valoriser en carburants dits biocarburants de seconde génération.

La biomasse végétale utilisée pour fabriquer les biocarburants présente l'avantage de pouvoir être produite localement, de manière renouvelable et en utilisant pour sa croissance le dioxyde de carbone

présent dans l'atmosphère. Elle permet donc d'améliorer le bilan des émissions de gaz à effet de serre engendrées par la production et la consommation énergétique.

Les biocarburants ou agro carburants de première génération sont obtenus à partir de cultures oléagineuses (colza, tournesol), de cultures sucrières (betterave, canne à sucre) et de céréales [RSC92]. Leur production nécessite l'utilisation de grandes surfaces agricoles et de matières premières alimentaires. De ce fait, les biocarburants de première génération entrent directement en compétition avec la filière alimentaire et peuvent contribuer à accentuer la crise alimentaire. Afin de pallier ces problèmes, une deuxième voie d'obtention de biocarburants a été imaginée.
Les biocarburants de deuxième génération sont obtenus à partir de sous-produits agricoles et matériaux lignocellulosiques [DB06]. Ils permettent une valorisation de la totalité de la biomasse et conduisent donc à de meilleurs rendements. Leur principal avantage est qu'ils n'entrent pas en compétition avec la filière alimentaire. Le potentiel énergétique théorique de cette voie est énorme mais la disponibilité réelle des matières premières reste limitée et la non prise en compte d'un changement d'affectation des sols indirects (comme la déforestation) pour produire des lignocelluloses dédiés pourrait générer des effets négatifs sur l'émission de gaz à effet de serre. Dans notre travail nous nous intéressons à ce type de biocarburants (deuxième génération).

Le premier usage de la biomasse végétale est l'alimentation qui sert directement ou indirectement à nourrir les hommes. Depuis quelques années, une attention grandissante est portée à la biomasse végétale pour des usages non alimentaires [HFB03]. Dans le secteur de l'énergie, son utilisation est envisagée comme une voie possible de réduction des émissions de gaz à effet de serre et de diminution de la dépendance aux énergies fossiles. Elle apparaît également comme une source de carbone renouvelable dans les domaines de la chimie et des matériaux. Plus généralement, le développement de nouvelles filières de valorisation de la biomasse est considéré comme une opportunité pour stimuler les économies agricoles et rurales [HFB03, DPP10].

La biomasse végétale, majoritairement composée de lignocelluloses, est le premier constituant de la matière végétale. Les ressources lignocellulosiques exploitables sont très variées et proviennent essentiellement des sous-produits de l'agriculture (pailles de céréales, rafles de maïs, tiges de colza, bagasse de cannes à sucre, etc.), des résidus des exploitations forestières (substrats ligneux tels que feuillus et résineux…), des déchets de l'industrie du bois et du papier, mais également de cultures dédiées de plantes annuelles (triticales) ou d'espèces pérennes à rotation rapide (miscanthus, peuplier, eucalyptus, saule…). Ces cultures représentent le potentiel le plus important en biomasse et constituent actuellement un enjeu considérable, compte tenu de leur niveau de production élevé et de leur impact positif sur l'environnement [LBF10]. Les racines (dédiées ou non à la production de biomasse pour la bioénergie, chimie verte…) et au même titre que les feuilles ou autres résidus de culture des biomasses végétales des éléments qui participent à la fertilisation des sols [PVS09].

I.2.2. Rôle de la matière organique des sols

La matière organique des sols est issue des organismes vivants comme les plantes, les animaux et les microorganismes et du recyclage des substrats d'élevages (lisiers) ou de traitements (boues de station d'épuration) etc. On distingue dans le sol :

- La matière organique dite « endogène » est constituée d'une fraction non vivante encore appelée « humus » qui est considérée comme stable soit parce qu'elle forme avec les éléments minéraux des associations qui la protègent de la dégradation microbienne, soit parce qu'elle est, par nature, récalcitrante à la dégradation par les micro-organismes; l'autre fraction,

vivante, est constituée des organismes vivants du sol, et notamment de la biomasse microbienne qui représente 2 à 3 % du carbone total du sol.

- La matière organique dite « exogène » est la matière organique entrant dans les sols dite « fraîche » qui résulte soit des cycles naturels des plantes (résidus sénescents qui tombent à la surface du sol et parties souterraines issues des cultures récoltées), soit des pratiques agricoles (résidus de récolte laissés au champ), soit d'une gestion anthropique des déchets (épandage des effluents et des boues, de composts, fumiers, etc.). Les résidus des grandes cultures (blé, maïs, coton, riz, etc.) constituent cependant l'essentiel des apports exogènes de matière organique. Ces apports exogènes de matière organique subissent sous l'action des organismes du sol des transformations physiques (fragmentation) et biochimiques (minéralisation et humification) dont l'aboutissement est la libération de molécules simples assimilables par la flore microbienne du sol et par les plantes. Le recyclage de la matière organique endogène ou exogène va être alors contrôlé par de multiples facteurs. Ces facteurs peuvent dépendre de la nature des matières organiques (facteurs intrinsèques) et des conditions de l'environnement (facteurs extrinsèques) qui vont agir sur l'activité des organismes décomposeurs. Les principaux facteurs extrinsèques déterminant le recyclage des matières organiques dans les sols sont l'humidité, la température, le pH du sol, les conditions d'aération, la disponibilité en nutriments. Les caractéristiques intrinsèques de la matière organique exogène sont un des principaux facteurs qui influencent largement leur décomposition dans les sols.

I.2.3. Bioconversion des ressources lignocellulosiques pour la production de biocarburant

L'utilisation de la biomasse lignocellulosique pour la production d'éthanol-carburant présenterait de multiples avantages du point de vue environnemental (bilan en émissions de CO_2 plus favorable que l'éthanol issu des plantes sucrières ou amylacées, valorisation des co-produits et déchets) et socio-économiques (pas de compétition avec les surfaces agricoles à usages alimentaire ou agro-alimentaire, moindre coût de la matière première).

Selon la directive européenne n°2003/30/CE du 08 mai 2003, la définition des biocarburants est la suivante : « combustibles liquides ou gazeux utilisés pour le transport et produits à partir de la biomasse ». Il existe actuellement deux filières industrielles de production de biocarburants à partir des éléments de réserve des plantes [Bal07, NGR10] et dits de première génération (1G) :

- Le biodiesel constitué d'esters méthyliques d'huiles végétales (colza, tournesol, soja). Ce carburant est utilisé en mélange dans le gazole.
- L'éthanol, issu de la fermentation des sucres contenus dans les plantes sucrières (betterave, canne à sucre…) et amylacées (maïs, blé...). Il peut être directement incorporé à l'essence ou être transformé en Ethyl t-Butyl Ether (ETBE) avant incorporation.

Ces biocarburants de première génération assurent de substantielles réductions des émissions de CO_2 mais présentent l'inconvénient majeur de ne concerner qu'une partie de la plante, les éléments de réserve, et reste limitée par la compétition des matières premières avec les usages alimentaires et la disponibilité des surfaces de culture [Gab08].

Les biocarburants de deuxième génération (2G) sont quant à eux issus de la transformation de la plante entière ou de sa fraction lignocellulosique (non comestible). Les biocarburants 2G peuvent être produits selon deux voies principales : la voie biologique et la voie thermochimique. Dans la voie biologique, le polysaccharide de structure des lignocelluloses, la cellulose, est dégradée en glucose qui peut être transformé par des bactéries ou des levures en alcool et donc en biocarburant (éthanol).

I.2.4. Points de convergence et divergence entre processus de dégradation des biomasses végétales en milieu naturel et en milieu contrôlé

a) Dans le sol

L'incorporation de résidus végétaux dans les sols agricoles est principalement un moyen de maintenir la teneur en matière organique du sol, de renforcer l'activité biologique, d'améliorer les propriétés physiques des sols et d'augmenter la disponibilité des nutriments. Le taux de décomposition des matières organiques endogènes et exogènes dans les sols est lié à de nombreux facteurs, en fonction de la nature de la matière organique et des conditions environnementales qui influencent l'activité des microorganismes du sol impliqués dans le processus de décomposition. Les principaux facteurs déterminant la décomposition de la matière organique dans le sol sont :

- Le pH du sol : Le pH du sol influence les processus de décomposition des résidus grâce à son effet sur l'activité microbiologique et affecte aussi la disponibilité des nutriments. Par exemple il est connu que les champignons tolèrent davantage les milieux acides que les bactéries et sont donc plus abondants dans des sols acides, tels ceux de certaines forêts. Le pH a une influence sur trois composantes importantes de la fertilité d'un sol : la biodisponibilité des nutriments et éléments toxiques, l'activité biologique et la stabilité structurale [DFA13].
- La température et l'humidité du sol : La température et l'humidité du sol sont des facteurs importants qui influent sur le taux de décomposition des résidus car ils touchent directement l'activité microbienne. Une humidité adéquate permettra d'accélérer la vitesse de décomposition et la croissance de micro-organismes. L'influence de la température sur les activités microbiennes du sol a été soulignée par de nombreux chercheurs et, en particulier, par Dommergues (1962). Ces facteurs, par leurs actions sur l'intensité des processus minéralisateurs, déterminent la rapidité des cycles des éléments nutritifs, mis à la disposition des végétaux et la fertilité [Mou67].
- Les microorganismes du sol : Les micro-organismes sont les décomposeurs de matière organique dans le sol. Par conséquent, la diversité et l'activité de la communauté microbienne au cours de la décomposition des résidus végétaux a reçu beaucoup d'attention. Les micro-organismes forment un groupe très hétérogène, avec environ 6000 génomes bactériens différents de la taille du génome d'Escherichia coli comme unité dans 1 g de sol. Parmi d'autres facteurs, le temps de reproduction et la croissance rapide des microorganismes du sol pourraient expliquer cette grande diversité [SP72].
- La nature des résidus végétaux : La structure chimique des résidus influence la vitesse et le taux de minéralisation dans le sol. La biomasse aérienne est généralement moins récalcitrante à la décomposition que les racines. La nature récalcitrante des racines et des feuilles sénescentes limite leur décomposition ce qui permet une accumulation potentielle de C organique favorable au maintien de la fertilité des sols à long terme [Rob87, HB99].

b) Dans des milieux contrôlés

Le développement de bioprocédés basés sur la transformation de la biomasse végétale en alternative au carbone fossile est limité par deux verrous majeurs : la récalcitrance des lignocelluloses, qui augmente le coût des traitements, et la diversité des biomasses disponibles. La récalcitrance de la biomasse végétale est principalement due à la structure cristalline de la cellulose et à la présence de lignine, un polymère polyphénolique qui limite l'accès des enzymes cellulosiques à leur substrat [XLD12].

En particulier, une étape de prétraitement de la biomasse végétale est nécessaire pour améliorer l'hydrolyse enzymatique. Elle a pour but de rendre la cellulose plus accessible aux enzymes cellulolytiques qui sont capables de générer du glucose fermentescible en éthanol. Les prétraitements

permettent de rompre les liaisons entre polymères et éliminer tout ou partie des polymères pouvant inhiber l'action des cellulases (lignine, hémicellulose). La Figure 1.1 schématise l'impact des prétraitements sur les complexes lignocellulosiques.

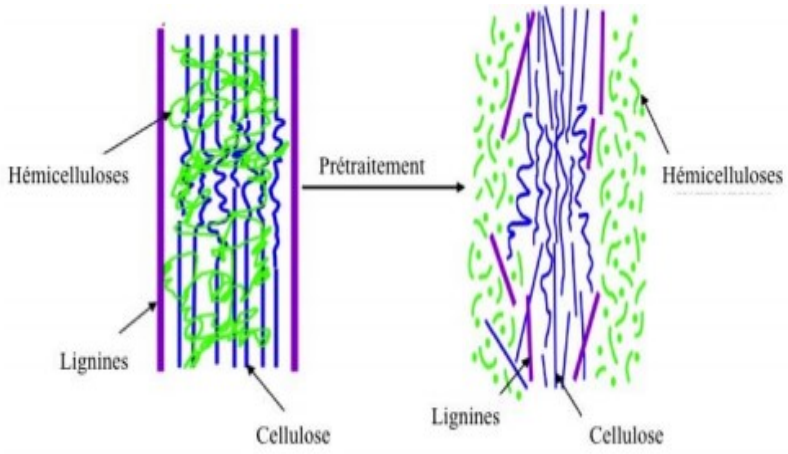


Figure 1.1. Représentation schématique de l'impact des prétraitements sur les complexes lignocellulosiques

Le choix du prétraitement est étroitement lié à la structure du substrat lignocellulosique et à la nature des enzymes utilisées lors de la saccharification. Un prétraitement idéal doit : (1) améliorer la dégradabilité enzymatique des sucres, (2) éviter de dégrader la cellulose, (3) ne pas engendrer des sous-produits capables d'inhiber les enzymes ou les levures chargées de la fermentation, (4) être peu gourmand en énergie, (5) être efficace sur de la biomasse de taille moyenne (non-broyée) et (6) être rentable [Bou08]. Les prétraitements peuvent être classés dans plusieurs catégories, par exemple chimiques (acide dilué, réactif alcalin), physiques (explosion à la vapeur). Les prétraitements les plus étudiés et les plus prometteurs sont [KBD09, BUA15] :

- Prétraitement à l'acide dilué : Ce type de prétraitement a pour objectif de solubiliser les hémicelluloses et rendre la cellulose plus accessible aux enzymes cellulolytiques [YW04, SRM08]. Les acides dilués tels que l'acide sulfurique (H_2SO_4) et l'acide nitrique (HNO_3) à de faibles concentrations (moins de 4 %) dégradent préférentiellement les xylanes pariétaux en xylose sans dégrader la cellulose. La sévérité de ce procédé dépend de plusieurs facteurs tels que le temps de réaction, la température et la concentration de l'acide.
- Prétraitement en milieu alcalin : Le principal objectif des prétraitements en milieux alcalins est de solubiliser la quasi-totalité des lignines [MWD05, KBD09]. Ces prétraitements représentent l'avantage de pouvoir se réaliser à température ambiante, évitant l'apparition de produits inhibiteurs, mais nécessitent une durée de traitement de quelques heures à quelques jours [KBD09]. Ce procédé a été largement étudié pour optimiser l'hydrolyse enzymatique des parois lignocellulosiques. Plusieurs solutions alcalines ont été appliquées telles que la soude, l'hydroxyde de potassium (KOH), l'hydroxyde de calcium ($Ca(OH)_2$), l'ammoniac (NH_3) et l'urée ($CO(NH2)_2$) [CBH97, CNH98]. Ces solutions alcalines vont, en plus d'extraire les lignines, éliminer les groupes acétates ou hydroxycinnamates substituant les hémicelluloses (essentiellement xylane), et rompre les liaisons esters au sein des parois.
- Prétraitements hydro thermiques en particulier « explosion à la vapeur » : Ce prétraitement repose sur deux étapes. La première consiste en un vapocraquage de la biomasse lignocellulosique à des températures allant de 160 à 260 °C (correspondant à des pressions de 10 à 50 bars) [MWD05, JVB10]. La vapeur diffuse à l'intérieur de la matrice pariétale et initierait

une hydrolyse partielle des hémicelluloses et la rupture des associations covalentes entre lignine et hémicellulose. La durée de cette étape varie de quelques secondes à quelques minutes. La deuxième étape consiste en une décompression explosive suite à une dépressurisation brutale, induisant un éclatement mécanique de la matrice pariétale [LHG07, HZ09, KBD09].

- Prétraitements biologiques : Ce type de prétraitement reste peu étudié par rapport aux procédés physicochimiques mais la nécessité de proposer des solutions plus respectueuses de l'environnement a fait renaître un intérêt croissant pour la solution biologique dans le contexte de la saccharification de lignocelluloses à destination énergétique [BKC10, DFM10, TSW05].

c) Divergence entre milieu naturel et milieu contrôlé

La biodégradation en milieu naturel tel que le sol consiste dans la décomposition de matières organiques par des micro-organismes comme les bactéries, les champignons ou les algues qui excrètent des enzymes dans le sol. Dans le cas de milieu contrôlé, la biodégradation des lignocelluloses en sucre fermentescible (bioéthanol) est effectuée sur des biomasses prétraitées par des cocktails d'enzymes que l'on obtient à partir de cultures de microorganismes.

La dégradation de la biomasse végétale dans les sols est un processus naturel. Le temps de décomposition (biodégradation) n'est pas le même suivant la nature des biomasses végétales, les conditions climatiques et peut varier de quelques jours à plusieurs milliers d'années. Le processus de biodégradation de la biomasse végétale dans le sol est un phénomène extrêmement lent, qui se déroule de manière continue sans réel arrêt du processus contrairement à la conversion chimique et enzymatique des biomasses en milieu contrôlé. Dans ce cas, c'est l'obtention du produit visé qui dicte la durée de la réaction en milieu contrôlé.

Dans les milieux contrôlés, certains composés peuvent être libérés et inhiber l'activité des enzymes. Par exemple, différents composés (furfural, acides, phénols …) sont formés au cours du prétraitement des biomasses lignocellulosiques et inhiber l'activité des cellulases [JAN13]. En revanche, ces inhibiteurs sont généralement métabolisés par les microorganismes présents dans le sol.

I.3. Caractérisation de la biomasse lignocellulosique et de sa dégradation

La décomposition des résidus végétaux dans le sol ainsi que la bioconversion de la biomasse végétale en alternative au carbone fossile sont limitées par la récalcitrance de la lignocellulose. De plus, la récalcitrance de la biomasse végétale à l'hydrolyse enzymatique est un problème industriel multifactoriel que l'on peut associer à une méconnaissance des rapports entre structure du substrat et efficacité enzymatique.

a) Fraction soluble

La fraction soluble des plantes comprend les composés qui, d'un point de vue méthodologique, sont extraits avec des solutions à faible agressivité chimique comme de l'eau à 20 ° C, l'eau à 100 ° C, un détergent à pH neutre ou d'autres détergents tels que l'acétone [Van63]. La fraction soluble est constituée de différents :

- acides aminés, peptides, protéines,
- sucres solubles (comme les sucres de stockage tels que l'amidon, métabolites)
- métabolites secondaires : pigments photosynthétiques, et d'autres molécules spécifiques (alcaloïdes, terpènes ...).

La composition et la proportion de la fraction soluble varient en fonction du type d'organe et le stade de maturité de la plante et représentent par exemple 25% et moins de 10% de la tige de blé à l'épiaison

(formation de l'épi) et maturité de la graine respectivement [BPC09]. En règle générale, la fraction soluble des résidus végétaux est rapidement assimilable par les microorganismes du sol lors de la décomposition à l'exception des composés tels que les tanins ou polyphénols [MHC07]. Cette fraction soluble est généralement transformée en composés inhibiteurs des activités enzymatiques lors des prétraitements physicochimiques de la biomasse [JAN13].

b) Fraction pariétale

La fraction pariétale de la biomasse lignocellulosique correspond aux parois cellulaires, parties externes des cellules végétales, et se caractérise par structure complexe, composée en majorité de trois fractions polymériques [Xu10, LWZ08, FSG10]:

- La cellulose : est le polymère dont la concentration est la plus abondante (35 à 50 % de la biomasse). Plus de cinq cents résidus glucosyles liés entre eux par une liaison β 1-4 forment ce polymère, lui conférant une structure linéaire. Ces chaînes s'associent entre elles pour donner des microfibrilles, structure cristalline qui constitue l'armature des parois cellulaires
- L'hémicellulose : est un polysaccharide non cellulosique se distinguant de la cellulose par le fait qu'elle est constituée de polymères hétérogènes (hétéropolysaccharides) à chaînes plus courtes et branchées. Elle représente de 25-35 % de la biomasse lignocellulosique. Cette fraction est essentiellement constituée de pentoses (comme la xylose) et de glucose.
- La lignine (un polymère complexe aromatique) : deuxième polymère naturel renouvelable le plus abondant sur terre, après la cellulose, ce polymère est troisième composant majeur de la biomasse lignocellulosique (5 à 25 %). Ce composé insoluble, composé d'unités phénylpropanes, est associé aux hémicelluloses par des liaisons chimiques covalentes pour former une matrice enrobant les micro-fibrilles de cellulose. C'est un polymère insoluble.

À l'intérieur de la lignocelluloses, ces trois macromolécules (la cellulose, l'hémicellulose et la lignine) s'entremêlent et forment une structure tridimensionnelle complexe et très résistante, maintenue par des liaisons covalentes et non covalentes. Chez les graminées, les acides phénoliques composés mineurs des parois végétales (moins de 3 %) ont néanmoins un rôle important dans la formation de liaisons covalentes entre la lignine et les hémicelluloses, contribuant ainsi à l'établissement d'une matrice lignine-hémicellulose enrobant les fibrilles de cellulose.

I.3.1. Caractérisation par chimie humide

La détermination quantitative des constituants majoritaires de la matière végétale brute constitue l'étape préliminaire indispensable à l'étude d'un procédé de fractionnement. En effet, la détermination du potentiel théorique de chacun des constituants est nécessaire à l'étude des bilans de matières.

Les teneurs en fraction solubles et pariétales peuvent être déterminées selon plusieurs méthodes. Une des méthodes les plus couramment utilisées est le fractionnement par la méthode de Van Soest qui permet de quantifier la fraction soluble dans une solution de détergent neutre et d'estimer la proportion en parois végétales, désignée par NDF (Neutral Detergent Fiber). De plus, le dosage de Van Soest permet de séparer les fractions « Acid Detergent Fiber » (ADF) et « Acid Detergent Lignin » (ADL) solubles dans des solutions neutres ou acides. Cette méthode, initialement mise au point pour prédire la digestibilité des fourrages par les animaux, est adaptée pour définir l'indice de stabilité biologique des amendements organiques. En effet, elle peut aboutir à la distinction entre les fibres dites solubles (certaines hémicelluloses, pectines) potentiellement plus digestibles par les microorganismes et les fibres insolubles (cellulose, hémicelluloses insolubles).

D'autres méthodes permettent de déterminer la fraction soluble sans extraire les polysaccharides solubilisés par la solution de détergent neutre de la technique Van Soest (pectines). Il s'agit le plus

souvent d'extraction par des solvants de type éthanol aqueux, le résidu alors obtenu correspondant aux parois végétales. Une hydrolyse acide séquentielle permet ensuite de quantifier les hémicelluloses et la cellulose, après dosage par chromatographie liquide des sucres monomères libérés par l'action des acides. Cette méthode est généralement mise en œuvre pour estimer le potentiel en glucose des biomasses lignocellulosiques en vue de la production de bioéthanol.

D'autres techniques de dégradation chimique permettent de caractériser la composition des polymères constitutifs des parois (nature et proportion en monomère). Les méthodes d'analyse par dégradation chimique précitées sont généralement longues (quelques heures) et dans certains cas comme l'extraction Van Soest nécessitent une certaine quantité d'échantillons (environ 1 gramme), ce qui peut être limitant pour caractériser une grande série d'échantillons. Le développement récent de plateformes analytiques basées sur la miniaturisation et la robotisation des protocoles ont permis d'augmenter le débit d'analyse, néanmoins ces systèmes analytiques ne sont pas largement accessibles.

I.3.2. Caractérisation par méthodes analytiques

Ces techniques nécessitent peu ou pas de préparation d'échantillon et fournissent des données qualitatives et quantitatives. On peut noter que le développement et la démocratisation de l'instrumentation utilisant les fibres optiques a permis des avancées considérables dans l'étude de processus chimiques. L'objectif de cette section est de fournir une analyse non exhaustive des méthodes spectroscopiques utilisées pour l'étude de la biomasse lignocellulosique.

a) Spectroscopie Raman

Quand une substance est mesurée par spectroscopie, la lumière interagissant avec l'échantillon peut être absorbée, transmise, diffusée. La spectroscopie Raman mesure la lumière diffusée à partir d'une molécule lorsqu'elle est irradiée par une source de lumière, généralement un laser [McC00, SD05]. Lorsque le laser interagit avec l'échantillon, les nuages d'électrons sont perturbés et vont exciter les molécules pendant une courte durée. On parle d'"état virtuel" car il n'y a pas assez d'énergie pour permettre à la molécule de rester à un niveau d'énergie plus élevé. Si l'énergie dispersée sous forme de photons diffère des photons incidents, un transfert d'énergie a eu lieu, conduisant à la diffusion Raman. Ce processus est de faible intensité puisqu'il concerne 1 pour 1.108 photons. Néanmoins, la spectroscopie Raman est non destructive, elle exige peu ou pas de préparation de l'échantillon et fournit une haute résolution spectrale. Elle n'est pas inhibée par la présence d'eau dans les échantillons et peut fournir des informations qualitatives et quantitatives abondantes. De nombreux travaux ont été développés autour de la spectroscopie Raman de substances lignocellulosiques [HDS10, Wor01, Aga99]. Un paramètre clé lors du développement expérimental d'une application Raman est la sélection de la longueur d'onde d'excitation.

Un inconvénient majeur est le suivant. De nombreux composés de la biomasse, tels que la lignine, présentent des contributions spectrales néfastes de fluorescence induites par l'excitation laser. Cette contribution peut dans ce cas perturber très sensiblement les signaux recueillis, le recours à une excitation laser de plus grande longueur d'onde.

b) Spectroscopie par Résonnance Magnétique Nucléaire (RMN)

Bien que la spectroscopie de RMN ne soit pas reconnue comme un technique rapide de traitement, sa prévalence dans l'analyse de la biomasse est importante [HSP09, AV99, ZLS09]. En spectroscopie RMN, les noyaux absorbent le rayonnement dans la gamme des radiofréquences du spectre électromagnétique [SHN98, SW00]. Les échantillons sont insérés dans des champs magnétiques

intenses. Ceux-ci vont faire interagir les noyaux différemment suivant l'état de leur moment magnétique. La différence de niveaux d'énergie des différents noyaux va permettre l'absorption ou l'émission de photons dans la gamme des radiofréquences. La RMN est décrite comme une méthode non destructive et non invasive. Puisque l'énergie d'excitation radiofréquence est faible et qu'elle ne peut apporter des changements physiques / chimiques moléculaires, elle constitue un excellent outil pour déterminer les structures moléculaires non modifiées [AV99]. La quantification peut être réalisée à partir des zones de non-chevauchement des spectres, puisque dans ce cas l'amplitude des pics est proportionnelle au nombre de noyaux responsables de celui-ci. Concernant les lignocelluloses, l'application de la RMN en phase solide est généralement moins précise que la RMN en phase liquide. Dans ce cas, il y a une préparation de l'échantillon pour parvenir à le solubiliser (on passe généralement par plusieurs étapes de broyages et/ou actions de liquides ioniques et ajout de fonction acétyles). Des exemples d'application de la RMN sur la biomasse lignocellulosique ont déjà été publiés [AV99, Mau02, RL10].

c) Spectrophotométrie U.V-Visible

La spectrophotométrie Ultraviolet-visible (UV-Visible) permet la mesure non destructive de transition électronique d'une molécule à des niveaux élevés d'énergie par le biais de l'absorption de lumière [Sch10]. L'analyse quantitative peut être réalisée en utilisant l'équation de Beer-Lambert, qui montre une dépendance linéaire entre les concentrations et l'absorbance pour des échantillons dilués. La spectrophotométrie UV-Visible est généralement utilisée conjointement avec des techniques chimiques par voie humide. Des mesures de teneurs en lignine par cette méthode montrent des performances similaires à celles obtenues à partir d'une analyse destructive de Klason de la lignine [KHW10]. Cependant, nous ne pouvons analyser que la fraction lignine et l'acide phénolique (ni cellulose et ni hémicellulose) qui n'absorbent pas en spectroscopie UV-Visible. Ainsi, comme dans le cas de la RMN en phase liquide, il faut solubiliser les échantillons par un réactif de type liquide ionique ou acétylation.

d) Spectroscopie de fluorescence

La spectroscopie de fluorescence a été appliquée pour l'étude non-destructive de la biomasse [GDR02, DKR07], son intérêt est grand car elle nécessite relativement peu de préparation de l'échantillon et l'instrumentation nécessaire est peu coûteuse. La méthode peut fonctionner en ligne ou hors ligne, et le système complet de mesure peut être implanté dans un appareil portatif. La spectroscopie de fluorescence mesure l'émission de rayonnement électromagnétique après qu'une molécule ait été excitée à des niveaux d'énergie très élevés. Bien que cette technique soit plus sensible que d'autres, la spectroscopie de fluorescence nécessite la présence d'un fluorophore, soit intrinsèque à la molécule, ou bien par marquage de l'échantillon avec une substance hautement fluorescente. Concernant les lignocelluloses, les spectres d'émission donnent généralement des spectres très larges, donc la déconvolution ne permet pas de remonter à une composition précise des lignocelluloses. En résumé c'est surtout une technique qui permet d'analyser la lignine.

e) Spectroscopie infrarouge FTIR

La spectroscopie infrarouge (IR) par transformée de Fourier (FTIR) est un autre outil d'analyse non destructif. Elle est souvent considérée comme complémentaire de la spectroscopie Raman puisque les modes de vibration qui sont actifs en Raman sont souvent faibles dans les spectres infrarouges, et vice-versa. En spectroscopie Raman, une modification de la polarisabilité (distorsion dans le nuage d'électrons) est primordiale pour obtenir le signal, tandis qu'un changement de dipôle est nécessaire pour un mode de vibration dans l'infrarouge. Une autre différence entre les deux techniques est que,

dans la spectroscopie IR, les molécules sont amenées à des niveaux d'énergie plus élevés, alors qu'en spectroscopie Raman, une molécule est amenée dans un état "virtuel" de courte durée et immédiatement relaxé. En raison des différences dans les spectres caractéristiques d'une molécule, l'utilisation conjointe des deux techniques peut fournir une meilleure analyse structurelle globale [HDS10, Coa00, SD05]. La spectroscopie FTIR a souvent été utilisée pour analyser toutes les structures (pas uniquement la lignine comme la spectroscopie UV) ou bien pour évaluer les changements qui se produisent dans la biomasse après application de traitements physiques et/ou chimiques. La FTIR peut concerner deux gammes de fréquences : le proche infrarouge (NIR) et le moyen infrarouge (MIR). Le principe de fonctionnement de la spectroscopie infrarouge par transformée de Fourier et son application sur des spectres MIR et NIR sera examiné en détail plus loin dans cette thèse.

f) Avantages et limites des différents types de spectroscopie

Dans cette section nous essayons de synthétiser ce que peuvent-nous apporter les différentes méthodes de spectroscopies, à savoir quels sont les avantages et les désavantages de chacune. Le Tableau 1.1 présente un résume synthétique.

Tableau 1.1. Avantages et désavantages des différentes méthodes analytiques non invasives d'analyse de la biomasse

Technique	Avantages	Désavantages
Spectroscopie de fluorescence	Coût faible, instrumentation simple Plus sélective que des méthodes d'absorbance	Nécessite d'avoir une molécule fluorophore, intrinsèque ou par ajout
Spectroscopie Raman	La quantification ne requiert pas de véritable standard de calibration N'est pas perturbée par la présence d'eau Pratiquement aucune préparation des échantillons Accepte différentes sources d'excitation Implantable in-situ Complémentaire de la FTIR	Signaux de faible intensité Peut-être masqué par des phénomènes de fluorescence Difficulté de se prémunir contre les autres sources de lumières qui viennent parasiter les mesures
Spectroscopie FTIR	Spectres d'intensité importante (plus que les spectres Raman) Complémentaire à la spectroscopie Raman Mesures relativement rapides (possibilité de répéter les mesures pour réduire le bruit)	Les spectres sont très sensibles à la présence d'eau Peut nécessiter des préparations spécifiques par exemple utilisation du KBr
Spectroscopie NMR	Permet d'avoir une description détaillée de la structure d'un échantillon Méthode non invasive	Nécessite souvent des temps longs d'analyse Les spectres présentent fréquemment des recouvrements de pics spécifiquement dans les régions des lignines Moins sensible que les autres méthodes spectrales
Spectrophotométrie UV-Visible	Complémentaire de la méthode FTIR Mesures rapides Coût peu élevé et instrumentation simple Précision et sélectivité Applicable sur une grande quantité de molécules	Etudes qualitatives limitées car faible différence entre minima et maxima d'absorption

I.4. Spectroscopie infrarouge (IR)

I.4.1. Généralité

a) Rayonnement et spectre électromagnétique

Le rayonnement électromagnétique, comme la lumière, est un phénomène vibratoire qui se propage sous la forme d'un transfert d'énergie entre deux ondes électromagnétiques. Sachant que l'onde électromagnétique est une perturbation périodique des champs électrique (E) et magnétique (B) associés à un flux continu de particules appelées photons (Figure 1.2), l'énergie d'un photon EP de fréquence v peut s'écrire :

$$EP = h\,v \qquad \text{(eq.1.1)}$$

où h est la constante de Planck (h = 6:6256 x 10-27 J.s),

$\vec{E}$: Champ électrique

$\vec{B}$: Champ magnétique

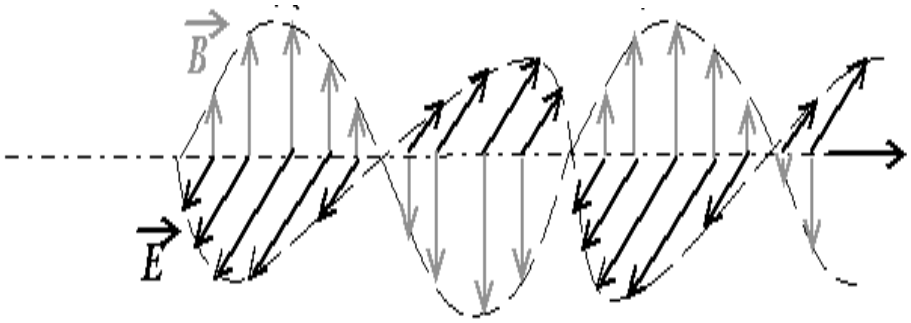


Figure 1.2. Représentation d'une onde électromagnétique

Le spectre électromagnétique est le résultat de la décomposition du rayonnement électromagnétique selon deux composantes en termes d'intensité et de fréquence v (nombre de vibrations par seconde exprimé en Hertz) ou de longueur d'onde λ (distance parcourue pendant une vibration exprimée en mètres) ou d'énergie des photons (EP).

Pour représenter un spectre, au lieu de la longueur d'onde, il est courant de manipuler le nombre d'onde, noté y, exprimé en cm^{-1} et défini comme l'inverse de la longueur d'onde λ [Lar11] :

$$y = \frac{\nu}{(c/n)} = \frac{1}{\lambda} \qquad \text{(eq.1.2)}$$

où c est la vitesse de la lumière et n l'indice de réfraction d'un milieu transparent. Dans le vide, n = 1 et c = 3.108 ms^{-1}.

Le spectre électromagnétique s'étend des ondes radio aux rayons gamma. La présente les différentes catégories d'ondes électromagnétiques en fonction de la fréquence ou du niveau d'énergie. Dans la suite de ce chapitre, nous allons nous intéresser au domaine spectral infrarouge parce que les techniques de spectroscopie infrarouge sont des méthodes d'analyse puissante de la composition

chimique d'un échantillon par mesure des phénomènes d'interactions entre onde et matière. Ces méthodes ont initialement permis de déterminer la structure moléculaire de nombreux composés issus de la chimie organique, et sont très utilisées pour la biomasse végétale.

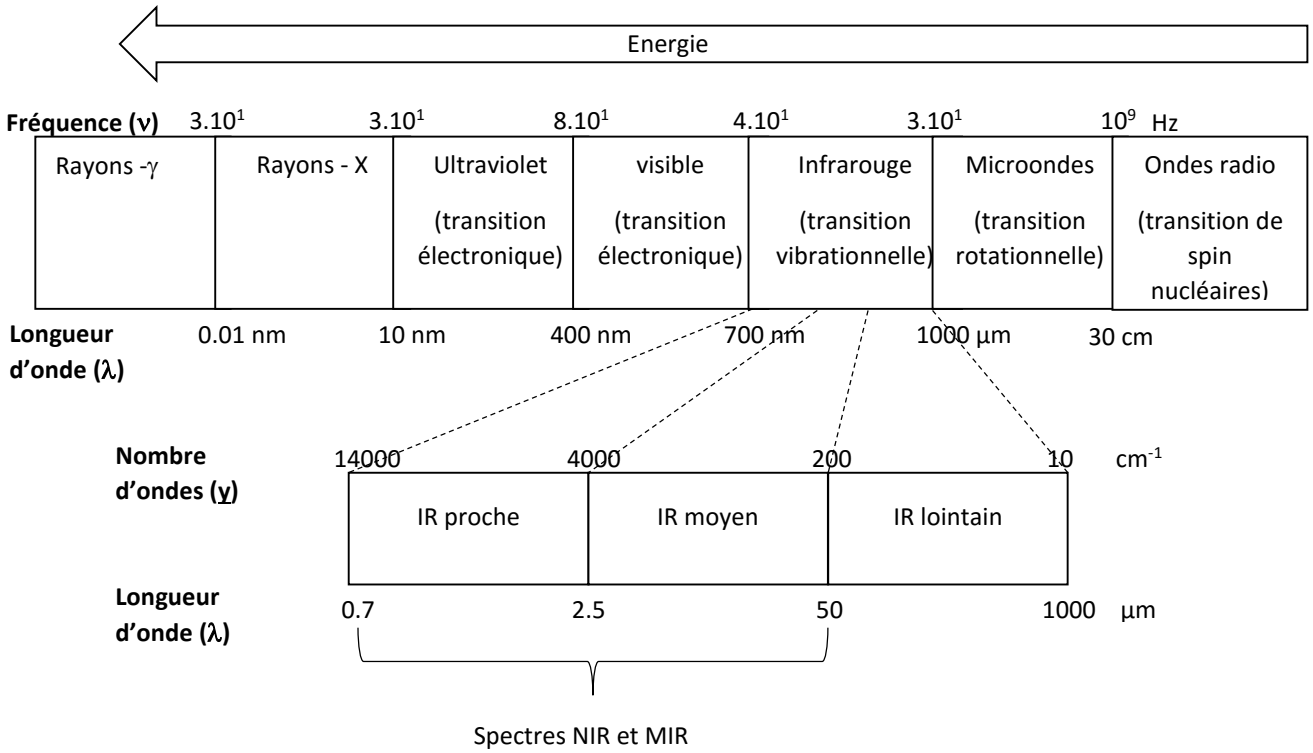


Figure 1.3. Les divers domaines spectraux du rayonnement électromagnétique

b) Définition et historique de la spectroscopie IR

La spectroscopie est la science qui traite des interactions des différents types de rayonnements avec la matière. La spectroscopie infrarouge (IR) est une classe de spectroscopie qui traite d'une région particulière du spectre électromagnétique et recouvre une large gamme de techniques. Historiquement, le chercheur Frederic Wilhelm Hershel a découvert le rayonnement infrarouge (IR) au début de XIXème siècle [Lar11]. Le rayonnement IR se situe entre la lumière visible et les micro-ondes. Il s'étend environ de 0.7 µm à 1000 µm de longueurs d'onde. La plus importante source naturelle de radiation IR est le soleil [GM87].

La spectroscopie IR a été la première technique de spectroscopie largement utilisée par les chimistes organiques. Dans les années 1930, la technique IR était expérimentalement limitée. Cependant, avec les développements conceptuels et expérimentaux, cette méthode est progressivement devenue une technique usuelle. Les premiers travaux importants en spectroscopie IR ont été réalisés dans l'industrie ainsi que dans les centres de recherches universitaires. Ces travaux, utilisant les modèles moléculaires et mécaniques, ont contribué à démontrer l'existence de modes de vibration pour différentes molécules [Lar11, KSA30, Col61].

La spectroscopie IR a été largement utilisée pour l'analyse quantitative et qualitative dans divers domaines tels que les industries alimentaires et pharmaceutiques puisqu'elle est très sensible et permet l'observation rapide de systèmes biologiques. Donnant des informations sur les modes de vibrations de l'échantillon analysé, la spectroscopie IR permet d'extraire des informations sur la composition moléculaire et la structure de l'échantillon [MRR02]. Le spectre infrarouge est traditionnellement découpé en trois parties présentées dans le Tableau 1.2.

Tableau 1.2. Longueurs d'ondes et nombres d'ondes de la région infrarouge du spectre électromagnétique

	Longueur d'onde $\underline{\lambda}$ (cm)	Nombre d'onde y (cm^{-1})
Proche Infrarouge	$0,7\ 10^{-4}$ à $2,5\ 10^{-4}$	14000 à 4000
Moyen Infrarouge	$2,5\ 10^{-4}$ à $5\ 10^{-3}$	4000 à 200
Lointain Infrarouge	$5\ 10^{-3}$ à 10^{-1}	200 à 10

c) Principe de la spectroscopie infrarouge IR

Le principe de la spectroscopie IR est d'envoyer un rayonnement IR sur un échantillon à tester. Certaines longueurs d'onde sont absorbées par les liaisons chimiques des molécules se trouvant dans l'échantillon [CDW75]. Il devient alors possible de générer un spectre IR absorbé, qui est le résultat du mécanisme de vibration moléculaire fondamental et qui se réfère à l'interaction de l'énergie et de la matière aux différentes longueurs d'ondes [Bru62]. Dans la suite de ce chapitre, nous présentons brièvement les différents modes des vibrations moléculaires.

Une molécule est un ensemble non rigide d'atomes liés les uns aux autres et qui peuvent être modélisés comme des ressorts de constante de raideur plus ou moins grande (les liaisons). Ces ressorts présentent des modes de vibrations qui apparaissent à des fréquences déterminées v.

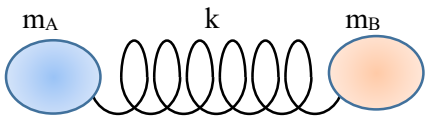


Figure 1.4. Modèle simple d'une molécule : deux masses m_A et m_B sont liées par un ressort caractérisé par une constante de raideur k.

La liaison covalente est ici modélisée par deux atomes de masses m_A et m_B liés par un ressort de constante de raideur k (Figure 1.4). La fréquence de vibration de la molécule v vérifie la loi de Hooke :

$$v = \frac{1}{2\pi}\sqrt{\frac{k}{\mu}} \text{ avec } \mu = \frac{m_A\, m_B}{m_A + m_B} \qquad \text{(eq.1.3)}$$

Lorsqu'un rayonnement infrarouge à fréquence de vibration de la molécule la frappe, l'amplitude de la vibration s'accroit et de l'énergie est absorbée. Ces vibrations ou modes de vibration donnent naissance aux 3 principaux types de bandes d'absorption [BM94, Her45] :

- les bandes de valence : elles sont caractéristiques d'un des modes de vibration fondamental d'une liaison ;
- les bandes harmoniques : leurs fréquences sont des multiples entiers de la fréquence d'une vibration fondamentale ;
- les bandes de combinaison : elles résultent de la combinaison (addition ou soustraction) des vibrations de plusieurs liaisons.

Il existe également les bandes nées de la résonance de Fermi qui sont une combinaison d'un mode fondamental et d'un mode harmonique. D'autre part, les fréquences de vibrations dépendent des masses des atomes et des forces des liaisons de covalence. De ces considérations se dégagent deux classes de vibrations moléculaires [BM94, Her45] :

- les vibrations de valence ou d'élongation (symétriques ou antisymétriques) qui font intervenir une ou des variations de longueurs de liaisons, les angles formant ces liaisons restant constants ;
- les modes de déformation pour lesquels, au contraire, les liaisons gardent leur longueur, mais les angles qu'elles forment varient. La Figure 1.5 représente les exemples de différents modes de vibration dans le cas des molécules triatomiques (H_2O et CO_2) [Soc80].

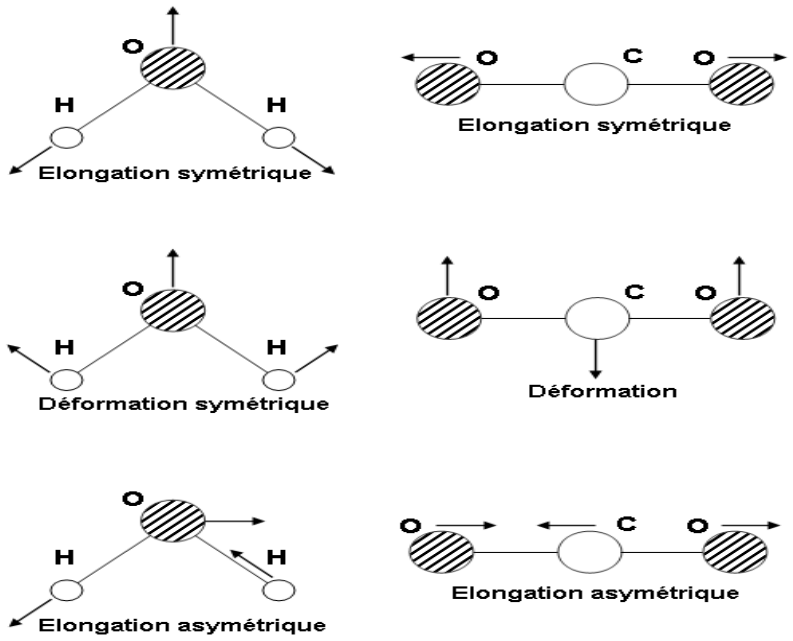


Figure 1.5. Exemple de différents modes de vibration des molécules d'H_2O et de CO_2 [Soc80]

Il existe deux types de dispositifs qui permettent de mesurer le spectre infrarouge. Le premier est dispersif (spectromètre IR à balayage) et le second est non dispersif (spectromètre IR à transformée de Fourier).
Comme nous l'avons mentionné précédemment la région étudiée du spectre infrarouge peut appartenir au moyen infrarouge ou bien au proche infrarouge. Dans la suite de ce chapitre, nous introduisons les spécificités de l'analyse de ces spectres et leurs applications pour l'étude de la biodégradation de la biomasse lignocellulosique.

I.4.2. Spectroscopie moyen infrarouge (MIR)

La spectroscopie moyen infrarouge (mid infrared en anglais, MIR) est basée sur l'absorption d'un rayonnement moyen IR sur le matériel analysé pour des nombres d'ondes allant de 400 à 4000 cm^{-1}. Les bandes d'absorption dans la région MIR résultent des modes de vibration. Ces modes peuvent être attribués à des groupements fonctionnels fondamentaux. L'identification des différentes bandes d'absorption se fait en référence aux données bibliographiques.
La spécificité chimique de la région moyenne infrarouge MIR permet l'identification de certains pics liés aux composants de l'échantillon où l'amplitude de chaque raie spectrale est un marqueur de la composition moléculaire et de la structure de celui-ci. La Figure 1.6 représente les spectres MIR pour deux échantillons de biomasse lignocellulosique : maïs et miscanthus.

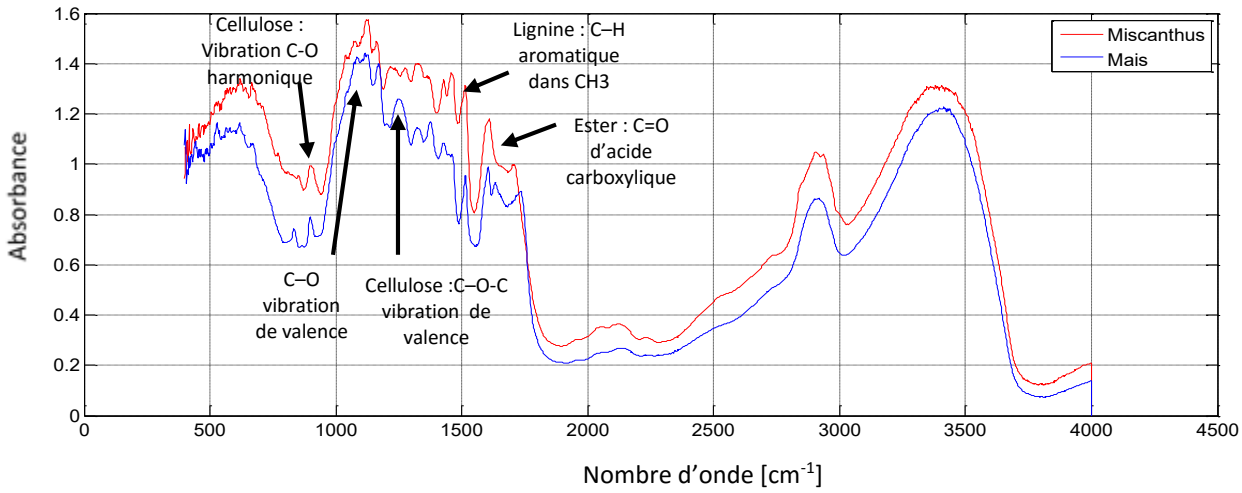


Figure 1.6. Exemple de spectres moyen infrarouge MIR pour deux types de biomasses végétales : maïs et miscanthus

Cette figure montre les variations de bandes spectrales (amplitude de pics et positions) qui ont des significations chimiques (modes de vibrations) communes ou spécifiques pour les deux types de biomasse lignocellulosique analysés. On peut observer que la plupart des pics sont centrés essentiellement dans la gamme allant de 800 à 1800 cm^{-1}.
La spectroscopie MIR est une technique d'identification de bandes spectrales qui correspondent aux groupes fonctionnels chimiques connus. Le Tableau 1.3 présente ces différentes bandes d'absorption, les modes de vibration et les attributions chimiques de la matière organique identifiée par la spectroscopie MIR.

La spectroscopie MIR est une technique largement utilisée dans une variété de secteurs, notamment l'agriculture [Ree94, ReeIII94, Ree96, RD97], les industries pharmaceutiques et agroalimentaires. Elle peut être utilisée à la fois dans l'analyse qualitative et quantitative d'un produit ou d'un procédé. Au cours des dernières années, la spectroscopie MIR a été utilisée dans le domaine alimentaire, notamment pour des études sur les huiles et graisses comestibles. Elle a été utilisée avec succès pour la surveillance du processus d'oxydation de l'huile de maïs [VSP06, TSK02, KHL01, MKD98, LKY95] et pour la classification de différentes huiles d'olive [CSO10].
Dans le domaine de l'agriculture, la spectroscopie MIR est récemment devenue la méthode de référence pour l'analyse rapide et fiable de matières agricoles comme les céréales, les aliments, les fourrages et les sols ainsi que pour la détermination de la composition de la biomasse lignocellulosique et son évolution chimique pendant la phase de biodégradation.
Dans les études de sol, la spectroscopie MIR montre un certain potentiel pour l'analyse efficace de leur composition [RWM06]. Elle a été appliquée pour la quantification du nitrate, du carbone total (C), des teneurs en azote (N) et la teneur en métaux lourds dans les sols [MRR02].
La spectroscopie MIR a été utilisée pour la classification des isolats fongiques et pour différencier la matière végétale contenant la croissance fongique à partir de matériau non infecté. Elle a également été aussi utilisée pour quantifier les acides gras (FA) dans une variété d'espèces de plantes fourragères [CRF07] et pour trouver des marqueurs infrarouges dans les plantes [CAD09].
Comme les parois des cellules végétales ont une structure très complexe, principalement constituée d'hydrates, de carbone et de protéines, la spectroscopie MIR a été utilisée pour la détermination rapide et fiable des principaux constituants des parois cellulaires des plantes et fruits au cours du leur développement [SCK15, LHG14]. Elle permet également l'identification de certains pics liés aux

composants de la paroi cellulaire [XYT13, WL99]. Elle a été mise en œuvre pour caractériser la chimie du bois afin de déterminer la teneur en lignine de la pâte, du papier et du bois [MO01]. Dans des études récentes, la spectroscopie MIR a été appliquée pour déterminer les quantités relatives de cellulose, d'hémicellulose et de lignine dans la biomasse lignocellulosique et pour étudier l'évaluation de la dégradation de la biomasse lignocellulosique [NDS10, CFA10, KCC12, ATM09].

Tableau 1.3. Principales bandes d'absorption dans l'infrarouge moyen(MIR) d'intérêt pour l'étude de la matière organique et des sols [SRP04]

Nombre d'onde cm^{-1}	Attribution	Vibration	Source
1710-1760	Groupe ester (C=O des acides carboxyliques)	élongation symétrique	Xylanes (hémicelluloses)
1660	Groupe C=O conjugués au cycle aromatique	élongation symétrique	Lignines
1600	C=C conjugués au cycle aromatique	Vibration du squelette Elongation Déformation	Lignines
1505	C=C conjugués au cycle aromatique	Vibration du squelette	Lignines
1460	Déformation dans le plan de groupes CH ; CH_3 et $–CH_2–$	Déformation, élongation	Lignines, hémicelluloses
1425	CH aromatique	Déformation, élongation	Lignines, cellulose polysaccharides
1370-1375	C–H aromatique dans CH_3 ; déformation vibration de CH, déformation dans le plan de groupes CH_2	Déformation, élongation	Lignine, cellulose Polysaccharides
1335	C-O cycle aliphatique	Déformation	Cellulose
1325	Vibration C-O et déformation dans le plan de groupes CH_2	Déformation	Lignines, polysaccharides
1275	Vibration C-O	élongation	Lignines
1240	Déformation des groupes acétyles (xylanes)	Déformation	Lignines
1160-1162	Vibration C-O-C des hémicelluloses et de la cellulose	élongation asymétrique	Cellulose, lignine, polysaccharides
1110	Vibration O-H de la cellulose et des hémicelluloses	élongation	polysaccharides
1050	Vibration C-O de la cellulose et des hémicelluloses	élongation	Cellulose, polysaccharides
890-830	Anomere C-groupes, déformation C-H, ring valence vibrations	élongation	Lignine, cellulose, hémicelluloses

I.4.3. Spectroscopie proche infrarouge (NIR)

La spectroscopie proche infrarouge (Near InfraRed, NIR) est basée sur l'absorption d'un rayonnement proche infrarouge par le matériel analysé pouvant aller de 4000 à 14000 cm^{-1}. Comme pour la spectroscopie MIR, l'absorption est modulée par les vibrations des liaisons chimiques de l'échantillon. Dans le domaine du proche infrarouge, les absorptions ne sont pas dues aux vibrations fondamentales des molécules mais à des vibrations harmoniques et à des vibrations de combinaisons. Il en résulte des bandes d'absorption très larges et un fort chevauchement des bandes d'absorption des différentes

liaisons chimiques (Figure 1.7). Il est donc difficile d'assigner des bandes d'absorption à des caractéristiques chimiques précises de l'échantillon comme dans l'infrarouge moyen.

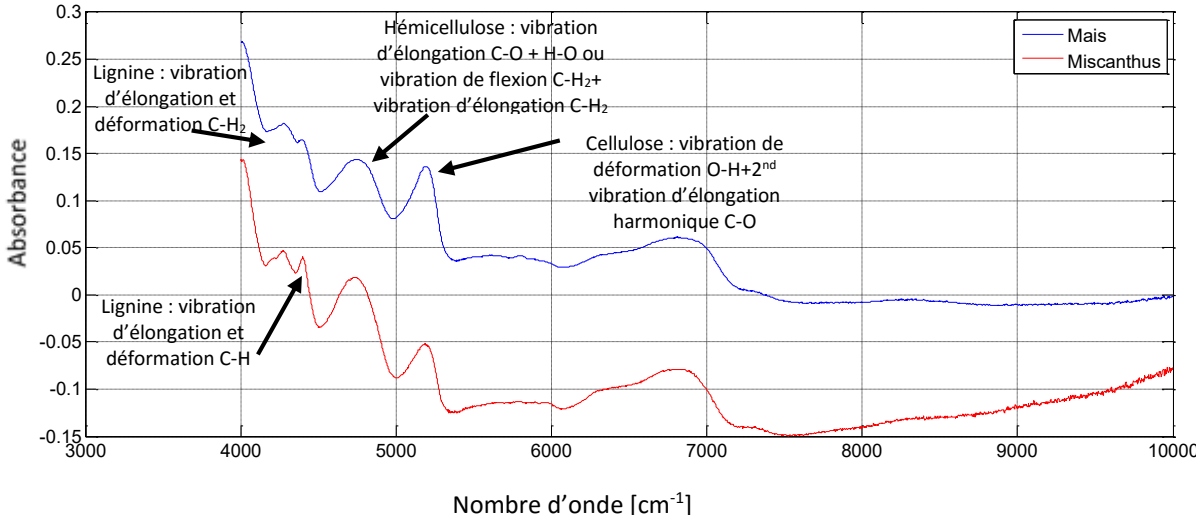


Figure 1.7. Exemple de spectres proche infrarouge NIR pour deux types de biomasses végétales : maïs et miscanthus

Sur la Figure 1.7 on observe les variations de bandes spectrales qui résultent des vibrations harmoniques et des vibrations de combinaisons pour les deux échantillons de biomasse lignocellulosique, maïs et miscanthus. Nous pouvons trouver que les bandes spectrales le plus importantes se situent essentiellement dans la gamme 4000 à 8000 cm^{-1}.

La spectroscopie NIR est couramment utilisée dans l'industrie. Elle sert à l'analyse quantitative et qualitative des produits pharmaceutiques, agricoles, polymères et l'analyse biologique [MCS02, RCM07, SSS03, Wil08]. La spectroscopie NIR est une technique d'analyse qui permet de connaître la composition chimique des aliments et des matières premières. Elle est également utilisée depuis plusieurs années comme méthode permettant l'évaluation de la qualité intrinsèque des fruits et légumes [Lan83, Del95, DPC98, SSK04]. Elle a été utilisée avec succès pour l'analyse de la composition et de la valeur nutritive de matières végétales telles que les matières grasses brutes, les protéines, les glucides et les fibres brutes [MAM91, MMA91, JGD92, GHJ99, AA97, BDH02, BBP02, OHA03].
Dans le domaine de l'agriculture, la spectroscopie NIR s'avère être un outil rapide et performant pour prédire des teneurs en C (carbone) et N (Azote) du sol [BBF06-BDC07], ainsi que pour prédire la minéralisation du carbone et de l'azote au cours de l'incubations de matériaux végétaux [BMB10, BSB05] ou la minéralisation de la matière organique du sol [TBJ09].

La spectroscopie NIR est également une technique rapide et non destructive pour l'analyse qualitative et quantitative des produits agricoles allant des fourrages aux céréales en passant par le tabac et le bois [WN01]. Elle a été utilisée pour évaluer la prédiction des composés de la biomasse lignocellulosiques (canne à sucre, miscanthus, fourrages, maïs, etc.) [TSH09, WS09, STB12]. Par exemple, Hames et al. [HTS03] ont employé les spectres NIR pour l'analyse de la composition de maïs. Les résultats obtenus suggèrent une bonne prédiction pour le glucane, le xylane, la lignine et les protéines. La spectroscopie NIR est également un outil puissant pour prédire d'autres propriétés, telles que l'humidité, ce qui est utile dans l'évaluation du traitement de la biomasse [LR05, LLC08]. Le Tableau 1.4 présente les bandes spectrales qui correspondent aux groupes fonctionnels chimiques rencontrés dans les composés organiques et les sols identifiés par la spectroscopie NIR.

Tableau 1.4. Principales bandes d'absorption dans l'infrarouge proche (NIR) d'intérêt dans l'étude de la matière organique et des sols [SRF11, FGP14]

Nombre d'onde cm^{-1}	Attribution	Source
4063	vibration d'élongation C-H + vibration de déformation C-H	Cellulose
4235	vibration de déformation O-H ou C-H + vibration d'élongation C-H ou C-H_2	Cellulose
4268	vibration d'élongation C-H + vibration de déformation C-H et 2nd harmonique C-H	Cellulose
4280	vibration d'élongation C-H + vibration de déformation C-H_2	Lignine
4283	vibration d'élongation C-H + vibration de déformation C-H	Cellulose et hémicellulose
4296-4288	vibration d'élongation C-H + vibration de déformation C-H	Hémicellulose
4365	vibration d'élongation C-O + H-O ou vibration de flexion C-H_2+ vibration d'élongation C-H_2	Cellulose
4401	vibration d'élongation C-H + vibration de déformation C-H	Hémicellulose
4404	vibration d'élongation C-H_2 + vibration de déformation C-H_2	Cellulose et hémicellulose
4411	vibration d'élongation H-O + vibration d'élongation C-O	Lignine
4546	vibration d'élongation C-H + vibration d'élongation C=O	Lignine
4686	vibration d'élongation C-H + vibration d'élongation C=O	Hémicellulose
4739	vibration de déformation O-H+ vibration d'élongation H-O	Cellulose
4780-4760 ; 4808	vibration de déformation O-H et C-H + vibration d'élongation O-H	Cellulose
5051 ; 5220-5150	vibration de déformation O-H+ vibration d'élongation H-O de H_2O	Eau
5495 ; 5464	vibration de déformation O-H+2nd vibration d'élongation harmonique C-O	Cellulose
5577 ; 5593	Premier vibration d'élongation harmonique C-H	Cellulose
5583 ; 5795	Premier vibration d'élongation C-H	Lignine
5618	Premier vibration d'élongation harmonique C-H_2	Cellulose
5816	Premier vibration d'élongation harmonique C-H	Cellulose /hémicellulose/lignine
5800 ; 5848 ; 5865	Premier vibration d'élongation harmonique C-H	Hémicellulose
5900 ; 5939	Premier vibration d'élongation C-H	Lignine
5950	Premier vibration d'élongation C-H	Hémicellulose
5980 ; 5974 ; 5963 ; 5978	Premier vibration d'élongation C-H	Lignine
6003	Premier vibration d'élongation C-H	Hémicellulose

I.4.4. Combinaison des spectres MIR et NIR

A cause de la nature complexe mais complémentaire des deux techniques spectroscopiques MIR et NIR, la combinaison des informations spectrales MIR et NIR est considérée comme l'un des défis actuels pour bon nombre d'applications en spectroscopie. Des méthodes ont été proposées pour combiner les spectres MIR et NIR telles que la combinaison par corrélation à deux dimensions (two dimensional correlation spectroscopy, IR 2D CoS), la combinaison par concaténation et la combinaison par produit extérieur [BMF02, IGG06].

La méthode la plus usuelle de combinaison utilisée est la concaténation des spectres MIR et NIR. Cette méthode permet de juxtaposer les informations MIR à côté des informations NIR. Cependant, cette technique de combinaison comporte des inconvénients qui sont essentiellement les problèmes de discontinuité et de différence d'échelle entre les spectres MIR et NIR [CSO10]. De plus, elle ne permet pas d'étudier l'interaction entre les informations spectrales à tous les nombres d'ondes MIR et NIR. La Figure 1.8 illustre le processus de combinaison par concaténation de deux spectres MIR et NIR.
Cette méthode de combinaison a été appliquée avec succès dans une variété de secteurs, telle que l'agriculture, l'industrie pharmaceutique et agroalimentaire. Récemment, les spectres MIR et NIR ont été combinés par concaténation pour prédire la quantité d'huile et d'eau dans des échantillons de fruits frais d'olive mais aussi pour différentier des échantillons de rhubarbe ainsi que pour déterminer les paramètres de qualité de la bière, tels que la teneur en alcool [DGL10, WMN14].

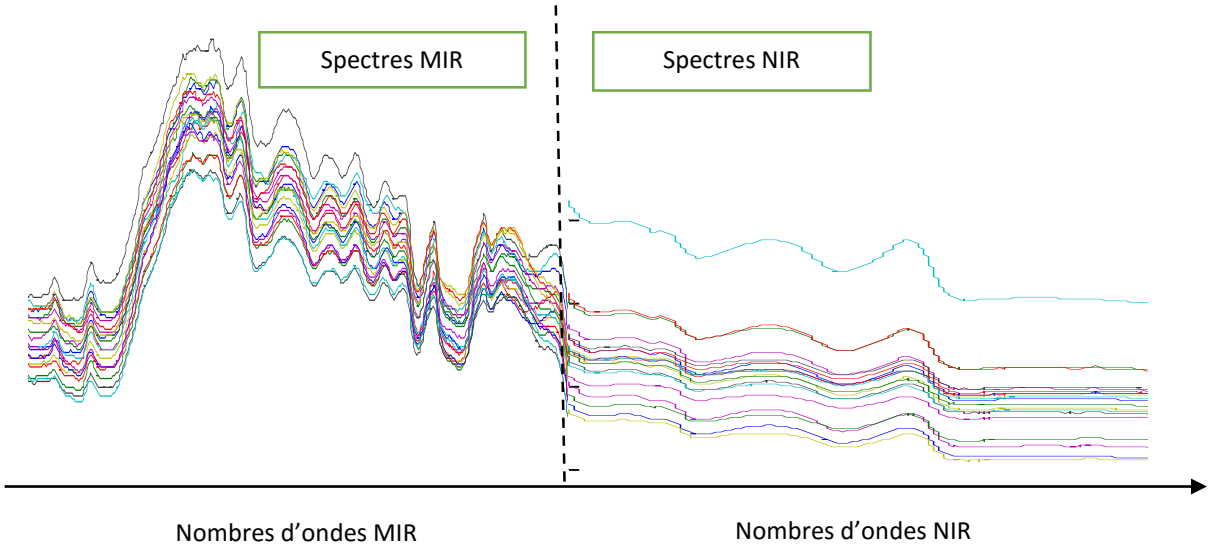


Figure 1.8. Représentation schématique de la combinaison de spectres NIR et MIR enregistrés sur des échantillons de biomasse lignocellulosique par concaténation.

La combinaison par corrélation à deux dimensions (IR 2DCoS) a été introduite afin de révéler les corrélations entre les changements de bandes spectrales et déconvoluer les pics qui se chevauchent. Cette technique s'est développée rapidement. Elle a été appliquée dans différentes applications agroalimentaires. Elle a été utilisée pour étudier les différences caractéristiques entre les échantillons de rhubarbe [ZWH09], et étudier les modifications structurelles de bois pendant le traitement thermique [PFN13]. Elle est également un outil très populaire dans l'analyse des spectres IR de protéines. Cependant, il nous est impossible d'utiliser cette méthode pour combiner les spectres MIR et NIR puisqu'elle permet seulement de corréler les bandes spectrales MIR ou NIR entre elles (mêmes types d'informations spectrales).

Le produit extérieur (Outer Product en anglais, OP) est une autre méthode de combinaison de spectres qui permet de combiner deux spectres de différents types et tailles en calculant leur produit extérieur. La combinaison par le produit extérieur entre les spectres MIR et NIR permet d'associer l'information spectrale de chaque nombre d'onde MIR avec l'information spectrale de chaque nombre d'onde NIR. Dans ce procédé, les spectres acquis dans les deux domaines MIR et NIR sont combinés en calculant la matrice de produit extérieur des deux vecteurs qui représentent chaque spectre (MIR et NIR) d'un

échantillon. Le produit extérieur de spectres MIR et NIR met en évidence l'information spectrale mutuelle à tous les nombres d'onde MIR et NIR car chaque nombre d'onde MIR se trouve lié à chaque nombre d'onde NIR.

Physiquement, cette méthode de combinaison met l'accent sur l'interaction entre les vibrations moléculaires fondamentales et harmoniques et les combinaisons de vibrations fondamentales. La Figure 1.9 illustre le processus de combinaison par l'OP de deux spectres MIR et NIR.

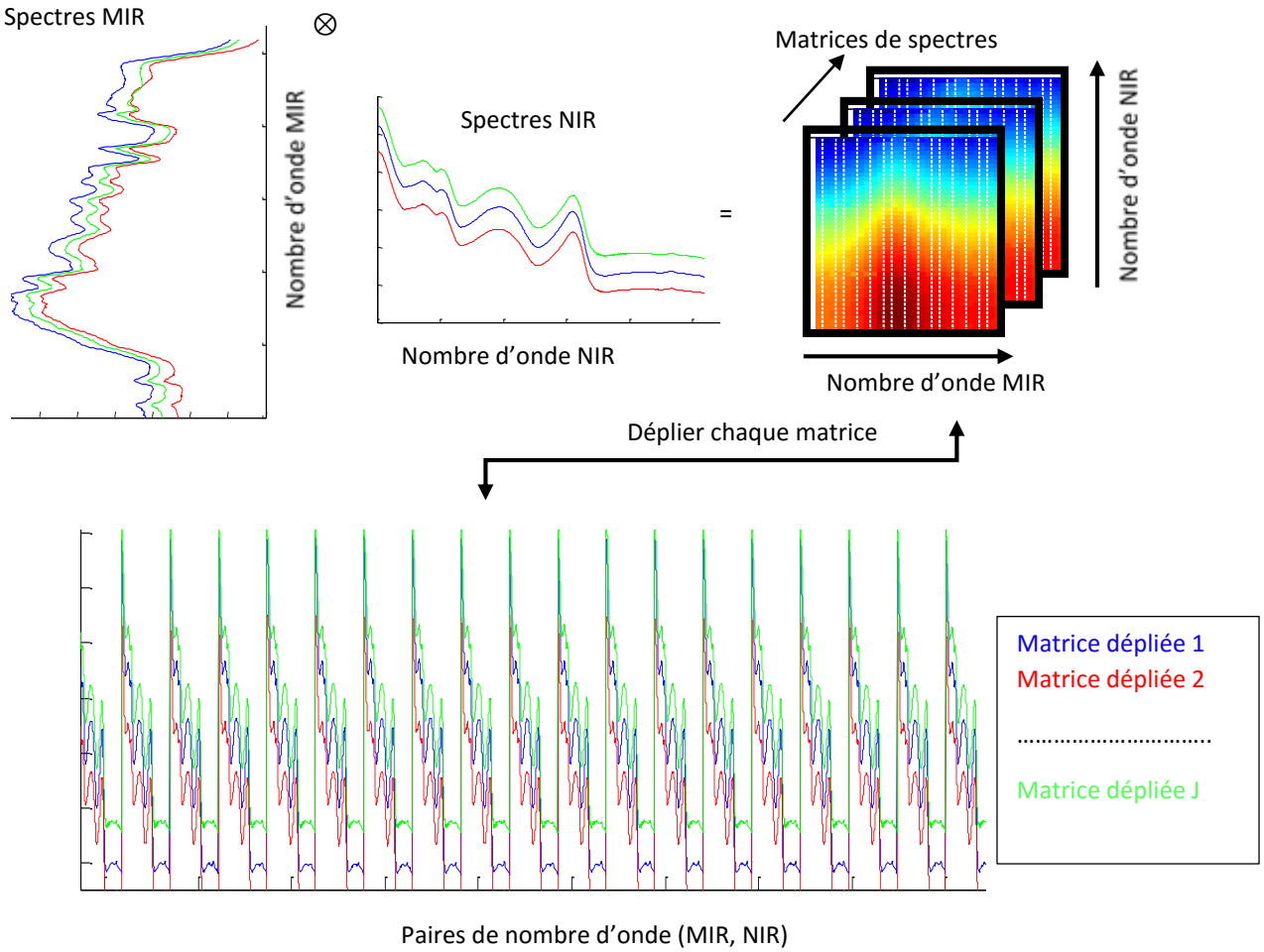


Figure 1.9. Représentation schématique de la combinaison de spectres NIR et MIR par le produit extérieur OP. Les spectres MIR, NIR, MIR⊗NIR ayant la même couleur correspondent au même échantillon.

La combinaison de spectres MIR et NIR par le produit extérieur (OP) a été récemment appliquée pour déterminer le degré de méthyl estérification des polysaccharides pectiques, pour prédire la teneur en sucre de la betterave, étudier l'influence de la température sur les spectres NIR de l'eau et plus récemment, pour déterminer la composition de la matière organique du sol [JPR05, JOF06]. Cependant, les études sur la combinaison entre les spectres MIR et NIR restent limitées. Cela peut s'expliquer par la nature et la difficulté d'interpréter les résultats obtenus. En effet, si on combine un spectre MIR qui a 500 nombres d'ondes avec un spectre NIR de 1000 nombres d'ondes, le résultat est une matrice de 500 x 1000 nombres d'ondes qui peut être dépliée sous la forme d'un vecteur de 500 000 nombres d'ondes combinés. Sur la matrice obtenue, il est possible de calculer des profils de variance [BMF02] ou appliquer des méthodes chimiométriques comme l'analyse en composantes principales [JPR05], PLS [JOF06] sur les vecteurs dépliés. Les résultats obtenus (facteurs PCA ou coefficients de régressions) étant remis sur une forme matricielle. Une autre approche a été d'utiliser les matrices non dépliées à l'aide de modèles de décomposition 3D [RB07]. Dans tous les cas, les

résultats obtenus sont analysés qualitativement et l'interprétation n'est pas toujours facile. Néanmoins, cette méthode de combinaison est très intéressante puisqu'elle met en évidence toutes les interactions entre les bandes et nous allons l'exploiter par la suite.

I.5. Instrumentation et acquisition

I.5.1. Instrumentation

a) Spectroscopie IR à balayage

Le spectromètre IR à balayage a émergé dans les années 1940. Les éléments principaux d'un spectromètre IR à balayage, conçu selon le schéma de principe représenté sur la Figure 1.10, sont une source de radiations IR, un système dispersif de séparation de radiations IR et un détecteur du signal.

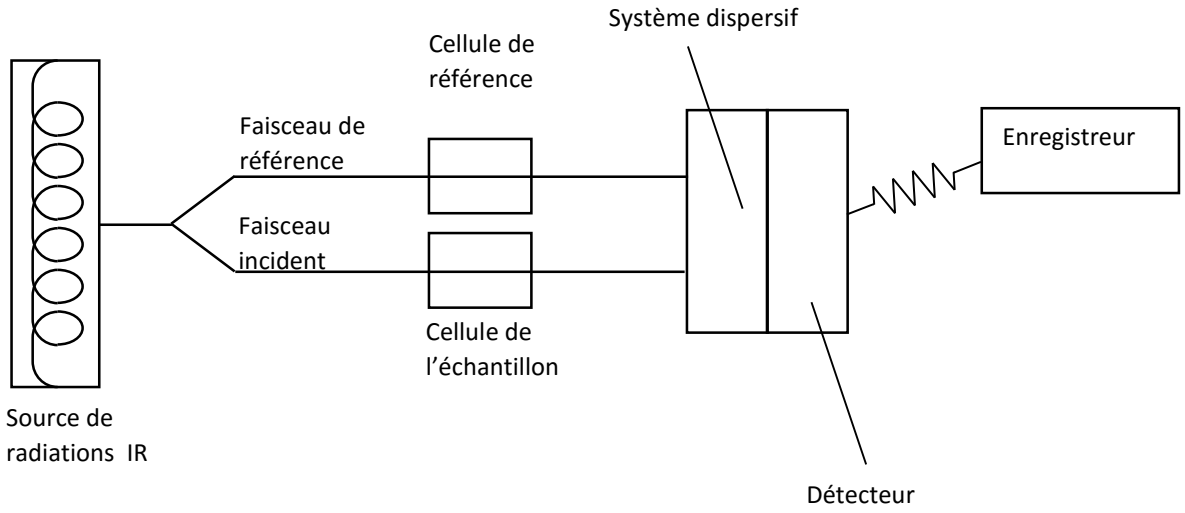


Figure 1.10. Schéma de principe d'un spectromètre IR dispersif

Les fréquences émises par la source infrarouge sont séparées à l'aide d'un prisme ou de réseaux (bloc de silice sur lequel on a gravé des traits métalliques en surface), éléments dispersifs plus efficaces.
Ce système a contribué à répandre l'utilisation de la spectroscopie infrarouge comme une technique d'analyse commune pour la caractérisation des composés organiques dans les laboratoires. Le spectromètre IR à balayage présente cependant des inconvénients :

- La durée des mesures : étant donné que l'instrument mesure chaque fréquence individuellement, l'enregistrement prend de 10 à 15 minutes,
- La relative insensibilité : la détection nécessite une quantité raisonnable de produits pour une analyse exploitable.
- La complexité mécanique : l'existence de certaines parties mobiles toutes sujettes à des problèmes de casse mécanique.

Pour résoudre les difficultés de lenteur de l'acquisition et de limitations des spectromètres dispersifs, il est apparu comme indispensable d'imaginer un dispositif mesurant toutes les fréquences simultanément. Ce dispositif est le spectromètre IR à transformée de Fourier (FT-IR).

b) Spectroscopie IR à transformée de Fourier (FT-IR)

Le spectromètre à transformée de Fourier (FT-IR) de type non dispersif a été développé dans les années 1960. Il comporte les éléments suivants : une source de radiations IR, l'interféromètre de Michelson,

un compartiment d'échantillon, un détecteur, un amplificateur, le convertisseur analogique-numérique qui interroge le détecteur à des intervalles réguliers et transforme le signal analogique en un signal numérique manipulable par le système informatique et un ordinateur [BrS11].
La radiation IR passe à travers l'échantillon après l'interféromètre et atteint le détecteur. Puis le signal est amplifié et converti en un signal digital par l'amplificateur et le convertisseur analogique-numérique, respectivement. Finalement, le signal est transféré à un ordinateur où la transformé de Fourier est calculée. La Figure 1.11 montre le schéma de principe d'un spectromètre FT-IR.

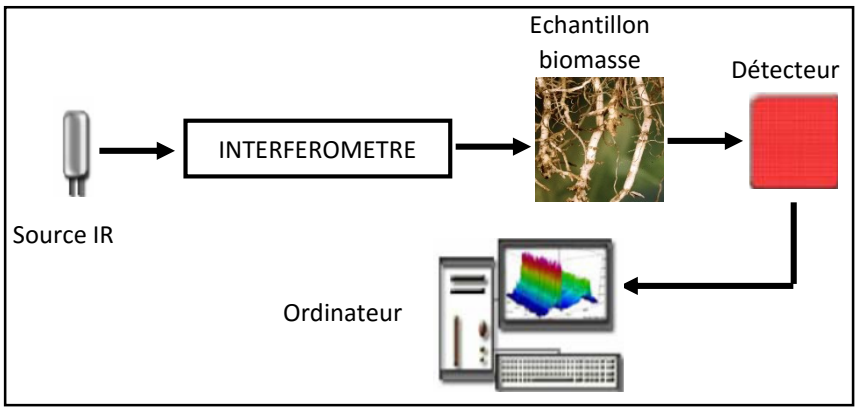


Figure 1.11. Schéma de principe d'un spectromètre FT-IR [NB76]

L'interféromètre de Michelson, qui est une pièce essentielle du spectromètre FT-IR, comporte deux miroirs perpendiculaires et un séparateur de faisceau. Un de ces miroirs est fixe (stationnaire) et l'autre est mobile. Le séparateur de faisceau est désigné pour transmettre la moitié de l'intensité lumineuse et réfléchir l'autre moitié. Par la suite, la lumière transportée et la lumière réfléchie frappent le miroir stationnaire et le miroir mobile, respectivement. Quand ils sont réfléchis par les deux miroirs, les deux faisceaux de lumière se recombinent ensemble au niveau du séparateur de faisceaux de manière cohérente. Ces faisceaux combinés sont envoyés vers l'échantillon. Le signal sortant de l'interféromètre créé par ces faisceaux combinés est enregistré par le détecteur pour constituer un interférogramme [BrS11]. La Figure 1.12 montre le schéma de principe de l'interféromètre de Michelson.

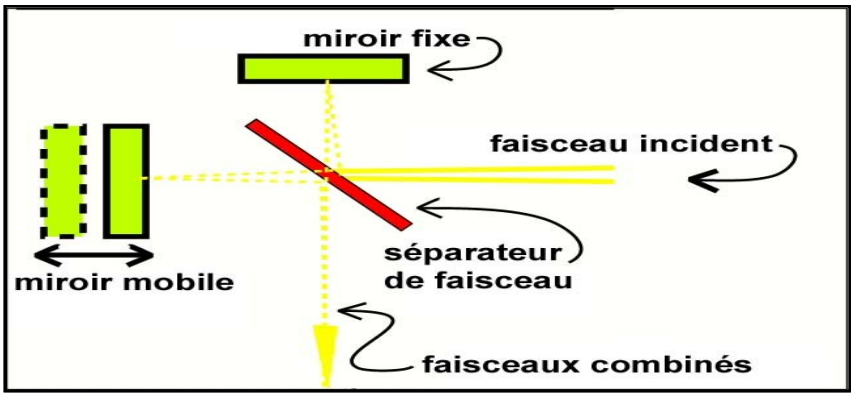


Figure 1.12. Schéma de principe de l'interféromètre de Michelson.

L'interférogramme (Figure 1.13) est ensuite converti en un spectre infrarouge grâce à la transformée de Fourier. Un exemple de spectre ainsi obtenu est montré sur la Figure 1.14.

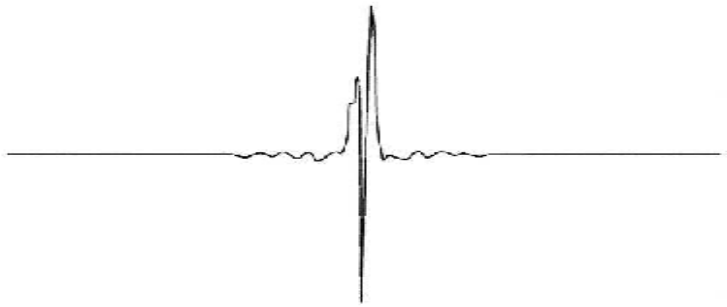


Figure 1.13. Interférogramme en sortie du détecteur

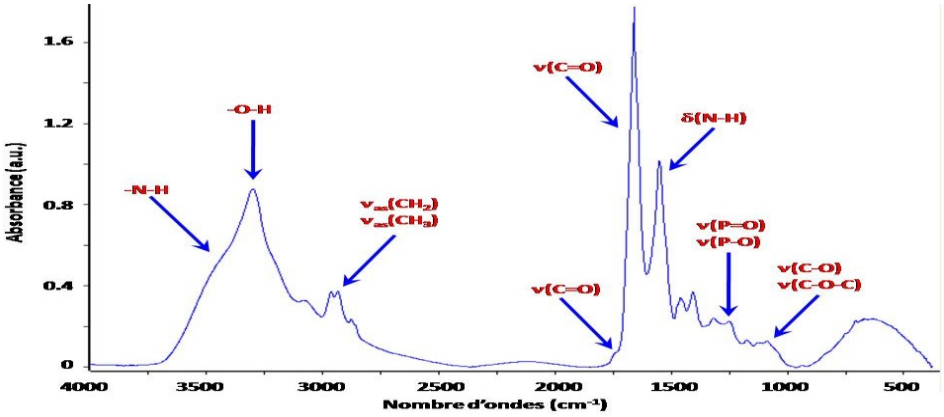


Figure 1.14. Spectres de différentes liaisons organiques

La transformée de Fourier est une procédure mathématique qui permet de décomposer un interférogramme en une somme de sinusoïdes représentant chacune une onde donnée. La fréquence et l'amplitude de ces ondes sont calculées à partir des données de l'interférogramme recueillies par le spectromètre. Cette technique permet de mesurer un spectre entier de l'échantillon.
La spectroscopie par transformée de Fourier a connu un réel intérêt, car elle présente de nombreux avantages qui sont :

- Rapidité de l'analyse : toutes les fréquences de la source infrarouge sont traitées ensemble sans sélection préalable, ce qui permet de capter le spectre entier en moins d'une seconde.
- Simplicité mécanique : la seule partie mobile de l'instrument est le miroir mobile.
- Sensibilité très largement améliorée par rapport aux systèmes dispersifs : la possibilité de réaliser plusieurs acquisitions permet d'améliorer considérablement le rapport signal sur bruit.
- Calibration interne : ces spectromètres sont auto-calibrés et ne nécessitent pas de calibration par l'utilisateur.
- Haute résolution spectrale.
- Peu ou pas de préparation de l'échantillon (qui peut être récupéré après l'analyse).
- Analyse moléculaire qualitative (nature des liaisons chimiques) et quantitative (à partir d'étalons).
- Analyse non destructive.
- La possibilité de traiter numériquement les spectres (déconvolution, interpolation, dérivée, filtrage, lissage, etc.) permet d'atteindre des fréquences plus précises et de diminuer la largeur des bandes d'absorption.

La spectroscopie par transformée de Fourier présente toutefois des inconvénients parmi lesquels le plus importantes sont la sensibilité à l'eau et au CO_2. Néanmoins, comme présenté par la suite, les techniques d'acquisition permettent de limiter leur effet.

I.5.2. Techniques d'acquisition des spectres FTIR

Trois techniques d'acquisition peuvent être utilisées :

- transmission,
- réflexion diffuse,
- réflexion totale (modes ATR).

Ces techniques sont présentées brièvement ci-après.

I.5.2.1. FT-IR par transmission (absorption)

Le procédé de mesure par transmission infrarouge est les plus utilisé [PNW10, KPH08, OIU91] en raison de sa simplicité de mise en œuvre. Son principe de mesure est représenté sur la Figure 1.15. Pour chaque longueur d'onde, la transmittance T est définie par :

$$T = I/I_0 \qquad \text{(eq.1.4)}$$

où I et I_0 représentent respectivement les intensités transmises de l'échantillon et d'une référence. La transmittance est souvent remplacée par son pourcentage (%T) ou par l'absorption A définie par :

$$A = \log (I_0/I) = \log (1/T) \qquad \text{(eq.1.5)}$$

Outre sa simplicité, le principal avantage du procédé par transmission est la possibilité d'utiliser différentes références, par exemple les échantillons peuvent être étudiés par transmission sous forme de films, d'émulsions ou de pastilles KBr (pastilles de bromure de potassium) lorsqu'ils sont des solides, purs ou en solution s'ils sont liquides et enfin à l'état de gaz ou de vapeurs.
La pastille KBr est l'une des techniques utilisées car très pratique et relativement rapide pour des applications en agriculture. Elle fournit des résultats de bonne qualité en mesures qualitatives et semi-quantitatives pour les spectres IR. Cependant, les désavantages de cette technique sont nombreux et en particulier elle nécessite une expérience et un « coup de main » particulier pour la préparation des échantillons. De plus, des problèmes physico-chimiques peuvent intervenir lors du pastillage dû aux effets conjugués de la pression et de l'échauffement : dégradation de certains échantillons fragiles.

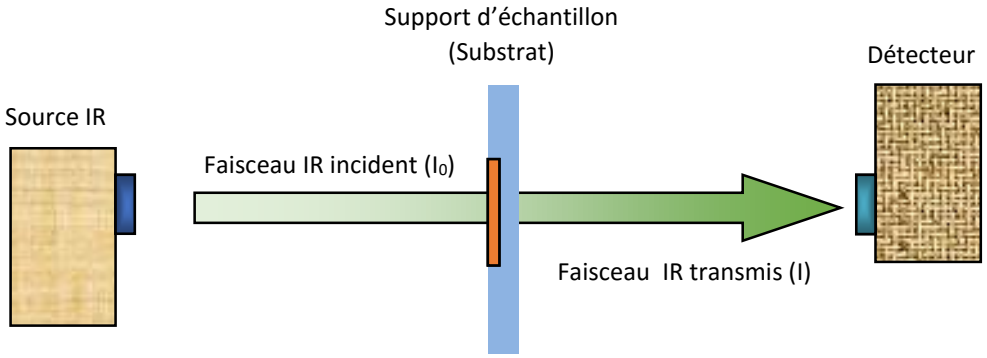


Figure 1.15. Schéma de principe d'une mesure en mode transmission

I.5.2.2. Les procédés par réflexion

a) Réflexion diffuse

La deuxième technique la plus courante d'enregistrement des spectres infrarouges est la réflexion. En mode réflexion diffuse [JA95, NFA93], le faisceau infrarouge incident arrive à la surface de l'échantillon avec un certain angle. Les molécules vont absorber aux longueurs d'onde et le signal va repartir vers le détecteur. La Figure 1.16 illustre le principe de la réflexion diffuse d'un échantillon.

L'inconvénient principal de ce mode est la perte d'une partie du rayonnement incident par diffusion au travers de l'échantillon. Cette technique de réflexion diffuse peut être utilisée pour des échantillons sous forme de poudres ou de solides ayant une surface rugueuse comme du papier, du tissu. Dans la technique de réflectance diffuse, la taille des particules et l'homogénéité des échantillons en poudre joue un rôle important sur la qualité du spectre. Afin d'obtenir un spectre de bonne qualité, l'échantillon doit être broyé extrêmement fin. De plus, ce mode d'analyse est totalement non destructif, aucun contact avec l'objet n'est requis. Nous avons utilisé cette méthode pour analyser nos échantillons de biomasse lignocellulosique parce qu'elle permet une analyse qualitative mais également quantitative de biomasse de végétaux. De plus, elle ne nécessite qu'une faible phase de préparation de l'échantillon contrairement à celle de transmission avec pastillage KBr.

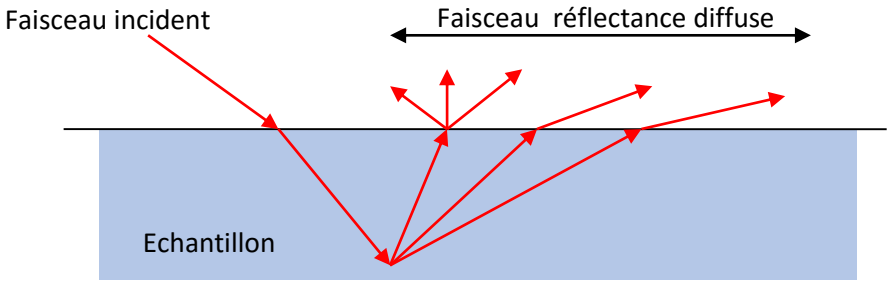


Figure 1.16. Illustration de la réflectance diffuse

b) Réflexion totale atténuée (Attenuated Total Reflectance, ATR)

La spectroscopie infrarouge à réflexion totale atténuée est basée sur un phénomène optique appelé réflexion totale interne. Ce phénomène se produit lorsque la lumière se propage à travers un milieu d'indice de réfraction élevé en interface avec un indice de réfraction plus faible [WMW98, DOR05, DOP08, HOS82, IT00] et uniquement si l'angle d'incidence θ est supérieur à l'angle d'incidence critique θ_c :

$$sin\,\theta_c = n_2/n_1 \qquad \text{(eq.1.6)}$$

où n_1 est l'indice de réfraction du premier milieu et n_2 l'indice du deuxième milieu. Habituellement, n_1 représente l'indice de réfraction du cristal ATR alors que n_2 est l'indice de réfraction de l'échantillon.

Le principe des dispositifs ATR est de faire subir au faisceau optique plusieurs réflexions à l'interface entre l'échantillon et un cristal parallélépipédique transparent en IR, d'indice de réfraction n_1 (cristal), supérieur à celui de l'échantillon d'indice n_2. Le cristal capte la quantité de rayonnement réfléchi par l'échantillon. Le principe de l'ATR est décrit sur la Figure 1.17.

L'ATR apporte une solution à une très grande variété de problèmes d'observation d'échantillons, et permet d'étudier les échantillons difficiles, voire impossibles à analyser par les méthodes spectroscopiques conventionnelles : échantillons opaques en raison de leur trop forte absorption (solution aqueuses, cristaux liquides visqueux, fibres textiles) ou de leur épaisseur.

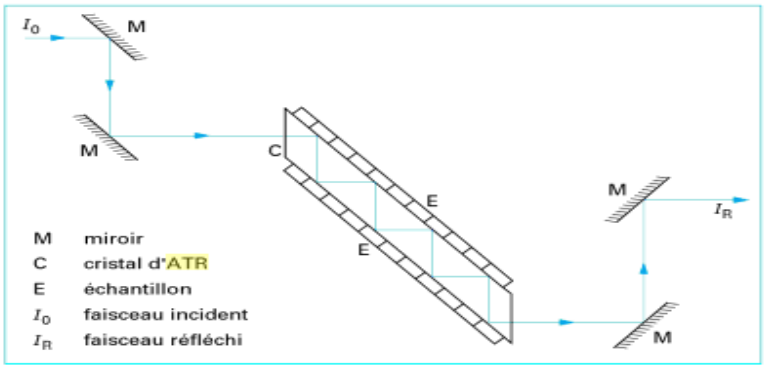


Figure 1.17. Schéma de principe de la réflexion totale atténuée [WMW98]

L'ATR est une méthode largement utilisée dans le domaine de l'agroalimentaire. Cependant cette technique ne permet d'examiner qu'une partie superficielle de l'échantillon. L'un des paramètres les plus importants de l'ATR sera donc l'homogénéité du contact physique entre l'échantillon et le cristal. Il est difficile de reproduire la même qualité de contact d'une expérience à l'autre, ce qui rend problématique l'obtention de résultats quantitatifs en ATR. En pratique, on ne peut malheureusement pas améliorer le contact à l'interface en exerçant une forte pression, car les cristaux utilisés en ATR sont fragiles.

I.6. Echantillons de lignocellulose étudiés, données spectrales et chimiques

Dans cette partie nous nous focalisons sur les types de biomasses lignocellulosiques retenus pour cette étude et les modalités de biodégradation en milieu naturel ou contrôlé ainsi que sur les paramètres d'acquisition de spectres IR et le mode de de préparation des échantillons, caractéristiques essentielles avant l'analyse spectrale en spectroscopie FTIR.

I.6.1. Biomasse lignocellulosique

Dans notre application, nous avons caractérisé deux principaux types des de biomasse lignocellulosiques en regard des domaines d'application ciblés : racines de maïs (décomposition dans le sol) et biomasse aérienne de miscanthus ou de peuplier (dégradation enzymatique pour la production de bioéthanol.

Afin d'étudier le choix des prétraitements de spectres IR et des gammes spectrales les mieux adaptés pour étudier la biomasse lignocellulosique, nous avons analysé des spectres de maïs et de miscanthus avant dégradation par les microorganismes du sol (racines) ou par les enzymes (biomasse aérienne). L'objectif est d'identifier les modalités permettant de discriminer en infrarouge des lignocelluloses en fonction de facteurs génétiques (racine de maïs) ou physiologiques comme la maturité des plantes (miscanthus) :

- Des génotypes de maïs (*Zea maïs*) regroupant des hybrides et 2 lignées consanguines (F2 et F292, lignées parentales) ont été cultivés au champ sur le domaine expérimental de Lusignan l'INRA (France) et ont été récoltés à maturité ensilage [MBB11]. Les racines d'un diamètre de de 2 à 3 mm sont brossées et sélectionnées pour l'application. L'ensemble des données des racines du maïs (16 échantillons) correspond à 3 groupes: hybrides (6 échantillons), F2 (5 échantillons), et F292 (5 échantillons).
- Le miscanthus (*Miscanthus giganteus × L*) a été cultivé à l'INRA Estrées-Mons (France). La biomasse aérienne de miscanthus a été récoltée à deux dates : Octobre 2007 (récolte précoce)

et Février 2008 (récolte tardive) [HRD10]. L'ensemble de données de la biomasse aérienne (12 échantillons) correspond à miscanthus récolté à deux dates de récolte : précoce (6 échantillons) et tardive (6 échantillons).

La composition proximale des racines du maïs et biomasse aérienne de miscanthus a été déterminée par chimie humide au cours de travaux précédents [MBB11, HRD10]. En bref, la fraction soluble a été quantifiée comme la fraction éliminée par une solution de détergent neutre (Van Soest, 1963). Le matériel restant correspond aux parois cellulaires (NDF, Neutral Detergent Fiber) a ensuite été utilisé pour quantifier la cellulose, hémicellulose, lignine Klason et les acides phénoliques liés en ester [HRD10]. Toutes les concentrations chimiques des racines du maïs et biomasse aérienne de miscanthus sont exprimées comme un pourcentage de la matière sèche brute tel que les concentrations totales de chaque échantillon est égale à 100 (Tableau 1.5).

Tableau 1.5. Composition chimique des racines de maïs et miscanthus biomasse (les données sont exprimées en% de la matière sèche brute)

			Soluble	Cellulose	Hemicellulose	Lignin	Phenolic esters
Maïs	Hybrid (H)	mean± SD	17.1±1.9	38.4±1.5	24.6±0.4	16.7±0.6	3.1 ±0.4
		max	20.4	41.0	25.6	17.6	3.8
		min	14.4	36.1	24.0	15.5	2.5
	F2 (n=5)	mean± SD	18.0±3.2	34.2±1.5	25.6±1.3	19.5±1.1	2.6±2.5
		max	22.2	36.6	27.4	20.7	3.2
		min	13.3	32.5	24.0	17.7	1.9
	F292 (n=5)	mean± SD	15.9±1.3	39.0±1.3	26.0±0.5	16.4±0.3	2.6±0.5
		max	17.9	40.8	26.9	16.9	3.0
		min	14.2	36.7	25.5	16.0	1.9
Miscanthus	Récolte précoce (n=6)	mean± SD	17.1±8.1	44.5±6.1	20.9±0.9	16.0±1.6	1.5±0.2
		max	25.0	50.5	22.3	18.6	1.8
		min	18.3	38.8	19.7	14.5	1.3
	Récolte tardive (n=6)	mean± SD	11.6±2.3	46.9±2.7	21.7±1.0	18.0±1.1	1.8±0.2
		max	14.5	49.3	23.6	19.7	2.0
		min	9.5	42.0	20.7	16.5	1.5

Après avoir déterminé les paramètres de l'analyse spectroscopique de maïs et de miscanthus en absence de biodégradation, nous avons développé des méthodes permettant de sélectionner des bandes spectrales IR discriminantes de résidus lignocellulosiques en fonction de leur niveau de dégradation. Dans ce cas, l'étude a porté sur des seules informations spectrales de lignocelluloses après différentes périodes de dégradation dans le sol ou par des enzymes :

- Des échantillons de racines de maïs (essai MACHINET [MBB11]) qui ont été mis à décomposer dans un sol puis enlevé manuellement du sol à différentes dates et lavés, séchés puis broyés avant la prise des spectres IR. Quatre génotypes de racines de maïs différents (F2 et F292, F2bm1 et F292bm3) ont été incubés dans un sol agricole pendant 112 jours à 15 ° C et maintenus à une humidité de -80kPa. Ces racines du maïs qui ont été retirées du sol et analysées à cinq périodes de biodégradation dans le sol : t_1=0, t_2=14, t_3=36, t_4=57, t_5=112 jours.
- Des échantillons de racines de maïs sont mélangés à une matrice sol et sans tri ultérieur (essai AMIN [Amin12]). Ainsi les spectres IR ont été réalisés dans ce cas sur un mélange de sol et racines de maïs à différentes dates de décomposition. Ces échantillons comportent quatre

traitements : trois racines de maïs de génotypes différents (F2, F2bm1, F292bm3) et un sol contrôle sans ajout de maïs nommé « C ». Les périodes de biodégradation dans le sol sont : t_1=0, t_2=27, t_3=58, t_4=84, t_5=120, t_6=478 jours. Les trois génotypes de maïs ont été mélangés à deux types de sol se différentiant, entre autre par leur pH (4.9 et 6.7). Une modalité supplémentaire avec et sans azote a été créé et nommée (+ (ajout) ; - (non ajout)). Chaque traitement comporte trois réplicats. La Figure 1.18 résume le design de l'expérience.

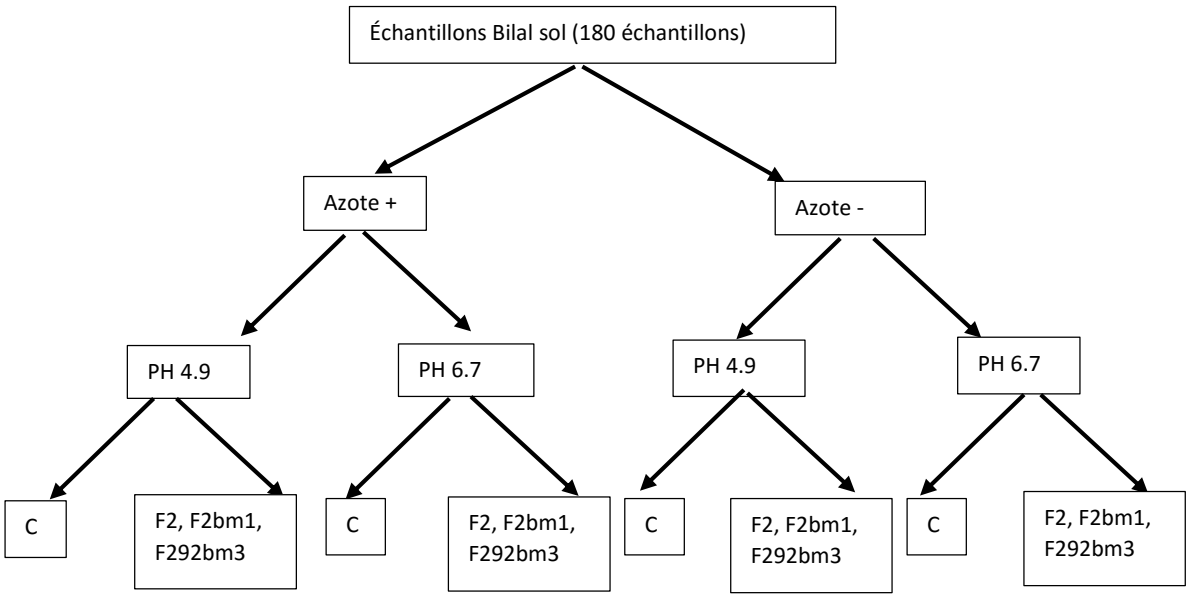


Figure 1.18. Les échantillons (Bilal sol) : racines de maïs

Au cours de l'expérimentation AMIN, différentes caractéristiques ont été déterminées à chaque période d'incubation [Amin12] :

- Activités enzymatiques associées à la biodégradation des racines dans le sol, exprimées en µmol/h/g (information biologique) : β-glucosidase (BG), N-acétyl glucosaminidase (NAG), l'activité phosphatase (PHOS), Netperoxydase (PER) et la phénol oxydase (PO).
- Deux variables supplémentaires et nécessaires pour caractériser le degré de décomposition complétent ce jeu de données. Il s'agit de la quantité de carbone minéralisé en gaz carbonique (information chimique) : C minéralisé cumulé (mgC-CO_2/kg sol) et vitesses de C minéralisée (mgC-CO_2/kg sol/jours).

Le Tableau 1.6 regroupe l'ensemble des les détails et les informations des activités enzymatiques.

- Des échantillons de miscanthus consistent en biomasse aérienne totale soumise à des prétraitements physico-chimiques (explosion à la vapeur) variant par leur degré de sévérité (température, pression,…)) selon des procédés développés par Procethol 2G, France (procédés confidentiels) puis lavés avec de l'eau distillée pour enlever les inhibiteurs d'enzymes potentiels. Les échantillons de miscanthus ont été incubés en présence de cocktail cellulasique commercial (Novozymes, Celltech, 30mg protéines/ g de cellulose) à 40 ° C dans du tampon citrate 50 mM à pH 5, et ont été recueillis à différentes périodes d'hydrolyse enzymatique : t_1=0, t_2=2, t_3=4, t_4=8, t_5=24, t_6=48, t_7=72 heures. Les durées d'incubation ont été identifiées par l'analyse du glucose libéré dans le milieu réactionnel (UMR FARE).

Tableau 1.6. Tableau récapitulatif des informations chimiques et biologiques

Activités enzymatiques	l'activité de glucosides beta BG (µmol/h/g)	l'activité de N-acétyl glucosaminidase (NAG)	phosphore extractible totale PHOS	Net peroxydase	Phénol oxydase	C minéralisé cumulé (mgC-CO_2/kg sol)	les vitesses de C minéralisée (mgC-CO_2/kg sol/jours)
Type d'échantillons	C, F2, F292bm3, F2bm1	C, F2, F292bm3, F2bm1	C, F2, F292bm3, F2bm1	C, F2, F292bm3, F2bm1	C, F2, F292bm3, F2bm1	C, F2, F292bm3, F2bm1	C, F2, F292bm3, F2bm1
pH du sol (deux sols différents)	6,7 et 4,9	6,7 et 4,9	6,7 et 4,9	6,7 et 4,9	6,7 et 4,9	6,7 et 4,9	6,7 et 4,9
La période T en jours	0, 27, 58, 84, 120, 478	0, 27, 58, 84, 120, 478	0, 27, 58, 84, 120, 478	0, 27, 58, 84, 120, 478	0, 27, 58, 84, 120, 478	27, 58, 84, 120, 478	27, 58, 84, 120, 478

- Des échantillons de peuplier consistent en biomasse aérienne totale soumise à des prétraitements physico-chimiques (explosion à la vapeur) variant par leur degré de sévérité (température, pression, …) selon des procédés développés par Procethol 2G, France (procédés confidentiels) puis lavés avec de l'eau distillée pour enlever les inhibiteurs d'enzymes potentiels. Les échantillons de peuplier ont été soumis à une hydrolyse enzymatique comme décrit précédemment.

I.6.2. Préparation des échantillons

a) Le séchage

L'eau présente des bandes spectrales relativement importantes et sa présence peut perturber l'analyse puisque les échantillons peuvent avoir des humidités relatives différentes lors des acquisitions spectrales. Une solution est d'analyser les échantillons tels qu'ils sont et de ne pas prendre en compte les bandes spectrales liées à l'eau dans les analyses ultérieures. Néanmoins, comme l'eau a une signature spectrale étendue, surtout en NIR, le risque est d'éliminer des informations spectrales intéressantes. Une autre solution est de sécher les échantillons de manière à avoir la même humidité relative. Le séchage est une opération permettant d'éliminer l'eau stockée dans l'échantillon. Au cours du séchage, l'humidité qui reste dans l'échantillon est vaporisée. La vitesse à laquelle il est possible de sécher un échantillon dépend de la vitesse à laquelle la vapeur d'eau diffuse de l'intérieur de l'échantillon vers l'extérieur. Les pièces peuvent être séchées soit statiquement à l'air libre ou dans des étuves, soit dans des séchoirs, soit par rayonnement (infrarouge) ou par convection (échangeurs). Les séchoirs et les étuves sont donc les plus recommandés. Ils sont plus rapides et ne nécessitent pas de stockage.

b) Le broyage

Le broyage est une étape critique de la préparation de l'échantillon pour l'analyse spectrale car il permet de réduire à l'état de poudre les échantillons préalablement séchés, donc d'homogénéiser l'information physico-chimique de l'échantillon, condition importante quant à la qualité des analyses infrarouges. Le choix d'un type de broyeur dépend de la nature du végétal à broyer (grains, tiges,

racines). Les modes de broyages proposés ainsi que les dimensions des cuves de broyeurs devront être adaptés. Il existe une grande variété de broyeurs : à couteaux, à marteaux, à fléaux, oscillo-vibrants (à bille), planétaires, centrifuges [Mar95]. Les broyeurs oscillo-vibrant (à bille) et centrifuge donnent de très bonnes performances puisque l'échantillon est en contact avec un unique matériau. Nous avons choisi ces deux types de broyeurs pour nos échantillons de biomasse lignocellulosique.

Le broyeur centrifuge est un appareil qui se compose d'un axe de rotation cylindrique dont les bords comportent des griffes en acier, autour desquelles on vient mettre en contact une grille cylindrique fixe dont le diamètre des trous correspond à notre choix de broyage. On place ensuite l'échantillon au centre de l'axe de rotation, la force centrifuge plaque les échantillons contre la grille (qui ne tourne pas puisqu'elle est fixe). Les échantillons sont donc écrasés entre la grille et les griffes en acier jusqu'à être suffisamment fin pour passer dans le diamètre des trous de la grille. Au-delà de la grille un bol permet de récupérer les échantillons broyés. La Figure 1.19 (a) représente le broyeur centrifuge.

Le deuxième type de broyeur que nous avons utilisé est le broyeur à bille qui est un appareil formé d'un axe animé horizontalement. L'échantillon est placé dans un réceptacle accompagné d'une bille et fermé par un couvercle. L'ensemble est fixé à l'axe du broyeur par un mors. Le mouvement horizontal de l'axe entraine l'écrasement de l'échantillon entre la bille et le réceptacle. Il est possible de choisir la durée du broyage de manière empirique en fonction de la résistance et du type d'échantillon à écraser. Son principal inconvénient est qu'il n'est pas possible de contrôler la finesse finale du broyage. La Figure 1.19 (b) représente le broyeur à bille.

Figure 1.19. (a) Broyeur centrifuge, (b) Broyeur à bille

Au cours des acquisitions au laboratoire de l'INRA de Reims sur les différents échantillons, les appareils mentionnés dans le Tableau 1.7 ont été utilisés avec les paramètres indiqués.

Tableau 1.7. Matériels utilisés durant l'expérimentation pour nos échantillons.

Echantillons\Appareils	Température et séchage	Broyage
Echantillons Bilal sol	Etuve à 50°C pendant 5 jours	Broyeur à bille (Durée de broyage 2 minutes à 80% d'amplitude)
Echantillons de racines de maïs (essai MACHINET [MBB11])	Etuve à 50°C ou 80°C	Broyeur centrifuge (à 80µm)
Echantillons de biomasse aérienne totale soumise à des prétraitements physico-chimiques (miscanthus et peuplier)	Etuve à 50°C pendant 2 jours	Broyeur à bille

c) Le spectromètre

Pour l'acquisition de nos spectres MIR et NIR nous avons utilisé le spectromètre « Thermo Scientific Nicolet iS50 FT-IR » qui offre des performances idéales dans le cadre de l'observation et l'identification de marqueurs sur des échantillons de biomasse lignocellulosique. La Figure 1.20 représente le spectromètre « Thermo Scientific Nicolet iS50 FT-IR ».

Figure 1.20. Spectromètre "Thermo Scientific Nicolet iS50 FT-IR"

Une fois les échantillons préparés, il est alors possible d'acquérir des spectres MIR et NIR. Pour cela, un certain nombre de paramètres doivent être choisis et les plus importants sont discutés ci-après.

I.6.3. Paramètres d'acquisition des spectres FTIR

Le choix des paramètres d'acquisition est primordial et indispensable au bon au fonctionnement d'un spectromètre par transformée de Fourier pour l'acquisition de spectres dans les meilleures conditions.

a) Nombre de scans

Le nombre de scans est l'un des paramètres expérimentaux les plus importants. L'appareil peut effectuer n fois l'enregistrement du spectre selon la volonté de l'utilisateur puis le logiciel calcule la moyenne des n spectres enregistrés. L'équation 1.6 indique que le rapport de signal sur bruit dans un spectre est proportionnel à la racine carrée du nombre de scans. Le rapport (S / N) après n scans est donné par [Gar96] :

$$\frac{S}{N} = n^{1/2} \frac{I_S}{I_N} \qquad \text{(eq.1.7)}$$

où I_S et I_N sont les intensités du signal et le bruit, respectivement.

Le nombre de scans affecte donc le rapport signal sur bruit (S / N) du spectre enregistré. Plus le nombre de scans n est élevé plus le rapport signal sur bruit est meilleur. Mais, nous ne pouvons pas choisir un nombre de scan élevé, à cause principalement de la possible modification des informations spectrales, notamment à cause de l'humidification de l'échantillon. Pour cela, nous avons effectué un test afin d'étudier l'effet du nombre de scan sur les spectres FTIR. Nous avons fixé la résolution à 4 cm^{-1} dans le spectromètre FTIR, ensuite nous avons fait varier le nombre de scan de 16 à 64.

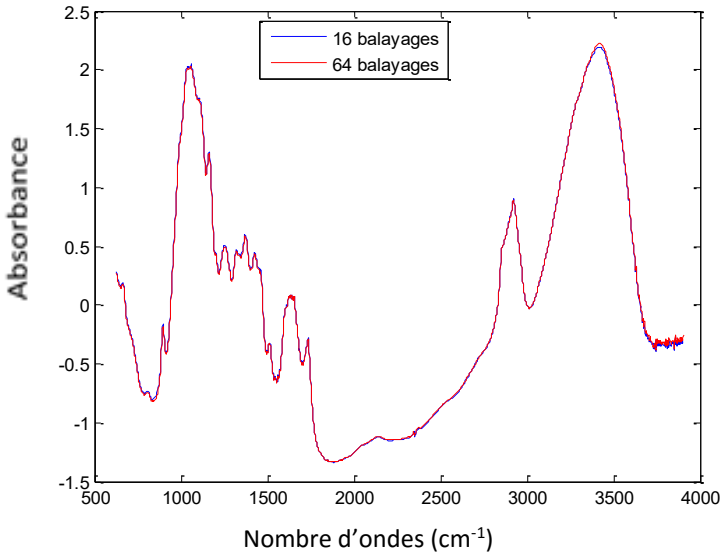


Figure 1.21. Comparaison des spectres pris avec 16 et 64 scans

Comme nous pouvons l'observer sur la Figure 1.21, les spectres acquis 64 fois puis moyennés représentent des spectres moins bruités même s'ils nécessitent 4 fois plus de temps pour obtenir un résultat. Cette étude demande à être approfondie afin de mieux comprendre l'influence de l'humidification sur les échantillons analysés et/ou le choix d'échantillons non séchés, mais celle-ci sort des objectifs de cette thèse.

En conclusion, pour toutes les acquisitions effectuées sur les différentes biomasses lignocellulosiques, nous avons choisi le paramètre suivant 64 fois pour le nombre de scans.

b) Résolution spectrale

La résolution est la capacité d'un instrument de mesure à séparer deux informations proches. En spectrométrie FTIR, la résolution correspond à la finesse des pics. Elle peut être réglée à des valeurs allant de 64 cm^{-1} en passant par 16 cm^{-1}, 8 cm^{-1}, 4 cm^{-1}, jusqu'à 2 cm^{-1}. Plus la valeur de résolution est petite, plus les spectres sont bien définis. Cependant, si la résolution est faible, l'intensité de la lumière entrant dans le détecteur est réduite et la quantité relative de bruit dans le spectre augmente. Par conséquent, il n'est pas souhaitable de réduire la résolution plus que nécessaire [DBL00]. Nous avons réalisé plusieurs tests sur différents échantillons afin de choisir la valeur de résolution optimale. Lors de ces tests, nous avons fixé le nombre de scan (voir définition ci-dessous) à 16 et nous avons fixé la résolution à 4 cm^{-1} puis à 8 cm^{-1}. La Figure 1.22 représente l'influence de la variation de la résolution spectrale pour des spectres FTIR acquis sur un échantillon de maïs.

Les spectres acquis avec une résolution de 4 cm^{-1} représentent des spectres plus riches en informations, les pics sont mieux séparés. Par contre on remarque l'influence du bruit sur ces spectres. Le choix d'une résolution de 8 cm^{-1} permet d'obtenir des spectres bien lissés, cependant nous pouvons remarquer une perte d'informations sur quelques nombres d'ondes.

En étudiant différents échantillons, nous avons convenu qu'il était souhaitable de fixer une résolution spectrale plus faible, de 4 cm^{-1} et d'augmenter le nombre de scans afin de réduire le bruit.

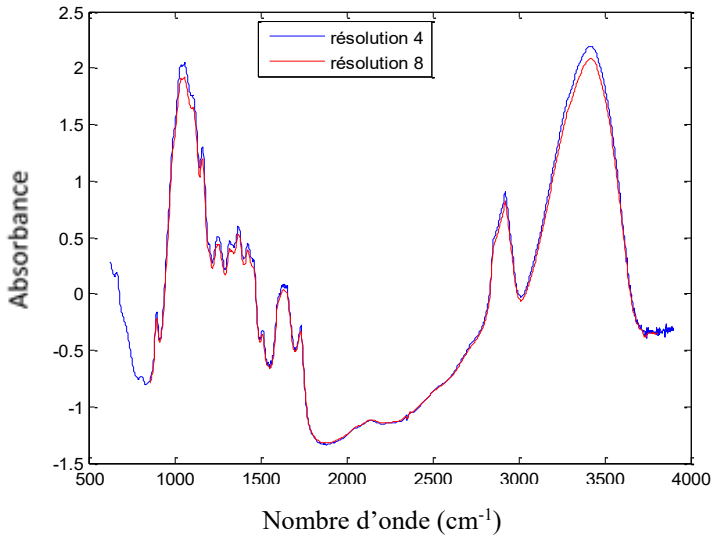


Figure 1.22. Comparaison entre deux spectres MIR pris avec une résolution 4 cm^{-1} et 8 cm^{-1}.

c) Fonction d'apodisation

Les spectres calculés peuvent faire apparaître des lobes à la base des « pics » qui constituent alors des « artéfacts numériques ». Pour limiter et réduire leur effet, nous pouvons appliquer une fonction de correction appelée fonction d'apodisation (ou également fonction de pondération) à l'interférogramme. Cette fonction multiplie l'interférogramme, ce qui a pour conséquence une réduction des lobes latéraux (amplitude, forme) dans un spectre, mais également un élargissement des « pics », donc une dégradation de la résolution spectrale. Cette technique étant très connue en analyse de Fourier, nous n'allons pas la détailler ici. Les divers types de fonctions d'apodisation existantes dans les logiciels d'acquisition fournis avec les spectromètres sont : boxcar, triangle, Happ-Genzel, Beer-Norton, Hanning et Bessel [NB76].

Pour comparer les performances des fonctions d'apodisation, nous nous sommes appuyés sur l'étude réalisée par Bretzlaff et Bahder [BB86] qui est basée sur le calcul : (i) des artefacts apparaissant dans les spectres obtenus par différence ; (ii) de la variation de l'absorbance apparente au sommet d'un pic d'absorption en fonction de l'absorbance vraie ; (iii) de la mesure intégrée de l'erreur due à ces artefacts pour une bande d'absorption à profil lorentzien. Cette étude nous a permis de choisir la fonction d'apodisation Happ-Genzel, qui est par ailleurs couramment employée pour différentes applications, puisqu'elle donnée le meilleur compromis entre la résolution d'un spectre et les effets des lobes latéraux.

I.7. Représentation mathématique des spectres FTIR

Après l'obtention de spectres infrarouges, il est nécessaire de formaliser les spectres acquis. Ceux-ci sont représentés sous la forme de vecteurs. Ainsi, chaque spectre IR est enregistré sur un vecteur de nombres d'ondes noté :

$$\underline{y}^{MIR}=[y_1^{MIR} \dots y_P^{MIR}]^T \in \mathbb{R}^P \quad , \qquad \text{(eq.1.8)}$$

$$\underline{y}^{NIR}=[y_1^{NIR} \dots y_Q^{NIR}]^T \in \mathbb{R}^Q \quad , \qquad \text{(eq.1.9)}$$

avec T désignant la transposée d'un vecteur, P et Q étant respectivement les dimensions des vecteurs de nombres d'ondes associés aux spectres enregistrés par le spectromètre FTIR en prenant en compte

les paramètres de résolution spectrale et le nombre de scans. $\underline{y}^{MIR}$ est le vecteur de nombres d'ondes associé au spectre MIR et $\underline{y}^{NIR}$ est le vecteur de nombres d'ondes associé au spectre NIR.
Le spectre enregistré sur le jième échantillon peut alors être écrit de la façon suivante :

$$\underline{x}_j^{MIR}(\underline{y}^{MIR}) = [x_{j1}^{MIR} \dots x_{jP}^{MIR}]^T \in \mathbb{R}^P , \quad \text{(eq.1.10)}$$

$$\underline{x}_j^{NIR}(\underline{y}^{NIR}) = [x_{j1}^{NIR} \dots x_{jQ}^{NIR}]^T \in \mathbb{R}^Q , \quad \text{(eq.1.11)}$$

Les spectres MIR et NIR enregistrés sur tous les échantillons J, peuvent être écrits sous forme de matrices de données :

$$X^{MIR}(\underline{y}^{MIR}) = \left[\underline{x}_1^{MIR}(\underline{y}^{MIR}) \quad \dots \underline{x}_j^{MIR}\left(\underline{y}^{MIR}\right) \dots \quad \underline{x}_J^{MIR}(\underline{y}^{MIR})\right] \in M_{P,J}(\mathbb{R}) \quad \text{(eq.1.12)}$$

$$X^{NIR}(\underline{y}^{NIR}) = \left[\underline{x}_1^{NIR}(\underline{y}^{NIR}) \quad \dots \underline{x}_j^{NIR}\left(\underline{y}^{NIR}\right) \dots \quad \underline{x}_J^{NIR}(\underline{y}^{NIR})\right] \in M_{Q,J}(\mathbb{R}) \quad \text{(eq.1.13)}$$

Il est également possible de formaliser les spectres combinés MIR et NIR pour le produit extérieur et par concaténation.

Le produit extérieur entre deux spectres $\underline{x}_j^{MIR}(\underline{y}^{MIR})$ et $\underline{x}_j^{NIR}(\underline{y}^{NIR})$ indexé par des paires de nombres d'ondes (MIR, NIR) est défini par la relation suivante :

$$\underline{y}^{MIR\otimes NIR} = [(y_1^{MIR}, y_1^{NIR})(y_1^{MIR}, y_2^{NIR}) \dots \ (y_P^{MIR}, y_Q^{NIR})]^T \in \mathbb{R}^{PQ} \quad \text{(eq.1.14)}$$

donne une matrice des informations spectrales exprimée par la relation suivante :

$$\boldsymbol{x}_j^{MIR\otimes NIR} = \underline{x}_j^{MIR}(\underline{y}^{MIR}) \otimes \underline{x}_j^{NIR}(\underline{y}^{NIR}) = \begin{bmatrix} x_{j1}^{MIR}[x_{j1}^{NIR} \dots x_{jQ}^{NIR}]^T \\ x_{j2}^{MIR}[x_{j1}^{NIR} \dots x_{jQ}^{NIR}]^T \\ \vdots \\ x_{jP}^{MIR}[x_{j1}^{NIR} \dots x_{jQ}^{NIR}]^T \end{bmatrix} \in M_{P,Q}(\mathbb{R}) \quad \text{(eq.1.15)}$$

Chaque matrice $\mathbf{x}_j^{MIR\otimes NIR}$ est dépliée pour produire un vecteur de grande dimension contenant toutes les combinaisons possibles de produits entre les deux ensembles de données.

$$\underline{\mathrm{x}}_j^{MIR\otimes NIR}(\underline{\mathrm{y}}^{MIR\otimes NIR}) = \left[x_{j1}^{MIR}[x_{j1}^{NIR} \dots x_{jQ}^{NIR}]^T \dots \ x_{jP}^{MIR}[x_{j1}^{NIR} \dots x_{jQ}^{NIR}]^T\right] \quad \text{(eq.1.16)}$$

Pour la méthode usuelle de combinaison, qui est la concaténation des spectres MIR et NIR, l'expression du spectre combiné pour l'échantillon j est donnée par la relation suivante :

$$\underline{x}_j^{MIR-NIR}(\underline{y}^{MIR-NIR}) = \left[[x_{j1}^{MIR} \dots x_{jP}^{MIR}]^T \ [x_{j1}^{NIR} \dots x_{jQ}^{NIR}]^T\right]^T \in \mathbb{R}^{\mathrm{P+Q}} \quad \text{(eq.1.17)}$$

Ce vecteur étant lui-même associé à la concaténation des nombres d'ondes MIR-NIR :

$$\underline{y}^{MIR-NIR} = \left[\underline{y}^{MIR^T} \ \underline{y}^{NIR^T}\right]^T \in \mathbb{R}^{P+Q} \quad \text{(eq.1.18)}$$

I.8. Conclusion

Nous avons présenté une étude bibliographique sur les différentes ressources de biomasses végétales, les processus de biodégradations, le rôle joué par ces éléments dans les sols et comment ces ressources peuvent s'avérer utiles pour la production de biocarburant. Nous avons détaillé la composition de la biomasse végétale en différentiant les fractions solubles et pariétales en insistant sur le rôle important des rapports de concentrations des éléments présents (cellulose, hémicellulose, lignine) dans cette dernière pour l'étude du processus de dégradation.

Dans une seconde partie, nous avons présenté les différentes techniques spectroscopiques particulièrement intéressantes, car non invasives et susceptibles de caractériser ce processus. Nous nous sommes arrêtés sur la spectroscopie infrarouge et notamment la spectroscopie moyenne et proche infrarouge. Comme les spectres MIR et NIR mettent en évidence des informations complémentaires, nous avons exposé comment ces informations peuvent être combinés. La dernière partie a été consacrée à la description échantillons de biomasse lignocellulosique utilisés, les méthodes de préparation, les paramètres d'acquisition des spectres FT-IR et leurs effets.

Chapitre 2 : Prétraitements de spectres et choix des gammes spectrales à l'aide de méthodes mathématiques de classification

II.1. Introduction

L'objectif de ce chapitre est d'identifier les prétraitements les plus adaptés et de choisir les gammes spectrales les plus pertinentes en s'appuyant uniquement sur l'information spectrale. Pour cela, nous allons nous concentrer sur des algorithmes de classification non supervisés qui permettent d'identifier ce type d'information en optimisant la classification de la biomasse lignocellulosique. En effet, étant donné que plusieurs spectres ont été acquis sur la même biomasse et que ce processus est répété à différentes périodes (par exemple récolte précoce / tardive), il est logique de considérer que le prétraitement le plus adapté pour la gamme spectrale la plus pertinente fournisse les classes les plus compactes et les plus séparées. Cette étude permet également de comparer la pertinence des informations spectrales par rapport aux informations chimiques qu'on peut mesurer sur les échantillons lignocellulosiques.

La section suivante présente un exposé non exhaustif des méthodes de prétraitement. Les bases théoriques des principales méthodes de prétraitement des spectres IR sont présentées. Nous essayons d'exposer à la fois les formulations mathématiques intrinsèques de ces méthodes ainsi que leurs applications pratiques à l'aide d'exemples, puis nous faisons une étude approfondie de ces méthodes appliquées aux spectres MIR et NIR enregistrés sur des données de biomasse lignocellulosique. Les interprétations chimiques et les caractéristiques physiques des différentes gammes spectrales MIR et NIR sont décrites dans la section suivante.

Nous introduisons ensuite le principe de fonctionnement des méthodes de classifications non supervisées, notamment la méthode de classification Fuzzy C-Means (FCM) et ses paramètres. Nous présentons les extensions de cette méthode basées sur la distance de Mahalanobis qui permettent de prendre en compte la répartition non sphérique des données, point faible de la FCM, notamment les algorithmes de Gustafson-Kessel, Gath-Geva, FCM-M, FCM-CM et FCM-SM. Nous montrons que les problèmes rencontrés par l'application de ces méthodes sont liés au type de données spectrales de grande dimension. Cette observation nous amène à proposer une nouvelle extension de l'algorithme FCM basée sur un facteur de covariance. Nous comparons les performances de l'algorithme proposé par rapport aux algorithmes existants sur des données simulées et réelles couramment utilisées dans divers tests, puis nous montrons son efficacité sur des spectres MIR et NIR enregistrés sur la biomasse lignocellulosique. Etant donné que nous disposons d'un nombre limité d'échantillons, nous proposons d'utiliser un ré-échantillonnage « Bootstrap » qui permet d'améliorer la précision de classification dans le cas d'un nombre réduit de spectres.

L'application de ces méthodes sur des spectres enregistrés sur la biomasse lignocellulosique est présentée dans la section suivante et le choix des prétraitements les plus adaptés et des gammes spectrales les plus pertinentes sont discutés. Une comparaison avec les résultats obtenus en appliquant la même méthodologie mais en utilisant l'information chimique est également discutée. La dernière section présente la conclusion et des perspectives d'amélioration.

II.2. Méthodes mathématiques de prétraitement

Les spectres IR sont affectés par : la taille des particules de l'échantillon qui dépend de la finesse de broyage (voir chapitre1), des perturbations rencontrées sur le chemin optique, de l'humidité de

l'échantillon, etc. Pour éliminer ou diminuer en partie l'influence de ces interférences, il est indispensable d'appliquer des prétraitements mathématiques sur ces spectres. Les méthodes de prétraitement se divisent en deux grandes catégories : correction de dispersion (scatter-correction) et dérivation spectrale [RVE09].
Dans cette partie, nous présentons les aspects théoriques des méthodes de prétraitements qui sont les plus couramment utilisées dans le domaine spectroscopique infrarouge.

II.2.1. Normalisation (SNV)

La normalisation (Standard Normal Variate en anglais, SNV) est l'une des méthodes de prétraitement les plus utilisées en spectroscopie infrarouge pour corriger des problèmes de dispersion [HL10, PW15]. Cette méthode consiste à corriger chaque spectre individuellement. L'intégralité du spectre est centrée sur sa moyenne et l'ensemble de l'intensité est réajusté grâce à l'écart type du spectre.
Dans une application de spectroscopie IR, nous pouvons construire des spectres corrigés par SNV pour chaque spectre d'origine $\underline{x}_j^{case} \in \mathbb{R}^O$ tel que "case" = MIR ou NIR et O = P ou Q pour tout j = 1…J. Le spectre corrigé par SNV est donné par la relation suivante :

$$\underline{x}_{SNV,j}^{case} = \frac{(\underline{x}_j^{case} - \overline{\underline{x}_j^{case}})}{\sigma(\underline{x}_j^{case})} \in \mathbb{R}^O \quad j=1…J \qquad \text{(eq.2.1)}$$

où $\overline{\underline{x}_j^{case}} = \frac{\sum_{i=1}^{O} \underline{x}_{ij}^{case}}{O}$ est la valeur moyenne du spectre $\underline{x}_j^{case}$ et $\sigma(\underline{x}_j^{case}) = \sqrt{\frac{\sum_{i=1}^{O} (\underline{x}_{ij}^{case} - \overline{\underline{x}_j^{case}})^2}{O-1}}$ son écart type [RVE09]. La Figure 2.1 représente le résultat obtenu avec la méthode de prétraitement SNV sur un spectre simulé x_j.

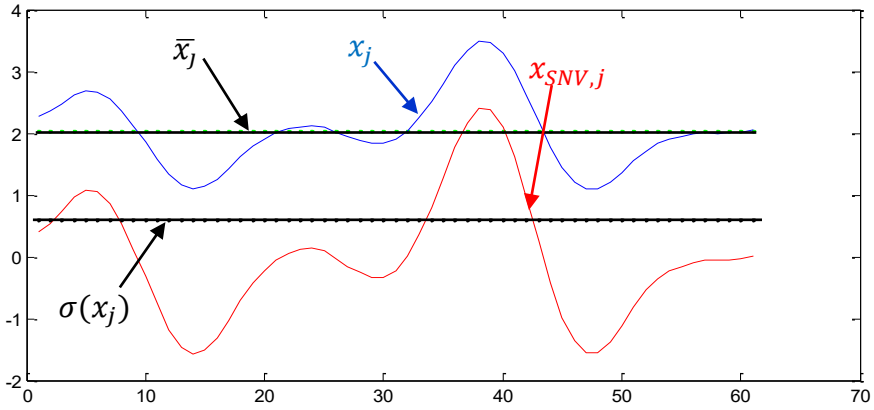


Figure 2.1. Exemple de spectre simulé (courbe en bleu) corrigé par la méthode SNV (courbe en rouge)

La correction SNV est très utile car elle permet de réduire les effets de diffusion. En effet, la diffusion peut être responsable d'une réduction sensible de l'intensité des spectres pour certaines longueurs d'ondes. Celle-ci peut varier considérablement entre deux échantillons (variation de la finesse de broyage, compacité, etc.) [LHV04, BDL99]. Cette correction peut être utilisée aussi comme un moyen de réduire la multi-colinéarité, les déviations verticales (effets additifs de la diffusion) et les variations de pentes des spectres par rapport à un spectre de référence (effets multiplicatifs de la diffusion) [BDL89].

II.2.2. Correction de la ligne de base (LB)

Dans les spectres IR, la ligne de base pourrait être le résultat de différents phénomènes. Le principal est l'effet de diffusion (méthode de transmission, méthode de réflexion, etc.). Par exemple, si l'échantillon a une surface rugueuse ou contient de la poudre de composé inorganique, la diffusion de la lumière infrarouge à la surface de l'échantillon ou à l'intérieur sera plus importante pour les longueurs d'ondes les plus courtes. Par conséquent, un spectre IR touché par un phénomène de diffusion va voir sa ligne de base diminuer en amplitude (si on affiche en transmission %T) vers les courtes longueurs d'onde (grandes valeurs de nombres d'ondes). Lors de la mesure du spectre d'un échantillon de transmission avec une surface lisse, la ligne de base peut aussi apparaître comme une forme d'onde sinusoïdale régulière. Ce sont des franges d'interférences dues à de multiples réflexions de lumière dans l'échantillon. L'estimation et ensuite la soustraction de cette ligne de base permet d'éliminer l'influence de ces phénomènes perturbateurs sur le potentiel informatif des spectres acquis et donne accès à un signal plus interprétable [MCB05]. Dans cette étude, la ligne de base a été modélisée et ensuite estimée par un polynôme de degré d. Les coefficients du polynôme ont été évalués par la minimisation d'une fonction de coût non quadratique [MCB05]. Le choix du degré du polynôme dépend des signaux analysés. Le spectre corrigé $\underline{x}_{LB,j}^{case}$ par cette méthode est donné par :

$$\underline{x}_{LB,j}^{case} = \underline{x}_{j}^{case} - P_j \text{ , } j=1...J \qquad \text{(eq.2.2)}$$

où P_j est le polynôme d'ordre « d » qui a été estimé par une méthode de minimisation semi-quadratique, $\underline{x}_{j}^{case}$ le spectre IR avec case = MIR ou NIR. Pour comprendre son fonctionnement, nous affichons sur la Figure 2.2 le même spectre simulé x_j, le polynôme P_j d'ordre « d » ajusté au spectre x_j par la minimisation d'une fonction de coût non quadratique et le spectre corrigé $x_{LB,j}$.

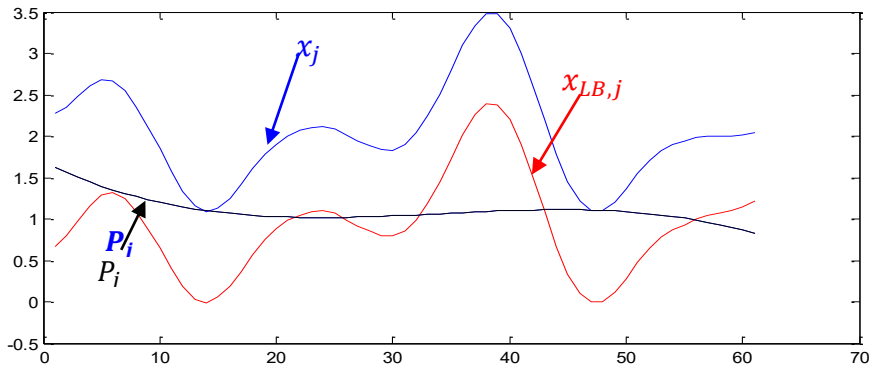


Figure 2.2. Exemple de spectre simulé (en bleu) et corrigé par la méthode LB (en rouge).

II.2.3. Dérivation

a) Principe de dérivation

La dérivation de spectres est un prétraitement très couramment appliqué en spectroscopie IR. La dérivée permet de séparer des pics ou des bandes d'absorption [OMS92, LPD86] qui se chevauchent alors que le bruit de fond basse-fréquence des spectres devient proche de zéro. La ligne de base se trouve de ce fait corrigée. Malheureusement, l'opération de différenciation amplifie le bruit haute-fréquence et augmente la complexité des spectres [SG64, LAB92],

Pour illustrer l'effet de la dérivation, supposons qu'un spectre déformé soit la somme d'un « signal utile » $\underline{x}_{UTIL,j}^{case}$ et d'une ligne de base formée par une droite [Ber05]. Un spectre à corriger $\underline{x}_j^{case}$ est alors représenté par une somme :

$$\underline{x}_j^{case} = b_0 + b_1 \underline{y}^{case} + \underline{x}_{UTIL,j}^{case} \qquad \text{(eq.2.3)}$$

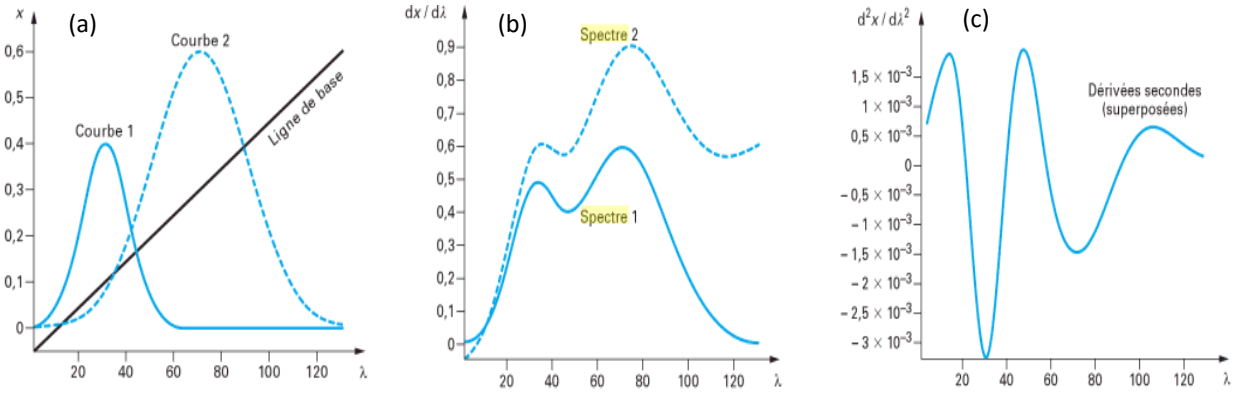
avec $\underline{y}^{\text{case}}$ le vecteur des nombres d'ondes tel que case = MIR ou NIR. La dérivation première de $\underline{x}_j^{case}$ en fonction de $\underline{y}^{\text{case}}$ donne :

$$\frac{d\,\underline{x}_j^{case}}{d\underline{y}^{case}} = b_1 + \frac{d\,\underline{x}_{UTIL,j}^{case}}{d\underline{y}^{case}} \qquad \text{(eq.2.4)}$$

Le terme b_0, représentatif de la « hauteur » de la ligne de base a disparu. De même, la dérivation seconde permet d'obtenir :

$$\frac{d^2\,\underline{x}_j^{case}}{d^2\underline{y}^{case}} = \frac{d^2\,\underline{x}_{UTIL,j}^{case}}{d^2\underline{y}^{case}} \qquad \text{(eq.2.5)}$$

Cette relation exprime le fait que, quelle que soit la ligne de base, fonction linéaire des nombres d'ondes, qui perturbe la mesure spectrale, son effet disparaît lorsque les spectres sont mis sous la forme de leurs dérivés secondes.
A titre d'illustration (Figure 2.3), on suppose que le spectre est composé de la somme de deux courbes « en cloche » formant le signal utile et d'une ligne de base linéaire dont les paramètres b_0 et b_1 peuvent varier de manière incontrôlée. Deux acquisitions du spectre peuvent donner donc deux signaux différents. En revanche, les spectres en dérivées secondes sont exactement superposés.



(a) On suppose que le spectre est la somme des courbes 1 et 2 et d'une ligne de base.

(b) L'enregistrement de deux spectres est perturbé par l'ajout de la ligne de base dont les paramètres peuvent varier.

(c) Les spectres mis sous la forme de leurs dérivées secondes sont superposés.

Figure 2.3. Effet de la dérivation seconde sur des spectres [Ber05]

Une approximation très simple de la dérivée première est donnée par :

$$\frac{d\,\underline{x}_j^{case}}{d\underline{y}^{case}} = \frac{(\Delta\,\underline{x}_j^{case})_l}{\Delta y_l^{case}} = \frac{x_{lj}^{case} - x_{l+1\,j}^{case}}{y_l^{case} - y_{l+1}^{case}},\ l=1\ldots O \qquad \text{(eq.2.6)}$$

La dérivée seconde peut être obtenue simplement par deux dérivations premières successives, soit :

$$\frac{d^2 \underline{x}_j^{case}}{d^2 \underline{y}^{case}} = \frac{x_{l-1\,j}^{case} - 2x_{lj}^{case} + x_{l+1\,j}^{case}}{(\Delta y_l^{case})^2},\ l=1\ldots O \qquad \text{(eq.2.7)}$$

La dérivée peut être également calculée par des méthodes numériques telles que celle décrite par Savitzky et Golay [SG64]. Cette méthode est basée sur la dérivée précédée d'un lissage et a été utilisée en traitement de spectres IR pour réduire les effets de la variation incontrôlée de la ligne de base et séparer les régions actives des spectres qui se chevauchent partiellement. Le lissage est une technique qui consiste à réduire les irrégularités et singularités d'une courbe. Il est fortement conseillé avant l'utilisation de dérivées afin de diminuer le bruit dû à la dérivation et qui se retrouve amplifié dans le spectre. Dans la suite, nous nous intéressons à présenter cette technique de lissage puis la méthode Savitzky et Golay.

b) Le lissage et la dérivation de Savitzky-Golay (SG)

Les spectres contiennent généralement du bruit qui est aléatoire. Le lissage permet de diminuer cette erreur non systématique. Savitzky et Golay ont montré que l'on peut dériver un ensemble de nombres entiers utilisables comme coefficients de pondération pour mener des opérations de lissage. L'utilisation de ces coefficients de pondération, souvent appelés nombres entiers de convolution, s'est avérée tout à fait équivalente à l'ajustage polynomial [SG64]. Une manière plus fine consiste à désigner l'intervalle I = [y_i^{case} - m ; y_i^{case} + m], comme étant l'intervalle de milieu y_i^{case} et de largeur 2m + 1. Nous considérons ensuite un polynôme P_i de degré d, avec d < 2m + 1. Pour chaque intervalle I, on effectue une régression afin de déterminer le polynôme P_i minimisant l'erreur au sens des moindres carrés (Figure 2.4). On définit la valeur lissée [AJT98] du spectre $\underline{x}_j^{case}$ comme :

$$x_{lissage,\,ij}^{case} = P_i\left(x_{ij}^{case}\right), \qquad \text{(eq.2.8)}$$

Si le polynôme est au moins de degré 1, on peut déterminer la dérivée première par Savitzky-Golay :

$$(x_{lissage,\,ij}^{case})' = P_i\left(x_{ij}^{case}\right)', \qquad \text{(eq.2.9)}$$

et s'il est au moins de degré 2, la dérivée seconde par Savitzky-Golay :

$$(x_{lissage,\,ij}^{case})'' = P_i(x_{ij}^{case})'', \qquad \text{(eq.2.10)}$$

La Figure 2.4 montre le processus et les étapes pour déterminer la dérivée première par Savitzky-Golay (SG) pour un spectre simulé $\underline{x}_j^{case}$. La courbe en noir représente le polynôme P_i estimé sur m points et la courbe en rouge représente la dérivée du spectre lissé $(\underline{x}_{lissage,j}^{case})'$. Les paramètres de la méthode sont : le nombre de points m, le degré du polynôme d et l'ordre de dérivation.

Cette méthode de prétraitement conduit à la suppression des effets de la variation incontrôlée de la ligne de base et peut être utilisée comme un moyen de réduire le décalage et la multi-colinéarité des données spectroscopiques. Par exemple elle supprime l'absorption du fond « background drift » due à la diffusion de la lumière par des particules.

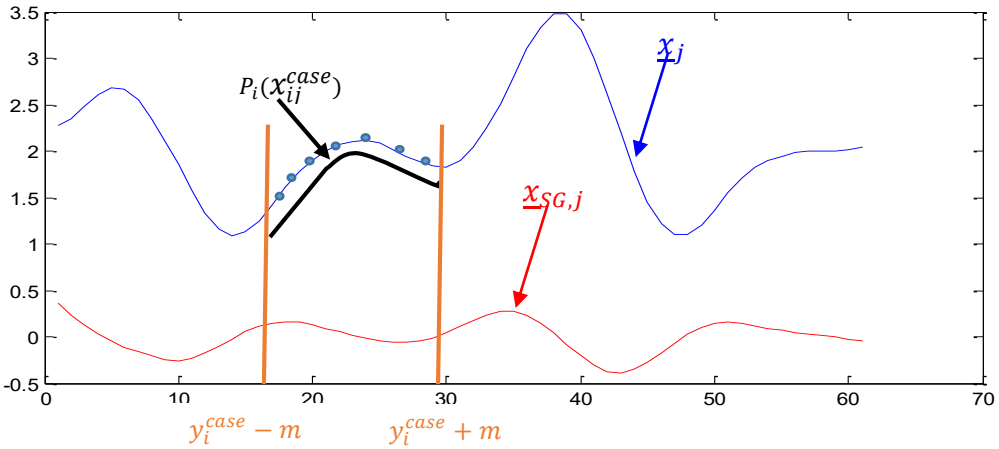


Figure 2.4. Estimation de la dérivée première par la méthode Savitzky-Golay (SG). Fenêtre de 2m+1 points (en bleu) et polynôme d'ordre k (en noir ; $P_i(\underline{x_j}^{case})$) utilisé pour l'opération de lissage. En rouge la dérivée par Savitzky-Golay ($\underline{x}_{lissage,ij}^{case}$)'

II.2.4. Multiplicative Scatter Correction (MSC)

La MSC est une méthode de correction de diffusion qui a été introduite par Martens et al. en 1983 [MJG83] et développée par Geladi et al. en 1985 [GMM85]. Elle repose sur l'idée de corriger le niveau de dispersion de tous les spectres des échantillons à partir d'un spectre « idéal » qui est généralement le spectre moyen. Le concept derrière la MSC est que les artefacts ou les imperfections (par exemple l'effet de dispersion indésirable) seront retirés de la matrice de données (spectres) avant la modélisation. La MSC comprend deux étapes :

- Estimation des coefficients de correction (contributions additive et multiplicative) : Chaque spectre $\underline{x_j}^{case}$ est alors estimé par rapport au spectre moyen de l'ensemble des spectres considéré $\underline{\overline{x^{case}}} = \frac{\sum_{j=1}^{J} \underline{x_j}^{case}}{J}$ par une méthode des moindres carrés.

$$\underline{x_j}^{case} = a_j + b_j \underline{\overline{x^{case}}} + e_j \qquad \text{(eq.2.11)}$$

- Correction du spectre enregistré.

$$\underline{x}_{MSC\ j}^{case} = \frac{\underline{x_j}^{case} - a_j}{b_j} = \underline{\overline{x^{case}}} + \frac{e_j}{b_j} \qquad \text{(eq.2.12)}$$

où e_j represente le spectre résidu, $\underline{x}_{MSC\ j}^{case}$ est le spectre corrigé et a_j et b_j sont les coefficients de correction qui peuvent être estimés par une méthode des moindres carrés.

Nous remarquons que Dhanoa et al. [DLS94] ont montré qu'il existe une similarité évidente entre la SNV et la MSC. Cette relation peut être présentée par l'approximation suivante :

$$\underline{x}_{MSC\ j}^{case} \approx \underline{x}_{SNV\ j}^{case} . \overline{\sigma(X^{case})} + \overline{\overline{X^{case}}} \qquad \text{(eq.2.13)}$$

où $\overline{\sigma(X^{case})}$ est l'écart type moyen entre tous les spectres et $\overline{\overline{X^{case}}}$ est la valeur moyenne de tous les spectres moyens, avec case = MIR ou NIR.

II.2.5. Extended Multiplicative Scatter Correction (EMSC)

La méthode EMSC est une extension de la méthode MSC qui a été développée par Martens et al [RVE09]. Elle repose sur l'idée de corriger le niveau de dispersion de tous les spectres des échantillons à partir de polynômes de degré k (k > 1), des facteurs de correction en fonction de l'échelle (nombre d'ondes) et l'utilisation d'une connaissance a priori d'un spectre de référence. Cette méthode de prétraitement a un plus grand potentiel pour éliminer les effets des interférences spectrales qui peuvent apparaître lorsque la longueur d'onde d'un élément non identifié présent dans l'échantillon est absorbée et entre dans la bande passante de l'absorption de l'élément qui nous intéresse.

L'EMSC comprend deux étapes :

- Estimation des coefficients de correction par une méthode des moindres carrés pour chaque spectre de l'échantillon :

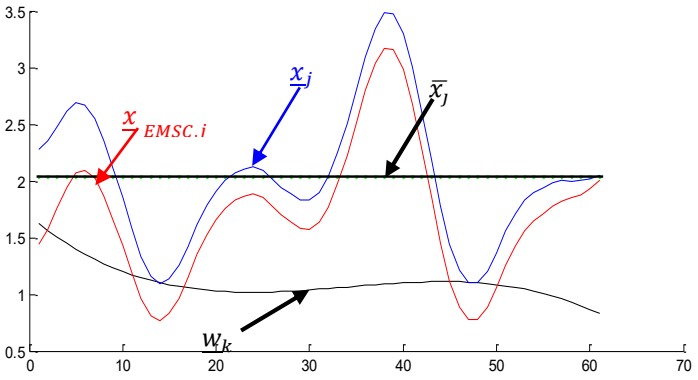
$$\underline{x}_j^{case} = a_j + b_j \overline{\underline{x}^{case}} + c_j w_k + e_j \qquad \text{(eq.2.14)}$$

où $\overline{\underline{x}^{case}}$ est comme précédemment le spectre moyen de l'ensemble des spectres considérés (spectre de référence), e_j est le spectre résidu qui représente l'information utile du spectre, $\underline{w}_k$ un polynôme de degré k calculé sur la vecteur de nombres d'ondes, et a_j, b_j et c_j sont les coefficients de correction estimés par une méthode des moindres carrés.

- Calcul du spectre prétraité par EMSC :

$$\underline{x}_{EMSC,j}^{case} = \frac{\underline{x}_j^{case} - a_j - c_j \underline{w}_k}{b_j}, \qquad \text{(eq.2.15)}$$

où $\underline{x}_{EMSC,j}^{case}$ est le spectre prétraité par EMSC du spectre à corriger (brut) $\underline{x}_j^{case}$. La Figure 2.5 illustre les différentes étapes du processus permettant de déterminer le spectre IR corrigé par la méthode de prétraitement EMSC. La méthode EMSC élimine les interférences spectrales multiplicatives de dispersion d'un spectre, par exemple ce prétraitement a permis d'enlever l'influence des rayons cosmiques dans le spectre IR [AK12]. Elle permet également d'enlever la courbure de degré k de la ligne de base grâce au polynôme w_k.



- $\underline{x}_j = a_j + b_j\,\overline{x}_j + c_j \underline{w}_k$
- w_k désigne un polynôme de degré k= 2
- $\overline{x}_j$ est le vecteur moyen de $\underline{x}_j$
- $\underline{x}_{EMSC,j}$ le spectre corrigé par EMSC.

Figure 2.5. Exemple de spectre simulé (en bleu) et spectre corrigé par EMSC (en rouge)

II.2.6. Méthodes de prétraitement combinées

Certains prétraitements énumérés ci-dessus peuvent être combinés. Par exemple, il est possible de corriger la ligne de base (prétraitement LB) dans un premier temps afin de supprimer des phénomènes intrinsèques étrangers à l'expérience (la lumière extérieure par exemple) puis de normaliser le résultat à l'aide de la SNV afin de réduire les variations d'intensité susceptibles d'apparaître à cause des

conditions d'acquisition qui ne sont pas parfaitement reproductibles. De même, la SNV peut être appliquée avant ou après un prétraitement de type SG. Le Tableau 2.1 regroupe une liste non exhaustive de publications dans lesquelles différentes méthodes de prétraitements ont été appliquées sur des spectres NIR et MIR de biomasse lignocellulosique.

Tableau 2.1. Liste de publications dans lesquelles différentes méthodes de prétraitements ont été utilisées sur des spectres MIR et NIR de biomasse lignocellulosique.

Auteurs	Type de spectre	Méthode de prétraitement utilisée
Lu Liu et al. [LYA10]	NIR	EMSC
Sanderson et al. [SAC96]	NIR	SNV
Torbjörn et al. [TLJ09]	NIR	SNV
Carballo-Meilan et al. [CGB14]	MIR	SNV
Casalea et al. [CSO10]	MIR, NIR	SG1 + SNV
Krasznai et al. [KCC12]	MIR	SG1 + SNV, SG2 + SNV, LB + SNV, MSC, LB
Parsons et al. [PCL14]	MIR	SG1 + SNV
Popescu et al. [PP13]	NIR	SG2
Allison et al. [AMH09]	MIR	SG1 + SNV
Allison et al. [ATM09]	MIR	SG1 + SNV, SG2 + SNV,
Baum et al. [BAM12]	NIR	MSC
Waruru et al. [WSN14]	NIR	SG2 + SNV
Bruun et al. [BJM10]	NIR	SG2, SNV, MSC,
Chazal et al. [CRD14]	MIR	SG2 + SNV
Chen et al. [CFA10]	MIR	LB+SNV, SG1+SNV, LB
Ferreira et al. [FGP14]	MIR, NIR	SG1 + SNV, SG2 + SNV
Dolezel-Horwath et al. [DHK05]	NIR	SG1 + SNV, SG2 + SNV
Chong et al. [CPS12]	NIR	MSC, SG2
Gholizadeh et al. [GBS13]	MIR, NIR	SNV, MSC, SG1, SG2
Wang et al. [WMN14]	MIR, NIR	SG1 + SNV
Yi-Wei et al. [YSC11]	MIR, NIR	SNV
Lupoi et al. [LSD14]	MIR, NIR	SG1 + SNV,
Calderon et al. [CAD09]	MIR, NIR	SNV
Bellon-Maurel et al. [BM11]	MIR, NIR	SNV, MSC, SG1 + SNV, SG2 + SNV,
Sabatier et al. [STB12]	NIR	SNV, MSC, SG1, SG2,
Rinnan et al. [RVE09]	NIR	SG1 + SNV, SG2 + SNV, SNV, MSC,
Ludwig et al. [Lud08]	MIR, NIR	MSC, SG1, SG2,
Liu et al. [LYS10]	NIR	EMSC, MSC, SG1 + SNV, SG2 + SNV,

À partir de cette étude, nous avons constaté que les méthodes de prétraitements le plus utilisées sont : la SNV, la MSC, l'EMSC, la LB suivie d'une SNV (LB+ SNV), la dérivée SG d'ordres 1 et 2 suivie d'une SNV (SG 1 et 2 + SNV).

II.3. Application des méthodes de prétraitement aux spectres IR

II.3.1. Prétraitement des spectres MIR et NIR

Dans les sections précédentes, nous avons expliqué le fonctionnement des différentes méthodes de prétraitement sur des données simulées et leur application pour l'analyse de la biomasse lignocellulosique. Dans cette section, nous allons appliquer les méthodes de prétraitement le plus utilisées sur des spectres MIR et NIR qui ont été recueillis sur un échantillon de biomasse lignocellulosique (maïs). La Figure 2.6 montre les spectres MIR et NIR bruts et prétraités par les

méthodes SNV, LB, MSC, EMSC, ainsi que les combinaisons : SG d'ordre 1et 2 suivies de la SNV et LB suivie de la SNV.

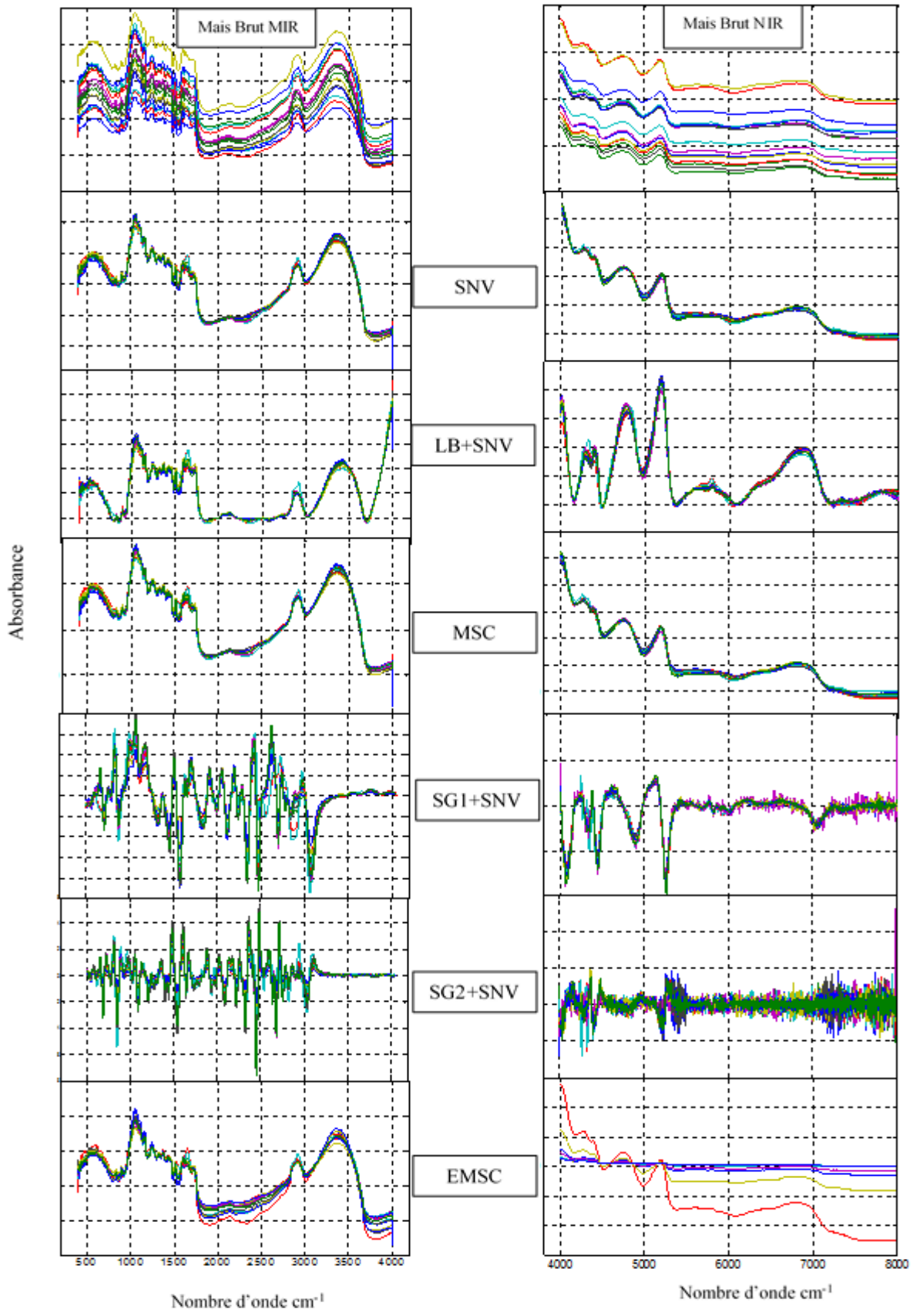


Figure 2.6. Application de méthodes de prétraitements sur des spectres MIR et NIR de biomasse lignocellulosique : SNV, LB suivie de SNV, MSC, EMSC, SG d'ordre 1 et 2 suivies de SNV.

D'après la Figure 2.6, nous pouvons confirmer qu'il existe une similarité entre les spectres MIR prétraités par les méthodes SNV et MSC, comme nous l'avions mentionné précédemment. Ces résultats montrent que les méthodes SNV ou LB suivies de SNV, MSC et EMSC permettent de supprimer les artéfacts de dispersion dans les spectres IR. Elles ont permis de réaligner les spectres en supprimant la ligne de base des spectres IR de biomasse lignocellulosique. Les méthodes SG d'ordre 1 et 2 suivies de la SNV mettent en évidence l'importance de pics IR, plus que les autres méthodes de prétraitements.

Pour les spectres NIR, nous pouvons également observer la similarité entre les méthodes SNV et MSC, avec l'émergence d'un petit nombre de pics spectraux. Les méthodes LB suivie de la SNV et SG d'ordre 1 suivie de la SNV permettent d'obtenir un nombre plus important de pics spectraux. Comme on peut l'observer, l'application des méthodes EMSC et SG d'ordre 2 suivie de la SNV aboutit à des spectres de mauvaise qualité.
Comme on peut le constater sur la Figure 2.6, les résultats des différentes méthodes de prétraitements sur les échantillons de biomasse lignocellulosique varient énormément et semblent dépendants de l'échantillon considéré.

II.3.2. Prétraitement des spectres combinés

Pour les spectres combinés par concaténation ou le produit extérieur, nous avons prétraité chaque spectre MIR et NIR séparément. Ensuite, nous avons combiné le spectre MIR avec le spectre NIR, spectres qui ont été enregistrés sur le même échantillon de biomasse lignocellulosique. On obtient des spectres prétraités MIR et NIR que l'on peut combiner soit par concaténation ou bien par le produit extérieur.

II.3.3. Choix des gammes spectrales

Un spectre MIR ou NIR est un intervalle qui contient l'ensemble des nombres d'ondes correspondants dont les amplitudes peuvent être reliées à l'évolution chimique des échantillons étudiés, par exemple au cours de la biodégradation. La Figure 2.7 montre les spectres MIR d'une biomasse lignocellulosique (maïs) enregistrés sur les nombres d'ondes allant de 400 à 4000 cm^{-1}. A partir de ces mesures, il est possible de définir des gammes spectrales d'intérêt avec leurs caractéristiques chimiques :
400 à 4000 cm^{-1} : Région spectrale totale MIR donnée par l'appareil.
900 à 1800 cm^{-1} : Gamme spectrale informative importante appelée "fingerprint" qui renseigne sur les vibrations principales des groupes chimiques des composés utiles dans les analyses de biomasse lignocellulosique. Elle est largement choisie dans la plupart des applications biomédicales et les sciences agricoles [LS15].
800 à 1800 cm^{-1} : Gamme spectrale qui contient en plus tous les pics correspondant aux groupes fonctionnels chimiques connus. Cette gamme permet d'inclure les pieds des pics IR. Elle est largement utilisée dans les applications de biomasse lignocellulosique et sols.
Nous avons remarqué que certaines équipes regardent les gammes spectrales 1800-2500, 700-4000 et 750-3700 cm^{-1}, les derniers incluant des bandes associées à la subérine, par exemple 2856 cm^{-1} (CH_3 stretching), 2930 cm^{-1} (asymmetric CH_2 stretching) et 2959 cm^{-1} (symmetric CH_2 stretching). Nous nous sommes concentrés dans cette thèse sur les gammes 800-1800 ; 900-1800 et 400-4000 cm^{-1} mais une étude élargie à celles-ci pourrait apporter des perspectives intéressantes [GBS13].

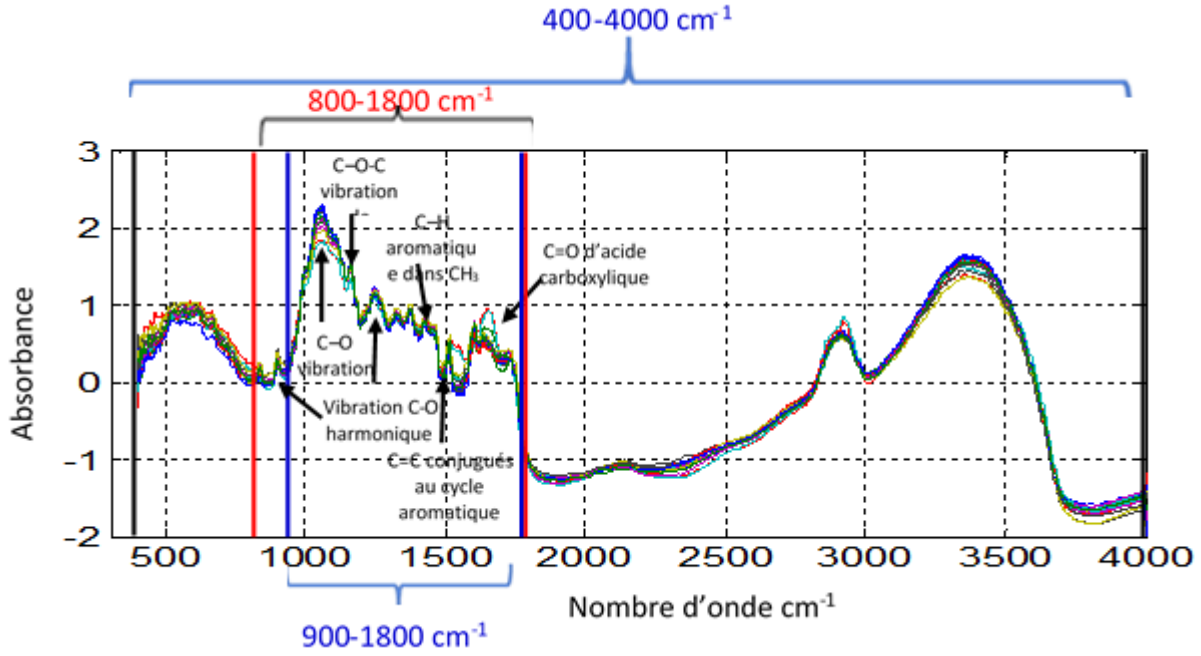


Figure 2.7. Spectres MIR d'une biomasse lignocellulosique (maïs) enregistrés sur la gamme spectrale 400-4000 cm^{-1}

La Figure 2.8 montre les spectres NIR enregistrés sur la même biomasse lignocellulosique. Les spectres ont été enregistrés sur des nombres d'ondes allant de 4000 à 12000 cm^{-1}, mais le bruit est très important à partir de 8000 cm^{-1}. Nous avons donc écarté cette gamme et nous avons choisi pour nos analyses 3 gammes :

- 4000 à 8000 cm^{-1} : Gamme spectrale sans bruit de l'équipement de mesure.
- 4000 à 6000 cm^{-1} : Gamme spectrale informative importante qui contient toutes les bandes correspondantes aux groupes fonctionnels chimiques connus dans les spectres NIR.
- 4200 à 7500 cm^{-1} : Gamme classique utilisée dans des travaux précédents [BM11, DGL10].

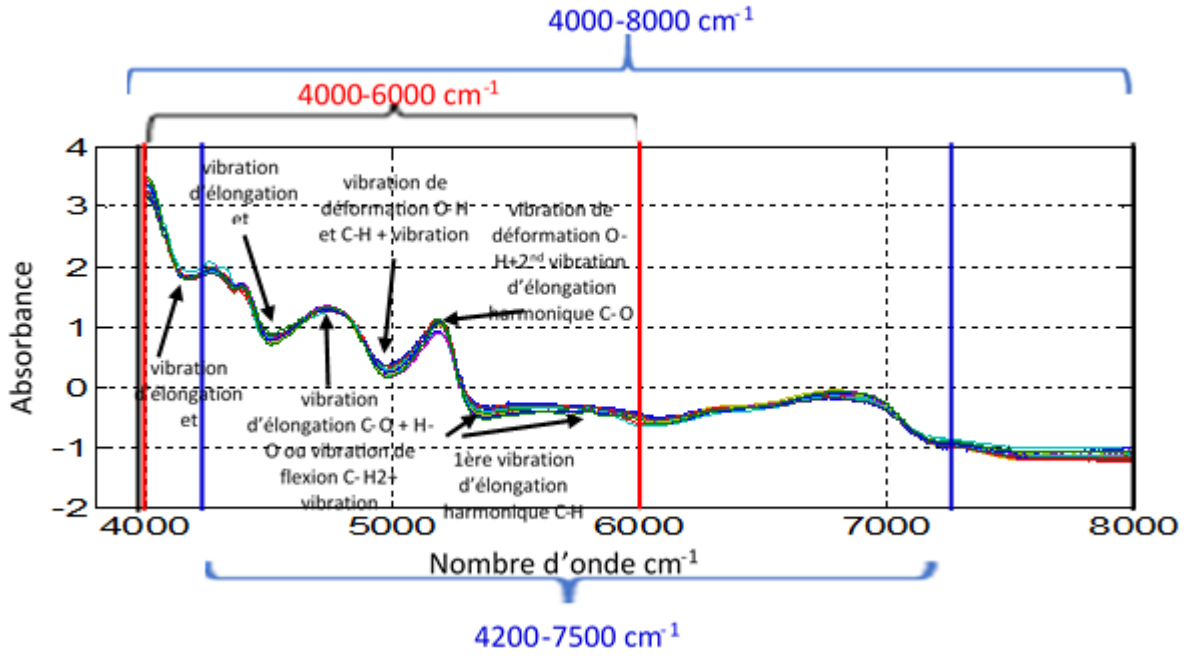


Figure 2. 8. Spectres NIR d'une biomasse lignocellulosique (maïs) enregistrés sur la gamme spectrale 4000-8000 cm^{-1}

Dans la suite de ce chapitre, nous allons nous intéresser à l'étude de l'influence des méthodes de prétraitement sur l'analyse des spectres IR et plus particulièrement les paramètres associés (par exemple : ordre du polynôme, degré de la dérivée, lissage, etc.), ainsi que le choix des gammes spectrales optimales. Pour cela, nous proposons d'utiliser des méthodes de classification non supervisées. Nous partons de l'hypothèse que le meilleur résultat de classification spectrale nous permet d'en déduire le meilleur jeu de paramètres et la/ou les meilleures gammes spectrales. Nous utilisons ces types de méthodes parce que nos échantillons de biomasse lignocellulosiques sont attribués à l'une ou l'autre classe. Par exemple, nous disposons de deux récoltes de miscanthus (parties aériennes) : 6 échantillons d'une récolte précoce et 6 échantillons d'une récolte tardive que nous cherchons à séparer en 2 classes. Pour cela, nous allons utiliser des méthodes de classification non supervisées qui visent à classer nos échantillons en classes compactes et séparées.

II.4. Méthodes de classification non supervisée

II.4.1. Principes de la classification non supervisée

L'objectif de la classification en spectroscopie est d'identifier les classes auxquelles appartiennent des spectres à partir de traits descriptifs (attributs, caractéristiques, etc.). Nous pouvons séparer les méthodes de classification en deux catégories : la classification supervisée et celle non supervisée. La classification non supervisée ou "clustering" est l'une des techniques fondamentales permettant de construire des classes de spectres IR similaires à partir d'un ensemble hétérogène de ces spectres. Elle vise à séparer automatiquement les spectres IR en classes naturelles, c'est-à-dire sans aucune connaissance préalable sur les classes excepté leur nombre. La classification supervisée quant à elle dispose au départ d'un échantillon dit d'apprentissage dont le classement est connu. Cet échantillon est utilisé pour l'apprentissage des règles de classement. Ces règles de classement sont déterminées par exemple par une approche probabiliste (maximum de vraisemblance, règle de Bayes…) [Mac67]. Dans le cadre de cette thèse, nous nous limiterons à l'étude d'une méthode de classification non-supervisée qui est couramment utilisée pour des données spectrales infrarouges.

Soit $X^{case} = \left[\underline{x}_1^{case} \quad \dots \ \underline{x}_j^{case} \ \dots \quad \underline{x}_J^{case}\right] \in M_{J,O}(\mathbb{R})$ une matrice formée de J spectres infrarouges avec $\underline{x}_j^{case} \in R^O$; case =MIR ou NIR ; O= P ou Q et M = $[\underline{m}_1, \dots \ \underline{m}_i \ \dots, \underline{m}_K]$ une matrice formée des K centres de chaque classe avec $\underline{m}_i \in \mathbb{R}^O$ et i=1…K. Ces spectres appartiennent à un ensemble de classes C = $\{c_1, \dots \ c_i \ \dots, c_K\}$ avec K $<$ J.

II.4.2. Algorithme Fuzzy C-Means (FCM)

Fuzzy C-Means (FCM) est un algorithme de classification non-supervisé issu de l'algorithme des K-moyens (K-means). FCM introduit la notion d'ensemble flou dans la définition des classes : chaque spectre IR $\underline{x}_j^{case}$ appartient à chaque classe avec un certain degré d'appartenance u_{ij} et toutes les classes sont caractérisées par leur centre de gravité m_i. Cet algorithme est moins sensible à l'initialisation aléatoire que le K-means [CKC11].

Comme les autres algorithmes de classification non supervisés, il utilise un critère de minimisation des distances intra-classe et de maximisation des distances inter-classes, mais en donnant un certain degré d'appartenance à chaque classe pour chaque spectre. Cet algorithme nécessite au préalable la connaissance du nombre de classes et génère les classes par un processus itératif en minimisant une fonction objectif qui est définie par [Bez81] :

$$J_{FCM} = \sum_{i=1}^{K} \sum_{j=1}^{J} u_{ij}^{m} d_2^2(\underline{x}_j^{case}, \underline{m}_i), \quad \text{(eq.2.16)}$$

avec J le nombre de spectres, K le nombre de classes, m ∈ [1; ∞[un coefficient flou et d(.) est la distance euclidienne entre le centre $\underline{m}_i$ et le spectre $\underline{x}_j^{case}$:

$$d_2^2(\underline{x}_j^{case}, m_i) = \left(\underline{x}_j^{case} - \underline{m}_i\right)^T \left(\underline{x}_j^{case} - \underline{m}_i\right), \quad \text{(eq.2.17)}$$

Les étapes de l'algorithme Fuzzy C-Means sont :

1. Initialiser la matrice d'appartenance

$$\text{U} = [u_{ij}^{(0)}]_{1 \le i \le K, 1 \le j \le J} \quad \text{(eq.2.18)}$$

avec des valeurs aléatoires comprises entre 0 et 1 telles que les éléments de U satisfont la contrainte (la somme des degrés d'appartenance d'un spectre à toutes les classes est égale à 1)

$$\sum_{i=1}^{K} u_{ij}^{(0)} = 1, \forall\, j = 1 \dots J \quad \text{(eq.2.19)}$$

2. Calculer les centres des classes m_i

$$\underline{m}_i{}^{(t)} = \sum_{j=1}^{J} \frac{u_{ij}^{(t-1)}}{\sum_{j=1}^{J} u_{ij}^{(t-1)}} \underline{x}_j^{case}, \forall\, i = 1 \dots K, \quad \text{(eq.2.20)}$$

où t représente l'index de l'itération de l'algorithme.

3. Réajuster les coefficients de la matrice d'appartenance suivant la position des centres de classes

$$u_{ij}^{(t)} = \left[\sum_{l=1}^{K} \left[\frac{d^2(\underline{x}_j^{case}, \underline{m}_i{}^{(t)})}{d^2(\underline{x}_j^{case}, \underline{m}_l{}^{(t)})} \right]^{\frac{1}{m-1}} \right]^{-1} \forall\, j = 1 \dots J, i = 1 \dots K, \quad \text{(eq.2.21)}$$

4. Répéter les étapes 2 et 3 tant que le critère d'arrêt n'est pas satisfait.
   L'algorithme s'interrompt suivant un critère d'arrêt fixé par l'utilisateur qui peut être choisi parmi les suivants : soit le nombre limite d'itérations est atteint, soit l'algorithme a convergé, c'est-à-dire que la différence entre la matrice d'appartenance actuelle et la matrice d'appartenance précédente est inférieure à une valeur de tolérance spécifiée ε [PYC04].

$$\left\| u_{ij}{}^{(t)} - u_{ij}{}^{(t-1)} \right\|^2 < \varepsilon \quad \text{(eq.2.22)}$$

Le changement du résultat de classification est directement lié aux paramètres du FCM, notamment au choix du coefficient flou m et de la distance d(.). Pour cela, nous allons analyser le choix de ces paramètres et ses effets dans la classification de données spectrales.
Si le coefficient flou $m = 1$, alors l'algorithme FCM est le même que l'algorithme K-means ; si $m \to \infty$ alors tous les spectres sont uniformément répartis entre toutes les classes. Généralement, m est choisi à égale 2.

L'utilité des distances est de pouvoir comparer les ressemblances et les différences entre les spectres. Cette opération est importante, notamment dans le domaine de la classification. Il est plus probable que deux spectres semblables soient dans une même classe que deux spectres dissemblables. L'algorithme FCM est basé sur la distance euclidienne qui est adaptée pour des données peu bruitées et qui se trouvent dans des classes ayant des formes sphériques. Pour notre application, les spectres peuvent être bruités et se retrouver regroupés dans des classes de formes non sphériques.
La Figure 2.9 nous permet d'avoir une représentation des classes de spectres MIR et NIR de données de biomasse lignocellulosique. Pour cela, nous avons appliqué la méthode d'analyse en composantes principales pour visualiser la distribution des spectres en fonction de leurs deux premières composantes principales. On constate que les spectres sont regroupés dans des classes de formes non sphériques.

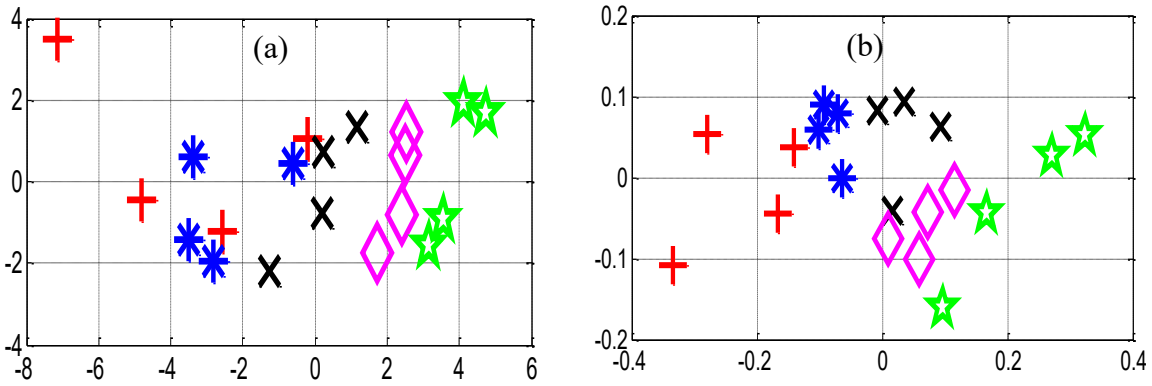


Figure 2.9. Classes formées par les spectres (a) MIR et (b) NIR après projection de spectres enregistrés sur une biomasse lignocellulosique sur les deux premières composantes principales. Les couleurs représentent les différents temps de dégradation.

Pour cette raison, nous avons considéré la distance généralisée :

$$d_A^2(\underline{x}_j^{case}, \underline{m}_i) = \left(\underline{x}_j^{case} - \underline{m}_i\right)^T A \left(\underline{x}_j^{case} - \underline{m}_i\right) \qquad \text{(eq.2.23)}$$

A est l'inverse de la matrice de variance des spectres, on obtient la distance diagonale, si A est l'inverse de la matrice de covariance des spectres alors on obtient la distance de Mahalanobis, et si A est la matrice identité, la distance d_A^2 serait égale à la distance euclidienne. De plus, il est à noter que la forme géométrique des classes associée à une distance de Mahalanobis ou diagonale est elliptique (représentation dans un espace à deux dimensions).
Nous décrivons dans la suite de ce chapitre une méthode de ré-échantillonnage qui va être combinée aux méthodes de classification (par exemple FCM). Nous sommes amenés à utiliser cette méthode parce que nous ne disposons que d'un nombre réduit de spectres. Ce rééchantillonnage permet d'améliorer la classification de nos spectres de biomasses.

II.4.3. Ré-échantillonnage « Bootstrap »

Le terme de ré-échantillonnage désigne un ensemble de méthodes qui consiste à faire de l'inférence statistique sur de « nouveaux » échantillons tirés à partir d'un échantillon initial. Cette approche permet d'estimer avec une meilleure précision la classification des échantillons de taille réduite (c.-à-d. ici un nombre réduit de spectres). Les méthodes les plus courantes sont celles de Monte-Carlo, méthodes bayésiennes (échantillonneur de Gibbs, algorithme de Metropolis-Hastings). La méthode

« bootstrap » ne nécessite pas d'information supplémentaire que celle disponible dans l'échantillon. Le principe général de cette méthode est de ré-échantillonner un grand nombre de fois les échantillons initiaux (les spectres IR), l'inférence statistique étant basée sur les résultats des échantillons ainsi obtenus.

On considère l'ensemble de spectres prétraités par un des prétraitements, $\underline{x}_1^{case}, \underline{x}_2^{case}, \dots, \underline{x}_j^{case}, \dots \underline{x}_J^{case}$ qui appartiennent aux classes $c(1), c(2), \dots, c(j), \dots c(J)$, avec $c(j) \in \{1, \dots, K\}$ et K le nombre de classes connues a priori et imposé par l'étude. Nous réalisons T tirages aléatoires dans cet ensemble. Pour chaque nouveau tirage, $t \in \{1 \dots T\}$ nous obtenons un nouvel ensemble de J spectres : $\underline{x}_{t(1)}^{case}, \underline{x}_{t(2)}^{case}, \dots, \underline{x}_{t(j)}^{case}, \dots \underline{x}_{t(J)}^{case}$ qui appartiennent aux classes $c(1), c(2), \dots, c(j), \dots c(J)$. La classification FCM est appliquée sur cet ensemble ré-échantillonné pour estimer K classes. Le résultat obtenu nous donne l'appartenance de chaque spectre $\underline{x}_{t(j)}^{case}$ aux classes $\hat{c}(1), \hat{c}(2), \dots, \hat{c}(j), \dots \hat{c}(J)$. Pour chaque tirage aléatoire, après défuzzification, nous mesurons le nombre d'échantillons bien classés, ce qui nous permet de calculer la moyenne de bonnes classifications sur l'ensemble des T itérations [RPC13].

$$E_{moy}^{p} = (1 - \frac{1}{T}\sum_{i=1}^{T}\frac{1}{J}\sum_{j=1}^{J}\left|sign\big(\, c(j) - \hat{c}(j)\big)\right|) \;\; en\,[\%] \qquad \text{(eq.2.24)}$$

Pour fixer les idées, nous prenons l'exemple suivant. Considérons un ensemble de six spectres appartenant à deux classes :

$\underline{x}_1^{case}, \underline{x}_2^{case}, \underline{x}_3^{case}$ ∈ à la classe 1
$\underline{x}_4^{case}, \underline{x}_5^{case}, \underline{x}_6^{case}$ ∈ à la classe 2

Cet ensemble est ré-échantillonné avec répétition, donnant l'ensemble $\underline{x}_1^{case}, \underline{x}_2^{case}, \underline{x}_3^{case}, \underline{x}_5^{case}, \underline{x}_6^{case}, \underline{x}_6^{case}$. Les trois premiers spectres appartiennent à la classe 1, les trois autres à la classe 2. Ce nouvel ensemble est soumis à la méthode de classification FCM, le nombre de classes étant connu à l'avance, K = 2. Supposons qu'on obtienne, après l'étape de défuzzification, le résultat suivant :

$\underline{x}_1^{case}, \underline{x}_2^{case}$ ∈ à la classe 1
$\underline{x}_3^{case}, \underline{x}_5^{case}, \underline{x}_6^{case}, \underline{x}_6^{case}$ ∈ à la classe 2

Un spectre est donc mal classé, donnant un pourcentage de spectres bien classés de 83,3%. Ce processus est répété T fois.
La combinaison de la méthode de ré-échantillonnage avec la méthode de classification permet d'obtenir une estimation plus précise des classifications des échantillons et limite l'influence de l'initialisation aléatoire dans l'algorithme FCM.

II.4.4. Autres méthodes de classification FCM

Comme nous l'avons mentionné précédemment, l'algorithme FCM s'appuie sur la distance euclidienne qui est adaptée pour les classes de structure sphérique. Puisque les spectres IR de nos données de biomasse lignocellulosique ont une répartition non sphérique, nous avons décidé de chercher d'autres métriques adaptées. La distance de Mahalanobis peut être utilisée pour couvrir les classes des données de structures non sphériques, notamment elliptiques [LJY09]. Mais, Krishnapuram et Kim (1999) ont souligné que la distance de Mahalanobis ne peut pas être utilisée directement dans l'algorithme FCM.

Pour cela, nous avons analysé d'autres approches de l'algorithme FCM. Celles-ci sont basées sur la notion de distance « adaptative » de Mahalanobis.

II.4.4.1. L'algorithme GK

Gustafson et Kessel (1978) ont généralisé l'algorithme FCM en employant une distance adaptative dans le but de détecter des classes de différentes formes géométriques dans un ensemble de données [GK78]. Dans ce cas, chaque classe possède une matrice de covariance floue. La distance est alors donnée par :

$$d^2_{V_i}(\underline{x}_j^{case}, \underline{m}_i) = \left(\underline{x}_j^{case} - \underline{m}_i\right)^T V_i \left(\underline{x}_j^{case} - \underline{m}_i\right) \quad \text{(eq.2.25)}$$

avec

$$V_i = |\Sigma_i|^{\frac{1}{o}} \Sigma_i^{-1} \quad \text{(eq.2.26)}$$

et Σ_i la matrice de covariance de l'i-ième classe :

$$\Sigma_i = \left[\sum_{i=1}^K \sum_{j=1}^J u_{ij}^m\right]^{-1} \sum_{i=1}^K \sum_{j=1}^J u_{ij}^m \left(\underline{x}_j^{case} - \underline{m}_i\right)\left(\underline{x}_j^{case} - \underline{m}_i\right)^T \quad \text{(eq.2.27)}$$

où |.| désigne le déterminant.
La fonction objectif de l'algorithme GK est définie par :

$$J_{GK} = \sum_{i=1}^K \sum_{j=1}^J u_{ij}^m d^2_{V_i}(S_j, \underline{m}_i) \quad \text{(eq.2. 28)}$$

Les étapes de l'algorithme GK sont identiques à l'algorithme FCM excepté l'étape 3.
Etape 3 : Calculer la matrice Vi et réajuster la matrice d'appartenance

$$\Sigma_i^{(t)} = \left[\sum_{i=1}^K \sum_{j=1}^J u_{ij}^{m(t-1)}\right]^{-1} \sum_{i=1}^K \sum_{j=1}^J u_{ij}^{m(t-1)} \left(\underline{x}_j^{case} - \underline{m}_i{}^{(t)}\right)\left(\underline{x}_j^{case} - \underline{m}_i{}^{(t)}\right)^T \quad \text{(eq.2.29)}$$

$$V_i^{(t)} = \left|\Sigma_i{}^{(t)}\right|^{\frac{1}{p}} \left[\Sigma_i{}^{(t)}\right]^{-1} \quad \text{(eq.2.30)}$$

$$u_{ij}{}^{(t)} = \left[\sum_{l=1}^K \left[\frac{(\underline{x}_j^{case} - \underline{m}_i{}^{(t)}) V_i{}^{(t)} \left(\underline{x}_j^{case} - \underline{m}_i{}^{(t)}\right)^T}{(\underline{x}_j^{case} - \underline{m}_l{}^{(t)}) V_i{}^{(t)} \left(\underline{x}_j^{case} - \underline{m}_l{}^{(t)}\right)^T}\right]^{\frac{1}{m-1}}\right]^{-1} \quad \forall\, j = 1 \ldots J, i = 1 \ldots K \quad \text{(eq.2.31)}$$

II.4.4.2. L'algorithme GG

L'algorithme de classification Gath Geva (GG) est une extension de l'algorithme GK. Cet algorithme est issu du FCM et fait intervenir une distance non euclidienne. Gath et Geva définissent une distance "exponentielle" afin d'adapter le FCM à des cas correspondant à des classes hyper-ellipsoïdales, avec des densités différentes pour les différentes classes, ou bien des distributions inégales de l'ensemble des points au sein des différentes classes. Cette distance est définie comme suit :

$$d_e^2(\underline{x}_j^{case}, m_i) = \frac{2\pi^p\sqrt{\Sigma_i}}{P_i} exp\left[\left(\underline{x}_j^{case} - \underline{m}_i\right)^T {\Sigma_i}^{-1} \left(\underline{x}_j^{case} - \underline{m}_i\right)\right] \quad \text{(eq.2.32)}$$

où Σ_i est la matrice floue de covariance de la ième classe définie par (2.27) et P_i est la probabilité de cette même classe.
La fonction objectif de l'algorithme GG est définie par :

$$J_{GG} = \sum_{i=1}^{K}\sum_{j=1}^{J} u_{ij}^m d_e^2(\underline{x}_j^{case}, \underline{m}_i) \quad \text{(eq.2.33)}$$

Les étapes 1 et 2 et 4 sont identiques à l'algorithme FCM. L'algorithme ne diffère qu'au niveau de l'étape 3.
Etape 3 : Calculer la matrice floue de covariances Σ_i, les probabilités P_i et réajuster la matrice d'appartenance

$${\Sigma_i}^{(t)} = \left[\sum_{j=1}^{J} u_{ij}^{m(t-1)}\right]^{-1} \sum_{j=1}^{J} u_{ij}^{m(t-1)} \left(\underline{x}_j^{case} - \underline{m}_i{}^{(t)}\right)\left(\underline{x}_j^{case} - \underline{m}_i{}^{(t)}\right)^T \quad \text{(eq.2.34)}$$

$$P_i{}^{(t)} = \left[\sum_{i=1}^{K}\sum_{j=1}^{J} u_{ij}^{m(t-1)}\right]^{-1} \sum_{j=1}^{J} u_{ij}^{m(t-1)} \quad \text{(eq.2.35)}$$

II.4.4.3. L'algorithme FCM-M
FCM-M est un algorithme Fuzzy C-Means qui est basé sur la distance adaptative de Mahalanobis. Afin de répondre aux limitations de l'algorithme GK et GG, le FCM-M ajoute un facteur de régulation de la matrice de covariance,$-ln\left|{\Sigma_i}^{-1}\right|$ pour chaque classe, et supprime la contrainte du déterminant de matrices de covariance présente dans l'algorithme GK. La fonction objectif du FCM-M est définie par :

$$J_{FCM-M} = \sum_{i=1}^{K}\sum_{j=1}^{J} u_{ij}^m d_M^2(\underline{x}_j^{case}, \underline{m}_i) \quad \text{(eq.2.36)}$$

avec d_M^2 une distance adaptative de Mahalanobis définie par :

$$d_M^2(\underline{x}_j^{case}, \underline{m}_i) = \begin{cases} \left(\underline{x}_j^{case} - \underline{m}_i\right)^T {\Sigma_i}^{-1} \left(\underline{x}_j^{case} - \underline{m}_i\right) - ln\left|{\Sigma_i}^{-1}\right|, si\ \left(\underline{x}_j^{case} - \underline{m}_i\right)^T {\Sigma_i}^{-1} \left(\underline{x}_j^{case} - \underline{m}_i\right) - ln\left|{\Sigma_i}^{-1}\right| \geq 0 \\ 0 \qquad\qquad si\ \left(\underline{x}_j^{case} - \underline{m}_i\right)^T {\Sigma_i}^{-1} \left(\underline{x}_j^{case} - \underline{m}_i\right) - ln\left|{\Sigma_i}^{-1}\right| < 0 \end{cases} \quad \text{(eq.2.37)}$$

Les étapes 2 et 4 sont identiques à l'algorithme FCM. Pour l'étape 1, il faut aussi calculer D et ${\Sigma_i}^{(0)}$

$$D = \sum_{i=1}^{K}\sum_{j=1}^{J} u_{ij}^{m(0)} \left[\left(\underline{x}_j^{case} - \underline{m}_i\right)^T \left(\underline{x}_j^{case} - \underline{m}_i\right)\right] > 0 \quad \text{(eq.2.38)}$$

$${\Sigma_i}^{(0)} = \left[\sum_{j=1}^{J} u_{ij}^{m(0)}\right]^{-1} \sum_{j=1}^{J} u_{ij}^{m(0)} \left(\underline{x}_j^{case} - \underline{m}_i{}^{(0)}\right)\left(\underline{x}_j^{case} - \underline{m}_i{}^{(0)}\right)^T \quad \text{(eq.2.39)}$$

$$\text{Si } \left|{\Sigma_i}^{(0)}\right| > D, \text{ ou } \left|{\Sigma_i}^{(0)}\right| < \frac{1}{D} \text{ alors } {\Sigma_i}^{(0)} = I \quad \text{(eq.2.40)}$$

Etape 3 : Calculer la matrice floue de covariances Σ_i, et réajuster la matrice d'appartenance

$${\Sigma_i}^{(t)} = \left[\sum_{j=1}^{J} u_{ij}^{m(t-1)}\right]^{-1} \sum_{j=1}^{J} u_{ij}^{m(t-1)} \left(\underline{x}_j^{case} - \underline{m}_i{}^{(t)}\right)\left(\underline{x}_j^{case} - \underline{m}_i{}^{(t)}\right)^T \quad \text{(eq.2.41)}$$

$$\text{Si } \left|\Sigma_i{}^{(t)}\right| > D, \text{ ou } \left|\Sigma_i{}^{(t)}\right| < \frac{1}{D} \text{ alors } \Sigma_i{}^{(t)} = I \qquad \text{(eq.2.42)}$$

$$u_{ij}{}^{(t)} = \left[\sum_{l=1}^{K}\left[\frac{(\underline{x}_j^{case}-\underline{m}_i{}^{(t)})[\Sigma_i{}^{-1}]^{(t)}\left(\underline{x}_j^{case}-\underline{m}_i{}^{(t)}\right)^T - ln[\Sigma_i{}^{-1}]^{(t)}}{(\underline{x}_j^{case}-\underline{m}_l{}^{(t)})[\Sigma_l{}^{-1}]^{(t)}\left(\underline{x}_j^{case}-\underline{m}_l{}^{(t)}\right)^T - ln[\Sigma_i{}^{-1}]^{(t)}}\right]^{\frac{1}{m-1}}\right]^{-1} \quad \forall\, j = 1 \dots J, i = 1 \dots K \qquad \text{(eq.2.43)}$$

On peut remarquer que l'algorithme FCM est un cas particulier de FCM-M, lorsque les matrices de covariance sont égales à la matrice identité.

II.4.4.4. L'algorithme FCM-CM

FCM-CM est un algorithme Fuzzy C-Means toujours basé sur la distance de Mahalanobis. Il permet d'améliorer la stabilité des résultats de classification de l'algorithme de FCM-M en remplaçant toutes les matrices de covariance par une seule matrice de covariance commune à toutes les classes dans la fonction objectif de FCM-M. La fonction objectif du FCM-CM est définie comme suit :

$$J_{FCM-CM} = \sum_{i=1}^{K}\sum_{j=1}^{J} u_{ij}^{m} d_{CM}^2(\underline{x}_j^{case}, \underline{m}_i) \qquad \text{(eq.2.44)}$$

avec d_{CM}^2 la distance définie par l'équation 2.38 mais on remplace la matrice $\Sigma_i{}^{-1}$ par Σ^{-1}

Les étapes 1, 2, 4 de l'algorithme FCM-CM sont les mêmes que celle de l'algorithme FCM.
Etape 3 : Calculer les matrices floues de covariances $\Sigma^{(t)}$ et réajuster la matrice d'appartenance

$$\Sigma^{(t)} = \left[\sum_{i=1}^{K}\sum_{j=1}^{J} u_{ij}^{m(t-1)}\right]^{-1} \sum_{i=1}^{K}\sum_{j=1}^{J} u_{ij}^{m(t-1)} (\underline{x}_j^{case} - \underline{m}_i{}^{(t)})\left(\underline{x}_j^{case} - \underline{m}_i{}^{(t)}\right)^T \qquad \text{(eq.2.45)}$$

$$\text{Si } \left|\Sigma^{(t)}\right| > D, \text{ou } \left|\Sigma^{-1(t)}\right| < \frac{1}{D} \text{ alors } \Sigma^{-1(t)} = I \qquad \text{(eq.2.46)}$$

$$u_{ij}{}^{(t)} = \left[\sum_{l=1}^{K}\left[\frac{(\underline{x}_j^{case}-\underline{m}_i{}^{(t)})[\Sigma^{-1}]^{(t)}\left(\underline{x}_j^{case}-\underline{m}_i{}^{(t)}\right)^T - ln[\Sigma^{-1}]^{(t)}}{(\underline{x}_j^{case}-\underline{m}_l{}^{(t)})[\Sigma^{-1}]^{(t)}\left(\underline{x}_j^{case}-\underline{m}_l{}^{(t)}\right)^T - ln[\Sigma^{-1}]^{(t)}}\right]^{\frac{1}{m-1}}\right]^{-1} \quad \forall\, j = 1 \dots J, i = 1 \dots K \qquad \text{(eq.2.47)}$$

II.4.4.5. L'algorithme FCM-SM

FCM-SM est une autre méthode de classification floue toujours basée sur la distance de Mahalanobis. Elle a été proposée pour normaliser chaque critère de la fonction objectif dans l'algorithme FCM-CM. Pour cela, toutes les matrices de covariance sont remplacées par les matrices de corrélations correspondantes. La fonction objectif du FCM basée sur la distance adaptative de Mahalanobis (FCM-SM) est la suivante :

$$J_{FCM-SM} = \sum_{i=1}^{K}\sum_{j=1}^{J} u_{ij}^{m} d_{SM}^2(\underline{z}_j^{case}, \underline{m}_i) \qquad \text{(eq.2.48)}$$

avec $\underline{z}_j^{case}$ le spectre normalisé du $\underline{x}_j^{case}$ et d_{SM}^2 une distance adaptative définie par :

$$d_{SM}^2(\underline{z}_j^{case},\underline{m}_i) = \begin{cases} \left(\underline{z}_j^{case} - \underline{m}_i\right)^T R^{-1} \left(\underline{z}_j^{case} - \underline{m}_i\right) - ln|R^{-1}|, si \left(\underline{z}_j^{case} - \underline{m}_i\right)^T R^{-1} \left(\underline{z}_j^{case} - \underline{m}_i\right) - ln|R^{-1}| \geq 0 \\ 0 \qquad\qquad si \left(\underline{z}_j^{case} - \underline{m}_i\right)^T R^{-1} \left(\underline{z}_j^{case} - \underline{m}_i\right) - ln|R^{-1}| < 0 \end{cases} \qquad \text{(eq.2.49)}$$

Les étapes 2 et 4 sont les mêmes que celles de l'algorithme FCM. Pour l'étape 1, en plus d'initialiser la matrice d'appartenance, nous devons calculer R :

$$R^{(0)} = \left[\sum_{i=1}^{K}\sum_{j=1}^{J} u_{ij}^{m(0)}\right]^{-1} \sum_{i=1}^{K}\sum_{j=1}^{J} u_{ij}^{m(0)} (\underline{z}_j^{case} - \underline{m}_i{}^{(0)})\left(\underline{z}_j^{case} - \underline{m}_i{}^{(0)}\right)^T \quad \text{(eq.2.50)}$$

$$\text{Si } \left|R^{(0)}\right| < \frac{1}{100} \text{ alors } R^{(0)} = I \quad \text{(eq.2.51)}$$

Etape 3 : Calculer la matrice floue de covariances $R^{(t)}$ et réajuster la matrice d'appartenance :

$$R^{(t)} = \left[\sum_{i=1}^{K}\sum_{j=1}^{J} u_{ij}^{m(t-1)}\right]^{-1} \sum_{i=1}^{K}\sum_{j=1}^{J} u_{ij}^{m(t-1)} (\underline{z}_j^{case} - \underline{m}_i{}^{(t)})\left(\underline{z}_j^{case} - \underline{m}_i{}^{(t)}\right)^T \quad \text{(eq.2.52)}$$

$$\text{Si } \left|R^{(t)}\right| < \frac{1}{100} \text{alors } R^{(t)} = I \quad \text{(eq.2.53)}$$

$$u_{ij}{}^{(t)} = \left[\sum_{l=1}^{K}\left[\frac{(\underline{z}_j^{case} - \underline{m}_i{}^{(t)})[R^{-1}]^{(t)}\left(\underline{z}_j^{case} - \underline{m}_i{}^{(t)}\right)^T - ln[R^{-1}]^{(t)}}{(\underline{z}_j^{case} - \underline{m}_l{}^{(t)})[R^{-1}](t)\left(\underline{z}_j^{case} - \underline{m}_l{}^{(t)}\right)^T - ln[R^{-1}](t)}\right]^{\frac{1}{m-1}}\right]^{-1} \quad \forall\, j = 1 \dots J, i = 1 \dots K \quad \text{(eq.2.54)}$$

Comme nous l'avons vu, l'algorithme FCM basé sur la distance euclidienne ne peut être utilisé que lorsque les classes sont sphériques. Les algorithmes GK et GG ont été développés pour des classes non-sphériques. Mais ces algorithmes ne réussissent pas à prendre en compte les relations entre les centres de clusters dans la fonction objectif. L'algorithme GK doit avoir une information à priori pour chaque classe de données, sinon, il ne peut être utilisé que pour la détection de classes de tailles approximativement identiques. L'algorithme GG doit disposer des probabilités des classes. Pour dépasser la limitation de l'algorithme GK et GG, les algorithmes FCM-M, FCM-CM et FCM-SM ont été proposés, mais ces algorithmes ne peuvent pas être utilisés pour les données infrarouges qui sont de grandes dimensions, car ils nécessitent de calculer le déterminant d'une matrice de covariance de grande dimension, ce qui est difficilement réalisable en raison des problèmes d'instabilité.

II.4.5. Nouvelle approche du FCM : le FCM-R

Comme les différents algorithmes présentés ci-dessus ne peuvent pas être appliqués sur nos spectres, nous proposons une nouvelle approche de l'algorithme FCM basée sur un facteur de covariance. Cet algorithme, que nous avons appelé FCM-R [RPV15], couvre les classes non sphériques et sphériques, tout en tenant compte de la grande dimensionnalité des spectres.

Les performances de notre algorithme seront comparées à celles des algorithmes FCM, GG, GK, FCM-M, FCM-CM, et FCM-SM sur des ensembles de données simulées et réelles qui sont classiquement utilisées pour la comparaison des méthodes de classifications basées sur la FCM. Nous comparons également les performances de l'algorithme FCM-R avec l'algorithme classique FCM sur les données spectrales réelles et artificielles obtenues par des générateurs aléatoires. Dans la suite de cette section, nous détaillons les différentes étapes du comportement de notre algorithme FCM-R.

II.4.5.1. Algorithme FCM-R

La fonction objective de l'algorithme FCM-R est :

$$J_{FCM-R} = \sum_{i=1}^{K}\sum_{j=1}^{J} u_{ij}^{m} \alpha_i \, (\underline{x}_j^{case} - \underline{m}_i)^T (\underline{x}_j^{case} - \underline{m}_i) \quad \text{(eq.2.55)}$$

avec α_i le facteur de covariance qui évalue la colinéarité entre les centres m_i et les spectres $\underline{x}_j^{case}$. Ce facteur prend également en compte la compacité des spectres dans les clusters.

Les étapes de l'algorithme FCM-R sont les mêmes que celles du FCM, à l'exception de l'étape 3 qui englobe cette fois le calcul des facteurs de covariance.
Etapes 3 : Calculer les facteurs de covariance ${\alpha_i}^{(t)}$

$$ {\alpha_i}^{(t)} = \left[\sum_{j=1}^{J} \left| cov(\underline{m_i}^{(t)}, \underline{x}_j^{case}) \right|\right]^{-1} \forall\, i = 1 \dots \mathrm{K} \qquad \text{(eq.2.56)} $$

et réajuster la matrice d'appartenance suivant la position des centres de classes

$$ {u_{ij}}^{(t)} = \left[\sum_{l=1}^{K} \left[\frac{{\alpha_i}^{(t)} (\underline{x}_j^{case} - \underline{m_i}^{(t)})^T (\underline{x}_j^{case} - \underline{m_i}^{(t)})}{{\alpha_l}^{(t)} (\underline{x}_j^{case} - \underline{m_l}^{(t)})^T (\underline{x}_j^{case} - \underline{m_l}^{(t)})}\right]^{\frac{1}{m-1}}\right]^{-1} \forall\, j = 1 \dots J, i = 1 \dots K \qquad \text{(eq.2.57)} $$

Nous notons que l'algorithme FCM est un cas particulier de la FCM-R, lorsque le facteur de covariance est égal à l'unité.

II.4.5.2. Analyse du comportement de l'algorithme FCM-R
Pour comparer les performances de l'algorithme proposé FCM-R avec les algorithmes FCM, GG, GK, FCM-M, FCM-CM et FCM-SM, nous nous concentrons ici sur un exemple pédagogique en considérant des données bidimensionnelles non-sphériques (elliptiques) appartenant à deux classes.

Un ensemble de données composé de 200 échantillons pour la classe c1 et 700 échantillons pour la classe c2 a été généré de façon aléatoire en utilisant des distributions gaussiennes bidimensionnelles. Les vecteurs moyens et les matrices de covariances sont m(1) = [3 ; 4] et σ(1) = [$\sigma_1^{(1)}$ 0 ; 0 $\sigma_2^{(1)}$], avec $\sigma_1^{(1)}$=0.5 et $\sigma_2^{(1)}$ = 1 pour la première classe, et m(2) = [0.5 ; 2] et σ(2) = [$\sigma_1^{(2)}$ 0 ; 0 $\sigma_2^{(2)}$], avec $\sigma_1^{(2)}$=2 et $\sigma_2^{(2)}$ = 0.2 pour la deuxième classe. La Figure 2.10 (a) représente cet ensemble de données dans lequel apparaissent les classes structurellement elliptiques et superposées avec des tailles asymétriques.

Les algorithmes de classification mentionnés ci-dessus ont été appliqués sur cet ensemble de données. Les paramètres choisis pour ce test sont les mêmes : le nombre de classes = 2, le coefficient flou m = 2, la tolérance $\varepsilon = 10^{-6}$. Pour ce test, étant donné le nombre important des données, nous n'avons pas effectué un sur-échantillonnage. Les résultats des classifications sont représentés dans les Figures 2.10 (b)-(h). Les coefficients d'appartenance représentent le background de chaque figure. Notons que le centre de la deuxième classe peut être identifié comme le point brillant du background de chaque figure et celui de la première classe comme le point le plus foncé.

Pour chaque algorithme, nous calculons les pourcentages de bonnes classifications après le processus de défuzzification, qui consiste à choisir la classe ayant la valeur d'appartenance la plus élevée. Les résultats obtenus sont présentés dans le Tableau 2.2. Nous en déduisons que les classes structurelles non-sphériques sont mieux classées par l'algorithme FCM-R que par les autres algorithmes de classification.

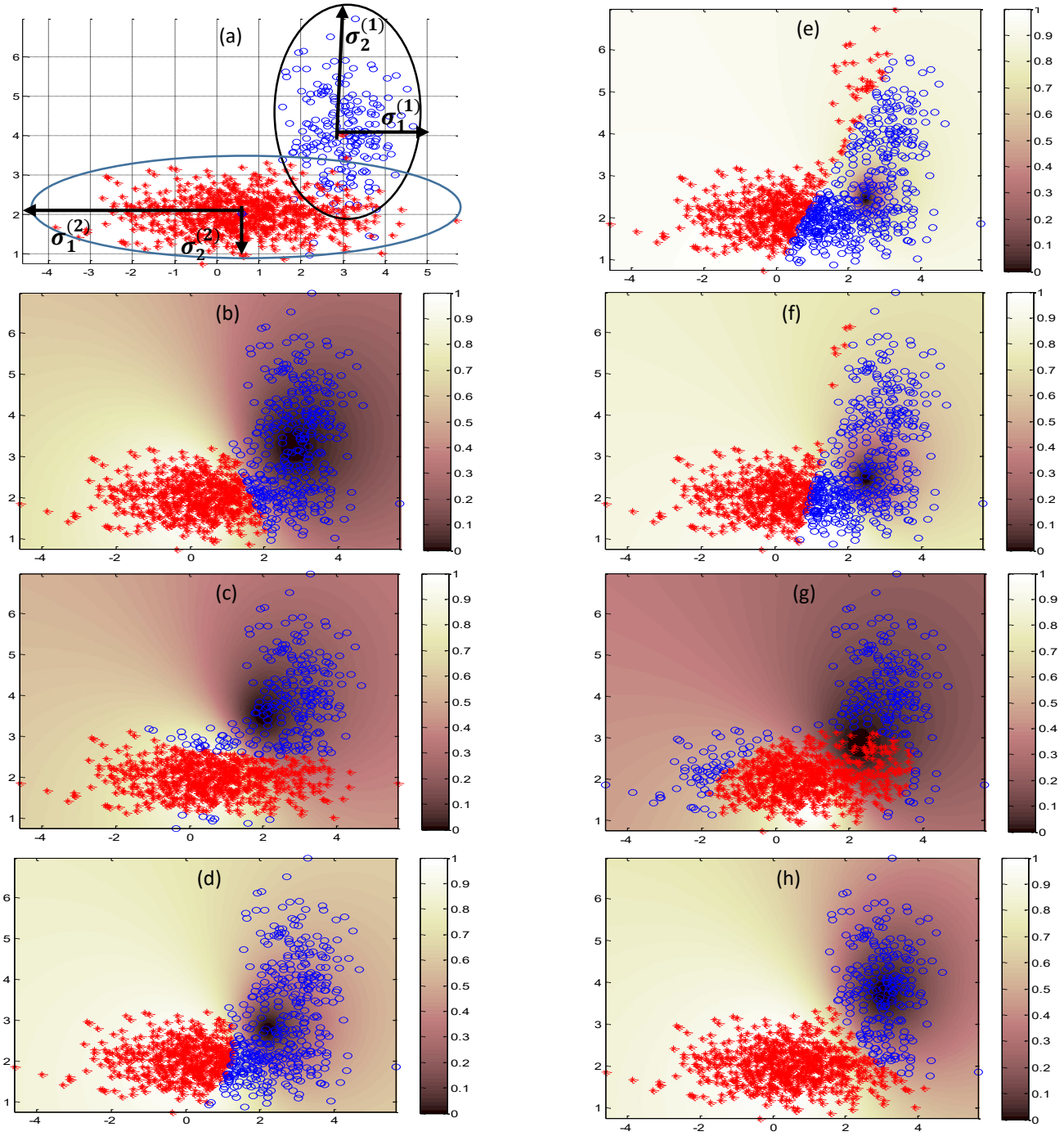


Figure 2.10 (a) Données générées aléatoirement : "o" représentent les 200 échantillons de la classe c_1 et "+" les 700 échantillons de la classe c_2. (b)-(h): Résultats des méthodes de classifications : (b) FCM, (c) GG, (d) GK, (e) FCM-CM, (f) FCM-SM, (g) FCM-M, (h) FCM-R [RPV15]

Pour confirmer que l'algorithme proposé FCM-R s'adapte aux données structurellement sphériques et non sphériques, nous avons étudié les performances des algorithmes de classifications FCM et FCM-R lorsque la forme géométrique des deux classes varie. Pour cela, nous avons généré plusieurs ensembles de données en faisant varier les ratios $\sigma_1^{(1)}/\sigma_2^{(1)}$ et $\sigma_1^{(2)}/\sigma_2^{(2)}$ de façon indépendante pour chaque classe c_1 et c_2. Les vecteurs moyens ont également été modifiés afin de maintenir le même taux de superposition des deux classes. Les tailles des classes ont été maintenues inchangées. Les classes c1 et c2 varient d'un ellipsoïde horizontal si $\sigma_1^{(.)}/\sigma_2^{(.)} < 1$ à un ellipsoïde vertical si $\sigma_1^{(.)}/\sigma_2^{(.)} >$

1, passant par la forme circulaire (sphérique) lorsque $\sigma_1^{(.)}/\sigma_2^{(.)} = 1$. Les Figures 2.11 (a)-(b) montrent les pourcentages de bonnes classifications des algorithmes FCM et FCM-R en fonction des ratios $\sigma_1^{(1)}/\sigma_2^{(1)}$ et $\sigma_1^{(2)}/\sigma_2^{(2)}$. D'après les deux figures, si les deux classes de données ont des formes sphériques, les algorithmes FCM et FCM-R donnent les meilleurs pourcentages de classification. Toutefois, pour les formes non sphériques, la précision de la classification diminue significativement pour l'algorithme FCM. L'algorithme proposé FCM-R donne les meilleurs pourcentages de bonnes classifications dans les deux cas : les classes structurellement sphériques et non sphériques.

Tableau 2.2. Pourcentages de bonnes classifications pour les différents algorithmes de classification sur l'ensemble des données représentées sur la Figure 2.10.

Algorithme	Pourcentage de bonne classification (%)
FCM	83.1 %
GG	92.2%
GK	75.1%
FCM-M	61.7%
FCM-SM	69.9%
FCM-M	88.2%
FCM-R	**96.4%**

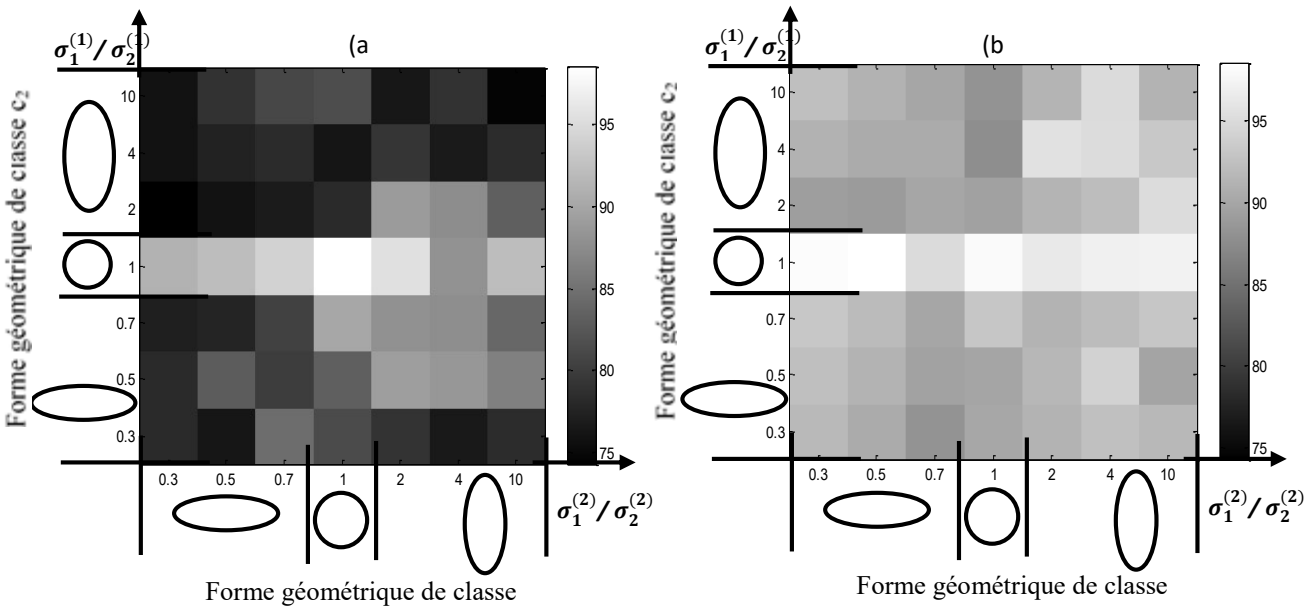


Figure 2.11. Pourcentages de bonnes classifications en fonction de la variation de la forme géométrique des deux classes pour : (a) l'algorithme FCM (b) l'algorithme proposé FCM-R [RPV15]

II.4.5.3. Benchmarks sur des données synthétiques et réelles connues

Nous testons ici les performances des algorithmes FCM, GK, GG, FCM-M, FCM-CM, FCM-SM, et FCM-R sur trois ensembles de données artificielles et quatre ensembles de données réelles. Les ensembles de données artificielles aléatoires sont DataSet-3-3, DataSet-4-3, et DataSet-5-2. Les ensembles de données réelles sont Iris, Wisconsin Breast Cancer (WBCD), Wisconsin Diagnostic Breast Cancer (WDBC), et Wine. Les informations sur les ensembles de données sont présentées dans le Tableau 2.3 [BKM98, Rez10].

Chaque algorithme de classification a été appliqué 100 fois pour les ensembles de données réelles avec des initialisations aléatoires. Pour les ensembles de données artificielles, les données ont été générées

100 fois et pour chaque réalisation, tous les algorithmes avec des initialisations aléatoires différentes ont été appliqués. Les paramètres choisis sont les mêmes que précédemment. Nous mesurons les moyennes des pourcentages de bonnes classifications sur les 100 réalisations. Les résultats finaux sont présentés dans le Tableau 2.4.

Tableau 2. 3. Informations sur les ensembles utilisés de données

Ensemble de données	Dimension	Classes	Taille des échantillons
Iris	4	3	150 (50 pour chaque classe)
WDBC	30	2	569 (357 pour c_1, 212 pour c_2)
WBCD	9	2	583 (444 pour c_1, 239 pour c_2)
Wine	13	3	178 (59 pour c_1, 71 pour c_2, 48 pour c_3)
DataSet-3-3	3	3	150 (50 pour chaque classe)
DataSet-4-3	3	4	400 (100 pour chaque classe)
DataSet-5-2	2	5	500 (100 pour chaque classe)

Tableau 2.4. Moyennes des pourcentages de bonnes classifications au cours des 100 réalisations pour les ensembles de données réelles et simulées.

	Données réelles				Données simulées		
Algorithme	Iris	WDBC	WBDC	Wine	DataSet-3-3	DataSet-4-3	DataSet-5-2
FCM	89.33%	84.53%	94.53%	68.52%	94.66%	100%	92.80%
GK	**90.00%**	72.20%	92.37%	60.97%	91.33%	99.52%	**93.44%**
GG	71.73%	80.57%	91.91%	61.04%	70.96%	48.98%	66.12%
FCM-M	89.32%	84.50%	73.16%	53.12%	67.24%	48.19%	59.96%
FCM-CM	89.33%	84.50%	95.44%	60.19%	69.98%	53.84%	59.49%
FCM-SM	89.33%	84.51%	76.32%	59.99%	73.25%	51.55%	60.52%
FCM-R	**90.00%**	**88.18%**	**96.13%**	**71.32%**	**95.34%**	**100%**	88.82%

Pour les données Iris, l'algorithme FCM-R donne 90,00% ce qui, avec le résultat de l'algorithme GK, représente le meilleur résultat. En ce qui concerne les deuxième, troisième et quatrième ensembles de données réelles, l'algorithme FCM-R donne les plus grandes valeurs moyennes de bonne classification (88,18% pour WDBC, 96,03% pour WBDC et 68,58% pour Wine). Dans le cas des données artificielles, le FCM-R donne les meilleurs résultats pour les DataSet-4-3 et 3-3-DataSet, (100% et 95,34% respectivement). Les résultats pour le DataSet-5-2 sont meilleurs avec l'algorithme GK. Ces résultats indiquent que les performances de l'algorithme proposé FCM-R sont globalement meilleures pour des données artificielles et réelles bien connues. A partir de ces tests, nous pouvons conclure que l'algorithme FCM-R proposé permet une meilleure classification puisqu'il permet de bien classer les clusters de structures non sphériques et sphériques et tolère des classes de tailles inégales.

Nous pouvons remarquer que les algorithmes GK, GG, FCM-M, FCM-CM et FCM-SM ne peuvent pas être utilisés pour la classification des spectres infrarouges de grande dimension puisque le calcul de déterminants de matrices est très coûteux en calcul et la solution numérique devient instable. Pour notre application, les vecteurs de la matrice Σ obtenue par les algorithmes GK, GG, FCM-M, et FCM-CM (ou la matrice R dans le FCM-SM) sont fortement corrélés. C'est une condition très défavorable qui ne permet pas de calculer l'inverse (Σ^{-1}) de la matrice Σ (problème de conditionnement). Par conséquent, il nous est impossible d'appliquer ces algorithmes sur des spectres IR.

II.4.5.4. Benchmarks sur des données IR

Nous testons ici les performances de l'algorithme FCM-R sur des spectres MIR prétraités par différentes méthodes de prétraitements pour les données de biomasse lignocellulosique. Nous

considérons ici les échantillons de maïs sans prise en compte de la cinétique de dégradation (détails dans la section 1.6). Ces échantillons doivent être classés par rapport au type de génotype.
Les méthodes de prétraitement qui ont été appliquées sur les spectres MIR sont : SNV, LB (d=5), LB (d=6) suivie de la SNV, SG (l=14, d=4) d'ordre 1 suivie SNV, SG (l=17, d=4) d'ordre 2 suivie SNV, MSC et EMSC (d=4). Les paramètres sont ceux optimaux qui ont été identifiés dans la section 2.5. Nous avons fait un ré-échantillonnage par bootstrap avec T = 1000.
Le but de ce test est de comparer la performance de classification de notre algorithme proposé avec l'algorithme classique FCM. Les Figure 2.12 et Figure 2.13 montrent les résultats représentés sous forme de boxplot des classifications de FCM-R et FCM pour différentes méthodes de prétraitements.

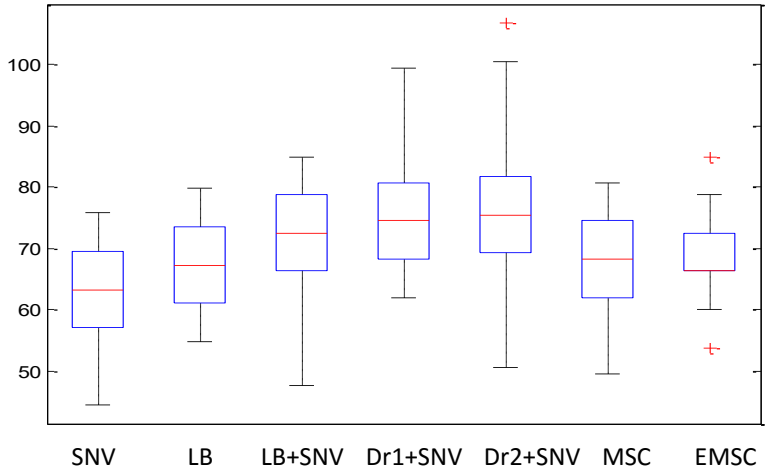


Figure 2.12. Pourcentages de bonnes classifications avec l'algorithme proposé FCM-R pour les spectres MIR prétraités par différentes méthodes de prétraitements.

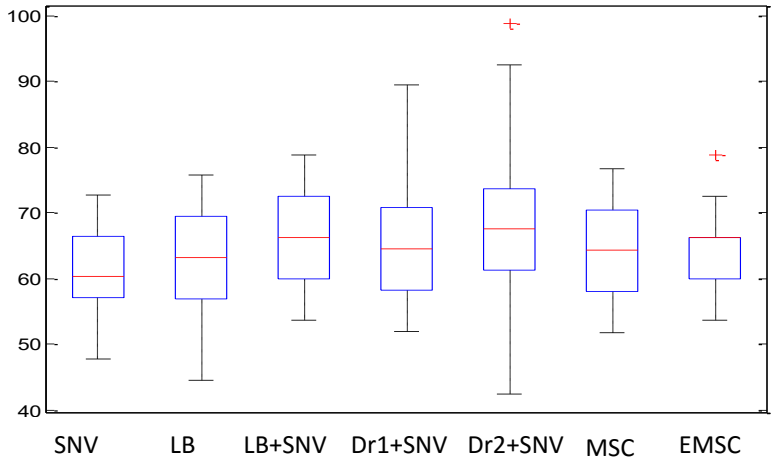


Figure 2. 13. Pourcentages de bonnes classifications avec l'algorithme FCM pour les spectres MIR prétraités par différentes méthodes de prétraitements.

D'après ces figures, nous avons trouvé que l'amplitude de variation des moyennes de bonnes classifications en fonction des méthodes de prétraitements est plus forte avec l'algorithme proposé FCM-R (environ 62-75%) qu'avec l'algorithme FCM (environ 61-68%). Ce résultat revient à la distribution de ces spectres IR qui peuvent être regroupés dans des classes de formes non sphériques, probablement à cause des prétraitements. De plus, nous avons trouvé que l'algorithme FCM-R que nous avons proposé donne de meilleures performances de classifications que l'algorithme FCM

classique sur toutes les méthodes de prétraitements testées, parce qu'il est mieux adapté à la fois aux classes sphériques et non sphériques.

II.5. Choix de méthodes prétraitements et gammes par classification non supervisée

Comme nous avons pu le voir dans la première partie de ce chapitre, le choix d'un prétraitement optimal est primordial pour l'étude ultérieure de spectres IR. Le but de cette partie est donc d'étudier l'influence des principales méthodes de prétraitement et des paramètres associés, ainsi que le choix des gammes spectrales lorsque l'on tente de classer les échantillons de la biomasse lignocellulosique. Cette méthodologie ainsi que certains résultats obtenus initialement avec l'algorithme K-means puis avec l'algorithme FCM ont été publiés dans les articles [RPC13, RPV14].
Dans la suite de cette partie, nous exposons les résultats de classification FCM-R pour des spectres MIR et NIR sur différentes gammes spectrales avec différentes méthodes de prétraitements pour deux jeux de spectres MIR et NIR enregistrés sur deux types différents de biomasses lignocellulosiques.

II.5.1. Application sur la biomasse lignocellulosique sans prise en compte la cinétique de dégradation

Dans un premier temps nous considérons le jeu composé de 16 spectres enregistrés sur des échantillons de maïs (racines) de trois génotypes différents : 5 de type F2, 5 de type F292 et 6 de type G que nous cherchons à séparer en 3 classes, soit les trois génotypes (voir la section « Biomasse lignocellulosique » du chapitre 1). L'algorithme FCM-R a été appliqué sur les spectres prétraités par différentes méthodes de prétraitements. Comme précédemment, nous avons choisi un coefficient m=2, quant au bootstrap le nombre de rééchantillonnages a été fixé à T=1000 et la tolérance $\varepsilon = 10^{-6}$.
Le deuxième jeu est composé de 12 spectres enregistrés sur des échantillons de miscanthus (parties aériennes) de deux récoltes : précoce (6 échantillons) et tardive (6 échantillons) que nous cherchons à séparer en 2 classes (voir la section « Biomasse lignocellulosique » du chapitre 1). Les résultats de classifications FCM-R pour les données de miscanthus sont présentés dans l'annexe du chapitre 2.

Au niveau des méthodes de prétraitements nous avons choisi la dérivée d'ordre 1 et 2 de type Savitzky-Golay (SG) avec lissage sur L = {8, 11, 14, 17 et 20} points et des polynômes d'ordre d = {3, 4, 5, 6,7}. Pour le prétraitement LB, l'ordre du polynôme varie dans une plage usuelle, d = {4, 5, 6, 7}. L'ordre du polynôme de la méthode EMSC est d = {3, 4, 5, 6, 7}.
En ce qui concerne les paramètres des algorithmes FCM-R, nous avons choisi un coefficient flou m=2, quant au bootstrap le nombre de ré-échantillonnages a été fixé à T=1000. Pour chaque étude, nous avons utilisé les méthodes suivantes de prétraitements et leurs combinaisons que nous avons identifiées précédemment, à savoir : SNV, MSC, LB d'ordre d, EMSC d'ordre d = {4, 5, 6, 7}, LB d'ordre d suivie d'une SNV, la dérivée SG d'ordres 1 et 2 suivie d'une SNV.

La Figure 2.14 représente les résultats de bonnes classifications en considérant les spectres MIR sur différentes gammes spectrales avec différentes méthodes de prétraitements pour les données biomasses lignocellulosiques de maïs. D'après cette figure, nous avons trouvé que l'amplitude de variation des moyennes de bonnes classifications en fonction des méthodes de prétraitements est plus forte avec la gamme spectrale 800-1800 cm^{-1} (environ 61-80%) qu'avec la gamme spectrale 400-4000 cm^{-1} (environ 55-72%). Dans cette figure, les valeurs indiquées par les flèches représentent les maximums des valeurs de bonnes classifications. Nous constatons que la gamme 800-1800 cm^{-1} donne les meilleurs résultats de classification pour les deux biomasses lignocellulosiques considérées (voir aussi l'annexe du chapitre 2), notamment pour deux méthodes de prétraitements LB d'ordre 6 suivie d'une SNV (LB 6 + SNV) et dérivée SG d'ordre 1 suivie d'une SNV (SG 1 + SNV).

Certaines explications peuvent être apportées :

- la gamme spectrale 800-1800 cm^{-1} permet d'inclure tous les pieds des pics IR qui informent sur les vibrations des groupes chimiques des composés. La gamme 900-1800 cm^{-1} est privée de cette information. Quant à la gamme totale 400-4000 cm^{-1}, celle-ci contient des bandes spectrales sans pics IR qui perturbent les résultats de classifications des échantillons de biomasses lignocellulosiques.
- Les meilleurs choix de prétraitements changent en fonction du type de données de la biomasse lignocellulosique. En effet, la méthode LB 6 + SNV et la dérivé SG d'ordre 1 suivie de la SNV sont les meilleurs prétraitements pour les spectres de maïs alors que la meilleure méthode est la LB pour les spectres de miscanthus (cf. annexe). Le choix d'une méthode de prétraitement est donc lié directement au type d'échantillon de biomasse lignocellulosique étudié. Il diffère également lorsque nous étudions le processus de dégradation.

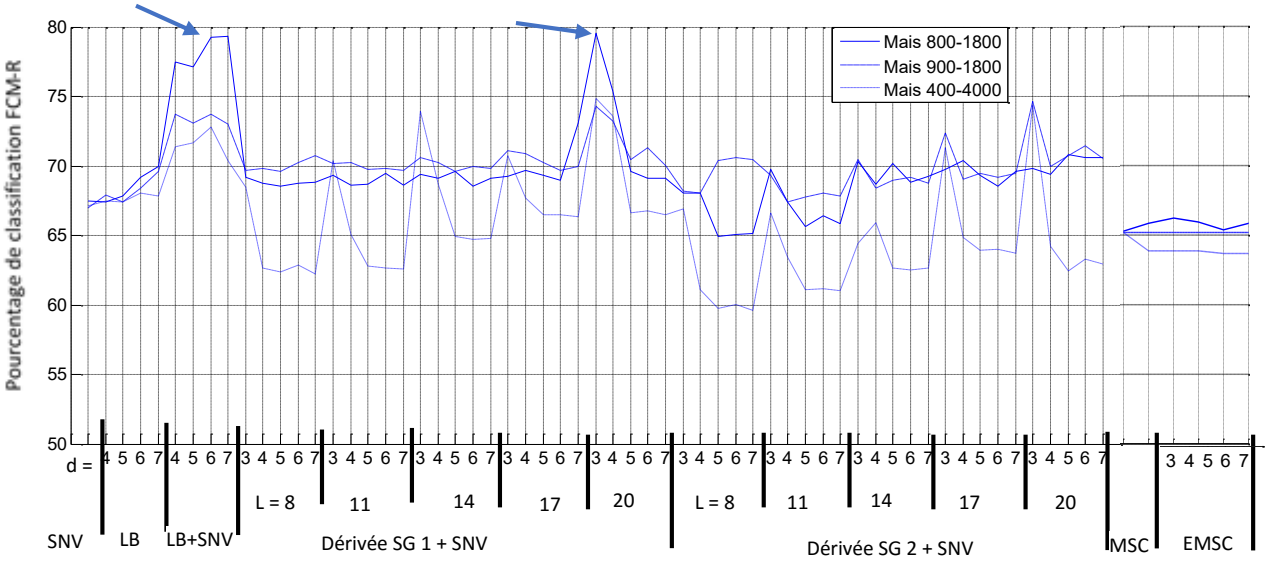


Figure 2.14. Pourcentages de bonnes classifications. FCM-R-bootstrap appliqué sur les spectres MIR enregistrés sur des échantillons de maïs. Gammes spectrales 800 –1800 cm^{-1} (continu), 900 – 1800 cm^{-1} (discontinu), 400 - 4000 cm^{-1} (pointillé) pour différentes méthodes de prétraitements.

La Figure 2.15 représente les résultats de bonnes classifications FCM-R pour les spectres NIR sur différentes gammes spectrales avec différentes méthodes de prétraitements pour les données de biomasses lignocellulosiques de maïs. Les valeurs indiquées par des flèches représentent les maximums des valeurs de bonnes classifications.

Nous pouvons constater que dans ce cas les méthodes de prétraitements influent peu sur les résultats de classifications, les différences entre les méthodes de prétraitement étaient moins importantes que précédemment. Néanmoins, le prétraitement SG 1 + SNV fournit également les meilleurs résultats. Pour les spectres NIR de miscanthus (cf. annexe 2), les dérivés SG d'ordre 1 et 2 suivies de la SNV donnent également les meilleurs résultats.

En ce qui concerne la gamme spectrale, les meilleurs résultats sont obtenus dans la gamme 4000-8000 cm^{-1} pour le maïs et 4000-6000 cm^{-1} pour le miscanthus (voir annexe 2), mais les valeurs sont très proches entre les deux gammes. Cela est dû à la nature des spectres NIR qui ne contiennent qu'un petit nombre de bandes d'absorption très larges (pas des pics spécifiques). La forme caractéristique

des spectres NIR influe peu sur les résultats de classification en donnant des résultats assez semblables pour les trois gammes.

En ce qui concerne de la comparaison entre MIR et NIR, nous avons trouvé que l'application de l'algorithme FCM-R sur les spectres MIR (environ 80% de bonnes classifications pour la méthode SG1 + SNV) donne des résultats meilleurs que sur les spectres NIR (environ 70% de bonnes classifications pour la méthode SG 2 + SNV).

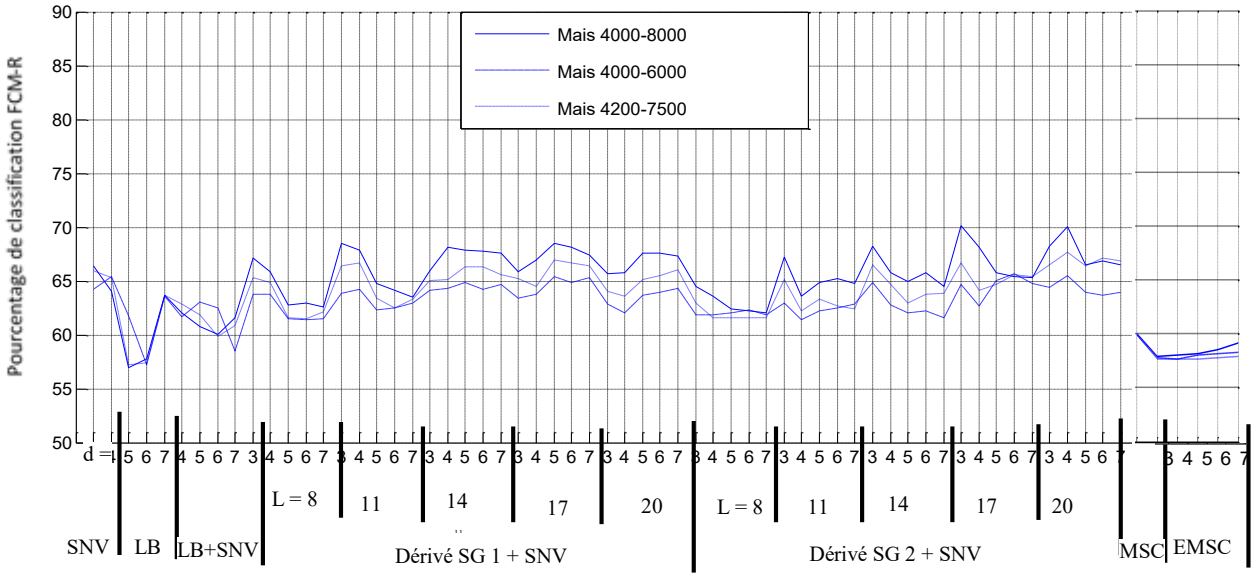


Figure 2.15. Pourcentages de bonnes classifications. FCM-R-bootstrap appliqué sur les spectres NIR enregistrés sur des échantillons de maïs. Gammes spectrales 4000 –8000 cm^{-1} (continu), 4000 – 6000 cm^{-1} (discontinu), 4200 - 7500 cm^{-1} (pointillé) pour différentes méthodes de prétraitements.

II.5.2. Comparaison des résultats obtenus par spectroscopie avec ceux par analyse chimique

Les compositions des racines de maïs et des parties aériennes de miscanthus (sans prise en compte de la cinétique de dégradation) peuvent être déterminées par analyse chimique comme indiqué dans le chapitre 1. La méthode consiste à quantifier la fraction soluble comme la fraction éliminée par une solution de détergent neutre (NDF). Le restant des fibres après l'étape NDF est ensuite utilisé pour quantifier la cellulose et l'hémicellulose (méthode basée sur la composition en monosaccharide après hydrolyse acide de NDF), la lignine et les esters (acides phénoliques) sont enfin estimés en utilisant la procédure détaillée par Huyen [HRD10].

Nous avons appliqué notre méthode de classification FCM-R-bootstrap sur les valeurs fournies par l'analyse chimique tout en tenant compte que les concentrations se réfèrent soit à de la matière brute (fraction soluble inclue) ou aux résidus NDF (fraction soluble non prise en compte). Nous avons appliqué la classification FCM-R-bootstrap sur les vecteurs composés des concentrations en : Cellulose, Hémicellulose, Lignine et Esters (CHLE). Nous présentons ici les résultats obtenus sur des échantillons bruts et NDF par analyse chimique, puis nous comparons uniquement avec les résultats de classification sur les spectres MIR et NIR des échantillons de maïs et miscanthus bruts.

La Figure 2.16 montre que les performances moyennes de la FCM-R-bootstrap, exprimées en pourcentage de bonnes classifications, varient de 70-77%. Les résultats obtenus affichent des variations significatives entre les échantillons de maïs et de miscanthus ainsi qu'entre des échantillons

bruts et NDF. La classification basée sur les concentrations comprenant les fractions solubles (brutes) est significativement moins performante dans le cas du maïs alors qu'on observe l'inverse pour le miscanthus. Ce résultat peut être expliqué par la variation de la fraction soluble dans le miscanthus en fonction de la date de récolte. La récolte précoce de miscanthus contient une quantité plus élevée de fraction soluble alors que les racines de maïs montrent moins cette variation.

La classification FCM-R sur les spectres MIR des racines de maïs fournit des résultats allant de 72 à 80% sur les fichiers bruts (respectivement 84 à 87 % pour le miscanthus voir annexe 2). En comparaison (Figure 2.16), les résultats de classification chimique sont de 70% pour les échantillons bruts de maïs (respectivement 77 % pour les échantillons NDF) et de 75% pour le miscanthus brut (respectivement 68% pour le NDF). Nous pouvons en conclure que la classification des végétaux est meilleure en utilisant l'information spectrale MIR que celle chimique. Cependant si l'on regarde les résultats obtenus sur les spectres NIR les classifications ne sont pour les spectres bruts que de 67% pour le maïs et de 62 % pour le miscanthus. La classification par rapport aux spectres NIR se situe donc en retrait par rapport à celle réalisée à partir des analyses chimiques.

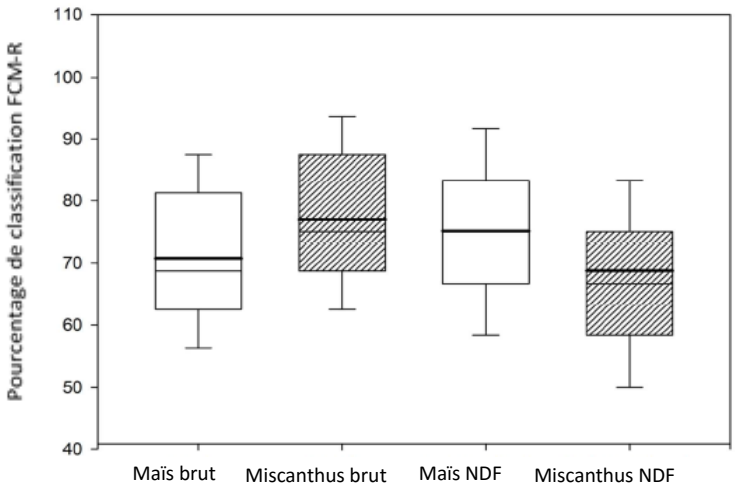


Figure 2.16. Pourcentages de bonnes classifications. FCM-R-bootstrap appliqué sur les concentrations chimiques CHLE pour les échantillons de biomasse lignocellulosique : maïs et miscanthus.

II.5.3. Application sur la biomasse lignocellulosique avec prise en compte la cinétique de dégradation

Dans un second temps, nous nous intéressons à la biodégradation. Nous considérons 4 échantillons de maïs (racines), de quatre génotypes différents (F2 et F292, F2bm1 et F292bm3), qui représentent la biomasse lignocellulosique que nous avons analysé pendant 5 périodes de biodégradation : t_1=0, t_2=14, t_3=36, t_4=57, t_5=112 jours. Nous avons détaillé ces échantillons de biomasse lignocellulosique dans le chapitre 1 (section « Biomasse lignocellulosique »). Des spectres MIR (400 - 4000 cm^{-1}) et NIR (4000 - 8000 cm^{-1}) ont été enregistrés sur ces échantillons. Nous disposons donc de 20 spectres que nous cherchons à séparer en 5 classes correspondant aux périodes de biodégradation. Nous avons utilisé les mêmes paramètres FCM-R-bootstrap que précédemment (m=2, T = 1000, tolérance $\varepsilon = 10^{-6}$).

La Figure 2.17 représente les résultats de bonnes classifications en considérant les spectres MIR pour différentes gammes spectrales avec différentes méthodes de prétraitements. Dans cette figure, la valeur indiquée par la flèche représente le maximum obtenu. Nous avons trouvé que la gamme 800-1800 cm^{-1} donne également les meilleurs résultats de classifications avec la méthode de

prétraitements SG d'ordre 1 suivie de la SNV. Si l'on compare ces résultats avec les résultats de classifications pour les échantillons de biomasse lignocellulosiques sans dégradation, nous retrouvons la même meilleure méthode de prétraitements SG d'ordre 1 + SNV. Cependant les échantillons sans biodégradation admettent également la LB 6 + SNV comme méthode optimale de prétraitement. Nous pouvons en conclure que le choix du prétraitement peut différer également lorsque nous étudions un processus de dégradation.

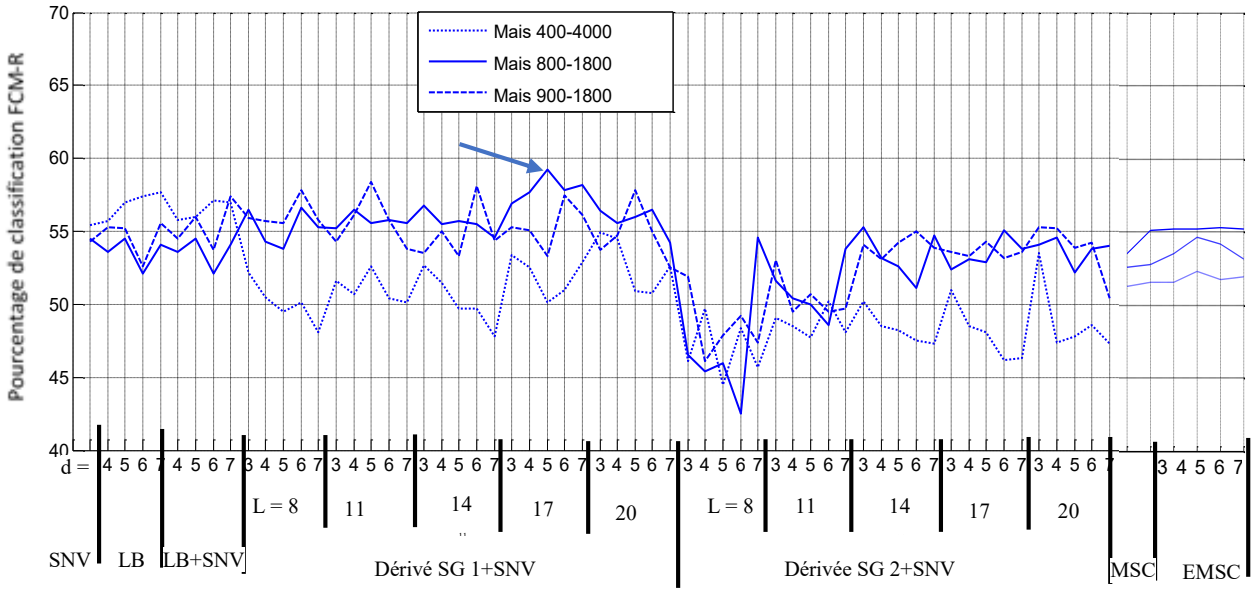


Figure 2.17. Pourcentages de bonnes classifications. FCM-R-bootstrap appliqué sur les spectres MIR enregistrés sur des échantillons de maïs subissant une biodégradation. Gammes spectrales 800 –1800 cm^{-1} (continu), 900 – 1800 cm^{-1} (discontinu), 400 - 4000 cm^{-1} (pointillé) pour différentes méthodes de prétraitement.

La Figure 2.18 montre les résultats de classification par FCM-R pour les spectres NIR, la valeur indiquée par la flèche représentant également le maximum de bonnes classifications. Nous trouvons que la gamme 4000-6000 cm^{-1} donne les meilleurs résultats de classification avec la méthode de prétraitement SG d'ordre 1 suivie de la SNV. Mais les résultats de classification dans la gamme 4000-6000 cm^{-1} sont très proches de ceux obtenus sur les autres gammes du NIR, en raison de la nature spectroscopique du proche infrarouge. En comparaison avec les résultats de classifications pour les échantillons de biomasse lignocellulosiques sans dégradation, nous retrouvons la même méthode de prétraitements (les dérivés SG d'ordre 1 + SNV) pour les deux (les échantillons maïs avec et sans biodégradation).

En ce qui concerne la comparaison entre les spectres MIR et NIR, nous avons trouvé que l'application de l'algorithme FCM-R sur des spectres MIR donne environ 58% de bonnes classifications pour la méthode SG 1 + SNV, alors que celle-ci est d'environ 56 % pour la méthode SG 1 + SNV pour les spectres NIR.

Nous pouvons en conclure que globalement la meilleure méthode de prétraitement pour l'étude des échantillons de biomasse lignocellulosique et de leur biodégradation des spectres MIR et NIR est la dérivée SG d'ordre 1 suivie de la SNV.

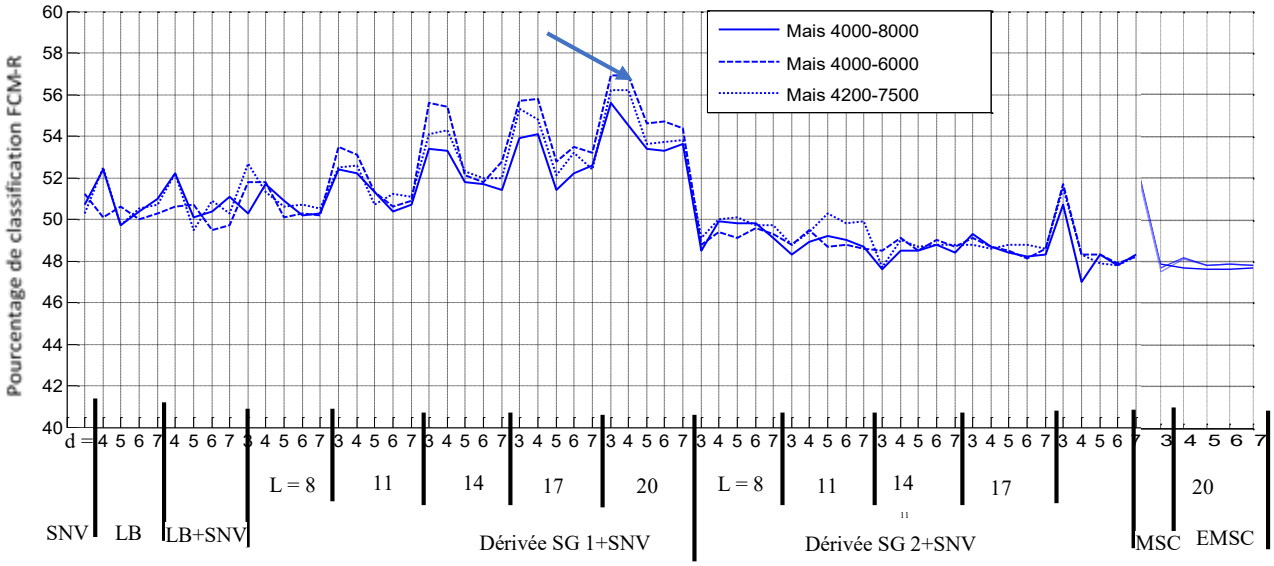


Figure 2.18. Pourcentages de bonnes classifications. FCM-R-bootstrap appliqué sur les spectres NIR enregistrés sur des échantillons de maïs subissant une biodégradation. Gammes spectrales 4000- 8000 cm^{-1} (continu), 4000-6000 cm^{-1} (discontinu), 4200-7500 cm^{-1} (pointillé) pour différentes méthodes de prétraitements.

II.6. Conclusion

Nous avons vu dans ce chapitre quelles sont les différentes méthodes de prétraitement des spectres IR, puis nous nous sommes orientés vers l'utilisation de méthodes de classification pour déterminer le ou les prétraitements optimaux de nos données. Nous avons proposé d'utiliser uniquement l'information spectrale MIR et NIR à travers des algorithmes de classification non supervisés qui permettent de réaliser ce choix en optimisant la classification de la biomasse lignocellulosique.

Pour cela, nous avons choisi le Fuzzy C Means (FCM) que nous avons modifié. En effet, l'un des paramètres les plus importants étant la métrique entre deux spectres, nous avons proposé une nouvelle extension de l'algorithme FCM, appelée FCM-R, basée sur une distance Euclidienne pondérée par un facteur de covariance. Cet algorithme dépasse, que ce soit sur données synthétiques ou réelles, les performances de l'algorithme FCM. Nous avons étudié ensuite l'influence des principales méthodes de prétraitement et ses paramètres associés lors de classification des spectres MIR et NIR qui ont été acquis sur les échantillons de biomasses lignocellulosiques de racines de maïs et de miscanthus.

Nos résultats montrent que la spectroscopie IR est une méthode robuste qui donne des résultats bons et cohérents sur différents types de végétaux. Nous avons déterminé les gammes spectrales optimales. Pour les spectres MIR la gamme spectrale 800-1800 cm^{-1} se démarque des autres gammes en termes de classification, les prétraitements LB+SNV et SG1 + SNV étant optimaux.

Nous avons également montré que la gamme spectrale 4000-6000 cm^{-1} du NIR associée à la méthode de prétraitement SG d'ordre 1 suivie d'une SNV donne également les meilleurs résultats de classification FCM-R. Pour cette dernière les performances observées sur cette gamme restent néanmoins très proches de celles obtenues avec d'autres gammes. Nous avons également montré que l'information spectrale est pertinente par rapport à l'information chimique classiquement utilisée dans ce type d'application (concentrations des différents constituants de la biomasse).

Chapitre 3 : Sélection de bandes spectrales discriminantes dans le processus de biodégradation

III.1. Introduction

L'objectif de ce chapitre est de sélectionner de bandes spectrales sans *a priori* complémentaires permettant de discriminer la biomasse lignocellulosique au cours du processus de biodégradation. Pour cela, nous proposons la stratégie suivante. Nous cherchons d'abord des méthodes mathématiques qui nous permettent de sélectionner des bandes spectrales discriminantes par rapport à la cinétique de biodégradation. Nous nous intéressons ensuite à la possibilité de combiner les informations spectrales MIR et NIR par concaténation ou par produit extérieur avec pour objectif d'améliorer la discrimination des échantillons lors du processus de biodégradation.

La section suivante analyse différentes méthodes de sélection de variables en mettant l'accent sur les avantages et inconvénients de celles-ci. Nous détaillons ensuite les étapes et le choix des paramètres pour une méthode de sélection automatique des bandes utilisant un algorithme génétique. Ensuite, nous développerons une nouvelle méthode basée sur une approche d'optimisation qui requiert moins des paramètres et qui permet d'extraire également les poids des bandes spectrales. Après avoir présenté les principes des méthodes de validations des résultats telles que l'analyse en composantes principales et l'indice de Dunn, nous proposons une méthodologie qui permet de sélectionner des bandes spectrales discriminantes sur des spectres MIR ou NIR séparés ainsi que sur des spectres combinés par concaténation ou par produit extérieur. Cette méthodologie est validée dans la dernière section sur des données simulées afin de tester les performances potentielles de sélection de bandes spectrales. Les résultats de son application sur des données réelles de différentes biomasses lignocellulosiques sont présentés et discutés par la suite.

III.2. Méthodes de sélection de bandes spectrales

La sélection de variables spectrales discriminantes est une étape importante dans l'analyse des données spectroscopiques [VLT14]. Le choix d'un petit nombre de variables sélectionnées dans l'ensemble original de variables devrait faciliter l'analyse et l'interprétation de données spectrales. En effet, la sélection de variables en analyse multivariée est une étape critique dans l'analyse de données car l'élimination des variables non informatives permet de fournir de meilleurs résultats de prédiction avec des modèles plus simples. Pour notre application, les variables sélectionnées correspondent aux « bandes spectrales » ou « nombres d'ondes » puisque nous traitons des données spectroscopiques IR. Le but de cette sélection de variables revient donc à identifier un sous-ensemble de nombres d'ondes dans une région spectrale qui minimise les erreurs lorsqu'elles sont utilisées pour effectuer des opérations comme des déterminations quantitatives ou bien des discriminations entre différents échantillons ou périodes de dégradation.

Un effort considérable a été mené pour développer des procédures qui identifient objectivement les nombres d'ondes qui contribuent à des informations utiles et/ou éliminent les nombres d'ondes contenant souvent du bruit ou des données inutiles. Plusieurs approches existent pour la sélection d'un sous-ensemble de nombres d'ondes optimal. Ainsi, pour des processus de calibration, nous pouvons lister : l'algorithme des projections successives (Successive Projections Algorithm, SPA) [ASG01, GAF08], l'algorithme génétique (Genetic Algorithm, GA) [PFL00], l'algorithme de sélection régressif (Backward Selection Algorithm, BSA), l'algorithme de sélection progressif (Forward Selection Algorithm, FSA), l'algorithme de sélection pas à pas (stepwise regression) [Aka69], la méthode d'élimination de variables non informatives (Uninformative Variable Elimination, UVE) [CM96, CLS08],

la méthode d'échantillonnage adaptatif concurrentiel repondéré (Competitive Adaptive Reweighted Sampling, CARS), etc. [DK92, LLX09, ZLW12]. On peut citer d'autres méthodes de sélection de variables spectrales telles que le modèle ASM (All Subset Models), qui identifie le meilleur sous-ensemble optimal en testant toutes les combinaisons possibles, mais ne peut pas être appliqué en pratique sur des données de grande taille telles que les spectres IR [SVP06]. La méthode de recherche séquentielle SS (Sequential Search), qui est une méthode de méta-heuristique, n'est pas non plus adaptée aux données spectrales IR de grande taille [Mil02, HTF09]. D'autre part, il existe plusieurs approches de sélection de variables basées sur la méthode des moindres carrés partiels (Partial least square, PLS) telles que les : Loading Weights (LW), Regression Coefficients (RCβ), la projection de variable d'importance (VIP), Uninformative Variable Elimination in PLS (UVE-PLS), Iterative Predictor Weighting (IPW-PLS), la régression PLS par intervalle (iPLS), Moving Window Partial Least Squares Regression (MWPLS), la méthode Backward (descendante) avec la régression iPLS (BiPLS), Forward iPLS (FiPLS), la régression PLS par sélection interactive de variables (IVS-PLS), Soft-Threshold PLS (ST-PLS), Powered PLS (PPLS) etc. Le succès de ces approches de sélection de variables spectrales est généralement évalué empiriquement par les statistiques du modèle multivarié résultant (par exemple la racine de l'erreur quadratique moyenne de prédiction RMSEP) [BS11, MLS12, VB10, Mao03, KJ97, SS89, YH98]. Le problème pour notre application dans l'utilisation de ces méthodes, qui sont toutes basées sur la PLS, est le besoin des connaissances *a priori* complémentaires. Or, dans ce chapitre, nous recherchons à effectuer une analyse de la biomasse lignocellulosique par rapport au processus de biodégradation en s'appuyant uniquement sur les informations spectrales.

Dans ce qui suit, nous présentons ci-dessous les principes de fonctionnement de quelques algorithmes utilisés couramment pour parcourir l'espace des modèles et effectuer une sélection des variables en mettant l'accent sur les avantages et inconvénients de celles-ci.

- L'algorithme de sélection régressif (Backward Selection Algorithm, BSA).

La procédure d'élimination progressive démarre en estimant les paramètres du modèle complet incluant toutes les variables explicatives que l'on a sélectionnées et jugées pertinentes à introduire [Mil02]. A chaque étape, la variable ayant la plus grande p-valeur (test de Fisher ou de Student) est éliminée du modèle si cette valeur est supérieure au seuil fixé a priori (en général 10% ou 5%). La procédure s'arrête lorsque les variables restantes dans le modèle ont toutes une p-valeur plus petite que le seuil [HTF09]. En pratique, l'élimination progressive nécessite plus d'opérations parce qu'elle doit généralement écarter plus de variables dépendantes.

- L'algorithme de sélection progressif (Forward Selection Algorithm, FSA)

Comme précédemment, il faut choisir au départ des variables explicatives que l'on juge pertinentes. A chaque étape, une variable est ajoutée en commençant par la variable ayant la plus petite valeur (p-value) obtenue en réalisant l'ensemble des modèles de régression linéaire simple. Ensuite, on évalue l'apport spécifique de chacune des variables non encore introduites dans le modèle qui contient déjà la ou les variable(s) retenue(s) dans les étapes précédentes et on introduit la variable dont l'apport spécifique est le plus important [Mil02, HTF09]. L'introduction d'une nouvelle variable dans le modèle ne se fait que si la p-valeur correspondante est inférieure à un seuil fixé a priori (en général 10% ou 5%). La procédure s'arrête lorsque toutes les variables sont introduites ou lorsqu'on ne peut plus introduire de nouvelles variables selon le critère choisi (p-valeurs supérieures au seuil) [HTF09, Efr60]. Cette méthode ne nous donne aucune garantie de trouver les modèles les mieux ajustés aux données pour la sélection de variables.

- L'algorithme de sélection pas à pas (Stepwise Regression, SR)

Il s'agit d'une amélioration de la méthode FSA. A chaque étape de la procédure, on examine à la fois si une nouvelle variable doit être ajoutée selon un seuil d'entrée fixé et si une des variables déjà incluses doit être éliminée selon un seuil de sortie fixé. Cette méthode permet de retirer du modèle d'éventuelles variables qui seraient devenues moins indispensables du fait de la présence de celles nouvellement introduites. La procédure s'arrête lorsqu'aucune variable ne peut être rajoutée ou retirée du modèle selon les critères choisis [Aka69, BS11].

Ces méthodes (BSA, FSA et SR) ont l'avantage d'être faciles à utiliser et permettent de traiter le problème de la sélection de variables de façon systématique. Un autre avantage est que nous pouvons sélectionner au départ un nombre important ou uniquement quelques bandes spectrales connues a priori, par exemple liées à la lignine, à l'hémicellulose, etc., que nous jugeons pertinentes par rapport au processus étudié. L'inconvénient majeur est que les variables sont retenues ou éliminées du modèle sur la base de critères uniquement statistiques sans tenir compte du contexte ni de l'objectif de l'étude, la notion de perturbation et de modification de l'effet n'est pas prise en compte dans ces méthodes de sélection. Ainsi, ces méthodes ont besoin d'un vecteur de variables dépendantes pour réaliser l'ensemble des modèles de régression linéaire simple.

- Les projections des variables d'importances (Variable Importance in Projection, VIP)

Les projections des variables d'importances (VIP) est une méthode de sélection de variables basée sur la régression PLS [WJC93]. L'algorithme VIP peut parfois conduire à des modèles plus robustes pour la prédiction de variables dépendances et la sélection de variables. Mathématiquement, la méthode VIP consiste à faire une projection des variables dépendantes sur les variables latentes. Cependant, comme nous l'avons mentionné, cette méthode est basée sur la régression PLS qui a besoin de connaissances à priori complémentaires.

- Algorithme des projections successives (Successive Projections Algorithm, SPA)

Il s'agit d'une technique de sélection de variables conçue pour minimiser les problèmes de colinéarité en régression linéaire multiple. La SPA emploie une opération de projection simple dans un espace vectoriel pour sélectionner des sous-ensembles de variables avec un minimum de colinéarité. Ce principe aboutit à une nouvelle variable sélectionnée qui est l'une des variables issue du sous-ensemble de variables qui a la valeur maximale de sa projection orthogonale dans le sous-espace. Cette variable optimale est choisie par évaluation des performances de prédiction du modèle de régression linéaire multiple.

Brièvement, les données spectrales sont disposées en une matrice $X^{case} = \left[\underline{x}_1^{case} \quad \dots \ \underline{x}_j^{case} \ \dots \quad \underline{x}_J^{case}\right] \in M_{J,O}(\mathbb{R})$ avec case = MIR ou NIR et O=P ou Q et $\underline{x}_j^{case} \in \mathbb{R}^{\mathrm{O}}$ correspond au j$^{\text{ième}}$ vecteur de données IR. Soit M = min (O-1, J) le nombre maximum de variables sélectionnées. La première étape consiste en des projections réalisées sur la matrice X, qui génèrent k chaînes de M variables. La deuxième étape consiste à évaluer des sous-ensembles de variables sélectionnées dans la première étape. Un sous-ensemble de variables de dimension M x J est testé et la meilleure variable du sous-ensemble est sélectionnée. Le critère de sélection adopté est la racine de l'erreur quadratique moyenne (RMSE). Le désavantage principal de cet algorithme est la nécessite d'utiliser une méthode de régression linéaire multiple qui exige des informations complémentaires, par exemple des informations chimiques.

Pour cela, nous proposons de traiter ce problème par l'utilisation de méthodes génétiques qui ont déjà été utilisées avec succès dans plusieurs applications telles que la réduction de la dimension dans

l'analyse des données hyper-spectrales, y compris des applications liées à l'agriculture, ainsi que pour l'évaluation de la biodégradation par les spectres NIR et MIR [NF77, YH98].

Dans la suite de ce chapitre, nous allons nous intéresser au principe et aux différentes étapes d'un algorithme génétique, puis, en proposant des choix pertinents pour ces étapes et pour les paramètres mis en jeu, nous allons montrer comment on peut l'utiliser afin de sélectionner des nombres d'ondes discriminantes dans le processus de biodégradation.

III.3. Algorithme génétique (AG)

III.3.1. Généralités

Les algorithmes génétiques sont des méthodes stochastiques basées sur une analogie avec des systèmes biologiques. Ils reposent sur un codage des variables en structures chromosomiques et prennent modèle sur les principes de l'évolution naturelle pour déterminer une solution optimale. Ils ont été initialement développés par Holland [Hol89] et popularisés par Goldberg [Gol89]. Ces algorithmes sont caractérisés par une grande robustesse et possèdent la capacité d'éviter les minima locaux pour effectuer une optimisation globale. Les algorithmes génétiques sont une méthode d'optimisation utile dans les cas non linéaires. Ces algorithmes s'appuient sur le paradigme darwinien de l'évolution génétique d'une population.

L'idée est de générer des populations de N solutions, chaque solution de la population étant représentée sous la forme d'un « chromosome ». Chaque chromosome est lui-même formé d'un nombre restreint, noté par la suite « L », de nombres d'ondes (bandes spectrales) sélectionnés et positionnés comme des « gènes » dans le chromosome. A chaque étape, l'algorithme évalue les chromosomes à travers une fonction fitness. Il conserve les chromosomes ayant les meilleures valeurs de fitness pour la génération suivante. Il combine également les meilleurs chromosomes dans l'étape de croisement, puis il fait subir des mutations aux chromosomes restants [YYS14]. L'algorithme génétique se décompose donc en différentes étapes : l'initialisation des paramètres et de la population, puis les évaluations, les sélections, les recombinaisons, les mutations jusqu'à convergence. Ces étapes et leurs différentes modalités de réalisation sont décrites dans de nombreux articles comme [Hol89, YYS14]. La Figure 3.1 montre les étapes de l'algorithme génétique.

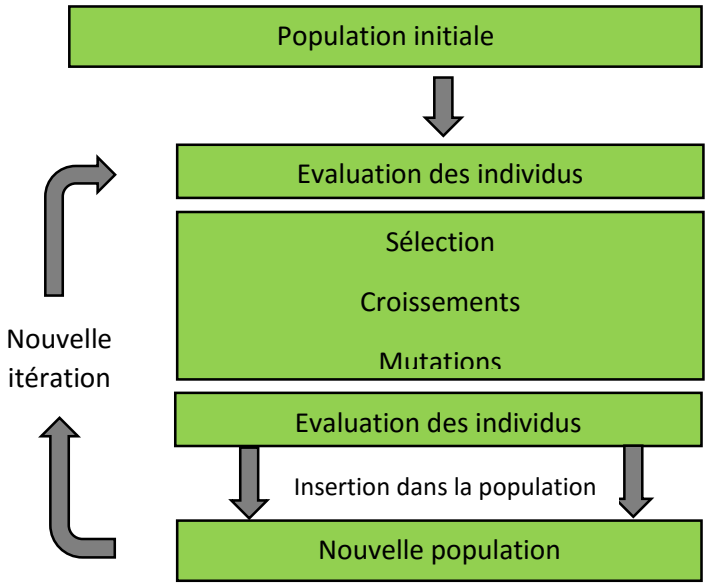


Figure 3.1. Principe de l'algorithme génétique (AG) [Hol89].

Les étapes de l'algorithme AG sont les suivantes :

1. Initialisation des paramètres : la taille des chromosomes L, la taille de la population N, le choix de fonction fitness et objectif, les paramètres pour la sélection, le croisement et la mutation, et les paramètres d'arrêt. En plus de ces paramètres, il existe différentes façons de sélectionner les chromosomes dans chaque génération :
   - La première façon consiste à prendre le nombre de chromosomes sélectionnés pour effectuer les croisements et la mutation égal à N. Ensuite, on effectue les croisements et mutations sur tous les chromosomes dans la population. L'inconvénient de la méthode est qu'elle ne prend pas en compte l'étape de sélection des chromosomes qui ont les meilleures valeurs de fitness.
   - La deuxième façon consiste à choisir N/2 chromosomes pour faire les croisements et N/2 chromosomes pour faire les mutations. Cette méthode présente les mêmes inconvénients que la précédente.
   - La troisième façon consiste à diviser la population N en trois parties : l'une contenant les chromosomes ayant les meilleures valeurs de fitness qui seront gardés inchangés dans la génération suivante, l'une contenant les chromosomes qui vont effectuer les croisements et la dernière pour les chromosomes qui vont subir des mutations. Cette technique augmente la rapidité de convergence de l'algorithme puisqu'elle permet de sélectionner les chromosomes qui ont les meilleures valeurs de fitness et qu'elle n'autorise des croisements et des mutations que sur une partie de la population.

2. Évaluation : la performance de chaque chromosome est évaluée dans la population initiale suivant la valeur de leur fonction fitness. Cette fonction est la mesure de qualité du chromosome. Il existe différentes fonctions fitness permettant l'évaluation : la fonction basée sur le principe de longueur de description minimale (MDL) ; les indices de validités ; le Feature Evaluation Index ; les Ranking Evaluation Functions ; la Fuzzy fitness Function ; la fonction de pénalité ; la fonction homogénéité, etc. [Deb02].
   Les valeurs fitness obtenues pour les N chromosomes sont ensuite ordonnées de manière croissante. Une fonction objectif est ensuite appliquée pour chaque chromosome dans la population. Elle s'appuie sur l'utilisation de l'opérateur position pour chaque chromosome : le chromosome qui a la plus petite valeur de la fonction fitness se trouve en première position. Il existe différentes fonctions objectifs :
   - La fonction proportionnelle qui est l'inverse de valeur fitness pour chaque chromosome. Cette fonction ne traite pas le problème de valeurs fitness très voisines.
   - La fonction rang (Rank) qui est l'inverse de la racine des valeurs de fitness pour chaque chromosome. Cette fonction permet de supprimer l'effet de l'écart des valeurs fitness. De plus, cette fonction est une solution optimale pour le problème de valeurs fitness qui sont très rapprochées [Whi94].

3. Sélection : Le rôle de l'opérateur de sélection est de filtrer la population de manière à conserver les chromosomes possédant de « bonnes » caractéristiques génétiques. La seule mesure dont on dispose de la qualité d'un chromosome est sa performance, liée à la valeur de la fonction fitness. Cette étape est utilisée pour choisir les nouveaux parents (ceux qui ont les meilleurs résultats de fitness) pour effectuer ensuite des croisements et obtenir une nouvelle population. Il y a plusieurs méthodes de sélection :

- La sélection par roulette (Roulette Wheel Selection ou RWS), introduite par Goldberg en 1989 [Gol89]. La population est représentée comme une roue de roulette où chaque chromosome est représenté par une portion dont la largeur est proportionnelle à sa valeur de fitness. L'un des inconvénients de ce type de sélection est de choisir presque toujours le même chromosome s'il en existe un bien meilleur que les autres, ce qui cause une perte de diversité dans la population.
- L'échantillonnage universel stochastique (Stochastic Universal Sampling ou SUS) proposé par Baker en 1987 [Bak87] utilise aussi une roulette partagée en autant de portions (individus), de largeur proportionnelle aux valeurs de fitness. Mais cette fois, les chromosomes sélectionnés sont désignés par un ensemble d'indicateurs équidistants où une rotation de la roue de roulette sélectionne tous les parents simultanément. Cette méthode a comme avantage de ne pas avoir de biais d'estimation et présente une dispersion minimale.
- La sélection par rang (Ranking selection), comporte deux étapes (Whitley, 1989). D'abord tous les chromosomes de la population sont rangés selon leurs valeurs de fitness. Le rang est fait dans l'ordre décroissant (ou croissant), selon que l'on veut minimiser (ou maximiser) la fonction de fitness. Normalement, les chromosomes de moins bonne qualité obtiennent un rang faible (à partir de 1). Mais cette méthode peut aboutir à une convergence plus lente, parce que les meilleurs chromosomes ne diffèrent pas tellement les uns des autres [Whi94].

4. Croisement (Crossover) : Cette étape est utilisée pour recombiner les meilleurs gènes pour obtenir des enfants potentiellement supérieurs. Il y a plusieurs méthodes de croisement :
   - Croisement simple à un point : l'opérateur de croisement le plus simple est le croisement en un point. La première étape consiste à choisir aléatoirement un point (gène) de coupure pour partager chaque parent en deux parties. Puis le premier enfant est construit en utilisant la première partie du premier parent et la deuxième partie du deuxième parent. A l'inverse, le deuxième enfant est une concaténation de la seconde partie du premier parent et de la première partie du second parent. L'inconvénient de cet opérateur est qu'il a de fortes chances de produire des enfants invalides (potentiellement inférieur), par duplication et/ou omission de certains éléments de la permutation.
   - Croisement multi points : cet opérateur est une généralisation du croisement à un point avec n points de croisement. Ces points (gènes) choisis aléatoirement coupent chaque parent en n+1parties. Elle a le même problème que le croisement simple à un point.
   - Croisement uniforme : consiste à choisir les valeurs des gènes des chromosomes fils à partir de leurs homologues dans les chromosomes parents, et cela suivant un processus de décision aléatoire et équiprobable. Le croisement uniforme est un bon opérateur pour éviter des problèmes de locus (partie invariable) des gènes.

5. Mutation : L'opérateur de mutation a pour rôle d'assurer la diversité des solutions pour sortir des minima locaux. Elle consiste à modifier un ou plusieurs gènes d'un individu sélectionné par l'étape de sélection. Il existe plusieurs types de mutation :
   - L'opérateur mutation aléatoire : consiste à changer les valeurs d'un certain nombre de gènes choisis aléatoirement parmi ceux portés par tous les chromosomes de toute la population. Cet opérateur n'est pas toujours bien adapté aux différents types de problèmes existants tels que les données de variables sensibles [FLK03].

- L'opérateur de mutation gaussienne : la mutation la plus employée est la mutation gaussienne, qui consiste à rajouter un bruit gaussien au vecteur des variables.

6. On répète les étapes 2 à 5 jusqu'à ce que le nombre d'itérations maximal T soit atteint, ou l'on s'arrête quand la variation moyenne pondérée des T dernières valeurs de fitness est inférieure à une tolérance ϵ (fixée) [Zel12].

III.3.2. Approche proposée pour des données spectrales IR

Nous rappelons que les données spectrales IR peuvent être conditionnées sous forme matricielle X^{case}, où X^{case} =$\left[\underline{x}_1^{case} \quad \dots \ \underline{x}_j^{case} \dots \ \underline{x}_J^{case}\right] \in M_{J,O}(\mathbb{R})$ est la matrice formée de J spectres infrarouges avec case = MIR, NIR MIR-NIR et MIR⊗NIR et $\underline{x}_j^{case}$ = $[\underline{x}_{j1}^{case}, \dots \underline{x}_{jO}^{case}]$T ∈ $\mathbb{R}^{\mathrm{O}}$ le jième spectre avec O = P, Q, P+Q ou PxQ. Chaque spectre est enregistré sur le vecteur de nombre d'ondes $\underline{y}^{case}$=$[y_1^{case} \dots y_O^{case}]^T \in \mathbb{R}^{\mathrm{O}}$. Ces spectres appartiennent à un ensemble de classes C = $\{c_1, \dots c_k, \dots c_K\}$ avec K < J, où K désigne le nombre de classes.

Nous proposons d'utiliser la version suivante de l'algorithme AG dont les particularités sont :

1. Initialisation des paramètres :
   - la taille des chromosomes L (le nombre correspondant de gènes soit le nombre d'ondes à sélectionner),
   - la taille de la population N (le nombre de chromosomes par génération),
   - la probabilité de croisement pc,
   - le nombre d'élites N_e (le nombre de chromosomes ayant les meilleures valeurs de fitness dans la génération actuelle qui sont garantis de survivre à la génération suivante),
   - la fraction de croisement Fc (≤1) qui est un facteur limitant le nombre de chromosomes sélectionnés pour effectuer les croisements,
   - les critères d'arrêt de l'algorithme à savoir : le nombre maximal d'itérations T, le nombre de générations minimal avant interruption T' (T'≤T) et la tolérance ϵ.

   Le nombre de parents sélectionnés pour le croisement et la mutation est donné par :

$$N_p = (F_c + 1)N - 2N_e = 2N_c + N_m \qquad \text{(eq.3.1)}$$

   tandis que le nombre d'enfants N_c qui sont créés par croisement par :

$$N_c = F_c * \mathrm{N} - N_e \qquad \text{(eq.3.2)}$$

   Le nombre de parents sélectionnés uniquement pour le croisement est donc de $2\,N_c$. On en déduit par soustraction que le nombre de chromosomes N_m qui vont subir une mutation est égal à :

$$N_m = N_p - 2\,N_c \qquad \text{(eq.3.3)}$$

   La population totale pour chaque génération vérifie N = $N_e + N_c + N_m$.

2. Initialisation de la population :
   Les chromosomes sont générés aléatoirement pour former une population initiale P(0) = $\left\{\underline{z}_i = [z_{i1} \dots z_{il} \dots z_{iL}]^T \in \mathbb{R}^{\mathrm{L}}\right\}_{i=1}^{N}$ telle que chaque gène z_{il} est un nombre d'onde (bande

spectrale) choisi aléatoirement dans le vecteur $\underline{y}^{case}$. Chaque $\underline{z}_i$ est donc un vecteur formé de L nombres d'ondes aléatoirement sélectionnés dans le vecteur $\underline{y}^{case}$.

3. Évaluation : A chaque génération, on évalue la performance de chaque chromosome $\underline{z}_i$ de la population par une fonction fitness F(.) qui assigne à tout chromosome une valeur F_i : $F_i = F(\underline{z}_i)\ \forall\ i = 1 \dots N$. Pour notre problème de minimisation, la valeur opposée ou inverse de la fonction fitness est choisie pour évaluer la qualité des chromosomes. Plus les valeurs obtenues de $F(\underline{z}_i)$ sont petites, plus le chromosome $\underline{z}_i$ aura des chances d'être sélectionné soit comme chromosome garanti de survivre à la génération suivante ou bien comme chromosome parent. Les valeurs Fi obtenues pour l'ensemble des chromosomes $\underline{z}_i$ sont ensuite ordonnées de manière croissante. Une fonction objectif (rang) F_{obj} est ensuite appliquée pour chaque chromosome $\underline{z}_i$. Elle s'appuie sur l'utilisation de l'opérateur position pour chaque chromosome $\underline{z}_i$: le chromosome qui a la plus petite valeur de la fonction fitness se trouve en première position [RZ13]:

$$F_{obj}(i) = \frac{1}{\sqrt{position(\underline{z}_i)}} \qquad \text{(eq.3.4)}$$

La fonction fitness est probablement la partie la plus importante d'un algorithme génétique. Le rôle d'une fonction fitness est de mesurer la qualité du chromosome dans la population conformément à l'objectif d'optimisation. Comme nous essayons de classer les spectres de la biomasse lignocellulosique au cours de K périodes du processus de biodégradation, nous proposons de tester différentes fonctions fitness en s'appuyant sur des indices de validité qui visent à avoir les clusters les plus compacts possible et les plus séparés [YYS14, XXW12]. Ces indices sont des fonctions statistiques-mathématiques bien connues et largement utilisées pour évaluer et mesurer la qualité des classes obtenues, comme : Davies Bouldin (DB), Calinski-Harabasz (CH), Xie Beni (XB), l'indice de séparation (SI), l'indice de Statistique Silhouette (SIL), l'indice de Fisher (FI), etc. Toutes ces fonctions visent à estimer la séparabilité et la compacité des classes [AGM13]:

- L'indice Xie Beni (XB) quantifie le rapport de la variation totale dans les clusters et la séparation des clusters. Il donne de bonnes réponses sur un large choix de classes pour K$\geq$ 2. Il est basé sur les propriétés de compacité et de séparation de l'ensemble des données [MB03]:

$$XB(\underline{z}_i) = \frac{\sum_{k=1}^{K}\sum_{j=1}^{J}(\mu_{kj})^2 \left\|\underline{x}_j^{t=t_k}(\underline{z}_i) - \underline{m}_k(\underline{z}_i)\right\|^2}{J\ min_{i,k} \left\|\underline{x}_j^{t=t_k}(\underline{z}_i) - \underline{m}_k(\underline{z}_i)\right\|^2} \qquad \text{(eq.3.5)}$$

avec $\underline{x}_j^{t=t_k}(\underline{z}_i)$ les amplitudes du spectre $\underline{x}_j^{case}$ enregistré pour la période de biodégradation t_k sélectionnées aux nombres d'ondes z_{il} ($1 \leq l \leq L$), μ_{kj} le coefficient d'appartenance du spectre j à la classe k et $\underline{m}_k(\underline{z}_i)$ le centre de la classe t_k :

$$\underline{m}_k(\underline{z}_i) = \frac{1}{card(t_k)} \sum_{j=1,..J} \underline{x}_j^{t=t_k}(\underline{z}_i) \qquad \text{(eq.3.6)}$$

- L'indice de séparation (SI) mesure le rapport de la somme de la compacité et la séparation des groupes [YYS14]:

$$SI(\underline{z}_i) = \frac{\sum_{k=1}^{K}\sum_{j=1}^{J}(\mu_{kj})^2 \|\underline{x}_j^{t=t_k}(\underline{z}_i) - \underline{m}_k(\underline{z}_i)\|^2}{J\, min_{k,k'} \|\underline{m}_k(\underline{z}_i) - \underline{m}_{k'}(\underline{z}_i)\|^2} \quad \text{(eq.3.7)}$$

avec les mêmes paramètres que ceux décrits pour l'indice Xie-Beni, la différence étant dans la normalisation représentée par la distance minimale de séparation entre les centres de classes utilisées.

- L'indice Calinski et Harabasz (CH) mesure le rapport de la dispersion des spectres dans un cluster sur la base des distances entre les points dans un cluster et son centre. On peut remarquer que l'indice CH a obtenu les meilleures performances de classification parmi 30 indices dans l'étude de comparaison de Milligan et Cooper [MC85]. Il peut être défini comme :

$$CH(\underline{z}_i) = \frac{\frac{1}{K-1}Tr(S_B)}{\frac{1}{J-K}Tr(S_W)} \quad \text{(eq.3.8)}$$

où $Tr(S_B)$ est la trace des matrices inter-clusters :

$$Tr(S_B) = \sum_{k=1}^{K} J_k \|\underline{m}_k(\underline{z}_i) - \underline{m}(\underline{z}_i)\|^2 \quad \text{(eq.3.9)}$$

et $Tr(S_w)$ la trace des matrices intra-clusters :

$$Tr(S_W) = \sum_{k=1}^{K}\sum_{j=1}^{J} \|\underline{x}_j^{t=t_k}(\underline{z}_i) - \underline{m}_k(\underline{z}_i)\|^2 \quad \text{(eq.3.10)}$$

avec J_k le nombre de spectres appartenant à la période de biodégradation t_k et m le spectre moyen pour l'ensemble de données, mais en considérant les nombres d'ondes $\underline{z}_i$.

- L'indice de Fisher (FI) est le rapport entre la diffusion au sein de la classe et la dispersion de la classe [CP04].

$$FI(\underline{z}_i) = \frac{Tr(S_B)}{Tr(S_W)} \quad \text{(eq.3.11)}$$

avec les mêmes paramètres que ceux décrits pour l'indice Calinski-Harabasz. Nous notons que les indices FI et CH sont très proches, mais l'indice FI étant une mesure bien connue, elle a été ajoutée dans notre étude.

- L'indice de silhouette (SIL) mesure la valeur de similarité de chaque spectre par rapport aux autres spectres dans son cluster en comparaison avec les spectres dans les autres clusters. Il est définit comme étant la moyenne de la silhouette s_j [XXW12] :

$$SIL(\underline{z}_i) = \frac{1}{J}\sum_{j=1}^{J} s_j = \frac{1}{J}\sum_{j=1}^{J} \frac{b_j - a_j}{max(a_j, b_j)} \quad \text{(eq.3.12)}$$

où

$$a_j = \frac{1}{card(t_k)-1}\sum_{j'=1}^{card(t_k)} \|\underline{x}_j^{t=t_k}(\underline{z}_i) - \underline{x}_{j'}^{t=t_k}(\underline{z}_i)\|^2 \quad \text{(eq.3.13)}$$

est la distance moyenne du spectre $\underline{x}_j^{t=t_k}$ enregistré à la période de biodégradation t_k par rapport à tous les autres spectres $\underline{x}_{j'}^{t=t_k}$ enregistrés à la même période de biodégradation et

$$b_j = min_{k',k\neq k'} \frac{1}{card(t_{k'})} \sum_{j'=1}^{card(t_{k'})} \|\underline{x}_j^{t=t_k}(\underline{z}_i) - \underline{x}_{j'}^{t=t_{k'}}(\underline{z}_i)\|^2 \qquad \text{(eq.3.14)}$$

est la distance moyenne du spectre $\underline{x}_j^{t=t_k}$ enregistré à la période de biodégradation t_k par rapport à tous les autres spectres $\underline{x}_j^{t=t_{k'}}$ enregistrés à une autre période de biodégradation.

- L'indice Davies Bouldin (DB) est le rapport de la somme des dispersions au sein de la classe à la séparation entre les classes. La dispersion au sein de classe c_k est calculée comme [Sus04, Ban01, Sar97, Alf08]:

$$S_k(\underline{z}_i) = \sqrt{\frac{1}{card(c_k)} \sum_{\underline{x}_j(\underline{z}_i)\epsilon c_k} d(\underline{x}_j(\underline{z}_i) - \underline{m}_k(\underline{z}_i))} \qquad \text{(eq.3.15)}$$

où $card(C_k)$ est le nombre de spectres qui appartiennent à la kième classe c_k et $d(\underline{x}_j(\underline{z}_i) - \underline{m}_k(\underline{z}_i))$ est la distance euclidienne entre le spectre $\underline{x}_j(\underline{z}_i)$ qui appartient à la classe c_k et son centre $\underline{m}_k(\underline{z}_i)$ (qui est la moyenne des spectres de la classe k). S_k représente la moyenne des distances entre les spectres appartenant à la classe c_k et son centre $\underline{m}_k(\underline{z}_i)$. On peut alors calculer la dispersion au sein de la classe K par rapport à la séparation avec les autres classes comme :

$$R_k(\underline{z}_i) = max_{k',k\neq k'}\{\frac{S_k(\underline{z}_i)+S_{k'}(\underline{z}_i)}{d_{kk'}}\} \; \forall \; k' = 1 \dots K \qquad \text{(eq.3.16)}$$

où $d_{kk'}$ représente la distance entre les centres des deux classes c_k et $c_{k'}$. L'indice Davies-Bouldin (DB) est alors calculé comme suit :

$$\text{DB}(\underline{z}_i) = \frac{1}{K}\sum_{k=1}^{K} R_k(\underline{z}_i) \; \forall \; k = 1 \dots K \qquad \text{(eq.3.17)}$$

La partie discussion sur le choix de fonction fitness pour les spectres IR de biomasse lignocellulosique est réalisée dans la section suivante « Paramètres de l'algorithme génétique ». Nous avons trouvé que l'indice Davies Bouldin (DB) est le plus adapté pour l'AG sur les spectres IR des échantillons de biomasse lignocellulosique lors du processus de biodégradation. Nous notons qu'on peut utiliser notre métrique que nous avons présentée dans le chapitre 2, mais nous préférons utiliser des métriques usuelles dans notre choix de fonction fitness pour ne pas avoir plusieurs nouveaux facteurs dans cette étude.

4. Sélection : Nous avons choisi la sélection de type « stochastic universal sampling » car cette méthode a comme avantage de ne pas avoir de biais d'estimation, et une dispersion minimale. Ce type de sélection réalise le calcul sur N chromosomes en une seule passe [RZ13].

Premièrement, on calcule la probabilité p_i de la sélection du chromosome $\underline{z}_i$ et la probabilité cumulée q_i:

$$p_i = \frac{F_{obj}(i)}{\sum_{i=1}^{N} F_{obj}(i)} \qquad q_i = \sum_{m=1}^{i} P_m \qquad \text{(eq.3.18)}$$

Ensuite, on génère un nombre aléatoire uniforme r ϵ [0,$\frac{1}{N_p}$] avec N_p le nombre de chromosomes qui vont être sélectionnés : N_p (cf. eq. 3.1) :

- Si r $< q_1$ alors on sélectionne le chromosome $\underline{z}_1$, ensuite on sélectionne les chromosomes $\underline{z}_i$ tels que $q_{i-1} <$ r $+\frac{1}{N_p} < q_i$. Ce qui revient à sélectionner N_p-1 chromosomes en appliquant la formule suivante :

  $\forall\, h = 1 \dots N_p - 1$, si $q_{m-1} <$ r $+ h\frac{1}{N_p} < q_m$, alors on sélectionne le chromosome $\underline{z}_m$.
- Si r $> q_1$ alors on sélectionne les chromosomes $\underline{z}_i$ tels que $q_{i-1} <$ r $< q_i$. Ce qui revient à sélectionner N_p chromosomes en appliquant la formule suivante :

  $\forall\, h = 0 \dots N_p - 1$, et $q_{m-1} <$ r $+h\frac{1}{N_p} < q_m$ alors on sélectionne le chromosome $\underline{z}_m$.

5. Croisement (Crossover) : Nous avons choisi la méthode de croisement uniforme qui a donné de bons résultats dans la majorité des cas [Pic10].
   Le nombre de parents qui vont être sélectionnés pour le croisement est égal à $2N_c = 2F_c$*N – $2N_e$. Ces $2N_c$ chromosomes sont pris deux à deux et recombinés pour obtenir N_c enfants. Tous les gènes de l'enfant sont sélectionnés de manière aléatoire soit à partir du premier ou du deuxième parent sur la base d'une probabilité déterminée p_c qui est la probabilité de sélection qui s'applique au niveau de chaque gène des deux parents. Par exemple, si pc = 0.8, alors le gène 1 du parent 1 à 80% de chance d'être sélectionné pour créer l'enfant et le gène 1 du parent 2 à 20%, idem pour les autres gènes.

   Mutation : Nous avons choisi l'opérateur de mutation Gaussien car il produit les meilleurs résultats pour la plupart des fonctions fitness [Hin95]. En pratique, la mutation gaussienne produit des petits changements aléatoires dans les chromosomes. Pour cela, on ajoute un nombre aléatoire généré par une distribution gaussienne de moyenne nulle pour toutes les positions des gènes dans le chromosome choisi (et ce pour chaque chromosome de l'ensemble N_m destiné à subir une mutation). Les valeurs des positions des gènes sont ensuite arrondies à l'entier le plus proche. Par exemple: soit un gène de taille L= 4. Ce gène peut être représenté par les 4 positions entières des bandes spectrales sélectionnées dans le spectre (la position n'est pas la valeur du nombre d'onde mais l'index désignant la position du nombre d'onde). Par exemple pour un chromosome formé de 4 gènes :

4	268	459	532

   Après mutation nous obtenons :

4.23456	269.11123	458.29453	531.87654

   soit

4	269	458	532

La variance de cette distribution est le paramètre que l'on appelle « scale » qui vaut 1 à la première génération, mais ce paramètre peut être contrôlé au cours des générations suivantes par un autre paramètre qui est « shrink ». La formule permettant le contrôle est la suivante :

$$\text{scale (t+1)} = \text{scale(t)} - \text{shrink*scale(t)*}(\frac{1}{T}).$$

où i désigne la génération précédente et $1 \leq t \leq T-1$
Une valeur faible de « shrink » va produire une décroissance faible et très progressive de l'amplitude de la mutation sur les indices des positions des gènes, alors qu'une valeur forte va induire une disparition rapide de l'effet de cette même étape.

La population suivante sera donc formée de N_e chromosomes ayant les meilleures valeurs de fitness extraits de la génération précédente, de N_c enfants qui sont créés par croisement et de N_m chromosomes qui ont subi une mutation.

6. On répète les étapes 3 à 6 jusqu'à ce que le nombre d'itérations maximal T soit atteint, ou l'on s'arrête quand la variation moyenne pondérée des T dernières valeurs de fitness est inférieure à une tolérance ϵ (fixée) [Zel12].

En résumé, nous pouvons dire que les algorithmes génétiques sont des méthodes aléatoires basées sur les probabilités et un principe d'optimisation. La population évolue par sélection (seulement les individus les mieux adaptés survivent), par croisement (génèrent des enfants comportant des séquences de chromosomes de leur parents) et par mutation d'une partie de leur génome. Ce type d'algorithme est très adapté à la minimisation d'une fonction présentant plusieurs minima (locaux ou globaux). Le coût de calcul est cependant très important car tous les individus de chaque génération doivent être évalués, faisant autant de fois appel au calcul de la fonction à optimiser.
Dans la suite de ce chapitre, nous aborder le choix des paramètres les mieux adaptés pour des données infrarouges.

III.3.3. Paramètres de l'algorithme génétique

Nous avons des paramètres initiaux pour l'AG comme la fraction de croisement F_c, le nombre d'élites N_e, la taille de la population N, le nombre maximum de générations T et la taille des chromosomes L. Dans notre cas, nous avons choisi :

- La fraction de croisement $F_c = 0.8$, car cette valeur donne dans la plupart des cas les meilleurs résultats [Aya13] et elle permet d'obtenir de meilleures valeurs de fitness que les autres fractions de croisement.
- Le nombre d'Elite $N_e = 2$ car cette valeur est la plus utilisée dans la plupart des applications qui utilisent un AG [Ver11].
- La tolérance $\varepsilon = 10^{-6}$.

Nous avons trouvé que ces paramètres sont souvent utilisés et bien adaptés aux spectres IR des échantillons végétaux. Ces paramètres ont montré leur efficacité pour différencier des espèces de végétaux [UGS12] et pour optimiser la production de biodiesel [RJR09].

Pour choisir L et N, il n'existe pas de méthode clairement définie. Nous sommes donc obligés de déterminer de façon itérative les valeurs optimales de ces paramètres. Pour cela nous itérons pour différentes valeurs de tailles des chromosomes (L=3 jusqu'à 10) et nous faisons de même pour le

nombre de population N, (N=50 jusqu'à 500). Nous choisissons ensuite les valeurs de L et N optimales à partir des meilleures valeurs de fitness obtenues [UGS12, Jef04] :

$$L = min_L\{min_N\{F(\underline{z}_i)\}\}\,;\, i = 1 \dots N \qquad \text{(eq.3.19)}$$

et

$$N = min_N\{min_L\{F(\underline{z}_i)\}\};\, i = 1 \dots N \qquad \text{(eq.3.20)}$$

L'algorithme génétique AG permet de sélectionner des nombres d'ondes les plus discriminants qui ont les mêmes poids, c'est-à-dire que l'AG sélectionne un chromosome de taille L ($L \ll O$) qui a donné la valeur optimale de fonction fitness, et les gènes (nombres d'ondes) de ce chromosome ont les mêmes poids. L'AG nécessite beaucoup de paramètres tels que L, N, Ne et F_c etc. Pour cela nous proposons une nouvelle approche qui permet de sélectionner un vecteur de poids w qui met en évidence les nombres d'ondes les plus discriminants pour les spectres IR. Cette approche ne nécessite pas d'initialisation de paramètres. Dans la suite de ce chapitre, nous nous intéressons à présenter cette nouvelle approche de sélection des bandes qui est basée sur le principe d'optimisation avec contrainte non linéaire.

III.4. Nouvelle approche par optimisation PQS

III.4.1. Généralités

Après avoir vu la possibilité de sélectionner des bandes spectrales par AG, nous proposons ici une autre approche par optimisation non linéaire avec contraintes pour sélectionner les bandes spectrales les plus discriminantes des biomasses lignocellulosiques en fonction de la biodégradation. Il y a plusieurs types d'algorithmes qui permettent de résoudre des problèmes d'optimisation mathématiques : Les méthodes de points intérieurs, la programmation quadratique séquentielle (PQS ; sequential quadratic programming en anglais), la méthode des pénalités et les algorithmes à régions de confiance, etc.

Nous avons utilisé la programmation quadratique séquentielle qui est une méthode des plus efficaces et automatique pour obtenir une solution numérique à un problème d'optimisation non linéaire avec contraintes. Ainsi, les algorithmes de programmation quadratique séquentielle présentent l'avantage d'avoir une propriété de convergence quadratique, contrairement aux autres algorithmes d'optimisation qui ont une convergence linéaire. La convergence quadratique permet d'apprécier l'efficacité des algorithmes plus que la convergence simple.

Le principe général des méthodes d'optimisation est de résoudre numériquement le problème suivant :

$$\begin{cases} trouver\ \underline{z}_i^* \in R^L\ et\ f\colon \mathbb{R}^L \to \mathbb{R} \\ \underline{z}_i^* = \underset{z_i \in \mathbb{R}^L}{argmin}\, f(X^{case}(\underline{z}_i)) \end{cases} \qquad \text{(eq.3.21)}$$

avec $\underline{z}_i = [z_{i1} \dots z_{il} \dots z_{iL}]^T \in \mathbb{R}^L$ l'ensemble de nombres d'ondes sélectionnées et $\underline{z}_i^* \in \mathbb{R}^l$ la solution optimale du problème. Comme nous l'avons précédemment mentionné, l'algorithme génétique permet la sélection des nombres d'ondes les plus discriminants en utilisant une fonction fitness de type indice de validité qui permet de mesurer la séparabilité des classes. Nous proposons d'utiliser le même type fonction pour ces méthodes afin de pouvoir comparer ces deux algorithmes plus facilement.

III.4.2. Approche proposée pour des données spectrales IR

La méthode d'optimisation proposée consiste à minimiser une fonction objectif permettant de sélectionner un vecteur de poids w qui met en évidence les nombres d'ondes les plus discriminants et qui permet la meilleure séparabilité entre les classes. La fonction F en fonction du vecteur des poids $\underline{w} = [w_1, \dots, w_i, \dots, w_O]^T \in \mathbb{R}^O$ qui multiplie la matrice de spectre X^{case} forme une nouvelle fonction objectif à minimiser. Le problème de cette approche d'optimisation revient ici à résoudre numériquement le problème suivant pour trouver la pondération optimale w_{opt} :

$$\begin{cases} trouver\ \underline{w} \in R^O,\ F : \mathbb{R}^O \to \mathbb{R} \\ 0 \leq w_i\ \forall\, i = 1 \dots O\ (avec\ d'autres\ contraintes) \\ \underline{w}_{opt} = \underset{\underline{w} \in \mathbb{R}^O}{argmin}\, F\left(\underline{w} * X^{case}(\underline{y}^{case})\right) = \underset{\underline{w} \in \mathbb{R}^O}{argmin}\, F\left(X_w^{case}(\underline{y}^{case})\right) \end{cases} \quad \text{(eq.3.22)}$$

avec $X^{case}(\underline{y}^{case})$ la matrice de spectres en fonction du vecteur de nombres d'ondes $\underline{y}^{case}$, $X_w^{case}(\underline{y}^{case})$ est la matrice X^{case} multipliée par le vecteur de poids w et F la fonction objectif utilisée pour l'optimisation. Comme précédemment, compte tenu de notre application, nous pouvons choisir comme fonction objectif un des indices de validité. Par souci de similarité avec l'AG, nous avons choisi par la suite l'indice Davies Bouldin (voir eq. 3.17) comme fonction objectif.

La contrainte imposée, c.-à-d. des valeurs positives du vecteur des poids, est tout à fait logique dans notre application puisque nous cherchons à estimer des bandes discriminantes. D'autres contraintes peuvent être imposées, comme la norme L_1 :

$$\sum_i |w_i| = 1 \quad \text{(eq.3.23)}$$

qui est une approche classique des représentations parcimonieuses (en anglais 'sparse'). Cette contrainte est pratique quand on veut limiter le nombre de bandes spectrales qu'on veut identifier. En effet, contrairement à l'AG pour lequel on imposait le nombre de bandes spectrales à identifier, ici il n'y a aucun critère a priori. La contrainte L_1 permet de résoudre les problèmes de fonctions objectifs non convexes et non dérivables. De même, nous avons la possibilité de choisir d'autres contraintes telles que la norme L_2 :

$$\sum_i {w_i}^2 = 1 \quad \text{(eq.3.24)}$$

qui permet de résoudre les problèmes d'optimisation convexes (bien adapté pour une fonction objectif convexe) et d'obtenir une solution stable mais moins robuste que la contrainte L_1 ou la norme L_∞ :

$$\sum_i |w_i| < \infty \quad \text{(eq.3.25)}$$

Une contrainte max-norme L_∞ permet de résoudre de nombreux problèmes tels que l'optimisation de matrice de rang faible (vecteurs qui sont linéairement dépendants) et la réduction de la dimensionnalité de données. La norme L_∞ permet d'améliorer la rapidité de la convergence de l'algorithme.

III.4.3. Paramètres de l'approche d'optimisation

Nous avons des paramètres initiaux pour l'approche d'optimisation comme le nombre maximum d'itérations T et la tolérance de terminaison. Nous avons choisi les valeurs T=1000 et ε = 10-6 parce que celles-ci sont fréquemment utilisées dans la plupart des applications d'optimisation.

Les deux méthodes (AG et optimisation par PQS) sont complémentaires, la première assure une recherche globale et robuste même sur les spectres de grandes dimensions (comme les spectres combinés par l'OP) et l'autre (PQS) calcule l'optimum global sur des spectres de taille réduite de façon plus rapide avec moins d'initialisation de paramètres. Dans la suite de ce chapitre, nous expliquons comment ces méthodes peuvent être appliquées sur des spectres MIR et NIR et leurs combinaisons pour sélectionner les bandes spectrales optimales.

III.5. Application des algorithmes AG et PQS aux données spectroscopiques IR

III.5.1. Principes

Les algorithmes proposés (AG et PQS) peuvent être appliqués sur des matrices de données $\{\mathrm{x}_j{}^{case}(\mathrm{y}^{case})\}_{j=1}^{J}$ avec case = {MIR, NIR, MIR-NIR, or MIR⊗NIR}. Néanmoins, du point de vue pratique, le deuxième algorithme ne peut pas être appliqué si les spectres sont de grandes tailles. En effet, pour notre application, le produit extérieur conduit à des spectres de dimensions PQ = 500 000 et l'estimation d'un vecteur de poids de cette dimension ne peut pas être réalisé avec des moyens de calcul classiques en utilisant l'algorithme PQS. Comme perspective, nous comptons par la suite étendre cette approche pour des spectres de grande dimension par exemple à l'aide des projections aléatoires.

Lorsqu'on utilise l'AG, nous obtenons un chromosome optimal $\mathrm{z}_{opt}^{case} = [z_{opt,1} \dots z_{opt,l} \dots z_{opt,L}]^T \in \mathbb{R}^{\mathrm{L}}$, avec z_{ol} un nombre d'onde pour les régions MIR, NIR et la combinaison MIR-NIR ou une paire de nombres d'ondes (MIR, NIR) pour la combinaison par le produit extérieur (OP). Le chromosome optimal permet d'extraire une nouvelle matrice de spectres sous dimensionnée $\{\mathrm{x}_j{}^{case}(\mathrm{z}_o^{case})\}_{j=1}^{J}$ sur laquelle nous pouvons appliquer des méthodes d'analyse de données.

Lorsqu'on utilise l'optimisation, nous obtenons un vecteur des poids w_{opt} qui, en multipliant la matrice des spectres, donne des spectres pondérés sur lesquels nous pouvons appliquer des méthodes d'analyse de données.

Après l'application de l'algorithme génétique ou de l'approche d'optimisation, nous aurons besoin de quantifier et qualifier les résultats obtenus. Pour cela, nous utilisons la méthode de décomposition l'analyse en composantes principales pour afficher une représentation des classes sous la forme de scores plots, et l'indice Dunn DI pour quantifier numériquement la valeur du résultat.

III.5.2. Qualification à l'aide de analyse en composantes principales (ACP)

L'analyse en composantes principales (ACP) est un grand classique de « l'analyse des données » pour la réduction des dimensions des données analysées et a servi de fondement théorique à d'autres méthodes de statistique multidimensionnelle dites factorielles. Elle est largement utilisée dans de nombreuses applications, principalement pour l'analyse de la relation entre les composés et pour illustrer les différences spectrales entre des échantillons. L'ACP a pour but d'obtenir la meilleure représentation possible (au sens de la variance) de n individus dans un sous-espace de dimension réduite, par exemple un plan [MR93]. Autrement dit, on cherche à définir des nouvelles combinaisons linéaires des variables initiales qui feront perdre le moins d'information possible. Ces variables seront

appelées « composantes principales » et les axes qu'elles déterminent : « axes principaux ». La projection des données sur le plan (ou l'hyperplan) formé des axes principaux permet d'obtenir ce que l'on appelle les « Scores ». La projection des variables sur les axes principaux s'appelle les « Loadings ».

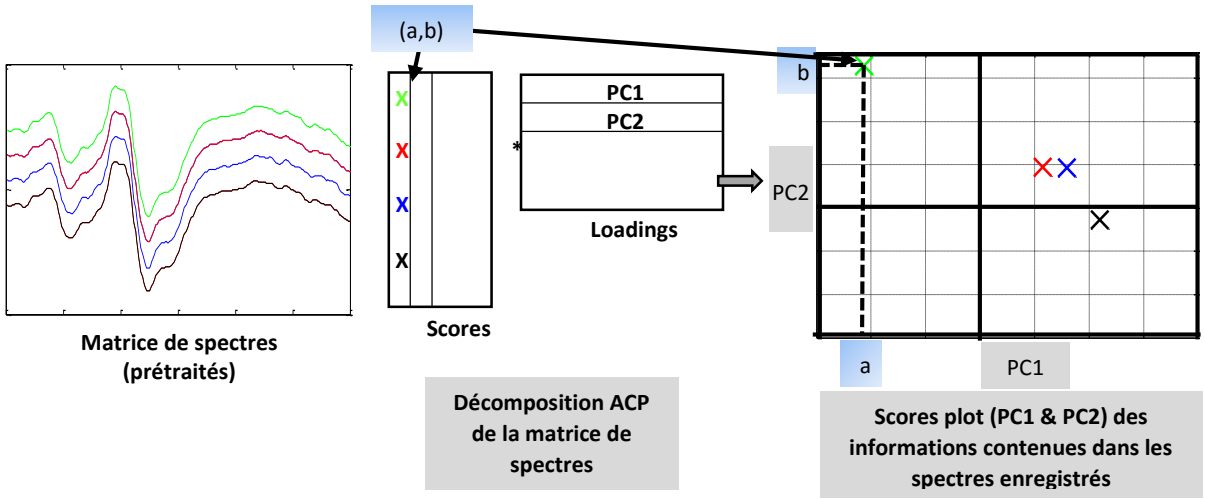


Figure 3.2. Décomposition ACP de la matrice de spectres X^{case} et représentation des scores plots suivant les 2 premières composantes (PC1 & PC2).

Mathématiquement, l'analyse en composantes principales représente un simple changement de base qui permet de passer d'une représentation dans la base canonique des variables initiales à une représentation dans la base des facteurs définis par les vecteurs propres de la matrice des corrélations [MR93]. L'ACP peut être relié à la décomposition en valeurs singulières SVD, quand les composants principaux sont calculés à partir de la matrice de covariance des données X^{case} :

$$X^{case} = UDV^T \qquad \text{(eq.3.26)}$$

Les colonnes de la matrice V représentent les composantes principales (PC ; Loadings) et T désigne la transposée de la matrice. La matrice $U * D$ contient les scores des composantes principales (scores) qui sont les coordonnées dans l'espace des PC. Les scores indiquent donc la localisation de l'échantillon le long des PC [DYX11].
L'affichage des scores (score plot) peut ainsi être utilisé comme un indicateur de la discrimination des échantillons analysés par rapport au processus de biodégradation. Ainsi, les scores d'ACP permettent d'analyser la séparabilité des échantillons selon la cinétique de biodégradation. La Figure 3.2 montre la décomposition ACP de la matrice des spectres X^{case} et la représentation des scores [MKB79].

III.5.3. Quantification à l'aide de l'indice de Dunn (DI)

Pour quantifier la séparabilité des échantillons par rapport à la cinétique de biodégradation et donc évaluer numériquement les méthodes, l'indice de Dunn a été utilisé pour estimer le rapport entre la distance minimale intra-cluster (entre deux classes différentes) à la distance maximale inter-cluster (la distance maximale entre deux spectres d'une même classe).
Soit C = $\{c_1, \dots c_k, \dots c_K\}$ un ensemble de clusters (classes), δ : C x C $\rightarrow$ $\mathbb{R}^+$ une mesure de distance de cluster à cluster, et Δ : C $\rightarrow \mathbb{R}^+$ une mesure de distance maximale entre deux individus d'une même classe. Ces distances sont souvent choisies comme étant euclidiennes [RRG12]. Nous notons qu'on peut également utiliser notre métrique que nous avons présentée dans le chapitre 2 au lieu d'une

métrique euclidienne. Toutefois, nous allons utiliser par la suite la métrique euclidienne, qui est couramment utilisée pour réaliser cette mesure, afin de ne pas introduire un paramètre d'analyse supplémentaire.
On peut calculer la valeur scalaire DI par [AGM13] :

$$DI = \frac{min_{i\neq j}\{\delta(c_i,c_j)\}}{max_{1\leq l\leq K}\{\Delta(c_l)\}} \qquad \text{(eq.3.27)}$$

avec

$$\delta(c_i,c_j) = min_{\underline{x}_m^{case}\epsilon c_i,\, \underline{x}_n^{case}\epsilon c_j}\{d(\underline{x}_m^{case},\underline{x}_n^{case})\},$$
$$\Delta(c_l) = max_{\underline{x}_m^{case},\, \underline{x}_n^{case}\epsilon c_l}\{d(\underline{x}_m^{case},\underline{x}_n^{case})\}.$$

Le DI est un nombre positif et une valeur élevée indique une meilleure séparation des clusters.

III.6. Méthodologies proposées

Dans cette section, nous présentons les méthodologies permettant de sélectionner les nombres d'ondes par l'algorithme génétique ou par l'approche d'optimisation pour les spectres MIR, NIR ou MIR-NIR. Nous présentons également l'OP-AG qui permet de combiner les spectres MIR et NIR par le produit extérieur MIR⊗NIR et d'appliquer ensuite l'AG pour sélectionner les bandes spectrales optimales.

La Figure 3.3 montre l'application de la méthode sur des spectres MIR, NIR ou MIR-NIR suivie d'une analyse ACP sur la matrice des nombres d'ondes sélectionnés par une approche d'optimisation (couleurs d'intensité différente en fonction des pondérations) ou respectivement par l'AG (dans ce cas les intensités sont toutes les mêmes). L'information utile issue de l'ACP est ensuite représentée sous la forme de scores plots (PC1 & PC2). Nous affichons la représentation de scores plots (PC1 & PC2) sur la matrice des informations spectrales aux nombres d'ondes sélectionnés (MIR et/ou NIR). La représentation des scores plots en termes de PC1 et PC2 donne une indication qualitative du processus de biodégradation. Pour quantifier la séparabilité par rapport à la cinétique de biodégradation, nous calculons l'indice DI sur la matrice des informations des paires de nombres d'ondes sélectionnées par l'AG ou pondérées par l'approche d'optimisation.

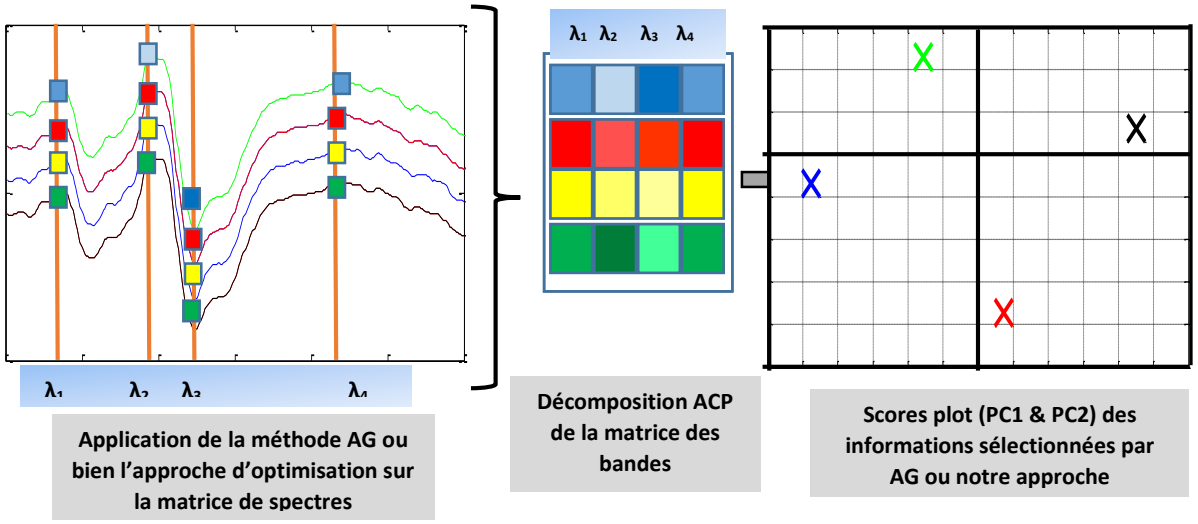


Figure 3.3. Sélection de bandes spectrales par AG ou PQS, décomposition ACP des informations sélectionnées et représentation de scores plot en termes de PC1 & PC2.

Nous avons proposé une méthodologie permettant de combiner les spectres MIR et NIR par le produit extérieur (OP) et l'application de l'AG pour la sélection des bandes qui a été publié récemment [RPC15]. Cette méthodologie est affichée sur la Figure 3.4. Comme nous l'avons dit, la méthode par optimisation ne peut pas être appliquée sur des données de grande taille, mais le principe de fonctionnement exposé ici est tout à fait transposable à cette méthode.

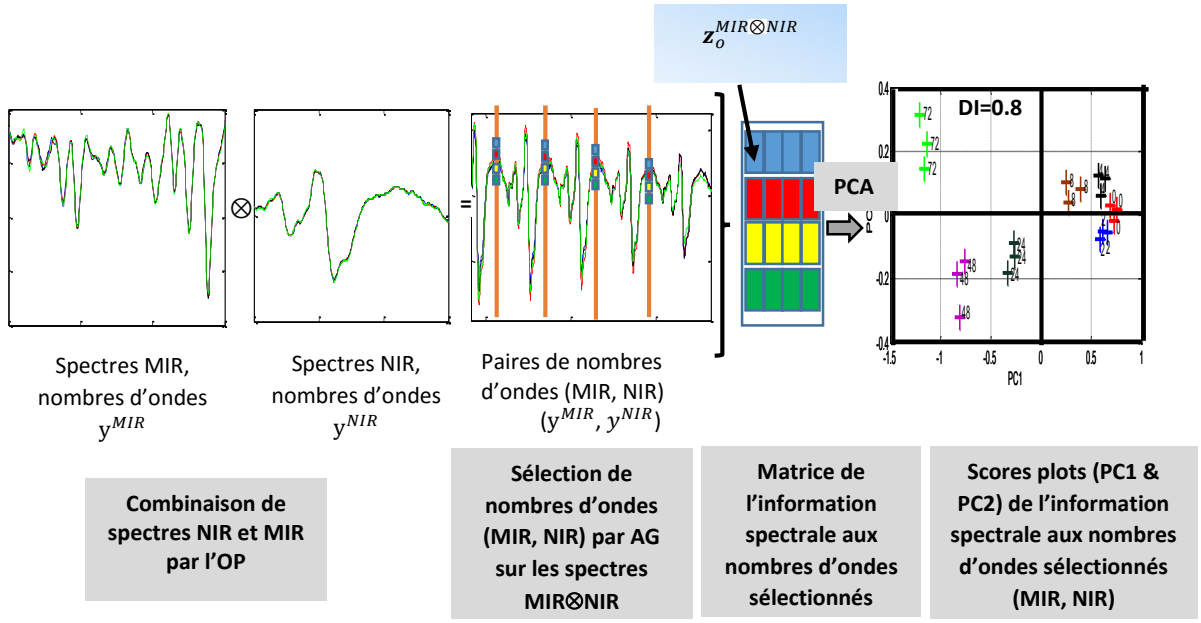


Figure 3.4. Synoptique de la méthodologie d'analyse par OP-AG suivie d'une ACP.

La première étape consiste à appliquer la méthode de combinaison par le produit extérieur OP entre les spectres MIR et NIR. Nous obtenons une matrice de « super spectres » de très grande dimension, PQ. Ensuite, l'algorithme génétique (AG) est appliqué sur cette matrice de données pour réduire sa dimension en identifiant des paires de nombres d'ondes discriminants (MIR, NIR). On obtient une sous matrice de données que nous analysons par ACP. Pour quantifier la séparabilité par rapport à la cinétique de biodégradation, nous calculons l'indice DI sur la matrice des informations des paires de nombres d'ondes sélectionnées par l'AG.

Pour valider le fonctionnement des algorithmes, nous allons les tester sur des données simulées. Dans la suite, nous exposons la construction de ces données et les résultats de sélection des bandes spectrales discriminantes obtenus.

III.6.1. Validation sur spectres simulés

L'objectif principal est la sélection automatique des nombres d'ondes les plus discriminants pour les spectres infrarouge. Les méthodes sont évaluées sur des données simulées avant d'appliquer ces méthodes sur des données IR réelles. L'objectif est de déterminer les performances ainsi que la précision dans la sélection des nombres d'ondes. Pour cela, nous prenons un exemple de 20 spectres simulés $\underline{x_j}$ ($\forall\, j = 1, \dots, 20$) qui ont les propriétés suivantes :

$$\underline{x_j}(z_l) \neq \underline{x_{j'}}(z_l)\ \forall\ j \neq j'\ et\ z_l = \{1030, 1295, 1620\ cm^{-1}\},$$

$$\underline{x_j}(z_l) = \underline{x_{j'}}(z_{l'})\ \forall\ z_l = z_{l'}\ et\ z_l \neq \{1030, 1295, 1620\ cm^{-1}\},$$

La Figure 3.5 montre les variations des trois nombres d'ondes dans ces spectres simulés. Les variations aux nombres d'ondes indiqués ont été réalisées en considérant cinq classes et en tirant aléatoirement avec des probabilités gaussiennes de moyennes différentes et variances unitaires des valeurs d'amplitudes qui ont été rajoutés à un spectre IR réel.

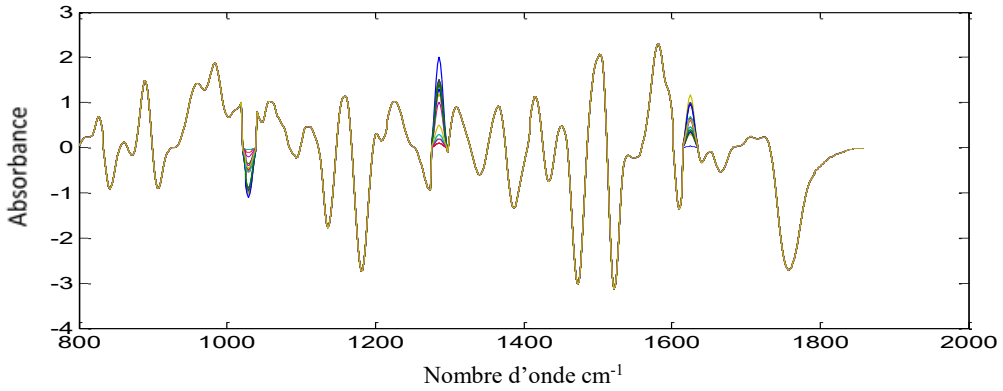


Figure 3.5. 20 spectres simulés qui diffèrent en 3 nombres d'ondes.

III.6.2. Résultats obtenus avec AG

Il s'agit de tester la robustesse de l'algorithme génétique par rapport aux paramètres choisis compte tenu des initialisations aléatoires. Pour les choix des paramètres de l'algorithme génétique appliqué sur les spectres simulés, nous avons choisi d'imposer le nombre de chromosomes N=1000 mais de faire varier les nombres de gènes (nombres d'ondes) L = 3, 5, 7 et 9. Nous avons fixé le nombre de générations T=100 et nous avons répété l'algorithme 10 fois avec des initialisations aléatoires. Nous avons choisi l'indice Davies Bouldin (DB) comme fonction fitness dans l'AG (voir section 3.2 de ce chapitre) et considéré donc K=5 classes puisque nous savons que les données ont été générés comme appartenant à 5 classes. Les fonctions de croisement et mutations choisies sont celles décrites plus haut dans le chapitre. La Figure 3.6 présente l'histogramme des fréquences de sélection des nombres d'ondes comme résultat de la sélection des nombres d'ondes par l'AG pour les nombres de gènes L=3. Les figures qui montrent les résultats d'AG pour L= 5, 7 et 9 sont fournies en annexe 3.

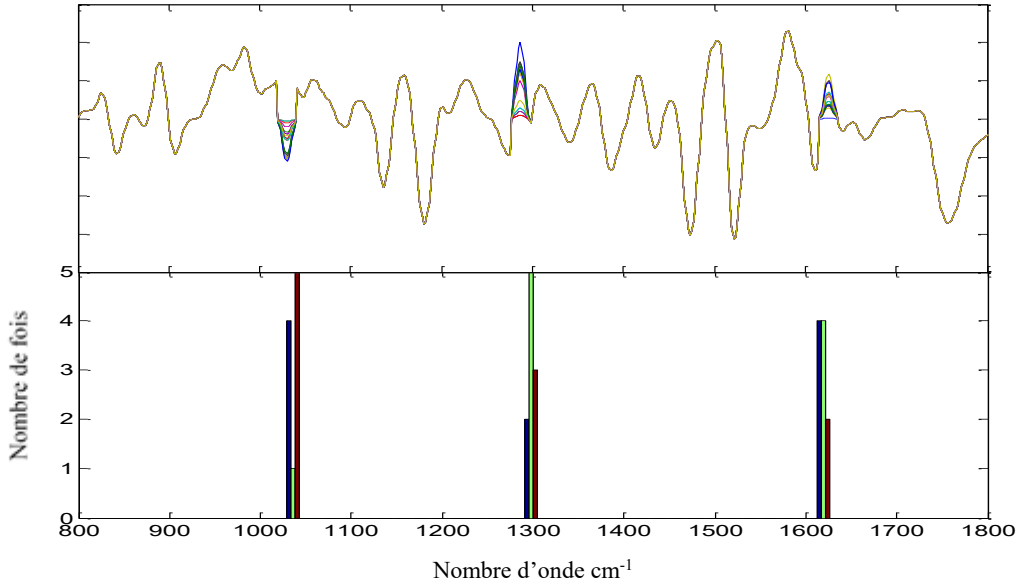


Figure 3.6. Histogramme des nombres d'ondes sélectionnés par l'AG appliqué sur des spectres simulés pour L=3 obtenu pour 10 répétitions.

D'après la Figure 3.6 et les figures fournies en annexe, nous pouvons constater que :

- L'algorithme génétique (AG) avec la fonction fitness Davies-Bouldin (DB) permet l'identification des nombres d'ondes correspondant aux changements des amplitudes.
- Les choix de nombres de gènes L affecte peu la performance de sélection des nombres d'ondes pour l'algorithme génétique puisque un gène va être sélectionné plusieurs fois dans le chromosome optimal. Mais nous pouvons résoudre ce problème grâce au critère de choix optimal du nombre N et de la taille des chromosomes L (voir équations 3.19 et 3.20).

III.6.3. Résultats obtenus avec l'approche par optimisation PQS

Les performances ainsi que la précision de la méthode d'optimisation PQS dans la sélection des nombres d'ondes ont été déterminées en utilisant les mêmes spectres simulés que précédemment dans la partie validation de l'algorithme génétique AG. Pour cela, nous avons testé l'approche d'optimisation avec les contraintes sur la norme L_1 qui a fourni les meilleurs résultats de sélection des nombres d'ondes pour les spectres IR (voir la section 4.3 de ce chapitre). En effet, dans ce cas simple il s'agit de 3 zones bien localisées, soit environ 3x10 nombres d'ondes qui varient par rapport aux 1000 nombres d'ondes disponibles. Il est donc tout à fait normal qu'une norme de type L_1 donne les meilleurs résultats.

La Figure 3.7 montre le vecteur de poids w_i identifié par l'approche d'optimisation proposée. Ce vecteur représente la pondération des nombres d'ondes sélectionnés.

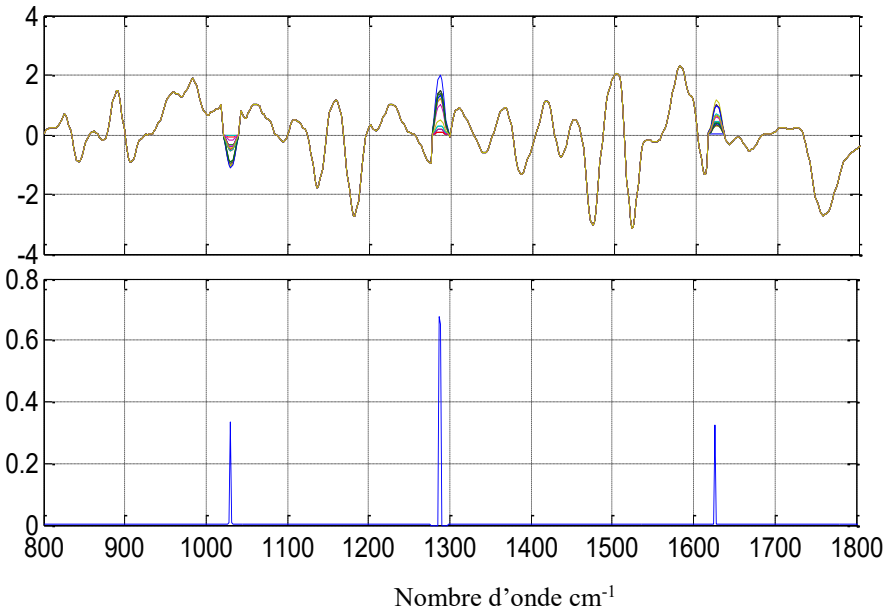


Figure 3.7. Les nombres d'ondes sélectionnés par l'approche d'optimisation proposée avec les contraintes L_1 appliquées sur les spectres simulés.

D'après le Figure 3.7, nous pouvons constater que :

- L'approche d'optimisation permet l'identification de nombres d'ondes qui correspondent aux changements d'amplitudes. Les poids ont des valeurs proportionnelles par rapport aux variations d'amplitudes de ces spectres.

- L'approche d'optimisation résout avec succès les problèmes d'initialisation des paramètres dans l'AG, notamment le choix de L dans l'AG.

Après la validation des algorithmes sur des données simulées, nous allons étudier le choix des paramètres dans l'AG sur les données IR des lignocelluloses, telle que le choix de fonction fitness, les étapes de croisement et mutation et les tailles de chromosomes et de population. Pour cela, nous étudions ces choix de paramètres sur des données spectrales (MIR et NIR) pour des échantillons de biomasse lignocellulosique.

III.7. Application à la biodégradation de biomasse lignocellulosique

III.7.1. Jeux de données

Nous rappelons que les types d'échantillons de biomasse lignocellulosique utilisés dans l'analyse de l'algorithme génétique AG et l'approche d'optimisation sont : des racines de maïs issues de deux lignées parentales distinctes (F2 et F292) et deux mutants de ces lignées (F2bm1 et F292bm3), analysées sur 5 périodes de biodégradation dans le sol : t_1=0, t_2=14, t_3=36, t_4=57 et t_5=112 jours.
Par ailleurs, nous disposons de 3 échantillons de miscanthus et 3 échantillons de peuplier, de génotypes différents, qui sont analysés à différents stades de dégradation par un cocktail enzymatiques: t_1=0, t_2=2, t_3=4, t_4=8, t_5=24, t_6=48, t_7=72 heures. Plus de détails sont fournis dans la section « Biomasse lignocellulosique » du chapitre 1.

Les spectres ont été acquis sur tous les échantillons en utilisant les spectroscopies moyennes infrarouges (MIR) et proches infrarouges (NIR). Pour les spectres MIR, nous avons choisi la gamme spectrale 800 – 1800 cm^{-1}, ce qui correspond aux vibrations principales des groupes fonctionnels chimiques de composés qui sont utiles dans les analyses de la biomasse lignocellulosique. Pour les spectres NIR, nous avons choisi la gamme spectrale 4000 - 6000 cm^{-1} qui représente la gamme spectrale non bruitée. Ces gammes spectrales de MIR et NIR ont été choisies en fonction des résultats obtenus de classification pour les échantillons de maïs par rapport à la cinétique de biodégradation, cf. section 5.3 du chapitre 2.
Les spectres ont été prétraités par filtrage Savitzky-Golay (SG) de 1er ordre avec un lissage sur 17 points et un polynôme d'ordre 4, suivi d'une normalisation de type Standard Normal Variate (SNV). Cette technique de prétraitement a été trouvée comme étant la plus efficace pour discriminer les échantillons de racines de maïs pour les spectres NIR et MIR cf. section 5.3 dans le chapitre 2. Nous combinons par concaténation ou par le produit extérieur la gamme 800-1800 cm^{-1} de MIR et la gamme 400-6000 cm^{-1} de NIR. Chaque spectre est prétraité séparément.

Dans la suite de ce chapitre, nous n'allons présenter uniquement les résultats de l'algorithme génétique et de l'approche d'optimisation sur les échantillons des racines de maïs. Les autres résultats pour les échantillons de miscanthus et peuplier sont présentés dans l'annexe 3.

III.7.2. Paramètres de l'AG

Pour les paramètres de l'AG, nous évaluons l'algorithme génétique pour différentes tailles de chromosomes, L = 3, 4, 5 ... 8, et différentes tailles de la population, N = 50, 100, ..., 500 pour les spectres NIR, MIR et les spectres concaténés MIR-NIR et N = 10000, 20000, …, 50000 pour les spectres de MIR⊗NIR combinée par l'OP. Les valeurs minimales de la fonction fitness indiquent les L et N optimaux selon les équations 3.19 et 3.20. Les autres paramètres que nous avons utilisés dans l'AG : la fraction de croisement $F_c = 0.8$, le nombre d'élites $N_e = 2$, la tolérance $\varepsilon = 10^{-6}$, le nombre d'itérations T=1000. Le choix de ces valeurs a été expliqué précédemment.

La Figure 3.8 montre les valeurs de fitness Davies Bouldin (DB) des différentes L et N pour les spectres MIR et NIR. D'après cette figure, nous avons trouvé que la taille de population N = 300 et celle de chromosomes L = 4 donne la plus petite valeur de fitness pour les spectres MIR et N = 400 avec L = 4 pour les spectres NIR. De la même manière, nous pouvons calculer pour les autres échantillons de biomasse lignocellulosique tel que la miscanthus. Nous obtenons le Tableau 3.1 suivant qui montre les meilleurs choix de valeurs de N et L pour les deux biomasses étudiées.
Nous constatons que peu importe la biomasse étudiée, ces valeurs sont assez proches. Ainsi, pour les spectres MIR et NIR, 4 nombres d'ondes sont discriminants et pour les spectres combinés par le produit extérieur entre 4 et 6 couples permettent de décrire la dégradation.

Tableau 3.1. Valeurs de N et L pour les spectres MIR, NIR, MIR-NIR et MIR⊗NIR de biomasse lignocellulosique : maïs, miscanthus et peuplier.

	maïs				miscanthus				peuplier			
	MIR	NIR	MIR⊗NIR	MIR-NIR	MIR	NIR	MIR⊗NIR	MIR-NIR	MIR	NIR	MIR⊗NIR	MIR-NIR
N	300	400	30000	400	400	300	40000	400	300	300	30000	4000
L	4	4	4	4	4	4	4	8	4	4	6	6

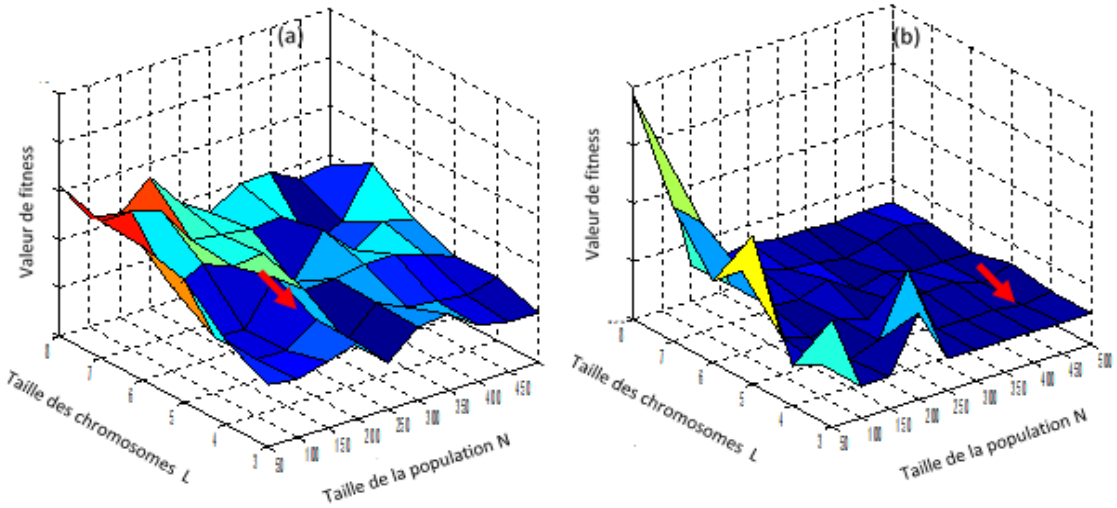


Figure 3.8. Racines de maïs. Valeurs de fitness de l'algorithme génétique avec la fonction Davies Bouldin pour différentes tailles de chromosomes L et différentes tailles de populations N pour : (a) les spectres MIR et (b) spectres NIR. Les valeurs minimales de la fonction fitness sont indiquées par des flèches rouges.

Nous nous sommes également intéressés à l'influence des fonctions fitness comme paramètres essentiels dans l'algorithme génétique, à savoir quelles sont ses performances en termes de classification des échantillons suivant la cinétique de biodégradation. Pour cela, nous avons appliqué l'algorithme génétique avec différentes fonctions fitness : Davis Bouldin (DB), l'indice Xie Beni (XB), l'indice Calinski-Harabasz (CH), l'indice Silhouette (SIL), l'indice Séparation (SI), et l'indice Fisher (FI). Ensuite, nous avons évalué les résultats de deux façons différentes par une représentation de scores plots en fonction des PC1 et PC2 et l'évaluation numérique à l'aide de l'indice DI.

Ce travail est détaillé dans la référence [RPB14], nous reprenons ici que les résultats les plus importants. La Figure 3.9 montre les représentations de scores plots (PC1 & PC2) de l' ACP appliquée

sur : (a) les spectres MIR enregistrés sur la gamme spectrale 800–1800 cm^{-1} ; (b)-(g) les informations spectrales MIR aux nombres d'ondes sélectionnés par l'AG avec les fonctions fitness suivantes: (b) l'indice Davis Bouldin (DB), (c) l'indice Xie Beni (XB), (d) Calinski-Harabasz (CH), (e) l'indice Silhouette (SIL), (f) l'indice de séparation (SI), et (g) l'indice de Fisher (FI). L'interprétation des bandes sélectionnées est présentée plus loin.

Les contributions importantes des valeurs singulières en pourcentages montrent que les deux premières composantes principales (PC1 & PC2) décrivent efficacement chaque ensemble de données considéré. Cette figure indique qualitativement que les nombres d'ondes sélectionnés par l'AG fournissent généralement une meilleure discrimination en fonction des périodes du processus de biodégradation que toute la gamme spectrale MIR. En effet, si l'on compare les Figure 3.9 (a) et Figure 3.9 (b), nous trouvons que les 5 classes correspondant aux périodes de biodégradation $\{t_1, t_2 \ldots . t_K\}$ sont plus faciles à séparer que dans la seconde Figure 3.9 (b).
De plus, la Figure 3.9 (b) montre que l'AG avec la fonction fitness Davies Bouldin fournit les meilleurs résultats. En effet, les cinq classes correspondant aux périodes de biodégradation sont très bien séparées par rapport aux autres fonctions fitness. L'indice Silhouette donne également un bon résultat, mais un échantillon de la classe t_0 peut être confondu avec un échantillon de la classe t_{36}. Nous observons des résultats identiques pour les spectres NIR enregistrés sur la même biomasse (racine des maïs) ainsi que sur des spectres MIR et NIR enregistrés sur d'autres types d'échantillons (miscanthus, peuplier). Ces résultats sont présentés dans l'annexe 3.

Pour quantifier la séparabilité des échantillons, nous avons calculé l'indice DI pour chaque résultat obtenu. Ces valeurs, présentées dans le Tableau 3.2, confirment que la fonction Davies Bouldin (DB) fournit le meilleur résultat de séparation des spectres à la fois pour les spectres MIR et NIR. De plus, nous avons obtenu des valeurs de DI pour les échantillons de maïs globalement plus faibles que pour les échantillons de miscanthus et de peuplier.

En ce qui concerne la comparaison entre MIR et NIR, nous pouvons affirmer que l'application de l'AG avec les différentes fonctions fitness sur les spectres MIR donne de meilleurs résultats que pour les spectres NIR.

Tableau 3.2. Valeurs de l'indice Dunn (DI) calculé sur les scores plots de PCA appliqués sur l'information spectrale MIR et NIR

		Sur gamme spectrale 800-1800 cm^{-1} pour MIR et 4000-6000 cm^{-1} pour NIR	aux nombres d'ondes sélectionnés par l'algorithme génétique avec les fonctions fitness :					
			Davis Bouldin (DB)	Xie Beni (XB)	Calinski-Harabasz (CH)	Silhouette (SIL)	Séparation (SI)	Fisher (FI)
Maïs	MIR	0.074	**0.173**	0.131	0.107	0.156	0.056	0.107
	NIR	0.050	**0.112**	0.055	0.063	0.101	0.064	0.063
Miscanthus	MIR	0.18	**0.59**	0.152	0.137	0.272	0.089	0.137
	NIR	0.164	**0.41**	0.121	0.124	0.241	0.074	0.124
Peuplier	MIR	0.118	**0.87**	0.174	0.135	0.39	0.110	0.135
	NIR	0.146	**0.75**	0.166	0.131	0.32	0.098	0.131

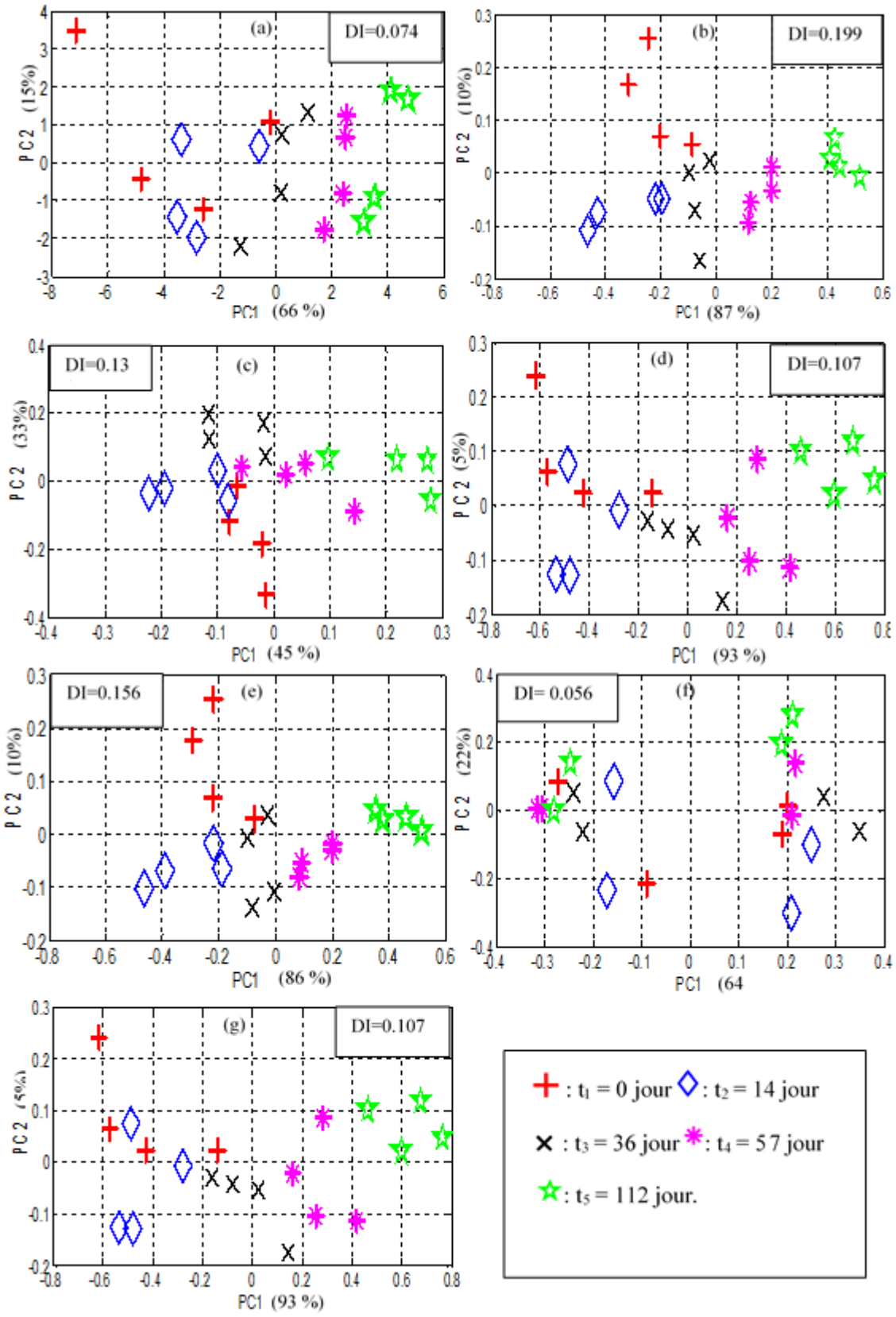


Figure 3.9. Racines de maïs. Scores plots représentant la discrimination selon les périodes du processus de biodégradation en termes de PC1 vs PC2. L'ACP appliquée sur : (a) les spectres enregistrées sur la gamme spectrale 800-1800 cm^{-1} du MIR, (b-g) les informations MIR sélectionnées aux nombres d'ondes identifiés par l'AG avec les fonctions fitness suivantes : (b) Davis Bouldin (DB), (c) Xie Beni (XB), (d) Calinski-Harabasz (CH), (e) Silhouette (SIL), (f) Séparation (SI), et (g) Fisher (FI).

III.7.2.1. Discussion sur les nombres d'ondes sélectionnés par l'AG avec différentes fonctions fitness

Comme nous venons de le voir, la sélection de nombres d'ondes par l'AG nous permet de séparer les différentes classes correspondant aux différents temps du processus de biodégradation. Cependant, il nous semble essentiel d'analyser quelles sont les valeurs des nombres d'ondes qui ont été sélectionnés avec les différentes fonctions fitness afin de déterminer leur sens physique. Le Tableau 3.3 représente les significations chimiques au voisinage de ces nombres d'ondes dans les régions MIR et NIR pour les échantillons de racines de maïs [KCC12, PP13, FGP14, SRF11, WSK08, FHX09]. Des études sur autres types de biomasse tels que la miscanthus et le peuplier sont présentées dans l'annexe 3.

Tableau 3.3. Significations chimiques des nombres d'ondes sélectionnés par l'algorithme génétique basés sur les différentes fonctions fitness pour les spectres MIR et NIR

	Davies-Bouldin (DB)	Calinski-Harabasz (CH)	Xie Beni (XB)	séparation index (SI)	Silhouette Index (SIL)	Fisher Index (FI)
MIR	• **858 cm^{-1}** = vibrations du squelette aromatiques combinés avec CH wag, • **953 cm^{-1}** = C-O-C étirement des polysaccharides • **1383 cm^{-1}** = cellulose avec lignine (aliphatiques CH étirement dans CH_3), • **1707 cm^{-1}** = hémicellulose (C=O étirement des cétones non conjugués, les carbonyles et en ester).	• **942 cm^{-1}** = C-O-C étirement des polysaccharides • **1382 cm^{-1}** = cellulose avec lignine (aliphatiques CH étirement dans CH_3), • **1529 cm^{-1}** = lignine (les vibrations du squelette aromatiques). • **1706 cm^{-1}** = hémicellulose (C=O étirement des cétones non conjugués, les carbonyles et en ester).	• **1365** et **1380 cm^{-1}** = cellulose avec la lignine (aliphatique CH étirement dans CH_3), • **1055 cm^{-1}** = cellulose (la vibration C–O–C) • **1546 cm^{-1}** = lignine (les vibrations du squelette aromatiques).	• **845** et **896 cm^{-1}** = vibrations du squelette aromatiques combinés avec CH wag, • **1098 cm^{-1}** = cellulose (C-O étirement), • **1321 cm^{1}** = lignine (les vibrations CH_2 étirement)	• **875 cm^{-1}** = vibrations du squelette aromatiques combinés avec CH wag, • **953 cm^{-1}** = C-O-C stretching des polysaccharides, • **1382 cm^{-1}** = cellulose avec lignine (aliphatiques CH étirement dans CH_3), • **1706 cm^{-1}** = hémicellulose (C=O étirement des cétones non conjugués, les carbonyles et en ester).	• **942 cm^{-1}** = C-O-C étirement des polysaccharides • **1382 cm^{-1}** = cellulose avec la lignine (aliphatiques CH étirement dans CH_3), • **1529 cm^{-1}** = lignine (les vibrations du squelette aromatiques). • **1706 cm^{-1}** = hémicellulose (C=O étirement des cétones non conjugués, les carbonyles et en ester).
NIR	• **4850 cm^{-1}** = CO avec CH_3 et l'étirement OH + déformation O-H	• **4308 cm^{-1}** = combinaison étirement C–H • **4651 cm^{-1}** = les bandes	• **4268** et **4279 cm^{-1}** = cellulose (étirement CH + déformation CH),	• **4092 cm^{-1}** aucune signification chimique • **4669** et **4696 cm^{-1}** = les bandes de la	• **4848** et **4850 cm^{-1}** = CO avec CH_3 et l'étirement OH +	• **4308 cm^{-1}** = combinaison étirement C–H, • **4651 cm^{-1}** = les bandes de la combinaison

	• **5195 cm^{-1}** = l'eau (O-H antisymétrique, vibration d'élongation + O-H vibration de déformation H_2O). • **5540 cm^{-1}** = cellulose avec lignine (1^{st} overtane CH_3 et -CH = CH-), • **5705 cm^{-1}** = Première harmonique du CH étirement.	de la combinaison de groupes OH et de CO • **4727 cm^{-1}** aucune signification chimique • **4850 cm^{-1}** = CO avec CH_3 et l'étirement OH + déformation O-H	• **4518 cm^{-1}** aucune signification chimique • **4804 cm^{-1}** = cellulose (étirement OH + déformation CH)	combinaison de groupes OH et de CO	déformation O-H • **5666 et 5712 cm^{-1}** = première harmonique du CH étirement.	de groupes OH et de CO • **4727 cm^{-1}** aucune signification chimique • **4850 cm^{-1}** = C = O avec CH_3 et l'étirement OH + déformation O-H

Pour les spectres MIR, nous avons trouvé que tous les nombres d'ondes sélectionnés par l'AG basé sur la fonction Davies Bouldin (DB) ont été identifiés au voisinage des principales vibrations des groupes fonctionnels chimiques des composés qui sont liés à la biomasse lignocellulosique (tableaux 3.3, A.3.1, A.3.2).

Si l'algorithme génétique est appliqué avec les fonctions Calinski-Harabasz (CH) et Fisher, nous observons que les nombres d'ondes 943, 1382 et 1706 cm^{-1} ont les mêmes significations chimiques que les nombres d'ondes 953, 1385 et 1709 cm^{-1} sélectionnés par l'AG avec la fonction Davies Bouldin (DB). Par contre 1529 cm^{-1} a une signification chimique différente du nombre d'onde 858 cm^{-1} choisi par l'AG avec la fonction DB qui donne la meilleure séparabilité des classes. Pour la fonction Xie-Beni (XB), les nombres d'ondes 1365 et 1380 cm^{-1} ont les mêmes significations que 1385 cm^{-1}. 1055 cm^{-1} correspond à la même liaison chimique que 953 cm^{-1}, mais le nombre d'onde 1546 cm^{-1} a une signification chimique différente du nombre d'onde 1709 cm^{-1} choisi par l'AG avec la fonction DB. La fonction de séparation (SI) a sélectionné les nombres d'ondes 1098, 1321, 845 et 896 cm^{-1} qui ont des significations chimiques voisines de 1383, 953 et 858 cm^{-1} (également choisi par DB), mais sont moins performantes pour discriminer les échantillons en fonctions des périodes de biodégradation (Figure 3.9). La fonction Silhouette (SIL) a sélectionné des nombres d'ondes qui ont les mêmes significations chimiques, mais sont moins pertinents pour la classification des échantillons (Figure 3.9) que les nombres d'ondes sélectionnés par la fonction Davies Bouldin (DB).

Pour les spectres NIR, tous les nombres d'ondes sélectionnés par l'AG basé sur la fonction Davies Bouldin (DB) ont des significations chimiques intéressantes. Les fonctions Calinski-Harabasz (CH) et Fisher ont sélectionnés les nombres d'ondes 4308 et 4651 cm^{-1} qui ne sont pas identiques, mais qui ont des significations chimiques voisines de 4850 cm^{-1}. Le nombre d'onde 4727 cm^{-1} n'a pas une signification chimique, tandis que 4850 cm^{-1} a également été trouvé par la fonction DB. La fonction Xie-Beni (XB) sélectionne le nombre d'onde 4518 cm^{-1} qui n'a pas de signification chimique. Les nombres d'ondes 4268, 4279 et 4808 cm^{-1} ont des significations chimiques différentes des nombres d'ondes choisis par DB qui permettent une meilleure séparabilité des classes (voir la Figure 3.9). La

fonction de séparation (SI) trouve le nombre d'onde 4092 cm^{-1} qui n'a pas de signification chimique mais les nombres d'ondes 4669 et 4696 cm^{-1} ont les mêmes significations chimiques que le nombre d'onde 4850 cm^{-1} choisi par la fonction DB. La fonction Silhouette (SIL) a choisi des nombres d'ondes qui ont les mêmes significations chimiques. On peut en tirer les mêmes conclusions que pour le spectre MIR.

En conclusion, nous pouvons dire que la fonction fitness Bouldin Davies (DB) permet d'obtenir la meilleure séparabilité de classes tout en réalisant l'identifiant des principales vibrations des groupes fonctionnels chimiques des composés liés à la biomasse lignocellulosique. Les autres fonctions fitness identifient partiellement les mêmes nombres d'ondes, ou des nombres d'ondes n'ayant pas de significations chimiques.

III.7.3. Résultats obtenus par l'AG pour l'analyse de la dégradation de la biomasse lignocellulosique

Nous rappelons les paramètres que nous avons utilisés dans l'algorithme génétique : la fraction de croisement $F_c = 0.8$, le nombre d'élites $N_e = 2$, la tolérance ε = 10^{-6}, le nombre d'itérations T=1000. Les choix de N et L sont présentés dans le Tableau 3.1. Dans cette section, nous ne montrons que les résultats obtenus par l'algorithme génétique avec la fonction fitness Davies Bouldin (DB) sur les échantillons des racines de maïs, les autres résultats pour les échantillons miscanthus et peuplier sont présentes dans l'annexe 3.

L'évaluation du processus de biodégradation est généralement faite en analysant les scores plots de l'ACP appliquée sur les gammes spectrales entières MIR / NIR ou sur les gammes spectrales sélectionnées par les spécialistes. Ici, nous avons appliqué l'ACP sur les gammes spectrales entières MIR et NIR, la combinaison des gammes MIR-NIR par concaténation, et la combinaison des spectres MIR⊗NIR par produit extérieur OP.

Nous avons également appliqué l'ACP sur les amplitudes des spectres aux nombres d'ondes qui ont été identifiés par l'AG avec la fonction fitness (DB) (MIR: 858, 953, 1383 et 1707 cm^{-1}, NIR: 4850, 5540, et 5705 cm^{-1}. Puis celles identifiées par la combinaison MIR-NIR par concaténation: 858; 1384; 5705 et 5541 cm^{-1}. Et enfin, les couples de nombres d'ondes identifiés par la combinaison MIR⊗NIR par la produit extérieur OP : 1385 x 4141 ; 920 x 4887; 956 x 4852 ; 922 x 5772 ; 1383 x 4659 ; 1248 x 4177 cm^{-1}).

Les Figure 3.10 montrent les scores plots (PC1 & PC2) de l'ACP obtenus pour les différents cas. Les pourcentages de contribution des valeurs singulières indiquées pour les composantes principales montrent que celles-ci décrivent efficacement chaque ensemble de données considéré.

La Figure 3.10 (d) montre que la combinaison de spectres MIR⊗NIR par le produit extérieur OP permet une meilleure discrimination au cours des périodes du processus de biodégradation que l'utilisation des gammes spectrales entières MIR, NIR ainsi que les gammes concaténées MIR-NIR (Figure 3.10 (a) (b) (c)).

Nous pouvons voir que les Figure 3.10 (e) (f) (g) (h) dans lesquelles les nombres d'ondes ont été sélectionnés par algorithme génétique avec la fonction fitness DB donnent qualitativement une meilleure séparabilité des échantillons que l'utilisation des bandes spectrales entières. La Figure 3.10 (h) indique que la sélection des nombres d'ondes par l'AG pour la combinaison des spectres MIR et NIR par le produit extérieur (OP) fournit les meilleurs résultats qualitatifs.

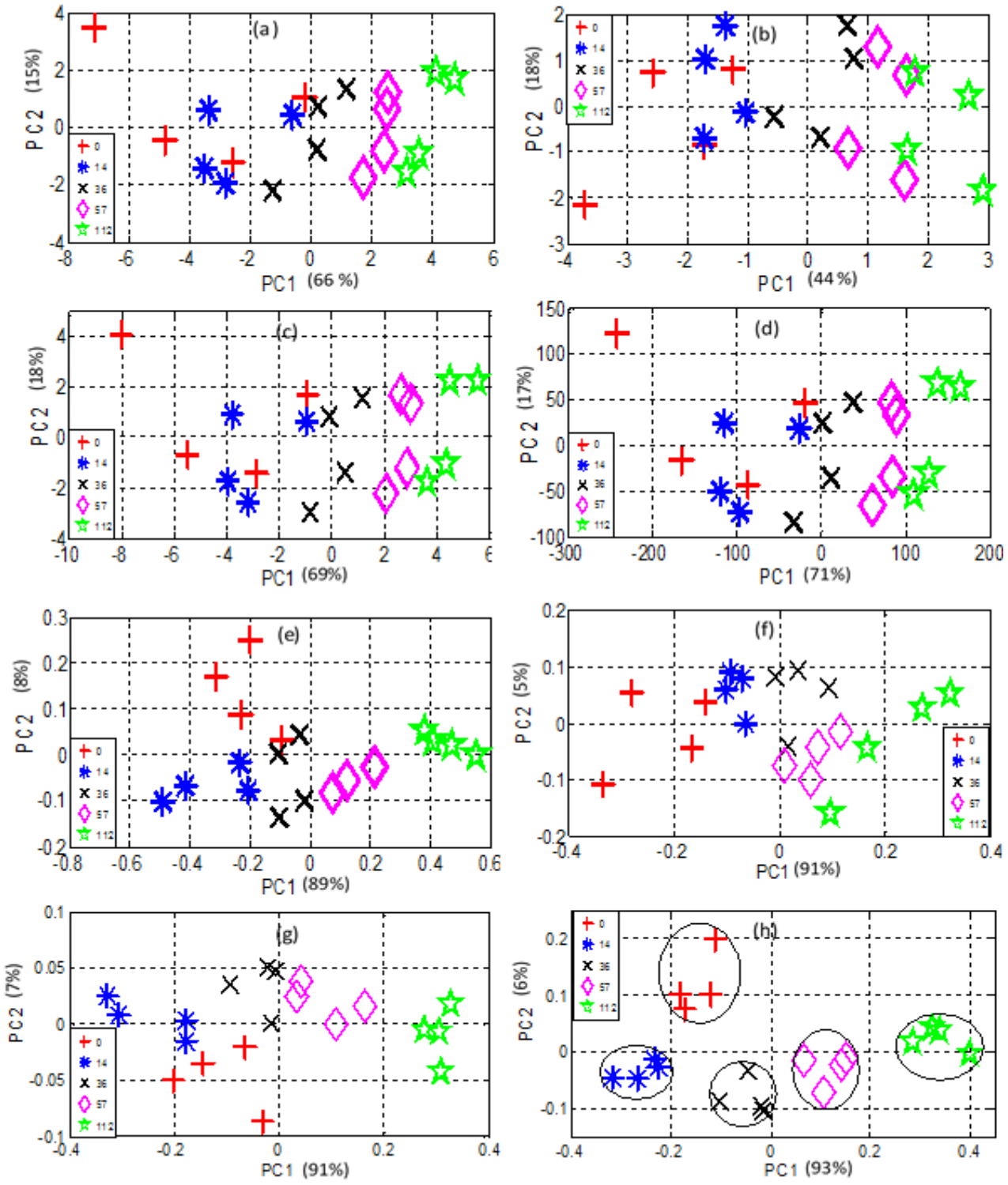


Figure 3.10. Racines de maïs : Scores plots montrant la discrimination suivant les périodes du processus de biodégradation en termes de PC1 vs PC2. La PCA a été appliquée sur: (a) la gamme spectrale de MIR 800-1800 cm^{-1}, (b) la gamme spectrale de NIR 4000-6000 cm^{-1}, (c) les gammes concaténées MIR-NIR, (d) les spectres combinés MIR⊗NIR par le produit extérieur OP, (e) les nombres d'onde sélectionnés par l'AG sur MIR, (f) les nombres d'ondes sélectionnés par l'AG sur NIR ,(g) les nombres d'ondes sélectionnés par l'AG sur les gammes MIR-NIR concaténées, (h) les nombres d'ondes sélectionnés par l'AG sur les spectres combinés MIR⊗NIR par le produit extérieur. La légende indique les temps de décomposition en jour.

Pour quantifier la séparabilité des classes selon les périodes du processus de biodégradation, nous avons calculé l'indice Dunn (DI) pour chaque score plot de l'ACP de la Figure 3.10. Ces valeurs présentées dans le Tableau 3.4 confirment les observations qualitatives et montrent une amélioration importante de la séparabilité lorsque nous appliquons l'AG avec la fonction fitness DB sur la combinaison des spectres MIR⊗NIR par le produit extérieur (OP). Nous obtenons pour celle-ci une valeur de l'indice (DI) remarquablement élevée [RPC15].

Tableau 3.4. Les valeurs de l'indice Dunn (DI) pour les biomasses lignocellulosiques. Une valeur plus élevée de DI signifie une meilleure discrimination dans le processus de biodégradation.

DI	MIR	NIR	MIR-NIR	MIR⊗NIR	MIR avec AG	NIR avec AG	MIR-NIR avec AG	MIR⊗NIR avec AG(OP-AG)
Maïs	0.074	0.050	0.120	0.127	0.173	0.11	0.19	**0.48**
Miscanthus	0.188	0.164	0.21	0.25	0.59	0.41	0.78	**0.88**
Peuplier	0.118	0.146	0.18	0.32	0.87	0.75	0.90	**1.32**

III.7.3.1. Analyse des nombres d'ondes sélectionnés par AG

L'algorithme génétique avec la fonction fitness Davies-Bouldin permet l'identification de nombres d'ondes dans les deux gammes spectrales MIR et NIR au voisinage des groupes fonctionnels chimiques qui peuvent ainsi être attribués à l'évolution chimique des échantillons étudiés au cours de la biodégradation. l'AG sélectionne quatre nombres d'ondes dans la gamme MIR: 858 cm^{-1} (proche des informations spectrales liées à la cellulose, hémicellulose et la lignine : Anomere C-groupes, la déformation CH, les vibrations de valence); 956 cm^{-1} (au voisinage de la signature moléculaire de C-O-C étirement des polysaccharides); 1383 cm^{-1} (proche des informations spectrales liées cellulose et lignine : aliphatique CH étirement dans CH_3, vibration déformation CH) ; 1707 cm^{-1} (au voisinage de la signature de l'hémicellulose : la liaison C=O étirement des cétones non conjugués, les carbonyles et les ester). Dans la gamme NIR, les informations relatives aux structures de lignocellulose sont également obtenues pour les nombres d'ondes sélectionnés tels que 4850 cm^{-1} (au voisinage de la signature de cellulose : vibration de déformation O-H et C-H + vibration d'élongation O-H et C=O) ; 5195 cm^{-1} (O-H antisymétrique, vibration d'élongation + O-H vibration de déformation H2O) ; 5540 cm^{-1} (cellulose : 1ère vibration d'élongation harmonique C-H) ; 5705 cm^{-1} (cellulose, hémicellulose et lignine : 1ère vibration d'élongation harmonique C-H). Sur la gamme de combinaison par la concaténation MIR-NIR, l'algorithme génétique a sélectionné les mêmes nombres d'ondes qu'il avait sélectionnés pour le MIR et le NIR : 858 ; 1384 ; 5705 et 5541 cm^{-1}.

Pour les spectres MIR⊗NIR combinés par le produit extérieur OP, l'AG a sélectionné les paires de nombres d'ondes suivants : (1385 x 4141) ; (920 x 4887) ; (956 x 4852) ; (922 x 5772) ; (1383 x 4659) ; (1248 x 4177). Les nombres d'ondes sélectionnés dans la région MIR : 920 et 922 cm^{-1} est dans le voisinage des informations spectrales liées à la cellulose, l'hémicellulose et la lignine (Anomere C-groupes, déformation CH, et la vibration de valence) ; 956 cm^{-1} au voisinage de C-O-C étirement des polysaccharides ; 1248 cm^{-1} est dû à la présence d'éthers ou d'esters d'acide (C-O étirement) ; et les nombres d'ondes 1383 et 1385 cm^{-1} correspondent au voisinage des informations spectrales liées à la cellulose et la lignine (aliphatique CH étirement dans CH_3, vibration déformation CH). Les nombres d'ondes sélectionnés dans la région NIR : 4141 et 4177 cm^{-1} correspondent à la lignine ; 4659 cm^{-1} correspond à l'hémicellulose (la combinaison CH étirement dans la molécule -CH = CH-). Les nombres d'ondes 4852 et 4887 cm^{-1} représentent la voisinage de la signature cellulose (vibration de

déformation O-H et C-H + vibration d'élongation O-H et C=O avec CH_3) ; et 5772 cm^{-1} correspond à la cellulose (la premier « overtone » du C-H étirement).

Les nombres d'onde sélectionnées par l'AG sur les spectres MIR⊗NIR sont très proches des principales vibrations des groupes fonctionnels chimiques des composés qui subissent une dégradation / conversion au cours de la biodégradation de la biomasse lignocellulosique. Cependant, nous pouvons nous interroger sur leur consistance puisque d'importantes variations apparaissent également à d'autres nombres d'ondes ou peuvent être attendues à d'autres nombres d'onde.

Pour discuter de ce problème, nous allons nous intéresser aux nombres d'ondes suivants : 1030 et 1450 cm^{-1}, indiqués par les flèches noires en pointillé sur la Figure 3.11. Nous avons choisi à priori ces deux nombres d'ondes qui correspondent aux fonctions chimiques et/ou variations importantes des spectres MIR en fonction du stade de dégradation qu'on peut visualiser. Nous choisissons également 953 et 1383 cm^{-1} indiquées par les flèches rouges sur la Figure 3.11, 956 et 1385 cm^{-1} indiquées par les flèches vertes et les couples (1385 x 4141) et (956 x 4852) cm^{-1}. Ces valeurs correspondent respectivement aux deux nombres d'onde sélectionnées par l'AG pour les spectres MIR, aux nombres d'ondes MIR sélectionnées par l'AG pour les spectres MIR⊗NIR combinés par l'OP, aux deux paires (MIR, NIR) de nombres d'ondes sélectionnés par l'AG pour les spectres MIR⊗NIR.

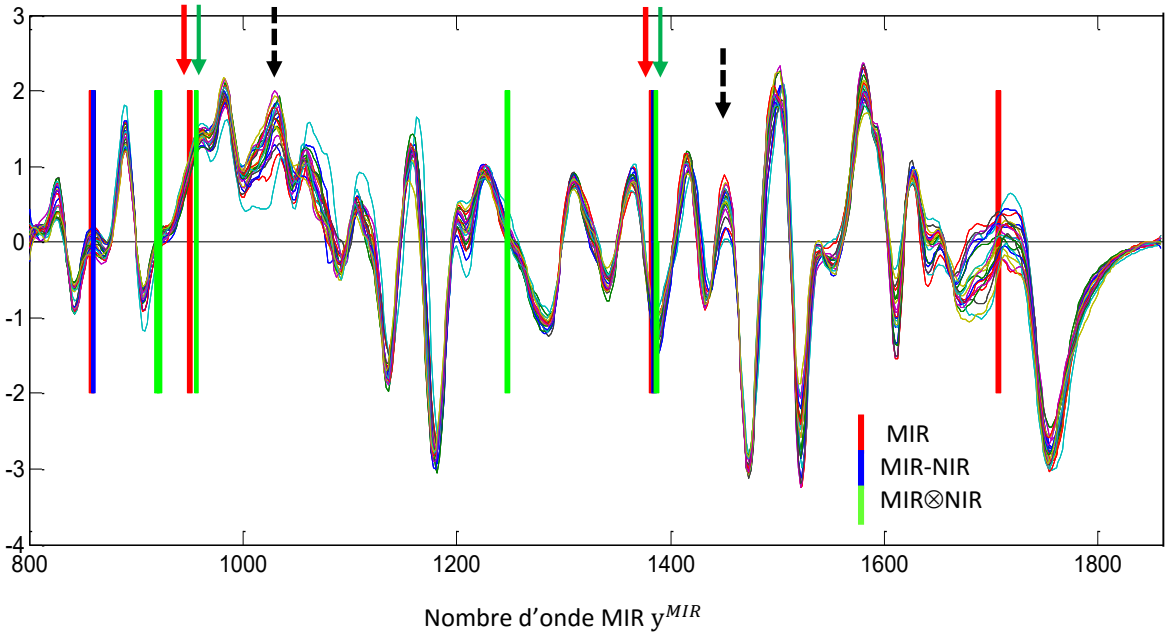


Figure 3.11. Spectres MIR enregistrés sur les J = 20 échantillons de racines de maïs prétraités par la dérivation Savitzky-Golay (SG) du 1ere ordre avec un lissage sur 17 points et un polynôme d'ordre 4, suivie d'une normalisation de type Standard Normal Variate (SNV). Les nombres d'ondes ont été sélectionnés par l'AG dans la gamme MIR, la combinaison de deux gammes MIR-NIR par concaténation, et le produit extérieur MIR⊗NIR.

La Figure 3.12 (a) montre que les résultats de classification des échantillons selon les périodes de biodégradation par rapport aux nombres d'ondes 1030 et 1450 cm^{-1} sont très médiocres. D'autre part, si nous combinons l'information spectrale MIR et NIR en utilisant le produit extérieur OP, mais que nous considérons uniquement l'information spectrale aux nombres d'onde MIR (Figure 3.12 (c)), la distribution de l'information spectrale met mieux en évidence le processus de dégradation que la simple utilisation des nombres d'ondes sélectionnés par l'AG sur la gamme MIR (Figure 3.12 (b)).

Finalement, si nous considérons l'information spectrale aux paires (MIR, NIR) de nombres d'ondes sélectionnés par l'AG sur les spectres MIR⊗NIR combinés par l'OP (Figure 3.12 (d)), les échantillons sont mieux clustérisés et le processus de dégradation est mieux mis en évidence. Cela signifie que les informations MIR ne doivent pas être considérées séparément des informations NIR. Cela explique aussi pourquoi les nombres d'ondes sélectionnés par l'AG dans la région MIR et la concaténation MIR-NIR ne correspondent pas à ceux sélectionnés sur les spectres combinés MIR⊗NIR par l'OP (voir le Figure 3.11).

Dans la suite de cette section, nous étudions les performances de l'algorithme génétique avec la fonction Davies Bouldin (DB) sur les spectres MIR⊗NIR combinés par l'OP prétraités par différentes méthodes de prétraitements.

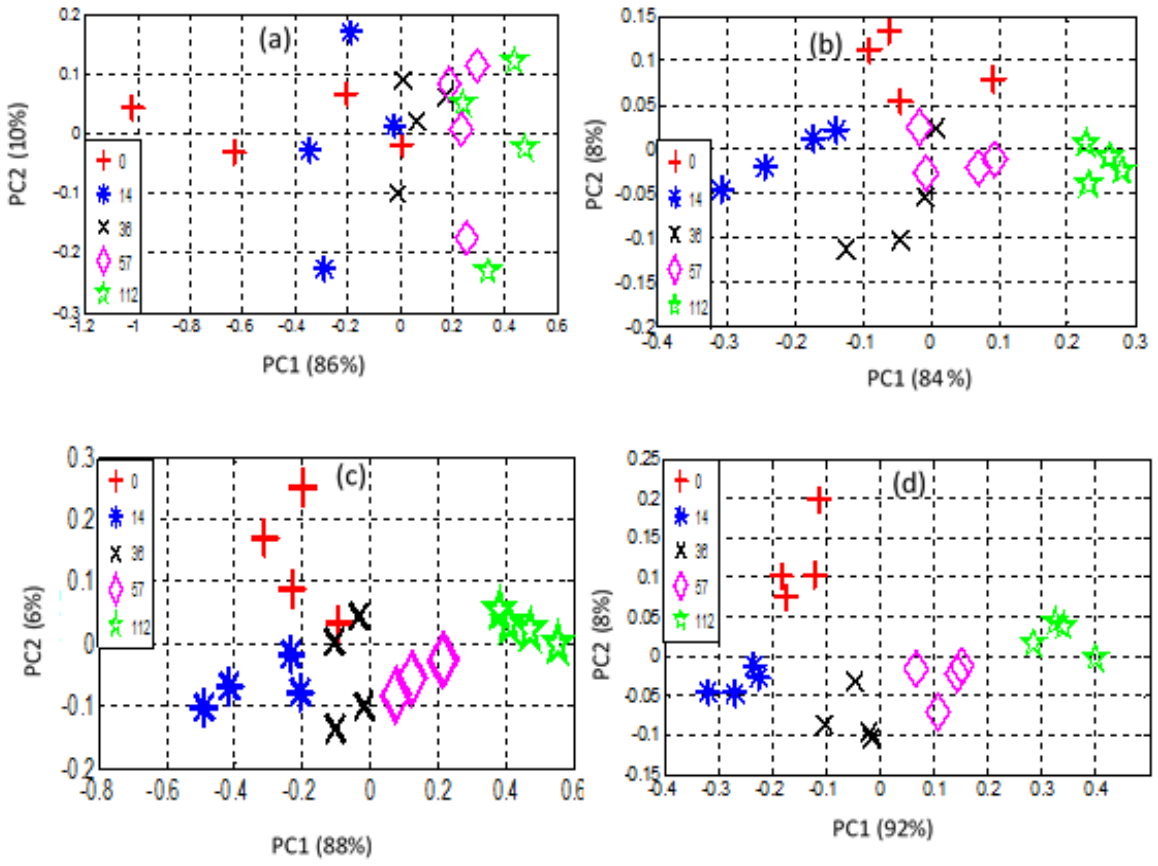


Figure 3.12. Biodégradation des racines de maïs. Représentations des scores plots de la distribution de l'information spectrale à: (a) 1030 et 1450 cm^{-1}, la variation la plus importante dans la région spectrale MIR; (b) 953 et 1383 cm^{-1}, deux nombres d'onde sélectionnées par l'AG de la gamme MIR; (c) 956 et 1385 cm^{-1}, correspondant uniquement à les nombres d'ondes MIR sélectionnées par l'AG des spectres MIR⊗NIR combinée par OP; (d) (1385 x 4141) et (956 x 4852) cm^{-1}, correspondant à deux paires(MIR, NIR) de nombres d'ondes sélectionnés par l'AG des spectres MIR⊗NIR combinés par l'OP.

III.7.3.2. Discussion sur la méthode de prétraitement des spectres combinés par le produit extérieur

Nous avons vu dans les chapitres précédents que le prétraitement des spectres est une étape importante qui va influencer sur le résultat final. Nous avons testé en plus de la méthode optimale identifiée dans le chapitre précédent (dérivée SG d'ordre 1 suivie de la normalisation SNV), les méthodes couramment utilisées dans ce type d'application, qui sont : la dérivation Savitzky-Golay (SG) d'ordre 2 suivie par la normalisation SNV, la MSC et l'EMSC.

Le Tableau 3.5 montre les valeurs de l'indice Dunn (DI) pour les scores plots obtenus pour chaque méthode de prétraitement sur les spectres combinés par le produit extérieur MIR⊗NIR. Ces valeurs confirment que la méthode de dérivation SG d'ordre 1 suivie par la SNV est la meilleure méthode de prétraitement sur ce type de spectres. Ce résultat confirme les conclusions précédentes obtenues sur les spectres IR de biomasse lignocellulosique lorsque nous avions appliqué les méthodes de classification non supervisées FCM.

Tableau 3.5. Valeurs de l'indice Dunn (DI) calculées pour les différentes méthodes de prétraitement. L'ACP a été appliqué sur l'information spectrale aux couples de nombres d'ondes sélectionnés par l'AG avec la fonction fitness Davies-Bouldin (DB) des spectres MIR⊗NIR combinés par le produit extérieur OP.

DI	SG 1 + SNV	SG 2 + SNV	MSC	EMSC
Racines de maïs	**0.48**	0.16	0.28	0.07

III.7.3.3. Discussion sur le choix de la fonction fitness optimale pour des spectres combinés par le produit extérieur

Comme nous l'avons mentionné auparavant, la fonction fitness est la partie la plus importante d'un algorithme génétique. Le rôle d'une fonction fitness est de mesurer la qualité du chromosome de la population selon l'objectif d'optimisation donné. Nous testons ici les différentes fonctions fitness basées sur des indices de validité que nous avons utilisées dans la section 3.2 (dans ce chapitre) pour mesurer la séparabilité des clusters pour les spectres combinés par le produit extérieur MIR⊗NIR de la biomasse lignocellulosique.

Pour comparer l'influence de la fonction fitness, nous avons calculé l'indice Dunn (DI) pour les scores plots obtenus pour chaque fonction fitness sur les spectres combinés par le produit extérieur MIR⊗NIR. Les valeurs de Tableau 3.6 indiquent que la fonction fitness Davies Bouldin (DB), qui maximise toutes les distances inter-cluster et minimise la distance intra-cluster pour chaque cluster, donne les meilleurs résultats, ce qui confirme aussi les résultats précédents sur l'efficacité de la fonction fitness Davies Bouldin (DB) pour les spectres MIR. Le prétraitement utilisé a été la dérivation SG d'ordre 1 suivie par la SNV.

Tableau 3.6. Valeurs de l'indice Dunn (DI) calculées pour différentes fonctions fitness. L'ACP a été appliqué sur l'information spectrale aux couples de nombres d'ondes sélectionnés par l'AG des spectres combinée par le produit extérieur MIR⊗NIR

Fonction fitness	Davis Bouldin (DB)	Xie Beni (XB)	Calinski-Harabasz (CH)	Silhouette (SIL)	Séparation (SI)	Fisher (FI)
Racines de mais	**0.48**	0.101	0.14	0.43	0.075	0.14

Nous allons dans la section suivante appliquer la méthodologie basée sur l'optimisation PQS associée aux différents choix de contraintes afin d'estimer ses performances dans la sélection des nombres d'ondes pour la classification d'échantillons pendant les périodes de biodégradations. Ces résultats sont comparés avec ceux obtenus par algorithme génétique.

III.7.4. Résultats de l'application de la méthode d'optimisation PQS pour l'analyse de la dégradation de la biomasse lignocellulosique

Après avoir testé l'algorithme génétique, nous allons nous intéresser à l'utilisation de la méthode d'optimisation PQS décrite en début de chapitre. La méthode a été appliquée sur les spectres MIR et NIR séparés et les spectres combinés par concaténation MIR-NIR avec les contraintes L_1, L_2 et L_∞. Les

spectres combinés par le produit extérieur sont de très grande taille et ne peuvent être étudiés par l'approche actuelle d'optimisation. Dans cette section, nous présentons les résultats obtenus sur les échantillons de dégradation de la biomasse lignocellulosique (racines de maïs) [RFG15].

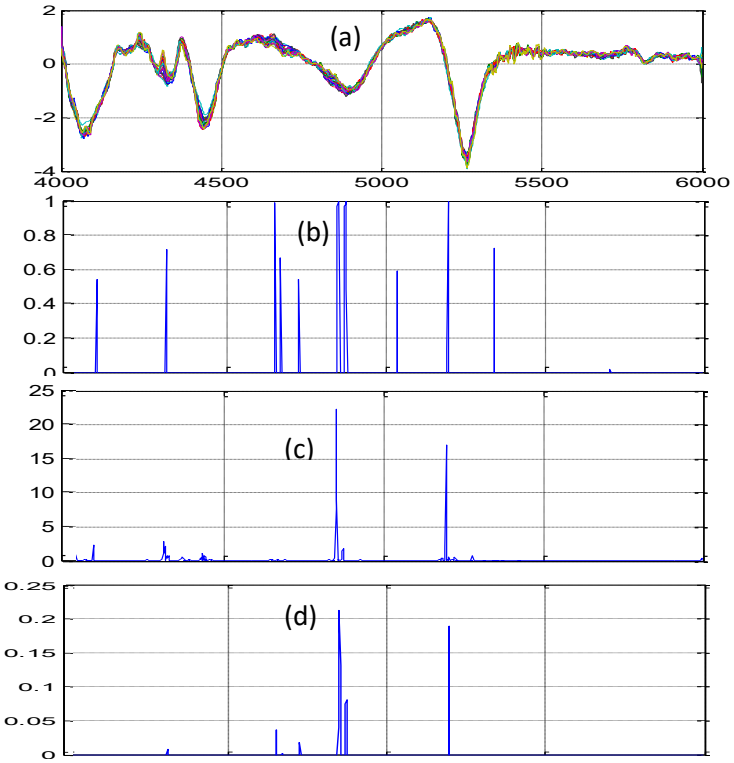


Figure 3.13. (a) : 20 spectres NIR prétraités par la méthode de dérivation Savitzky-Golay (SG) de 1er ordre avec un lissage sur 17 points et un polynôme d'ordre 4, enregistrés sur les quatre échantillons de maïs avec K = 5 périodes de biodégradation. Les poids w_i identifiés par l'approche d'optimisation proposée avec les contraintes (b) L_1 ; (c) L_∞ ; (d) L_2

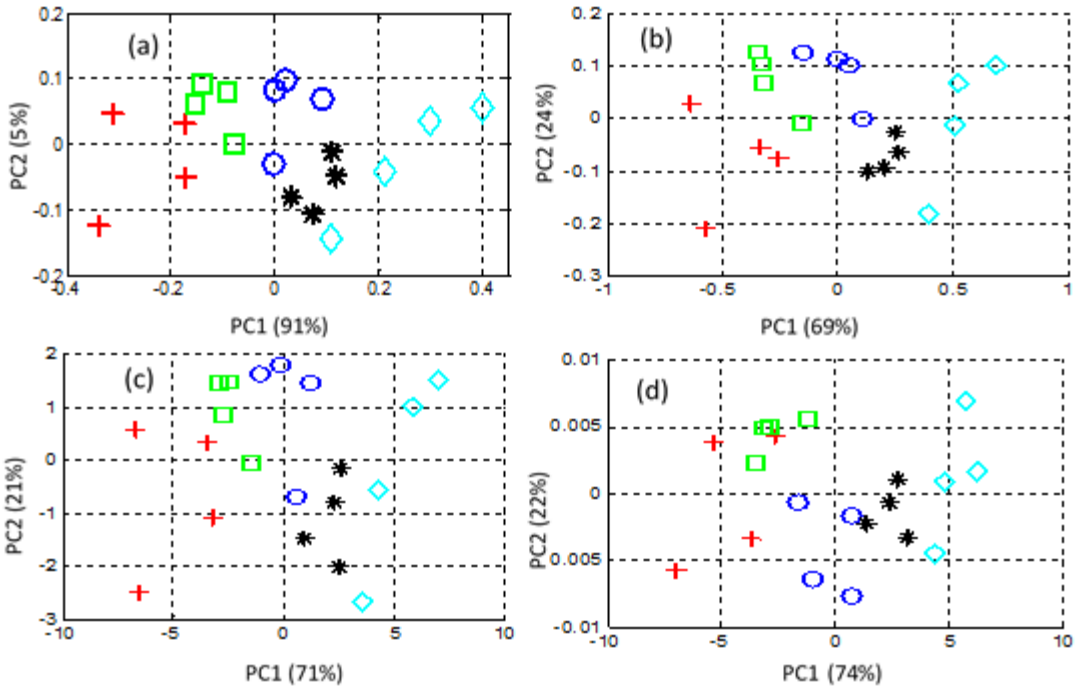


Figure 3.14. Scores plots obtenus sur les nombres d'ondes (NIR) identifiés par : (a) l'algorithme génétique ; (b) l'approche d'optimisation proposée avec (b) L_1; (c) L_∞ ; (d) L_2

La Figure 3.13 montre le vecteur de poids $\underline{w}$ identifié par l'approche d'optimisation proposée avec les contraintes L_1, L_2 et L_∞ appliquées sur les spectres NIR. Ce vecteur représente la pondération des nombres d'ondes sélectionnés. Nous affichons également la figure des scores plots de l'ACP (Figure 3.14) pour montrer la séparabilité des échantillons. De plus, pour quantifier les résultats de classifications, nous calculons l'indice de Dunn (DI).

Tableau 3.7. Valeurs de l'indice Dunn calculées sur les scores plots pour différents choix de normes. L'ACP a été appliqué sur l'information spectrale aux nombres d'ondes sélectionnés par l'optimisation PQS des spectres NIR.

DI	AG	L_1	L_2	L_∞
NIR	**0.11**	**0.23**	0.12	0.15

Tableau 3.8. Nombres d'ondes sélectionnés par l'optimisation PQS avec différents choix de contraintes sur les spectres NIR.

Bandes sélectionnées	AG	L_1	L_2	L_∞
NIR	4850 ; 5540 ; 5707	4158 ; 4335 ; 4612 ; 4684 ; 4785 ; 4850 ; 4885 ; 5442	4655 ; 4850 ; 4873 ; 5195 ;	4088 ; 4245 ; 4850 ; 5235

D'après les figures des scores plots, nous observons que l'approche d'optimisation PQS avec la contrainte L_1 (Figure 3.14 (a)) donne une meilleure discrimination des échantillons par rapport à la cinétique de biodégradation que les autres contraintes L_2 et L_∞ (Figure 3.14 (b) et Figure 3.14 (c) respectivement). Ce qui revient à dire que l'optimisation avec la contrainte L_1 permet d'obtenir la valeur minimale de la fonction objectif DB (la contrainte L_1 est bien adaptée pour les fonctions non convexes) dans notre l'algorithme pour les spectres NIR. Ce résultat est confirmé par les valeurs de l'indice Dunn (DI) regroupées dans le Tableau 3.7. Si nous réalisons la comparaison entre notre approche par optimisation PQS et l'algorithme génétique à partir des valeurs de DI obtenues, nous trouvons pour l'algorithme génétique une valeur de DI = 0.11 et pour la méthode d'optimisation DI = 0.23. L'approche d'optimisation avec la contrainte L_1 est donc plus discriminante que l'algorithme génétique parce qu'il prend en compte le vecteur de poids qui met en évidence les nombres d'ondes les plus discriminants dans la gamme spectrale NIR, ce qui permet une meilleure séparabilité entre les classes. De plus, l'algorithme d'optimisation a sélectionné un plus grand nombre de nombres d'ondes qui correspondent aux vibrations principales des groupes fonctionnels chimiques (Tableau 3.8).

Tableau 3.9. Valeurs de l'indice Dunn (DI) calculées sur les scores plots pour différents choix de normes. L'ACP a été appliquée sur l'information spectrale aux nombres d'ondes sélectionnés par l'approche d'optimisation des spectres MIR

DI	AG	L_1	L_2	L_∞
MIR	0.173	**0.19**	0.17	0.10

Tableau 3.10. Nombres d'ondes sélectionnés par l'approche d'optimisation avec différentes choix de contraintes sur les spectres MIR

Bandes sélectionnées	AG	L_1	L_2	L_∞
MIR	858 ; 953 ; 1383 ; 1707	853 ; 912 ; 1383 ; 1706	850 ; 914 ; 1385 ; 1707	912 ; 1383 ; 1706

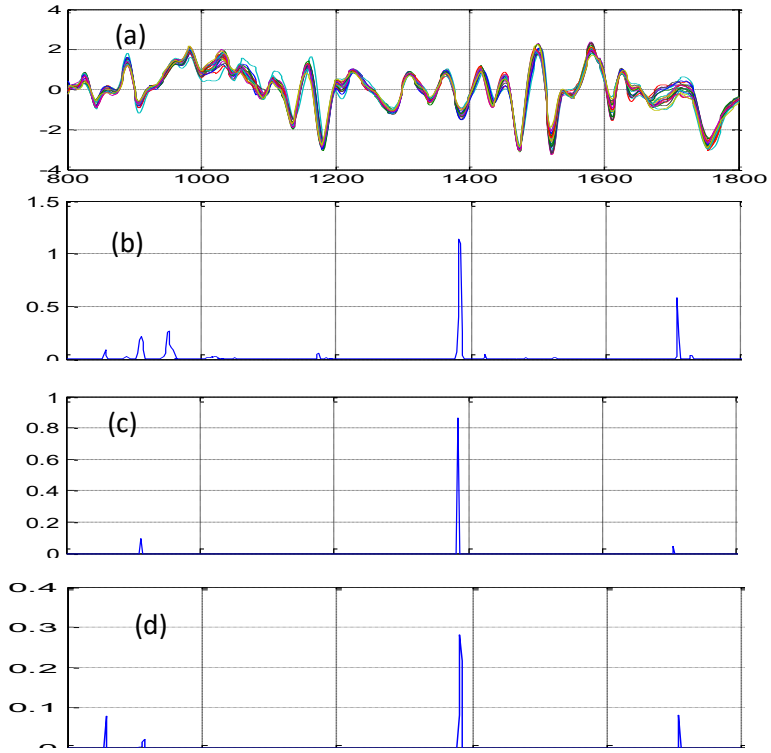


Figure 3.15. (a) : 20 spectres MIR prétraités par la méthode de dérivation Savitzky-Golay (SG) de 1er ordre avec un lissage sur 17 points et un polynôme d'ordre 4, enregistrés sur les quatre échantillons au K = 5 périodes de biodégradation. Les poids w_i identifiés par l'approche d'optimisation proposée avec les contraintes (b) L_1 ; (c) L_∞ ; (d) L_2

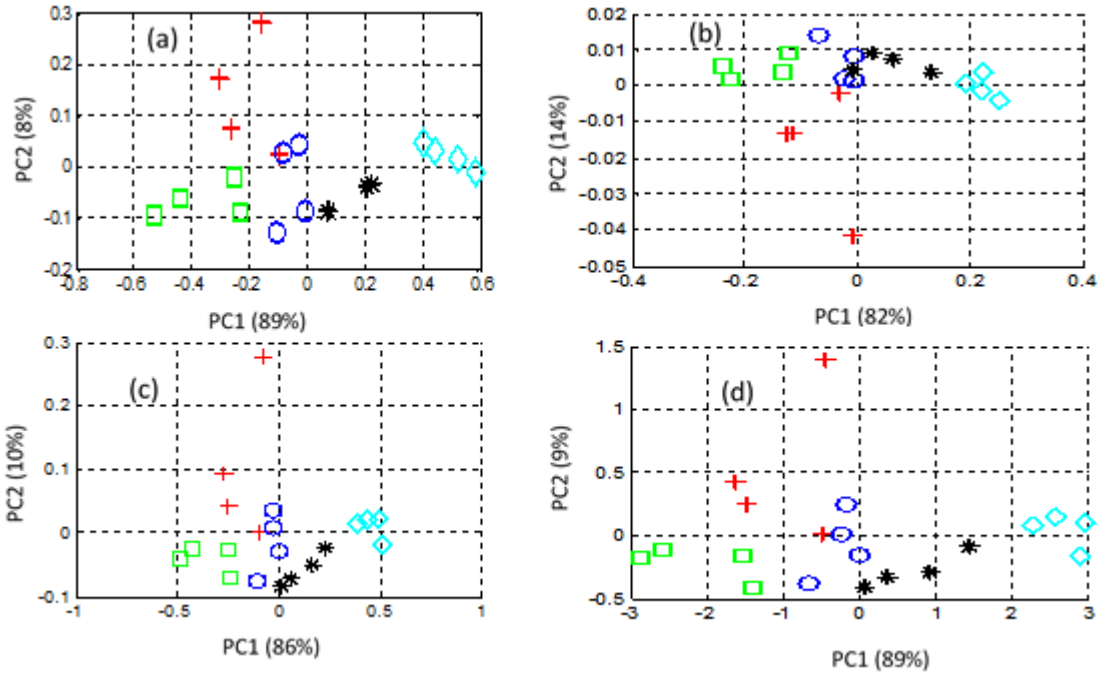


Figure 3.16. Scores plots obtenus sur les nombres d'ondes (MIR) identifiés par : (a) l'algorithme génétique ; (b) l'approche d'optimisation proposée avec (b) L_1; (c) L_∞ ; (d) L_2

D'après les figures des scores plots, les meilleurs résultats obtenus sont également ceux de la méthode d'optimisation correspondant à la contrainte L_1 (Figure 3.16 (a)). Ces résultats sont confirmés par le Tableau 3.8 dans lequel nous avons calculé les valeurs de l'indice Dunn (DI).

Si nous réalisons la comparaison entre l'approche par optimisation et l'algorithme génétique à partir des valeurs de DI obtenues, nous trouvons pour l'algorithme génétique une valeur de DI = 0.173 et pour la méthode d'optimisation DI = 0.19. La méthode d'optimisation avec contrainte L_1 est plus discriminante que l'algorithme génétique. Notre approche d'optimisation a sélectionné les mêmes nombres d'ondes que l'algorithme génétique avec une légère variation sur les valeurs de nombres d'ondes (Tableau 3.10).

Notre approche d'optimisation a donné des résultats meilleurs que ceux de l'algorithme génétique sur les spectres MIR et NIR séparées pour les données de biomasse lignocellulosique. Mais globalement l'application de l'algorithme génétique sur les spectres combinés MIR et NIR par la produit extérieur donne les meilleurs résultats. Comme perspective, nous allons adapter notre approche d'optimisation sur les spectres MIR⊗NIR.

III.8. Conclusion

L'implémentation d'un algorithme génétique (AG) avec l'indice Davies-Bouldin (DB) comme fonction fitness et le choix de paramètres adaptés nous a conduits à proposer une méthodologie pour la sélection des nombres d'ondes discriminants. Nous avons pu identifier de nombres d'ondes dans les deux gammes spectrales MIR et NIR qui peuvent être attribués à l'évolution chimique des échantillons étudiés au cours du processus de biodégradation de la biomasse lignocellulosique.

Nous avons montré que les informations spectrales MIR et NIR sélectionnées aux nombres d'ondes identifiés par l'AG donnent une meilleure discrimination des échantillons pour la cinétique de biodégradation que si on utilisait les informations enregistrées à tous les nombres d'ondes en MIR et NIR. D'autre part, l'analyse conjointe des spectres MIR et NIR fournit de meilleurs résultats que le traitement séparé des spectres des régions MIR ou NIR. La méthodologie proposée permet de sélectionner un nombre très réduit des couples de nombres d'ondes (MIR, NIR) par rapport au nombre très important des couples (dizaines de milles) que le produit extérieur donne.

Nous pouvons conclure que l'application de l'algorithme génétique sur les spectres MIR et NIR combinés par le produit extérieur donne la meilleure discrimination possible des échantillons par rapport à la cinétique de biodégradation, quelle que soit l'échelle de temps. L'explication principale réside dans le fait que le produit extérieur permet de mettre en évidence les interactions entre les vibrations moléculaires fondamentales, les harmoniques et leurs combinaisons.

Nous avons ensuite proposé une nouvelle approche d'optimisation basée sur la minimisation d'un vecteur de pondération afin d'examiner l'impact sur la qualité de la classification et la sélection de nombres d'ondes. Cet algorithme requiert moins des paramètres et permet d'extraire également les poids des nombres d'ondes discriminantes. Nous en avons conclu que les résultats obtenus par cette approche d'optimisation dépendent significativement des choix de contraintes. Les valeurs d'indice Dunn indiquent que la contrainte de normalisation L_1 fournit de bons résultats qui sont meilleurs en comparaison avec les résultats obtenus par l'algorithme génétique AG sur les spectres MIR et NIR séparés. L'inconvénient de l'approche optimisation est de ne pas être applicable sur les combinaisons

MIR⊗NIR à cause d'un problème de dimension trop importante, mais cela ouvre la voie à des travaux futurs, par exemple en utilisant des projections aléatoires pour réduire la dimension.

Nous pouvons conclure que les méthodologies proposées, basées sur l'AG et sur l'optimisation PQS sont complémentaires. La première assure une recherche globale et robuste même sur les spectres de grandes dimensions, comme les spectres combinés par l'OP, l'autre calcule un optimum de façon rapide. Cependant, il pourrait être intéressant d'ajouter à cette étude la prise en compte de connaissances a priori (informations chimiques ou biologiques) afin de pouvoir déterminer un modèle de prédiction du processus de biodégradation de la biomasse lignocellulosique. Ce travail est abordé dans le dernier chapitre de cette thèse.

Chapitre 4 : Modélisation mathématique de la biomasse lignocellulosique s'appuyant sur l'information spectrale et chimique

IV.1. Introduction

Dans les chapitres précédents, nous avons développé des méthodes basées sur des algorithmes de classification, génétiques et d'optimisation en utilisant uniquement les informations spectrales IR. Ces méthodes ont été développées pour de sélectionner les prétraitements optimaux, les gammes et les bandes spectrales discriminantes par rapport au processus de biodégradation de la biomasse lignocellulosique.

Dans ce chapitre, nous allons analyser les échantillons de biomasse lignocellulosique par des méthodes mathématiques basées sur les principes de modélisation et d'optimisation en utilisant les informations spectrales mais également des informations complémentaires : analyses chimiques et biologiques. Pour cela, nous allons nous appuyer sur des méthodes de régression pour modéliser les liens entre les données spectrales et les informations complémentaires et sur les algorithmes génétiques pour sélectionner les bandes spectrales discriminantes. L'objectif est d'améliorer la calibration entre les informations complémentaires mesurées et les informations prédites tout en gardant la complémentarité MIR-NIR. La pierre angulaire de ce chapitre concerne donc l'utilisation des informations complémentaires, telles que les analyses chimiques concernant le taux de minéralisation des biomasses en gaz carbonique et des analyses biologiques (ici des enzymes mesurées dans les sols au cours de la dégradation de la lignocellulose), tout en combinant les informations spectrales dans les deux gammes MIR et NIR mais en sélectionnant les couples de bandes spectrales les plus discriminantes.

Pour modéliser les liens entre les données spectrales et les informations complémentaires, nous introduisons dans la section suivante, des méthodes de modélisation mathématiques telles que les régressions et plus particulièrement la régression linéaire multiple, la régression en composantes principales et la régression des moindres carrés partiels. Nous exposons les avantages et les inconvénients de chaque méthode de régression. Ensuite, nous présentons la possibilité d'associer la méthode de régression des moindres carrés avec les algorithmes génétiques dans le but de sélectionner les bandes spectacles discriminantes qui permettent de bien modéliser les données chimiques et biologiques. Nous exposons une autre méthode de sélection des bandes basée sur la régression des moindres carrés qui correspond à l'importance des variables dans la projection (Variable Importance in Projection ; VIP) et nous comparons ses performances avec la méthode précédente. Finalement, nous présentons une nouvelle méthodologie qui associe l'algorithme génétique avec la méthode des moindres carrés partiels. Il s'agit de combiner les informations spectrales MIR et NIR par le produit extérieur, puis de modéliser les liens entre les informations spectrales et celles chimiques en optimisant la sélection des couples de bandes spectrales discriminantes par l'algorithme génétique. Les résultats de ces méthodes sont discutés dans la dernière section.

IV.2. Modélisation

La modélisation est une méthode utilisée pour extraire ou prévoir des informations en construisant un modèle optimal reliant les données mesurées aux données observées. Les outils utilisés sont très nombreux et assez différents allant des réseaux de neurones à la régression (régression linéaire

multiple, régression de moindres carrées partiels, régression sur composantes principales), en passant par l'analyse discriminante [LL98].
Les méthodes de régression sont des techniques chimiométriques largement utilisées et efficaces pour la prédiction et l'étalonnage (calibration), notamment dans l'agroalimentaire. La régression permet de modéliser, d'examiner et d'explorer les relations entre les variables indépendantes (les spectres IR) et dépendantes (les informations chimiques et biologiques) [CK10]. L'intérêt principal des méthodes de régression réside dans leur capacité à prendre en compte des données de grande dimension et même de très grande dimension lorsque le nombre de variables est largement plus grand que le nombre de spectres.
D'autres méthodes basées sur les réseaux de neurones artificiels ont été développées. Celles-ci sont relativement complexes en termes de mise en œuvre, de configuration, de formation et d'estimation des paramètres par rapport aux méthodes de régression et leur application reste plus limitée. De plus, la calibration d'un réseau de neurones (minimisation de l'erreur) n'est pas toujours directe : le nombre de paramètres est souvent plus grand que le nombre d'observations [BJC01]. Pour ces différentes raisons, nous avons décidé d'utiliser les méthodes de régression pour modéliser les relations entre les données spectrales de biomasse lignocellulosiques et les informations complémentaires (chimiques et biologiques).

IV.3. Méthodes de régression

La régression est une approche qui permet d'établir une relation entre un certain nombre de variables indépendantes (ou prédicteurs, en fait les variables d'entrée) et des variables dépendantes (les variables de sortie). La régression peut se mettre mathématiquement sous la forme :

Soit Y = $\begin{bmatrix} Y_1 \\ \vdots \\ Y_j \\ \vdots \\ Y_J \end{bmatrix} \in M_{J,M}(\mathbb{R})$ avec $Y_j = [Y_{j1} \dots Y_{jt} \dots Y_{jm}] \in \mathbb{R}^{\mathrm{m}}$ un vecteur contenant les variables dépendantes, et

$X^{case} = \begin{bmatrix} x_{11}^{case} & \cdots & x_{1O}^{case} \\ \vdots & \ddots & \vdots \\ x_{J1}^{case} & \cdots & x_{JO}^{case} \end{bmatrix} \in M_{J,O}(\mathbb{R})$ la matrice contenant les variables d'entrée telles que case = MIR, NIR, MIR-NIR ou MIR⊗NIR et O = P, Q, P+Q, PxQ pour tout j = 1…J, avec J = nombre d'échantillons.

L'objectif de la régression est de trouver la relation entre X^{case} et Y.
Dans le cas linéaire, on cherche à relier chaque variable dépendante j (observation Y_j) aux variables indépendantes ou explicatives x_{jl}^{case}, $l \in [1, O]$ par une relation linéaire du type :

$$Y_{jt} = b_{1t}x_{j1}^{case} + b_{2t}x_{j2}^{case} + \dots b_{Ot}x_{jO}^{case} + e_{jt} \Leftrightarrow Y_{jt} = \sum_{l=1}^{O} b_{lt}x_{jl}^{case} + e_{jt} \qquad \text{(eq.4.1)}$$

où e_{jt} est l'erreur du modèle régressif qui explique ou résume l'information manquante lors de l'évaluation de Y_{jt} à partir de $\underline{x_j^{case}}$, et b_{lt} sont les coefficients de régression.

L'hypothèse de linéarité ne peut évidemment pas être confirmée de manière théorique, mais de petites variations autour sont généralement bien tolérées par les modèles. Si les relations liant les variables d'entrées/sorties ne sont pas tout à fait linéaires, le modèle calculé avec l'hypothèse de

linéarité est quand même adéquat. De manière plus générale, en utilisant le produit matriciel, on peut écrire :

$$Y = X^{case}B + E \qquad \text{(eq.4.2)}$$

Cela équivaut à :

$$\begin{bmatrix} Y_{11} & Y_{12} & \dots & Y_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ Y_{j1} & Y_{j2} & \dots & Y_{jm} \\ \vdots & \vdots & \vdots & \vdots \\ Y_{J1} & Y_{J2} & \dots & Y_{Jm} \end{bmatrix} = \begin{bmatrix} x_{11}^{case} & x_{12}^{case} & \dots & x_{1O}^{case} \\ \vdots & \vdots & \vdots & \vdots \\ x_{j1}^{case} & x_{j2}^{case} & \dots & x_{jO}^{case} \\ \vdots & \vdots & \vdots & \vdots \\ x_{J1}^{case} & x_{J2}^{case} & \dots & x_{JO}^{case} \end{bmatrix} * \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ b_{l1} & b_{l2} & \dots & b_{lm} \\ \vdots & \vdots & \vdots & \vdots \\ b_{O1} & b_{O2} & \dots & b_{Om} \end{bmatrix} + \begin{bmatrix} e_{11} & e_{12} & \dots & e_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ e_{j1} & e_{j2} & \dots & e_{jm} \\ \vdots & \vdots & \vdots & \vdots \\ e_{J1} & e_{J2} & \dots & e_{Jm} \end{bmatrix}$$

La régression consiste à trouver les coefficients de la matrice de régression B en réalisant l'approximation (l'estimateur de B) :

$$\hat{B} = arg\,min \|Y - X^{case}B\|_2^2 \qquad \text{(eq.4.3)}$$

Différentes méthodes de modélisation peuvent être utilisées pour calculer $\hat{B}$. C'est l'inversion de X^{case} qui pose généralement problème, car la matrice n'est pas forcément carrée et/ou inversible. Les méthodes les plus utilisées pour résoudre ce problème sont : la régression linéaire multiple, la régression en composantes principales et la régression des moindres carrés partiels.

IV.3.1. Régression linéaire multiple (MLR)

La régression linéaire multiple (Multiple Linear Regression en anglais, MLR) est une généralisation à n variables de la régression linéaire simple. Pour rappel, la régression linéaire simple ne prend en compte qu'une seule variable explicative pour une seule variable expliquée [GK86]. Dans le cas de J échantillons pour O variables indépendantes, nous pouvons distinguer 3 cas :

1. O $<$ J : Il y a plus de variables que d'échantillons. Dans ce cas, il y a une infinité de solutions pour $\widehat{B}$. Ce n'est pas l'objectif de la régression.
2. O = J : Le nombre d'échantillons est égal au nombre de variables. C'est une situation très peu probable. Cependant elle donne une solution unique pour $\hat{B}$, en notant que le rang de la matrice des variables indépendantes doit être maximal.
3. O $>$ J : Il y a plus d'échantillons que de variables dépendantes. Cela ne permet pas d'avoir une solution exacte pour $\hat{B}$. Cette solution peut être obtenue en minimisant l'erreur d'estimation de $\hat{B}$. Généralement, pour minimiser celle-ci, on utilise la méthode des moindres carrés, dont la solution est donnée par :

$$\hat{B} = \left(X^{case^T}X^{case}\right)^{-1}X^{case^T}Y \qquad \text{(eq.4.4)}$$

Cependant, l'inverse de $X^{case^T}X^{case}$ n'existe pas forcement : colinéarité des éléments, singularité de la matrice, déterminant nul [BKW05].

La MLR est la plus simple des régressions qui permet de relier un ensemble de prédicteurs à une variable dépendante. Malgré sa simplicité, la MLR a quelques défauts. Tout d'abord, la détermination du modèle est très instable lorsque le nombre d'observations est inférieur au nombre de variables :

c'est une situation pourtant rencontrée fréquemment dans notre application IR de biomasse lignocellulosique. Ensuite, on se heurte à l'impossibilité de prendre en compte les données manquantes, et cela oblige souvent le praticien à rejeter certaines observations qui peuvent néanmoins se révéler utiles pour le modèle, des informations importantes peuvent être contenues dans les variables disponibles. Enfin, la MLR présente une grande sensibilité aux données colinéaires au sein des variables explicatives.

Pour pallier ce problème lié à l'application de la MLR, on fait appel à d'autres méthodes alternatives. Parmi celles-ci, la régression en composantes principales et la régression au sens des moindres carrées. Les deux techniques sont décrites dans la suite de ce chapitre.

IV.3.2. Régression en composantes principales (RCP)

Pour s'affranchir du problème de la colinéarité des variables prédictives, il est possible d'associer l'analyse en composantes principales (ACP) avec la régression. En effet, en effectuant une ACP avant la régression, on projette les variables initiales dans un nouvel espace où elles sont orthogonales entre elles. Ne sont retenues que les composantes apportant le plus de variance. Ce nombre de composantes, encore appelé nombre de facteurs, détermine la complexité du modèle. Cette méthode appelée régression en composantes principales (Principal component Regression, RCP) est très efficace, mais elle ne tient pas compte, dans la projection, des variables dépendantes [GK86]. Il peut aussi arriver que certaines variations non corrélées avec les variables de sortie soient retenues dans les composantes principales, et inversement, que de l'information pertinente soit rejetée. Les bases de projection ne seront alors pas optimales. On lui préfère donc une méthode un peu plus avancée qui est la régression des moindres carrés partiels.

IV.3.3. Régression des moindres carrées (PLS)

La régression PLS (Partial Least Squares regression en anglais) est une méthode d'analyse des données proposée par Wold [Wol66, WM83]. La régression PLS est une méthode de détermination d'un modèle prédictif permettant de passer outre les limitations rencontrées précédemment. En fait, on peut voir la régression PLS comme une généralisation de la régression linéaire multiple et de la régression en composantes principales. On répond au problème des données manquantes, ainsi qu'à la colinéarité éventuelle des prédicteurs, et le fait d'avoir beaucoup plus de prédicteurs que d'observations n'est plus un problème. Cette méthode connaît un très grand succès en chimiométrie, particulièrement dans les applications concernant des données de spectrométrie. La PLS est devenue un outil standard pour la modélisation des relations linéaires entre des mesures multi variées [Ten98]. Elle est aussi implantée dans de nombreux logiciels commerciaux sous une forme conviviale.

IV.3.3.1. Objectif et l'approche statistique

Les difficultés liées à la grande dimension des matrices peuvent se résoudre par des techniques de réduction de dimension, comme la régression PLS [HTF09, Hos88, Ten98]. Le but est de trouver une transformation linéaire de X qui compresse les données, i.e. qui réduise la dimension ou de manière équivalente le nombre de variables.

On cherche une matrice de poids $W = [w_1 \dots w_K] \in M_{O,K}(\mathbb{R})$ avec K le nombre de composantes recherchées (son choix sera explicité), et on définit de nouvelles variables, appelées composantes, variables latentes (VL) ou scores, notées $(T_1 \dots T_K)$, orthogonales entre elles telles que $T_k = X^{case} w_k$

pour tout $j = 1 \dots J$. Les observations de ces nouvelles variables sont ensuite regroupées dans la matrice $T = [t_1 \dots t_K] \in M_{J,K}(\mathbb{R})$ définie par :

$$T = X^{case} W, \qquad \text{(eq.4.5)}$$

En particulier, chaque colonne de T est définie par $t_k = X^{case} w_k$ pour $k = 1 \dots K$. Cette transformation peut être vue comme la recherche d'un sous-espace (celui engendré par les colonnes de T) sur lequel on projette X^{case}, d'où la réduction de dimension.
Cette projection dépendra de la manière dont on définit les poids W. La finalité étant de construire des nouvelles variables qui résument l'information contenue dans X^{case} et qui peuvent expliquer Y, via un modèle linéaire pour la prédiction ou la caractérisation des réponses. Dans la pratique, le calcul des poids W se fera par la résolution d'un problème d'optimisation, i.e. le calcul d'un optimum (maximum ou minimum suivant la technique considérée) d'une fonction objectif.
On pourra alors s'intéresser au modèle linéaire à sa version matricielle :

$$Y = TQ + E \qquad \text{(eq.4.6)}$$

avec $Q \in M_{K,m}(\mathbb{R})$ la matrice des coefficients de régression (poids factoriels) pour T et avec E le terme d'erreur du modèle de même dimension que Y. Ces poids sont calculés de telle façon que chacun d'entre eux maximise la covariance entre les réponses et les composantes [GK86, Ten98]. La régression de Y sur T devient possible par la réduction de dimension et par l'orthogonalité des composantes (absence de multi-colinéarité). On remonte au modèle initial par la relation :

$$B = WQ, \qquad \text{(eq.4.7)}$$

obtenu en injectant le produit $T = X^{case} W$ dans ce nouveau modèle.

La régression PLS cherche des poids W qui maximisent la covariance entre Y et les nouveaux axes T, construisant ainsi des nouvelles variables résumant l'information contenue dans X^{case} et expliquant au mieux Y [WSE01]. Une matrice supplémentaire des poids factoriels P donnant le modèle factoriel X^{case}=TP+F, où F est la partie inexpliquée des composantes de X^{case}, est alors nécessaire pour compléter la description de la régression PLS.

Dans tout notre exposé, la matrice de prédicteurs Y sera supposée être les informations complémentaires (chimiques et biologiques) centrées et réduites (normalisées). La matrice d'observations (ou réponse) X^{case} sera formée des spectres de biomasse lignocellulosique prétraités par la méthode Savitzky-Golay (SG) d'ordre 1 suivi de la normalisation SNV. Ce prétraitement permet de normaliser les données spectrales.

IV.3.3.2. Principe de la PLS

L'idée de la régression PLS [HTF09, Hos88, Ten98] est de modifier l'analyse en composantes principales (ACP) pour obtenir une technique qui tiendrait compte de l'information contenue dans la réponse Y et pas seulement dans les prédicteurs X^{case}. Le problème d'optimisation se retrouve modifié ainsi que sa résolution (théorique et numérique). La régression PLS induit également la nécessité d'une ré-estimation des variables à chaque itération qu'on appellera « déflation ».

a) Modèle PLS

On cherche une transformation linéaire des prédicteurs X^{case} ainsi que des réponses Y, sous les formes respectives $T_k = X^{case}w_k$ et $C_k = Yu_k$ pour $k = 1 \dots K$. Soit matriciellement, $T = X^{case}W$ et $C = YU$ avec les notations :

- K nombre de composantes ou dimension de la réduction ;
- $W = [w_1 \dots w_K] \in M_{O,K}(\mathbb{R})$ poids des X^{case} ; avec $w_k = [w_{1k} \dots w_{lk} \dots w_{Ok}] \in \mathbb{R}^O$ est la vecteur de poids pour $k = 1 \dots K$ et $l = 1 \dots O$
- T = $(T_1 \dots T_K)$ composantes PLS ou scores (parfois notés X-scores), dont les observations sont regroupées dans la matrice $T = [t_1 \dots t_K] \in M_{J,K}(\mathbb{R})$; avec $t_k = [t_{1k} \dots t_{jk} \dots t_{Jk}] \in \mathbb{R}^J$ est la vecteur de kième variable latente.
- $U_{OxK} = [u_1 \dots u_K]$ poids des Y ;
- C = $(C_1 \dots C_K)$ Y-scores, dont les observations sont regroupées dans la matrice $C = [c_1 \dots c_K] \in M_{J,K}(\mathbb{R})$.

Cette transformation devra maximiser la covariance entre les nouvelles variables créées T et C, ainsi on cherchera des poids w et u pour X et Y respectivement qui maximisent :

$$cov(T, C) = cov(X^{case}w, Yu) = \frac{1}{J-1} w^t X^{case^t} Yu \qquad \text{(eq.4.8)}$$

On a alors le problème d'optimisation suivant :

$$arg \max_{w,u} \left(w^t X^{case^t} Yu \right), \qquad \text{(eq.4.9)}$$

sous les contraintes $w^t w = u^t u = 1$ et $w^t X^{case^t} X^{case} w_l = 0$ pour $l = 1 \dots k-1$ (orthogonalité entre composantes), lors du calcul de la kième composante T_k.

b) Mise en application

La méthode mise en place est itérative. Chaque étape se décompose en deux phases :

- construction d'une nouvelle composante T_k, par le calcul de w_k (ainsi que de U_k et c_k);
- déflation : on retire de X et Y la partie qui a été déjà expliquée en les remplaçant par les résidus respectifs de leur régression sur la composante T_k nouvellement créée.

Ainsi, à chaque étape k, on résout le problème d'optimisation non pas directement sur X et Y mais sur leurs versions « déflatées » notées respectivement X^{k-1} et Y^{k-1}, construites récursivement par la régression linéaire suivante :

$X^{k-1} = T_k \tilde{p}_k^t + X^k$

$Y^{k-1} = T_k \tilde{q}_k^t + Y^k$

où les coefficients de régression des prédicteurs X^{k-1} et réponses Y^{k-1} sur la nouvelle composante T_k sont notés respectivement $\tilde{p}_k^t$ et $\tilde{q}_k^t$, de sorte que l'étape de déflation s'écrira matriciellement :

$X^k = X^{k-1} - t_k \tilde{p}_k^t$

$Y^k = Y^{k-1} - t_k \tilde{q}_k^t$

On aura [Hos88] :

$\tilde{p}_k^t = (t_k^t, t_k)^{-1} t_k^t X^{k-1}$ i.e. $\tilde{p}_k = \frac{(X^{k-1})^t t_k}{t_k^t t_k}$ (vecteur colonne de dimension O) ;

$\tilde{q}_k^t = (t_k^t, t_k)^{-1} t_k^t Y^{k-1}$ i.e. $\tilde{q}_k = \frac{(Y^{k-1})^t t_k}{t_k^t t_k}$ (vecteur colonne de dimension m) avec $t_k^t t_k \epsilon R$.

Comme introduit précédemment, on notera $\tilde{P} \in M_{K,O}(\mathbb{R})$ et $\tilde{Q} \in M_{K,m}(\mathbb{R})$ les matrices regroupant respectivement les $\tilde{p}_k$ et $\tilde{q}_k$ avec $\tilde{P} = [\tilde{p}_1, \dots, \tilde{p}_K]^t$ et $\tilde{Q} = [\tilde{q}_1, \dots, \tilde{q}_K]^t$.

On définit également P $\in M_{K,O}(\mathbb{R})$ et Q $\in M_{K,m}(\mathbb{R})$ les matrices des coefficients de régression respectifs de X^{case} et Y sur les composantes T, i.e. à partir des modèles respectifs : $X^{case} = TP + F$ et $Y = TQ + E$. On aura par régression linéaire : $P = (T^t T)^{-1} T^t X^{case}$ et $Q = (T^t T)^{-1} T^t Y$.

c) Algorithmes PLS

Le schéma de l'algorithme PLS est donné par :

Initialisation : $X^0 = X^{case}$; $Y^0 = Y$

Itération : Pour $k = 1,2, \dots, K$

1) $c_k =$ première colonne de Y^{k-1}(calcul des Y -scores)
2) Répéter jusqu'à convergence de w_k :
   i. $w_k = \frac{(X^{k-1})^t c_k}{c_k^t c_k}$, normer w_k à 1.
   ii. $t_k = \frac{X^{k-1} w_k}{w_k^t w_k}$ (calcul des composantes PLS)
   iii. $u_k = \frac{(Y^{k-1})^t t_k}{t_k^t t_k}$, normer u_k à 1
   iv. $c_k = \frac{Y^{k-1} u_k}{u_k^t u_k}$
3) $p_k = \frac{(X^{k-1})^t t_k}{t_k^t t_k}$, (coefficients de régression de X^{k-1} sur t_k)
   $X^k = X^{k-1} - t_k p_k^t$ (déflation de X^{case})
4) $u_k = \frac{(Y^{k-1})^t t_k}{t_k^t t_k}$ (coefficients de régression de Y^{k-1} sur t_k)
   $Y^k = Y^{k-1} - t_k q_k^t$ (déflation de Y)

IV.3.4. Analyse statistique de modèles prédictifs

L'objectif de la régression est d'estimer ou d'exprimer des variables de réponses avec un ensemble de variables indépendantes. En d'autres termes, c'est une méthode avec laquelle on fait de la prédiction. Quel que soit le type d'approche choisi pour effectuer de la régression, il faut évaluer les performances et la viabilité du modèle. De cette manière, nous pouvons l'apprécier à sa juste valeur et le comparer à d'autres modèles ou d'autres approches. Nous allons présenter succinctement les indices généralement utilisés pour mesurer la performance d'un modèle. Dans le protocole expérimental nous reviendrons plus en détail sur le formalisme de ces différents indices.

La régression est un processus en deux étapes : la première consiste à construire le modèle, c'est-à-dire déterminer le vecteur de régression. C'est l'étape d'étalonnage. La deuxième consiste à vérifier que notre modèle réagit bien face à de nouvelles données. C'est l'étape dite de test. Pour cela, nous avons utilisé la méthode de validation Leave-One-Out qui est un cas particulier de la validation croisée. Cette méthode est une alternative très populaire pour estimer correctement l'erreur de prédiction du modèle PLS. C'est aussi le moyen le plus utilisé pour déterminer le nombre de composantes à inclure dans la régression PLS.

Pour minimiser l'influence du choix du partitionnement de l'ensemble des données, la validation croisée subdivise l'ensemble d'entrainement initial en k sous-ensembles disjoints $D_1, \dots, D_k$ de même taille. L'entrainement et le test sont effectués k fois. A l'itération i le sous-ensemble D_i est réservé pour

le test et le reste des données est utilisé pour entrainer le modèle. La précision finale du modèle est égale à la moyenne des k précisions du test. La méthode Leave-One-Out est un cas particulier de la validation croisée où k = N. Á chaque itération, le modèle est entrainé sur N − 1 exemples et testé sur l'exemple exclu de l'entrainement. On obtient à la fin N précisions, la précision du modèle est égale à leur moyenne.

En premier lieu, la performance du modèle peut s'évaluer en calculant l'erreur commise. Celle-ci pour chaque prédiction effectuée peut être évaluée, c'est ce qu'on appelle les résidus. Pour évaluer l'erreur sur l'ensemble des prédictions, on calcule l'erreur quadratique moyenne (Mean Squared Error en anglais, MSE) : c'est la moyenne des résidus élevée au carré. Enfin, par analogie à l'analyse de variance, on prend la racine carrée de cette erreur (Root Mean Squared Error en anglais, RMSE). Cette erreur peut être calculée lors de l'étalonnage, on l'appelle Root Mean Squared Error of Prediction (RMSEP), ou lors du test, Root Mean Squared Error of Validation (RMSECV).

Dans notre application, nous avons évalué la valeur de RMSECV, celle-ci est largement utilisée comme critère pour juger de la performance d'un modèle de calibrage multivarié dans de nombreux articles [TWP14, VLT14, HS03]. La RMSECV est une mesure fréquemment utilisée pour les différences entre les valeurs prédites par un modèle ou un estimateur et les valeurs réellement observées à partir des objets estimés. Ainsi, pour éviter des conclusions injustifiées sur les performances d'un modèle de calibrage multivarié, la RMSECV a été calculée en utilisant l'équation suivante :

$$RMSECV = \sqrt{\frac{\sum_{j=1}^{J}(\hat{Y}_j - Y_j)^2}{J}} \qquad \text{j= 1 … J} \qquad \text{(eq.4.10)}$$

avec $Y_j \in \mathbb{R}^m$ les concentrations mesurées, J le nombre de variables indépendantes et $\hat{Y}_j \in \mathbb{R}^m$ les variables dépendantes prédites par la PLS en utilisant donc l'équation suivante :

$$\hat{Y}_j = q_0 + q_1 t_{1j} + \dots + q_K t_{Kj} \qquad \text{j = 1 … J} \qquad \text{(eq.4.11)}$$

avec t_{1j} à t_{Kj} les variables latentes (les scores des composantes principales (PC) 1 à K pour la variable dépendante prédite ($\hat{Y}_j$) et q_0 à q_K sont les coefficients de régression PLS. Pour choisir le nombre optimal K de variables latentes, nous avons utilisé le critère de détection du premier minimum de RMSECV (i.e. nous avons appliqué la PLS sur tous les variables et nous avons choisi ensuite le premier minimum de RMSECV qui dépend de K) comme il est courant de le rencontrer dans de nombreux articles [TWP14, VLT14, HS03].

Un autre critère pour juger de la qualité du modèle est le calcul du coefficient de détermination communément noté R^2 qui « mesure la proportion de variation totale sur la moyenne des échantillons expliquée par la régression » [DS81]. Le R^2 peut être vu comme la corrélation qui existe entre les variables prédites et les variables effectivement mesurées. C'est une valeur comprise entre 0 (peu de corrélation) et 1 (corrélation forte).

$$R^2 = 1 - \frac{\sum_{j=1}^{J}(\hat{Y}_j - Y_j)^2}{\sum_{j=1}^{J}(Y_j - \bar{Y})^2} \qquad \text{(eq.4.12)}$$

où $\bar{Y}$ est la valeur moyenne des variables dépendantes Y.

$$\bar{Y} = \frac{\sum_{j=1}^{J} Y_j}{J} \quad \text{(eq.4.13)}$$

Un bon modèle prédictif est un modèle ayant une erreur d'étalonnage et de prédiction peu élevées et proches l'une de l'autre, ainsi qu'un coefficient de détermination voisin de 1 [TYL07]. Ce critère permet de mesurer la qualité d'ajustement entre les valeurs réellement observées et les valeurs prédites par un modèle d'étalonnage. Le coefficient de détermination (R^2) et l'erreur RMSECV ont été utilisés pour évaluer la qualité des modèles de PLS. Le coefficient RMSECV a été choisi pour évaluer l'erreur du modèle relatif et nous permet de comparer les performances de ce même modèle sur des données différentes (ex spectres MIR et NIR).

D'autre part, la précision d'un modèle de régression PLS pour prédire correctement les valeurs de variables dépendantes est souvent évaluée en utilisant le pourcentage d'erreur relative de prédiction (relative error percent of prediction ; REP). Il a été utilisé comme une mesure permettant de comparer les performances entre les méthodes de régressions PLS. La REP (%) est calculée selon l'équation suivante :

$$REP(\%) = \frac{100}{\bar{Y}} * [\frac{1}{J}(\hat{Y}_j - Y_j)^2]^{0.5} \quad \text{(eq.4.14)}$$

Dans notre application, nous allons utiliser ces trois paramètres statistiques RMSECV, R^2 et $REP(\%)$ qui sont les plus couramment utilisés pour mesurer la performance de modèle PLS. Un modèle sera considéré comme supérieur à un autre si ses RMSECV et REP(%) sont plus petits et que son R^2 est plus grand.

IV.4. Méthodes de sélection de variables

Dans les chapitres précédents, nous avons appliqué des méthodes de sélection de variables telle que l'algorithme génétique sur des échantillons de biomasse lignocellulosique en utilisant uniquement les informations spectrales IR, afin de sélectionner les bandes spectrales discriminantes par rapport au processus de biodégradation. Dans cette partie, nous allons analyser les échantillons de biomasse lignocellulosique en utilisant les informations spectrales IR et les informations complémentaires (chimiques et biologiques). Pour cela, nous allons associer des méthodes de sélection de variables avec la méthode de régression des moindres carrées afin de sélectionner les bandes spectrales discriminantes par rapport au processus de biodégradation afin d'améliorer la calibration entres les informations complémentaires mesurées et les informations complémentaires prédites.

En étalonnage multivarié, la sélection de variables permet d'identifier et d'éliminer les variables qui pénalisent les performances d'un modèle dans la mesure où elles peuvent être bruitées, redondantes ou corrélées [MVB97]. Les procédures de sélection de variables présentent un intérêt particulier en ce qui concerne les données spectroscopiques. En effet, le nombre de variables est en général très important vis-à-vis du nombre d'échantillons présents dans la matrice X^{case} (O >> J). Habituellement, ce problème de dimension est géré par l'utilisation de méthodes factorielles de régression telle que la PLS [BSS96]. La sélection de variables a prouvé être une solution satisfaisante pour simplifier la complexité du modèle et pour améliorer les capacités prédictives des variables dépendantes.

De nombreuses méthodes de sélection existent pour lesquelles le choix d'inclure ou non des variables est effectué en minimisant les erreurs de prédiction des modèles de régression construits. Parmi ces méthodes, on peut noter par exemple la régression PLS par élimination des variables non informatives

(UVE-PLS) [FLC04], la méthode Backward (descendante) avec la régression PLS (BVE-PLS), la régression PLS par intervalle (iPLS) [NSW00], la régression PLS par sélection interactive de variables (IVS-PLS) [AJS03], l'importance des variables dans la projection (VIP) [Hoo11], la sélection de variables par la méthode des algorithmes génétiques (AG-PLS) [CLS08].

Les méthodes de sélection UVE-PLS, BVE-PLS et IVS-PLS nécessitent soit de longs temps de calcul et l'optimisation des nombreux paramètres ou bien, comme dans le cas de la méthode iPLS, une sélection des bandes qui dépend fortement de la taille de l'intervalle défini. De plus, ces méthodes présentent souvent des problèmes au niveau de l'étalonnage ou de la validation des données de grande dimension. Celle-ci est fortement influencée par la magnitude des coefficients de régression des variables et par la sélection d'un grand nombre de variables non pertinentes ou redondantes [GA07].

Pour cela, nous proposons de traiter ce problème par l'utilisation de méthodes génétiques AG avec la méthode de régression PLS qui ont déjà été utilisées avec succès dans plusieurs applications telles que l'amélioration de calibration, la prédiction d'informations complémentaires et la réduction de la dimension dans l'analyse des données hyper-spectrales.

Dans la suite de ce chapitre, nous allons nous intéresser au principe et aux différentes étapes d'un algorithme génétique associé à la régression (AG-PLS). Puis, nous allons proposer les choix pertinents de ces étapes et les paramètres pour les données spectrales IR afin de sélectionner des nombres d'ondes discriminantes dans le processus de biodégradation.

IV.4.1. Combinaison de l'algorithme génétique avec la régression PLS (AG-PLS)

a) Généralités

D'un point de vue plus conceptuel, une procédure de sélection de variables inclut en premier le choix d'un algorithme pour réaliser l'optimisation. Concernant le choix de l'algorithme d'optimisation, des algorithmes stochastiques sont en général utilisés lorsque l'on travaille sur des données de grande dimension en particulier des données spectroscopiques. On s'intéressera notamment aux algorithmes génétiques (AG) qui sont très utilisés pour la sélection de variables en étalonnage multivarié tel que la régression PLS. Ce sont des outils d'optimisation qui réalisent une recherche aléatoire et globale dans un espace de grande dimension. La méthode AG-PLS est une approche hybride qui combine l'AG comme méthode d'optimisation avec la méthode PLS comme méthode statistique robuste pour sélectionner les variables qui contribuent de manière significative à la prédiction. L'AG-PLS présente une supériorité sur les autres méthodes multivariées en raison de sa capacité à sélectionner les nombres d'ondes dans le calibrage PLS en utilisant un algorithme génétique sans perte de capacité de prédiction, tout en fournissant des informations utiles sur le système chimique [XJP10].

L'idée de l'algorithme AG-PLS est de construire aléatoirement une population de N solutions. Chaque solution de la population étant représentée sous la forme d'un « chromosome ». Chaque chromosome est lui-même formé d'un nombre restreint, noté par la suite « L », de nombres d'ondes sélectionnés et positionnés comme des « gènes » dans le chromosome. Cette population peut être représentée comme une matrice de taille N x L.

L'étape suivante correspond à l'évaluation des chromosomes dans cette population à travers une fonction fitness, c'est-à-dire que nous appliquons la méthode de régression PLS pour chaque chromosome, nous allons alors obtenir N modèles de PLS qui sont évalués par une fonction fitness précise. Ensuite, l'algorithme AG-PLS conserve les chromosomes ayant les meilleures valeurs de fitness

pour la génération suivante. L'algorithme permet alors de combiner les meilleurs chromosomes dans l'étape de croisement, puis de faire subir des mutations aux chromosomes restants [YYS14]. L'algorithme AG-PLS se décompose donc en différentes étapes : l'initialisation des paramètres et de la population, puis les évaluations, les sélections, les recombinaisons, les mutations jusqu'à convergence. Les choix de ces étapes ont été détaillés dans le chapitre 3.

En spectroscopie IR, l'AG-PLS est le plus souvent appliqué pour la sélection d'observations ou de variables dans la matrice d'échantillons X^{case} en optimisant un critère qui correspond en général à l'erreur quadratique moyenne de validation croisée (RMSECV). La sélection de variables par AG-PLS a été appliquée en spectroscopie proche infrarouge pour déterminer, par exemple, la quantité d'octane dans les échantillons de gasoil ou le taux d'humidité dans les céréales, ou bien la teneur en coton dans des mélanges de fibres coton/polyester et coton/viscose [AJS03, DDR07].

b) Approche proposée pour des données spectrales IR

Nous rappelons que les données spectrales IR peuvent être conditionnées sous forme matricielle X^{case}, où $X^{case} = \left[\underline{x}_1^{case} \ \dots \ \underline{x}_j^{case} \ \dots \ \underline{x}_J^{case}\right] \in M_{O,J}(\mathbb{R})$ est la matrice formée de J spectres infrarouges avec case = MIR, NIR MIR-NIR et MIR⊗NIR et $\underline{x}_j^{case} = [\underline{x}_{j1}^{case}, \dots \underline{x}_{jO}^{case}]^{\mathrm{T}} \in \mathrm{R}^{\mathrm{O}}$ le j[ième] spectre avec O = P, Q, P+Q ou PxQ. Chaque spectre est enregistré sur le vecteur de nombre d'ondes $\underline{y}^{case} = [y_1^{case} \ \dots \ y_O^{case}]^T \in R^{\mathrm{O}}$.

De plus, Y = $\begin{bmatrix} Y_1 \\ \vdots \\ Y_j \\ \vdots \\ Y_J \end{bmatrix} \in M_{J,M}(\mathbb{R})$ est la matrice formée de J enzymes chimiques et biologiques avec $Y_j = [Y_{j1} \dots Y_{jt} \dots Y_{jm}] \in \mathbb{R}^{\mathrm{m}}$ un vecteur contenant les valeurs de cet enzyme.

Nous proposons d'utiliser la version suivante de l'algorithme AG-PLS dont le principe est le suivant :

1. Initialisation de la population :
   Les paramètres initialisés sont : la taille du chromosome L, la taille de la population N. Les chromosomes sont générés de manière aléatoire pour former une population initiale : $P(0) = \{z_n = [z_{n1} \dots z_{nl} \dots z_{nL}] \in \mathbb{R}^L\}_{n=1}^N$ où chaque z_{il} est un nombre d'onde choisi de façon aléatoire dans le vecteur de nombre d'ondes $\underline{y}^{case}$.

2. Évaluation : Nous avons appliqué la modèle PLS qui a été présenté dans la section 3.3 (dans ce chapitre) pour chaque chromosome dans la population $\underline{z}_n$. Ensuite, ce chromosome $\underline{z}_n$ est évalué par une fonction fitness F(.), qui assigne une valeur F_n : $F_n = F(\underline{z}_n) \ \forall \ n = 1 \dots N$. Comme nous essayons de mesurer la prédiction et la calibration du modèle PLS, nous pouvons utiliser différentes fonctions fitness en s'appuyant sur des paramètres statistiques d'étalonnage de la PLS tels que le RMSEP, REP, R^2 et RMSECV. Pour notre application, nous utilisons l'erreur quadratique moyenne de validation croisée (RMSECV) de la régression PLS (équation 4.10) comme une fonction fitness pour chaque chromosome puisqu'elle permet d'indiquer le pouvoir de prédiction et valider le modèle de calibration PLS [Hoo11, MLS12]. Le RMSECV est évalué pour chaque chromosome de la population, avec élimination des chromosomes ayant des valeurs de RMSECV élevées.

$$F_n = F(\underline{z}_n) = min_{VL \le L}\{RMSECV(\underline{z}_n)\} \ \forall \ n = 1 \dots N \qquad \text{(eq.4.15)}$$

avec

$$RMSECV(\underline{z}_n) = \sqrt{\frac{\sum_{j=1}^{J}(\hat{Y}_j(\underline{z}_n) - Y_j(\underline{z}_n))^2}{J}} \quad j= 1 \dots J \qquad \text{(eq.4.16)}$$

avec $Y_j(\underline{z}_n) \in \mathbb{R}^L$ les concentrations mesurées et $\hat{Y}_j(\underline{z}_n) \in \mathbb{R}^L$ les concentrations prédites par la méthode PLS en utilisant donc l'équation (4.11). Plus les valeurs obtenues de $F(\underline{z}_n)$ sont petites, plus le chromosome $\underline{z}_n$ aura une bonne modélisation (PLS) et des chances d'être sélectionné soit comme chromosome garanti de survivre à la génération suivante ou bien comme chromosome parent.

3. Création d'une population future en utilisant les opérateurs de sélections, croisement et mutations. Nous avons choisi la sélection de type "stochastic universal sampling" car cette méthode a comme avantage de ne pas avoir de biais d'estimation, et une dispersion minimale, la méthode de croisement uniforme et l'opérateur de mutation Gaussien. Nous notons que ces opérateurs de croisement et mutation sont les mêmes qui ont été utilisés dans le chapitre 3.

4. Vérification des conditions d'arrêt de l'algorithme. Si la nouvelle population ne satisfait pas ces conditions, les étapes 2 à 4 sont répétées afin de générer une nouvelle population.

Pour choisir les paramètres optimaux L et N pour notre application, il n'existe pas de méthode clairement définie. Comme dans le chapitre 3 (section 3.3), nous itérons pour différentes valeurs de tailles des chromosomes (L= L_{min} jusqu'à L_{max}) et nous faisons de même pour la taille de population N, (N = N_{min} jusqu'à N_{max}). On choisit après L et N optimum à partir des meilleures valeurs de fitness obtenues [UGS12, Jef04]. Dans notre application, pour chaque valeur de L et N, nous calculons les valeurs fitness RMSECV pour chaque chromosome dans la population. Nous choisissons ensuite les valeurs de L et N à partir des valeurs minimales de RMSECV. Nous notons que la robustesse de l'algorithme génétique par rapport aux paramètres des initialisations aléatoires a été testée dans le chapitre 3 (section 6.1).

$$L = min_{L_{min} \le L \le L_{max}}\{min_{N_{min} \le N \le N_{max}}\{RMSECV(\underline{z}_n)\}\};\ n = 1 \dots N \qquad \text{(eq.4.17)}$$

$$N = min_{N_{min} \le N \le N_{max}}\{min_{L_{min} \le L \le L_{max}}\{RMSECV(\underline{z}_n)\}\};\ n = 1 \dots N \qquad \text{(eq.4.18)}$$

D'autres méthodes de sélections peuvent être appliquées sur les données spectroscopiques IR. Une de ces méthodes les plus populaires de sélection de variables est la projection des variables d'importances dans la projection (VIP) [FPT15]. Elle est un indicateur du pouvoir de modélisation d'une covariable prédictive. La méthode VIP a été largement utilisée dans différents domaines et donc pour une grande variété de types de données [EHJ95, YZL06, EGJ04, FMM15]. Cette méthode a été utilisée avec succès pour sélectionner les bandes spectrales des spectres MIR et NIR dans des échantillons de biomasse végétale [CCC08]. De plus, l'algorithme VIP peut parfois conduire à des modèles plus robustes pour la prédiction de propriétés biologiques du sol (meilleurs R^2, besoin de moins de composantes PLS pour construire les modèles). Le VIP a montré sa capacité à prédire la concentration d'huile dans le tournesol et le soja [Ash13, OEP12]. Nous présentons brièvement cette méthode que nous allons utiliser pour analyser les performances de l'algorithme génétique AG-PLS.

IV.4.2. Méthode des projections des variables d'importances

La projection des variables d'importances (VIP) a été publiée par Wold en 1993. C'est une méthode de sélection de variables basée sur la régression PLS [Wol95, WJC93]. Elle consiste à faire une projection des variables dépendantes sur les variables latentes. La méthode VIP sélectionne les variables en calculant le score VIP pour chaque variable. Le score VIP de la variable l est calculé comme suit [MMS11] :

$$VIP = \sqrt{\frac{\sum_{k=1}^{K} R^2(Y_j,t_k)(\frac{w_{kL}}{\|w_k\|})^2}{(\frac{1}{p})\sum_{k=1}^{K} R^2(Y_j,t_k)}} \qquad \text{(eq.4.19)}$$

où w_{kl} est la $l$ème élément de la vecteur de poids w_k dans l'algorithme PLS et $R^2(Y_j, t_k)$ est la fraction de la variance de la variable dépendante Y_j expliquée par la composante k.

La VIP permet de hiérarchiser les variables selon leur pouvoir explicatif sur les variables indépendantes. La variable qui a la plus grande valeur de score VIP montre qu'elle est plus pertinente pour prédire la variable dépendante.

Dans la suite du chapitre, nous nous intéressons à exposer une nouvelle méthodologie dite « OP-AG-PLS » qui permet de combiner les informations spectrales MIR et NIR par le produit extérieur et de sélectionner les bandes discriminantes par l'algorithme AG-PLS.

IV.4.3. Méthodologie proposée (OP-AG-PLS)

Dans le chapitre précédent, nous avons vu qu'en combinant les informations spectrales MIR et NIR par la produit extérieur (OP), nous pouvons sélectionner des couples de nombres d'ondes (MIR, NIR) qui permettent d'améliorer la classification de la biomasse lignocellulosique par rapport à la cinétique de biodégradation. Dans cette partie, nous partons du même principe : nous cherchons à ce que l'algorithme génétique associé à la régression des moindres carrés partiels (AG-PLS) appliquée sur les spectres MIR et NIR combinés par le produit extérieur (OP) identifie des couples de nombres d'ondes (MIR, NIR). Le but est d'améliorer les calibrations entre les valeurs des variables biologiques et chimiques mesurées et les valeurs prédites par AG-PLS. La figure 4.1 montre les étapes de la méthode OP-AG-PLS proposée.

L'application de l'algorithme AG-PLS sur les spectres MIR⊗NIR a pour but de sélectionner des couples de nombres d'ondes (MIR, NIR) qui permettent une meilleure calibration par la PLS du point de vue de l'erreur RMSECV. Premièrement, nous combinons les J spectres de MIR ∈ $M_{J,P}(\mathbb{R})$ avec J spectres de NIR ∈ $M_{J,Q}(\mathbb{R})$ par le produit extérieur (OP). Nous obtenons alors une matrice de J spectres de MIR⊗NIR ∈ $M_{J,PQ}(\mathbb{R})$ qui est de très grande dimension. Ensuite, nous appliquons la méthode de sélection AG-PLS sur la matrice MIR⊗NIR. La première étape consiste à choisir les N chromosomes de taille L (les couples de nombres d'ondes) de manière aléatoire. Nous obtenons alors N matrices d'information spectrales aux couples de nombres d'ondes sélectionnés (MIR, NIR), chacune de ces matrices comporte J lignes et L colonnes (∈ $M_{J,L}(\mathbb{R})$). Nous réalisons ensuite N régressions PLS avec d'une part la matrice des variables mesurées (enzymes) et d'autre part les informations spectrales aux couples de nombres d'ondes sélectionnés. Ensuite, nous sélectionnons les L couples des nombres d'ondes (MIR, NIR) qui donnent la valeur minimale de RMSECV et effectuons les opérations de croisement et mutation sur cette matrice. Après la convergence, nous calculons le coefficient de détermination R^2 pour juger la qualité du résultat et mesurons la corrélation entre les variables prédites et les variables effectivement mesurées.

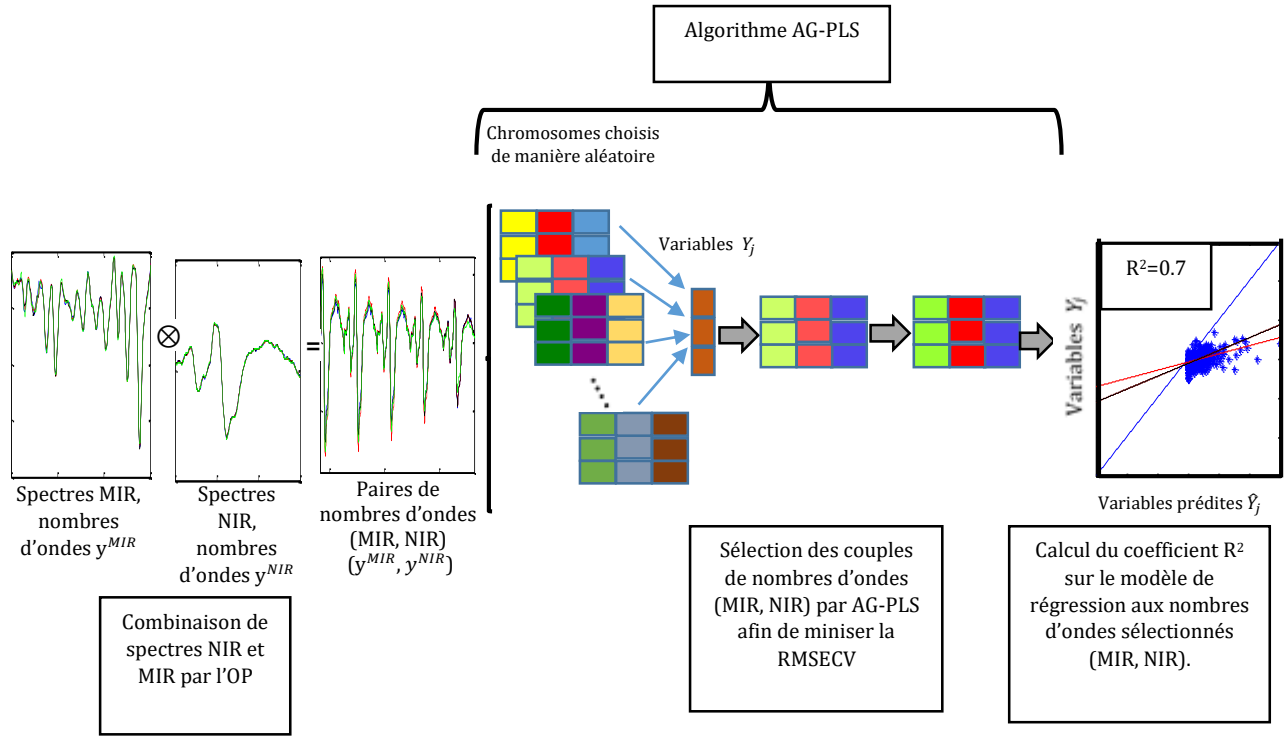


Figure 4.1. Synoptique de la méthodologie proposée OP-AG-PLS.

IV.5. Application à l'analyse de la dégradation de la biomasse lignocellulosique

Le but de cette étude est d'analyser quelle méthode est la plus adaptée pour prédire les valeurs prises par un ensemble de variables dépendantes, les activités enzymatiques et le taux de minéralisation en CO_2, à partir d'un groupe de variables indépendantes, les spectres IR des racines de maïs. Nous souhaitons prédire chaque variable chimique et biologique séparément.

Pour chaque application, nous mesurons le coefficient de détermination R^2, le pourcentage d'erreur relative de prédiction REP (%) et la racine de l'erreur quadratique moyenne de validation RMSECV (voir les équations (4.10), (4.12) et (4.14)) qui mesure l'ajustement du modèle à un ensemble d'observations qui n'a pas été utilisé lors de la modélisation. Plus les valeurs de RMSECV et REP sont faibles, meilleur est le résultat, alors que nous avons l'inverse pour R^2. Ces valeurs (R^2, REP et RMSECV) sont classiquement utilisées dans la littérature pour évaluer la qualité et les performances des modèles PLS. Dans les tableaux suivants, nous nous intéressons donc aux valeurs minimales de RMSECV et REP et aux valeurs maximales de R^2.

IV.5.1. Jeux de données

Nous disposons de 180 échantillons de maïs (racines), de génotypes différents (F2, F2bm1, F292bm3 et C), qui représentent la biomasse lignocellulosique à analyser aux périodes de biodégradation dans le sol : t_1=0, t_2=27, t_3=58, t_4=84, t_5=120, t_6=478 jours (voir le chapitre 1 pour plus de détails).

Les spectres ont été acquis sur ces échantillons dans les gammes spectrales MIR 400 - 4000 cm^{-1} et NIR 4000 - 8000 cm^{-1}. En s'appuyant sur les résultats présentés dans le chapitre 2, nous avons choisi la gamme spectrale 800 – 1800 cm^{-1} pour les spectres MIR, et 4000 – 6000 cm^{-1} pour les spectres NIR. Les spectres ont été prétraités avec les méthodes suivantes : filtrage Savitzky-Golay (SG) du 1ere ordre avec lissage sur 17 points suivi d'une normalisation de type Standard Normal Variate, (SNV). Ce

prétraitement est celui identifié dans le chapitre 2 comme étant optimal pour l'étude de la biomasse lignocellulosique et de sa dégradation. De plus, il permet d'obtenir une matrice centrée et normalisée de spectres, qui est une contrainte de l'algorithme PLS.

Pour les activités enzymatiques, nous disposons de cinq types des activités mesurées en µmol/h/g à chaque temps de décomposition des racines : l'activité β-glucosidases (BG), l'activité du N-acétyl glucosaminidase (NAG), l'activité phosphatase PHOS, la peroxydase nette (NP) et le phénol oxydase (PO). Pour caractériser la dégradation, nous avons deux variables sont quantifiées pour chaque temps de décomposition : C minéralisé cumulé (CM en mgC-CO_2/kg sol) et les vitesses de C minéralisé (VCM en mgC-CO_2/kg sol/jours).

IV.5.2. Application de la méthode PLS

Dans cette partie, nous appliquons la méthode de régression PLS sur les spectres MIR et NIR séparés, et sur les spectres combinées MIR-NIR et MIR⊗NIR. Nous notons que J = 180 le nombre d'échantillons de maïs sol et m=7 le nombre de caractérisation (variables chimiques et biologique) utilisées dans la PLS.

Le Tableau 4.1 montre les résultats de R^2, REP et RMSECV obtenus ainsi que le nombre optimal de variables latentes. Comme l'objectif est de prédire chaque variable séparément, nous obtenons pour chacune d'elles des valeurs différentes.

Tableau 4.1. Valeurs de RMSECV, REP, R^2 et du nombre de variables latentes VL pour la prédiction par la méthode PLS de différentes enzymes. Nous utilisons les informations spectrales MIR, NIR, et les informations combinées MIR-NIR et MIR⊗NIR.

PLS	MIR				NIR				MIR-NIR				MIR⊗NIR			
	REP%	RMSECV	R^2	VL	REP%	RMSECV	R^2	VL	REP%	RMSECV	R^2	VL	REP%	RMSECV	R^2	VL
BBG	6.8	0.38	0.45	4	7.5	0.40	0.42	3	6.3	0.29	0.49	4	5.2	0.26	0.59	5
NAG	7.2	0.30	0.40	5	7.4	0.32	0.38	3	6.7	0.27	0.46	3	5.7	0.26	0.57	5
PHOS	6.9	0.21	0.33	5	6.6	0.22	0.31	4	6.3	0.20	0.42	4	5.5	0.18	0.52	5
NP	7.3	0.34	0.30	3	7.9	0.38	0.29	3	7.0	0.32	0.31	2	6.1	0.26	0.32	2
PO	7.7	0.36	0.31	3	8.2	0.38	0.30	3	6.9	0.34	0.32	3	6.4	0.28	0.34	3
CM	6.5	0.31	0.50	4	6.7	0.34	0.49	3	6.2	0.28	0.58	4	5.5	0.24	0.65	5
VCM	6.7	0.31	0.52	3	6.9	0.32	0.50	3	6.3	0.28	0.62	4	5.8	0.24	0.70	5

Si l'on raisonne sur les spectres séparés (MIR et NIR), le résultat de prédiction sur les spectres MIR est meilleur que celui sur les spectres NIR. Cependant, la gamme du proche infrarouge (NIR) a souvent été préférée à l'infrarouge moyen (MIR), puisque l'analyse NIR nécessite peu d'échantillons et peu de préparations et semble être mieux adaptée pour une analyse « de terrain » [VLT14]. La reproductibilité de la mesure en NIR est par contre assez mauvaise. Il y a plusieurs études qui portent sur la dégradation dans le sol qui confirment nos résultats de calibration PLS. Par exemple, les résultats de l'estimation des valeurs des teneurs en C organique, après un traitement de type PLS, sont meilleurs dans le cas du MIR que dans le NIR [RWM06]. De même, une autre étude a montré que la spectroscopie MIR peut être utilisée pour déterminer la composition de céréales broyées avec une précision égale ou meilleure à celle obtenue par spectroscopie NIR [Coc04]. De plus, le spectre NIR est beaucoup plus complexe à interpréter puisqu'il correspond aux harmoniques et/ou aux combinaisons de bandes fondamentales de vibrations moléculaires [WN87], ce qui rend le traitement statistique PLS de ces spectres difficile. Le spectre MIR correspond uniquement à des bandes fondamentales de vibrations moléculaires, ce qui permet d'améliorer la calibration de la PLS.

Nous observons que la simple combinaison des informations spectrales MIR et NIR par une concaténation améliore le résultat par rapport aux bandes spectrales MIR ou NIR prises en compte séparément.

L'application de la PLS sur les spectres combinés par le produit extérieur MIR⊗NIR donne des valeurs REP et RMSECV qui sont globalement plus petites et des valeurs R^2 relativement plus grandes que si la PLS était appliquée sur les spectres seuls MIR et NIR ou sur les spectres concaténés MIR-NIR et ce pour toutes les variables chimiques et biologiques testées. Ceci démontre une fois de plus que la prise en compte de l'information spectrale mutuelle à tous les nombres d'onde MIR et NIR en utilisant le produit extérieur OP permet une meilleure prédiction de l'évolution des variables biologiques au cours des périodes du processus de biodégradation.

Ces résultats confirment ceux obtenus dans le chapitre 3 sur la discrimination de résidus selon leur stade de dégradation biologique sans prise en compte des informations complémentaires (avec uniquement des informations spectrales) et sur des spectres réalisés sur des biomasses, alors qu'ici les spectres ont été pris sur de sols.

IV.5.3. Application de la méthode VIP

Comme nous l'avons mentionné, la VIP est une méthode de sélection de variables basée sur la régression PLS couramment utilisée. Dans ce cas, les valeurs de RMSCV, REP et R^2 sont calculées en utilisant la matrice des informations spectrales aux nombres d'ondes sélectionnés. Le Tableau 4.2 montre les résultats de la méthode VIP pour les spectres MIR et NIR séparés et les spectres combinés MIR-NIR et MIR⊗NIR.

Tableau 4.2. Valeurs de REP (%), RMSECV, R^2 et du nombre de variables latentes VL pour les spectres MIR, NIR, MIR-NIR et MIR⊗NIR déterminées après application de la méthode VIP.

VIP	MIR				NIR				MIR-NIR				MIR⊗NIR			
	REP%	RMSECV	R^2	VL	REP%	RMSECV	R^2	VL	REP%	RMSECV	R^2	VL	REP%	RMSECV	R^2	VL
BBG	6.7	0.37	0.36	4	7.3	0.39	0.33	4	6.3	0.27	0.50	3	5.1	0.26	0.60	5
NAG	7.1	0.31	0.38	3	7.3	0.32	0.34	4	6.9	0.28	0.43	3	5.6	0.25	0.59	4
PHOS	6.8	0.20	0.29	3	6.9	0.22	0.26	3	6.5	0.22	0.36	3	5.6	0.20	0.49	4
NP	7.1	0.34	0.36	4	7.6	0.37	0.31	4	6.9	0.30	0.38	3	5.8	0.24	0.52	4
PO	7.6	0.35	0.35	3	8.0	0.37	0.32	4	6.7	0.32	0.40	3	6.2	0.26	0.53	4
CM	6.5	0.32	0.48	5	6.6	0.34	0.42	5	6.3	0.30	0.42	5	5.2	0.23	0.68	4
VCM	6.7	0.31	0.46	5	6.8	0.32	0.45	5	6.4	0.29	0.55	6	5.8	0.24	0.62	4

Ces résultats montrent que l'application de la méthode VIP sur les spectres combinés MIR⊗NIR donne, comme dans la partie précédente, de meilleurs résultats de calibration. Cependant, cette méthode est peu pratique parce que le nombre de bandes spectrales sélectionnées est très grand pouvant aller jusqu'à plus de 6000 sur les spectres combinés MIR⊗NIR. Ceci ne permet pas d'identifier de manière spécifique des bandes spectrales liées à la composition chimique qui sont les plus discriminantes pour qualifier la cinétique de biodégradation.

Par ailleurs, nous avons trouvé que l'application de la VIP sur les spectres MIR donne de meilleures performances de prédiction par rapport à l'application de la VIP sur les spectres NIR. La discussion de ce résultat est identique à celle présentée précédemment.
Nous pouvons également remarquer qu'il y a ici une grande différence du nombre de bandes L sélectionnées de l'ordre de 50 pour les spectres MIR et de 250 pour les spectres NIR. Ceci revient à la nature complexe du spectre NIR qui rappelons-le correspond aux harmoniques et aux combinaisons

de bandes fondamentales de vibrations moléculaires, alors que le spectre MIR correspond uniquement à des bandes fondamentales de vibrations moléculaires.

Nous avons également trouvé que l'application de la VIP sur les spectres MIR et NIR séparés donne une performance de prédiction moins bonne que l'application de la PLS. En revanche, l'application de la VIP sur les spectres combinés MIR-NIR et MIR⊗NIR donne des résultats d'étalonnage légèrement supérieurs à l'application de la PLS à cause de la nature complémentaire des techniques spectroscopiques MIR et NIR.

Ces résultats nous encouragent à privilégier l'utilisation de la combinaison de spectres MIR et NIR par le produit extérieur, quelles que soient les méthodes de sélection de variables (VIP ou PLS), par rapport à l'utilisation de spectres MIR et NIR séparés.

IV.5.4. Application des méthodes AG-PLS et OP-AG-PLS

Nous allons ici utiliser l'algorithme génétique avec la méthode PLS (AG-PLS) comme méthode de sélection de bandes. Comme précédemment cette méthode a été appliquée aux spectres MIR et NIR séparés et les spectres combinées MIR-NIR et MIR⊗NIR.

Tableau 4.3. Valeurs de REP (%), RMSECV, R^2 et du nombre de variables latentes VL pour les spectres MIR, NIR, MIR-NIR et MIR⊗NIR appliquées dans la méthode AG-PLS.

	AG-PLS												OP-AG-PLS			
	MIR				NIR				MIR-NIR				MIR⊗NIR			
	REP%	RMSECV	R^2	VL	REP%	RMSECV	R^2	VL	REP%	RMSECV	R^2	VL	REP%	RMSECV	R^2	VL
BBG	5.1	0.17	0.49	14	5.3	0.18	0.45	14	4.5	0.16	0.55	13	4.1	0.13	0.64	10
NAG	5.5	0.18	0.45	10	5.6	0.20	0.43	11	5.3	0.16	0.49	10	4.9	0.12	0.72	9
PHOS	5.1	0.17	0.42	12	5.2	0.19	0.36	11	5.0	0.15	0.46	11	4.7	0.11	0.64	8
NP	5.0	0.19	0.39	10	5.1	0.21	0.35	12	4.8	0.17	0.42	9	4.5	0.13	0.64	12
PO	5.2	0.18	0.38	9	5.2	0.18	0.37	10	5.0	0.14	0.46	9	4.8	0.11	0.66	13
CM	4.6	0.14	0.54	13	4.7	0.15	0.52	14	4.4	0.14	0.60	14	4.1	0.11	0.74	14
VCM	4.7	0.13	0.58	14	4.8	0.14	0.54	14	4.5	0.12	0.65	13	4.1	0.10	0.80	15

Il est facile à partir des résultats montrés dans le Tableau 4.3 de voir que notre méthode OP-AG-PLS, qui est l'application de l'AG-PLS sur les spectres combinés MIR⊗NIR, donne les meilleures valeurs de RMSCV, REP et R^2 par comparaison à celles obtenues par la simple application de l'algorithme AG-PLS sur les spectres séparés MIR et NIR ou sur les spectres concaténés MIR-NIR.

D'autre part, nous avons trouvé que l'application de l'AG-PLS sur les spectres MIR donne toujours de meilleurs résultats de prédiction que l'application de l'AG-PLS sur les spectres NIR. La discussion de ce résultat est toujours à rapprocher de celle présentée dans la section 5.2.

Lorsqu'on compare l'AG-PLS à la méthode VIP, nous observons une diminution importante des nombres d'ondes (ou couples de nombres d'ondes) sélectionnés. De plus, nous avons trouvé que l'AG-PLS permet l'identification de nombres d'ondes dans les deux gammes spectrales MIR et NIR qui correspondent aux groupes fonctionnels chimiques qui peuvent être impliqués dans l'évolution chimique des échantillons étudiés au cours de la biodégradation. Les nombres d'ondes sélectionnées par l'AG-PLS correspondent principalement aux fractions pariétales de la biomasse lignocellulosique (cellulose, l'hémicellulose et la lignine).

Si nous comparons les nombres d'ondes sélectionnés par l'AG sur les échantillons de biomasse lignocellulosique sans prise en compte des informations chimiques (présentés dans le chapitre 3) et les bandes sélectionnées par l'AG-PLS sur les échantillons de biomasse lignocellulosique, mais cette fois avec la prise en compte des informations chimiques, nous trouvons l'AG-PLS sélectionne une quantité plus importante de nombres d'ondes que l'algorithme AG sans utiliser la régression PLS. Comme ce nombre dépend significativement des informations biologiques à prédire, nous pouvons en conclure que plus la modélisation de l'évolution d'une variable biologique (enzyme par exemple) s'éloigne d'un modèle linéaire, plus il faudra un nombre important de nombres d'ondes pour décrire ce modèle. A l'opposé, quand aucun modèle n'est à construire, comme c'était le cas présenté dans le chapitre précédent, un nombre réduit de nombres d'ondes permet de classer la biomasse lignocellulosique.

IV.5.5. Comparaison et discussion

IV.5.5.1. Nombre de variables latentes

Les méthodes présentées ci-dessus montrent des différences dans le nombre de variables latentes obtenues pour chaque méthode. Par exemple, nous avons trouvé que le nombre de variables latentes pour la prédiction de l'activité phosphatase par la méthode PLS sans AG est de 5 alors qu'avec l'algorithme génétique AG-PLS celui-ci est de 13. En effet, le critère de choix du nombre optimal de variables latentes VL est celui pour lequel on obtient le premier minimum de la valeur RMSECV. Pour illustrer cette discussion, nous avons appliqué les méthodes PLS et AG-PLS sur les spectres MIR pour l'enzyme PHOS. La Figure 4.2 représente les valeurs de RMSECV pour différentes variables latentes.

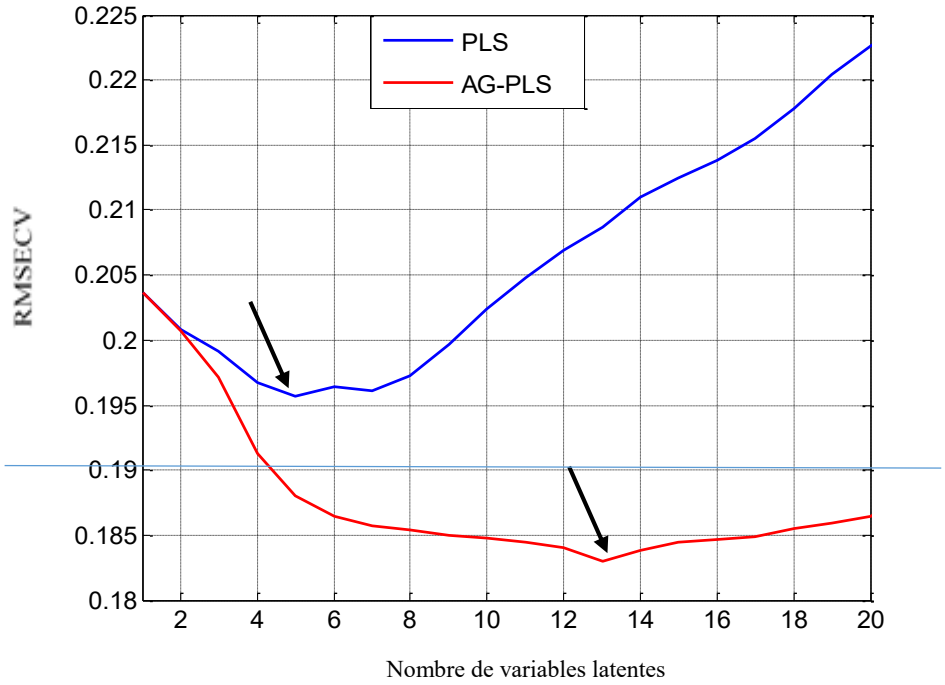


Figure 4.2. Courbes des valeurs RMSECV en fonction du nombre de variables latentes déterminées par les méthodes PLS (courbe en bleu) et AG-PLS (courbe en rouge) obtenues pour la modélisation de l'enzyme PHOS à partir de spectres MIR.

Pour la méthode PLS, on obtient le minimum de la valeur RMSECV (0.196) pour 5 variables latentes. L'AG-PLS appliqué sur 13 variables latentes donne une plus petite valeur de RMSECV (0.184).

L'évolution des courbes montre que les valeurs de RMSECV obtenues avec l'AG-PLS sont, à partir de 4 variables latentes, inférieures à celles obtenues par la méthode PLS. On peut donc en conclure que l'AG-PLS est plus performante que la méthode PLS. Si nous avions comparé les valeurs RMSECV de chaque méthode sur les mêmes nombres de variables latentes (5 ou 13), nous aurions également trouvé que l'AG-PLS donnait des valeurs RMSECV plus petites que la PLS. Toutefois, ce type de comparaison rend l'analyse des résultats plus complexe et nous avons préféré calculer la valeur optimale du nombre de variables latentes pour chaque méthode comme étant celle qui minimise les valeurs RMSECV. Cette discussion montre qu'il est possible de comparer les résultats de différentes méthodes de sélections basées sur la PLS en utilisant des nombres de variables latentes (VL) différents.

IV.5.5.2. Comparaison par rapport aux erreurs de modélisation

Pour synthétiser les résultats obtenus avec les méthodes PLS, VIP et AG-PLS, les valeurs de RMSECV et R^2 ont été reportées sur les Figure 4.3 et Figure 4.4 en considérant les spectres MIR, NIR, MIR-NIR et MIR⊗NIR et ce pour chaque activité enzymatique.

On observe que la méthode OP-AG-PLS donne les plus grandes valeurs de R^2 (les valeurs les plus petites de RMSECV, cf. Figure 4.4) en comparaison avec les autres méthodes PLS et VIP. L'OP-AG-PLS apparaît donc comme une méthode très performante pour la calibration et la prédiction des informations chimiques et biologiques à partir de données de biomasse lignocellulosique.

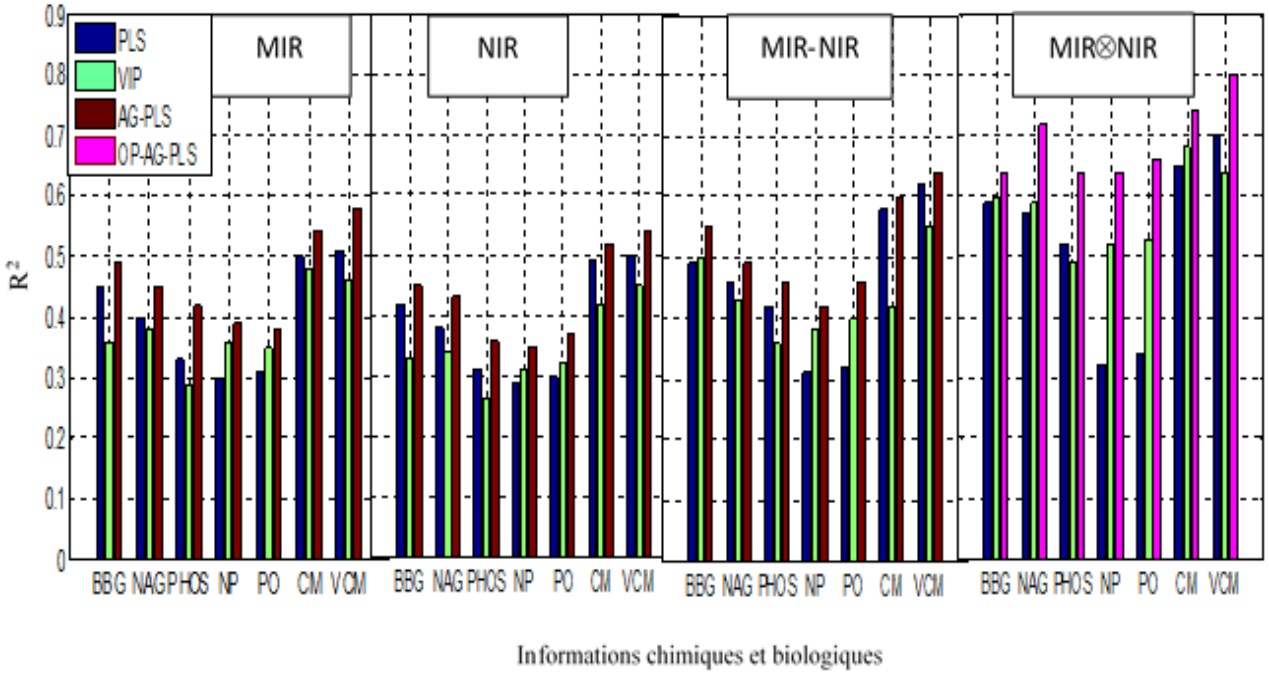


Figure 4.3. Valeurs de R^2 pour les méthodes PLS, VIP et AG-PLS appliquées sur les spectres MIR, NIR, MIR-NIR et MIR⊗NIR avec les informations chimiques (CM et VCM) et biologiques (activités enzymatiques BBG, NAG, PHOS, NP, PO) (voir section 4.4.1 pour les définitions).

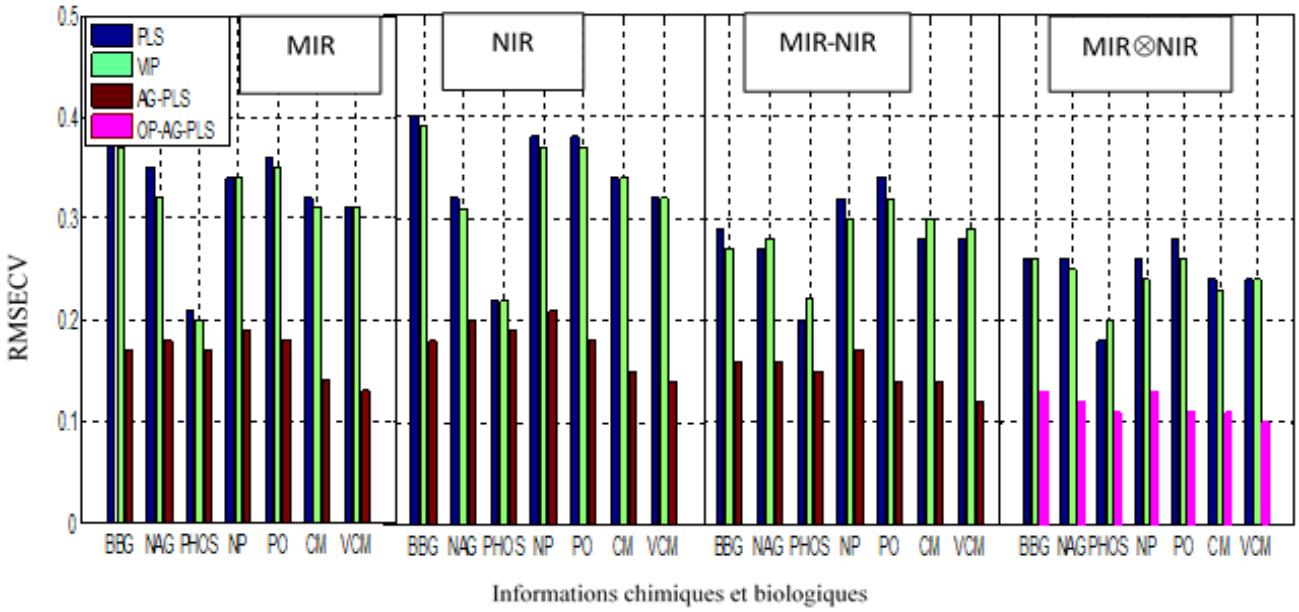


Figure 4.4. Valeurs de RMSECV pour les méthodes PLS, VIP et AG-PLS appliquées sur les spectres MIR, NIR, MIR-NIR et MIR⊗NIR avec les informations chimiques (CM et VCM) et biologiques (activités enzymatiques BBG, NAG, PHOS, NP, PO)

IV.5.5.3. Comparaison par rapport au pouvoir prédictif

Pour quantifier les résultats des méthodes de régression PLS et l'algorithme AG-PLS, nous avons construit les résultats des prédictions des informations chimiques et biologiques à partir des spectres MIR, NIR, MIR-NIR et MIR⊗NIR enregistrés sur la racines à différents stades de dégradation dans le sol. Pour chaque représentation de la Figure 4.5, la droite bleue représente $\hat{Y} = Y$ (donc le résultat idéal). Pour rappel, $\hat{Y}$ sont les informations chimiques ou biologiques prédites et Y les valeurs mesurées. La droite rouge représente $\hat{Y}_1 = aY + b$, avec $\hat{Y}_1$ les valeurs prédites avec le modèle PLS. La droite noire représente $\hat{Y}_2 = aY + b$, avec $\hat{Y}_2$ les enzymes prédites par la méthode AG-PLS. Plus les lignes rouge ou noire se rapprochent de la ligne bleue, meilleurs sont les résultats de prédiction.

Pour l'ensemble des représentations regroupées dans la Figure 4.5 que les lignes noires sont toujours plus proches des lignes bleues que les lignes rouges. Cela signifie que la méthode AG-PLS donne de meilleurs résultats de prédiction des informations chimiques et biologiques que la méthode PLS. L'AG-PLS améliore donc la calibration en identifiant des nombres d'ondes qui sont plus spécifiques à l'évolution des informations chimiques et biologiques mesurées. Ce résultat est en concordance avec ceux obtenus pour la calibration du modèle et pour la prédiction (valeurs de R^2, RMSECV et REP, cf. Tableau 4.1, Tableau 4.2, Tableau 4.3, Figure 4.3 et Figure 4.4), nous trouvons qu'il existe une bonne corrélation. En effet, plus les valeurs de R^2 sont élevées et plus les valeurs de REP et RMSECV sont faibles (dans nos tableaux), plus les figures montrent des meilleurs résultats de calibration. Par ailleurs, le produit extérieur des informations spectrales MIR et NIR (OP-AG-PLS) améliore également la prédiction des informations chimiques et biologiques, en accord avec les résultats précédents

D'autre part, la comparaison les niveaux de prédiction des informations montre que les informations chimiques : C minéralisé cumulé (CM en mgC-CO_2/kg sol) et vitesses de C minéralisée (VCM en mgC-CO_2/kg sol/jours) sont mieux prédites que les activités enzymatiques, car le carbone C du CO_2 est majoritairement lié au carbone C des constituants lignocellulosiques [Amin12].

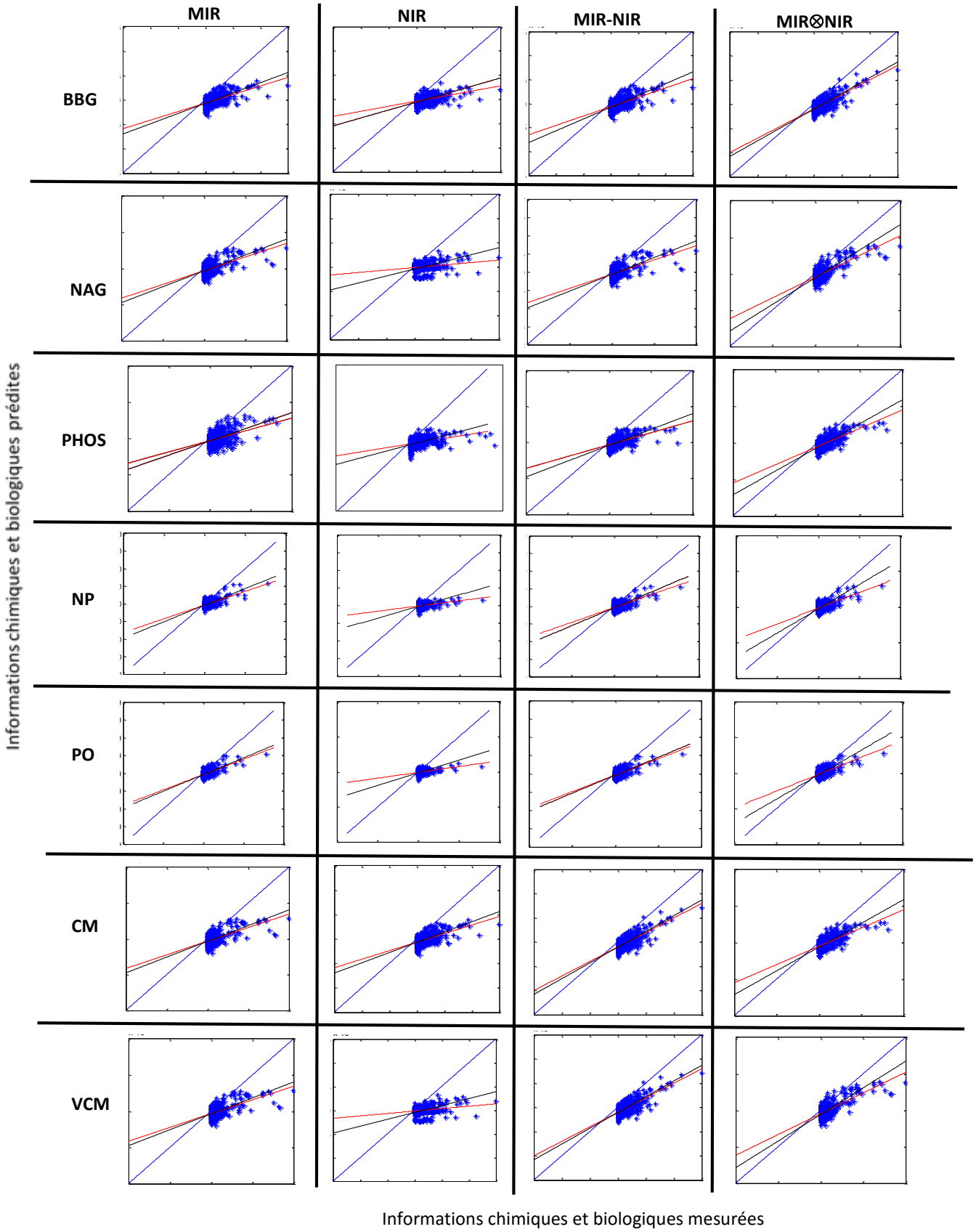


Figure 4.5. Résultats (informations prédites vs enzymes mesurées) en utilisant les spectres MIR, NIR, MIR-NIR et MIR⊗NIR.

IV.6. Conclusion

Dans ce chapitre nous avons présenté différentes méthodes de régression et nous nous sommes attardés sur la méthode de régression des moindres carrés partiels. En effet, cette dernière présente l'avantage d'introduire la notion de connaissance a priori qui est, pour notre application, la connaissance des valeurs des données enzymatiques et des données de minéralisation de C de biomasse lignocellulosique.

Nous avons présenté une méthode classique (VIP) qui permet de sélectionner des nombres d'ondes dans le cadre de l'application de l'algorithme PLS. Nous avons ensuite proposé d'implémenter l'algorithme génétique (AG) présenté au chapitre précédent en le combinant avec la régression PLS. Pour cela, nous avons proposé d'utiliser l'erreur quadratique moyenne de validation croisée RMSECV comme fonction fitness. Nous obtenons alors l'algorithme (AG-PLS) qui, appliqué sur les spectres combinés par le produit extérieur (OP), conduit à une nouvelle méthode appelée OP-AG-PLS.

Nous avons appliqué ces différentes méthodes sur des spectres NIR et MIR séparés, puis sur des spectres NIR et MIR regroupés par une simple concaténation et enfin sur des spectres NIR et MIR reliés par le produit extérieur. Nos premières conclusions montrent que les régressions (PLS, VIP et AG-PLS) sur les spectres NIR donnent de meilleurs résultats que sur les spectres MIR. L'algorithme génétique AG-PLS fournit quant à lui les meilleures performances (valeurs de R^2 maximales et RMSECV minimales) comparées aux deux méthodes plus classiques PLS et VIP à la fois sur les spectres MIR et NIR séparés. Les résultats obtenus avec l'AG-PLS sont également meilleurs pour les spectres MIR et NIR fusionnés ou reliés par le produit extérieur (OP). L'algorithme OP-AG-PLS permet, à partir d'un nombre extrêmement restreint de facteurs (de 9 à 16 bandes spectrales sélectionnées), d'obtenir les meilleurs résultats toutes méthodes confondues. Nous avons également montré que l'OP-AG-PLS permet d'améliorer considérablement les performances de prédiction des données enzymatiques et chimiques dans un processus global de dégradation de biomasse lignocellulosique.

Conclusion générale et perspectives

Le travail effectué au cours de cette thèse de doctorat a permis de montrer l'importance du rôle des mathématiques appliquées et du traitement du signal pour l'étude de la dégradation de la biomasse végétale, la source de carbone renouvelable la plus abondante de la planète. En effet, la matière lignocellulosique, principal constituant de la biomasse végétale, est généralement quantifiée par des méthodes chimiques. . Ces méthodes sont destructives, et souvent couteuses en temps d'analyse pour parvenir à caractériser spécifiquement les composés ciblés. Une alternative pour analyser la matière lignocellulosique est la spectroscopie IR à transformée de Fourier (FTIR). La gamme moyen IR (MIR MIR) est sensible à la fois à des constituants organiques et inorganiques tandis que la gamme proche IR (NIR) à des composants minéraux et organiques.

À cause de la nature complexe, mais complémentaire, des informations spectrales MIR et NIR, leur combinaison a été considérée comme l'un des défis pour bon nombre d'applications. Après avoir étudié les méthodes classiques utilisées, nous avons proposé de combiner les informations spectrales par le produit extérieur. Cet opérateur mathématique permet d'associer l'information spectrale de chaque nombre d'ondes MIR avec l'information spectrale de chaque nombre d'ondes NIR. De ce fait, cette méthode de combinaison met l'accent sur l'interaction entre les vibrations moléculaires fondamentales et harmoniques et les combinaisons de vibrations fondamentales.
D'autre part, puisque ces spectres sont complexes, nous avons montré qu'on pouvait utiliser et développer des outils de traitement du signal afin d'identifier des marqueurs spectroscopiques qui permettent de discriminer le niveau de dégradation de la matière lignocellulosique. Les outils mathématiques ont été également utilisés pour construire des modèles fonctionnels permettant de prédire la dégradation de la biomasse végétale à partir des informations spectrales. Ces outils ont pris en compte la cinétique de dégradation et non pas uniquement le point final, comme la majorité des travaux effectués jusqu'à maintenant.

Dans le deuxième chapitre, nous avons étudié les différentes méthodes de prétraitement des spectres IR et de choix des gammes spectrales. L'objectif était d'identifier celles les mieux adaptées à l'étude de la biomasse lignocellulosique et notamment de sa dégradation. Pour cela, nous avons envisagé d'utiliser un algorithme de classification non supervisé Fuzzy C Means (FCM) qui permet de réaliser ce choix en optimisant la classification de la biomasse lignocellulosique. Toutefois, la distance euclidienne utilisée par défaut ne couvre que les données de classes sphériques. Or, comme nos spectres IR enregistrés sur la biomasse lignocellulosique ont une répartition non sphérique, nous avons proposé un nouvel algorithme FCM basé sur un facteur de covariance. Nous avons montré sur des données synthétiques et réelles que cet algorithme, que nous avons appelé FCM-R, convient à des données appartenant à des classes non sphériques et sphériques, tout en tenant compte de la grande dimensionnalité des spectres IR. Cet algorithme a été utilisé ensuite pour étudier l'influence des principales méthodes de prétraitement et leurs paramètres et ainsi pour sélectionner la solution optimale par rapport à la dégradation de la biomasse lignocellulosiques. Ainsi, pour les spectres MIR, cette approche a permis de sélectionner les prétraitements LB suivie d'une SNV ainsi que SG d'ordre 1 suivie d'une SNV et une gamme spectrale optimale de 800-1800 cm^{-1}. Pour les spectres NIR, la gamme spectrale optimale a été la 4000-6000 cm^{-1} et la méthode de prétraitement la SG d'ordre 1 suivie d'une SNV.

Dans le troisième chapitre, nous nous sommes intéressés à développer des méthodes permettant de sélectionner des bandes spectrales IR discriminant la matière lignocellulosiques en fonction de son

niveau de dégradation. En utilisant uniquement les informations spectrales, nous avons proposé dans un premier temps une méthode basée sur un algorithme génétique (AG). Nous avons montré qu'en choisissant les étapes adaptées et l'indice Davies-Bouldin comme fonction fitness, cette méthode permet l'identification de nombres d'ondes dans les deux gammes spectrales MIR et NIR qui correspondent aux groupes fonctionnels chimiques et qui peuvent être attribués à des composés lignocellulosiques qui subissent une évolution au cours de la dégradation de la biomasse lignocellulosique. Ensuite, nous avons montré que l'analyse conjointe des spectres MIR et NIR par le produit extérieur fournit un meilleur résultat que le traitement séparé des spectres MIR ou NIR, permettant de mieux discriminer la biomasse lignocellulosique au cours du processus de dégradation. Dans un deuxième temps, nous avons proposé une nouvelle approche d'optimisation basée sur la minimisation d'une fonction objectif. Cette approche permettant de sélectionner un vecteur qui met en évidence les nombres d'ondes les plus discriminants mais également leurs poids. Cette méthode requiert moins des paramètres, tout en étant plus rapide, que la méthode précédente tout en étant plus rapide, et donne des bons résultats en comparaison avec ceux obtenus par l'AG sur les spectres MIR et NIR séparés. Cependant, l'inconvénient principal de cette approche est qu'elle ne peut pas être appliquée sur des spectres combinés par le produit extérieur à cause d'un problème de dimensionnalité.

Dans le quatrième chapitre, nous avons développé des modèles capables de prédire des informations reflétant l'état de dégradation de la biomasse. Pour cela, en plus de l'information spectrale, nous avons considéré également l'information chimique. Dans un premier temps, nous avons combiné la méthode classique de régression des moindres carrés partiels (PLS) avec l'approche de sélection de bandes spectrales par AG. Nous avons montré que l'algorithme développé, l'AG-PLS, ainsi que les méthodes classiques de PLS ou VIP (projection des variables d'importances) donnent de meilleurs résultats sur les spectres MIR que sur les spectres NIR. De plus, l'algorithme AG-PLS fournit les meilleures performances (valeurs de R^2, RMSECV REP) par rapport aux méthodes classiques, peu importe la gamme spectrale. Nous avons proposé ensuite une extension de cet algorithme AG-PLS en combinant les informations spectrales MIR et NIR par produit extérieur (OP). Nous avons conclu qu'en s'appuyant sur des couples des bandes spectrales MIR – NIR, la nouvelle approche OP-AG-PLS permet une calibration optimale, ce qui améliore considérablement les performances de prédiction des activités enzymatiques et taux de minéralisation dans le processus de dégradation de la biomasse lignocellulosique.

Ce travail permet d'envisager plusieurs perspectives.
Afin d'identifier des bandes spectrales IR discriminantes de résidus lignocellulosiques en fonction de leur niveau de dégradation à l'aide de méthodes mathématiques sur les spectres de grandes dimensions comme les spectres combinés par l'OP, il serait intéressant de proposer une autre approche d'optimisation basée sur la minimisation d'un vecteur de pondération, appliquée sur les spectres combinés MIR⊗NIR. Cette méthode nous permettrait d'extraire des informations spectrales plus complètes en sélectionnant des couples des bandes spectrales (MIR, NIR) de façon rapide ainsi que leurs poids respectifs.

D'autre part, il pourrait également être intéressant d'appliquer des méthodes de décomposition tensorielle pour la séparation de sources telles que les algorithmes de décomposition d'un tableau de données IR tri-dimensionnel comme par exemple les méthodes Candecomp/Parafac (CP), la décomposition PARALIND ou la décomposition CP quadri-linéaire. L'idée est d'extraire des bio-marqueurs liés aux caractéristiques intrinsèques de la biomasse lignocellulosique et à sa dégradation,

par exemple pour évaluer la courbe de minéralisation du carbone ou pour identifier des signaux (sources) liés aux systèmes biologiques qui peuvent être les bases d'un modèle fonctionnel de déstructuration de tissus végétaux.

Bibliographie

[AA97] M. Amari, A. Abe. Application of near infrared reflectance spectroscopy to forage analysis and prediction of TDN contents. JARQ-Jpn Agr Res Q. Vol. 31, pp. 55-63, 1997.

[AGM13] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M.Prez, I. Perona. An extensive comparative study of cluster validity indices. Pattern Recognition. Vol. 46, pp. 243 – 256, 2013.

[Aga99] U.P. Agarwal. An overview of Raman spectroscopy as applied to lignocellulosic materials. In: Argyropoulos DS Advances in lignocellulosic characterization. TAPPI, Atlanta, pp 201–225, 1999.

[AJS03] C. Abrahamsson, J. Johansson, A. Sparen, F. Lindgren. Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets. Chemometrics and intelligent Laboratory Systems. Vol. 69, pp. 3-12, 2003.

[AJT98] D. Arvid Skoog, F. James Holler, A. Timothy. Nieman. Principes d'analyse instrumentale. Fifith edition. By Harcourt brace & company. 1998.

[AK12] N.K. Afseth, A. Kohler. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. Chemometrics and Intelligent Laboratory Systems. Vol. 117, pp. 92–99, 2012.

[Aka69] H. Akaike. Fitting autoregressive models for prediction. Annals of the Institute of Statistical Mathematics. Vol. 21 N°1, pp. 243-247, 1969.

[Alf08] R. Alfred. A Genetic Based Feature Construction Method for Data Summarization. Chapter: Advanced Data Mining and Applications. Vol. 5139 of the series Lecture Notes in Computer Science, pp. 39-50, 2008.

[AMH09] G. Allison, C. Morris, E. Hodgson, J. Jones, M. Kubacki, T. Barraclough, N. Yates, I. Shield, A. Bridgwater, I. Donnison. Measurement of key compositional parameters in two species of energy grass by Fourier transform infrared spectroscopy. Bioresource Technology. Vol. 100, pp. 6428–6433, 2009.

[Amin12] B.A.Z. Amin. Rôle des enzymes lignocellulosiques dans le processus de biodégradation de résidus végétaux dans les sols : Influence de la qualité des résidus sur l'efficacité des enzymes et leur dynamique. Thèse de doctorat, Université de Reims Champagne Ardenne, France, 2012.

[ASG01] M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvão, T. Yoneyama, H. C. Chame, V. Visani. The Successive Projections Algorithm for variable selection in spectroscopic multicomponent. Chemometrics and Intelligent Laboratory Systems. Vol. 57 N°2, pp. 65-73, 2001.

[Ash13] Q. Ashton Acton. Issues in Fossil Fuel Energy Technologies; Chapitre 2: Oil and gas research. Edition 2013.

[ATM09] G.G. Allison, S.C. Thain, P. Morris, C. Morris, S. Hawkins, B. Hauck, T. Barraclough, N. Yates, I. Shield, A.V. Bridgwater, I.S. Donnison. Quantification of hydroxycinnamic acids and lignin in perennial forage and energy grasses by Fourier-transform infrared spectroscopy and partial least squares regression. Bioresource Technology. Vol. 100, pp. 1252–1261. 2009.

[AV99] R.H. Atalla, D.L. VanderHart. The role of solid-state carbon¬13 NMR spectroscopy in studies of the nature of native celluloses. Solid State Nucl Magn Reson. Vol. 15 N°1, pp. 1–19, 1999.

[Aya13] A.R. Ayad. Parametric analysis for genetic algorithms handling parameters, Alexandria Engineering Journal. Vol. 52, pp.99–111, 2013.

[Ban01] S. Bandyopadhyay. Nonparametric Genetic Clustering: Comparison of Validity Indices. IEEE Transactions on Systems, Man, and Cybernetics: Applications and Reviews. Vol. 31 N°1, pp. 120-125, 2001.

[Bak87] J. E. Baker. Reducing bias and inefficiency in the selection algorithm. In J. J. Grefenstette (ed), Proceedings of the International Conference on Genetic Algorithms, pages14-21, 1987.

[BAM12] A. Baum, J. Agger, A. S. Meyer, M. Egebo J. Mikkelsen. Rapid near infrared spectroscopy for prediction of enzymatic hydrolysis of corn bran after various pretreatments. New Biotechnology. Vol. 29 N°3, 2012.

[Bal07] D. Ballerini. Les biocarburants de première génération: l'éthanol et le biodiesel. 2007.

[BBF06] B.G. Barthès, D. Brunet, H. Ferrer, J.L. Chotte, C. Feller. 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. Vol. 14, pp. 341-348, 2006.

[BB86] R.S. Bretzlaff, T.B. Bahder. Apodization effects in Fourier transform infrared difference spectra. Rev. Phys. Appl. Vol. 21, pp. 833-844, 1986.

[BBP02] K. Brinkmann, L. Blaschke, A. Polle. Comparison of different methods for lignin determination as a basis for calibration of near-infrared reflectance spectroscopy and implications of lignoproteins. J Chem Ecol. Vol. 28, pp. 483-501, 2002.

[BDC07] D. Brunet, B.G. Barthès, J.L. Chotte, C. Feller. 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. Vol. 139, pp. 106-117, 2007.

[BDH02] H. Baillères, F. Davrieux, F. Ham-Pichavant. Near infrared analysis as a tool for rapid screening of some major wood characteristics in a eucalyptus breeding program. Ann For Sci. Vol. 59, pp. 479-490, 2002.

[BDL89] R.J. Barnes, M.S. Dhanoa, S.J. Lister. Standard normal variate transformation and detrending of near infrared diffuse reflectance spectra, Appl.Spectrosc. Vol. 43, pp. 772-777, 1989

[BDL99] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Correction of the description of Standard Normal Variate (SNV) and De-Trend transformations in Practical Spectroscopy with Applications in Food and Beverage Analysis, 2nd. Edition, J. Near Infrared Spectrosc. Vol. 1, pp. 185-186. 1999.

[Ber05] D. Bertrand. Étalonnage multidimensionnel : application aux données spectrales. Techniques de l'Ingénieur, pp. 1-20, 2005.

[Bez81] J.C. Bezdek. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press: New York, 1981.

[BJC01] N. Büchmann, H. Josefsson, I. Cowe, Performance of European Artificial Neural Network (ANN) Calibrations for Moisture and Protein in Cereals Using the Danish Near- Infrared Transmission (NIT) Network, Cereal Chemistry. Vol. 78 N°5, pp. 572–577, 2001.

[BJM10] S. Bruuna, J. W. Jensen, J. Magid, J. Lindedam, S. B. Engelsen. Prediction of the degradability and ash content of wheat straw from different cultivars using near infrared spectroscopy. Industrial Crops and Products. Vol. 31, pp. 321–326, 2010

[BKC10] J.S. Bak, M.D. Kim, I.G. Choi, K.H. Kim. Biological pretreatment of rice straw by fermenting with Dichomitus squalens. New biotechnology. Vol. 27 N°424, 2010

[BKM98] C. Blake, E. Keogh, C. J. Merz. UCI Repository of Machine Learning Databases, Dept. Inf. Comput. Sci., Univ. California, Irvine, CA, 1998. From http://www.ics.uci.edu/ mlearn/MLRepository.html

[BKW05] D. A. Belsley, E. Kuh, R. E. Welsch. Book Review: Régression Diagnostic-Identifying influential data and sources of collinearity. Biometrical Journal. 2005.

[BM94] C.N. Banwell, E.M. Mccash. Fundamentals of Molecular Spectroscopy. McGraw Hill. 1994.

[BM11] V. Bellon-Maurel, A. McBratney. Near infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils e Critical review and research perspectives. Soil Biology & Biochemistry. Vol. 43, pp. 1398-1410, 2011.

[BMB10] S.K. Borgen, L. Molstad, S. Bruun, T.A. Breland, L.R. Bakken, M.A. Bleken. Estimation of plant litter pools and decomposition-related parameters in a mechanistic model. Plant and Soil. Vol. 338, pp. 2015-222, 2010.

[BMF02] A.S. Barros, I. Mafra, D. Ferreiraa,, S. Cardoso, A. Reis, J.A. Lopes da Silva, I. Delgadillo, D.N. Rutledge, M.A. Coimbra. Determination of the degree of methyl esterification of pectic polysaccharides by FT-IR using an outer product PLS1 regression. Carbohydrate Polymers. Vol. 50, pp. 85-94, 2002.

[Bou08] H. Boussarsar. Application de Traitements Thermique et Enzymatique de Solubilisation et Saccharification de la Fraction Hemicellulosique en Vue de la Valorisation de la Bagasse de Canne à Sucre. Doctorat d'Etat, Université de Reims Champagne-Ardenne, France, 2008.

[BPC09] I. Bertrand, M. Pervot, B. Chabbert. Soil decomposition of wheat internodes of different maturity stages: Relatives impact of the soluble and structural fractions. Bioressource Technology. Vol. 100, pp. 155-163, 2009.

[Bru62] W. Brugel. An Introduction to Infrared Spectroscopy. Methuen & Co. Ltd., 1962.

[BrS11] P. Brian, C. Smith. Fundamentals of Fourier Transform Infrared Spectroscopy, Second Edition. CRC Press, 2011.

[BS11] R. M. Balabina, S. V. Smirnov. Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data. Analytica Chimica Acta. Vol. 692, pp. 63–72, 2011.

[BSB05] S. Bruun, B. Stenberg, T.A. Breland, J. Gudmundsson, T.M. Henriksen, L.S. Jensen, A. Korsaeth, J. Luxhoi, F. Palmason, A. Pedersen, T. Salo. Empirical predictions of plant material C and N mineralization

patterns from near infrared spectroscopy, stepwise chemical digestion and C/N ratios. Soil Biology & Biochemistry. Vol. 37, pp. 2283-2296, 2005.

[BSS96] A.S. Bangalore, R.E. Shaffer, G.W. Small, M.A. Arnold. Genetic algorithm-based method for selecting wavelengths and model size for use with partial least-squares regression: application to near-infrared spectroscopy. Analytical Chemistry. Vol. 68 N°23, pp. 4200-4212, 1996.

[BUA15] R. Biswas, H. Uellendahl, B.K. Ahring. Wet Explosion: a Universal and Efficient Pretreatment Process for Lignocellulosic Biorefineries. Bioenerg. Res. 2015.

[CAD09] F. Calderón, V. Acosta-Martinez, D. Douds, J. B. Reeves, M. F. Vigil. Mid-Infrared and Near-Infrared Spectral Properties of Mycorrhizal and Non-mycorrhizal Root Cultures. Applied Spectroscopy. Vol. 63 N°5, pp. 494-500, 2009.

[CBH97] V.S. Chang, B. Burr, M.T. Holtzapple. Lime pretreatment of switchgrass. Applied biochemistry and biotechnology. Vol. 63 N°3, 1997.

[CCC08] L. Cecillon, N. Cassagne, S. Czarnes, R. Gros, J. Brun. Variable selection in near infrared spectra for the biological characterization of soil and earthworm casts. Soil Biology & Biochemistry. Vol. 40, pp. 1975–1979, 2008

[CDW75] N.B. Colthup, L.H. Daly, S.E. Wiberley. Introduction to Infrared & Raman Spectroscopy. Academic Press, 1975.

[CFA10] H. Chena, C. Ferrari, M. Angiuli, J. Yao, Costantino Raspi, Emilia Bramanti. Qualitative and quantitative analysis of wood samples by Fourier transform infrared spectroscopy and multivariate analysis. Carbohydrate Polymers. Vol. 82, pp. 772–778, 2010.

[CGB14] A. Carballo-Meilan, A. M. Goodman, M. G. Baron, J. Gonzalez-Rodriguez. A specific case in the classification of woods by FTIR and chemometric: discrimination of Fagales from Malpighiales. Cellulose. Vol. 21, pp. 261–273, 2014.

[CK10] H. Chun, S. Keles. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society. Vol. 72, pp. 3–25, 2010.

[CKC11] C. Cheng Hung, S. Kulkarni, B. Chen Kuo, “A New Weighted Fuzzy C-Means Clustering Algorithm for Remotely Sensed Image Classification”, IEEE Journal of Selected Topics in Signal Processing. Vol. 5, pp. 543-553, 2011.

[CLS08] W. Cai, Y. Li, X. Shao. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. Chemometrics and Intelligent Laboratory Systems. Vol. 90 N°2, pp.188-194, 2007.

[Coa00] J. Coates. Interpretation of infrared spectra, a practical approach. In: R.A. Meyers. 2000.

[Coc04] H. Cocchi. Soil conservation, output diversification and farm income: Evidence from hillside farmers in Central America. Ph.D. Dissertation, University of Connecticut, 2004.

[Col61] N.B. Colthup. Vibrating molecular models: Frequency shifts in strained ring double bonds. Journal of Chemical Education. Vol. 38 N°8, pp. 394-396, 1961.

[CM96] V. Centner, D. Massart. Elimination of Uninformative Variables for Multivariate Calibration. Anal. Chem. Vol. 68, pp. 3851-3858, 1996.

[CNH98] V.S. Chang, M. Nagwani, M.T. Holtzapple. Lime pretreatment of crop residues bagasse and wheat straw. Applied biochemistry and biotechnology. Vol. 74 N°135, 1998.

[CP04] L. Chiang, R. Pell. Genetic algorithms combined with discriminant analysis for key variable identification. Journal of Process Control. Vol. 14, pp. 143–155, 2004.

[CPS12] B. Chong, D. E. Purcell, M. G. O’Shea. Diffuse Reflectance, Near-Infrared Spectroscopic Estimation of Sugarcane Lignocellulose Components—Effect of Sample Preparation and Calibration Approach. Bioenerg. Res. 2012.

[CRD14] R. Chazal, P. Robert, S. Durand, M. Devaux, L. Saulnier, C. Lapierre, F. Guillon. Investigating Lignin Key Features in Maize Lignocelluloses Using Infrared Spectroscopy. Applied Spectroscopy. Vol. 68 N°12, pp. 1342-1347, 2014.

[CRF07] F.J. Calderon, J.B. Reeves III, J.G. Foster, W.M. Clapham, J.M. Fedders, M.F. Vigil, W.B. Henry. Comparison of Diffuse Reflectance Fourier Transform Mid-Infrared and Near-Infrared Spectroscopy with Grating-Based Near Infrared for the Determination of Fatty Acids in Forages. Journal of Agricultural and Food Chemistry. Vol. 55, pp. 8302-8309, 2007.

[CSO10] M. Casalea, N. Sinelli, P. Oliveri, V. Di Egidio, S. Lanteri. Chemometrical strategies for feature selection and data compression applied to NIR and MIR spectra of extra virgin olive oils for cultivar identification. Talanta. Vol. 80, pp. 1832–1837, 2010.

[CSJ04] J.C. Clifton-brown, P.F. Stampfl, M.B. Jones. Miscanthus biomass production for energy in Europe and its potential contribution to decreasing fossil fuel carbon emissions. Global Change Biology. Vol. 10 N°509, 2004.

[DB06] A. Demirbas, M. Balat. Recent advances on the production and utilization trends of bio-fuels: A global perspective. Energy Conversion and Management. Vol. 47 N°16, pp. 2371-2381, 2006.

[DBL00] C. Dorrer, N. Belabas, J.P. Likforman, M. Joffre. Spectral resolution and sampling issues in Fourier-transform spectral interferometry. Journal of the Optical Society of America B, Vol. 17 N°10, pp. 1795-1802, 2000.

[Deb02] K. Deb. Multi-Objective Optimization Using Evolutionary Algorithms. Wiley, 2002.

[Del95] S.R. Delwiche. Single wheat kernel analysis by near-infrared transmittance. Protein content. Cereal Chem. Vol. 72, pp. 11–6, 1995.

[DDR07] A. Durand, O. Devos, C. Ruckebusch, J.P. Huvenne. Genetic algorithm optimisation combined with partial least squares regression and mutual information variable selection procedures in near-infrared quantitative analysis of cotton–viscose textiles. Analytica Chimica Acta. Vol. 595, pp. 72–79, 2007.

[DFA13] A. Dhyèvre, A. Foltete, D. Aran, S. Muller, S. Cotelle. Effets du PH du sol sur le test de génotoxicité Vicia-micronoyaux. Etude et Gestion des sols. Vol. 29, pp. 107-115, 2013.

[DFM10] A.A. Dias, G.S. Freitas, G.S.M. Marques, A. Sampaio, I.S. Fraga, M.AM. Rodrigues. Enzymatic saccharification of biologically pre-treated wheat straw with white-rot fungi. Bioresource technology. Vol. 101 N°6045, 2010.

[DGL10] N. Dupuy, O. Galtier, Y. Le Dréau, C. Pinatel, J. Kister, J.Artaud. Chemometric analysis of combined NIR and MIR spectra to characterize French olives. Lipid Sci. Technol. Vol. 112, pp. 463–475, 2010.

[DHK05] E. Dolezel-Horwath, T. Hutter, R. Kessler, R. Wimmer. Feedback and feedforward control of wet-processed hardboard production using spectroscopy and chemometric modelling. Analytica Chimica Acta. Vol. 544, pp. 47–59, 2005.

[DK92] S. Derksen, H. J. Keselman. Backward, forward and stepwise-automated subset selection algorithms: Frequency of obtaining authentic and noise variables. British Journal of Mathematical and Statistical Psychology. Vol. 45 N°2, pp. 265–282, 1992.

[DKR07] D. Djikanovic, A. Kalauzi, K. Radotic, C. Lapierre, M. Jeremic. Deconvolution of lignin fluorescence spectra: a contribu¬tion to the comparative structural studies of lignins. Russ J Phys Chem A. Vol. 81 N°9, pp. 1425–1428, 2007.

[DLS94] M.S. Dhanoa, S.J. Lister, R. Sanderson, R.J. Barnes, The link between Multiplicative Scatter Correction (MSC) and Standard Normal Variate (SNV) transformations of NIR spectra, J. Near Infrared Spectrosc. Vol. 2, pp. 43-47, 1994

[DOR05] J.M. Delgado, J.M. Orts, A. Rodes, ATR-SEIRAS study of the adsorption of acetate anions at chemically deposited silver thin film electrodes, Langmuir. Vol. 21, pp. 8809-8816, 2005.

[DOP08] J.M. Delgado, J.M. Orts, J.M. Perez, A. Rodes, Sputtered thin-film gold electrodes for in situ ATR-SEIRAS and SERS studies, Journal of Electroanalytical Chemistry. Vol. 617, pp. 130-140, 2008.

[DPC98] S.R. Delwiche, R.O. Pierce, O.K. Chung, B.W. Seabourn. Protein content of wheat by near-infrared spectroscopy of whole grain: Collaborative study. J AOAC Int. Vol. 81, pp. 587–603. 1998.

[DPP10] C. Dumas, S. Perez, E. Paul, X. Lefebvre. Combined thermophilic aerobic process and conventional anaerobic digestion: Effect on sludge biodegradation and methane production." Bioresource Technology. Vol. 101 N°8, pp. 2629-2636, 2010.

[DS81] N.R. Draper, H. Smith. Applied Régression Analysis (2nd Ed.), 1. USA: John Wiley &. Sons. 1981.

[DYX11] Y. W. Dong, S. Q. Yang, C. Y. Xu, Y. Z. Li, W. Bai, Z. N. Fan, Y. N. Wang, Q. Z. Li. Determination of Soil Parameters in Apple-Growing Regions by Near- and Mid-Infrared Spectroscopy. Pedosphere. Vol. 21 N°5, pp. 591-602, 2011.

[Efr60] A. M. Efroymson. Multiple regression analysis, in Mathematical methods for digital computers. Wiley, New York, 1960.

[EGJ04] L. Eriksson, J. Gottfries, E. Johansson, S. Wold. Time-resolved QSAR: an approach to PLS modelling of three-way biological data. Chemometr. Intell. Lab. Syst. Vol. 73 N°1, pp. 73–84, 2004.

[EHJ95] L. Eriksson, J.M. Hermens, E. Johansson, H.M. Verhaar, S. Wold. Multivariate analysis of aquatic toxicity data with PLS. Aquat. Sci. Vol. 57 N°3, pp. 217–241, 1995.

[ELB14] N. Eloutassi, B. Louaste, L. Boudine et A. Remmal. Valorisation de la biomasse lignocellulosique pour la production de bioéthanol de deuxième génération. Revue des Energies Renouvelables. Vol. 17 N°4, pp. 600 – 609, 2014.

[FGP14] D. Ferreira, O. Galao, J. Pallone, R. Poppi. Comparison and Application of Near-Infrared (NIR) and Mid–Infrared (MIR) Spectroscopy for Determination of Quality Parameters in Soybean Samples". Food Control. Vol. 35, pp. 227-232, 2014.

[FHX09] P. Fu, S. Hu, L. Sun, J. Xiang, T. Yang, A. Zhang, J. Zhang. Structural evolution of maize stalk/char particles during pyrolysis. Bioresource Technology. Vol. 100, pp. 4877–4883, 2009.

[FLC04] M. Forina, S. Lanteri, M.C. Cerrato Oliveros, C. Pizzaro Millan. Analytical and Bioanalytical Chemistry. Vol. 54 N°3, pp. 413-419, 2004.

[FLK03] H. Fang, S. Liang, A. Kuusk. Retrieving leaf area index using a genetic algorithm with a canopy radiative transfer model. Remote Sens. Vol. 85, pp. 257–270, 2003.

[FMM15] M. Farrés, B. Martrat, B.D. Mol, J.O. Grimalt, R. Tauler. Extraction of climatic signals from fossil organic compounds in marine sediments up to 11.7 Ma old (IODP-U1318). Anal. Chim. Acta. Vol. 879, pp. 1–9, 2015.

[FPT15] M. Farrés, S. Platikanov, S. Tsakovski, R. Tauler. Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. Chemometrics. 2015.

[FSG10] J.C. Frigon, R. Serge. R. Guiot. Biomethane production from starch and lignocellulosic crops: a comparative review. Biofuels, Bioproducts and Biorefining. Vol. 4 N°4, pp. 447–458, 2010.

[GA07] J. Ghasemi, S. Ahmadi. Combination of Genetic Algorithm and Partial Least Squares for cloud point prediction of nonionic surfactants from molecular structures. Annali di Chimica. Vol. 97, pp. 69-83, 2007.

[Gab08] B. Gabrielle. Intérêts et limites des biocarburants de première génération. Journal de la Société de Biologie. Vol. 202 N°161, 2008.

[GAF08] R. K. H. Galvão, M. C. U. Araújo, W. D. Fragoso, E. C. Silva, G. E. José, S. F. C. Soares, H. M. Paiva. A Variable Elimination Method to Improve the Parsimony of MLR Models Using the Successive Projections Algorithm. Chemometrics and Intelligent Laboratory Systems. Vol. 92 N°1, pp. 83–91, 2008.

[Gar96] J.L. Gardette. Caractérisation des polymères par spectrométrie optique. Techniques de l'ingénieur. Analyse et caractérisation. Vol. 5 N°3762, pp. 1-14, 1996.

[GBS13] A. Gholizadeh, L. Borůvka, M. Saberioon, R. Vašát. Visible, Near Infrared, and Mid-Infrared Spectroscopy Applications for Soil Assessment with Emphasis on Soil Organic Matter Content and Quality: State-of-the-Art and Key Issues. Applied Spectroscopy. Vol. 67 N°12, pp. 1349-1362, 2013.

[GDR02] L.E. Gollapalli, B.E. Dale, D.M. Rivers. Predicting digestibility of ammonia fiber explosion (AFEX)-treated rice straw. Appl Biochem Biotechnol. Vol. 98-100, pp. 23–35, 2002.

[GHJ99] D. Gillon, C. Houssard, R. Joffre. Using near-infrared reflectance spectroscopy to predict carbon, nitrogen and phosphorus content in heterogeneous plant material. Oecologia; Vol. 118, pp. 173-82, 1999.

[GK78] D. E. Gustafen, W. C. Kessel. Fuzzy clustering with a fuzzy covariance matrix. IEEE. pp: 761-766, 1978.

[GK86] P. Geladi, B. R. Kowalski. Partial least-squares regression: a tutorial. Analytica Chimica Acta. Vol. 185, pp. 1-17, 1986.

[GM87] B. George, P. McIntyre. Infrared spectroscopy. John Wiley & Sons: London, 1987.

[GMM85] P. Geladi, D. MacDougal, H. Martens. Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat. Applied Spectroscopy. Vol. 39, pp. 491-500, 1985.

[Gol89] D. E. Goldberg. Genetic Algorithms in Search, Optimization, and Machine Learning. Reading, MA: Addison-Wesley, 1989.

[HB99] T. M. Henriksen, T. A. Breland. Evaluation of criteria for describing crop residue degradability in a model of carbon and nitrogen turnover in soil. Soil Biology and Biochemistry. Vol. 31, pp.1135-1149, 1999.

[HDL08] E. A. Heaton, F. G. Dohleman, S. P. Long. Meeting US biofuel goals with less land: the potential of Miscanthus. Global Change Biology. Vol. 14, pp.2000-2014, 2008.

[HDS10] C. Heitner, D.R. Dimmel, J.A. Schmidt. Lignin and lignans: advances in chemistry. CRC, Boca Raton, pp 103–136, 2010.

[Her45] G. Herzberg. Molecular Spectra and Molecular Structure. In Infrared and Raman Spectra of Polyatomic Molecules. D. Van Nostrand Company, Inc., 1945.

[HFB03] M. Hoogwijka, A. Faaij, R. Broeka, G. Berndes, D. Gielen, W. Turkenburg. Exploration of the ranges of the global potential of biomass for energy. Biomass and Bioenergy. Vol. 25, pp. 119 – 133, 2003.

[Hin95] R. Hinterding. Gaussian Mutation and Self-adaption for Numeric Genetic Algorithms. IEEE, pp.384-389, 1995.

[HL10] S. Hou, L. Li. Rapid Characterization of Woody Biomass Digestibility and Chemical Composition Using Near-infrared Spectroscopy. Journal of Integrative Plant Biology. 2010.

[Hol89] J. Holland. Adaptation in natural and artificial systems. University of Michigan Press, 1975.

[Hoo11] H. Noorizadeh. Application of GA-PLS and GA-KPLS Calculations for the prediction of the retention indices of essential oils. Quim. Nova. Vol. 34 N°8, pp. 1398-1404, 2011

[Hos88] A. Höskuldsson. PLS regression methods. Journal of chemometrics. Vol. 2, pp. 221–228, 1988.

[HOS82] A. Hatta, T. Ohshima, W. Suëtaka. Observation of the enhanced infrared absorption of p-nitrobenzoate on Ag island films with an ATR technique, Applied Physics A: Materials Science & Processing. Vol. 29, pp. 71-75, 1982.

[HRD10] T.L.N. Huyen, C. Remond, R.M. Dheilly, B. Chabbert. Effect of harvesting date on the composition and saccharification of Miscanthus x giganteus. Bioresource Technology. Vol. 101 N°21, pp. 8224-8231, 2010.

[HS03] P.M. Hansen, J.K. Schjoerring. Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. Remote Sensing of Environment. Vol. 86, pp.542–553. 2003

[HSP09] B.B. Hallac, P. Sannigrahi, Y. Pu, M. Ray, R.J. Murphy, A.J. Ragauskas. Biomass characterization of Buddleja davidii: a potential feedstock for biofuel production. J Agric Food Chem. Vol. 57 N°4, pp. 1275– 1281, 2009.

[HZ09] A. Hendriks, G. Zeeman. Pretreatments to enhance the digestibility of lignocellulosic biomass. Bioresource Technology. Vol. 100 N°10, 2009.

[HTF09] T. Hastie, R. Tibshirani, J. H. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, second edition, 2009.

[HTS03] B.R. Hames, S.R. Thomas, A.D. Sluiter, C.J. Roth, D.W. Templeton. Rapid biomass analysis. Appl Biochem Biotechnol. Vol. 105, pp. 5–16. 2003

[IGG06] F. Inon, S. Garrigues, M. Guardia. Combination of mid- and near-infrared spectroscopy for the determination of the quality properties of beers. Analytica Chimica Acta. Vol. 571, pp. 167–174, 2006.

[IT00] T. Imae, H. Torii. In situ investigation of molecular adsorption on Au surface by surface-enhanced infrared absorption spectroscopy. Journal of Physical Chemistry B. Vol. 104, pp. 9218-9224, 2000.

[JA95] E. Johnson, R. Aroca. Surface-enhanced infrared spectroscopy of monolayers. The Journal of Physical Chemistry. Vol. 99, pp. 9325-9330, 1995.

[JAN13] L.J. Jönsson, B. Alriksson, N.O. Nilvebrant. Bioconversion of lignocellulose: inhibitors and detoxification. Biotechnology for Biofuels. Vol. 6 N°16, 2013.

[Jef04] N. O. Jeffries. Performance of a genetic algorithm for mass spectrometry proteomics. BMC Bioinformatics. Vol. 5, 2004.

[JGD92] R. Joffre, D. Gillon, P. Dardenne, R. Agneessens, R. Biston. The use of near-infrared reflectance spectroscopy in litter decomposition studies. Ann Sci For. Vol. 49, pp. 481-488, 1992.

[JOF06] B. Jaillais, M.A. Ottenhof, I.A. Farhat, D.N. Rutledge. Outer-product analysis (OPA) using PLS regression to study the retrogradation of starch. Vibrational Spectroscopy. Vol. 40, pp. 10–19, 2006.

[JPR05] B. Jaillais, R. Pinto, A.S. Barros, D.N. Rutledge. Outer-product analysis (OPA) using PCA to study the influence of temperature on NIR spectra of water. Vibrational Spectroscopy. Vol. 39, pp. 50–58, 2005.

[JVB10] N. Jacquet, C. Vanderghem, C. Blecker et M. Paquot. La Steam Explosion: Application en tant que Prétraitement de la Matière Lignocellulosique', Biotechnologie, Agronomie, Société et Environnement, Vol. 14 N°2, pp. 561 - 566, 2010.

[KBD09] P. Kumar, D.M. Barrett, M.J. Delwiche, P. Stroeve. Methods for Pretreatment of Lignocellulosic Biomass for Efficient Hydrolysis and Biofuel Production. Industrial & Engineering Chemistry Research. Vol. 48 N°3713, 2009.

[KCC12] D. J. Krasznai, P. Champagne, M. F. Cunningham. Quantitative characterization of lignocellulosic biomass using surrogate mixtures and multivariate techniques. Bioresource Technology. Vol. 110, pp. 652–661, 2012.

[Khe09] A. Khelfa. Etude des Etapes Primaires de la Dégradation Thermique de la Biomasse Lignocellulosique. Doctorat d'Etat, Université Paul Verlaine, Metz, France, 2009.

[KHL01] L. Kupper, H.M. Heise, P. Lampen, A.N. Davies, P. McIntyre. Authentication and quantification of extra virgin olive oils by attenuated total reflectance infrared spectroscopy using silver halide fiber probes and partial least-squares calibration. Applied Spectroscopy. Vol. 55 N°5, pp. 563-570, 2001.

[KHW10] L.M. Kline, D.G. Hayes, A.R. Womac, N. Labbe. Simplified determination of lignin content in hard and soft woods via UV¬spectrophotometric analysis of biomass dissolved in ionic liquids. BioResources. Vol. 5 N°3, pp. 1366–1383, 2010.

[KJ97] R. Kohavi, G. H. John. Wrappers for feature subset selection. Artif. Intell. Vol. 97 N°2, pp. 273–324, 1997.

[Koe75] J. L. Koenig. Application of Fourier Transform Infrared Spectroscopy to Chemical Systems. Applied Spectroscopy. Vol. 29 N°4, pp. 293-308, 1975.

[KPH08] J. Kundu, F. Le, P. Nordlander, N.J. Halas, Surface enhanced infrared absorption (SEIRA) spectroscopy on nanoshell aggregate substrates, Chemical Physics Letters. Vol. 452, pp. 115-119, 2008.

[KSA30] C.F. Kettering, L.W. Shultz, D.H. Andrews. A Representation of the Dynamic Properties of Molecules by Mechanical Models. Phys. Rev. Vol. 36, pp. 531, 1930.

[LAB92]. S. Lorilee, L. Arakaki, D. Burns, Multispectral Analysis for Quantitative Measurements of Myoglobin Oxygen Fractional Saturation in the Presence of Hemoglobin Interference, Applied Spectroscopy. Vol. 46, pp. 1919-1928, 1992.

[Lan83] E. Lanza. Determination of moisture, protein, fat, and calories in raw pork and beef by near infrared spectroscopy. J Food Sci. Vol. 48, pp. 471–4, 1983

[Lar11] P. J. Larkin. IR and Ramman Spectroscopy Principales and Spectral Interpretation. USA: Elsevier. 1st Edition, 2011.

[LBF10] G. Lorenzini, C. Biserni,G. Flacco. Solar Thermal and Biomass Energy. WIT Press. USA. 2010.

[LHG14] A. Largo-Gosens, M. Hernández-Altamirano, L. García-Calvo, A. Alonso-Simón, J. Álvarez, J.L. Acebes. Fourier transform mid infrared spectroscopy applications for monitoring the structural plasticity of plant cell walls. Plant science. Vol. 5 N°303, 2014.

[LHG07] J. Li, G. Henriksson, G. Gellerstedt. Lignin depolymerization/repolymerization and its critical role for delignification of aspen wood by steam explosion. Bioresource technology. Vol. 98 N°3061, 2007.

[LHV04] J. Luypaert, S. Heuerding, Y. Vander Heyden, D.L. Massart. The effect of preprocessing methods in reducing interfering variability from near-infrared measurements of creams. Journal of Pharmaceutical and Biomedical Analysis. Vol. 36, pp. 495–503, 2004.

[LJY09] H. Liu, B. Jeng, J. Yih, Y. Yu. Fuzzy C-Means Algorithm Based on Standard Mahalanobis Distances. Huangshan, P. R. China, pp. 422-427, 2009.

[LKY95] Y.W. Lai, E.K. Kemsley, R.H. Wilson. Potential of Fourier transform-infrared spectroscopy for the authentication of vegetable oils Journal of Agricultural and Food Chemistry. Vol. 42, pp. 1154-1159, 1995.

[LL98] P. Lantéri, R. Longeray. Chimiométrie : Outils du XXème siècle, méthode du XXIème siècle. Chemometrics from Experimental Design up to Data Processing. Vol. 26 N°8, pp. 13-78, 1998.

[LL11] K.A. Lê Cao, C. Le Gall. Integration and variable selection of omics data sets with PLS: a survey. Journal de la Société Française Statistiques. Vol. 152 N°2, 2011.

[LLC08] N. Labbé, S.H. Lee, H.W. Cho, M.K. Jeong, N. André. Enhanced discrimination and calibration of biomass NIR spectral data using non-linear kernel methods. Bioresour Technol. Vol. 99, pp. 45–52, 2008.

[LLX09] H. Li, Y. Liang, Q. Xu, D. Cao. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration, Anal Chim Acta. Vol. 648 N°1, pp. 77-84, 2009

[LPD86] P. Levillain, D. Pompeydie. Derivative spectrophotometry principles, advantages and limitations, applications, Analysis. Vol. 14, pp. 1-20, 1986.

[LR05] T.A. Lestander, C. Rhén. Multivariate NIR spectroscopy models for moisture, ash and calorific content in biofuels using bi-orthogonal partial least squares regression. Analyst. Vol. 130, pp. 1182–1189, 2005.

[LS15] R. Lal, B.A. Stewar. Soil-Specific Farming: Precision Agriculture. CRC Press. Vol. 431, pp. 119. 2015.

[LSD14] J. S Lupoi, S. Singh, M. Davis, D. J Lee, M. Shepherd, B. Simmons, R. J Henry. High-throughput prediction of eucalypt lignin syringyl/guaiacyl content using multivariate analysis: a comparison between mid-infrared, near-infrared, and Raman spectroscopies for model development. Biotechnology for Biofuels. 2014.

[Lud08] B. Ludwig. Use of mid-infrared spectroscopy in the diffuse-reflectance mode for the prediction of the composition of organic matter in soil and litter, J. Plant Nutr. Soil Sci. Vol.171, pp.384–391, 2008.

[LWZ08] Q. Liu, S. Wang, Y. Zheng, Z. Luo, K. Cen. Mechanism study of wood lignin pyrolysis by using TG–FTIR analysis. J. Anal. Appl. Pyrolysis. Vol. 82 N°1, pp. 170–177, 2008.

[LYA10] L. Liu, X. Ye, A.R. Womac, S. Sokhansanj. Variability of biomass chemical composition and rapid analysis using FT-NIR techniques. Carbohydrate Polymers. Vol. 81, pp. 820–829, 2010.

[LYS10] L. Liu, X. Philip Ye, A. M. Saxton, A. Womac. Pretreatment of near infrared spectral data in fast biomass analysis. Journal of Near Infrared Spectroscopy. Vol. 18 N°5, pp. 317–331, 2010.

[Mac67] J. B. MacQueen. Some Methods for classification and Analysis of Multivariate Observations. Proc. of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, Vol. 1, pp. 281–297, 1967.

[MAM91] T.M. McLellan, J.D. Aber, M.E. Martin. Determination of nitrogen, lignin, and cellulose content of decomposing leaf materials by near-infrared reflectance spectroscopy. Can J For Res. Vol. 21, pp. 1684-1688, 1991.

[Mar95] B. Markert. Sample preparation (cleaning, drying, and homogenization) for trace element analysis in plant matrices. Science of the Total Environment. Vol. 176, pp. 45-61, 1995.

[Mao03] K. Z. Mao. Orthogonal Forward Selection and Backward Elimination Algorithms for Feature Subset Selection. IEEE transactions on systems, man, and cybernetics-part B: cybernetics.2003.

[Mau02] S.L. Maunu. NMR studies of wood and wood products. Prog Nucl Magn Reson Spectrosc. Vol. 40 N°2, pp. 151–174, 2002.

[MB03] U. Maulik, S. Bandyopadhyay. Fuzzy Partitioning Using a Real-Coded Variable-Length Genetic Algorithm for Pixel Classification. IEEE Transactions on geoscience and remote sensing. Vol. 41 N°5, 2003.

[MBB11] G.E. Machinet, I. Bertrand, Y. Barriere, B. Chabbert, S. Recous. Impact of plant cell wall network on biodegradation in soil: Role of lignin composition and phenolic acids in roots from 16 maize genotypes. Soil Biology & Biochemistry. Vol. 43 N°7, pp. 1544-1552, 2011.

[MCB05] V. Mazet, C. Carteret, D. Brie, J. Idier, B. Humbert. Background removal from spectra by designing and minimising a non-quadratic cost function. Chemometrics and Intelligent Laboratory Systems. Vol. 76, pp. 121–33, 2005

[MC85]G. W. Milligan, M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. Psychometrika. Vol. 50, pp. 159–179, 1985.

[McC00] R.L. McCreery. Raman spectroscopy for chemical analysis. Wiley, New York, pp. 448, 2000.

[MCS02] W. McClure, B. Crowell, D. Stanfield, S. Mohapatra, S. Morimoto, G. Batten. Near infrared technology for precision environmental measurements: Determination of nitrogen in green-and dry-grasstissue .J. Near Infrared Spectrosc. Vol. 10 N°3, pp. 177–185, 2002.

[MHC07] R. Mutabaruka, K. Hairiah, G. Cadish. Microbial degradation of hydrolysable and condensed tannin polyphenol-protein complexes in soils from different land-use histories. Soil Biology and Biochemistry. Vol. 39, pp. 1479-1492, 2007.

[Mil02] A. Miller. Subset selection in regression, 2nd edition. Chapman & Hall/CRC, 2002.

[MJG83] H. Martens, S.A. Jensen, P. Geladi, Multivariate linearity transformations for near infrared reflectance spectroscopy, in: O.H.J. Christie (Editor), Proc. Nordic Symp. Applied Statistics, Stokkland Forlag, Stavanger, Norway. pp. 205–234, 1983.

[MLS12] T. Mehmood, K. H. Liland, L. Snipen, S. Sæbø. A review of variable selection methods in Partial Least Squares Regression. Chemometrics and Intelligent Laboratory Systems. Vol. 118, pp. 62–69, 2012.

[MMA91] T.M. McLellan, M.E. Martin, J.D. Aber, J.M. Melillo, K.J. Nadelhoffer, B. Dewey. Comparison of wet chemistry and near-infrared reflectance measurements of carbon-fraction chemistry and nitrogen concentration of forest foliage. Can J For Res. Vol. 21, pp. 1689-1693, 1991.

[MO01] A.K. Moore, N.L. Owen. Infrared spectroscopic studies of solid wood. Applied Spectroscopy Reviews. Vol. 36, pp. 65–86, 2001.

[MMS11] T. Mehmood, H. Martens, S. Sæbø, J. Warringer, L. Snipen. A Partial Least Squares algorithm for parsimonious variable selection. Algorithms for Molecular Biology. Vol. 6 N°27, 2011

[Mou67] C. Moureaux. Influence de la température et l'humidité sur les activités biologiques de quelques sols ouest-africains. Cah. O.R.S.T.O.M., sér. Pédol. Vol. 5 N°4, pp. 393-420, 1967.

[MKB79] K.V. Mardia, J.T. Kent, J.M. Bibby. Multivariate Analysis. Academic Press, 1979.

[MKD98] N.A. Marigheto, E.K. Kemsley, M. Defernez, R.H. Wilson, J. Am. Comparison of mid-infrared and raman spectroscopies for the authentication of edible oils Journal of the American Oil Chemists Society. Vol. 75, pp. 987-992, 1998.

[MR93] A. Mackiewicz, W. Ratajczak. Principal Components Analysis (PCA). Computers & Geosciences. Vol. 19 N°3, pp. 303-342, 1993.

[MRR02] G. W. McCarty, J. B. Reeves III, V. B. Reeves, R. F. Follett, J. M. Kimble. Mid-Infrared and Near-Infrared Diffuse Reflectance Spectroscopy for Soil Carbon Measurement. Soil Science Society of America Journal. Vol. 66, pp. 640–646, 2002.

[MWD05] N. Mosier, C. Wyman, B. Dale, R. Elander, Y.Y. Lee, M. Holtzapple. Features of promising technologies for pretreatment of lignocellulosic biomass. Bioresource Technology. Vol. 96 N°673, 2005.

[MVB97] D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, P. J. Lewi. Handbook of chemometric and Qualimetrics: part A. Elsevier Science. Vol. 20, 1997

[NB76] R. Norton, H. Reinhard Beer. New Apodizing Functions for Fourier Spectrometry. Journal of the Optical Society of America. Vol. 66 N°3, pp. 259–264, 1976.

[NCB92] J. Niemeyer, Y. Chen, J. M. Bollay. Characterization of humic acids, composts, and peat by diffuse reflectance fourier-transform infrared spectroscopy. Soil Sci.Soc.Am. Vol. 56, pp. 135-140. 1992.

[NDS10] K. Nkansah, B. Dawson-Andoh, J. Slahor. Rapid characterization of biomass using near infrared spectroscopy coupled with multivariate data analysis. Bioresource Technology. Vol. 101, pp. 4570–4576, 2010.

[NFA93] Y. Nishikawa, K. Fujiwara, K. Ataka, M. Osawa, Surface-enhanced infrared external reflection spectroscopy at low reflective surfaces and its application to surface analysis of semiconductors, glasses, and polymers, Analytical Chemistry. Vol. 65, pp. 556-562, 1993.

[NF77] P. M. Narendra, K. Fukunaga. A branch and bound algorithm for feature subset selection. IEEE Trans. Comput. Vol. 26, pp. 917–922, 1977.

[NGR10] S.N. Naik, V.V. Goud, P.K. Rout, A.K. Dalai. Production of first and second generation biofuels: A comprehensive review. Renewable and Sustainable Energy Reviews. Vol. 14 N°578, 2010.

[NSW00] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen. Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. Applied Spectroscopy. Vol. 54 N°3, pp. 413-419, 2000.

[OD06] M.J. O'donohue, P. Debeire, Fractionnement de la biomasse lignocellulosique en synthon, la chimie vert, Lavoisier, 2006.

[OEP12] A. Oussama, F. Elabadi, S. Platikanov, F. Kzaiber, R. Tauler. Detection of Olive Oil Adulteration Using FT-IR Spectroscopy and PLS with Variable Importance of Projection (VIP) Scores. Journal of the American Oil Chemists' Society. Vol. 89 N°10, pp. 1807-1812, 2012.

[OHA03] K. Ono, M. Hiraide, M. Amari. Determination of lignin, holocellulose, and organic solvent extractives in fresh leaf, litterfall, and organic material on forest floor using nearinfrared reflectance spectroscopy. J For Res. Vol. 8, pp. 191-198, 2003.

[OIU91] K. Masatoshi Osawa, Masahiko Ikeda, Hiroshi Uchihara, Ryujiro Nanba, Surface Enhanced Infrared Absorption Spectroscopy: Mechanism and application to trace analysis, in, Analytical Sciences. pp. 503-506, 1991.

[OMS92]. Y. Ozaki, T. Miura, K. Sakurai, T. Matsunaga, Nondestructive Analysis of Water Structure and Content in Animal Tissues by FT-NIR Spectroscopy with Light-Fiber Optics. Part I: Human Hair, Applied Spectroscopy. Vol. 46, pp. 875-878, 1992.

[PCL14] S. A. Parsons, R. A. Congdon, I. R. Lawler. Determinants of the pathways of litter chemical decomposition in a tropical region. New Phytologist. Vol. 203, pp. 873–882, 2014.

[PFL00] O. Polgár, M. Fried, T. Lohner, I. Bársony. Comparison of algorithms used for evaluation of ellipsometric measurements: Random search, genetic algorithms, simulated annealing and hill climbing graph-searches. Surface Science. Vol. 457, pp. 157-177, 2000.

[PFN13] M.C. Popescu, J. Froidevaux, P. Navi, C.M. Popescu. Structural modifications of Tilia cordata wood during heat treatment investigated by FT-IR and 2D IR correlation spectroscopy. Journal of Molecular Structure. Vol. 1033, pp. 176-186, 2013.

[Pic10] S. Picek. Comparison of a Crossover Operator in Binary-coded Genetic Algorithms, WSEAS Trans. on Computers. Vol. 9 N°9, pp. 1064-1073, 2010.

[PNW10] A. Pucci, F. Neubrech, D. Weber, S. Hong, T. Toury, M.L. de la Chapelle, Surface enhanced infrared spectroscopy using gold nanoantennas, Physica Status Solidi B-Basic Solid State Physics. Vol. 247, pp. 2071-2074, 2010.

[PP13] C. Popescu, M. Popescu. A near infrared spectroscopic study of the structural modifications of lime (Tilia cordata Mill.) wood during hydrothermal treatment. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy. Vol. 115, pp. 227–233, 2013.

[PVS09] A. Prechtel, M. von Lützow, B. Schneider, O. Bens, C. G. Bannick, I. Kögel-Knabner, R. F. Hüttl. Organic carbon in soils of Germany: Status quo and the need for new data to evaluate potentials and trends of soil carbon sequestration. Journal of Plant Nutrition and Soil Science. Vol 172 N° 5, pp. 601-614, 2009.

[PW15] C. Payne, E. J Wolfrum. Rapid analysis of composition and reactivity in cellulosic biomass feedstocks with near infrared spectroscopy. Biotechnol Biofuels. Vol. 8 N°43, 2015.

[PYC04] H. Park, S. Yoo, S. Cho. A Fuzzy Clustering Algorithm for Analysis of Gene Expression Profiles. Trends in Artificial Intelligence Lecture Notes in Computer Science. Vol. 3157, pp 967-968, 2004.

[RB07] D.N. Rutledge, D. Bouveresse. Multi-way analysis of outer product arrays using PARAFAC. Chemometrics and Intelligent Laboratory Systems. Vol. 85, pp. 170–178, 2007.

[RCM07] Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, N. Jent. A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. J. Pharm. Biomed. Anal. Vol. 44 N°3, pp. 683–700, 2007.

[RD97] J.B. Reeves III, S.R. Delwiche. Determination of protein in ground wheat samples by mid-infrared diffuse reflectance spectroscopy. Applied Spectroscopy. Vol. 51 N°8, pp. 1200–1204, 1997.

[Ree94] J.B. Reeves III. Near versus mid-infrared diffuse reflectance spectroscopy for the quantitative determination of the composition of forages and by-products. Journal of Near Infrared Spectroscopy. Vol. 2, pp. 49–57, 1994

[ReeIII94] J.B Reeves III. Near versus mid-infrared spectroscopy for the quantitative analysis of chlorite treated forages and by-products. Journal of Near Infrared Spectroscopy. Vol. 2, pp. 153–162, 1994.

[Ree96] J.B. Reeves III. Improvement in Fourier near- and mid-infrared diffuse reflectance spectroscopic calibrations through the use of a sample transport device. Applied Spectroscopy. Vol. 50 N°8, pp. 965–969, 1996.

[Ree10] J. B. Reeves III. Near-versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis. Geoderma. Vol. 158, pp. 3–14. 2010.

[Rez10] B. Rezaee. A cluster validity index for fuzzy clustering", Fuzzy Sets and Systems. Vol. 161, pp. 3014-3025, 2010.

[RJR09] M. Rajendra, P. C. Jena, H. Raheman. Prediction of optimized pretreatment process parameters for biodiesel production using ANN and GA. Fuel. Vol. 88, pp. 868–875, 2009.

[RL10] J. Ralph, L.L. Landucci. NMR of lignins. In: Heitner C, Dimmel DR, Schmidt JA (eds) Lignin and lignans: advances in chemistry. CRC, Boca Raton, pp. 137–244, 2010.

[RWM06] R.A. Viscarra Rossel, D.J.J. Walvoort, A.B. McBratney, L.J. Janik, J.O. Skjemstad. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma. Vol. 131, pp. 59 – 75, 2006.

[Rob87] A. I. Robertson. Decomposition of mangrove leaf litter in tropical Australia. Journal of Experimental Marine Biology and Ecology. Vol. 116, pp. 235-247, 1987.

[RPB14] A. Rammal, E. Perrin, B. Chabbert, I. Bertrand, B. Lecart, V. Vrabie. Using a Genetic Algorithm as an Optimal Band Selector in the Mid-Near Infrared: Evaluation of the Biodegradation of Maize Roots. Journal of Applied Science and Agriculture. Vol. 9 N°11, pp. 382-388, 2014.

[RFG15] A. Rammal, H. Fenniri, A. Goupil, V. Vrabie, I. Bertrand, B. Chabbert. Feature Selection Based On Weighted Distance Minimization, Application to Biodegradation Evaluation. The 23rd European Signal Processing Conference (EUSIPCO), Nice-France, 31 Aout- 04 Septembre 2015.

[RPC13] A. Rammal, E. Perrin, B. Chabbert, I. Bertrand, G. Mihai, V. Vrabie. Optimal preprocessing of Mid InfraRed spectra. Application to classification of lignocellulosic biomass: maize roots and miscanthus internodes. International Conference on Mass Data Analysis of Images and Signals (MDA), New York, July 2013.

[RPC15] A. Rammal, E. Perrin, B. Chabbert, I. Bertrand, B. Lecart, V. Vrabie. Evaluation of Lignocellulosic Biomass Degradation by Combining Mid- and Near-Infrared Spectra by the Outer Product and Selecting Discriminant Wavenumbers by a Genetic Algorithm. Applied Spectroscopy, Vol 69, N°11, pp. 1303-1312, 2015.

[RPV14] A. Rammal, E. Perrin, V. Vrabie, I. Bertrand, A. Habrant, B. Chabbert. Optimal Preprocessing And FCM Clustering Of MIR, NIR And Combined MIR-NIR Spectra For Classification Of Maize Roots. International Conference on e-Technologies and Networks for Development (IEEE – ICeND 2014). Beirut-Lebanon, 29 April-01 May, 2014

[RPV15] A. Rammal, E. Perrin, V. Vrabie, I. Bertrand, B. Chabbert. Weighted-Covariance Factor Fuzzy C-Means Clustering. International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (IEEE TAEECE 2015). Beirut-Lebanon, 29 April-01 May, 2015.

[RRG12] O.M. Rivera-Borroto, M. Rabassa-Gutiérrez, C. Grau-Ábalo Rdel, Y. Marrero-Ponce, J.M. García-de la Vega. Dunn's index for cluster tendency assessment of pharmacological data sets, Can. J. Physiol. Pharmacol. Vol. 90, pp. 425–433, 2012.

[RSC92] V. U. S. Rao, G. J. Stiegel, G. J. Cinquegrane, R.D. Srivastava. Iron-based catalysts for slurry-phase Fischer-Tropsch process: Technology review. Fuel Processing Technology. Vol. 30 N°1, pp. 83-107, 1992.

[RVE09] A. Rinnan, F. van den Berg, S.B. Engelsen. Review of the most common pre-processing techniques for near infrared spectra. Trends in analytical chemistry. Vol. 28 N°10, pp. 1201-1222, 2009.

[RZ13] A. Ranjini, B. Zoraida. Analysis of Selection Schemes for Solving Job Shop Scheduling Problem Using Genetic Algorithm. IJRET: International Journal of Research in Engineering and Technology. Vol. 2 N°11, pp. 2319-1163, 2013.

[SAC96] M. Sanderson, F. Agblevor, M. Collins, D. Johnson. Compositional analysis of biomass feedstocks by near infrared reflectance spectroscopy. Biomass Bioenergy. Vol. 11 N°5, pp. 365–90, 1996.

[Sar97] M. Sarkar. A clustering Algorithm using an evolutionary programming based approach. Pattern Recognition Letters. Vol.18, pp. 975-968, 1997.

[Sch10] J.A. Schmidt. Electronic spectroscopy of lignins. In: C. Heitner, D.R. Dimmel, J.A. Schmidt. Lignin and lignans: advances in chemistry. CRC, Boca Raton, pp. 49–102, 2010.

[SCK15] M. Szymanska-Chargot, M. Chylinska, B. Kruk, A. Zdunek. Combining FT-IR spectroscopy and multivariate analysisfor qualitative and quantitative analysis of the cell wallcomposition changes during apples development. Carbohydrate Polymers. Vol. 115, pp. 93–103, 2015.

[SD05] W. Smith, G. Dent. Modern Raman spectroscopy. Wiley, Chichester. 2005.

[SG64] A. Savitzky, M. J. E. Golay. Smoothing and Differentation of Data by Simplified Least Squares Procedure, Analytical Chemistry. Vol. 36, pp. 1627-1639, 1964.

[SHN98] D.A. Skoog, F.J. Holler, T.A. Nieman. Principles of instrumen¬tal analysis, 5th edn. Harcourt, Brace, & Company, Philadelphia. 1998.

[Soc80] G. Socrates. Infrared Characteristic Group Frequencies. John Wiley & Sons Ltd., 1980

[SP72] G. Stotzky, D. Pramer. Activity, Ecology, and Population Dynamics of Microorganisms in Soil. CRC Critical Reviews in Microbiology. Vol. 2 N°1, pp. 59-137, 1972.

[SRF11] M. Schwanninge, J. Rodrigues, K. Fackler. A review of band assignments in near infrared spectra of wood and wood components. Journal of near Infrared Spectroscopy. Vol. 19, pp. 287-308. 2011.

[SRM08] P. Sannigrahi, A.J. Ragauskas, S.J. Miller. Effects of two-stage dilute acid pretreatment on the structure and composition of lignin and cellulose in loblolly pine. BioEnergy Research. Vol. 1 N°205, 2008.

[SRP04] M. Schwanninger, J.C. Rodrigues, H. Pereira, B. Hinterstoisser. Effects of short-time vibratory ball milling on the shape of FT-IR spectra of wood and cellulose. Vibrational Spectroscopy. Vol. 36 N°1, pp. 23-40, 2004.

[SSK04] S. Saranwong, J. Sornsrivichai, S. Kawano. Prediction of ripe-stage eating quality of mango fruit from its harvest quality measured nondestructively by near infrared spectroscopy. Postharvest Biol Technol. Vol. 31, pp. 137–145, 2004.

[SS89] W. Siedlecki, J. Skansky. A note on genetic algorithms for large-scale feature selection. Pattern Recognit. Lett. Vol. 10 N°5, pp. 335–347, 1989.

[SSS03] C. Solberg, E. Saugen, L.P. Swenson, L. Bruun, T. Isaksson,. Determination of fat in live farmed Atlantic salmon using non-invasive NIR techniques. J. Sci. Food Agric. Vol. 83 N°7, pp. 692–696, 2003.

[STB12] D. Sabatier, L. Thuriès, D. Bastianelli, P. Dardenne. Rapid prediction of the lignocellulosic compounds of surgarcane biomass by near infrared reflectance spectroscopy: comparing classical and independent cross validation. Journal of Near Infrared Spectroscopy. Vol. 20, pp. 371-385, 2012.

[STW05] K. Stehfest, J. Toepel, C. Wilhelm. The application of micro-FTIR spectroscopy to analyze nutrient stress-related changes in biomass composition of phytoplankton algae. Plant Physiology and Biochemistry. Vol. 43, pp. 717–726, 2005.

[SW00] M.E. Smith, G.A. Webb. Solid state NMR. Nucl Magn Reson. Vol. 29, pp. 251–315, 2000.

[SVP06] M. Skurichina, S. Verzakov, P. Paclik, R.P.W. Duin, 2006. Effectiveness of Spectral Band Selection Extraction Techniques for Spectral Data. Structural, Syntactic, and Statistical Pattern Recognition, Lecture Notes in Computer Science. Vol. 4109, pp. 541-550, 2006.

[Sus04] M. Sushmita. An evolutionary rough partitive clustering. Pattern Recognition Letters. Vol. 25, pp. 1439-1449, 2004

[TBJ09] I.K. Thomsen, S. Bruun, L.S. Jensen, B.T. Christensen. Assessing soil carbon lability by near infrared spectroscopy and NaOCl oxidation. Soil Biology and Biochemistry. Vol. 41, pp. 2170-2177, 2009.

[Ten98] M. Tenenhaus, La régression PLS: Théorie et pratique. Edition Technip. Paris 1998.

[TLJ09] A. Torbjörn, B. Lestander, B. Johnsson, M. Grothage. NIR techniques create added values for the pellet and biofuel industry. Bioresource Technology. Vol. 100, pp. 1589–1594, 2009.

[TSW05] M. Taniguchi, H. Suzuki, D. Watanabe, K. Sakai, K. Hoshino, T. Tanaka. Evaluation of pretreatment with Pleurotus ostreatus for enzymatic hydrolysis of rice straw. Journal of bioscience and bioengineering. Vol. 100 N°637, 2005.

[TSK02] A. Tay, R. K. Singh, S. S. Krishnan, J. P. Gore. Authentication of olive oil adulterated with vegetable oils using Fourier transform infrared spectroscopy. Lebensmittel Wissenschaft und-Technologie. Vol. 35, pp. 99-103, 2002.

[TSH09] D.W. Templeton, A.D. Sluiter, T.K. Hayward, Hames BR, Thomas SR. Assessing corn stover composition and sources of variability via NIRS. Cellulose. Vol. 16, pp. 621–639, 2009

[TWP14] J. M. Triolo, A. J. Ward, L. Pedersen, M. M. Løkke, H. Qu, S. G. Sommer. Near Infrared Reflectance Spectroscopy (NIRS) for rapid determination of biochemical methane potential of plant biomass. Applied Energy. Vol. 116, pp.52–57, 2014

[TYL07] H. Tian, Y. Ying, H. Lu, X. Fu, H. Yu. Measurement of soluble solids content in watermelon by Vis/NIR diffuse transmittance technique. Journal of Zhejiang University - Science B. Vol. 8 N°2, pp. 105-110, 2007.

[UGS12] S. Ullah, T. A. Groen, M. Schlerf, A. K. Skidmore, W. Nieuwenhuis, C. Vaiphasa. Using a Genetic Algorithm as an Optimal Band Selector in the mid and Thermal Infrared (2.5–14 μm) to Discriminate Vegetation Species, Sensors.Vol.12, pp. 8755-8769, 2012.

[Van63] P.J. Van Soest. Use of detergents in the analys of fibrous feeds. II. A rapid method for the determination of fiber and lignin. Journal of the A.O.A.C. Vol. 46, pp. 829-835, 1963.

[VB10] R.A. Viscarra Rossel, T. Behrens. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma. Vol. 158, pp. 46–54, 2010.

[Ver11] R. Verma. Genetic Algorithm for Multiprocessor Task Scheduling, IJCSMS International Journal of Computer Science and Management Studies. Vol. 11 N°2, pp.181-185, 2011.

[VLT14] M. Vohland, M. Ludwig, S. Thiele-Bruhn, B. Ludwig. Determination of soil properties with visible to near- and mid-infrared spectroscopy: Effects of spectral variable selection. Geoderma. Vol. 223, pp. 88–96, 2014

[VRL91] P.J. Van Soest, J.B. Robertson, B.A. Lewis. Methods for Dietary Fiber, Neutral Detergent Fiber, and Nonstarch Polysaccharides in Relation to Animal Nutrition. Journal of Dairy Science. Vol. 74 N°10, pp. 3583–3597, 1991.

[VSP06] N. Vlachos, Y. Skopelitis, M. Psaroudaki, V. Konstantinidou, A. Chatzilazarou, E. Tegou. Applications of Fourier transform-infrared spectroscopy to edible oils. Analytica Chimica Acta. Vol. 573, pp. 459–465, 2006.

[Whi94] D. Whitley. A Genetic Algorithm Tutorial. Statistics and Computing. Vol. 4, pp. 65-85, 1994. [Wil08] P. Williams. Near-infrared technology getting the best out of light. A short course in the practical implementation of near-infrared spectroscopy for the user. In: A Short Course Heldin Conjunction with the 13th ANISG Conference. Australian Near Infrared Spectroscopy Group, Department of Primary Industries—Hamilton Centre. VIC, Canada. 2008.

[WJC93] S. Wold, A. Johansson, M. Cochi. PLS-partial least squares projections to latent structures. ESCOM Science Publishers: Leiden, pp. 523–550, 1993.

[WL99] D.L. Wetzel, S.M. LeVine. Imaging molecular chemistry with infrared microscopy. Science. Vol. 285 N°5431, pp. 1224-1225, 1999.

[WM83] S. Wold, H. Martens, H. Wold. The multivariate calibration method in chemistry solved by the PLS method. In Lecture Notes in Mathematics. Vol. 973, pp. 286-293. 1983

[WMN14] Y. Wang, M. Mei, Y. Ni, S. Kokot. Combined NIR/MIR analysis: A novel method for the classification of complex substances such as Illicium verum Hook. F. and its adulterants. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy. Vol. 130, pp. 539–545, 2014.

[WMW98] H.D. Wanzenbock, B. Mizaikoff, N. Weissenbacher, R. Kellner, Surface enhanced infrared absorption spectroscopy (SEIRA) using external reflection on low-cost substrates, Fresenius Journal of Analytical Chemistry. Vol. 362, pp. 15-20, 1998

[WN01] P. Williams, K. Norris (Eds.). Near-infrared technology in the agricultural and food industries, 2nd ed. St. Paul, Minn.: American Association of Cereal Chemists. 2001.

[WN87] P.C. Williams, K. Norris. Near Infrared Technology in the Agricultural and Food Industries. American Assoc. of Cereal Chemist, Inc., St. Paul, Minnesota, USA, 1987.

[Wol66] H. Wold. Estimation of Principal Components and Related Models by Iterative Least Squares. Multivariate Analysis, New York: Academic Press. 1966.

[Wol95] S. Wold. PLS for multivariate linear modeling. Chemometric methods in molecular design. Vol. 2, 1995.

[Wor01] J.J. Workman. Infrared and Raman spectroscopy in paper and pulp analysis. Applied Spectroscopy Rev. Vol. 36 N°3, pp. 139–168, 2001.

[WS09] E.J. Wolfrum, A.D. Sluiter. Improved multivariate calibration models for corn stover feedstock and dilute-acid pretreated corn stover. Cellulose. Vol. 16, pp. 567–576, 2009.

[WSK08] F. Westad, A. Schmidt, M. Kermit. Incorporating chemical band-assignment in near infrared spectroscopy regression models. Journal of near infrared spectroscopy. Vol. 16, pp. 265-273, 2008. [WSN14] B. K. Waruru, K. D. Shepherd, G. M. Ndegwa, P. T. Kamoni, A. M. Sila. Rapid estimation of soil engineering properties using diffuse reflectance near infrared spectroscopy. biosystems engineering. Vol. 121, pp. 77-85, 2014.

[WSE01] S. Wold, M. Sjöström, L. Eriksson. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems. Vol. 58 N°2, pp. 109–130, 2001.

[XJP10] Z. Xiaobo, Z. Jiewen, M. J.W. Povey, M. Holmes, M. Hanpin. Variables selection methods in near-infrared spectroscopy. Analytica Chimica Acta. Vol. 667, pp. 14–32, 2010.

[XLD12] X. Zhao, L. Zhang, D. Liu. Biomass recalcitrance. Part I: the chemical compositions and physical structures affecting the enzymatic hydrolysis of lignocellulose. Biofuels, Bioprod. Bioref. 2012.

[Xu10] F. Xu. Structure, ultrastructure, and chemical composition. In Cereal Straw as a Resource for Sustainable Biomaterials and Biofuels: Chemistry, Extractives, Lignins, Hemicelluloses and Cellulose. London, Elsevier, pp. 9-47, 2010.

[XXW12] R. Xu, J. Xu, D. Wunsch. A Comparison Study of Validity Indices on Swarm-Intelligence-Based Clustering. IEEE Transactions on systems, man, and cybernetic. Vol. 42, pp. 1243-1256, 2012.

[XYT13] F. Xu, J. Yu, T. Tesso, F. Dowell, D. Wang. Qualitative and quantitative analysis of lignocellulosic biomass using infrared techniques: A mini-review. Applied Energy. Vol. 104, pp. 801–809, 2013.

[YH98] J. Yang and V. Honavar. Feature subset selection using a genetic algorithm. IEEE Intell. Syst. Vol. 13, pp. 44–49, 1998.

[YSC11] D. Yi-Wei, Y. Shi-Qi, X. Chun-Ying, L. Yu-Zhong, B. Wei, F. Zhong-Nan, W. Ya-Nan, L. Qiao-Zhen. Determination of Soil Parameters in Apple-Growing Regions by Near- and Mid-Infrared Spectroscopy. Pedosphere. Vol. 21 N°5, pp. 591–602, 2011.

[YW04] B. Yang, C.E. Wyman. Effect of xylan and lignin removal by batch and flowthrough pretreatment on the enzymatic digestibility of corn stover cellulose. Biotechnology and Bioengineering. Vol. 86 N°88, 2004.

[YYS14] M. Yang, Y. Yang, T. Su, 2014. An Efficient Fitness Function in Genetic Algorithm Classifier for Landuse Recognition on Satellite Images. The Scientific World Journal. 2014.

[YZL06] J. Yang, X. Zhao, X. Liu, C. Wang, P. Gao, J. Wang, L. Li, J. Gu, S. Yang, G. Xu. High performance liquid chromatographymass spectrometry for metabonomics: potential biomarkers for acute deterioration of liver function in chronic hepatitis B. J. Proteome Res. Vol. 5 N°3, pp. 554–561, 2006.

[Zel12] I. Zelinka. Handbook of Optimization: From Classical to Modern Approach, Springer, 2012.

[ZLS09] A. Zhang, F. Lu F, R. Sun, J. Ralph. Ferulate-coniferyl alcohol cross-coupled products formed by radical coupling reactions. Planta. Vol. 229 N°5, pp. 1099–1108, 2009.

[ZLW12] K. Zheng, Q. Li, J. Wang, J. Geng, P. Cao, T. Sui, X. Wang, Y. Du. Stability competitive adaptive reweighted sampling (SCARS) and its applications to multivariate calibration of NIR spectra. Chemometrics and Intelligent Laboratory Systems. Vol. 112, pp. 48–54, 2012.

[ZWH09] Z. Zhang, F Wang, B. Harrington. Two-Dimentional Mid- and Near-Infrared Correlation Spectroscopy for Rhubarb Identification. 2nd Int. Congress CISP. 2009.

[ZZL12] X. Zhao, L. Zhang, D. Liu. Biomass recalcitrance. Part I: the chemical compositions and physical structures affecting the enzymatic hydrolysis of lignocellulose. Biofuels, Bioprod. Bioref. 2012.

Annexes

Annexe du chapitre 2

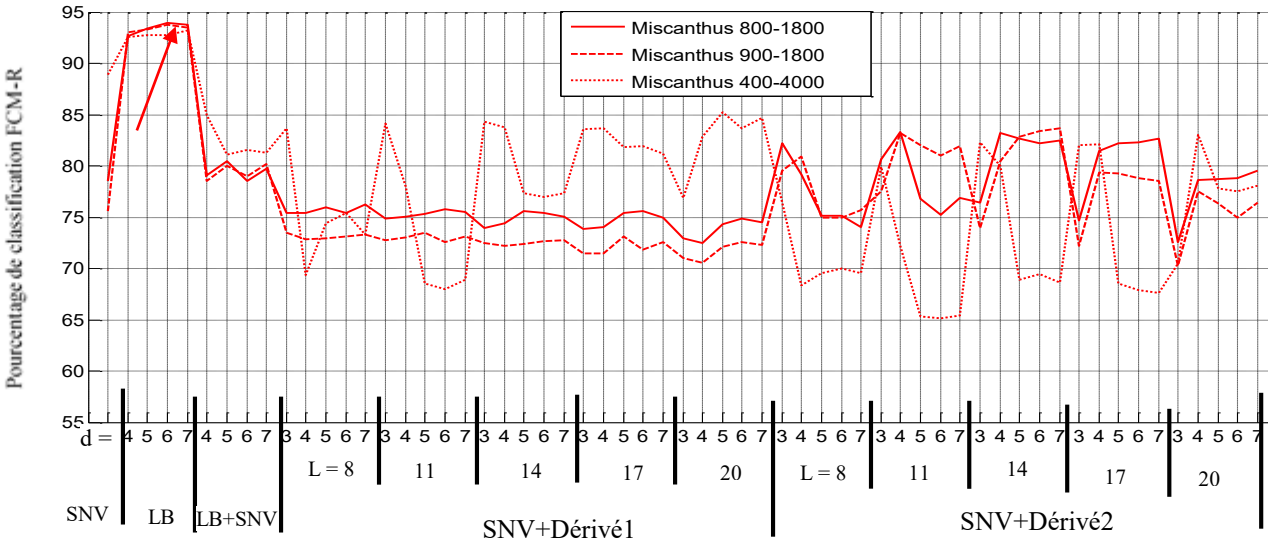


Figure A.2.1. Pourcentages de bonnes classifications. FCM-R-bootstrap appliqué sur les spectres MIR enregistrés sur des échantillons de miscanthus. Gammes spectrales 800 –1800 cm^{-1} (continu), 900 – 1800 cm^{-1} (discontinu), 400 - 4000 cm^{-1} (pointillé) pour différentes méthodes de prétraitements.

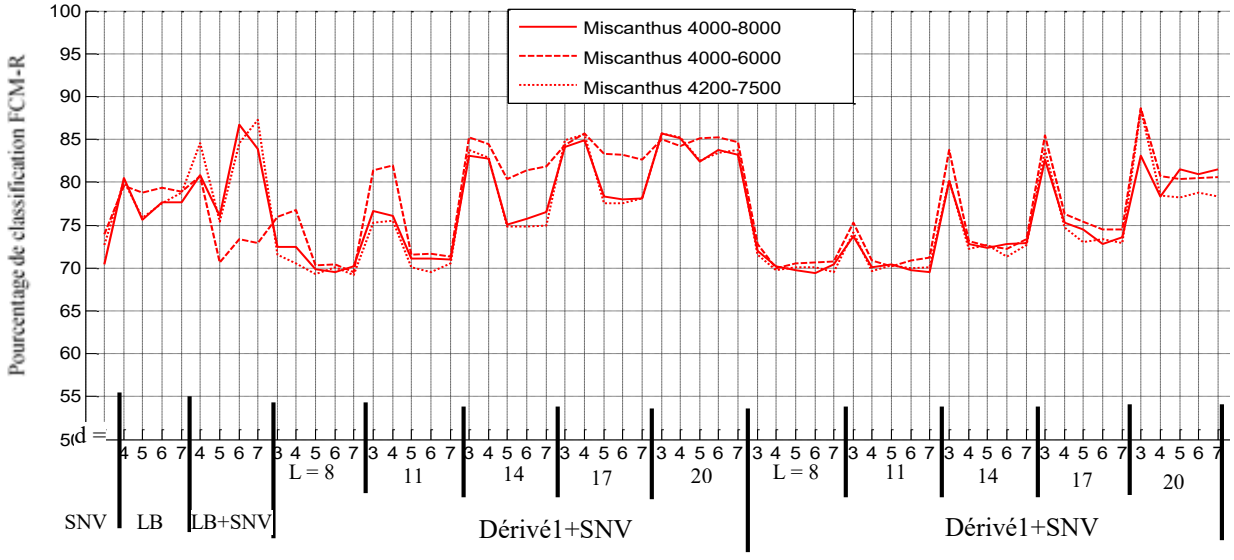


Figure A.2.2.Pourcentages de bonnes classifications. FCM-R-bootstrap appliqué sur les spectres NIR enregistrés sur des échantillons de miscanthus. Gammes spectrales 4000 –8000 cm^{-1} (continu), 4000 – 6000 cm^{-1} (discontinu), 4200 - 7500 cm^{-1} (pointillé) pour différentes méthodes de prétraitements

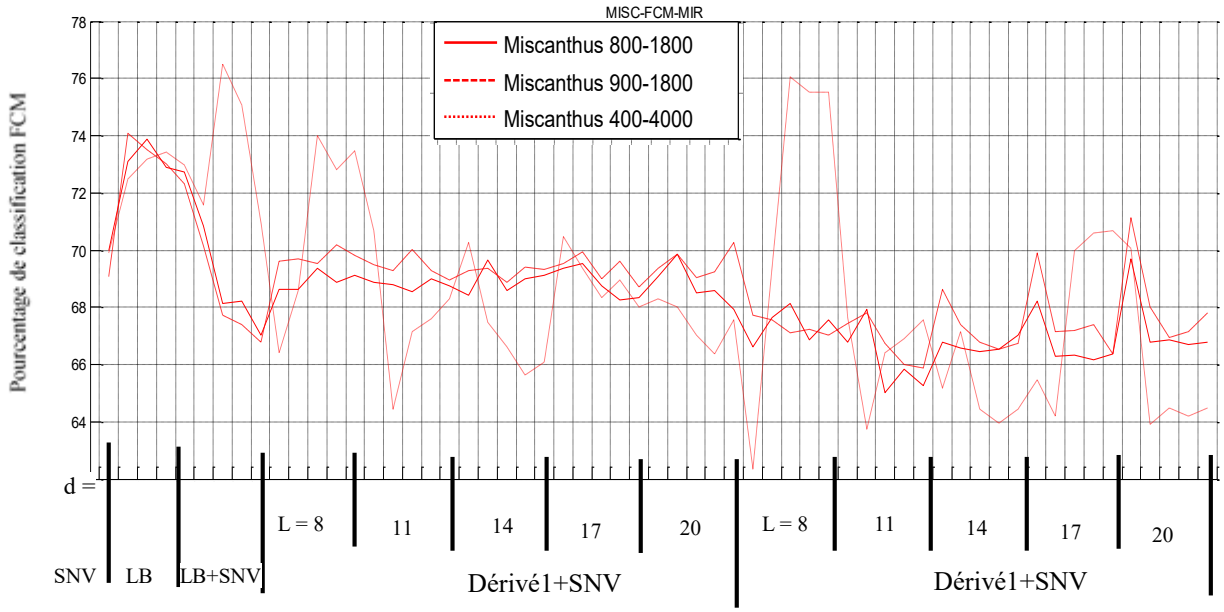


Figure A.2. 3.Pourcentages de bonnes classifications. FCM bootstrap appliqué sur les spectres MIR enregistrés sur des échantillons de miscanthus. Gammes spectrales 800 –1800 cm^{-1} (continu), 900 – 1800 cm^{-1} (discontinu), 400 - 4000 cm^{-1} (pointillé) pour différentes méthodes de prétraitements.

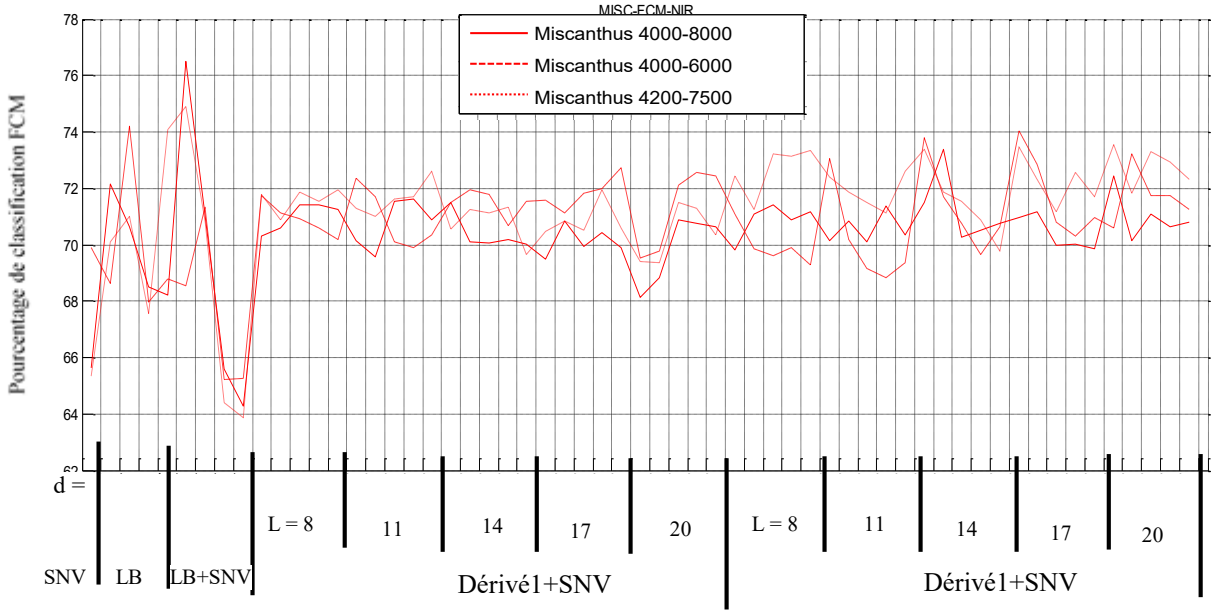


Figure A.2.4.Pourcentages de bonnes classifications. FCM bootstrap appliqué sur les spectres NIR enregistrés sur des échantillons de miscanthus Gammes spectrales 4000 –8000 cm^{-1} (continu), 4000 – 6000 cm^{-1} (discontinu), 4200 - 7500 cm^{-1} (pointillé) pour différentes méthodes de prétraitements

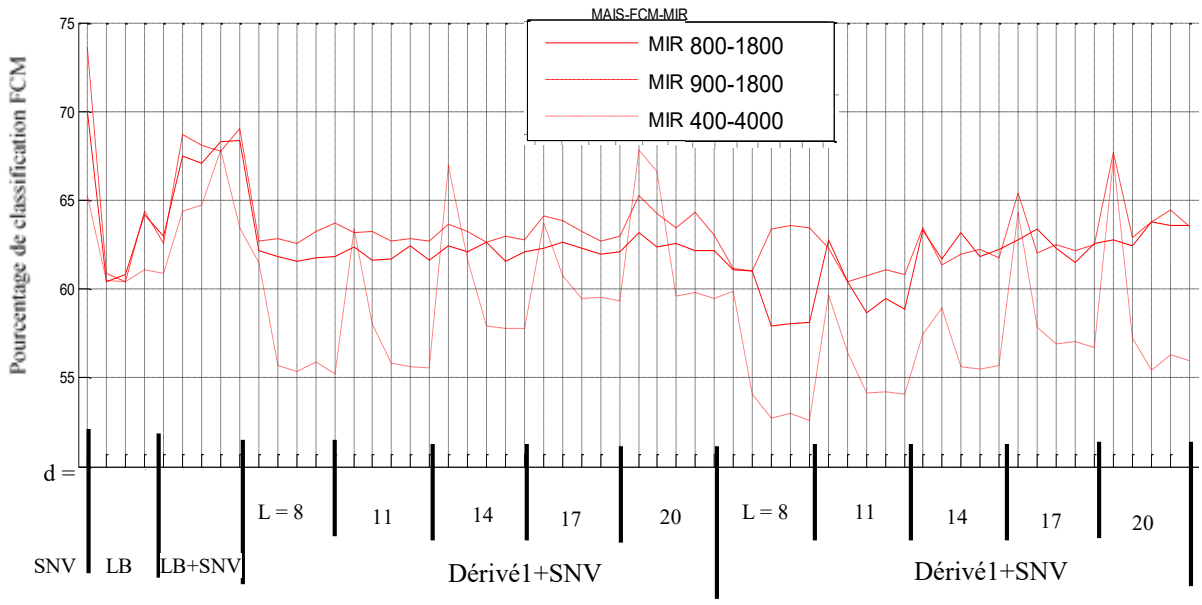


Figure A.2.5.Pourcentages de bonnes classifications. FCM bootstrap appliqué sur les spectres MIR enregistrés sur des échantillons de maïs. . Gammes spectrales 800 –1800 cm^{-1} (continu), 900 – 1800 cm^{-1} (discontinu), 400 - 4000 cm^{-1} (pointillé) pour différentes méthodes de prétraitements.

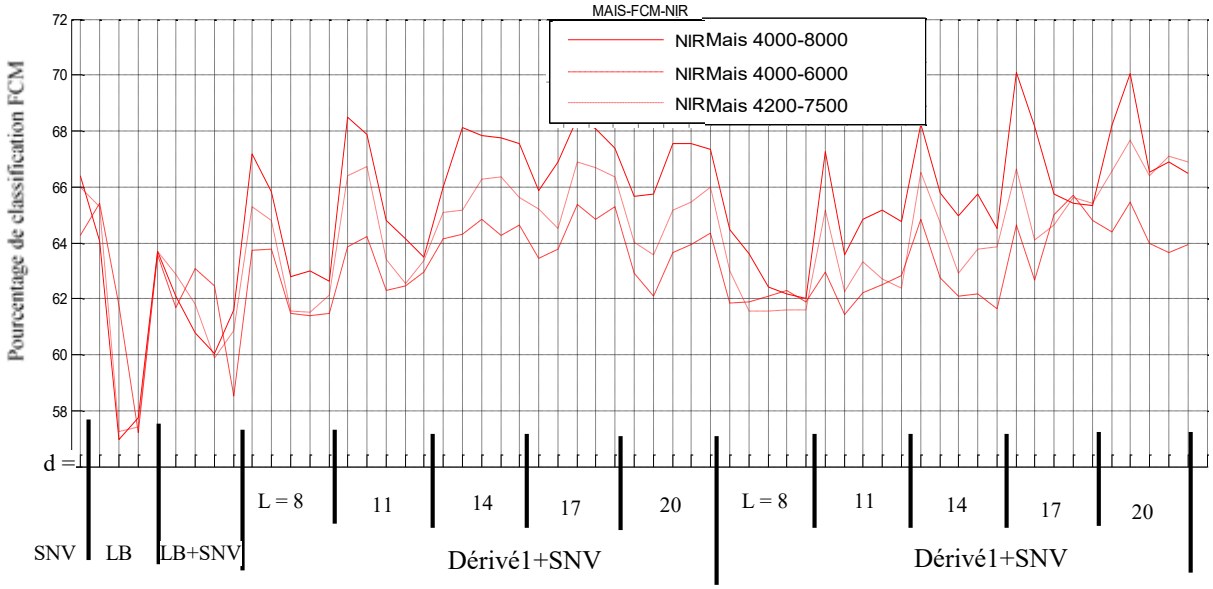


Figure A.2.6.Pourcentages de bonnes classifications. FCM bootstrap appliqué sur les spectres NIR enregistrés sur des échantillons de maïs. Gammes spectrales 4000 –8000 cm^{-1} (continu), 4000 – 6000 cm^{-1} (discontinu), 4200 - 7500 cm^{-1} (pointillé) pour différentes méthodes de prétraitements

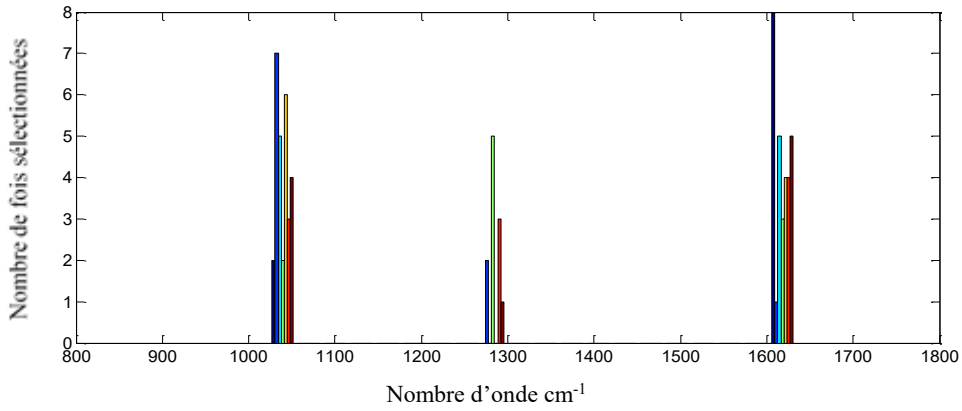


Figure A.3.1. Histogramme des nombres d'ondes sélectionnés par l'AG appliqué sur des spectres simulés pour L=5.

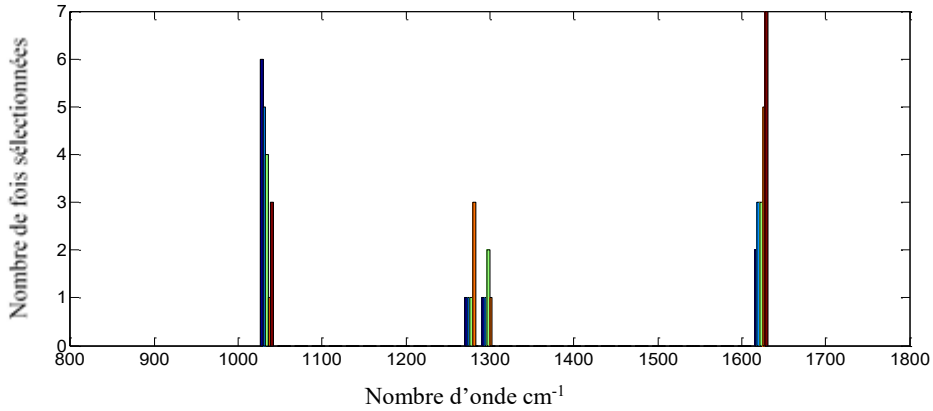


Figure A.3.2. Histogramme des nombres d'ondes sélectionnés par l'AG appliqué sur des spectres simulés pour L=7.

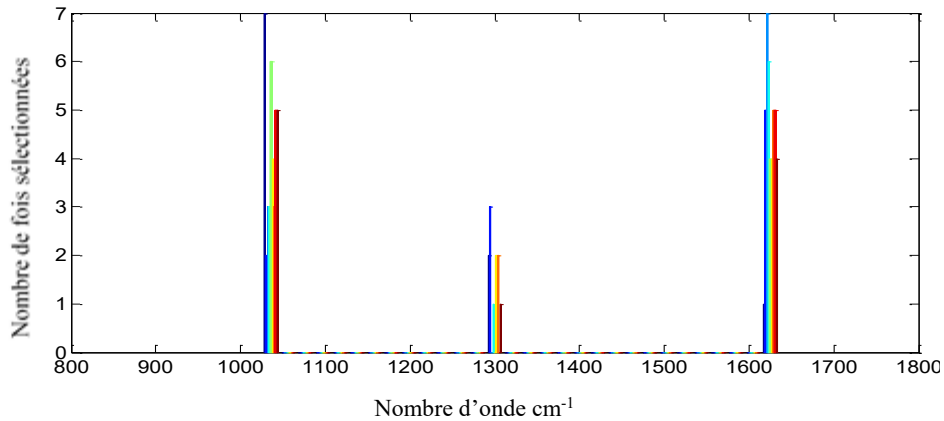


Figure A.3.3. Histogramme des nombres d'ondes sélectionnés par l'AG appliqué sur des spectres simulés pour L=9.

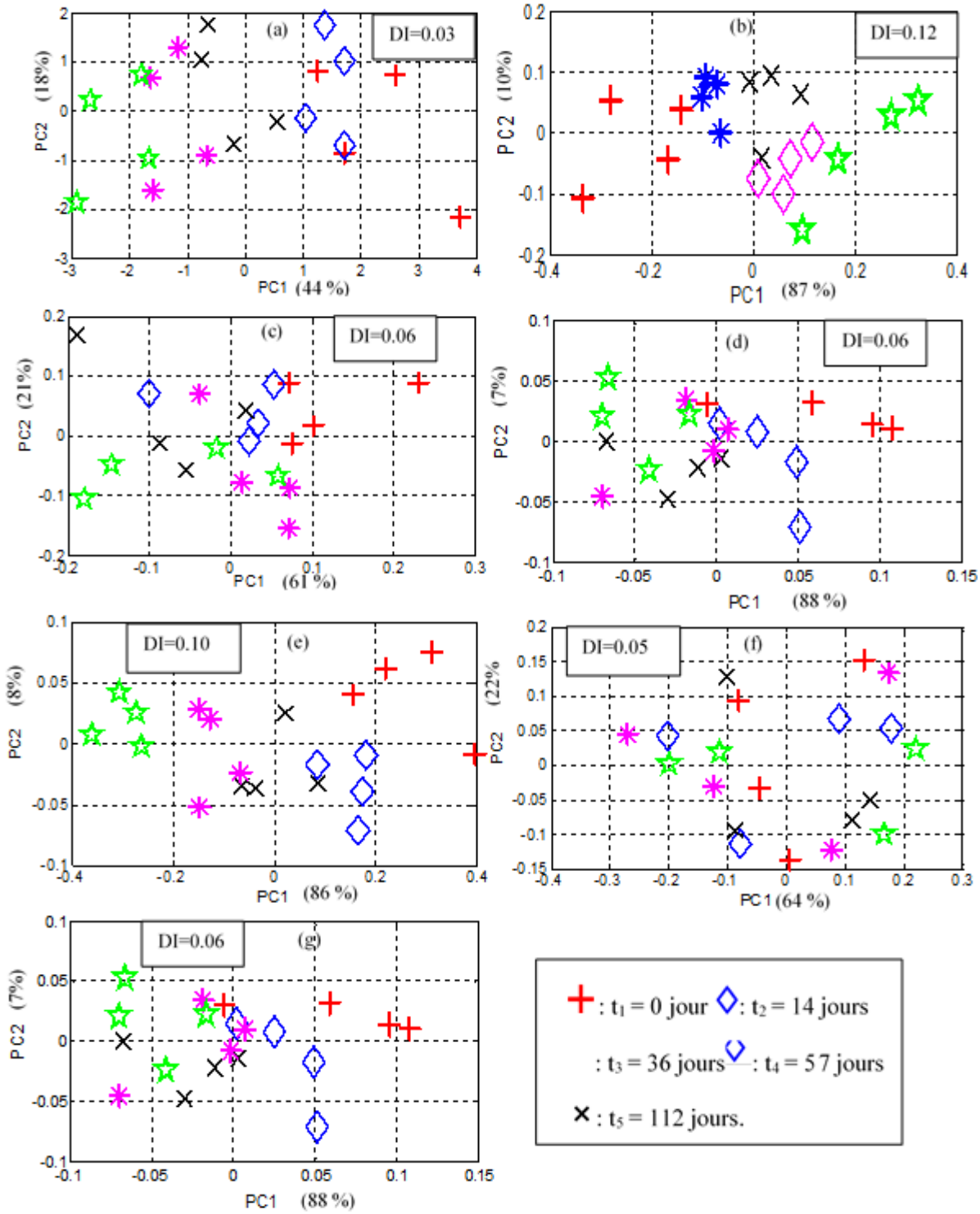


Figure A.3.4. Racines de maïs. Scores plots représentant la discrimination selon les périodes du processus de biodégradation en termes de PC1 vs PC2. L'ACP appliquée sur : (a) les spectres enregistrées sur la gamme spectrale 4000-6000 cm^{-1} du MIR, (b-g) les informations NIR sélectionnées aux nombres d'ondes identifiés par l'AG avec les fonctions fitness suivantes: (b) Davis Bouldin (DB), (c) Xie Beni (XB), (d) Calinski-Harabasz (CH), (e) Silhouette (SIL), (f) Séparation (SI), et (g) Fisher (FI).

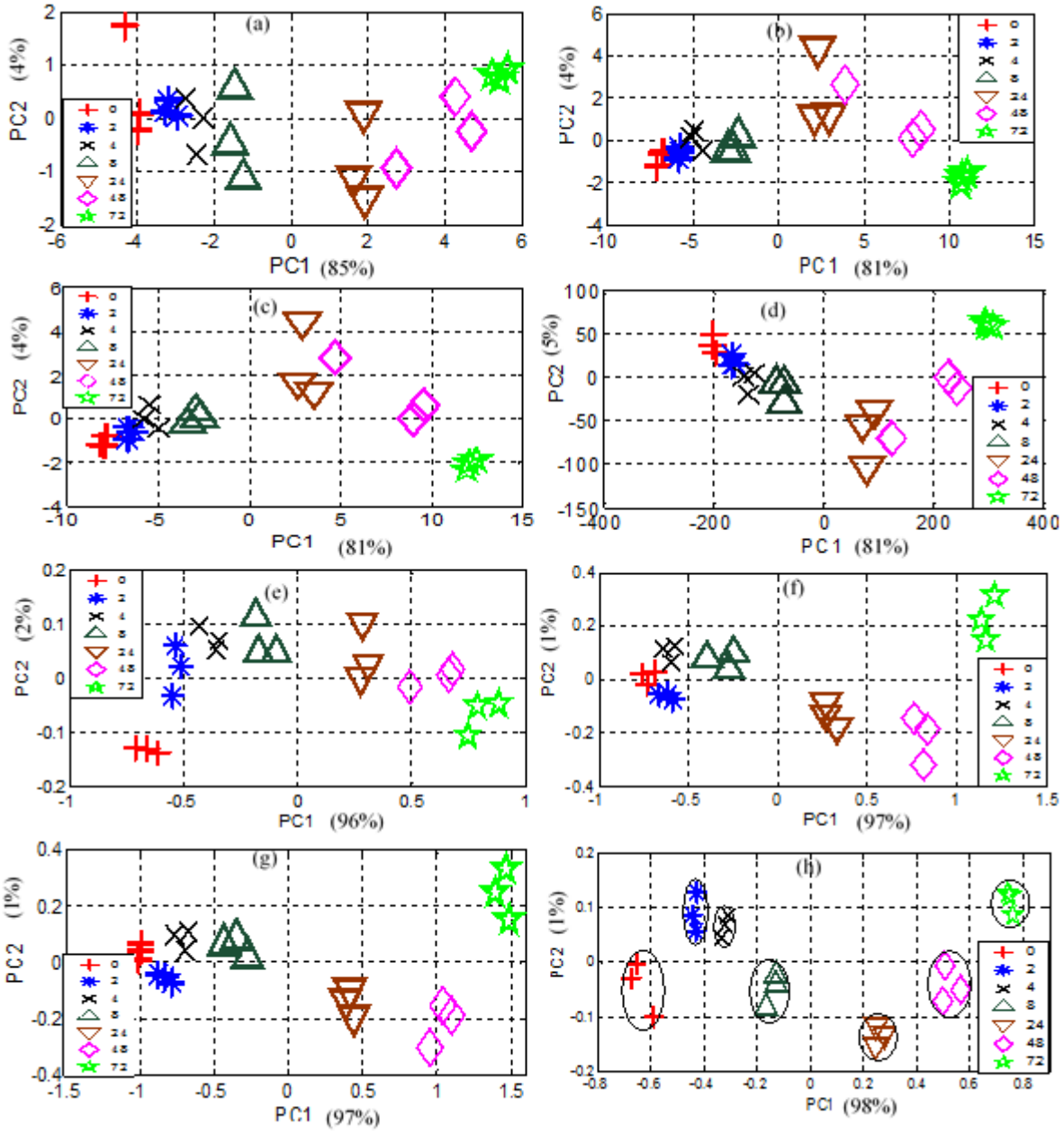


Figure A.3.5. Échantillons de miscanthus : Scores plots montrant la discrimination suivant les périodes du processus de biodégradation en termes de PC1 vs PC2. La PCA a été appliquée sur: (a) la gamme spectrale de MIR 800-1800 cm^{-1}, (b) la gamme spectrale de NIR 4000-6000 cm^{-1}, (c) les gammes concaténées MIR-NIR, (d) les spectres combinés MIR⊗NIR par le produit extérieur OP, (e) les nombres d'onde sélectionnés par l'AG sur MIR (883; 1489; 1641 et 1705 cm^{-1}), (f) les nombres d'ondes sélectionnés par l'AG sur NIR (4850, 5540, et 5705 cm^{-1}),(g) les nombres d'ondes sélectionnés par l'AG sur les gammes MIR-NIR concaténées (885, 1489, 1643, et 1708 cm^{-1}), (h) les nombres d'ondes sélectionnés par l'AG sur les spectres combinées MIR⊗NIR par le produit extérieur OP (833 x 5496), (877 x 4844), (1328 x 5340), and (1563 x 4889) cm^{-1}. La légende indique les temps de décomposition en heure.

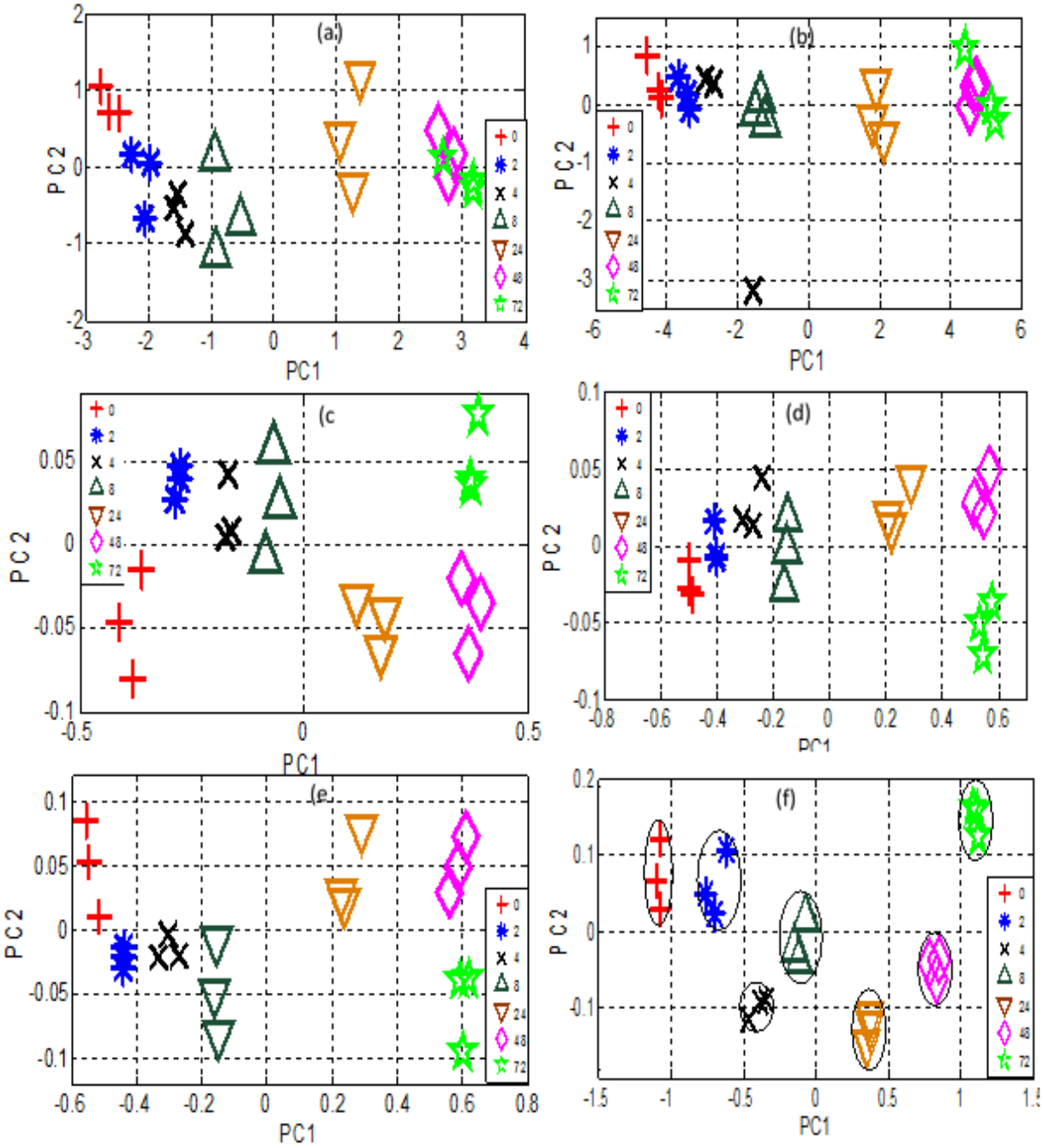


Figure A.3.6. Échantillons de Peuplier : Scores plots montrant la discrimination suivant les périodes du processus de biodégradation en termes de PC1 vs PC2. La PCA a été appliquée sur: (a) la gamme spectrale de MIR 800-1800 cm^{-1}, (b) la gamme spectrale de NIR 4000-6000 cm^{-1}, (c) les nombres d'onde sélectionnés par l'AG sur MIR (891;1485;1620 et 1687 cm^{-1}), (d) les nombres d'ondes sélectionnés par l'AG sur NIR (4470; 4885; 5593 et 5720 cm^{-1}),(e) les nombres d'ondes sélectionnés par l'AG sur les gammes MIR-NIR concaténées (885, 1489, 1643, et 1708 cm^{-1}), (f) les nombres d'ondes sélectionnés par l'AG sur les spectres combinées MIR⊗NIR par le produit extérieur OP (850 x 5908; 937 x 4748; 910 x 4709; 1481 x 4098; 1415 x 4462; 1675 x 5357 cm^{-1}). La légende indique les temps de décomposition en heure.

Tableau A.3.1. Significations chimiques des nombres d'ondes sélectionnés par l'algorithme génétique basés sur la fonction fitness Davies Bouldin pour les spectres MIR, NIR, MIR-NIR et MIR⊗NIR enregistrés sur les échantillons de miscanthus

MIR	NIR	MIR-NIR	MIR⊗NIR
• 883 cm^{-1}: le cellulose, hemicellulose, et lignine (Anomere C-groups, CH deformation, ring valence vibration). • 1489 cm^{-1}: les hemicelluloses et lignine (C–H deformations; asymmetric in $–CH_3$ and $–CH_2$) • 1641 cm^{-1}: le cellulose (H–OH stretching, C=C aromatic stretching vibrations) • 1705 cm^{-1}: Hemicellulose (C=O stretching in unconjugated ketones, carbonyls and in ester groups).	• 4323 cm^{-1}: le cellulose (OH stretching, C=O stretching + CH_2 bending) • 4438 cm^{-1} : la lignine (O-H stretching + C-O stretching). • 4734; 4711 cm^{-1}: cellulose (C-O stretching+ O-H deformation). • 5879 cm^{-1}: hemicellulose (1st overtone C-H stretching + C=O stretching).	• 1217 cm^{-1}: la présence d'éthers ou esters (C–O stretching. • 1637 cm^{-1}: le cellulose (H–OH stretching, C=C aromatic stretching vibrations) • 1725 cm^{-1}: l'hemicellulose (C=O stretching in unconjugated ketones, carbonyls and in ester groups). • 4485 cm^{-1}: la lignine (O-H stretching + C-O stretching). • 4888 cm^{-1}: le cellulose (OH stretching + C-H deformation). • 5824 cm^{-1}: hemicellulose (1st overtone C-H stretching + C=O stretching).	• (1328 x 5340): 1328 cm^{-1}: le cellulose et lignine (Aliphatic CH stretching in CH_3, CH deformation vibration). 5340 cm^{-1}: le hemicellulose (2nd overtone C=O stretching). • (1563 x 4889) • 1563 cm^{-1}: la lignine (Aromatic skeletal vibrations). 4889 cm^{-1}: la cellulose et le combinasion OH stretching + C-H stretching. • (877 x 4844); 877 cm^{-1}: la cellulose, hemicellulose, et lignine (Anomere C-groups, CH deformation, and ring valence vibration). 4844 cm^{-1}: cellulose the combination OH stretching + C-H stretching • (833 x 5496); 833 cm^{-1}: la cellulose, hemicellulose, et lignine (Anomere C-groups, CH deformation, and ring valence vibration). 5496 cm^{-1}: cellulose (O-H stretching + 2nd overtone C=O stretching).

Tableau A.3.2. Significations chimiques des nombres d'ondes sélectionnés par l'algorithme génétique basés sur la fonction fitness Davies Bouldin pour les spectres MIR, NIR, MIR-NIR et MIR⊗NIR enregistrés sur les échantillons de peuplier

MIR	NIR	MIR-NIR	MIR⊗NIR
• 891 cm^{-1} : le cellulose, hemicellulose, et lignine (Anomere C-groups, CH deformation, ring valence vibration). • 1488 cm^{-1} le hemicellulose, et lignine (C–H deformations;	• 4470 cm^{-1}: la lignine (O-H stretching + C-O stretching). • 4885 cm^{-1}: le bande de combinaison OH stretching + O-H déformation, combinaison de C=O avec CH_3.	• 891 cm^{-1}: le cellulose, hemicellulose, and lignin (Anomere C-groups, CH deformation, ring valence vibration). • 1620 cm^{-1}: la lignine (Aromatic skeletal vibrations C=O stretching, C=C aromatic stretching vibrations).	• (850 x 5908): 850 cm^{-1}: the vibrations of aromatic skelet on combined with CH wag 5908 cm^{-1}: la lignine (C-H stretching). • (937 x 4748):937 cm^{-1}: C-O-C stretching des polysaccharides (C-O-C stretching of the polysaccharides). 4748 cm^{-}

asymmetric in $–CH_3$ and $–CH_2–$). • 1620 cm^{-1}: le lignine (Aromatic skeletal vibrations C=O stretching, C=C aromatic stretching vibrations). • 1687 cm^{-1}: C=O stretching in carbonyls + C=C stretching in Olefinic.	• 5593 cm^{-1}: l'hémicellulose (1^{st}overtone de la C-H stretching). • 5720 cm^{-1}: la cellulose (1^{st}overtone de la C-H stretching).	• 1688 cm^{-1}: C=O stretching in carbonyls + C=C stretching in Olefinic. • 4468 cm^{-1}: la lignine (O-H stretching + C-O stretching). • 4885 cm^{-1}: le bande de combinaison OH stretching + O-H déformation, combinaison de C=O avec CH_3. • 5720 cm^{-1}: la cellulose (1^{st}overtone de la C-H stretching).	[1] : la cellulose (O-H stretching + O-H deformation). • (910 x 4709) 910 cm^{-1}: the cellulose, hemicellulose, and lignin (Anomere C-groups, CH deformation, ring valence vibration). 4709 cm^{-1} the cellulose (O-H stretching + O-H deformation). • (1487 x 4098) 1487 cm^{-1} les hemicelluloses et la lignine (C–H deformations; asymmetric in $–CH_3$ and $–CH_2–$).4098 cm^{-1} correspond to carbohydrates / lignin (C-C stretching + C-H stretching). • (1415 x 4462) 1415 cm^{-1}: le cellulose et la lignine (Aromatic skeletal vibrations combined with C–H in plane deformation, CH_2 scissoring). 4462 cm^{-1}: lignin (O-H stretching + C-O stretching). • (1675 x 5357) 1687 cm^{-1}: C=O stretching in carbonyls + C=C stretching in Olefinic. 5357 cm^{-1}: the first overtone of CH_2, CH_3 et -CH=CH-molecular;

Titre: Mathématiques appliquées et traitement du signal pour l'évaluation de la dégradation de la biomasse lignocellulosique

Résumé

Dans cette thèse nous proposons de mettre en œuvre des méthodes des mathématiques appliquées et du traitement du signal pour l'étude à partir de spectres infrarouges (IR) de l'évaluation de la dégradation de la biomasse lignocellulosique. Nous présentons tout d'abord une nouvelle méthode de classification floue fondée sur une optimisation de type non supervisée, basée sur le facteur de covariance qui permet de classer des données IR de forme sphérique ou non sphérique afin d'identifier les méthodes de prétraitement et de choix de gammes spectrales les mieux adaptées. Nous développons des outils mathématiques et des algorithmes innovants permettant de combiner des informations spectrales moyen IR (MIR) et proche IR (MIR) afin d'identifier des marqueurs spectroscopiques discriminants de résidus lignocellulosiques en fonction de leur niveau de dégradation. Pour cela, nous proposons une méthode d'optimisation stochastique basée sur un algorithme génétique avec paramètres adaptés. Nous montrons que l'analyse conjoints des spectres MIR et NIR fusionnés par le produit extérieur permet de mieux discriminer la biomasse lignocellulosique au cours du processus de dégradation qu'un traitement séparé. Nous proposons ensuite une nouvelle approche d'optimisation non linéaire basée sur la sélection d'un vecteur qui met en évidence les poids des bandes spectrales. Enfin, nous développons une méthode de modélisation mathématique basée sur l'extension de l'algorithme AG-PLS en combinant les informations spectrales MIR et NIR par le produit extérieur (OP-AG-PLS). Cette méthode permet d'améliorer les performances de prédiction de l'état de dégradation de la biomasse.

Mots clés : Mathématiques appliquées, Classification non supervisée, Optimisation non linéaire, Analyse numérique matricielle, Algorithme génétique, Modélisation mathématique, Spectroscopie MIR-NIR, Extraction des bandes spectrales discriminantes, Régression moindre carrés partiels, Combinaison spectres, Produit extérieur, Biomasse lignocellulosique, Processus de dégradation.

Title: Applied mathematics and signal processing for the evaluation of lignocellulosic biomass degradation.

Abstract

In this thesis we propose to implement methods of applied mathematics and signal processing for the study of the evolution of plant biomass during the biodegradation process. The degradation of plant biomass is identified by FTIR spectroscopy, particularly in the MIR and NIR ranges. We proposed a new unsupervised classification method of Fuzzy C-Means based on the covariance factor to classify the IR data with spherical and not spherical form to identify the pre-treatment methods and the choice of spectral ranges that are the best adapted for our study. We have developed mathematical tools and innovative algorithms to combine these spectral information and identifying infrared spectroscopic markers that are discriminative in the lignocellulosic residues according to their level of degradation. For this, we have proposed a stochastic optimization method based on a genetic algorithm by choosing the appropriate parameters. We have shown that the joint analysis of the MIR and NIR spectra by the outer product (OP) provides better results than the separate analysis for the discrimination of the lignocellulosic biomass during the degradation process. Then, we proposed a new nonlinear optimization approach based on the built of vector which highlights the weight of spectral bands. Finally, we have developed a mathematical modelisation based on the extension of the GA-PLS algorithm combining the MIR and NIR spectral information by outer product (OP-GA-PLS) which significantly improves the prediction performance of the state of degradation of biomass.

Key Word: Applied mathematics, Unsupervised classification, Non-linear optimization, Numerical matrix analysis, Genetic algorithms, Mathematical modeling, MIR-NIR Spectroscopy, Extraction of discriminating spectral bands, Partial least squares, Combination spectra, Outer product, Lignocellulosic biomass, Degradation processus.

Discipline: AUTOMATIQUE, SIGNAL, PRODUCTIQUE, ROBOTIQUE

Centre de Recherche en Science et Technologie de l'Information et de la Communication (CRESTIC) - EA 3804
UFR Sciences Exactes et Naturelles
Moulin de la Housse
51687 REIMS cedex 2