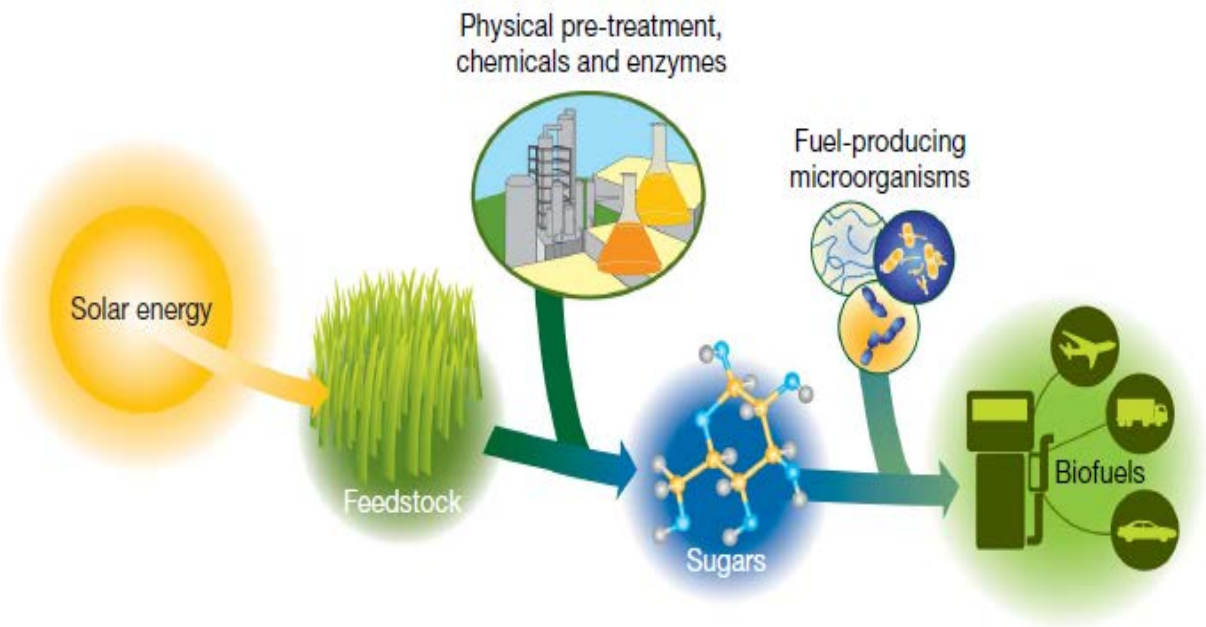
UNIVERSITE D'ORLEANS

INRA
Institut National de la Recherche Agronomique

RAPPORT DE STAGE 2012

MASTER 1 STATISTIQUES ET RECHERCHE OPERATIONNELLE



Construction de modèles de prédiction des propriétés du bois de peuplier à partir de données de spectrométrie en moyen infrarouge.

MURCIANO Violaine

Liste des Abréviations :

INRA : Institut National de la Recherche Agronomique

UR AGPF : Unité de Recherche Amélioration Génétique et Physiologie Forestières

EPST : Etablissement Public à caractère Scientifique et Technologique

EFPA : Écologie, Prairie et Milieux Aquatiques

EA : Environnement et Agronomie

GA : Génétique Animale

EPCS : Etablissement Public de Coopération Scientifique

CIRAD : Centre de coopération internationale en recherche agronomique pour le développement

UMR : Unité Mixte de Recherche

CNRS : Centre National de la Recherche Scientifique

IRP : l'Institut de Recherche pour le Développement

ONF: Office National des Forêts

ZF : Zoologie Forestière

GBFor: Génétique et Biomasse Forestière

ARCHE: Arbres et Réponses aux contraintes Hydriques et Environnementales

FTIR: Fourier Transformed InfraRed

ACP: Analyse en Composantes Principales

PLS: Partial Least Squares

PCR: Principal Component Regression

LOO: Leave One Out

CV: Cross Validation

CARS: Competitive Adaptive Reweighted Sampling

SNV: Standard Normale Deviate

MCCV: Monte Carlo Cross Validation

Remerciements

Je tiens à remercier l'ensemble de l'équipe de l'Unité de Recherche AGPF de l'INRA d'Orléans pour son accueil et son aide qui a contribué au bon déroulement de mon stage.

Je remercie plus particulièrement Gilles PILATE, directeur de l'UR AGPF pour m'avoir accueilli au sein de sa structure.

Je remercie vivement Vincent SEGURA, mon tuteur de stage, d'avoir pris le temps de répondre à mes questions, mes doutes. Pour m'avoir guidé progressivement tout au long de ma mission.

Je remercie également toutes les personnes travaillant sur ce projet pour leur disponibilité afin que je comprenne au mieux ce contexte biologique.

Enfin je remercie Laurent DELSOL professeur encadrant du stage ainsi que toute l'équipe pédagogique du master « Statistiques et Recherche Opérationnelle » de la Faculté des Sciences d'Orléans pour les enseignements m'ayant permis d'effectuer ce stage.

Introduction

En réfléchissant à mon avenir professionnel, des questions me viennent, en quoi ce métier de statisticien consiste-t-il concrètement ? Quelles tâches lui sont demandées, compétences attendues au niveau professionnel mais aussi personnel ? Ainsi le stage proposé durant cette première année de master est pour moi une réelle opportunité de trouver des réponses à ces questions.

Les statistiques sont utilisées dans de nombreux domaines. J'ai ainsi eu la chance d'effectuer mon stage de première année de master dans l'Unité de Recherche Amélioration Génétique et Physiologie Forestières (UR AGPF) de l'Institut National de la Recherche Agronomique (INRA) d'Orléans du 1er mai au 31 aout 2012.

Mon sujet de stage si situe dans un contexte de réduction des émissions des gaz à effet de serre et plus particulièrement la production de biocarburants à partir de bois de peuplier. Ce stage avait donc pour objectif d'effectuer des calibrations pour les propriétés chimiques du bois de peuplier.

Je vais tout d'abord dans ce rapport présenter la structure d'accueil, l'INRA ainsi que le contexte et les objectifs du stage. Puis, je présenterai le matériel et les méthodes que j'ai utilisés durant ce stage, en détaillant notamment les analyses statistiques mises en œuvre. Ensuite, j'exposerai les résultats que j'ai obtenus. Et enfin, je terminerai par une discussion sur les méthodes utilisées et les résultats obtenus.

1. Contexte et Objectifs

1.1. Contexte du stage

1.1.1. Présentation générale de l'INRA

Etablissement public à caractère scientifique et technologique (EPST), l'INRA est placé sous la double tutelle du ministère de l'agriculture, de l'agroalimentaire et de la forêt. Premier institut de recherche agronomique en Europe, deuxième dans le monde, l'INRA mène des recherches finalisées pour une alimentation saine et de qualité, pour une agriculture compétitive et durable, et pour un environnement préservé et valorisé.

C'est en 1946, juste après la guerre, que L'INRA a été créé dans un contexte de reconstruction du pays et de modernisation de l'agriculture française. A l'heure d'aujourd'hui les objectifs ont évolué et ont désormais une dimension mondiale. En effet, les recherches de l'INRA sont motivées par l'évolution permanente des questionnements scientifiques ainsi que les défis planétaires posés à l'agronomie et l'agriculture tels que le changement climatique, l'épuisement des ressources fossiles et la nutrition humaine.

L'INRA compte 8488 agents titulaires dont 1837 chercheurs, 2590 ingénieurs et 4061 techniciens, mais également 2000 thésards. Sont aussi accueillis chaque année environ 1989 stagiaires et 1800 chercheurs et étudiants étrangers.

L'INRA fait partie d'Agreenium, qui est un Etablissement Public de Coopération Scientifique (EPCS) et qui comprend aussi le Centre de coopération internationale en recherche agronomique pour le développement (CIRAD) et les écoles nationales d'agronomie. Il entretient de nombreuses collaborations et échanges avec la communauté scientifique internationale avec des pays d'Europe, Asie, Amérique et Afrique.

Cet établissement public qu'est l'INRA est composé de 14 départements scientifiques et 19 centres régionaux. On y trouve 213 unités de recherche dont 112 Unités Mixtes de Recherche (UMR) qui associent donc l'Inra à d'autres organismes de recherche (comme par exemple le Centre National de la Recherche Scientifique (CNRS), le CIRAD, ou l'Institut de Recherche pour le Développement (IRD)) et d'enseignement (notamment des Universités et Grandes Ecoles). L'INRA comprend aussi 49 unités expérimentales qui représentent une surface totale d'environ 10 000 ha ainsi que 94 000 animaux en élevage. En 2010, L'INRA possédait un budget de 813,9 millions d'Euros. Cet institut détient donc un dispositif unique de recherche dans les sciences du vivant.

1.1.2. Centre INRA d'Orléans

C'est en 1972 que le site d'Ardon fut choisi pour y implanter un centre de l'INRA. Si au tout début ce centre ne comptait qu'un bungalow, aujourd'hui il est composé de 4 domaines de recherche et de 6 unités différentes (Figure 1). Ces domaines sont les suivants :

- **La sélection des arbres forestiers** dont les différents critères sont une croissance optimale, une certaine qualité du bois et une bonne résistance aux parasites. Dans ce domaine, diverses

connaissances y sont développées telles que la génétique, la physiologie des arbres, les biotechnologies et le perfectionnement des techniques de diffusion des variétés améliorées.

- **La biologie des insectes forestiers ravageurs** ainsi que leur épidémiologie et leur relation avec les arbres-hôtes.
- **La maîtrise des érosions et pollutions, l'évaluation des risques agro climatiques** en se basant sur la modélisation des comportements physiques et hydriques des sols, en relation avec leur histoire et leur cartographie.
- **L'amélioration génétique des performances des troupeaux et la qualité de leurs produits** comme la productivité numérique et les qualités bouchères des ovins, la prolificité et la qualité de la viande des porcins, le rendement fromager des caprins ainsi que la qualité maternelle et l'aptitude bouchère de la race charolaise.

Ces différentes thématiques donnent au centre d'Orléans une orientation en matière d'**environnement** et de **développement durable**. Il a par ailleurs déployé une politique d'accueil et de partenariat avec différents organismes (Office National des Forêts (ONF), ArboCentre) qui favorise le lien entre la recherche, le développement et l'aide à la décision publique.

Le centre est composé de 6 unités:

- 3 unités de recherche : AGPF, Zoologie Forestière (ZF) et Science du Sol ;
- 1 unité de service : Infosol ;
- 2 unités expérimentales : Génétique et Biomasse Forestière (GBFor) et une autre située à Bourges.

Ces unités appartiennent à 3 départements scientifiques (Figure 1) : Écologie, Prairie et Milieux Aquatiques (EFPA) ; Environnement et Agronomie (EA); et Génétique Animale (GA).

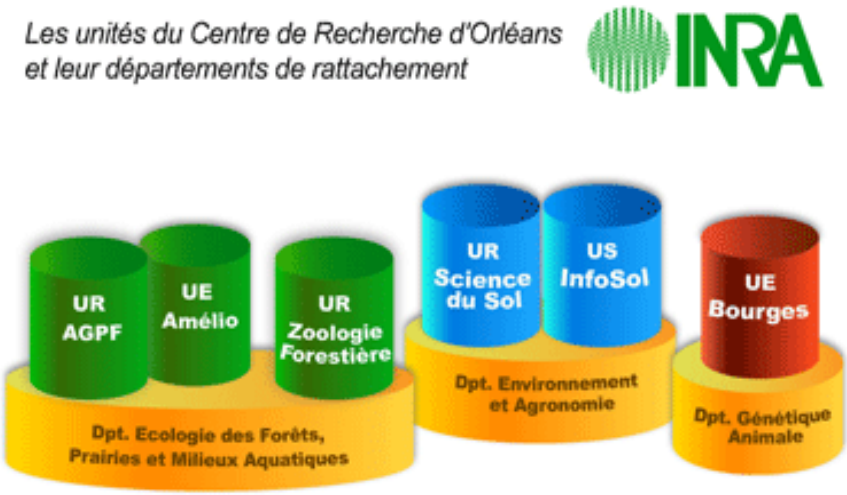


Figure 1. Les 6 unités du centre INRA d'Orléans avec leurs départements de rattachement

Le site d'Orléans compte environ 200 agents titulaires dont 38 % sont des Chercheurs et Ingénieurs, 49 % de Techniciens et 13 % d'Administratifs. Une cinquantaine de non titulaires sont accueillis chaque année, beaucoup sont des étudiants Doctorants ou Masters mais aussi des post-doctorants et chercheurs étrangers. Le budget du centre en 2005, hors salaires, s'élevait à 5,9 millions d'Euros dont les origines sont diverses (Ministères, collectivités territoriales, Europe, autres

organismes de recherche, partenaires, …).

1.1.3. L'Unité de Recherche AGPF

Dirigée par Gilles Pilate, l'Unité de recherche AGPF, fait partie du Département EFPA. Cette unité dans laquelle on étudie les arbres forestiers, rassemble différentes compétences en génétique, génomique et physiologie. Elle a pour objectif de valoriser les ressources génétiques forestières afin d'avoir une production durable de bois d'œuvre et de biomasse tout en prenant en compte l'impact écologique des populations domestiquées sur l'écosystème et le contexte climatique changeant. Des programmes d'amélioration génétique sont effectués sur 6 espèces forestières : Douglas, Mélèze, Pin Sylvestre, Frêne, Merisier et Peuplier.

L'UR AGPF collabore avec 3 unités locales, l'Unité Expérimentale GBFor, le Conservatoire génétique des Arbres Forestiers de l'ONF et l'équipe ARCHE (Arbres et Réponses aux contraintes Hydriques et Environnementales) de l'Université d'Orléans.

Trente-cinq personnes travaillent dans cette unité, dont la moitié sont des chercheurs et l'autre moitié est des ingénieurs, techniciens et agents administratifs. Une vingtaine de stagiaires, CDD et MOO y sont accueillis chaque année.

Le projet de l'Unité est organisé en 2 axes de recherche complémentaires. L'axe A est la recherche d'indicateurs d'adaptabilité au niveau des individus et des populations et l'axe B concerne la gestion durable de la diversité génétique dans les écosystèmes forestiers spontanés et cultivés. Les recherches menées dans le premier axe visent à obtenir une meilleure compréhension de l'élaboration du phénotype à la fois pour mieux comprendre l'architecture des caractères complexes c'est-à-dire identifier leurs déterminants moléculaires et génétiques importants pour la production de biomasse mais également étudier les variations du phénotype dans une large gamme de variation de milieux. Pour le second axe, les recherches ont pour objectif de maintenir la compétitivité des systèmes de production tout en gérant durablement les écosystèmes forestiers.

Mon projet se situe dans l'axe A, dans la partie « Peut-on prédire les caractéristiques de la biomasse en associant informations génomiques et phénotypage haut débit ? »

1.2. Contexte du projet

Dans un contexte général de réduction des émissions de gaz à effets de serre, la biomasse lignocellulosique constitue une ressource d'intérêt pour la production d'énergie au sens large et plus particulièrement de biocarburants. Afin de mobiliser cette ressource, un système sylvicole original, le taillis à courte et à très courte rotation (Figure 2) a été proposé chez le peuplier, notamment grâce aux travaux conduits dans les années 80-90 à l'INRA d'Orléans. Ces travaux ont par ailleurs révélé l'importance considérable que revêt le choix de matériel végétal sur le rendement en biomasse et que les clones de peuplier homologués pour une production de bois d'œuvre ne sont peut-être pas les mieux adaptés pour les cultures en taillis.

a b

Figure 2. Peupliers cultivés en (a) taillis à courte rotation et (b) taillis à très courte rotation.

En effet, jusqu'à présent les programmes de sélection n'ont pas pris en compte des critères visant à optimiser la quantité et la qualité de la production lignocellulosique, tels que l'aptitude au rejet de souche, la tolérance à la compétition et les propriétés chimiques du bois en vue de la production de bioéthanol. Pour le peuplier, la quantité et la qualité de la lignine semblent influencer largement le rendement de saccharification. La qualité de la cellulose est également déterminante dans la composition en sucres du produit de la saccharification. Il est donc indispensable d'évaluer la variabilité des caractéristiques chimiques du bois en vue de leur amélioration.

La caractérisation des propriétés du bois par des techniques de chimie classique est une approche extrêmement coûteuse et laborieuse. En vue d'évaluer à moindre coût les nombreux échantillons des programmes d'amélioration génétique une méthode indirecte a été proposée : la spectrométrie infra rouge. Il s'agit d'une méthode basée sur l'absorption du rayonnement infrarouge par les liaisons chimiques de la matière organique qui permet d'estimer la quantité et la qualité de ces liaisons. La mise en œuvre de cette méthode, pour l'évaluation indirecte des propriétés du bois, nécessite la construction de modèles de calibration pour les caractères cibles tels que les teneurs en lignines et cellulose. Cette modélisation s'effectue à partir de données chimiques et spectrales acquises sur un sous-échantillon de référence par des techniques de régression sur composantes principales (Principal Component, PC) ou sur variables latentes (Partial Least Squares, PLS).

Dans le cadre d'un projet de thèse dont l'objectif est d'identifier des régions génomiques impliquées dans la variabilité des caractères d'intérêt pour la production de biomasse lignocellulosique de peuplier, des premiers modèles de calibration ont été construits pour les teneurs en extractibles, lignines et cellulose du bois de peuplier. Ces calibrations ont été établies avec des spectres acquis dans le proche infrarouge. Tandis que cette technique détecte les bandes harmoniques et de combinaisons des vibrations fondamentales de la matière, il est possible dans le moyen infra rouge de détecter directement les bandes des vibrations fondamentales des constituants de la matière. La spectrométrie en moyen infrarouge s'avère ainsi plus prometteuse que la spectrométrie en proche infrarouge pour prédire les propriétés chimiques du bois.

Très peu d'études de ce genre ont été menées jusqu'à présent sur le peuplier. On trouve quelques travaux en proche infrarouge sur l'Eucalyptus (Baillères *et al.*, 2002 ; Giordanego, 2004), mais aussi sur d'autres espèces forestières (Schwanninger *et al.*, 2011)

Sur le peuplier, une recherche a été conduite avec de la spectroscopie dans le moyen

infrarouge (Zhou *et al.*, 2011). Pour le taux de lignine, ils ont effectué plusieurs modèles de calibration avec différents prétraitements des données spectrales et en sélectionnant des nombres d'onde. Ils ont également identifié certains nombres d'onde spécifiques des différents composants chimiques du bois de peuplier en s'aidant des données de la littérature, telle que les nombres d'onde 1593 et 1506 cm^{-1} qui correspondent à la lignine. Cependant leur étude ne permet pas de conclure sur la supériorité attendue du moyen infrarouge par rapport au proche infrarouge.

1.3. Objectifs

Dans ce contexte, l'objectif du stage était d'effectuer des calibrations pour les propriétés chimiques du bois de peuplier en utilisant des données spectrales acquises en moyen infrarouge et de les comparer aux modèles obtenus avec les spectres acquis en proche infrarouge (travaux effectués dans l'UR AGPF mais pas encore valorisés).

Pour cela, après une phase d'exploration et de prétraitement des données, des modèles de calibration ont été établis par des approches de régression sur variables latentes avec validations croisées. Par ailleurs, des techniques de sélection de variables ont été entreprises afin d'améliorer la qualité des modèles de calibration mais aussi d'identifier les nombres d'onde les plus pertinents car impliqués dans la prédiction des caractères chimiques d'intérêt. Ces résultats ont alors été confrontés à ceux obtenus en proche infrarouge afin de juger de la supériorité éventuelle de la spectrométrie en moyen infrarouge pour établir des modèles de calibrations pour les propriétés chimiques du bois.

2. Matériel et méthodes

2.1. Matériel végétal

Afin d'évaluer l'intérêt relatif de la spectroscopie en moyen infrarouge pour la calibration, les échantillons de peuplier utilisés pour ce projet sont les même que ceux qui ont été utilisés pour la construction des modèles obtenus avec les spectres acquis en proche infrarouge. Il s'agit d'échantillons de peuplier noir (*Populus nigra*) qui est une espèce dominante le long de nos fleuves et rivières. Cette espèce est étudiée dans l'UR AGPF pour son intérêt en amélioration génétique car elle est utilisée comme parent d'hybrides cultivés pour la production de bois et de biomasse.

Les échantillons de cette étude proviennent d'une famille de pleins frères constitués de 308 génotypes. Cette famille est issue de croisements contrôlés entre 2 parents ('71072-501' et 'BDG') dont on peut voir l'origine sur la Figure 3. Le parent femelle '71072-501', échantillonné en 1972 vient d'une population de Savoie. Le parent mâle 'Blanc de Garonne' ('BDG') vient de Garonne.
La famille complète a été produite au cours de 2 croisements successifs effectués respectivement en 1994 et en 2002.

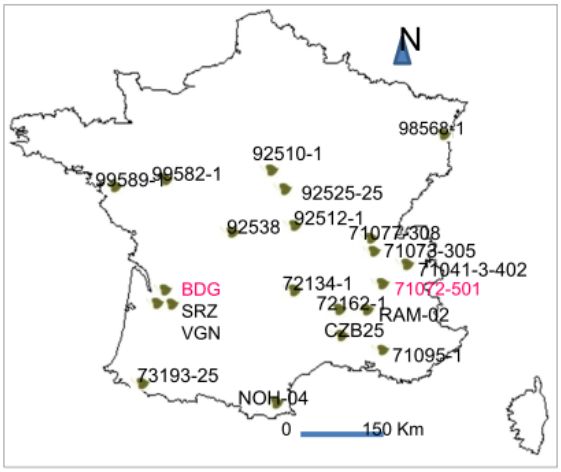


Figure 3 : Origine géographique des parents de la famille d'étude (en rose)

Le déroulement de ces croisements est le suivant :

- Tout d'abord des rameaux florifères sont récoltés en janvier février.
- Les rameaux sont ensuite mis dans l'eau afin que les bourgeons se développent
- On récolte ensuite le pollen des fleurs mâles (Figure 4)

Figure 4 : Rameaux florifères mâles dans l'eau laissant échapper le pollen que l'on récupère au sol.

- Ce pollen est déposé au pinceau sur les fleurs femelles au stade de maturité qui dure 3-4 jours. (Figure 5)

Figure 5 : Application du pollen sur les chatons femelles.

- On attend 3 à 4 semaines que les fleurs se développent (Figure 6)

Figure 6 : Développement des capsules.

- Une fois que les fleurs se sont développées, elles éclatent et libère du coton dans lequel se trouve des graines.

- On sème donc chaque graine qui est donc un individu de la famille et on suit son développement. (Figure 7)

Figure 7 : Plantation des graines récoltées.

Les 479 échantillons de l'étude consistent en 271 génotypes fois 1,77 répétitions, c'est-à-dire que parmi les 308 génotypes de la famille initialement plantée en 2004 à raison de 2 répétitions dans la pépinière de l'INRA d'Orléans certains étaient morts lorsque les échantillons ont été collectés en mars 2010. L'échantillonnage a consisté à prélever une section de tige de 50 cm sur le brin dominant de chaque cépée vivante. Ces échantillons ont ensuite été séchés, broyés au broyeur à couteaux et tamisés afin d'atteindre une poudre de granulométrie comprise entre 50µm et 1 mm, nécessaire pour effectuer les analyses des constituants de la paroi en laboratoire.

2.2. Les tests en laboratoire

2.2.1. Objectifs et protocoles expérimentaux

La Figure 8 illustre de façon schématique les différentes analyses effectuées sur les 479 échantillons de poudre de bois.

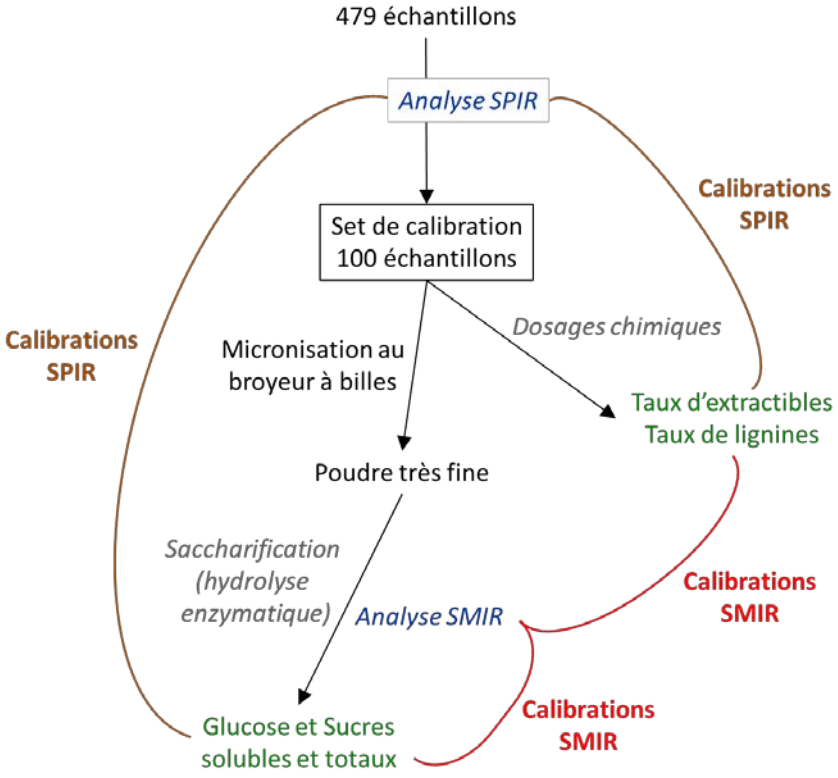


Figure 8 : Représentation schématique des différentes analyses faites sur les échantillons.

Les 479 échantillons ont d'abord été analysés par spectrométrie en proche infrarouge afin de constituer un set de calibration constitué de 100 échantillons qui soit représentatif de la gamme de variabilité spectrale. La spectrométrie infrarouge est une méthode de dosage indirecte qui repose sur l'absorption du rayonnement infrarouge par la matière organique aboutissant ainsi à l'obtention de spectres (Figures 9 et 10).

Figure 9 : Spectromètre avec le module pour le proche infrarouge et au-dessus le module pour le moyen infrarouge

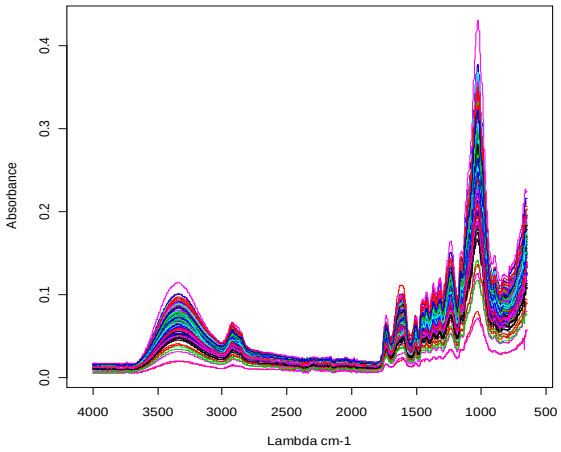


Figure 10 : Spectres des 100 échantillons de référence en moyen infrarouge

Divers tests en laboratoire ou analyses de références ont ensuite été réalisés sur ces échantillons de peuplier afin de pouvoir établir des modèles de calibration par analyse statistique.

Je ne vais pas décrire précisément tous les protocoles des tests en laboratoire étant donné qu'ils ne sont pas dans ma spécialité. J'ai pu tout de même assister à quelques analyses en laboratoire (dosages chimiques et prises de spectres) afin de mieux me rendre compte du travail effectué pour la collecte des données.

Brièvement, les premières analyses en laboratoire ont consisté à doser les teneurs en extractibles et lignines des échantillons. Ces mesures permettent par différence de déduire la teneur en holocellulose des échantillons (Taux d'holocellulose = 100 –(Taux d'extractibles + Taux de lignines). Ensuite, les poudres des échantillons ont été micronisées au broyeur à billes afin d'effectuer d'une part des dosages de saccharification (effectués à l'UMR Biotechnologie des champignons filamenteux à Marseille), et d'autre part d'acquérir les spectres en moyen infrarouge (contrainte technique liée à la très faible quantité de matière analysable). Les spectres en proche et moyen infrarouge combinés aux données de références permettent d'établir par analyse statistique des modèles de calibrations qui par la suite peuvent être utilisé pour prédire les dosages chimiques sur les autres échantillons. En effet, les dosages chimiques sont très longs et très coûteux contrairement aux tests spectroscopiques qui sont beaucoup moins long à effectuer (environ 200 échantillons en une journée).

Au cours de mon stage, j'ai travaillé spécifiquement sur les spectres acquis en moyen infrarouge avec comme principal objectif de confronter les calibrations ainsi obtenues à celle déjà effectuées au sein de l'unité AGPF sur les mêmes échantillons mais en proche infrarouge.

2.2.2. Les Données

Nous possédons donc deux types de données, les données spectrales et les données chimiques.

Les données spectrales, obtenues par spectrométrie moyen infrarouge des 100 échantillons de référence sont sous forme de fichier (un pour chaque échantillon). Chacun de ces fichiers contient le spectre de l'échantillon c'est-à-dire la valeur d'absorbance entre 4000 et 650 cm^{-1} avec un pas de 1 cm^{-1}.

Afin de pouvoir les utiliser sur le logiciel R, on utilise une fonction readSP.pe (gentiment fournie par la société Perkin Elmer fabriquant du spectromètre) qui nous permet de les lire. On sauvegarde ensuite toutes les données dans un fichier texte nommé data_MIR.txt.
On obtient ainsi une matrice à 3350 lignes qui correspondent aux nombres d'onde et 100 colonnes pour les 100 échantillons. Ces données sont représentées graphiquement en Figure 11.

Les données chimiques sont sous forme d'un tableau de 21 colonnes et 100 lignes. Chaque ligne correspond à un échantillon. Dans les 6 premières colonnes on retrouve des informations sur les échantillons comme leur différents identifiants attribués lors des différentes étapes de manipulations :

- **MIR_id** : il s'agit de l'identifiant de l'échantillon attribué lors de l'analyse en spectroscopie moyen infrarouge.
- **clone** : c'est le numéro de clone d'où proviens l'échantillon.
- **rep** : prend la valeur 1 ou 2 suivant qu'il s'agit de la première ou de la deuxième répétition du clone.
- **Echantillon** : numéro de l'échantillon.
- **ref_1 et ref_2** sont des références attribuées aux échantillons lors des différents dosages.

Les 15 autres colonnes correspondent aux différents composants chimiques obtenus suite aux dosages ou par calcul :

- **tx_extract_sec** : taux d'extractibles (molécules libres qui se trouvent dans la structure poreuse du bois), on les extrait avec des solvants.
- **tx_lign_tot_sec** : taux de lignines total qui comprend la lignine Klason (insoluble) et la lignine soluble.
- **tx_holo_calc_sec** : taux d'holocellulose calculé en soustrayant le taux de lignines total et le taux d'extractibles à la matière sèche totale.
- **Sucres_sol** : Sucres dosés avant l'attaque enzymatique
- **Gluc_sol** : Glucose dosé avant l'attaque enzymatique
- **NonGluc_sol** : Sucres qui ne sont pas du Glucose avant l'attaque enzymatique (Sucre_sol - Gluc_sol)
- **Gluc_sol_prop** : Proportion de Glucose avant attaque enzymatique (Gluc_sol / Sucres_sol)
- **Sucres_tot** : Sucres dosés après l'attaque enzymatique
- **Gluc_tot** : Glucose dosé après l'attaque enzymatique
- **NonGluc_tot** : Sucres qui ne sont pas du Glucose après l'attaque enzymatique (Sucres_tot - Gluc_tot)
- **Gluc_tot_prop** : Proportion de Glucose après l'attaque enzymatique (Gluc_tot / Sucres_tot)
- **Sucres_hydrol** : Sucres libérés par l'attaque enzymatique (Sucres_tot - Sucres_sol)
- **Gluc_hydrol** : Glucose libéré par l'attaque enzymatique (Gluc_tot - Gluc_sol)
- **NonGluc_hydrol** : Sucres qui ne sont pas du Glucose libérés par l'attaque enzymatique (NonGluc_tot - NonGluc_sol)
- **Gluc_hydrol_prop** : Proportion de Glucose libéré par l'attaque enzymatique (Gluc_tot_prop - Gluc_sol_prop).

2.3. Méthodes Statistiques

Pour toute l'analyse statistique, j'ai utilisé le logiciel R. Durant mon stage, j'ai appliqué plusieurs méthodes statistiques.

Dans un premier temps, j'ai utilisé des outils statistiques « simple » pour une première analyse des données afin de détecter d'éventuels problèmes, de mieux les connaître.

Tout d'abord, pour l'analyse des données spectrales, j'ai fait les moyennes et écart-types des spectres puis des Analyses en Composantes Principales (ACP) afin de détecter d'éventuels « outliers » ou valeurs aberrantes mais aussi pour évaluer la qualité de représentation des données. J'ai également effectué divers prétraitements des spectres qui permettent d'améliorer le signal et de condenser les données.

Pour les données chimiques, j'ai dans un premier temps réalisé les histogrammes de chaque composant dans le but d'étudier leur répartition. Ensuite j'ai calculé la matrice de corrélation pour détecter de possibles liens entre les différentes variables. J'ai également représenté les relations entre ces différentes variables par ACP.

Dans un deuxième temps, afin de trouver un lien entre nombres d'onde et composants du bois, j'ai combiné les données spectrales et chimiques en faisant au préalable des corrélations paramétriques et des corrélations de rang de Spearman.

Ensuite j'ai effectué des modèles de calibration au moyen de régressions sur composantes principales (PCR) et sur variables latentes (ou Partial Least Squares (PLS)). Pour déterminer le nombre de composantes ou de variables latentes optimal à inclure dans le modèle de régression, j'ai fait des validations croisées. Afin d'affiner et d'améliorer les modèles de calibration, j'ai filtré les données pour d'éventuels « outliers » et j'ai effectué une sélection de nombres d'onde au moyen de l'algorithme CARS (competitive adaptive reweighted sampling).

J'ai appliqué cette démarche aux 15 variables chimiques en faisant varier certains critères de sélection dans l'objectif d'identifier pour chacune des variables le meilleur modèle. J'ai réalisé l'ensemble de ces analyses sur toute la gamme spectrale moyen infrarouge mais aussi en me focalisant seulement sur la région 1800 - 900 cm^{-1} connue pour comprendre les bandes d'absorption correspondant aux composés chimiques à calibrer. Lorsque tous les modèles ont été trouvés, j'ai pu les comparer d'une part entre eux et d'autre part avec les modèles précédemment sélectionnés en proche infrarouge.

Dans ce qui suit je vais détailler les méthodes statistiques que j'ai utilisées pour effectuer les calibrations. Le code des fonctions utilisées est présenté en Annexe 1.

2.3.1. Analyses sur les données spectrales

Tout d'abord, j'ai découpé le spectre de 1800 à 900 cm^{-1} car en observant le spectre entier (4000 à 650 cm^{-1}), on remarque que c'est l'intervalle où il y a le plus de variations et donc de potentiels nombres d'onde identifiables (Figure 11). Il s'agit également de la région qui comprend les principales bandes d'absorptions correspondant aux molécules chimiques à calibrer (Faix, 1991 ; Zhou *et al.*, 2011).

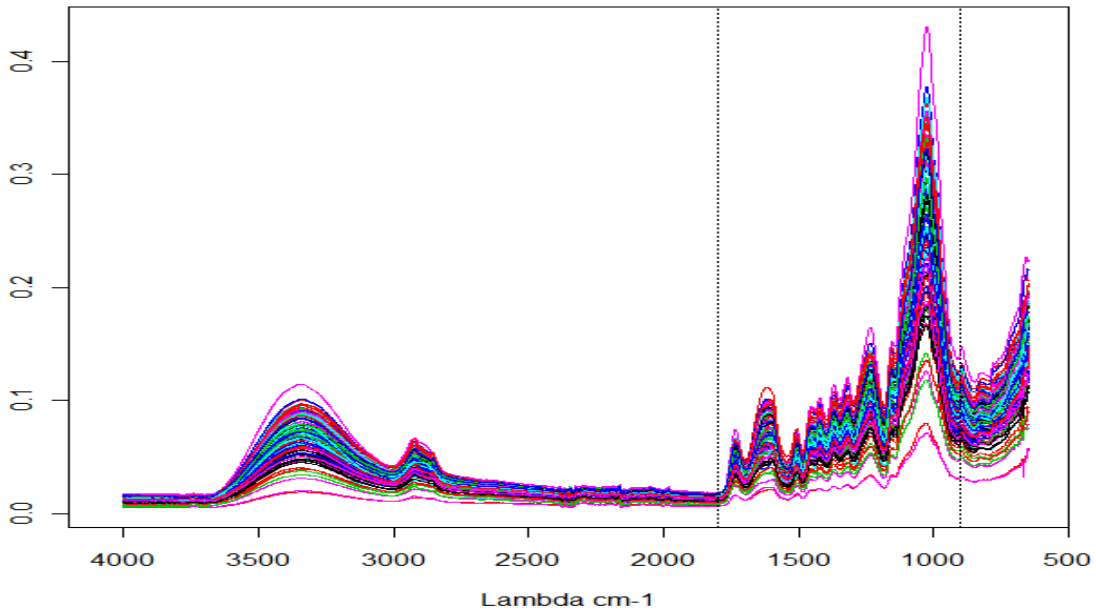


Figure 11 : Spectres entiers avec la partie sélectionnée (1800 à 900 cm^{-1}) entre les pointillés.

Ensuite, j'ai travaillé sur les données spectrales en y appliquant différents prétraitements car les données brutes peuvent être entachées par des défauts liés à la présence de bruit aléatoire ou affectés par les propriétés physiques de l'échantillon comme la taille et la distribution des particules. Les prétraitements ont donc pour objectifs d'améliorer le signal et de condenser les données. Ils sont donc indispensables pour une meilleure analyse des données (Bertrand et Dufour, 2006).

Les prétraitements que j'ai appliqués sont les suivants :

- La **normalisation** permet de réduire les variations incontrôlées de l'intensité générale des spectres, de les ramener à la même échelle pour les comparer sur le plan quantitatif et qualitatif (Figure12). Ici, nous avons utilisé comme normalisation le centrage réduction, cette transformation est aussi appelée en chemométrie Standard Normal Variate (SNV) ou variation normale standardisée. Les données transformées sont données par :

$$X_{norm} = (x - \bar{x} / \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}})$$

Avec n le nombre de nombre d'onde et $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

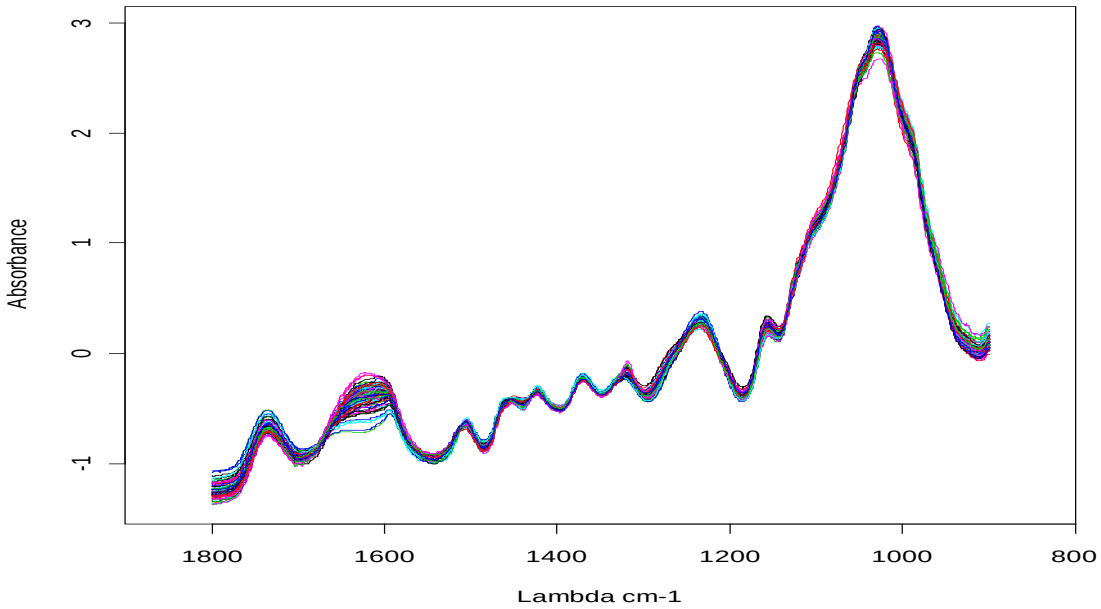


Figure 12: Spectre découpés et normalisés.

- La **dérivation** réduit les effets incontrôlés de la ligne de base pour rendre certaines informations du spectre plus distinctes en augmentant virtuellement la résolution du spectre. J'ai effectué des dérivées premières et secondes. La dérivée première sert à mesurer les pentes d'absorption en chaque point du spectre et la dérivée seconde permet de mesurer la concavité afin de localiser la position des maxima d'absorption (Figure 13).

Pour calculer ces dérivées, j'ai utilisé la méthode de Savitzky-Golay qui consiste à adapter un polynôme au spectre par une régression polynômiale dans un petit intervalle et d'utiliser les dérivées du polynôme pour estimer les dérivées du spectre. Les coefficients du polynôme sont calculés par la méthode des moindres carrés. On utilise pour cela, la fonction *sgolayfilt* de la bibliothèque *signal* du logiciel R. Cette fonction a pour paramètres : x le signal à filtrer, p l'ordre du polynôme, n le nombre de points pour dériver qui doit être impair, m l'ordre de la dérivée et ts le facteur d'échelle lié au pas de nombre d'onde (facteur d'échelle de temps). Ici, le paramètre ts est la différence entre le nombre d'onde maximal et le nombre d'onde minimal divisée par le nombre de nombres d'onde moins 1.

Les paramètres p, n et m sont différents pour calculer la dérivée première et la dérivée seconde. Pour la dérivée première, nous avons pris p égal à 2 et m égal à 1, tandis que pour la dérivée seconde, p égal à 3 et m égal à 2. En ce qui concerne le paramètre n, j'ai effectué une analyse de sensibilité en fonction du nombre de point utilisés afin qu'il ne manque pas trop d'information mais qu'il n'y ait pas trop de bruits non plus. Pour la dérivée première, j'ai pris 25, 51, 75 et 101 points et après analyse (moyenne, écart-type et ACP), j'ai choisi de garder 75 points. En effet, pour n égal à 25 et à 51, il y avait beaucoup de variance et donc beaucoup de bruit mais pour n grand égal à 101, il y avait très peu de variance et donc trop d'information avait été enlevé. De même, pour la dérivée seconde, j'ai pris 31, 67, 101 et 129 points et après analyse, j'ai conservé 101 points.

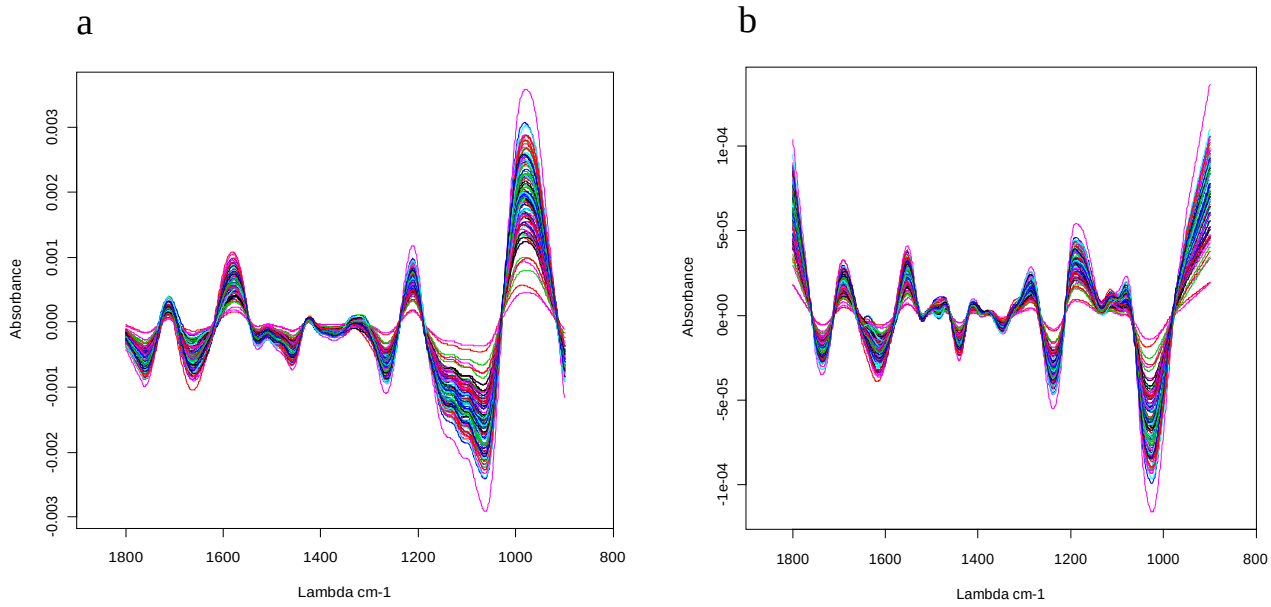


Figure 13 : Spectre découpés et dérivés 1 fois (a) et 2 fois (b).

J'ai fait ensuite des combinaisons de ces prétraitements : dérivée première puis normalisation, dérivée seconde puis normalisation, normalisation puis dérivée première et normalisation puis dérivée seconde (voir graphiques en Annexe 2). Pour ces différents prétraitements, j'ai calculé les moyennes et écart-types. J'ai ensuite effectué des ACP afin de détecter d'éventuels « outliers » avec la fonction prcomp et la fonction dudi.pca de la bibliothèque ade4.

2.3.2. Analyses sur les données issues des dosages chimiques

J'ai d'abord effectué des histogrammes pour chaque composant afin d'étudier sa répartition au moyen de la fonction *hist* de R.

Puis j'ai calculé la matrice de corrélation afin de détecter d'éventuels liens entre les différentes variables au moyen de la fonction *cor*.

J'ai ensuite appliqué une ACP afin de détecter d'éventuels « outliers » mais aussi pour identifier certains groupes de variables corrélées. Pour les ACP j'ai utilisé la fonction *prcomp* et la fonction *dudi.pca* de la bibliothèque *ade4.*

2.3.3. Combinaison des données spectrales et des variables chimiques

Après avoir fait l'analyse des données séparément, j'ai fait la combinaison de celles-ci afin de trouver un lien entre spectres et composants du bois.

Pour cela j'ai tout d'abord calculé des corrélations paramétriques et corrélations de Spearman (utilise les rangs plutôt que les valeurs exactes) pour les différentes variables chimiques et les différents prétraitements des données spectrales au moyen de la fonction *cor*.

On utilise ici une corrélation non paramétrique car l'on ne peut pas vérifier les conditions d'applications des corrélations paramétriques sur tous les nombres d'onde du spectre. Le coefficient

de corrélation de Spearman compare l'ordre dans lequel apparaissent les valeurs, c'est-à-dire leur rang et ensuite on calcule simplement le coefficient de corrélation paramétrique (ou de Pearson) sur les rangs. Ici nous n'avons pas détecté de différences significatives entre les 2 types de corrélations.

Pour construire les modèles de calibration, nous ne pouvons pas appliquer de régression linéaire multiple car la matrice X des données contient plus de variables que d'observations. Comme l'estimation des coefficients des moindres carrés implique le calcul de l'inverse de la matrice $X'X$ et que cette matrice est singulière, c'est-à-dire non inversible, cette méthode n'est donc pas applicable. On utilise alors des régressions sur variables latentes telles que les régressions PC ou PLS. Une autre justification importante sur l'utilisation de ces méthodes est la présence de fortes colinéarités dans les données spectrales.

Les régressions PC et PLS sont basées sur le même principe qui consiste à remplacer la matrice des données prédictives X de n lignes et p colonnes par une matrice C dérivée de X de n lignes mais k colonnes avec k inférieur à p. Les colonnes de C doivent être des combinaisons linéaires des variables d'origine. On a donc :

C = X V, où V est la matrice p×k des coefficients qui définissent les combinaisons linéaires.

Ensuite on effectue une régression linéaire multiple sur C et non sur X. Le problème est donc de déterminer V. La façon de calculer V est ce qui différencie une régression PC d'une régression PLS.

2.3.3.1. PCR

On calcule la matrice V par ACP. A partir du tableau de données X centré, on applique une ACP qui nous donne la matrice C dont les colonnes sont appelées les composantes principales orthogonales entre elles. La matrice V composée des vecteurs qui définissent les coefficients des combinaisons linéaires est alors la matrice des vecteurs propres de $X'X$. Cette matrice V remplie la propriété suivante :

$VV' = I$, où I est la matrice identité.

Le nombre de colonnes, k, de V désigne le rang de la matrice X. La matrice $X'X$ étant singulière, cela implique que les valeurs propres $\lambda_n, \lambda_{n+1}, \ldots, \lambda_p$ sont nulles.

Ainsi le principe de cette méthode est de construire un modèle de régression de y sur les composantes principales restantes.

En considérant le modèle de régression linéaire $y = X\beta + e$, on peut le récrire de la façon suivante :

$y = XVV'\beta + e = C\alpha + e$, avec $\alpha = V'\beta$.

Il s'agit maintenant d'un modèle de régression linéaire multiple où les variables prédictives sont les composantes principales de X, dans lequel α est estimable avec le procédé des moindres carrés. En pratique, les composantes principales sont incluses dans le modèle de régression selon leur valeur propre les unes après les autres et la détermination du nombre de composantes optimal se fait par validation croisée (cf. chapitre 2.3.3.3).

J'ai effectué les régressions PC au moyen de la fonction *pcr* de la librairie R *pls*.

2.3.3.2. PLSR

La régression PLS qui signifie partial least squares (moindres carrés partiels) mais aussi

projection to latent structure (projection sur des structures latentes) a été introduite par Wold en 1966 puis a subi de nombreuses modifications, évolutions. Il s'agit de la méthode la plus couramment utilisée pour les analyses chemométriques. Intuitivement, cette technique consiste à construire des variables latentes selon un critère de maximisation de leur covariance avec la variable réponse.

La matrice X de dimension n × p est constituée des vecteurs x_j des variables explicatives. On va donc chercher le vecteur $b_h = (b_{h1}, \ldots, b_{hp})$ contenant les coefficients de régression du modèles à h composantes.

La première étape consiste à déterminer la première composante t_1, combinaison linéaire des p variables de x_j dont les coefficients sont dans le vecteur $w_1 = (w_{11}, \ldots, w_{1p})$.

On doit donc avoir $t_1 = Xw_1$, avec $w_1 = \frac{X'y}{\|X'y\|}$.

On effectue ensuite une régression simple de y sur t_1 :
$y = c_1 t_1 + y_1$, où y_1 est le vecteur des résidus et $c_1 = \frac{y't_1}{t_1't_1}$ le coefficient de régression.
On peut donc en déduire une première équation de régression :
$y = c_1w_{11}x_1 + \ldots + c_1w_{1p}x_p + y_1 = b_{11}x_1 + \ldots + b_{1p}x_p$, avec $b_1 = c_1w_1$.

La deuxième étape consiste à construire la deuxième composante t_2 non corrélée avec t_1 et expliquant le résidu y_1, donc on a $t_2 = X_1w_2$ et $w_2 = \frac{X_1'y_1}{\|X_1'y_1\|}$

Pour calculer les résidus x_{1j} de X_1, on fait une régression linéaire des x_j sur t_1 :
$x_j = p_{1j}t_1 + x_{1j}$,
d'où $X_1 = X - t_1p_1'$, avec $p_1 = \frac{X't_1}{t_1't_1}$.
Puis on effectue une régression de y sur t_1 et t_2 :
$y = c_1t_1 + c_2t_2 + y_2$, où c_1 est le coefficient de la première étape et $c_2 = \frac{y_1't_2}{t_2't_2}$ coefficient de la régression simple $y_1 = c_2t_2 + y_2$.

Les étapes suivantes suivent cette même procédure. Le nombre de composantes $t_1,\ldots,t_H$ à retenir est déterminé par une validation croisée.

J'ai effectué les régressions PLS au moyen de la fonction *plsr* de la librairie R *pls*

2.3.3.3. Validation croisée

On utilise cette méthode pour déterminer le nombre de composantes à retenir dans les régressions.

La validation croisée consiste à découper le jeu de données en k sous-échantillons tirés aléatoirement. Des k sous-échantillons un seul est retenu comme donnée de validation. Les k-1 autres sont utilisés pour prédire. Le processus de validation croisée est répété k fois afin que chacun des k sous-échantillons soient utilisés comme donnée de validation. On fait ensuite une moyenne des k résultats obtenus pour avoir une estimation. Ainsi chaque échantillon sert à la fois pour la prédiction et pour la validation. Dans le cas particulier où k est égal au nombre de données on parle de validation croisée Leave One Out (LOO). Cette approche présente l'avantage d'être exhaustive

sur le jeu de données et ainsi les résultats qui sont issus de la validation croisée ne sont pas dépendants de l'échantillonnage effectué. Une autre façon de stabiliser les résultats de la validation croisée lorsque k n'est pas égal au nombre d'échantillons consiste à répéter la procédure de nombreuses fois. Cette approche est appelée « Monte-Carlo Cross Validation » (MCCV).

En pratique, on introduit successivement les composantes en ordre décroissant de leur valeur propre et on estime par validation croisée le « PRediction Error Sum of Squares » (PRESS) :

PRESS = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

Ensuite, on calcule le critère R de Wold donné par :

$R = PRESS_{m+1} / PRESS_m$, où $PRESS_m$ est le PRESS estimée dans le modèle comprenant les m premières composantes.

Enfin, on sélectionne le nombre de composantes minimum pour lesquelles le critère R de Wold est supérieur ou égal à une valeur de seuil. En général on utilise un seuil R = 1 qui consiste à sélectionner le nombre de composantes permettant d'obtenir le premier minimum de PRESS.

2.3.3.4. Filtrage des données

Cela consiste à détecter les valeurs aberrantes ou « outliers » et les enlever pour faire l'analyse sur les données sans « outliers ».

A cet effet, une fonction R a été créée, nommée : drop_outliers_LOO et présentée en Annexe 1. Cette fonction procède de façon itérative. A chaque itération une régression PLS est effectuée et le meilleur modèle est identifié selon le seuil choisi pour le critère de Wold par validation croisée LOO. Dans le modèle ainsi sélectionné, un z-score est calculé pour chacune des observations de la façon suivante : $z_i = |y_i - \hat{y}_i|/s$, où s est l'écart-type résiduel de prédiction ($s = \sqrt{PRESS/n}$). Une observation est alors déclarée comme un « outlier » si la probabilité que son z-score suive une loi normale est inférieure à un seuil prédéfini. Les « outliers » sont ainsi éliminés du jeu de données pour l'itération suivante. La procédure cesse à l'itération pour laquelle il n'y a plus d'« outliers » à enlever dans le jeu de données.

En appliquant cette fonction à un jeu contenant analyses de références et données spectrales, on récupère pour la variable à calibrer la liste d'observations « outliers » et le jeu de données filtré, c'est-à-dire dans lequel les « outliers » ont été enlevés.

2.3.3.5. CARS

La méthode « Competitive Adaptive Reweighted Sampling » (CARS), proposée par Li *et al.* (2009), est une technique permettant de sélectionner une combinaison optimale de nombres d'onde clés parmi toutes les données spectrales. Ces nombres d'onde clés sont sélectionnés à partir de leurs coefficients calculés par régression PLS.

Voici la description de la méthode CARS telle que proposée par Li *et al.* (2009). Tout d'abord, on choisit un nombre d'itérations N à effectuer. A la première itération on débute avec tous les nombres d'onde de la gamme spectrale. On applique une régression PLS avec validation croisée et on sélectionne le modèle minimisant l'erreur de prédiction. En utilisant les p coefficients de la régression pour le modèle sélectionné (p étant le nombre de variables, donc le nombre de nombres d'onde), on définit un poids, w, afin de pouvoir évaluer l'importance de chaque nombre d'onde :

$b = [b_1, \ldots, b_p]^t$, les coefficients, et

$w_i = \frac{|b_i|}{\sum_{i=1}^{p}|b_i|}$, i = 1, 2, …, p, les poids normalisés

On applique ensuite une “Exponentially Decreasing Function” (EDF) afin de mettre en application la sélection de nombres d’onde. Cette fonction consiste à calculer, à la j^{ème} itération, la proportion r_j de nombres d’onde à conserver (ceux qui ont le poids w le plus élevé) :

$r_j = ae^{-ki}$, où a et k sont des constantes déterminées par :a = $\left(\frac{p}{2}\right)^{1/(N-1)}$ et k = $\frac{\ln(p/2)}{N-1}$.

La Figure 14 illustre la fonction exponentielle décroissante EDF divisée en 2 parties, la première consistant en une sélection rapide et la deuxième en une sélection plus raffinée.

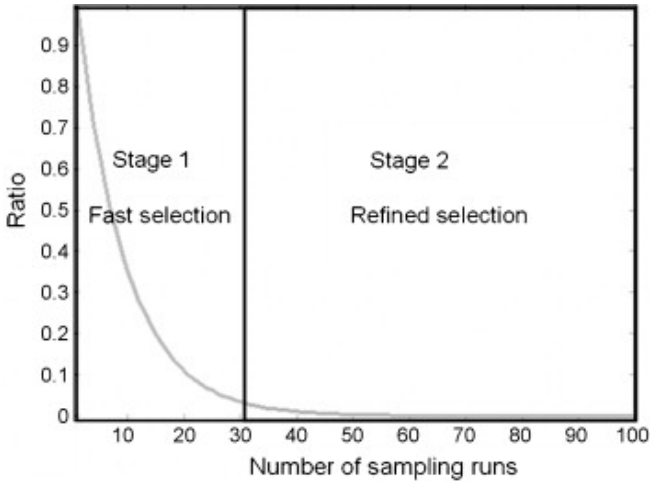


Figure 14 : Illustration de l’EDF (Figure tirée de Li et al. (2009)).

Enfin, la dernière étape qui est la méthode « adaptive reweighted sampling » (ARS) suit le principe de « survie des plus fort » qui est la base de la théorie d’évolution de Darwin. Cette technique permet de sélectionner les nombres d’onde à conserver pour l’itération suivante, leur nombre total étant déterminé par l’EDF. La Figure 15 permet d’illustrer le principe de la méthode ARS.

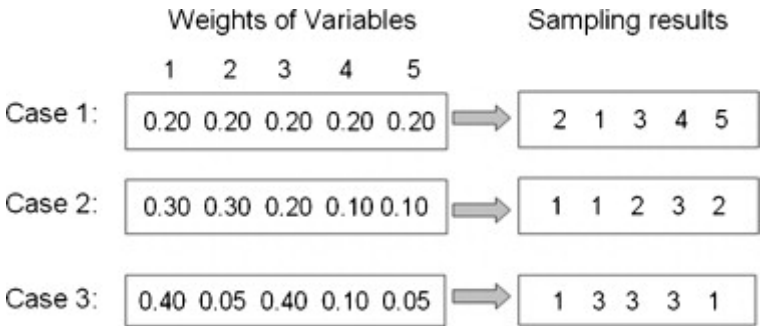


Figure 15 : Illustration de la méthode ARS (Figure tirée de Li et al. (2009)).

Dans cette figure, on suppose que l’on a 5 variables pondérées qui sont soumises à cinq expériences d'échantillonnage pondéré aléatoire avec remplacement. Dans le cas 1, chaque variable à un poids égal à 0,20 ce qui signifie qu’elles peuvent être échantillonnées avec une même probabilité. Ainsi le résultat idéal est que chaque variable est échantillonnée une seule fois. Dans le cas 2, les variables 1 et 2 ont le poids le plus élevé (0,30) alors que les variables 4 et 5 ont le poids le plus faible (0,10). Donc les variables 1 et 2 sont échantillonnées 2 fois, la variable 3 une fois et les variables 4 et 5 ne sont pas échantillonnées et éliminées. Le cas 3 est similaire au cas 2, seules les variables 1 et 3 qui ont un poids nettement plus élevé que les trois autres sont échantillonnées, alors que les variables 2, 4 et 5 sont éliminées.

Dans mon travail, j'ai utilisé une version simplifiée de la fonction qui n'inclue pas la dernière partie ARS par souci de reproduire exactement le résultat de sélection des variables avec les mêmes paramètres. Cette version simplifiée est recommandé pour l'étude scientifique par les auteurs de la fonction, et elle consiste à ne conserver à chaque itération que les k nombres d'onde dont le poids est le plus élevé. En outre, la fonction a été modifiée pour effectuer à chaque itération la sélection de modèle de régression PLS par validation croisée LOO et en utilisant un seuil prédéfini pour le critère de Wold, tels que décrits précédemment. Cette fonction, nommée cars_LOO, est présentée en Annexe 1.

Une fois les N itérations effectuées, on sélectionne les nombres d'onde de l'itération pour laquelle l'erreur de prédiction est minimale.

2.3.3.6. Etablissement des modèles de calibration

Pour chacune des variables de référence à calibrer, les modèles finaux sont alors construits par régression PLS avec validation croisée de type MCCV pour chaque prétraitement des données spectrales en utilisant les données filtrées pour les « outliers » et les nombres d'onde (grâce à l'utilisation des fonctions drop_outliers_LOO et cars_LOO).

Les modèles ainsi construits pour chaque variable de référence sont comparés du point de vue :

- de leurs statistiques : coefficient de détermination de validation croisée (R^2), erreur moyenne standard de validation croisée (RMSECV) et écart résiduel de prédiction (RPD) qui est le ratio entre l'écart-type de la variable de référence et l'erreur standard de validation croisée. Voici les formules des coefficients R^2 et RMSECV en fonction du coefficient PRESS :

  $RMSECV = \sqrt{\frac{PRESS}{n}}$

  $R^2 = 1 - \frac{PRESS}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
- du nombre de composantes ;
- du nombre d'« outliers ».

En fonction de ces différents indicateurs certains paramètres peuvent alors être ajustés, notamment les seuils pour la détection d'« outliers » et pour le critère de Wold afin d'identifier de meilleurs modèles potentiels.

2.3.3.7. Test de Chi 2 sur les nombres d'onde sélectionnés par CARS

Les modèles de calibrations ont été effectués pour toutes les mesures de références en utilisant le spectre découpé mais aussi le spectre entier ayant subi les divers prétraitements. Dans l'objectif de comparer ces 2 modalités, notamment du point de vue des nombres d'onde sélectionnés par l'algorithme CARS, des test du Chi 2 ont été entrepris. Il s'agissait alors de tester si les nombres d'onde sélectionnés par CARS sur le spectre MIR entier étaient surreprésentés dans la région de spectre MIR découpé.

Soit p_0 la proportion de nombre d'onde dans la zone du spectre découpé et p la proportion de nombre d'onde sélectionnés par CARS dans la zone découpée. On a comme hypothèses :

H0 : $p = p_0$

H1 : $p \neq p_0$

On utilise pour cela un test du Chi 2 à l'aide de la fonction R *chisq.test*. Cette fonction renvoi une p-valeur, si celle-ci est inférieure au niveau de significativité (0,01 ou 0,05) alors on rejette l'hypothèse nulle.

3. Résultats

3.1. Analyses exploratoires

3.1.1. Analyses des données spectrales

Pour chaque prétraitement, une ACP a été réalisée. Aucune variable aberrante n'est ressortie de ces analyses. Le pourcentage d'inertie expliqué par un axe ou un plan permet d'évaluer en quelque sorte la quantité d'information recueillie par cet axe ou ce plan.

On peut remarquer que le pourcentage d'inertie capturé par le premier axe et le premier plan diffèrent selon les prétraitements appliqués aux spectres. Pour le spectre brut, presque toute l'information est expliquée par le premier axe (96,6 % de l'inertie), chose que l'on ne retrouve pas pour les autres. Le pourcentage d'inertie capturé par le premier plan pour les spectres dérivé 1 et 2 fois demeure toutefois assez élevé (73,7 % et 70,1 % respectivement), tandis que pour les 5 autres prétraitements il diminue (entre 43,6 % et 45,7 %). Ces pourcentages d'inertie capturés par les axes sont visibles sur les nuages de points représentés pour les spectres brut et normé en Figure 16.

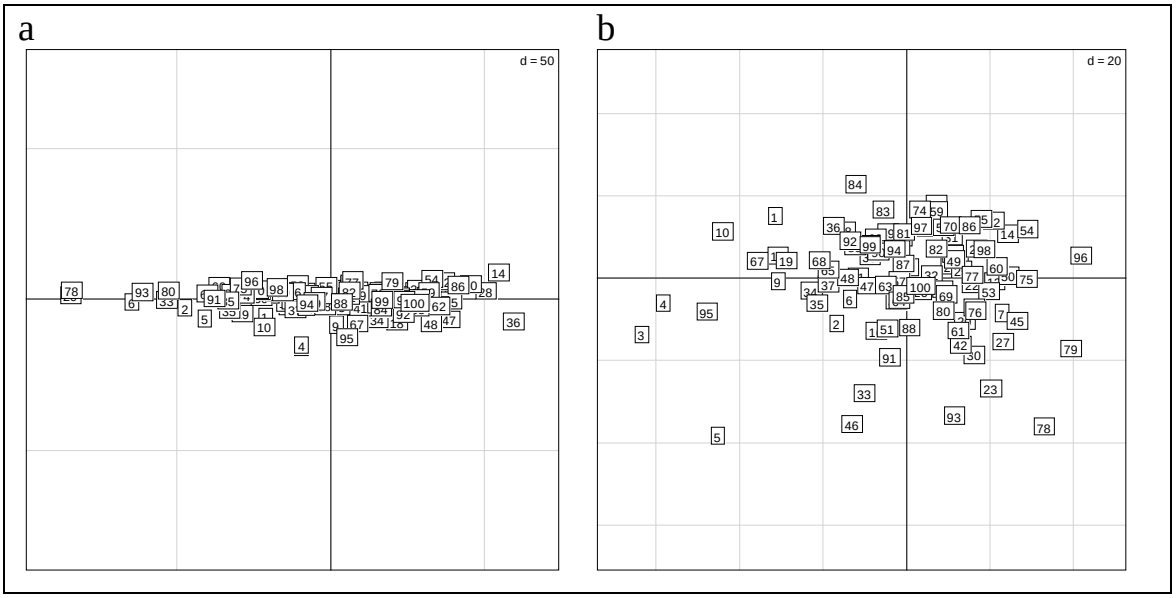


Figure 16 : Nuages de points de l'ACP des spectres brut (a) et normé (b)

On peut également voir la qualité de représentation des variables grâce aux cercles de corrélation. Par exemple pour les données spectrales brutes, le cercle des corrélations est représenté en Figure 17.

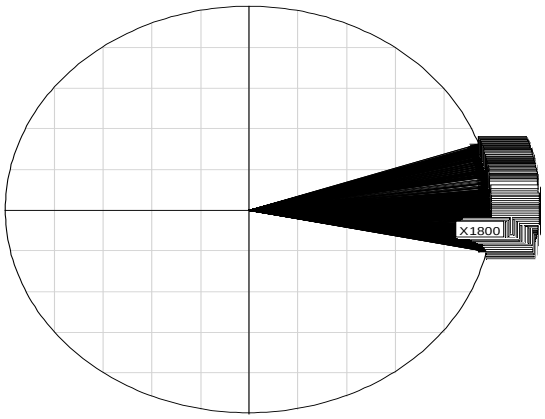


Figure 17 : Cercle de corrélation ACP spectre brut

On voit bien que les variables sont toutes corrélées avec l'axe 1 ce qui était attendu puisque ce dernier capture 96,6 % de l'inertie.

D'autre part on a pu constater que pour le spectre brut, il y a une corrélation entre la moyenne et l'écart-type des valeurs de spectre (Figure 18).

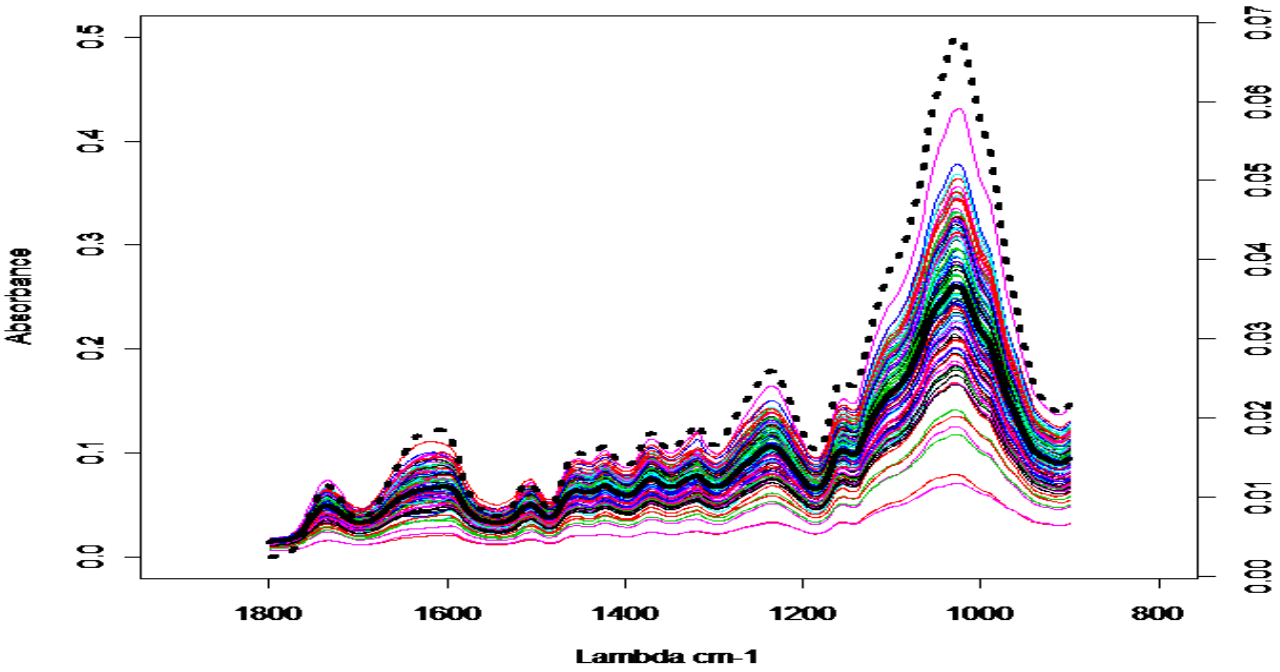


Figure 18 : Représentation spectres bruts avec sa moyenne (trait noir épaissi) et son écart-type (pointillés)

En revanche, pour le spectre normé, comme l'inertie capturée ne se concentre pas complètement sur le premier axe, le cercle des corrélations est très différent, (Figure 19).

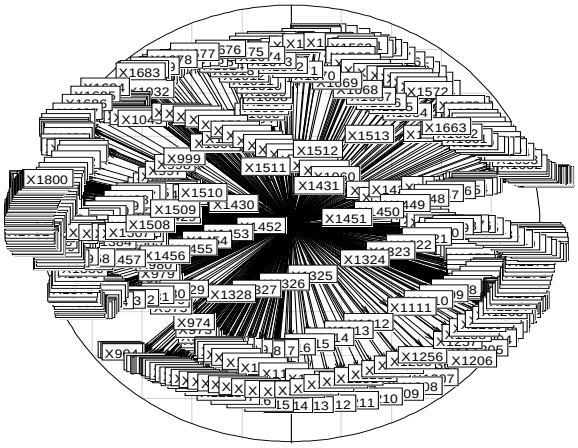


Figure 19 : Cercle de corrélation ACP spectre normé

On peut ici remarquer une plus grande diversité, en effet certaines variables sont bien représentées d'autres beaucoup moins bien, certaines sont corrélées avec l'axe 1 tandis que d'autres le sont avec l'axe 2. Pour les autres prétraitements, on obtient des résultats assez similaires (présentés en annexe 3). En outre, la corrélation entre moyenne et écart-type observée pour le spectre brut n'est pas détecté pour aucun prétraitement.

3.1.2. Analyses des données issues des dosages chimiques

Les histogrammes effectués sur les données chimiques ont permis de montrer que pour la variable NonGluc_hydrol il semble y avoir un « outlier », il s'agit de l'échantillon 11, (Figure 20). Mis à part ce dosage, les autres semblent bien répartis.

Certaines variables semblent gaussiennes (tx_extract_sec, tx_lign_tot_sec, tx_holo_calc_sec) mais de nombreuses variables (Gluc_sol, NonGluc_sol, NonGluc_tot,…) notamment parmi celles dosées à Marseille sont plutôt distribuées en forme de L. Il semble en effet que seulement les ratios (_prop) ne soient pas distribués en forme de L probablement à cause du calcul (les histogrammes de toutes les variables sont présentés en Annexe 4).

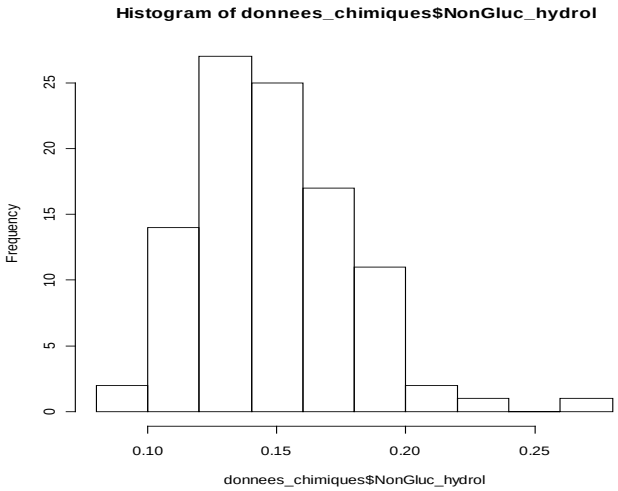


Figure 20 : Histogramme pour NonGluc_hydrol

L’analyse des corrélations entre les variables montre qu’il y a de fortes corrélations entre les différents dosages chimiques :

- Très forte corrélation entre : NonGluc_sol et Sucres_sol (0,996)

- Fortes corrélations entre : Gluc_sol et Sucres_sol (0,947)
  NonGluc_tot et Sucres_sol (0,914)
  NonGluc_sol et Gluc_sol (0,916)
  NonGluc_tot et NonGluc_sol (0,904)
  Gluc_sol_prop et Gluc_tot_prop (0,923)
  Gluc_tot et Gluc_hydrol (0,949)
  Gluc_hydrol_prop et Gluc_tot_prop (0,913)

- Assez fortes corrélations entre : tx_holo_calc_sec et tx_extract_sec (-0,817)
  Sucres_sol et tx_extract_sec (0,849)
  NonGluc_sol et tx_extract_sec (0,853)
  Gluc_sol_prop et tx_extract_sec (-0,865)
  Gluc_tot_prop et tx_extract_sec (-0,846)
  Gluc_sol_prop et Sucres_sol (-0,876)
  Gluc_tot_prop et Sucres_sol (-0,842)
  NonGluc_tot et Gluc_sol (0,888)
  Gluc_sol_prop et NonGluc_sol (-0,899)
  Gluc_tot_prop et NonGluc_sol (-0,854)
  NonGluc_tot et Gluc_sol_prop (-0,813)
  NonGluc_tot et Sucres_tot (0,810)
  Sucres_hydrol et Gluc_tot_prop (0,817)
  Gluc_hydrol et Gluc_tot (0,877)
  Gluc_tot_prop et NonGluc_tot (-0,874)
  Gluc_hydrol et Sucres_hydrol (0,862)

L’inertie de L’ACP effectuée sur les variables chimique se répartie de la manière suivante sur les axes : 59,2% des variables sont expliquées par l’axe 1, 23,6% par l’axe 2 et 8,6 par l’axe 3. On a choisi de garder les 2 premiers axes pour représenter les corrélations entre les variables mais aussi les observations. Cette représentation a permis de confirmer ce que l’on a observé avec les histogrammes, c'est-à-dire que l’échantillon 11 est un « outlier » (Figure 21).

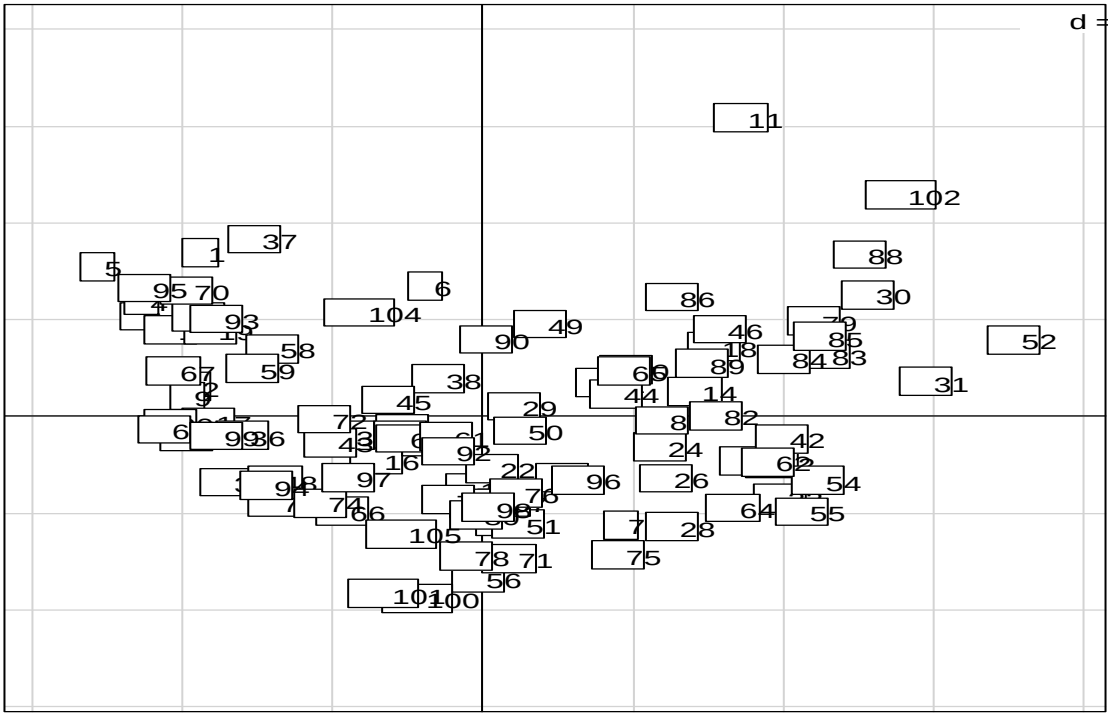


Figure 21 : Nuage de point de l'ACP

Le cercle des corrélations, présenté en Figure 22, confirme que l'existence de corrélations élevées qui permettent ainsi d'identifier certains groupes de variables:

- tx_extract_sec, NonGluc_sol, Sucres_sol, Gluc_sol et NonGluc_tot sont assez bien représentées, corrélées positivement entre elles ainsi qu'avec l'axe 1.

- Gluc_hydrol_prop, Gluc_sol_prop et Gluc_tot_prop corrélées positivement entre eux mais négativement avec l'axe 1.

- Gluc_tot et Sucres_hydrol sont corrélées positivement entre eux et avec l'axe 2.
  C'est aussi le cas pour NonGluc_hydrol et Sucres_tot qui forment également un groupe.

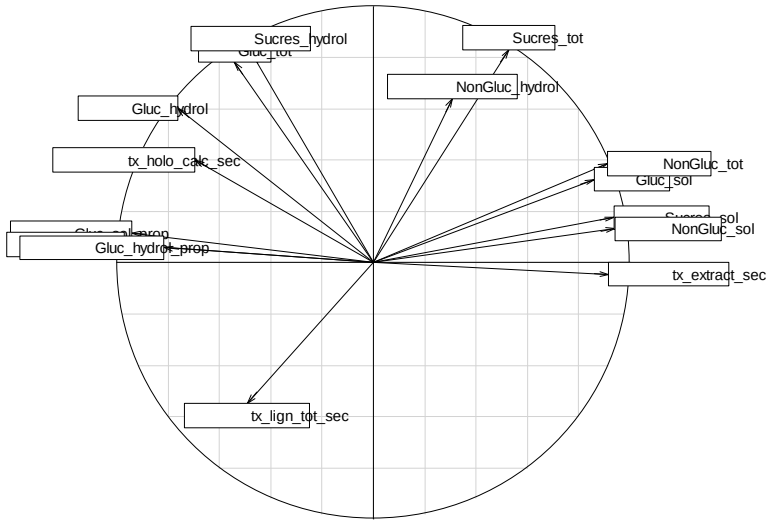


Figure 22 : Cercle de corrélation ACP

On a aussi regardé la contribution des variables à la formation des axes, qui est donnée pour les 2 premiers axes dans le Tableau 1.

	Composante 1 (59,2 %)	Composante 2 (23,6%)
tx_extract_sec	950	6
tx_lign_tot_sec	273	845
tx_holo_calc_sec	544	453
Sucres_sol	989	88
Gluc_sol	833	299
NonGluc_sol	998	49
Gluc_sol_prop	1002	39
Sucres_tot	312	1948
Gluc_tot	330	1724
NonGluc_tot	936	424
Gluc_tot_prop	1059	15
Sucres_hydrol	259	1924
Gluc_hydrol	658	1023
NonGluc_hydrol	106	1156
Gluc_hydrol_prop	751	10

Tableau 1 : *Contribution des variables de référence aux 2 premiers axes de l'ACP*.

On constate que NonGluc_sol, Gluc_sol_prop et Gluc_tot_prop contribuent fortement à l'axe 1 ; tandis que Sucres_tot, Gluc_tot et Sucres_hydrol contribuent fortement à l'axe 2.

Le taux de lignine est assez isolé sur le cercle des corrélations, corrélé négativement avec l'axe 2 alors que toutes les autres variables ne sont pas corrélées ou sont corrélées positivement avec cet axe. Ce résultat confirme l'attendu, à savoir que les lignines ont un effet négatif sur la saccharification.

3.1.3. Analyses conjointe des données spectrales et chimiques.

Comme première analyse pour la combinaison des données spectrales et chimiques, j'ai effectué des calculs de corrélations entre chaque variable chimique de référence et les données spectrales (brutes, mais aussi ayant subi les divers prétraitements). Ces résultats sont présentés sous forme de graphiques et pour quelques variables à titre d'exemple en Annexe 5.

D'une manière générale, on obtient des corrélations très hétérogènes en fonction de la gamme spectrale mais aussi du prétraitement et de la variable de référence considérée. Dans certains cas on observe d'assez bonne corrélations mais n'atteignant jamais 0,9. Quels que soient les prétraitements, on peut tout de même identifier les composantes chimiques pour lesquelles de fortes corrélations avec le spectre ont été observées : tx_extract_sec, Sucres_sol, NonGluc_sol, Gluc_sol_prop et Gluc_tot_prop. A l'inverse NonGluc_hydrol présente les plus mauvaises corrélations avec le spectre quel que soit le prétraitement. On peut aussi remarquer que pour les spectres normés, dérivés 1 fois et dérivés 2 fois, les corrélations sont à peu près égales et nettement plus élevées que celles observées avec le spectre brut (les corrélations maximales étant de 0,45 environ).

Ces analyses montrent les potentialités des spectres à prédire les variables chimiques, mais soulignent aussi la nécessité d'utiliser des outils multivariés plus sophistiqués afin d'obtenir des modèles de prédiction de bonne qualité.

3.2. Les modèles de calibration

3.2.1. Mise en point de la stratégie d'établissement des modèles

Afin de me familiariser avec les régressions PC et PLS mais également de les comparer pour choisir la meilleure méthode, je les ai appliquées pour les différents prétraitements et les variables de références tx_extract_sec et tx_lign_tot_sec avec des validations croisées LOO et à 3, 4 et 5 segments.

J'ai ainsi pu remarquer que la régression PC donne de moins bon résultats que la PLSR. En effet, pour le même nombre de composantes sélectionnées, le R^2 de validation croisée de la PCR est plus faible que celui de la PLSR ou alors pour un R^2 de validation croisée équivalent, la PCR sélectionne plus de composantes (Tableau 2).

	tx_extract_sec				tx_lign_tot_sec			
	PCR		PLS LOO		PCR		PLS LOO	
	R^2cv	Nb comps	R^2cv	Nb comps	R^2 cv	Nb comps	R^2cv	Nb comps
Brut	0,68	8	0,71	8	0,44	8	0,51	8
Normé	0,69	5	0,73	8	0,42	8	0,60	8
Dérivée 1	0,68	5	0,71	9	0,42	5	0,52	7
Dérivée 2	0,69	5	0,71	7	0,40	5	0,52	7
Dérivée 1 + Norme	0,71	6	0,73	8	0,41	5	0,60	7
Dérivée 2 + Norme	0,72	6	0,72	4	0,42	6	0,57	7

Tableau 2 : Résultats des régressions PC et PLS avec validation croisée LOO

En ce qui concerne les différentes méthodes de validations croisées pour la régression PLS, je n'ai pas pu distinguer de meilleure méthode (Tableaux 2 et 3). Alors, pour les mêmes variables que précédemment, j'ai aussi effectués des régressions PLS avec validation croisée de type MCCV avec 4 segments et 1000 répétitions. En comparant les résultats avec ceux obtenus précédemment, j'ai pu nettement m'apercevoir que l'approche MCCV est meilleure. En effet, pour les mêmes variables, on obtient un R^2 plus élevé et moins de composantes (Tableau 3).

	tx_extract_sec tx_lign_tot_sec							
	3 segments		4 segments		5 segments		MCCV	
	R^2cv	comps	R^2cv	comps	R^2cv	comps	R^2cv	comps
Brut	0,66 0,44	7 6	0,67 0,35	9 8	0,71 0,43	9 4	0,76 0,64	5 6
Normé	0,74 0,54	7 8	0,73 0,63	7 9	0,73 0,58	6 7	0,82 0,72	7 7
Dérivée 1	0,67 0,49	7 7	0,71 0,50	8 7	0,71 0,50	8 7	0,74 0,67	4 7
Dérivée 2	0,69 0,48	6 7	0,65 0,50	7 8	0,71 0,48	4 8	0,76 0,67	4 7
Dérivée 1 + Norme	0,73 0,52	3 6	0,71 0,58	8 7	0,73 0,56	3 6	0,79 0,71	5 7
Dérivée 2 + Norme	0,70 0,55	4 6	0,73 0,55	5 7	0,70 0,56	3 8	0,76 0,69	3 7

Tableau 3 : Résultats des régressions PLS avec différentes validations croisées

Ensuite, j'ai filtré ces données (toujours pour les mêmes variables) afin d'enlever les « outliers ». Sur ces données filtrées, j'ai appliqué des régressions PLS LOO et MCCV. On peut remarquer la même tendance que pour les données complètes, c'est-à-dire que les validations croisées MCCV fournissent des meilleurs résultats que les validations croisées LOO (Tableau 4). On constate également que sur ces données filtrées, on obtient de bien meilleurs R^2 et un nombre de composantes équivalent ou un peu plus élevé pour certains prétraitements (Tableau 4).

	tx_extract_sec				tx_lign_tot_sec			
	PLS LOO		PLS MCCV		PLS LOO		PLS MCCV	
	R^2cv	Nb comps	R^2cv	Nb comps	R^2cv	Nb comps	R^2cv	Nb comps
Brut	0,79	9	0,88	9	0,51	8	0,64	6
Normé	0,84	9	0,89	8	0,67	8	0,79	8
Dérivée 1	0,82	9	0,88	9	0,59	7	0,71	7
Dérivée 2	0,77	7	0,83	7	0,56	7	0,69	7
Dérivée 1 + Norme	0,85	9	0,90	9	0,67	7	0,77	7
Dérivée 2 + Norme	0,79	4	0,83	4	0,69	7	0,78	7

Tableau 4 : Résultats PLSR LOO et MCCV sur les données filtrées

Afin de déterminer le nombre d'itérations à faire pour la sélection de nombre d'onde avec la technique CARS, je l'ai testé sur tx_extract_sec avec le spectre normé. Pour cela j'ai appliqué la méthode CARS avec 100, 150, 200, 250, 300, 350, 400, 450, 500, 600 et 1000 itérations et j'ai comparé l'erreur standard minimum de validation croisée ainsi obtenue, la meilleure méthode étant celle qui est la plus faible. Ce coefficient décroit lorsque le nombre d'itérations augmente (Figure 23).

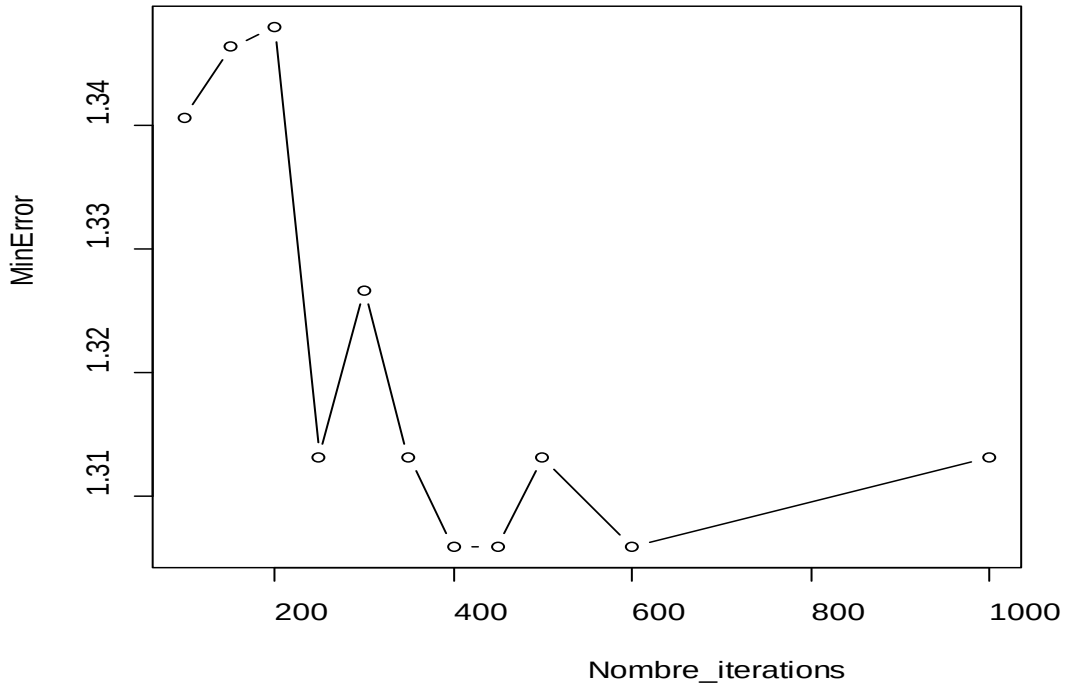


Figure 23 : Erreur standard minimum en fonction du nombre d'itérations

A partir de 400 ou 500 itérations, on obtient résultat égal à celui correspondant à 1000 itérations pour un temps de calcul beaucoup plus faible. J'ai donc décidé, dans l'objectif de sélectionner le meilleur modèle, de tester à chaque fois avec 400 et 500 itérations et de sélectionner

le meilleur des 2.

Après avoir effectué une analyse CARS sur les 2 mêmes jeux de données, j'y ai appliqué une régression PLS avec validations croisées LOO et MCCV. On constate que l'on a toujours de meilleurs résultats pour MCCV que pour LOO mais également que l'on a moins de composantes que pour l'analyse sur les données filtrées mais parfois avec un R^2 de validation croisée un peu plus faible (Tableau 5).

	tx_extract_sec				tx_lign_tot_sec			
	PLS LOO		PLS MCCV		PLS LOO		PLS MCCV	
	R^2cv	Nb comps	R^2cv	Nb comps	R^2cv	Nb comps	R^2cv	Nb comps
Brut	0,73	9	0,76	4	0,55	6	0,63	5
Normé	0,77	6	0,77	3	0,65	7	0,73	6
Dérivée 1	0,72	10	0,73	3	0,56	4	0,61	4
Dérivée 2	0,72	3	0,73	2	0,59	6	0,66	5
Dérivée 1 + Norme	0,74	2	0,77	2	0,62	5	0,70	5
Dérivée 2 + Norme	0,75	8	0,77	2	0,61	5	0,69	4

Tableau 5 : Résultats PLSR LOO et MCCV sur les données CARS

On a donc décidé de combiner les 2 méthodes CARS et le filtrage dans l'espoir d'obtenir d'encore meilleurs résultats. J'ai donc toujours sur les 2 variables de références tx_extract_sec et tx_lign_tot_sec, appliqué CARS sur les données filtrées pour les « outliers ». Puis j'ai effectué des régressions PLS avec validations croisées LOO et MCCV sur le jeu de données ainsi obtenu. La combinaison des 2 techniques permet en effet d'obtenir de meilleurs résultats que les analyses avec seulement CARS ou seulement les données filtrées (Tableau 6). Cependant on ne distingue pas de différences selon l'ordre dans lesquels on applique CARS et le filtre pour les « outliers ». On a donc décidé pour la suite de faire CARS sur les données filtrées.

	tx_extract_sec tx_lign_tot_sec							
	CARS sur filtre				Filtre sur CARS			
	PLS LOO		PLS MCCV		PLS LOO		PLS MCCV	
	R^2cv	comps	R^2cv	comps	R^2cv	comps	R^2cv	comps
Brut	0,81 0,55	5 6	0,84 0,66	5 6	0,82 0,59	9 6	0,88 0,69	8 6
Normé	0,86 0,72	5 5	0,89 0,78	5 5	0,87 0,72	6 7	0,89 0,79	6 7
Dérivée 1	0,84 0,61	4 5	0,87 0,69	4 5	0,72 0,61	7 4	0,75 0,66	3 4
Dérivée 2	0,80 0,61	5 6	0,84 0,67	5 5	0,74 0,65	5 6	0,77 0,71	3 6
Dérivée 1 + Norme	0,88 0,72	5 3	0,90 0,74	6 4	0,78 0,71	2 5	0,80 0,77	2 5
Dérivée 2 + Norme	0,82 0,72	7 5	0,81 0,75	3 4	0,85 0,78	6 5	0,88 0,81	6 4

Tableau 6 : Résultats PLSR LOO et MCCV pour données filtre sur CARS et CARS sur filtre

Pour chaque composant chimique, j'ai ainsi appliqué les régressions PLS avec validations croisées LOO et MCCV sur les données filtrées pour les « outliers » seulement mais aussi sur les données filtrées pour les « outliers » et pour les nombres d'onde sélectionnés avec la technique CARS. Une fois toutes ces analyses effectuées, pour la sélection des modèles j'ai utilisé uniquement

les régressions MCCV sur les données filtrées auxquelles on a appliqué la méthode CARS car on constate que c'est la méthode qui permet toujours d'obtenir les meilleurs résultats. J'ai également voulu m'assurer que l'on ne perdait pas d'information en ayant sélectionné seulement une partie du spectre j'ai donc effectuée ces mêmes analyses pour le spectre entier. Je me suis aperçue que pour certains dosages ces analyses sur le spectre entier donnaient de meilleurs résultats. Ainsi pour la sélection de modèles, j'ai pris en compte les analyses par régression PLS avec MCCV sur données filtrées pour les « outliers » et avec nombres d'onde sélectionnés selon CARS. Lorsque cela était nécessaire, dans le but d'obtenir les meilleurs modèles, j'ai aussi fait varier 2 critères, le seuil de détection des « outliers » et celui du critère de Wold pour la sélection du nombre de composantes.

3.2.2. Sélection des modèles

Je vais maintenant décrire la démarche que j'ai suivie pour obtenir les meilleurs modèles possibles. Toutes les analyses qui suivent sont des régressions PLS MCCV sur les données filtrées pour les « outliers » et les nombres d'onde (avec CARS). Seuls les seuils pour la détection d'« outliers » et pour le choix du nombre de composantes selon le critère de Wold varient.

Dans un premier temps, j'ai effectué l'analyse pour toutes les mesures de références et tous les prétraitements des spectres entiers et découpés avec un seuil de détection d'« outliers » égal à 0,01 et un seuil pour le critère de Wold égal à 1. Ensuite, en fonction des valeurs des différents indicateurs de qualité des modèles (R^2 de validation croisée, nombre de composantes, d'« outliers »), j'ai éventuellement ajusté les valeurs des seuils au cas par cas dans l'objectif d'obtenir à chaque fois le meilleur modèle possible.

Voici le bilan des modèles ainsi sélectionnés pour chacune des variables de référence. Ces modèles sont également présentés dans le Tableau 7 :

- **tx_extract_sec :** Pour cette variable, j'ai obtenu de très bons résultats avec les critères initiaux et je n'ai pas eu besoin de les faire varier aussi bien pour le spectre entier que pour le spectre découpé. En effet, pour le spectre découpé, j'ai obtenu un modèle à 4 composantes avec un R^2 environ égal à 0,84, 8 « outliers » et comme prétraitement la dérivée première.
  Pour le spectre entier, le modèle sélectionné est à 3 composantes avec un R^2 de 0,83, 8 « outliers » et on l'a obtenu en effectuant une dérivée $1^{ère}$ puis une normalisation comme prétraitement.
- **tx_lign_tot_sec :** Pour le spectre découpé, avec les critères initiaux, on obtient comme meilleur modèle, un modèle à 4 composantes avec un R^2 égal à 0,71 et 6 « outliers » ce qui n'est pas satisfaisant pour le R^2. Comme on a pour les 8 prétraitements assez peu d'« outliers », j'ai augmenté un peu le seuil de détection d'« outliers » à 0,015 dans l'objectif d'augmenter également le R^2. Ainsi avec ce nouveau seuil j'obtiens un bon modèle à 6 composantes avec un R^2 de 0,82 et 13 « outliers » pour le spectre dérivée 1 fois puis normalisé.
  Pour le spectre entier, et les critères initiaux, le meilleur modèle a 6 composantes, un R^2 égal à 0,78 et 8 « outliers ». Là aussi j'ai augmenté le critère à 0,015 et j'ai ainsi obtenu de bien meilleurs modèles (R^2 environ égal à 0,86) mais dont le nombre de composantes était un peu élevé (7). J'ai donc aussi diminué le seuil pour le critère de Wold à 0,95 et j'ai ainsi obtenu un modèle à 6 composantes avec un R^2 égal à 0,83 et 12 « outliers » pour le prétraitement dérivée $1^{ère}$ puis normalisation.
- **tx_holo_calc_sec :** Pour le spectre découpé avec les critères initiaux, le R^2 obtenu n'était pas très satisfaisant (0,75), le nombre d'« outliers » étant assez bas, j'ai augmenté leur seuil de détection à 0,015. Là encore avec seulement 8 « outliers » et un R^2 à 0,76, j'ai décidé de ré-augmenter le critère jusqu'à 0,02. J'ai de la sorte obtenu un bon modèle pour le

prétraitement norme puis dérivée 1ère avec 5 composantes, un R^2 égal à 0,80 et 13 « outliers ».

Pour le spectre entier, comme précédemment avec les critères initiaux le R^2 était seulement de 0,71 avec uniquement 6 « outliers ». J'ai donc augmenté le seuil à 0,015, pour obtenir avec 12 « outliers » un modèle dont le R^2 était de 0,76. J'ai essayé d'obtenir un meilleur modèle en prenant un seuil égal à 0,02, mais je n'ai pas obtenu de meilleurs résultats. Je suis donc passée à un seuil de 0,025 et là j'ai pu sélectionner un bon modèle pour le prétraitement normalisation avec 4 composantes, un R^2 égal à 0,83 et 18 « outliers ».

- **Sucres_sol :** Pour le spectre découpé avec les critères initiaux, les R^2 sont déjà très bon mais il y a trop d'« outliers » et de composantes. J'ai alors diminué le seuil de détection d'« outliers » à 0.001. J'ai ainsi obtenu des nombres d'« outliers » très faible avec toujours des R^2 assez élevé mais le nombre de composantes était toujours peu trop élevé. J'ai donc diminué le seuil pour le critère de Wold à 0,95 et j'ai sélectionné un très bon modèle sans « outlier », avec 5 composantes, un R^2 égal à 0.90 et aucun prétraitement.

  Pour le spectre entier c'est la même chose que pour le spectre découpé pour les critères initiaux, j'ai donc de la même manière diminué le seuil d'« outliers » à 0,001. Comme les R^2 étaient encore très élevé on a pu se permettre de diminuer encore le seuil à 0,0001 pour ainsi obtenir un bon modèle sans « outliers » avec 6 composantes et un R^2 assez élevé à 0,89 pour le prétraitement dérivée 2nde.
- **Gluc_sol :** Pour le spectre découpé et les critères initiaux, les R^2 obtenus sont très bons mais il y a trop d'« outliers » et de composantes. J'ai donc diminué le seuil à 0,001. J'ai pu sélectionner un bon modèle sans « outlier », avec 5 composantes, un R^2 égal à 0,92 et comme traitement la normalisation.

  Pour le spectre entier, avec les critères initiaux, les R^2 sont élevés mais il y a trop d'« outliers » on a donc diminué le seuil à 0,001. J'ai ainsi pu obtenir des modèles sans « outliers » avec un bon R^2 mais dont le nombre de composantes restait élevé. J'ai en conséquence diminué le seuil du critère de Wold à 0,95 et j'ai ainsi sélectionné un modèle sans « outliers » avec 6 composantes, un R^2 égal à 0,93 avec normalisation puis dérivée 1ère comme prétraitement.
- **NonGluc_sol :** Pour le spectre découpé avec les critères initiaux, les modèles obtenus avaient un R^2 très élevé mais un nombre de composantes et d'« outliers » également très élevé, j'ai donc diminué le seuil de détection d'« outliers » jusqu'à 0,0001 afin d'obtenir des modèles sans « outliers » et avec des R^2 encore assez élevé. J'ai ainsi pu sélectionner un modèle sans prétraitement ni « outliers » et avec 6 composantes et un R^2 égal à 0,90.

  Pour le spectre entier c'est le même procédé que pour spectre découpé, en abaissant le seuil de détection d'« outliers » à 0,0001, j'ai pu sélectionner un modèle sans « outliers » ni prétraitement et avec 6 composantes et un R^2 égal à 0,90.
- **Gluc_sol_prop :** Pour le spectre découpé comme pour le spectre entier et avec les critères initiaux, on obtient de bons modèles mais avec des « outliers » j'ai donc diminué le seuil à 0,001 pour ne plus avoir d'« outliers ». Ainsi pour le spectre découpé, j'ai sélectionné un modèle sans « outliers » avec dérivée 2nde puis normalisation comme prétraitement 4 composantes et un R^2 égal à 0,85. Pour le spectre entier, j'ai sélectionné un modèle avec normalisation comme prétraitement, aucun « outlier », 4 composantes et un R^2 égal à 0,87.
- **Sucres_tot :** Pour le spectre découpé, avec les critères initiaux, les R^2 obtenus ne sont pas assez bon. Le nombre d'« outliers » étant assez bas, j'ai augmenté le seuil afin d'obtenir un meilleur R^2 tout en gardant un nombre de composantes convenable. Ainsi, j'ai augmenté le seuil de détection d'« outliers » à 0,015 puis 0,02 mais ce n'est qu'avec 0,025 que j'ai pu obtenir un bon modèle. Avec la normalisation comme prétraitement, j'ai sélectionné un modèle avec 16 « outliers », 6 composantes et un R^2 égal à 0,80.

  Pour le spectre entier j'ai également obtenu des modèles avec peu d'« outliers » mais des R^2 peu élevés. J'ai donc augmenté le seuil directement à 0.02. Ainsi j'ai sélectionné un modèle

avec 5 composantes 16 « outliers » et un R^2 égal à 0,80, obtenu avec normalisation puis dérivée 2nde comme prétraitement.

- **Gluc_tot :** Pour cette variable il a été plus compliqué de trouvé un modèle correct. En effet pour le spectre découpé, j'ai dû faire varier le seuil de 0,01 (valeur initiale) puis à 0,001, 0,005, 0,0075 et enfin à 0,008 car pour 0,01 il y avait trop d'« outliers » et le R^2 était bon puis en passant à 0,001 le R^2 était devenu trop bas et il n'y avait presque plus « outliers ». J'ai donc augmenté progressivement ce seuil afin d'obtenir un bon compromis entre un nombre un peu plus élevé d'« outliers » mais correct et un R^2 assez bon. C'est en diminuant légèrement le seuil du critère de Wold à 0,90 et avec un seuil de détection d'« outliers » à 0,008 que j'ai obtenu le meilleur modèle pour le prétraitement normalisation. C'est un modèle à 5 composantes un R^2 égale à 0,85 et 12 « outliers ».
  Pour le spectre entier, de la même manière et pour les mêmes raisons j'ai fait varier le seuil de 0,01 à 0,001 puis 0,005, 0,0075, 0,008, 0,009 et enfin à 0,0095 avec un seuil du critère de Wold à 0,90 pour obtenir un bon modèle à 6 composantes, 14 « outliers », un R^2 égal à 0,85 et comme prétraitement la normalisation puis dérivée 2nde.
- **NonGluc_tot :** Pour le spectre découpé et les critères initiaux, j'ai obtenu de bons modèles avec peu de composantes (aux alentours de 3), des R^2 assez élevé environ 0,87 et assez peu d'« outliers » 3 ou 4. J'ai donc essayé de diminuer le seuil d'« outliers » afin de ne plus en avoir tout en gardant un R^2 assez élevé. Avec un threshold à 0,001, le nombre d'« outliers » a très peu diminué, j'ai donc pris 0,0005 pour obtenir des modèles sans « outliers ». J'ai ainsi pu sélectionner le modèle correspondant au prétraitement normalisation avec 4 composantes, un R^2 égal à 0,85 et aucun « outlier ».
  Pour le spectre entier, le seuil à 0,001 a suffi à obtenir des modèles sans « outlier ». J'ai donc sélectionné le modèle correspondant au prétraitement dérivée 1ère puis normalisation, avec 5 composantes, un R^2 égal à 0,83 et aucun « outlier ».
- **Gluc_tot_prop :** Pour le spectre découpé et entier, avec les critères initiaux, j'ai obtenu des modèles avec très peu d'« outliers » et de bon R^2. J'ai diminué le seuil d'« outliers » à 0,001 et ainsi pu obtenir dans les 2 cas des modèles sans aucun « outlier ».
  Ainsi pour le spectre découpé j'ai sélectionné le modèle du prétraitement normalisation avec 3 composantes et un R^2 égal à 0,86 ; et pour le spectre entier, le modèle sélectionné est celui qui correspond au traitement normalisation puis dérivée 2nde avec 4 composantes et un R^2 égal à 0,86 également.
- **Sucres_hydrol :** Cette variable a été assez dure à calibrer et je n'ai pas réussi à obtenir d'aussi bons modèles que pour les variables précédentes. En effet, pour le spectre découpé avec les critères initiaux, il y avait très peu d'« outliers » mais les R^2 étaient très faible (0,43 au mieux). J'ai donc augmenté le seuil de détection d'« outliers » à 0,02, 0,025, 0,03 et enfin 0,035 afin d'identifier des modèles dont le R^2 était assez élevé. J'ai dû m'arrêter à cette valeur de seuil car le nombre d'« outliers » devenait beaucoup trop élevé. J'ai ainsi sélectionné le modèle correspondant au prétraitement dérivée 1ère avec 3 composantes, 27 « outliers » et un R^2 peu élevé égal à 0,70.
  Pour le spectre entier, c'est le même constat pour les critères initiaux. En augmentant le seuil à 0.02, j'ai pu sélectionner un modèle avec 5 composantes un R^2 égal à 0.64 et 14 « outliers » pour le traitement normalisation puis dérivée 1ère. J'ai essayé d'augmenter encore le seuil à 0.025 mais je n'ai pas obtenu de meilleur modèle.
- **Gluc_hydrol :** Pour le spectre découpé et les critères initiaux, les R^2 obtenus sont assez élevés mais il y a trop d'« outliers ». J'ai donc diminué un peu le seuil à 0,005. Là j'ai sélectionné le modèle correspondant au traitement normalisation avec 4 composantes, un R^2 de 0,77 et 4 « outliers ». Je n'ai pas diminué encore plus le seuil pour éviter d'obtenir un R^2 trop faible.
  Pour le spectre entier et les critères initiaux, les R^2 sont également assez élevé et tout comme le nombre d'« outliers ». J'ai donc diminué le seuil à 0,001 le nombre d'« outliers » était

alors faible mais le R^2 également. J'ai alors augmenté légèrement à 0,005 et ainsi pu sélectionner un modèle. Il s'agit du modèle correspondant au prétraitement normalisation puis dérivée 2nde avec 5 composantes 4 « outliers »et un R^2 égal à 0,81.

- **NonGluc_hydrol :** Pour cette variable les modèles étaient très mauvais, je n'ai d'ailleurs pas réussi à en obtenir de convenables. Pour le spectre découpé et entier avec les critères initiaux, le nombre d'« outliers » est assez bon (8 environ) mais les R^2 sont très faibles (au mieux égal à 0,16). J'ai donc augmenté le seuil à 0,02, puis à 0,025 et enfin à 0,03 où j'ai dû m'arrêter car le nombre d'« outliers » devenait trop important. Pour le spectre découpé j'ai alors sélectionné le modèle correspondant au traitement dérivée 1ère avec 3 composantes, un R^2 égal à 0,70 mais 40 « outliers ». Pour le spectre entier, j'ai obtenu un encore moins bon modèle qui correspond au prétraitement dérivée 1ère avec 4 composantes, un R^2 égal à 0,46 et 34 « outliers ».
- **Gluc_hydrol_prop :** Pour le spectre découpé et entier avec les critères initiaux, le nombre d'« outliers » est assez faible mais le R^2 également (0,67 au plus haut). J'ai donc dans les 2 cas augmenté le seuil threshold à 0,02. Pour le spectre découpé, j'ai ainsi sélectionné le modèle avec le prétraitement dérivée 1ère puis normalisation avec 5 composantes, 16 « outliers »et un R^2 égal à 0,80. Pour le spectre entier, le modèle sélectionné correspond au prétraitement normalisation puis dérivée 1ère avec 3 composantes, 14 « outliers »et un R^2 égal à 0,78.

3.2.3. Classement des modèles sélectionnés et comparaison avec les modèles précédemment établis en proche infrarouge

Composé	type spectre	Prétraitement	paramètres	Nb comp	R² train	R² cv moy (sd)	RMSE cv moy (sd)	RPD cv moy (sd)	Nb Outliers	Nb lambda	Rang
Extractibles	NIR	Der2		5	0,87	0,83 (0,01)	1,12 (0,03)	2,46 (0,07)	3	131	1
	MIR découpé	Der1	SO: 0,01	4	0,87	0,84 (0,009)	1,07 (0,03)	2,49 (0,06)	8	119	2
	MIR entier	Der1Norm	SO: 0,01	3	0,86	0,83 (0,01)	1,07 (0,03)	2,48 (0,07)	8	100	3
Lignines	NIR	Der1		5	0,86	0,83 (0,01)	0,63 (0,02)	2,42 (0,06)	10	505	1
	MIR découpé	Der1Norm	SO: 0,015	6	0,86	0,82 (0,01)	0,67 (0,02)	2,37 (0,08)	13	30	3
	MIR entier	Der1Norm	SO: 0,015 ; W: 0,95	6	0,88	0,83 (0,01)	0,62 (0,02)	2,47 (0,09)	12	115	2
Holo	NIR	Der1		6	0,88	0,84 (0,01)	0,97 (0,03)	2,49 (0,08)	13	279	1
	MIR découpé	NormDer1	SO: 0,02	5	0,84	0,80 (0,01)	1,03 (0,03)	2,28 (0,07)	13	29	2
	MIR entier	Norme	SO: 0,025	4	0,85	0,83 (0,008)	1,01 (0,02)	2,44 (0,05)	18	10	3
Sucres_sol	NIR	Der1		5	0,91	0,88 (0,01)	0,03 (0,001)	2,92 (0,09)	0	72	3
	MIR découpé	Brut	SO: 0,001; W: 0,95	5	0,93	0,90 (0,007)	0,02 (0,0008)	3,20(0,11)	0	252	1
	MIR entier	Der2	SO: 0,0001	6	0,93	0,89 (0,01)	0,02 (0,001)	3,09 (0,14)	0	268	2
Gluc_sol	NIR	Der2Norm		4	0,87	0,85 (0,01)	0,01 (0,0002)	2,60 (0,06)	0	46	3
	MIR découpé	Norme	SO: 0,001	5	0,94	0,92 (0,004)	0,005 (0,0001)	3,56 (0,1)	0	72	1
	MIR entier	NormDer1	SO: 0,001; W: 0,95	6	0,95	0,93 (0,005)	0,004 (0,0001)	3,71 (0,13)	0	69	2
NonGluc_sol	NIR	Der1		4	0,88	0,86 (0,01)	0,02 (0,001)	2,70 (0,07)	0	157	3
	MIR découpé	Brut	SO: 0,0001	6	0,92	0,90 (0,008)	0,02 (0,0007)	3,11 (0,11)	0	314	2
	MIR entier	Brut	SO: 0,0001	6	0,93	0,90 (0,007)	0,02 (0,0007)	3,16 (0,11)	0	394	1
Gluc_sol_prop	NIR	Der2Norm		6	0,89	0,84 (0,01)	0,06 (0,002)	2,52 (0,10)	1	262	3
	MIR découpé	Der2Norm	SO: 0,001	4	0,88	0,85 (0,008)	0,05 (0,001)	2,62 (0,07)	0	52	2
	MIR entier	Norme	SO: 0,001	4	0,88	0,87 (0,006)	0,05 (0,001)	2,74 (0,06)	0	11	1
Sucres_tot	NIR	Der2Norm		5	0,83	0,78 (0,01)	0,03 (0,001)	2,17 (0,07)	20	251	3
	MIR découpé	Norme	SO: 0,025	6	0,85	0,80 (0,01)	0,03 (0,0008)	2,28 (0,07)	16	125	2
	MIR entier	NormDer2	SO: 0,02	5	0,84	0,80 (0,01)	0,03 (0,0007)	2,27 (0,06)	16	42	1
Gluc_tot	NIR	Der1		6	0,86	0,80 (0,02)	0,02 (0,001)	2,27 (0,09)	9	105	1
	MIR découpé	Norme	SO: 0,008; W: 0,90	5	0,91	0,85 (0,01)	0,02 (0,0008)	2,65 (0,13)	12	172	2
	MIR entier	NormDer2	SO: 0,0095; W: 0,90	6	0,91	0,85 (0,02)	0,02 (0,0009)	2,60 (1,13)	14	998	3
NonGluc_tot	NIR	Der1		4	0,88	0,85 (0,01)	0,03 (0,001)	2,61 (0,07)	4	236	3
	MIR découpé	Norme	SO: 0,0005	4	0,88	0,85 (0,008)	0,03 (0,0008)	2,62 (0,07)	0	42	1
	MIR entier	Der1Norm	SO: 0,001	5	0,87	0,83 (0,009)	0,03 (0,0008)	2,47 (0,06)	0	236	2
Gluc_tot_prop	NIR	Der1		5	0,89	0,85 (0,01)	0,03 (0,001)	2,63 (0,09)	1	80	3
	MIR découpé	Norme	SO: 0,001	3	0,88	0,86 (0,007)	0,03 (0,0008)	2,65 (0,06)	0	406	1
	MIR entier	NormDer2	SO : 0,001	4	0,89	0,86 (0,008)	0,03 (0,0009)	2,71 (0,07)	0	431	2
Sucres_hydrol	NIR	Norme		6	0,82	0,77 (0,02)	0,02 (0,001)	2,10 (0,07)	27	262	1
	MIR découpé	Der1	SO: 0,035	3	0,74	0,70 (0,02)	0,03 (0,0007)	1,85 (0,05)	27	16	2
	MIR entier	NormDer1	SO: 0,02	5	0,73	0,64 (0,02)	0,03 (0,001)	1,69 (0,05)	14	143	3
Gluc_hydrol	NIR	Der1Norm		6	0,87	0,83 (0,01)	0,02 (0,001)	2,41 (0,08)	9	67	2
	MIR découpé	Norme	SO: 0,005	4	0,81	0,77 (0,01)	0,02 (0,0006)	2,10 (0,05)	4	61	3
	MIR entier	NormDer2	SO: 0,005	5	0,85	0,81 (0,01)	0,02 (0,0006)	2,32 (0,07)	4	91	1
NonGluc_hydrol	NIR	NA/ Der2	NA/SO: 0,025	NA/2	NA/0,62	NA/0,57 (0,02)	NA/0,01 (0,0003)	NA/1,53 (0,04)	NA/39	NA/4	2
	MIR découpé	Der1	SO : 0,03	3	0,76	0,70 (0,02)	0,01 (0,0004)	1,86 (0,07)	40	79	1
	MIR entier	Der1	SO: 0,03	4	0,56	0,46 (0,04)	0,01 (0,0005)	1,37 (0,04)	34	44	3
Gluc_hydrol_prop	NIR	NormDer1		5	0,85	0,80 (0,01)	0,03 (0,001)	2,26 (0,07)	15	360	1
	MIR découpé	Der1Norm	SO: 0,02	5	0,84	0,80 (0,01)	0,03 (0,0008)	2,25 (0,06)	16	148	2
	MIR entier	NormDer1	SO: 0,02	3	0,82	0,78 (0,01)	0,03 (0,001)	2,15 (0,06)	14	91	3

NA: pas de modèle correct / modèle que j'ai sélectionné pour faire la comparaison.
SO: seuil pour la detection d'« outliers »
W: critère de Wold

Tableau 7 : modèles sélectionnés en proche et moyen infrarouge avec leur classement respectif pour chaque variable de référence

Dans le tableau ci-dessus j'ai résumé les modèles que j'ai sélectionnés en moyen infrarouge (MIR) pour le spectre entier et découpé mais également les modèles qui ont été sélectionné lors de l'étude en NIR (proche infrarouge) et que j'ai pu récupérer afin, pour chaque variable, de les classer du meilleur (rang 1) au moins bon (rang 3).

Pour le classement des modèles, j'ai adopté comme stratégie générale de comparer avant tout le R^2 de validation et le nombre d'« outliers », le meilleur modèle sera celui avec un R^2 assez élevé et le moins d'« outliers » possible. Si la distinction n'est pas possible, je regarde le nombre de composantes (le plus faible possible) et le RPD (le plus élevé possible).

Je vais à présent expliquer pour chaque variable le classement que j'ai effectué et croiser ces résultats avec les graphes des nombres d'onde sélectionnés présentés en Annexe 6.

- **tx_extract_sec :** Les 3 modèles ont un R^2 de validation croisée très similaire, mais diffèrent par le nombre d'« outliers ». En effet, le modèle obtenu en proche infrarouge a 5 « outliers » de moins que les 2 autres c'est pour cela que j'ai considéré qu'il était le meilleur. Les modèles en moyen infrarouge ont le même nombre d'« outliers » mais celui pour le spectre découpé à R^2 légèrement plus élevé que celui pour le spectre entier. Le modèle MIR découpé est donc meilleur que le MIR entier et on remarque que les nombres d'onde sélectionnés avec spectre entier se situent en grande partie dans la partie du spectre découpé mais aussi un peu à droite de cette zone.
- **tx_lign_tot_sec :** Les 3 modèles ont également environ le même R^2. Ce qui fait donc la différence est le nombre d'« outliers », le modèle en proche IR en ayant le moins, puis celui obtenu en MIR avec le spectre entier et enfin celui en MIR avec spectre découpé. Le modèle MIR entier est donc pour cette variable meilleur que le MIR découpé et on remarque que les nombres d'onde sélectionnés par le CARS se situent dans la partie du spectre découpé mais aussi à droite et à gauche de cette zone.
- **tx_holo_calc_sec :** J'ai estimé que le modèle NIR est le meilleur des 3 car son R^2 est plus élevé que celui des 2 autres et il présente le plus faible nombre d'« outliers ». Au rang 2, le modèle MIR spectre découpé a un R^2 un peu plus faible que celui du MIR spectre entier situé au rang 3 (de 0,03) mais a aussi 6 « outliers »de moins. Le modèle MIR découpé est donc meilleur que le MIR entier et on remarque que les nombres d'onde sélectionnés avec le spectre entier se situent surtout dans la partie du spectre découpé mais aussi un peu à droite de cette zone.
- **Sucres_sol :** Les trois modèles n'ont pas d'« outliers ». Ils se différencient donc pas leur R^2, le modèle du rang 1 (MIR découpé) étant celui dont le R^2 le plus élevé et celui du rang 3 (NIR) le R^2 le plus bas. Le modèle MIR découpé est meilleur que le MIR entier et on remarque que les nombres d'onde sélectionnés avec le spectre entier se situent dans la partie du spectre découpé et un peu à droite et à gauche de cette zone.
- **Gluc_sol :** Là aussi, les trois modèles n'ont pas d'« outliers », je les ai donc classés en fonction de leur R^2 mais aussi du nombre de composante lorsque les valeurs de R^2 étaient similaires. Le modèle NIR est nettement moins bon que ceux obtenus en MIR du point de vue du R^2, puis le modèle MIR découpé est meilleur que le MIR entier car pour des valeurs de R^2 similaires il a une composante de moins. On remarque que les nombres d'onde sélectionnés avec le spectre entier se situent dans la partie du spectre découpé et un peu à droite et à gauche de cette zone.
- **NonGluc_sol :** Les 3 modèles n'ont pas d'« outliers ». Le moins bon modèle est donc celui qui a le moins bon R^2, il s'agit là aussi du modèle pour le NIR. Les modèles en MIR ont le même R^2, le même nombre de composantes ce qui a donc fait la différence entre les 2 est le fait que le RPD du modèle sur spectre entier était plus élevé que celui obtenu pour le spectre découpé. On remarque que les nombres d'onde sélectionnés avec le spectre entier se situent en partie dans la zone de spectre découpé mais surtout à droite et à gauche de cette zone.
- **Gluc_sol_prop :** Le moins bon modèle a été obtenu en proche infrarouge car il présente 1 « outlier » tandis que les deux autres modèles n'en ont pas et il a aussi un R^2 moins élevé. Les modèles MIR ont le même nombre de composantes mais celui pour le spectre entier a un R^2 sensiblement plus élevé que celui pour le spectre découpé. On remarque que les nombres d'onde sélectionnés avec le spectre entier se situent dans la partie du spectre découpé mais également à gauche de cette zone.
- **Sucres_tot :** Au rang 3 se trouve le modèle en proche infrarouge, car il a le plus d'« outliers » et le moins bon R^2. Les 2 autres modèles ont le même nombre d'« outliers », le même R^2 et environ le même RPD. Cependant le modèle établi sur le spectre entier a une

composante de moins que celui effectué sur le spectre découpé. Le modèle MIR entier est donc meilleur que celui avec MIR découpé et on remarque que les nombres d'onde sélectionnés avec le spectre entier se situent dans la partie du spectre découpé mais également à gauche de cette zone.

- **Gluc_tot :** Le modèle du rang 1 est celui pour le proche infrarouge. Il se caractérise par un R^2 un peu plus faible que les modèles obtenus en moyen infrarouge mais surtout un nombre d'« outliers » plus faible. Les modèles en MIR ont le même R^2 mais celui sur spectre découpé a deux « outliers »de moins que celui sur spectre entier. Le modèle MIR découpé est donc meilleur que celui sur MIR entier et on remarque que les nombres d'onde sélectionnés avec le spectre entier se situent dans la partie du spectre découpé mais également à droite et à gauche de cette zone.
- **NonGluc_tot :** Les trois modèles ont environ le même R^2 mais le modèle en proche infrarouge a 4 « outliers »tandis que les 2 autres n'en ont aucun, il a donc été classé au rang 3. Le modèle MIR sur spectre découpé a un R^2 légèrement plus élevé que celui sur spectre entier et une composante de moins. Le modèle MIR découpé est donc meilleur que le MIR entier et on remarque que les nombres d'onde sélectionnés avec le spectre entier se situent dans la partie du spectre découpé et un peu à droite de cette zone.
- **Gluc_tot_prop :** Le moins bon modèle (rang 3) est celui obtenu en proche infrarouge. Il se caractérise par un R^2 légèrement plus élevé (de 0,01) que celui des modèles obtenu en MIR, un peu plus de composantes et un « outlier » tandis que les 2 autres n'en ont pas. Les modèles en MIR ont le même R^2 mais celui sur spectre découpé a une composante de moins. Le modèle MIR découpé est donc meilleur que le MIR entier et on remarque que les nombres d'onde sélectionnés avec le spectre entier se situent dans la partie du spectre découpé et un peu à droite et à gauche de cette zone.
- **Sucres_hydrol :** Le modèle en moyen infrarouge sur spectre entier est le moins bon (rang 3) car bien qu'il a moins d'« outliers » que les 2 autres modèles son R^2 est aussi beaucoup moins bon (0,64). Les 2 autres modèle ont le même nombre d'« outliers » mais celui en MIR avec spectre découpé (rang 2) a un R^2 égal à 0,70 ce qui est moins bon que le R^2 obtenu pour le modèle en proche infrarouge (0,77). Le modèle MIR découpé est meilleur que le MIR entier et on remarque que les nombres d'onde sélectionnés avec le spectre entier se situent dans la partie du spectre découpé et un peu à droite de cette zone.
- **Gluc_hydrol :** Le modèle MIR spectre découpé est celui qui a le R^2 le plus faible, il a donc été classé au rang 3. Ensuite, au rang 2 on trouve le modèle en proche infrarouge car bien qu'il a un R^2 plus élevé de 0,02 que celui en MIR spectre entier il a aussi 5 « outliers »de plus. Le modèle MIR entier est donc meilleur que le MIR découpé et on remarque que les nombres d'onde sélectionnés avec le spectre entier se situent dans la partie du spectre découpé mais également à droite et à gauche de cette zone.
- **NonGluc_hydrol :** N'étant pas de bonne qualité, j'ai classé les modèles sélectionnés ici uniquement pour en faire la comparaison en me basant sur le R^2. Le meilleur modèle est donc celui obtenir en MIR avec le spectre découpé, puis celui en NIR et enfin celui en MIR sur spectre entier. Cependant, on remarque que les nombres d'onde sélectionnés avec le spectre entier se situent dans la partie du spectre découpé, mais ce ne sont pas les mêmes.
- **Gluc_hydrol_prop :** Le moins bon modèle (rang 3) est celui obtenu en MIR sur spectre entier car il a seulement un ou deux « outliers »de moins que les 2 autres mais un R^2 plus faible. Les modèles du rang 1 et 2 ont les même valeurs de R^2 mais celui en proche infrarouge a un « outlier » de moins que celui en MIR spectre découpé, il a donc été classé au rang 1. Le modèle MIR découpé est meilleur que le MIR entier et on remarque que les nombres d'onde sélectionnés avec le spectre entier se situent dans la partie du spectre découpé et un peu à droite de cette zone.

On peut remarquer qu'en général lorsque le modèle MIR entier est meilleur que le MIR

découpé, il y a plus de nombres d'onde sélectionnés à gauche de la zone découpé. A droite de cette zone, on retrouve pour presque chaque variable des nombres d'onde sélectionnés.

Afin de tester si les nombres d'onde sélectionnés sur le spectre entier sont surreprésentés dans la région découpée, j'ai effectué pour chaque variable un test du Chi 2. Ceux-ci testent si la proportion de nombre d'onde sélectionnés par CARS (p) qui sont situés dans la partie du spectre découpé est égale à la proportion des nombres d'onde dans la zone découpée p_0 qui est égale à 0,269. Les résultats de ces tests sont présentés dans le Tableau 8.

dosages	**p-value**	**Proportion p**
tx_extract_sec	<2.2e-16	0,980
tx_lign_tot_sec	<2.2e-16	0,834
tx_holo_calc_sec	0.01809	0,600
Sucres_sol	<2.2e-16	0,896
Gluc_sol	<2.2e-16	0,826
NonGluc_sol	<2.2e-16	0,576
Gluc_sol_prop	1.652e-06	0,909
Sucres_tot	<2.2e-16	0,905
Gluc_tot	<2.2e-16	0,581
NonGluc_tot	<2.2e-16	0,894
Gluc_tot_prop	<2.2e-16	0,923
Sucres_hydrol	<2.2e-16	0,993
Gluc_hydrol	<2.2e-16	0,857
NonGluc_hydrol	<2.2e-16	1,000
Gluc_hydrol_prop	<2.2e-16	0,978

Tableau 8 : Récapitulatif des tests du Chi 2

Pour tous les dosages, la proportion de nombres d'onde sélectionnés par CARS et situés dans la région découpée est supérieure à la proportion attendue p0.

Pour un seuil alpha égal à 0,01 seule l'hypothèse nulle du test pour tx_holo_calc_sec n'est pas rejetée, tandis que pour un seuil égal à 0,05 pour toutes les variables l'hypothèse nulle est rejetée, c'est-à-dire que les nombres d'onde sélectionnés sont surreprésentés dans la région découpée.

On constate également que parmi les modèles MIR, les modèles correspondant au spectre découpé arrive dix fois en meilleure position que ceux sélectionnés avec le spectre entier. Si on compare les modèles NIR et MIR (sans distinction entre spectre découpé et entier), le moyen infrarouge occupe 9 fois le rang 1 contre 6 fois pour le proche infrarouge (Figure 24).

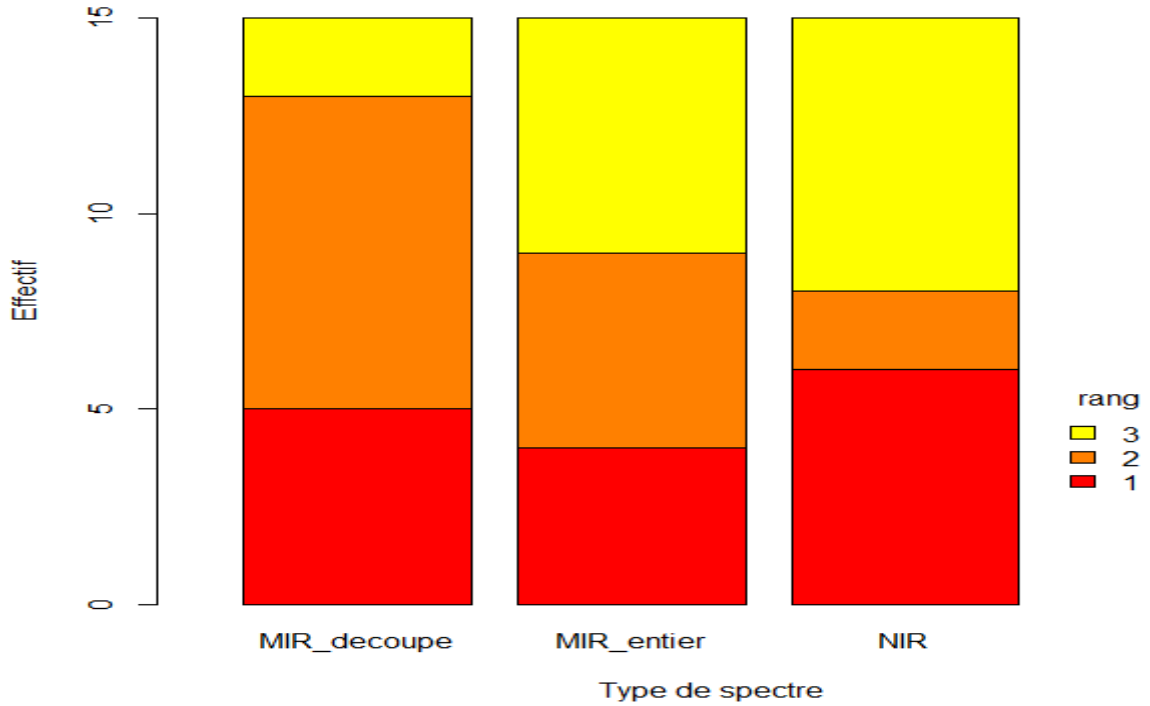


Figure 24 : Représentation graphique de la table de contingence des rangs en fonction des type de spectre pour l'ensemble des modèles de calibration sélectionnés.

4. Discussion

Faisons maintenant un bilan des résultats obtenus. On peut dire que nous avons atteint notre objectif de calibrer les données MIR pour les propriétés chimiques du bois mais aussi la saccharification (données provenant de Marseille) qui est une approche nouvelle sur le peuplier mais aussi en général (pour la saccharification).

Les seuls travaux similaires sur le peuplier pour des données MIR ont été effectués seulement pour la teneur en lignines du bois (Zhou *et al.*, 2011). Nous pouvons donc comparer nos résultats pour la teneur en lignines avec ceux de cet article, car les auteurs ont effectué une analyse similaire pour la teneur en lignine mais dans une autre espèce de Peuplier (hybrides interspécifiques). Ils ont tout d'abord en sélectionné les nombres d'onde dans la région 2000 – 700 cm^{-1} avec aucun prétraitement, dérivée première, normalisation et dérivée première puis normalisation et ont obtenu des modèles avec un R^2 de validation croisée compris entre 0,61 et 0,67 et entre 4 et 6 composantes. Ensuite, en faisant des restrictions automatiques des nombres d'onde et comme prétraitement une normalisation, ils ont obtenu un meilleur R^2 de validation de 0,806 mais avec beaucoup plus de composantes (12). Enfin, ils ont fait une restriction manuelle des nombres d'onde (1650 – 1380 cm^{-1}) avec normalisation, le R^2 de validation obtenu est alors de 0,734 et 3 composantes.

On peut dire que nous avons obtenu de meilleurs résultats pour la teneur en lignines. En effet, nos modèles ont des R^2 de validation croisée égal à 0,82 et 0,83 avec seulement 5 et 6 composantes tandis que pour un R^2 égal à 0,81 ils avaient 12 composantes et pour 6 composantes leur R^2 était moins bon. Nous avons cependant un peu plus d'« outliers » entre 10 et 13 % suivant les modèles (ce qui est encore convenable) alors qu'eux n'en ont que 8 %.

On a pu remarquer au cours de toutes ces analyses que les différentes variables ne sont pas toutes aussi facile à calibrer. En effet, en général les variables « _tot » et « _hydrol » ont été plus difficile à calibrer que les autres. Ceci peut être dû aux analyses chimiques effectuées, puisque ces données sont obtenues après une attaque enzymatique dont la répétabilité peut beaucoup varier. En particulier, nous n'avons pas réussi à obtenir de bons modèles pour NonGluc_hydrol. Pour cette variable on peut remarquer qu'il s'agit de celle qui a les corrélations les plus faibles avec les données spectrales et lors de l'ACP, elle est moins bien représentée sur le cercle de corrélation.

Je vais maintenant tenter de répondre à la question posée au début de ce rapport à savoir si les calibrations avec les spectres MIR sont meilleures que celle avec spectres NIR.

On ne peut pas réellement discerner quelle approche est la meilleure d'une manière générale. En effet, comme on peut le voir sur la Figure 24, les différents types de spectres arrivent environ le même nombre de fois au premier rang notamment compte tenu du nombre de données qui en plus ne sont pas indépendantes puisque les variables chimiques sont corrélées. On peut remarquer également que si NIR arrive le plus souvent en premier par rapport aux 2 autres c'est aussi le plus souvent le dernier et en ce sens il ne semble pas y avoir de réel avantage du NIR par rapport au MIR. On peut toutefois noter que MIR découpé arrive rarement dernier, seulement 2 fois et en ce sens pourrait donc présenter un avantage par rapport aux 2 autres.

Comparons désormais les 2 approches MIR découpé et entier. Les modèles du spectre découpé sont à 10 reprises meilleurs que ceux du spectre entier mais cela suffit-il à dire que cette approche est meilleure ?

L'intérêt du CARS est de sélectionner les nombres d'onde permettant d'obtenir une erreur de prédiction minimum. Si le CARS sélectionnait de façon optimale, les modèles sur le spectre entier devraient être aussi bon voir meilleur que ceux sur le spectre découpé, puisque le spectre entier englobe le spectre découpé. Or ce n'est plutôt pas le cas, car le spectre MIR découpé arrive plus souvent au rang 1 et au rang 2 que le spectre MIR entier. Cela souligne donc les limites du CARS et

l'intérêt de prédécouper le spectre.

En outre, les analyses Chi 2 confirment que les bandes sélectionnées sur le spectre entier sont surreprésentées dans la région découpée, confirmant l'intérêt du pré-découpage ainsi effectué. Toutefois, ce découpage pourrait être éventuellement amélioré, pour intégrer les nombres d'onde les plus faibles (à droite de la zone découpée) qui semblent contenir de l'information utile aux calibrations car fréquemment sélectionnées par le CARS sur spectre entier.

En ce qui concerne les prétraitements des spectres, nous avons essayé de savoir lesquels permettaient d'obtenir les meilleurs modèles. Comme le montre la Figure 25, on peut remarquer que 2 prétraitements se distinguent des autres à savoir la dérivée première (sélectionnée 11 fois) et la normalisation (sélectionnée 9 fois). Viennent ensuite et donc de façon logique la combinaison de ces 2 prétraitements, la dérivée 1ère puis normalisation (sélectionnée 6 fois) et la normalisation suivie de dérivation première (sélectionnée 5 fois). Les autres prétraitements sont sélectionnés entre 3 et 4 fois. On d se demander pour la dérivée seconde si nous avons sélectionné le bon nombre de point au cours de l'analyse de sensibilité car contrairement à la dérivée première, elle n'a été sélectionnée que 3 fois. En effet par manque de temps nous n'avons pas pu pousser cette analyse plus loin par exemple en effectuant les régressions puis en sélectionnant les meilleurs modèles. On peut aussi noter que nous avons sélectionné à 3 reprises des modèles dont les spectres n'ont subi aucun prétraitement. Ceci est surprenant car cela correspond au cas où le spectre est assez redondant comme nous l'a montré l'ACP dans laquelle presque toutes les variables étaient expliquées par le premier axe.

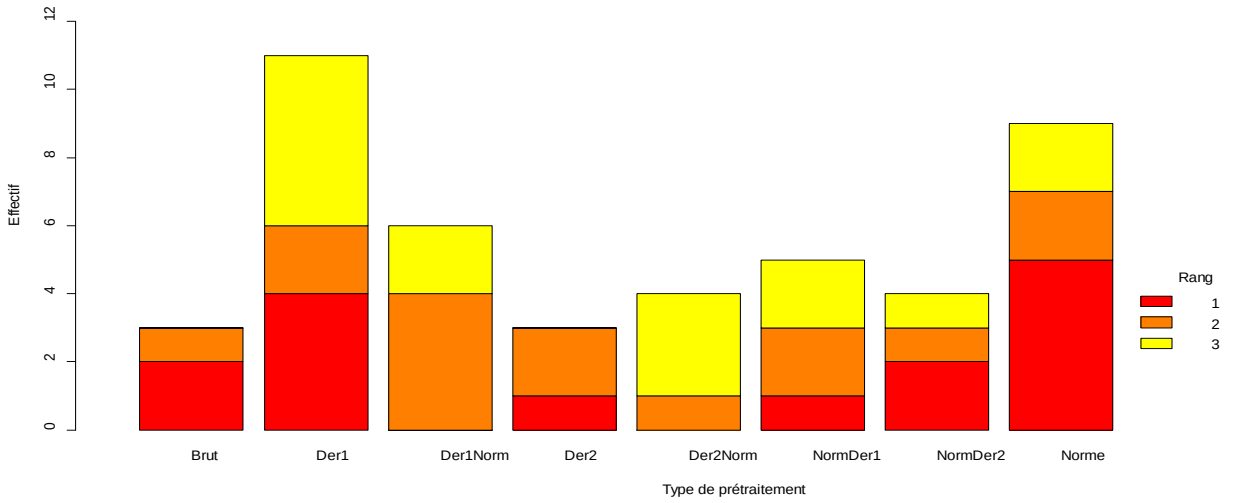


Figure 25 : représentation effectif dans chaque rang pour chaque type de prétraitement

Le fait de retirer des observations peut sembler étrange, pourquoi enlever des valeurs. Cette détection d'« outliers » s'explique par le fait que certains dosages effectués n'ont pas une bonne répétabilité, on s'attend donc à ce qu'il y ait des valeurs aberrantes. En les enlevant on améliore donc les modèles. Cependant nous ne voulons pas en enlever trop car le risque après serait d'avoir des modèles pas très représentatifs et pas efficace pour la prédiction. Le nombre de valeurs aberrantes que nous avons enlevé va de aucune à 40 qui est une valeur un peu extrême car il s'agit du modèle que nous n'avons pas réussi à calibrer et qui donc ne sera pas retenu pour la prédiction.

On peut envisager comme perspective à cette étude la validation des modèles. Ceci consiste à constituer un set de validation contenant quelques échantillons (un dizaine) bien répartis sur la gamme de variation des caractères prédits, effectuer les dosages sur ce set, les confronter aux valeurs prédites et estimer les statistiques R^2, RMSE et RPD de validation. Une fois ces modèles validés ils pourront être utilisés pour prédire les propriétés chimiques du bois ainsi que l'aptitude à la saccharification dans l'ensemble de la population d'étude afin d'effectuer des analyses génétiques.

Conclusion

Ce stage au sein de l'Unité de Recherche AGPF du centre INRA d'Orléans constitue ma première expérience professionnelle dans le domaine des statistiques.

La principale difficulté que j'ai rencontrée durant ce stage a été la compréhension du contexte de l'étude relevant du domaine de la biologie. Pour résoudre ce manque de connaissance en biologie forestière, j'ai lu de nombreux articles et documentations mais j'ai également pu voir quelques tests effectués en laboratoire avec des explications de la part de spécialistes.

Ce stage m'a permis d'obtenir des réponses à mes interrogations sur le métier de statisticien. J'ai pu voir ce qu'est le quotidien d'un statisticien, le travail effectué sur de grosses bases de données, les difficultés d'interprétation contrairement à l'université où nous travaillons sur de petites bases de données. J'ai ainsi pu me rendre compte que la réalité est plus complexe.

J'ai apprécié le travail qui m'a été confié, cela m'a permis d'enrichir mes connaissances en biologie forestière, sur la problématique de production de biomasse mais surtout sur diverses méthodes statistiques que je ne connaissais pas encore. J'ai pu également appliquer des outils statistiques vus à l'université et plus particulièrement le logiciel R.

Cette expérience m'a permis d'enrichir mes compétences professionnelles et humaines. Ce stage a confirmé mon choix d'exercer les statistiques et j'envisage de faire un prochain stage dans un autre domaine afin de découvrir d'autres secteurs.

En conclusion je peux dire que durant ce stage j'ai acquis de nouvelles compétences dans le domaine des statistiques.

Bibliographie

- Ballières H, Davrieux F et Ham-Pichavant F. 2002. Near infrared analysis as a tool for rapid screening of some major wood characteristics in a eucalyptus breeding program. Annals of Forest Science 59: 479-490.

- Bertrand D et Dufour E. 2006. La spectroscopie infrarouge et ses applications analytiques. Tec & Doc, 660 p.

- Faix O. 1991. Classification of lignins from different boitanical origins by FT-IR spectroscopy. Holzforschung 45: 21-27.

- Giordanengo T. 2004 Analyse de la variabilité génétique de quatre propriétés chimiques du bois d'hybrides d'Eucalyptus urophylla x E. grandis par utilisation de la Spectrométrie Proche Infrarouge. Rapport de Stage

- Li H, Liang Y, Xu Q et Cao D. 2009. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. Analytica Chimica Acta 648: 77-84.

- Schwanninger M, Rodrigues JC et Fackler K. 2011. A review of band assignments in near infrared spectra of wood and wood components. Journal of near infrared spectroscopy 19: 287-308

- Zhou G, Taylor G and Polle A. 2001. FTIR-ATR-based prediction and modelling of lignin and energy contents reveals independent intra-specific variation of these traits in bioenergy poplars. Plant Methods 7: 9.

- **Sites internet :**

  - ***INRA***
  http://www.orleans.inra.fr/
  http://www.inra.fr/

  - ***Logiciel R :***
  http://www.duclert.org/
  http://books.google.fr/books?id=59KNFF_8aMIC&pg=PT156&lpg=PT156&dq=couleurs+rgb+logiciel+R&source=bl&ots=p1_TrkRDT1&sig=l49CG_rDLUPzrxM0w0JmeUUPJc4&hl=fr&sa=X&ei=QKz-T8wgwczRBbjcrPcG&ved=0CG4Q6AEwCA#v=onepage&q=couleurs%20rgb%20logiciel%20R&f=false (pour les couleurs rgb)

  - ***Spectroscopie :***
  http://ebureau.univ-reims.fr/slide/files/quotas/SCD/theses/exl-doc/GED00000629.pdf
  http://www.csim.cnrs.fr/html/developpement%20application.pdf
  http://books.google.fr/books?id=buQHl2Rj9ycC&printsec=frontcover&dq=Etalonnage+multidimension-

nel:+application+aux+donn%C3%A9es+spectrales&hl=fr&sa=X&ei=HK0GUJuBA4rKtAaV-fXuBg&ved=0CD8Q6AEwAA#v=onepage&q=Etalonnage%20multidimensionnel%3A%20application%20aux%20donn%C3%A9es%20spectrales&f=false

- ***Statistiques :***

http://juan.rosas.free.fr/pdfs/pls.pdf
http://books.google.fr/books?id=OesjK2KZhsAC&printsec=frontcover&hl=fr&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false
https://www.rocq.inria.fr/axis/modulad/archives/numero-30/chavent-30/chavent-30.pdf
http://en.wikipedia.org/wiki/Cross-validation_%28statistics%29

- **Cours de Master 1 Statistiques et Recherche Opérationnelle :**
  - CHAUVEAU Didier, Statistiques descriptives, M1 2e semestre
  - DELSOL Laurent et EMILION Richard, Statistiques Générales, M1 2e semestre
  - JACQUOT Sophie, Analyse de données, M1 2e semestre

- **Toutes les photos du rapport proviennent de l'INRA.**

Annexes

Annexe 1 : code R

- Fonctions utilisées pour les calibrations :

Utilitaires

```
#détermination du nombre de composantes par le critère de Wold

nbcomp<-function(model,criterion)
        {R<-vector(mode='numeric',length=model$ncomp)
        for(i in 1:model$ncomp)
                {R[i]<-(model$validation$PRESS[i+1]/model$validation$PRESS[i])}
        if(max(R,na.rm=TRUE)<criterion) {model$ncomp} else {which(R>=criterion)[1]}
        }

#identification d' « outliers »

outlier<-function(model,threshold,criterion)
        {obs_pred_dif<-model$model$trait-model$validation$pred[,,nbcomp(model,criterion)]
        sigma<-vector(mode='numeric',length=length(obs_pred_dif))
        for (i in 1:length(obs_pred_dif)){sigma[i]<-sd(obs_pred_dif[-i])}
        zscore<-abs(obs_pred_dif)/sigma
        p_value<-1-pnorm(zscore)
        which(p_value<=threshold)
        }

#calcul de la statistique RPD

RPD<-function(model,criterion)
        {sd(model$model$trait)/(RMSEP(model,estimate='CV',ncomp=nbcomp(model,criterion))$v
al[2])
        }

#création d'un tableau de « sortie » pour les calibrations
outstat<-function(model,criterion)
        {c('ncomp'=nbcomp(model,criterion),
        'R2_train'=R2(model,estimate='train',ncomp=nbcomp(model,criterion))$val[2],
        'R2_cv'=R2(model,estimate='CV',ncomp=nbcomp(model,criterion))$val[2],
        'RMSE_cv'=RMSEP(model,estimate='CV',ncomp=nbcomp(model,criterion))$val[2],
        'RPD_cv'=RPD(model,criterion))
        }
```

Détection d'« outliers »

```
drop_outliers_LOO<-function(data_frame,trait,maxcomp=20,threshold=0.01,criterion=1,
        maxsteps=20)
        {outliers<-list()
        modeles<-list()
#step 1
        data_set<-data.frame('echantillon'=data_frame$Echantillon,
        'trait'=data_frame[,which(colnames(data_frame)==trait)],'MIRS'=data_frame$MIRS)
        modeles[[1]]<-plsr(trait~MIRS,data=data_set,ncomp=maxcomp,validation='LOO')
```

```
        outliers[[1]]<-data_set[outlier(modeles[[1]],threshold,criterion),]$echantillon
        cat(1)
#step 2
        data_set<-data_set[!data_set$echantillon %in% outliers[[1]],]
        modeles[[2]]<-plsr(trait~MIRS,data=data_set,ncomp=maxcomp,validation='LOO')
        outliers[[2]]<-c(outliers[[1]],data_set[outlier(modeles[[2]],threshold,criterion),]$echantillon)
        cat(2)
#steps 3 to i
        for (i in 3:maxsteps)
                {if (identical(outliers[[i-2]],outliers[[i-1]])) break else {
                data_set<-data_set[!data_set$echantillon %in% outliers[[i-1]],]
                modeles[[i]]<-plsr(trait~MIRS,data=data_set,ncomp=maxcomp,validation='LOO')
                outliers[[i]]<-c(outliers[[i-1]],
                data_set[outlier(modeles[[i]],threshold,criterion),]$echantillon)}}
                cat(i)}
        list('data'=data_set,'model_0'=modeles[[1]],'model_ok'=modeles[[length(modeles)]],'outliers
'=outliers[[length(outliers)]])}
```

Validations croisées Monte-Carlo

```
MCCV<-function(data_set,trait,Xmat,maxcomp=20,fold=10,iter=100,criterion=1)
        {ncomp<-maxcomp
        data_ok<-data_set
        colnames(data_ok)[which(colnames(data_ok)==trait)]<-'trait'
        colnames(data_ok)[which(colnames(data_ok)==Xmat)]<-'spectra'
        model<-mvr(trait~spectra,data=data_ok,ncomp,method='simpls')
        R2_train<-1-(colSums(model$residuals[,1,]^2)/(var(data_ok$trait)*(nrow(data_ok)-1)))
        MC_PRESS_mat<-matrix(NA,nrow=ncomp,ncol=iter)
        pred_mat<-array(NA,dim=c(nrow(data_ok),ncomp,iter))
        dimnames(pred_mat)[[1]]<-rownames(data_ok)
        dimnames(pred_mat)[[2]]<-paste(c(1:ncomp),"_comp",sep="")
        dimnames(pred_mat)[[3]]<-paste("iter_",c(1:iter),sep="")
        for (i in 1:iter)
                {CV<-crossval(model,segments=fold,data=data_ok,segment.type='random')
                MC_PRESS_mat[,i]<-CV$validation$PRESS
                pred_mat[,,i]<-CV$validation$pred[,1,]
                rm(CV)
                cat(i)}
        cat('\n')
        MC_R2_CV_mat<-1-(MC_PRESS_mat/(var(data_ok$trait)*(nrow(data_ok)-1)))
        MC_RMSEP_mat<-sqrt(MC_PRESS_mat/nrow(data_ok))
        MC_RPD_mat<-sd(data_ok$trait)/MC_RMSEP_mat
#nb comp
        R<-vector(mode='numeric',length=ncomp)
        for(i in 1:ncomp)
                {R[i]<-(rowMeans(MC_PRESS_mat)[i+1]/rowMeans(MC_PRESS_mat)[i])}
        if(max(R,na.rm=TRUE)>=criterion)
                {n_comp=which(R>=criterion)[1]}else{n_comp=length(R)}
#statistiques
#moyennes
        R2_mccv_mean<-rowMeans(MC_R2_CV_mat)
        RMSE_mccv_mean<-rowMeans(MC_RMSEP_mat)
```

```
        RPD_mccv_mean<-rowMeans(MC_RPD_mat)
#ecartypes
        R2_mccv_sd<-apply(MC_R2_CV_mat,1,sd)
        RMSE_mccv_sd<-apply(MC_RMSEP_mat,1,sd)
        RPD_mccv_sd<-apply(MC_RPD_mat,1,sd)
#moyenne des valeurs prédites à chaque iteration pour le nombre de composantes sélectionnées
        pred_mat_selcomp_ok<-rowMeans(pred_mat[,n_comp,])
#output
        output<-c('ncomp'=n_comp,'R2_train'=as.numeric(R2_train[n_comp]),
        'R2_MCCV_mean'=R2_mccv_mean[n_comp],'R2_MCCV_sd'=R2_mccv_sd[n_comp],
        'RMSE_MCCV_mean'=RMSE_mccv_mean[n_comp],
        'RMSE_MCCV_sd'=RMSE_mccv_sd[n_comp],
        'RPD_MCCV_mean'=RPD_mccv_mean[n_comp],
        'RPD_MCCV_sd'=RPD_mccv_sd[n_comp])

        list('model'=model,'output'=output,'predicted'=pred_mat_selcomp_ok)}
```

CARS

```
carspls_LOO<-function(X,y,nLV=2,iteration=50,criterion=1{
        Num_row<-nrow(X)
        Num_col<-ncol(X)
        RMSECV<-rep(0,iteration)
        NumLV<-rep(0,iteration)
        VarIndex<-1:Num_col
        subsetVariable<-1:Num_col
        Coef<-matrix(rep(0,Num_col*iteration),Num_col)
        Nvar<-rep(0, iteration)
        ratio0<-1
        ratio1<-2/Num_col
        b=log(ratio0/ratio1)/(iteration-1)
        a=ratio0*exp(b)
        for (iter in 1:iteration) {
        Xcal <- X[,subsetVariable]
        data.CARS.cal<-data.frame(indepdent=Xcal,response=y)
        nLV<-min(c(nLV,dim(Xcal)))
        ncomp=nLV
        CV<-plsr(response~.,data=data.CARS.cal,ncomp,validation='LOO',method='simpls')
        PRESS<-CV$validation$PRESS
        RMSECV.temp=sqrt(PRESS/Num_row)
        R<-vector(mode='numeric',length=nLV)
        for(i in 1:(nLV-1)) {R[i]<-(PRESS[i+1]/PRESS[i])}
        NumLV[iter]<-if(max(R)>=criterion)
                {which(R>=criterion)[1]}else{length(R)}#else{which.max(R)}
        RMSECV[iter]<-RMSECV.temp[NumLV[iter]]
        coef0<-rep(0,Num_col)
        coef.iter<-CV$coefficients[,,NumLV[iter]]
        coef0[subsetVariable]<-coef.iter
        Coef[,iter]<-coef0
        weight<-abs(coef0)
        weight.order<-order(weight,decreasing=TRUE)
        ratioVariable<-a*exp(-b*(iter+1))
```

```
        Nvar[iter]<-length(which(coef0!=0))
        K<-round(Num_col*ratioVariable)
        weight[weight.order[K+1:Num_col]]<-0
        subsetVariable<-which(weight!=0)
        screen.output<-paste("The", iter, "th CARS-PLS iteration finished.")
        print(screen.output)}
        MinError<-min(RMSECV)
        OPT.iter<-which(RMSECV==MinError)
        OPT.iter<-OPT.iter[length(OPT.iter)]
        SelectedVariables<-which(Coef[,OPT.iter]!= 0)
        CARS<-list(Coef=Coef,Nvar=Nvar,RMSECV=RMSECV,NumLV=NumLV,
        Optimal.iteration=OPT.iter,MinError=MinError,SelectedVariables=SelectedVariables)
        return(CARS)}
```

Graphique des nombres d'onde sélectionnés par CARS pour chaque variable (dosage

```
onde_plot<- function(dosage)
        {matplot(data_norm$lambda,data_norm$MIRS,type='l',lty=1,pch=0,
        xlab='Lambda cm-1',ylab='Absorbance',xlim=c(4000,500),xaxs='i')
        abline(v=dosage$entier[,1],col=rgb(0.1,0.8,0.1,alpha=0.8))
        abline(v=dosage$decoup[,1],col=rgb(0.8,0.1,0.1,alpha=0.2))
        abline(v=c(1900,800),lty = 'dotted')
    }
```

- Codes d'exécutions :

chargement des données

```
data_MIR=read.table('data_MIR.txt',header=T)
data_MIR_ok<-
data.frame(lambda=data_MIR[,1],MIRS=I(as.matrix(data_MIR[,2:ncol(data_MIR)])))

#graphe des spectres bruts allant de 4000 à 650 cm-1
matplot(data_MIR_ok$lambda,data_MIR_ok$MIRS,type='l',lty=1,pch=0,xlab='Lambda cm-1', ylab
= 'Absorbance',xlim=c(4200,450),xaxs='i')
```

découpe du spectre (1800 -> 900 cm-1)

```
data_sub<-subset(data_MIR_ok, lambda >= 900 & lambda <= 1800)

matplot(data_sub$lambda, data_sub$MIRS, type = 'l', lty = 1, pch = 0, xlab = 'Lambda cm-1', ylab
= 'Absorbance', xlim = c(1900,800), xaxs='i')
```

normalisation du spectre (centrage, réduction)

```
normalize <- function(x) ((x-mean(x))/sd(x))

data_norm<-
data.frame(lambda=data_sub$lambda,MIRS=I(as.matrix(apply(data_sub$MIRS,2,normalize))))

matplot(data_norm$lambda, data_norm$MIRS, type = 'l', lty = 1, pch = 0, xlab = 'Lambda cm-1',
ylab = 'Absorbance', xlim=c(1900,800),xaxs='i')
```

derivation des spectres
calcul de dérivées 1eres, filtre de savitzky-golay (bibliothèque signal)
fonction sgolayfilt, p=ordre du polynome, n=nombre de points pour dériver, m=ordre de la dérivée
ts=facteur d'échelle lié au pas de nombre d'onde

```
library(signal)
tsf<-(max(data_sub$lambda)-min(data_sub$lambda))/(length(data_sub$lambda)-1)

data_der1<-
data.frame(lambda=data_sub$lambda,MIRS=I(as.matrix(apply(data_sub$MIRS,2,function(x){sgol
ayfilt(x,p=2,n=75,m=1,ts=tsf)})))
matplot(data_der1$lambda, data_der1$MIRS, type = 'l', lty = 1, pch = 0, xlab = 'Lambda cm-1',
ylab = 'Absorbance', xlim=c(1900,800),xaxs='i')

data_der2<-
data.frame(lambda=data_sub$lambda,MIRS=I(as.matrix(apply(data_sub$MIRS,2,function(x){sgol
ayfilt(x,p=3,n=101,m=2,ts=tsf)})))
matplot(data_der2$lambda, data_der2$MIRS, type = 'l', lty = 1, pch = 0, xlab = 'Lambda cm-1',
ylab = 'Absorbance',xlim=c(1900,800),xaxs='i')

#der1 + norm
data_der1_norm<-
data.frame(lambda=data_der1$lambda,MIRS=I(as.matrix(apply(data_der1$MIRS,2,normalize))))
matplot(data_der1_norm$lambda,data_der1_norm$MIRS,type='l',lty=1,pch=0,xlab='Lambda cm-
1',ylab='Absorbance',xlim=c(1900,800),xaxs='i')

#der2 + norm
data_der2_norm<-
data.frame(lambda=data_der2$lambda,MIRS=I(as.matrix(apply(data_der2$MIRS,2,normalize))))
matplot(data_der2_norm$lambda,data_der2_norm$MIRS,type='l',lty=1,pch=0,xlab='Lambda cm-
1',ylab='Absorbance',xlim=c(1900,800),xaxs='i')

#norm + der1
data_norm_der1<-
data.frame(lambda=data_norm$lambda,MIRS=I(as.matrix(apply(data_norm$MIRS,2,function(x){s
golayfilt(x,p=2,n=75,m=1,ts=tsf)})))
matplot(data_norm_der1$lambda,data_norm_der1$MIRS,type='l',lty=1,pch=0,xlab='Lambda cm-
1',ylab='Absorbance',xlim=c(1900,800),xaxs='i')

#norm + der2
data_norm_der2<-
data.frame(lambda=data_norm$lambda,MIRS=I(as.matrix(apply(data_norm$MIRS,2,function(x){s
golayfilt(x,p=3,n=101,m=2,ts=tsf)})))
matplot(data_norm_der2$lambda,data_norm_der2$MIRS,type='l',lty=1,pch=0,xlab='Lambda cm-
1',ylab='Absorbance',xlim=c(1900,800),xaxs='i')
```

analyse des données spectrales (à adapter à chaque variable)

```
##stat descriptives (moy, sd)
stat_descr <-
data.frame(lambda=data_sub$lambda,MEAN=rowMeans(data_sub$MIRS),SD=apply(data_sub$MI
```

```
RS,1,sd))
matplot(data_sub$lambda,data_sub$MIRS,type='l',lty=1,pch=0,xlab='Lambda cm-
1',ylab='Absorbance',xlim=c(1900,800),ylim= c(0.0,0.5))
par(new=T)
plot(stat_descr$lambda,stat_descr$MEAN,type='l',xlim=c(1900,800),xlab='',ylab='',lwd=4,ylim=
c(0.0,0.5))
par(new=T)
plot(stat_descr$lambda,stat_descr$SD,type =
'l',xlim=c(1900,800),xlab='',ylab='',lwd=4,lty=3,axes=F)
axis(4)

##ACP spectre brut
pca<- prcomp(t(data_sub$MIRS),scale=TRUE)
summary(pca)
scores <- pca$x
loadings_ <- pca$rotation                #poids
plot(loadings_[,1], loadings_[,2],xlab='PC1',ylab='PC2')
plot(scores[,1], scores[,2],xlab='PC1',ylab='PC2')

matplot(data_sub$lambda,data_sub$MIRS,type='l',lty=1,pch=0,xlab='Lambda cm-
1',ylab='Absorbance',xlim=c(1900,800))
par(new=T)
plot(stat_descr$lambda,loadings_[,1],type = 'l',xlim=c(1900,800),xlab='',ylab='',lwd=4,axes=F)
axis(4)

##ACP avec dudi.pca
#brut
test<-matrix(t(data_sub$MIRS),nrow=100,ncol=901)
rownames(test)<-colnames(data_sub$MIRS)
colnames(test)<-(data_sub$lambda)
test_<-as.data.frame(test)

acp=dudi.pca(test)
pve=100*acp$eig/sum(acp$eig)
cumsum(pve)
comp=data.frame(acp$li$Axis1,acp$li$Axis2)
s.label(comp,xax=1,yax=2)
s.corcircle(acp$co,xax=1,yax=2)
```

analyse des données chimiques (à adapter à chaque variable)

```
#lire fichier données chimiques
donnees_chimiques=read.table('donnees_chimiques.txt',header=T)

#histogrammes pour voir la distribution des variables
hist(donnees_chimiques$tx_extract_sec)

#Corrélations
cor(donnees_chimiques[,c(7:21)])

#ACP
```

```
donnees_chimiques1 = donnees_chimiques[,-c(1,2,3,4,5,6)]
rownames(donnees_chimiques1)=donnees_chimiques[,4]
acp_dc1=dudi.pca(donnees_chimiques1)
pvedc1=100*acp_dc1$eig/sum(acp_dc1$eig)
cumsum(pvedc1)
s.label(acp_dc1$li[,1:2],xax=1,yax=2)
s.corcircle(acp_dc1$co,xax=1,yax=2)
iner<-inertia.dudi(acp_dc1,col.inertia=T,row.inertia=T)
iner$col.abs
```

Combinaison des données spectrales et des variables chimiques (à adapter à chaque variable)

```
###données spectrales normées
##tx_lign_tot_sec
#corrélation paramétrique
cor(donnees_chimiques[,8],t(data_norm$MIRS))
hist(cor(donnees_chimiques[,8],t(data_norm$MIRS)))

matplot(data_norm$lambda,data_norm$MIRS,type='l',lty=1,pch=0,xlab='Lambda cm-
1',ylab='Absorbance',xlim=c(1900,800),ylim= c(-1.5,3.0))
par(new=T)
plot(data_norm$lambda,abs(cor(donnees_chimiques[,8],t(data_norm$MIRS))),type =
'l',xlim=c(1900,800),xlab='',ylab='',lwd=4,ylim= c(0,0.6),axes=F)
axis(4)

#corrélation de spearman
cor(donnees_chimiques[,8],t(data_norm),method="spearman")
hist(cor(donnees_chimiques[,8],t(data_norm$MIRS),method="spearman"))

matplot(data_norm$lambda,data_norm$MIRS,type='l',lty=1,pch=0,xlab='Lambda cm-
1',ylab='Absorbance',xlim=c(1900,800),ylim= c(-1.5,3.0))
par(new=T)
plot(data_norm$lambda,abs(cor(donnees_chimiques[,8],t(data_norm$MIRS),method="spearman"))
,type = 'l',xlim=c(1900,800),xlab='',ylab='',lwd=4,ylim= c(0,0.6),axes=F)
axis(4)
```

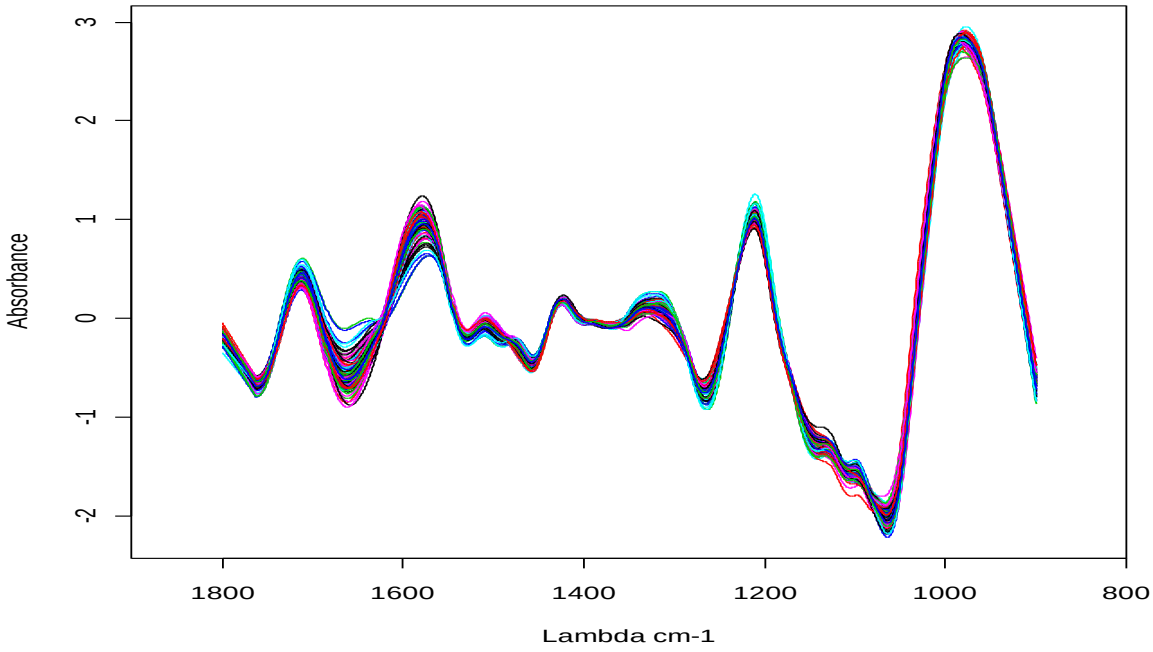
Calibration des données (à adapter à chaque variable)

```
#chargement analyses de référence
data_ref_ok<-read.table('donnees_chimiques.txt',sep='\t',header=TRUE)

#préparation des données pour calibration
data_calib<-merge(data_ref_ok,data_t,by='MIR_id')

# détection d'outliers en utilisant LOO CV
calib_data<- drop_outliers_LOO (data_frame=data_calib, trait=chem, maxcomp=20,
threshold=0.0005, criterion=1, maxsteps=100)

#Modèles avec données filtrées
data_calib_filt<-calib_data$data

##spectres bruts
```

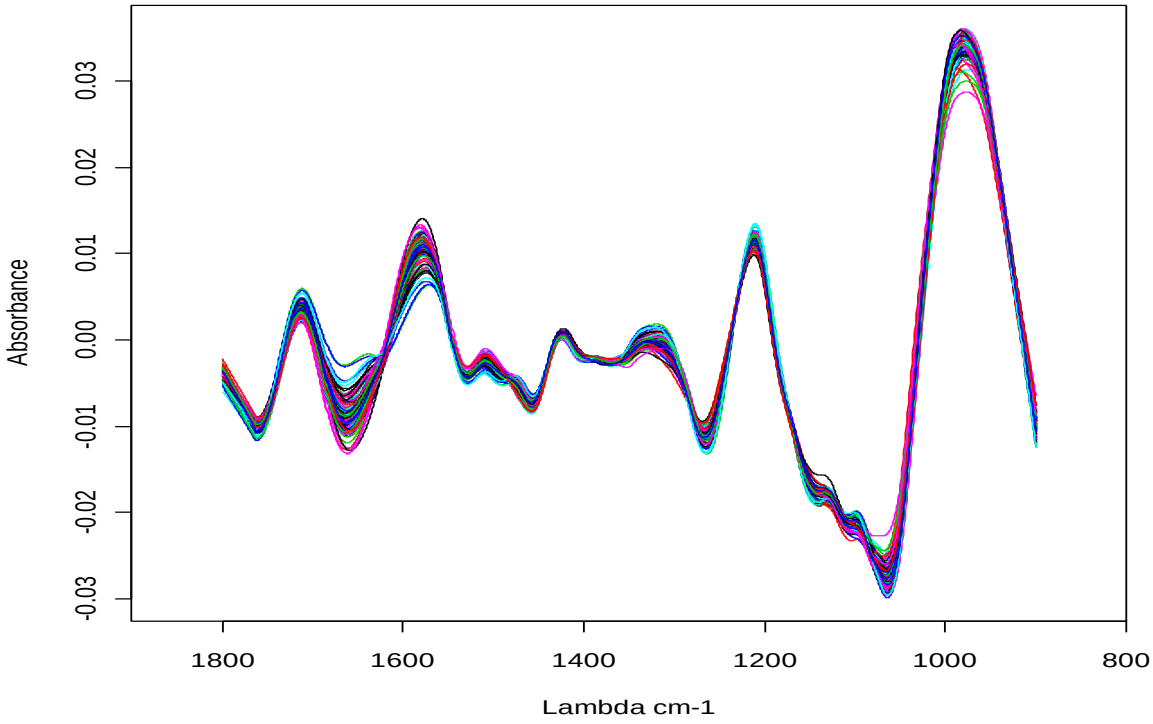
```
# régression PLS LOO
calibf_plsr<-plsr(trait~MIRS,data=data_calib_filt,ncomp=10,validation='LOO')
summary(calibf_plsr)
nbcomp(calibf_plsr,1)
plot(RMSEP(calibf_plsr))
plot(R2(calibf_plsr,estimate="CV"))
R2(calibf_plsr)
RPD(calibf_plsr,1)
par(mfrow=c(3,3))
plot(calibf_plsr,ncomp=c(1:9),line=T)
explvar(calibf_plsr)
par(mfrow=c(1,1))
matplot(data_sub$lambda,data_sub$MIRS,type='l',lty=1,pch=0,xlab='Lambda cm-
1',ylab='',xlim=c(1900,800),xaxs='i',axes=F)
axis(4)
par(new=T)
plot(calibf_plsr,"loadings",comps=1:3,legendpos="topleft",labels="numbers",xlab=',lwd=3,xlim=c(
1900,800))
calibf_plsr$validation$PRESS

#Régression PLS MCCV
calibf_data_MCCV<-
MCCV(data_set=data_calib_filt,trait=chem,Xmat='MIRS',maxcomp=20,fold=4,iter=1000,criterion
=1)
calibf_data_MCCV$output
plot(data_calib_filt[,colnames(data_calib_filt)=='trait'],calibf_data_MCCV$predicted,xlab='observe
d',ylab='predicted')
abline(0,1)

##CARS sur données filtrées
#spectre brut
calibf_data_CARS<-
carspls_LOO(X=data_calib_filt$MIRS,y=data_calib_filt[,colnames(data_calib_filt)=='trait'],nLV=2
0,iteration=400)
matplot(data_norm$lambda,data_norm$MIRS,type='l',lty=1,pch=0,xlab='Lambda cm-
1',ylab='Absorbance',xlim=c(1900,800),xaxs='i')
abline(v= as.integer(colnames(data_calib_filt$MIRS)[calibf_data_CARS$SelectedVariables]))

data_calibf_CARS<-data_calib_filt
data_calibf_CARS$MIRS<-data_calibf_CARS$MIRS[,calibf_data_CARS$SelectedVariables]
```

Annexe 2 : représentation des spectres pour différents prétraitements

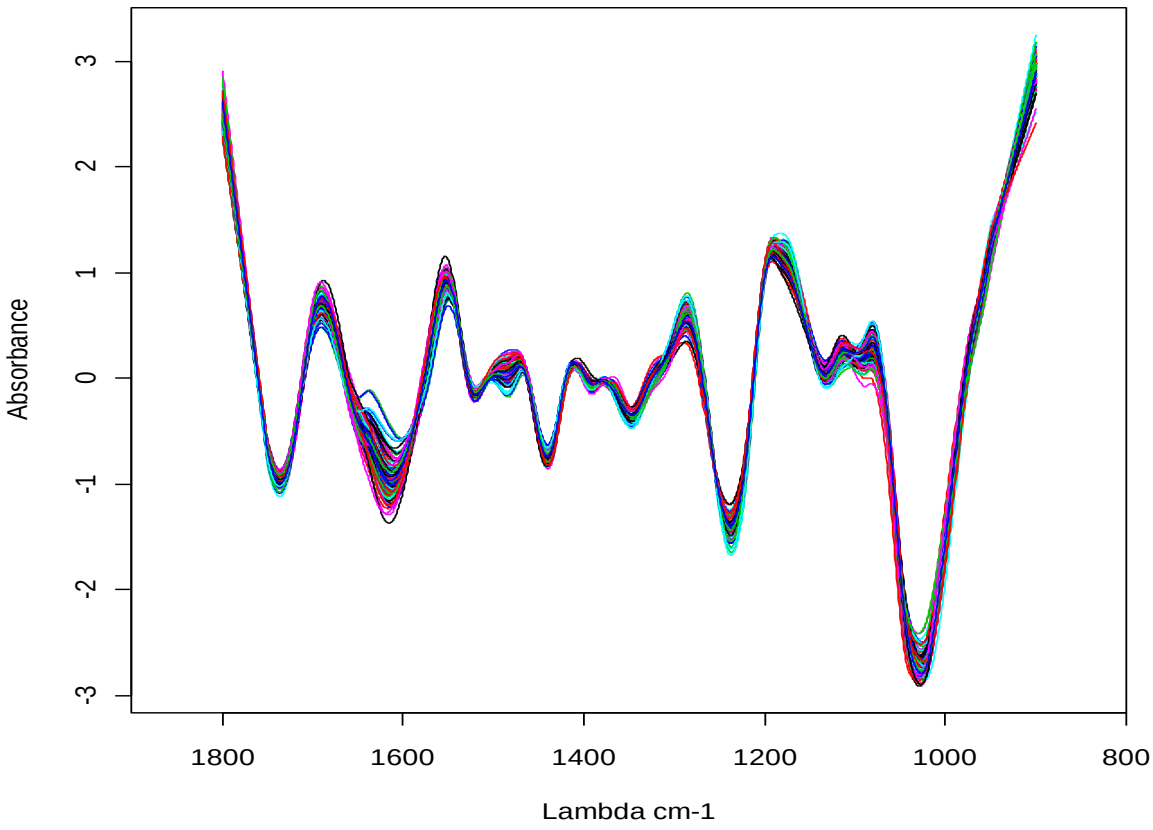
Dérivée première puis normalisation :



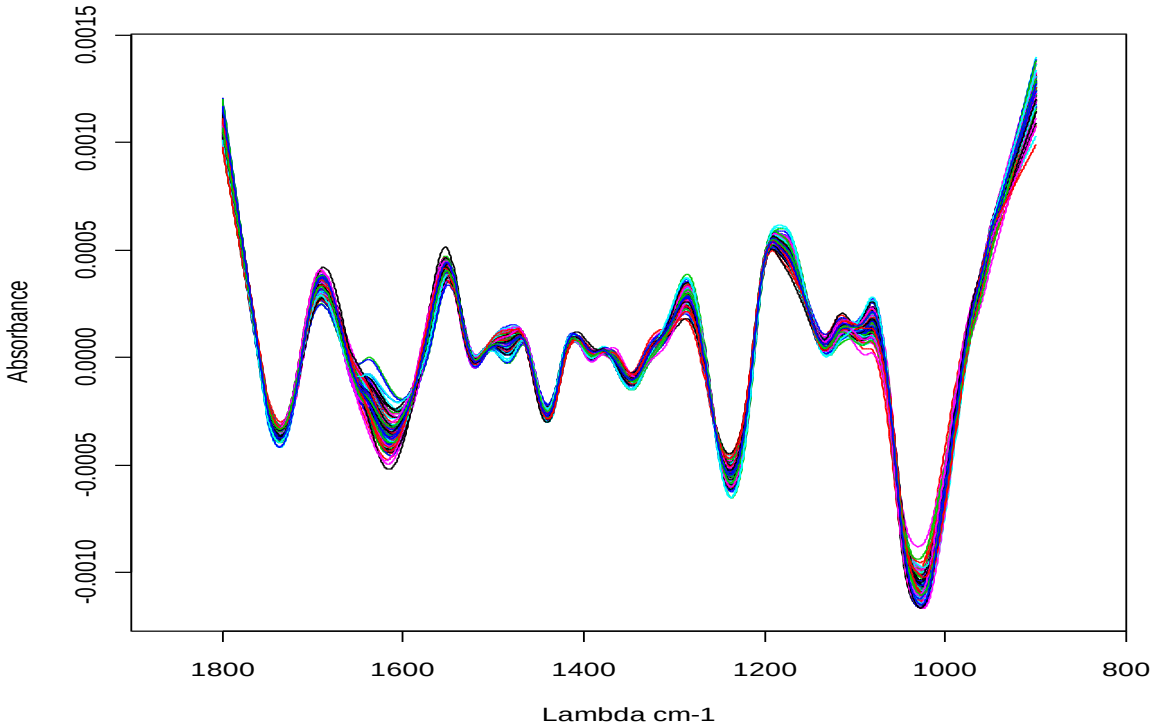
Normalisation puis dérivée première :



Dérivée seconde puis normalisation :

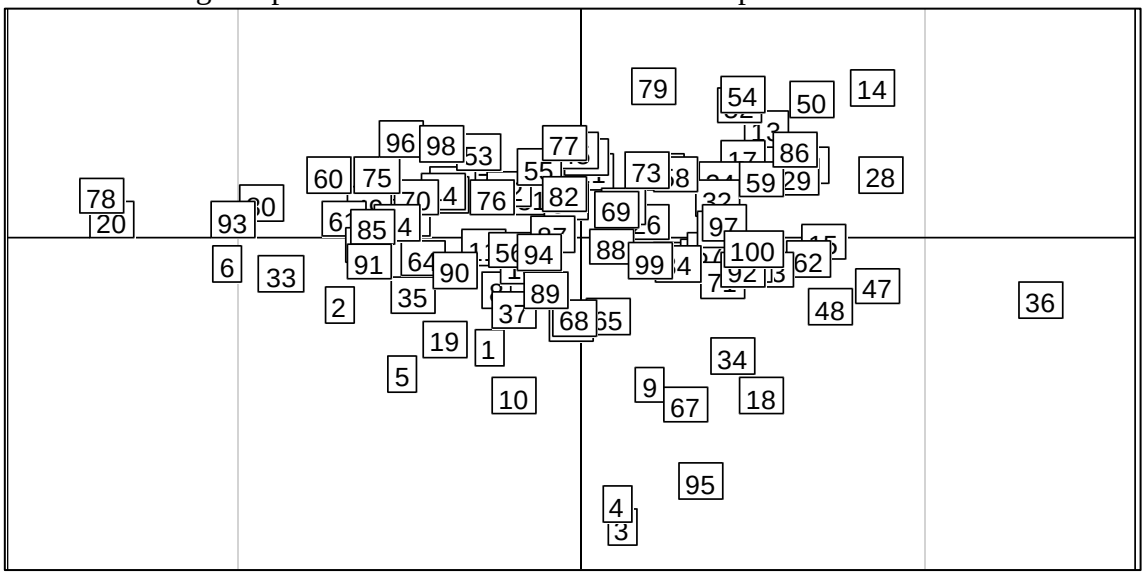


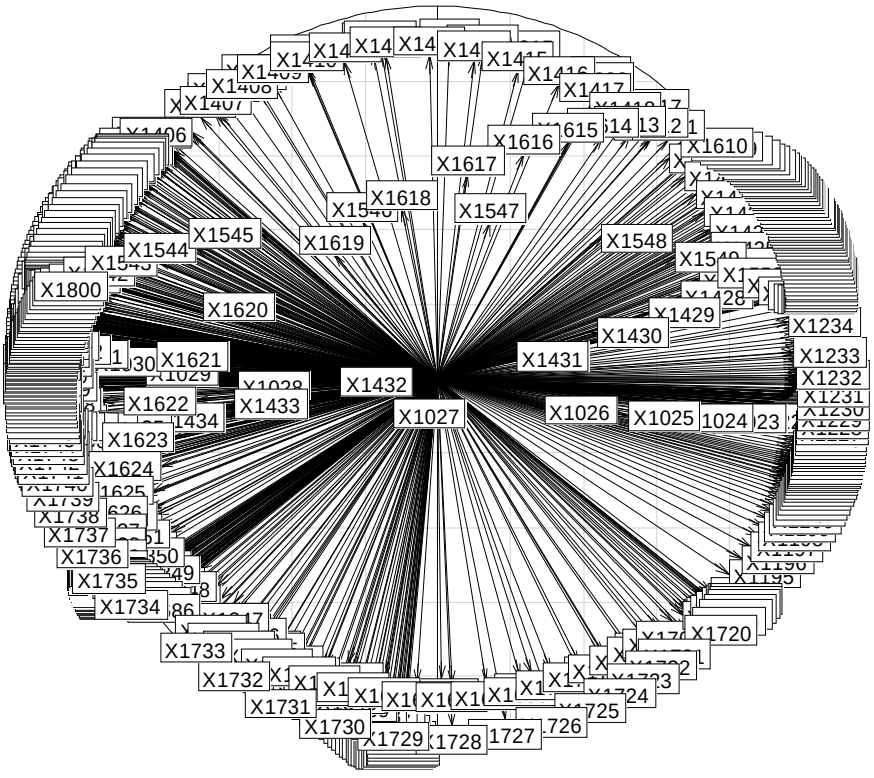
Normalisation puis dérivée seconde :



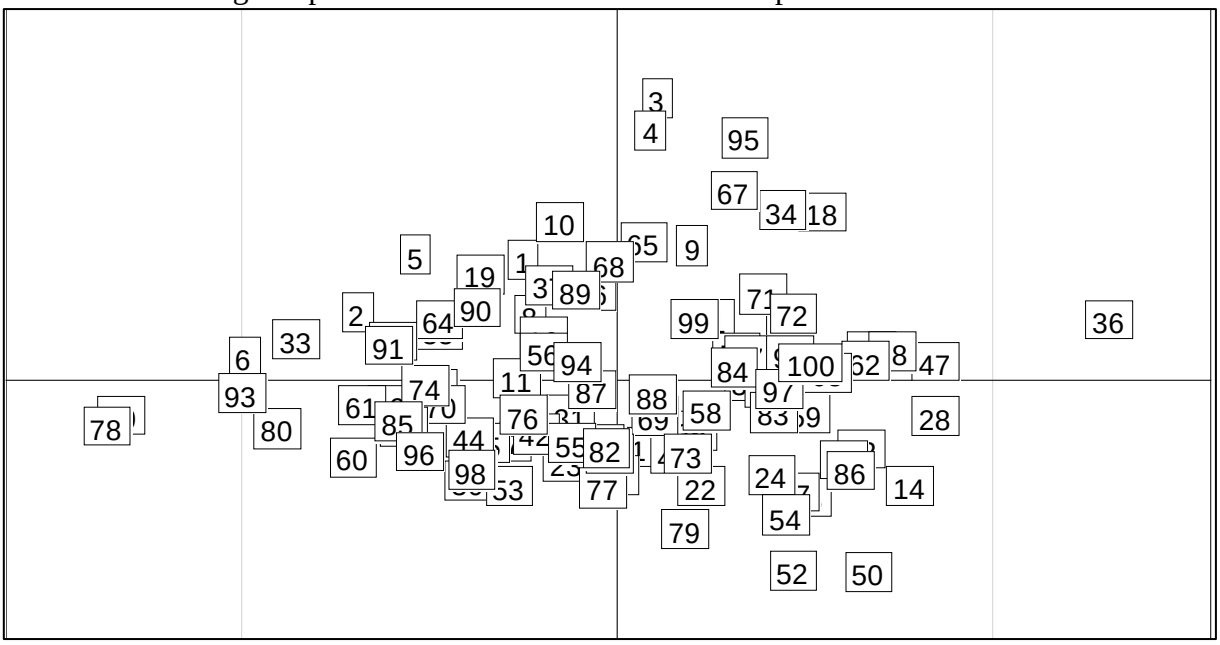
Annexe 3 : ACP sur les spectres

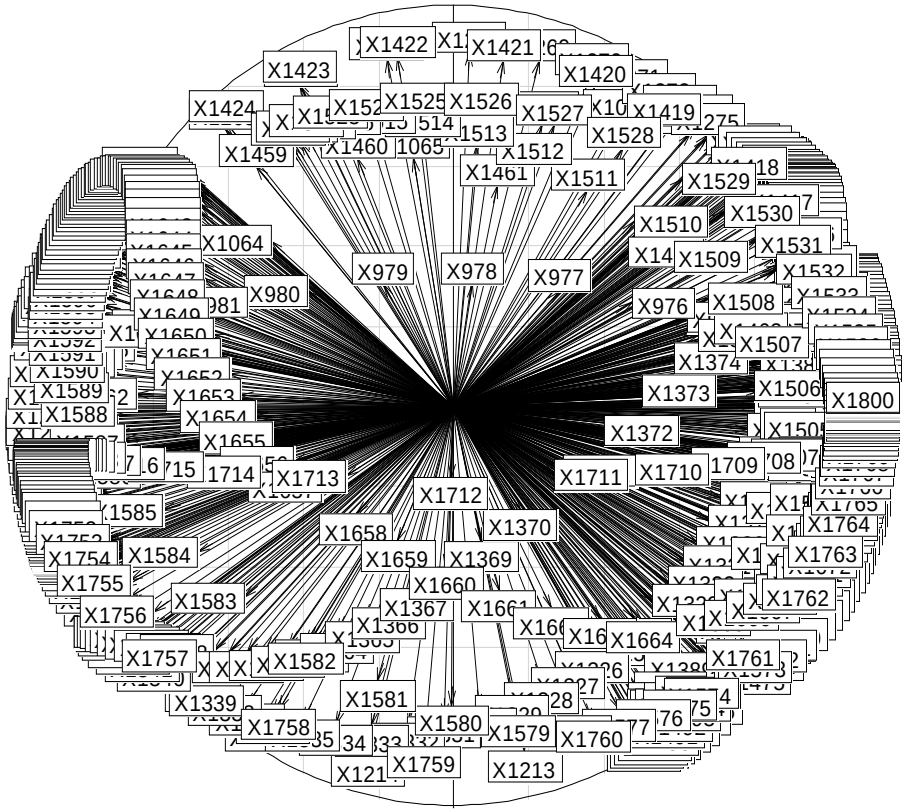
Nuage de points et cercle de corrélation ACP spectre dérivé 1 fois :



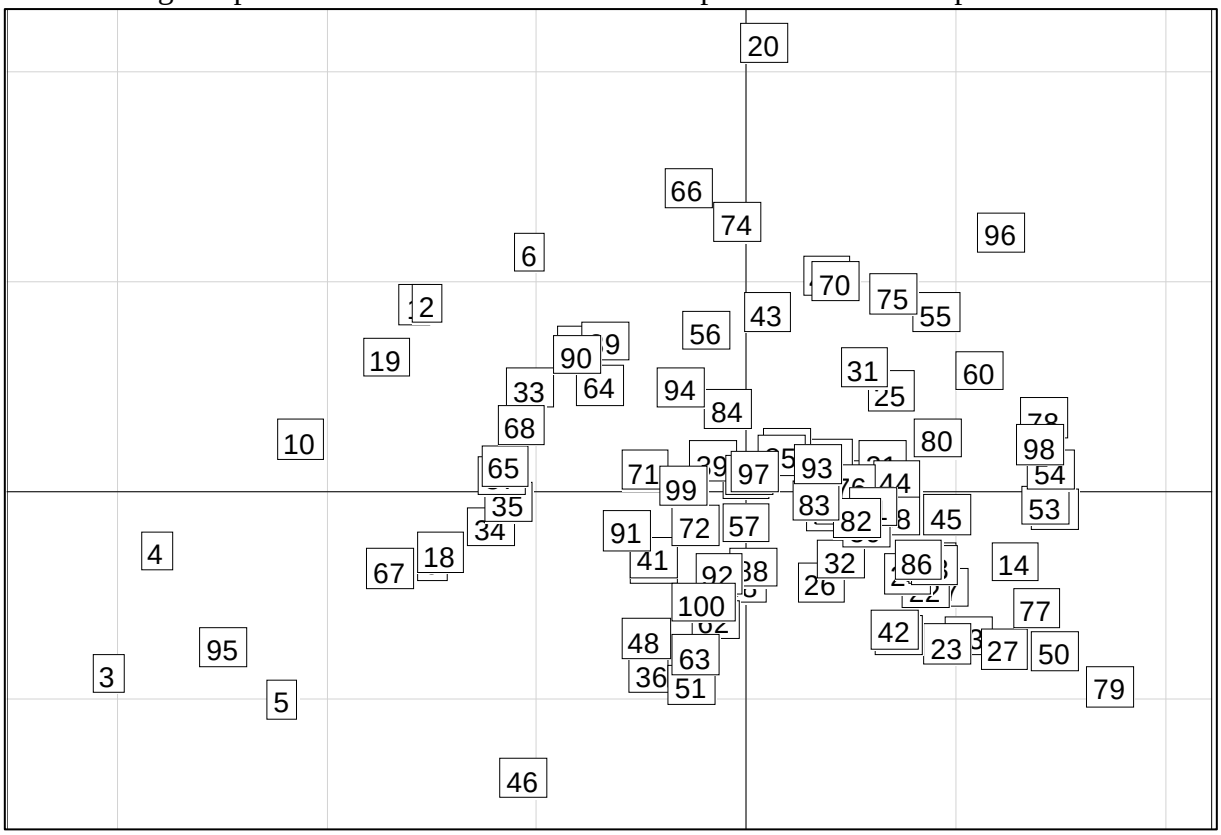


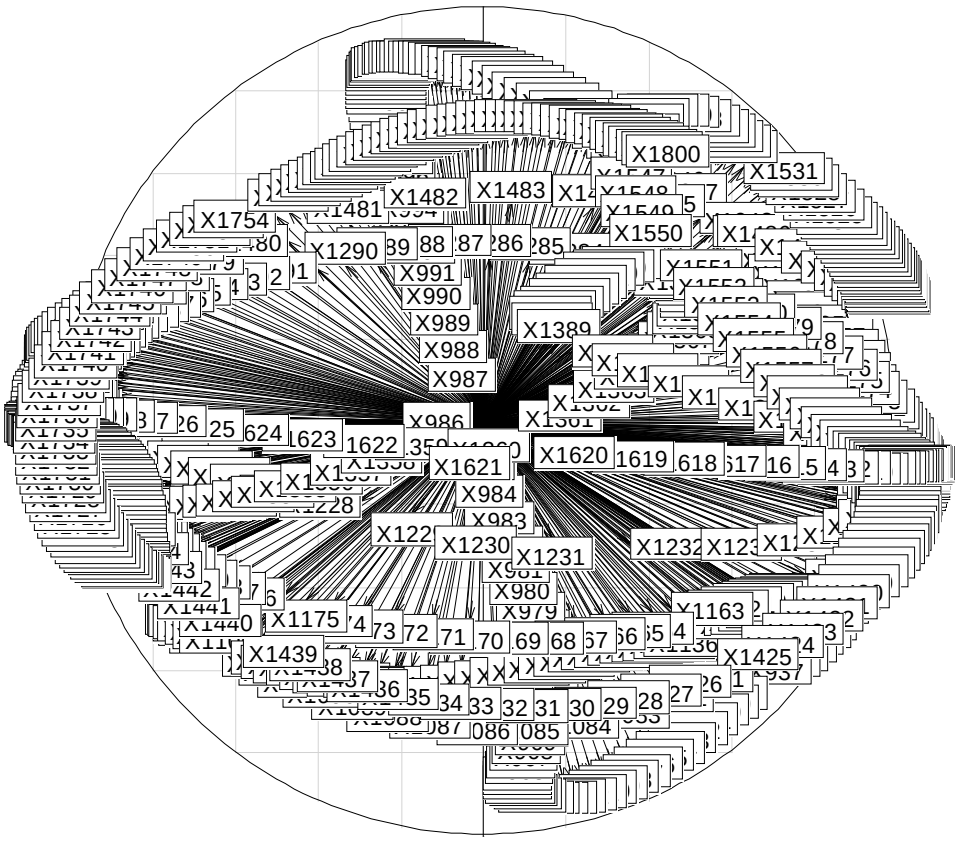
Nuage de points et cercle de corrélation ACP spectre dérivé 2 fois :



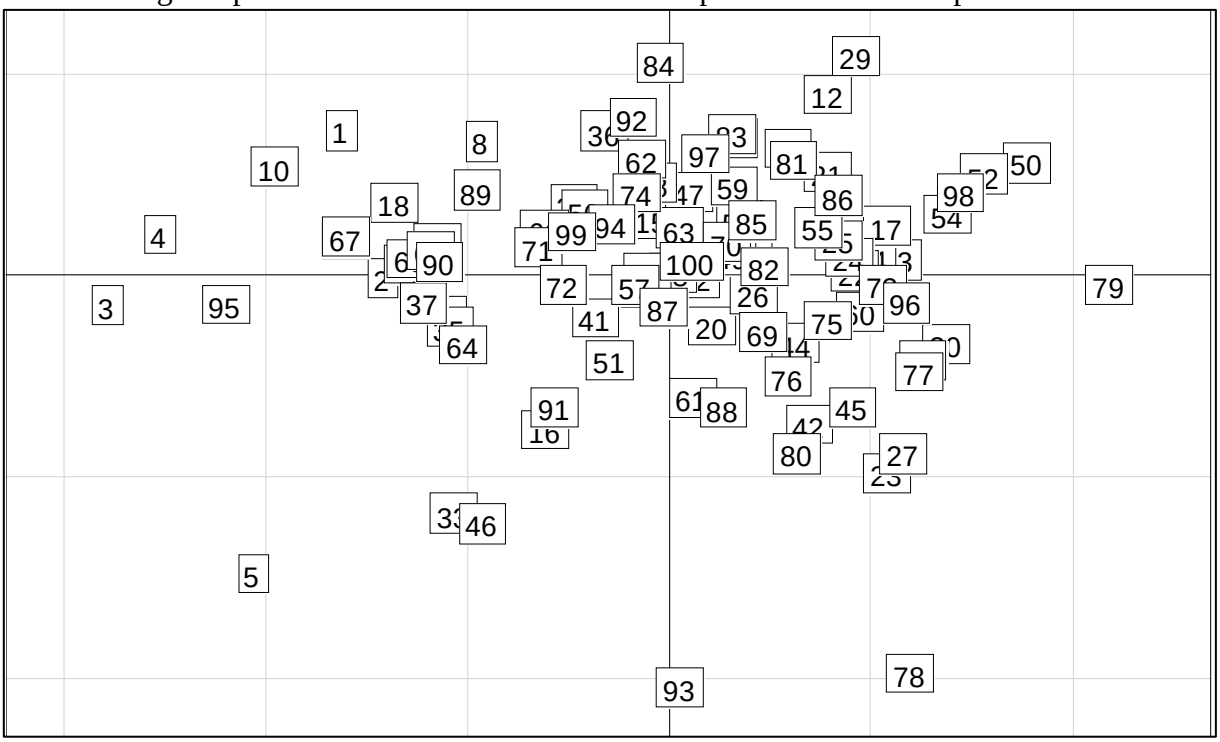


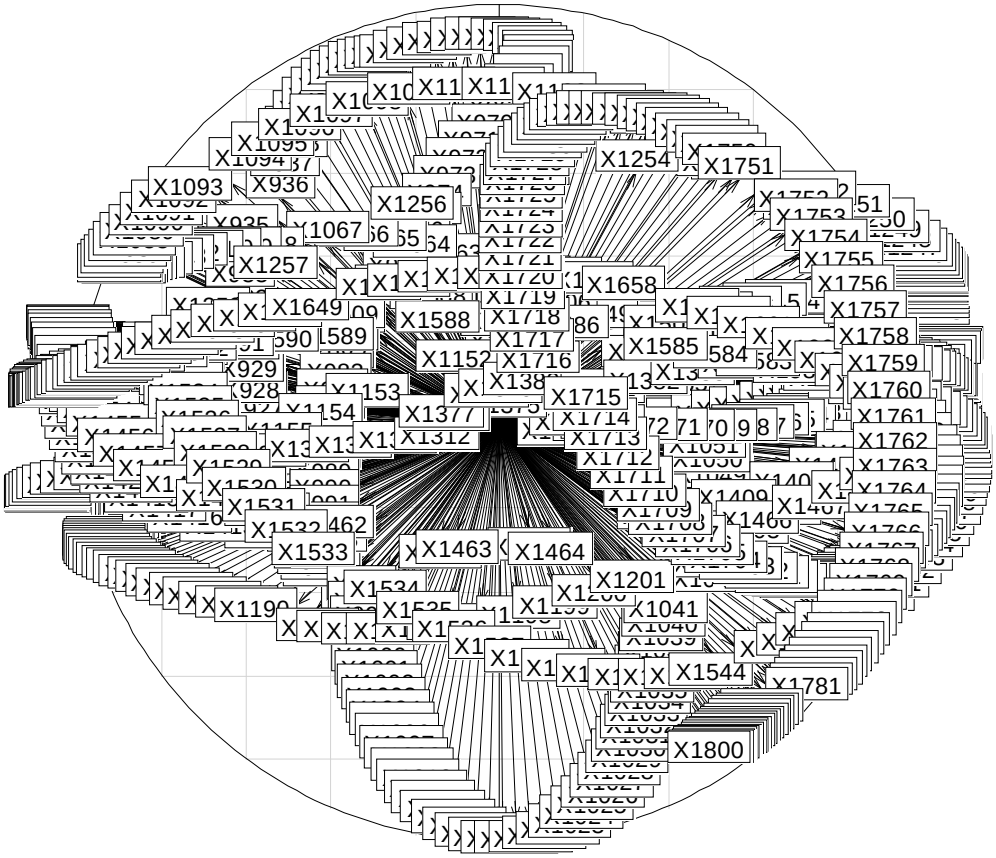
Nuage de points et cercle de corrélation ACP spectre dérivé 1 fois puis normé :



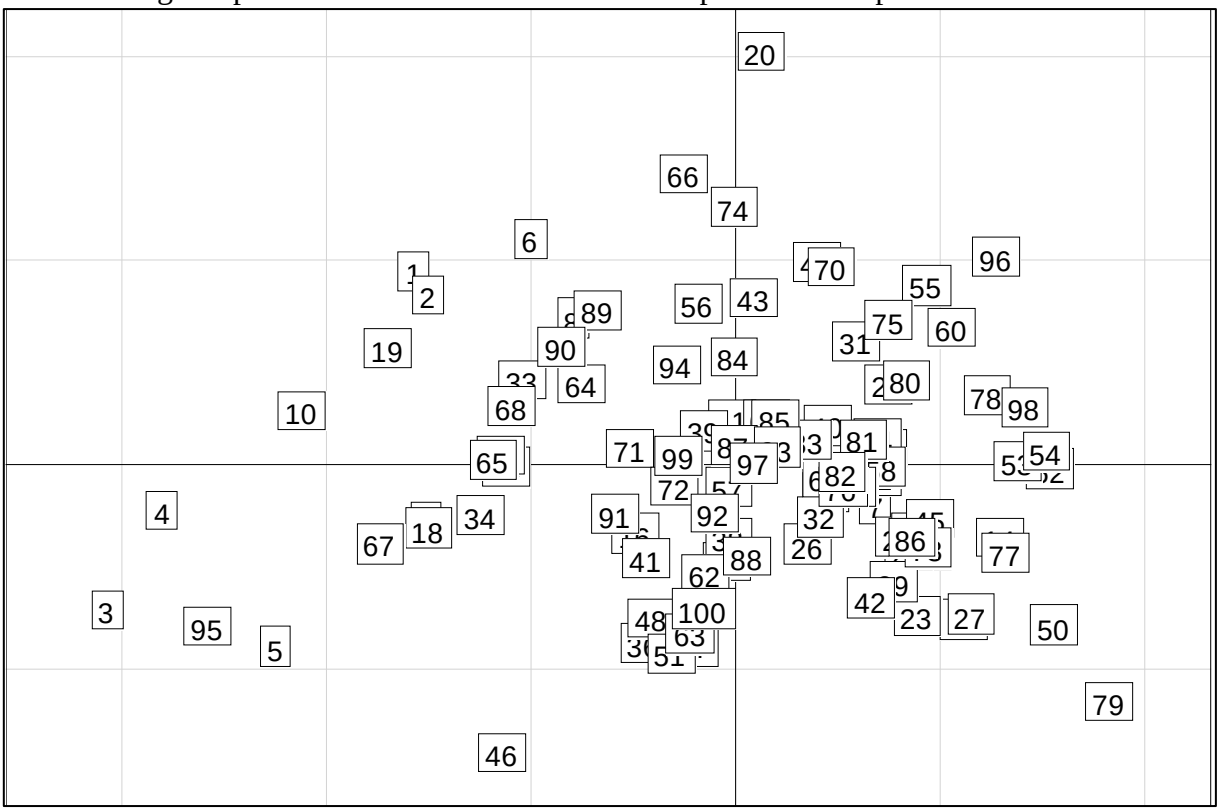


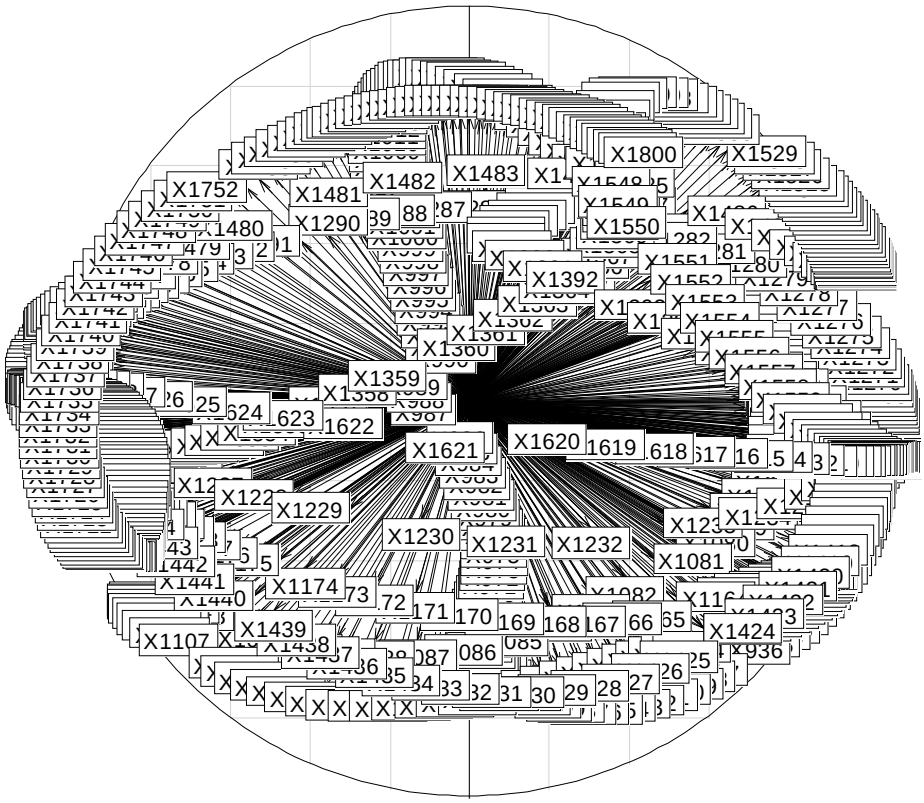
Nuage de points et cercle de corrélation ACP spectre dérivé 2 fois puis normé :



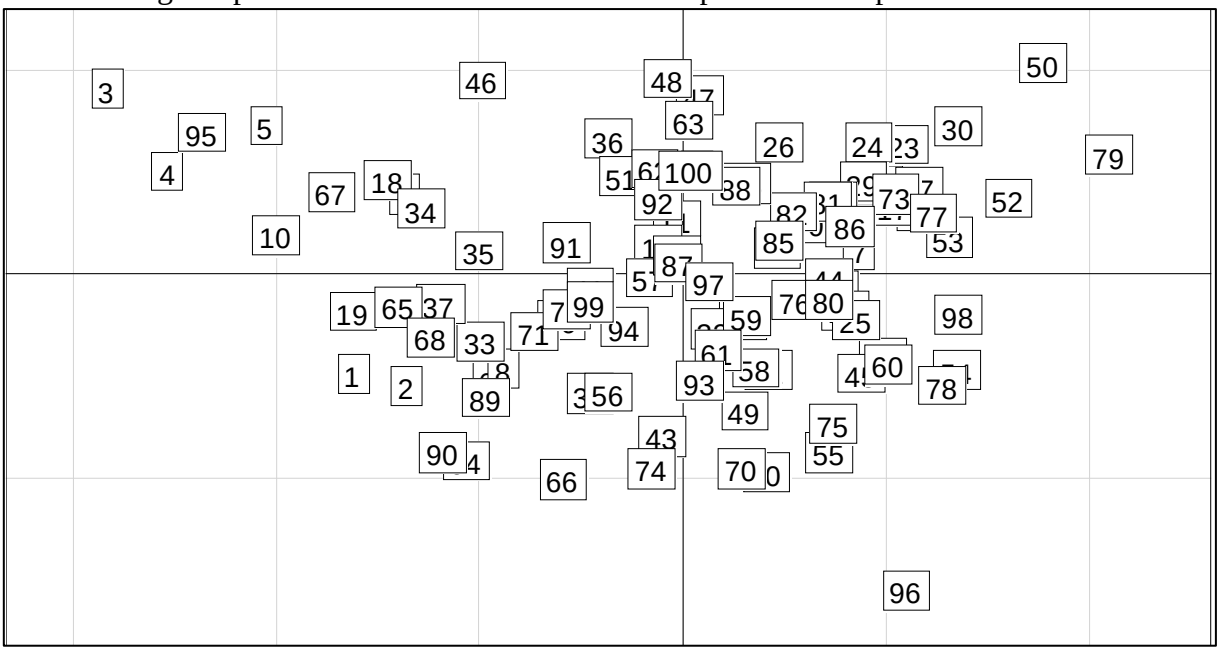


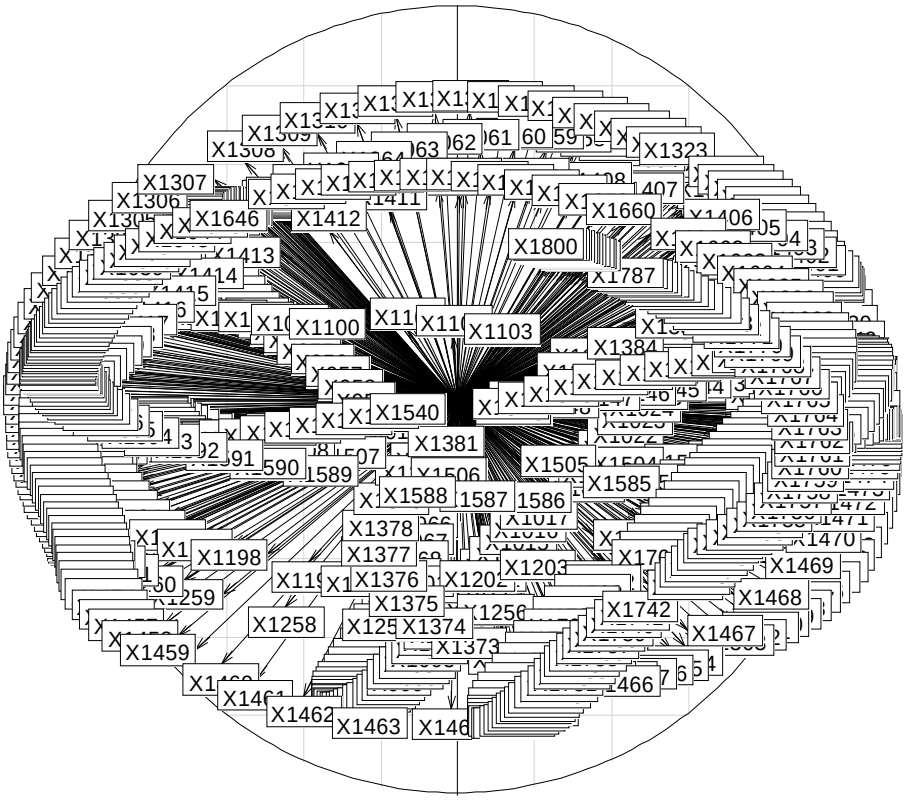
Nuage de points et cercle de corrélation ACP spectre normé puis dérivé 1 fois :



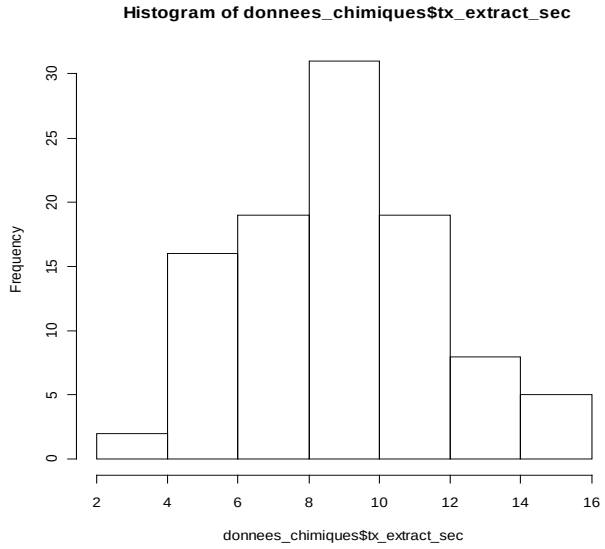


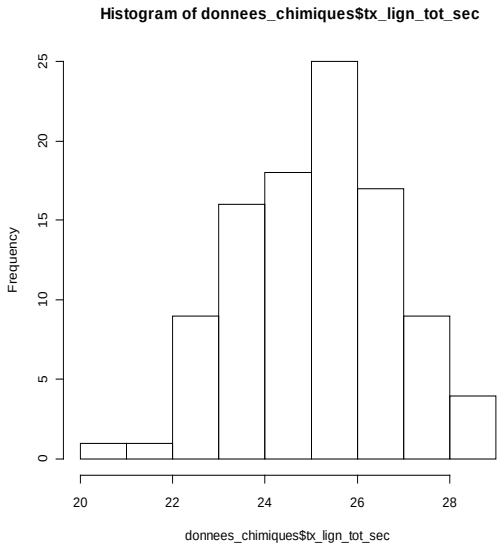
Nuage de points et cercle de corrélation ACP spectre normé puis dérivé 2 fois :

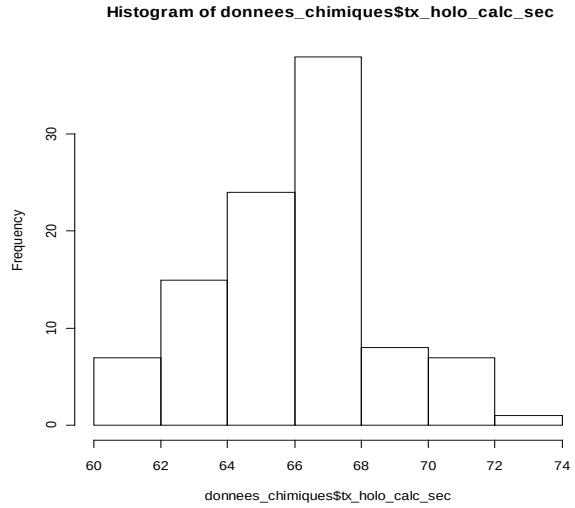


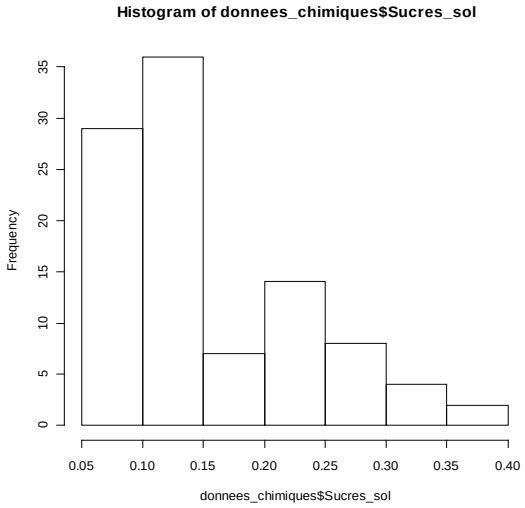


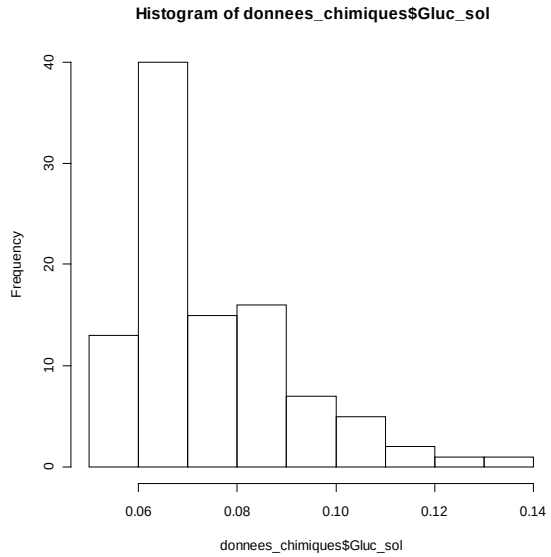
Annexe 4 : histogrammes des différents dosages chimiques

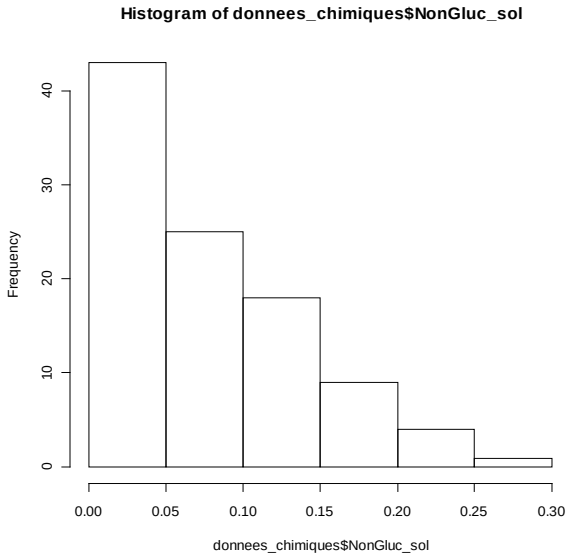


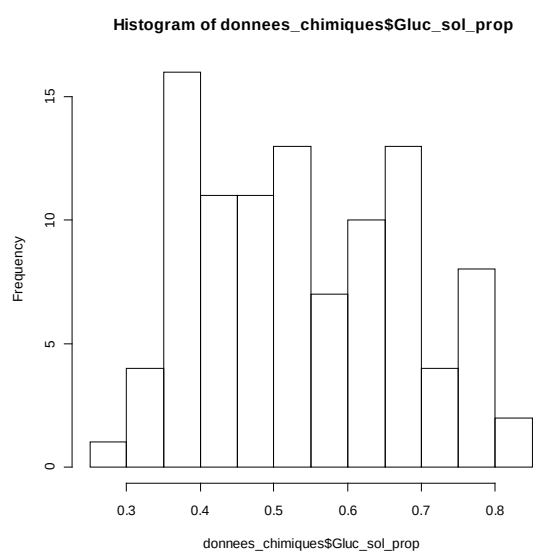


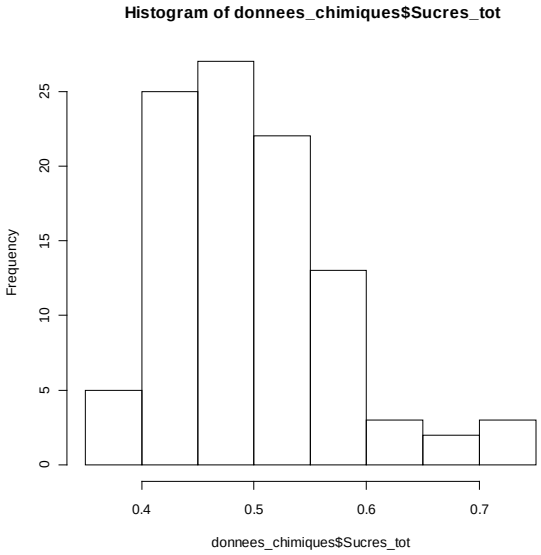


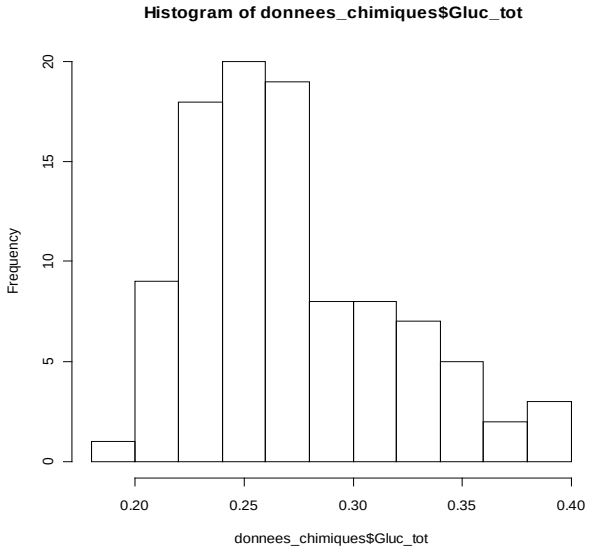


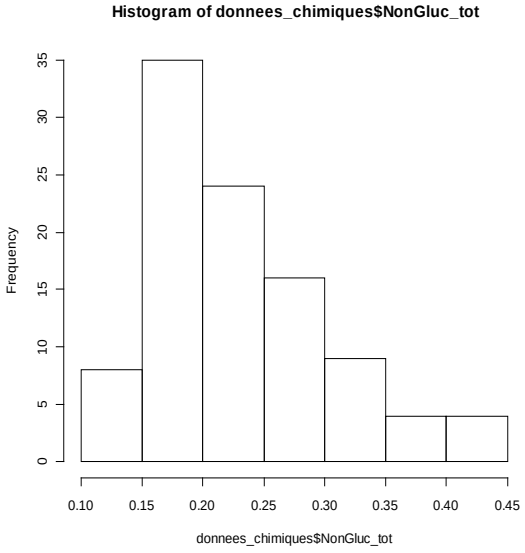


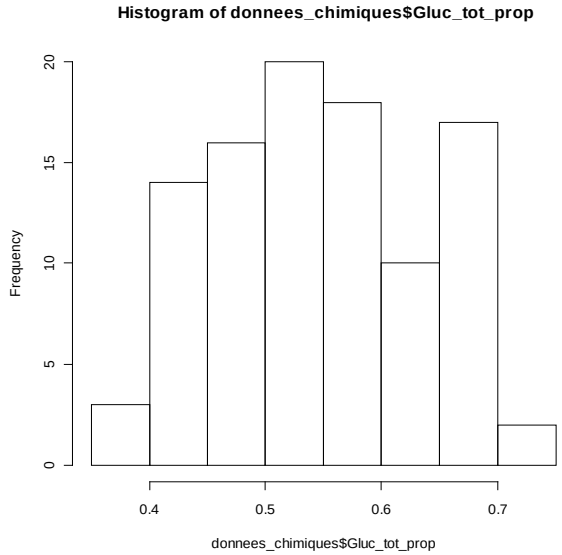


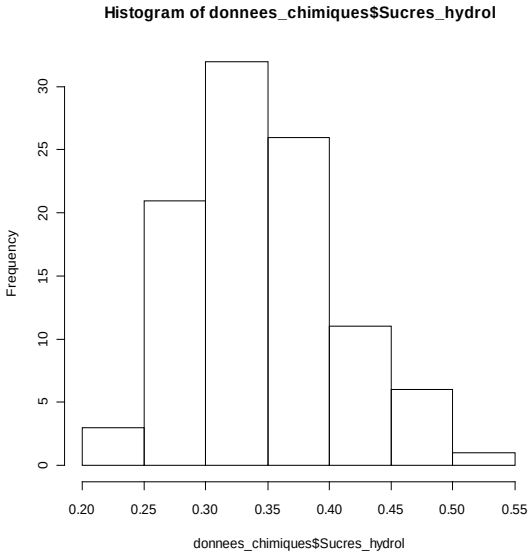

Histogram of donnees_chimiques$Gluc_sol_prop
Frequency
0
5
10
15
0.3
0.4
0.5
0.6
0.7
0.8
donnees_chimiques$Gluc_sol_prop

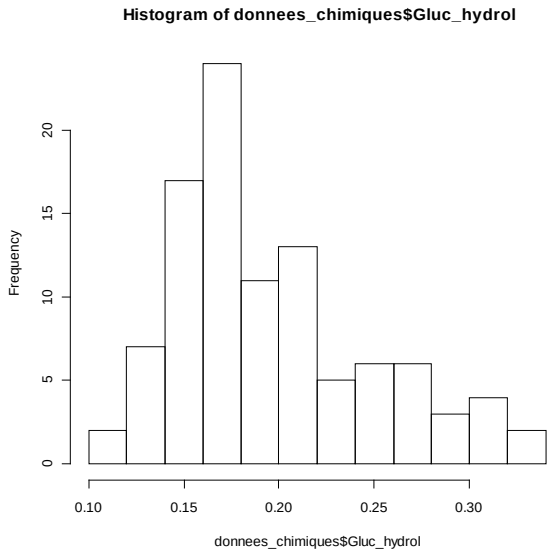

Histogram of donnees_chimiques$Sucres_tot
Frequency
0
5
10
15
20
25
0.4
0.5
0.6
0.7
donnees_chimiques$Sucres_tot

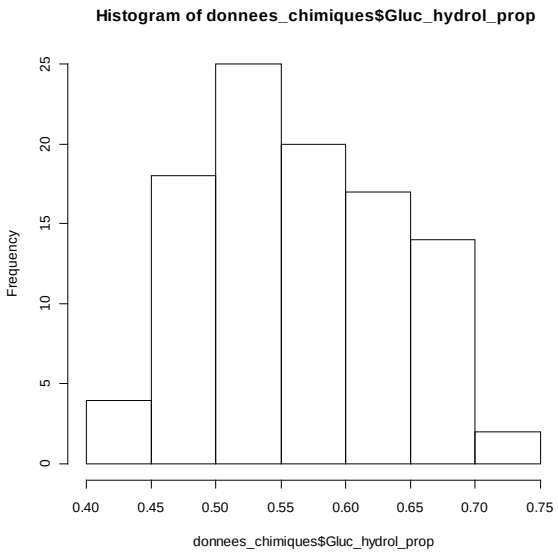

Histogram of donnees_chimiques$Gluc_tot
Frequency
0
5
10
15
20
0.20
0.25
0.30
0.35
0.40
donnees_chimiques$Gluc_tot


Histogram of donnees_chimiques$NonGluc_tot
Frequency
0
5
10
15
20
25
30
35
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
donnees_chimiques$NonGluc_tot


Histogram of donnees_chimiques$Gluc_tot_prop
Frequency
0
5
10
15
20
0.4
0.5
0.6
0.7
donnees_chimiques$Gluc_tot_prop


Histogram of donnees_chimiques$Sucres_hydrol
Frequency
0
5
10
15
20
25
30
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
donnees_chimiques$Sucres_hydrol

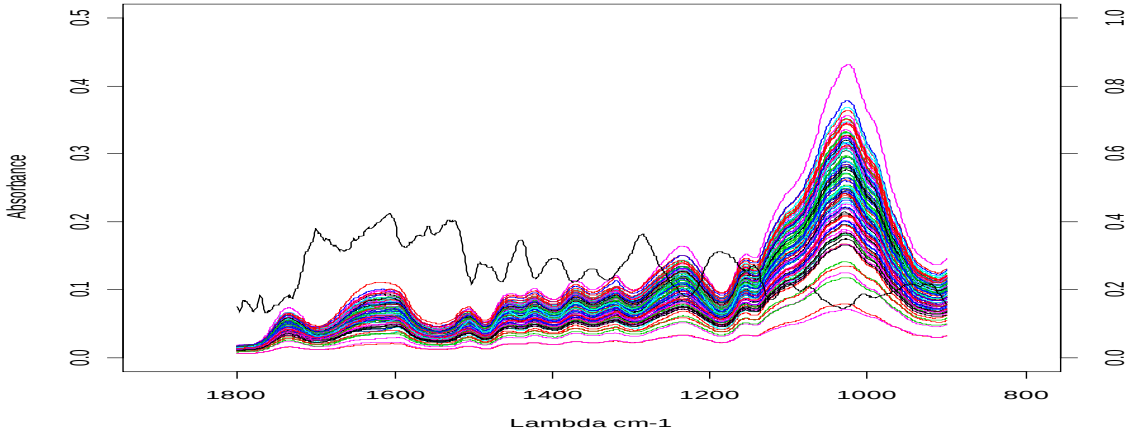

Histogram of donnees_chimiques$Gluc_hydrol
Frequency
0
5
10
15
20
0.10
0.15
0.20
0.25
0.30
donnees_chimiques$Gluc_hydrol


Histogram of donnees_chimiques$Gluc_hydrol_prop
Frequency
0
5
10
15
20
25
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
donnees_chimiques$Gluc_hydrol_prop

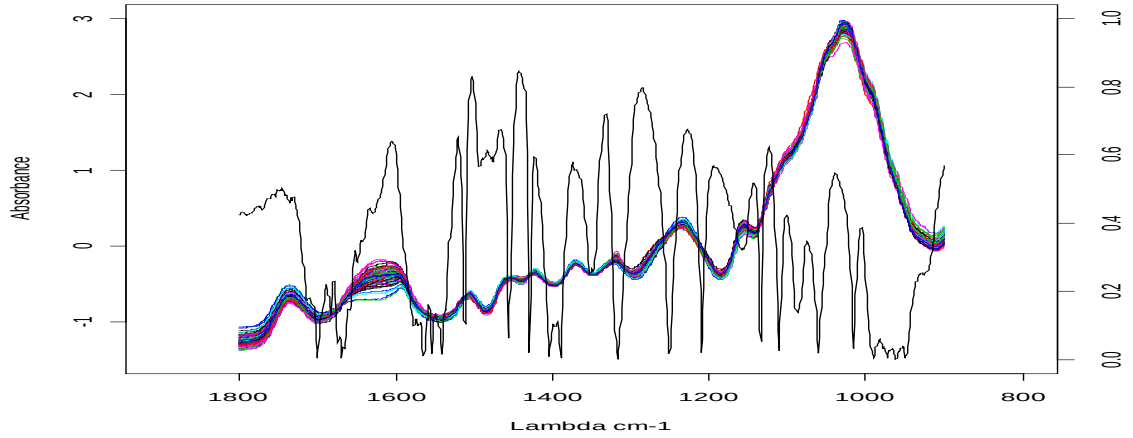
Annexe 5 : Exemples de graphes représentant les corrélations de Spearman entre quelques variables chimiques et les données spectrales (brutes, normées, dérivées 1 fois et dérivées 2 fois)

Taux d'extractibles :

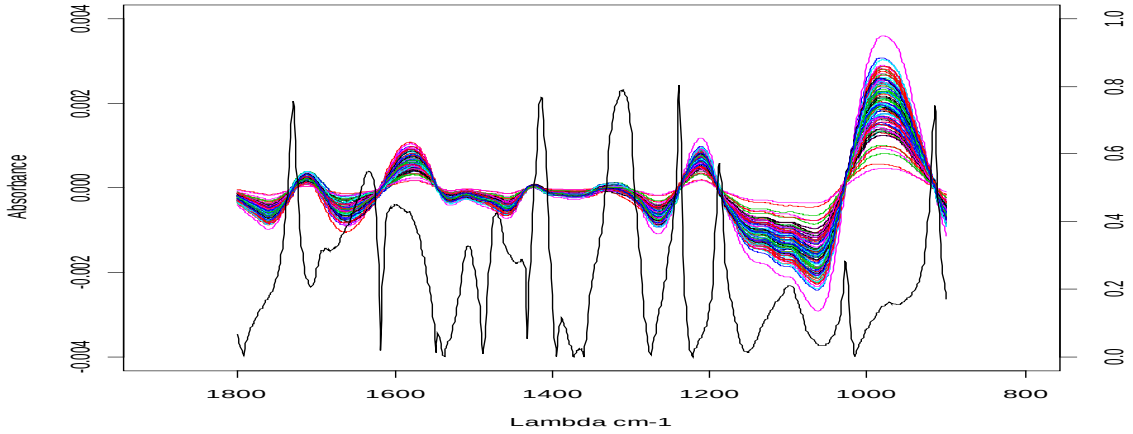
tx_extract_sec et données spectrales brutes

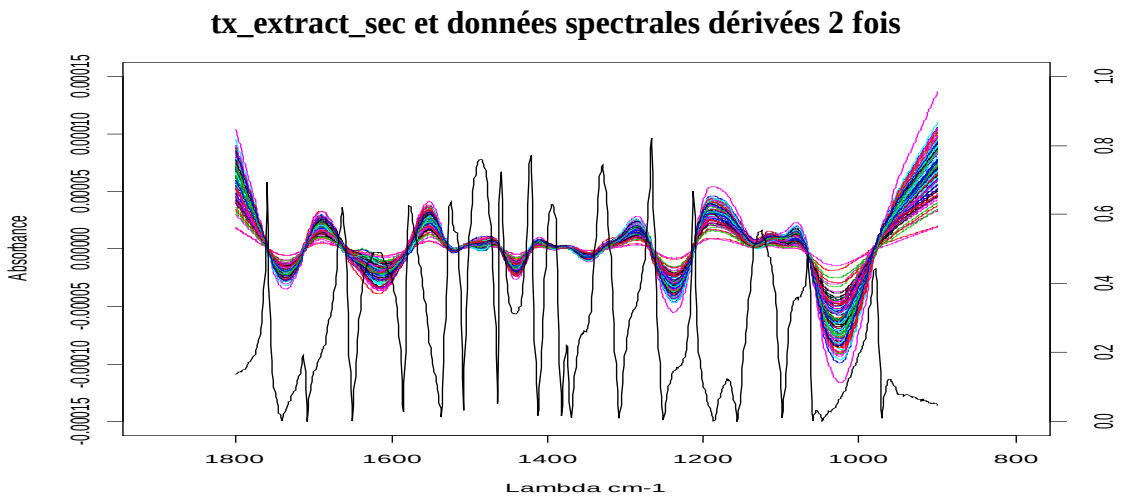


tx_extract_sec et données spectrales normalisées

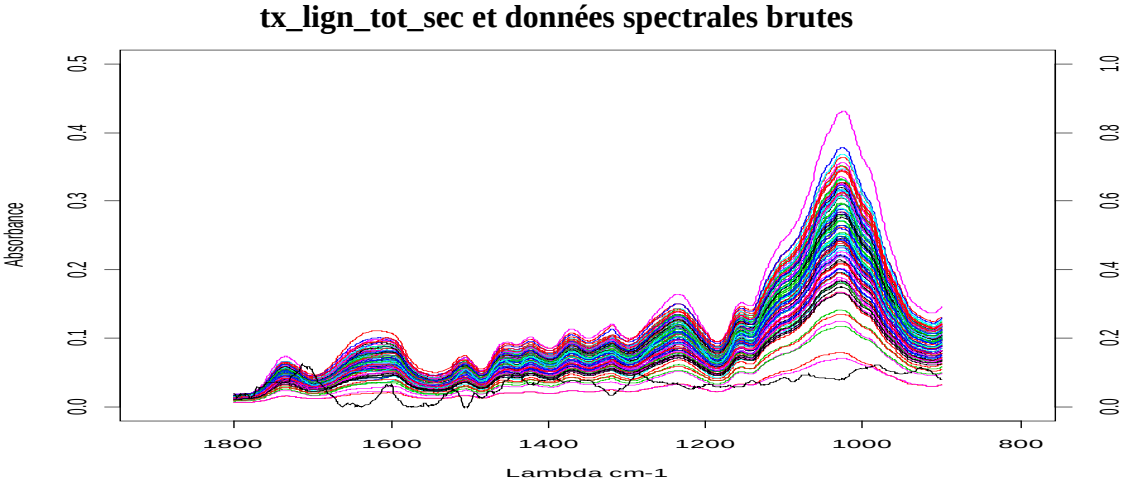


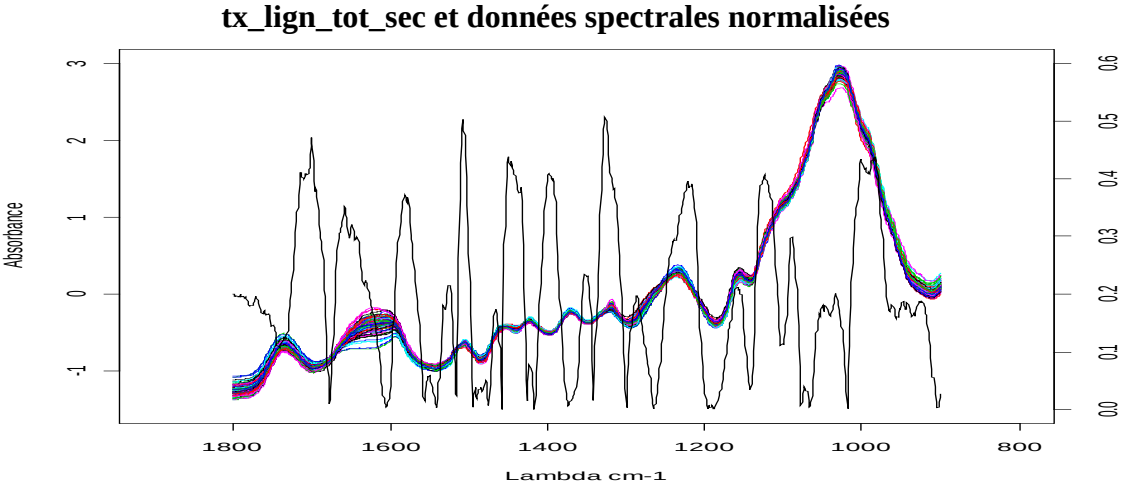
tx_extract_sec et données spectrales dérivées 1 fois



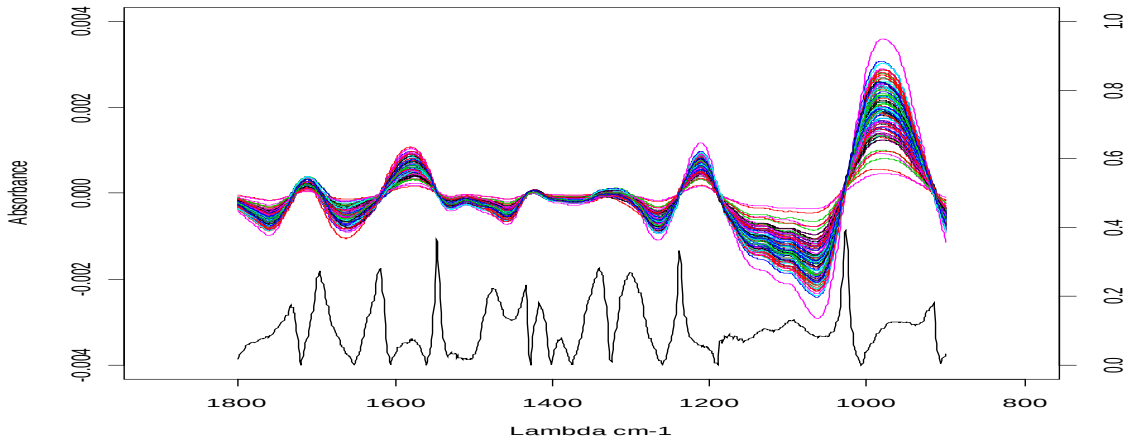


Taux de lignines :

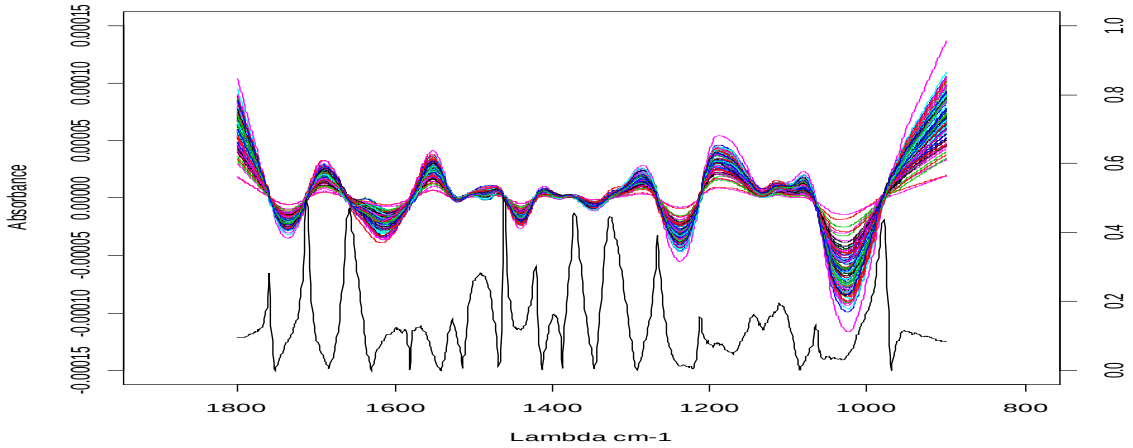




tx_lign_tot_sec et données spectrales dérivées 1 fois

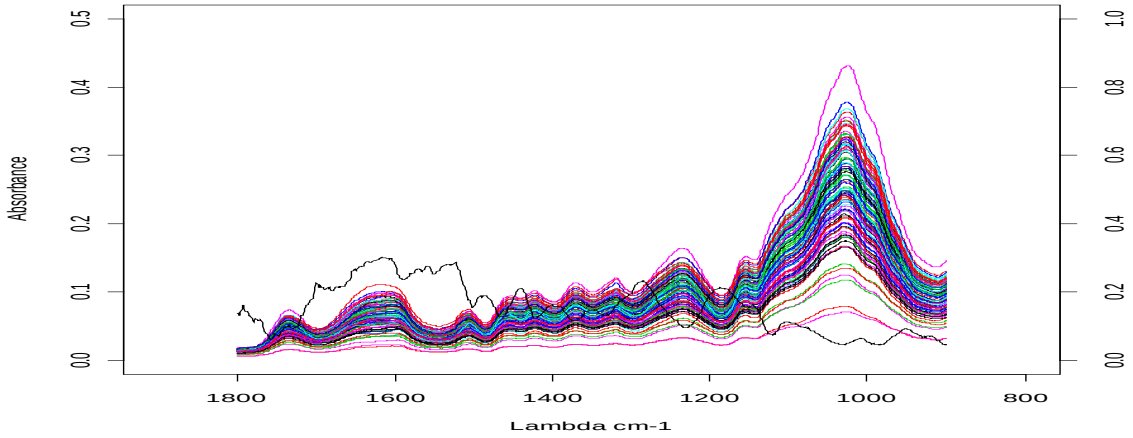


tx_lign_tot_sec et données spectrales dérivées 2 fois

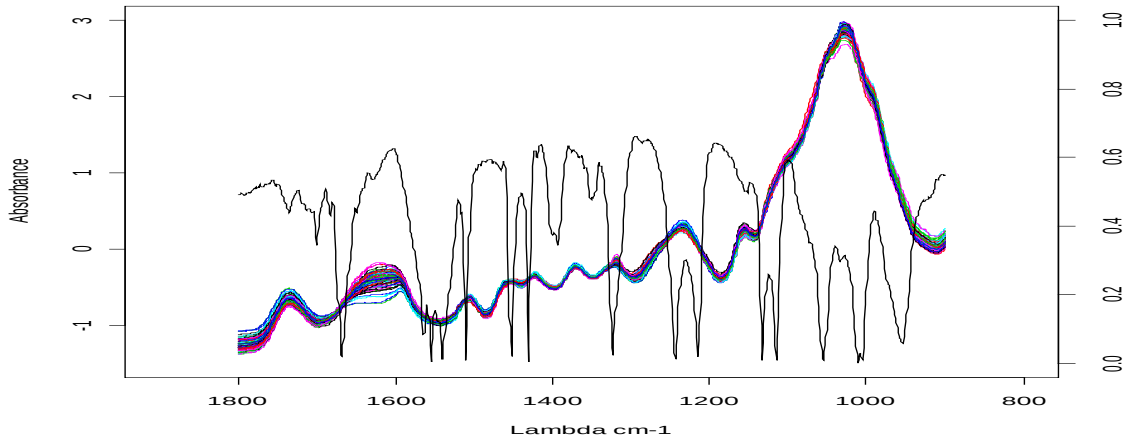


Glucose total :

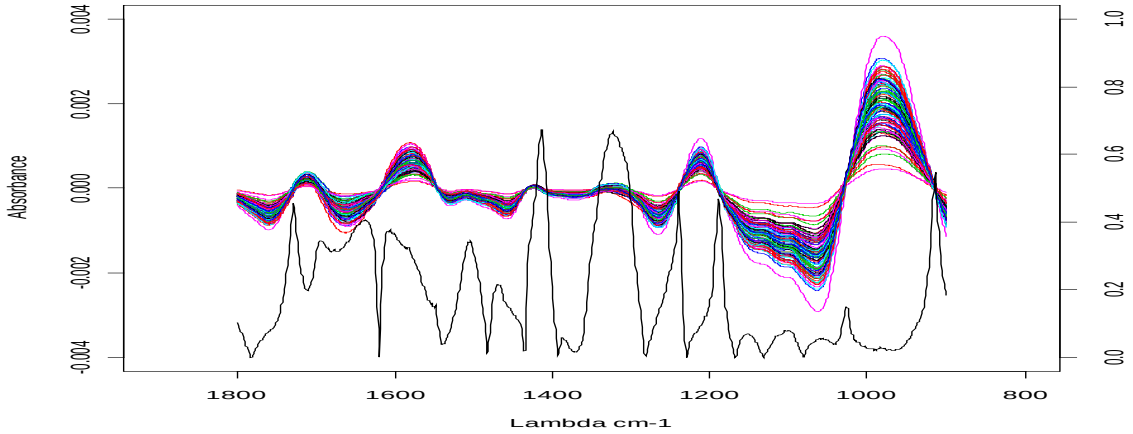
Gluc_tot et données spectrales brutes



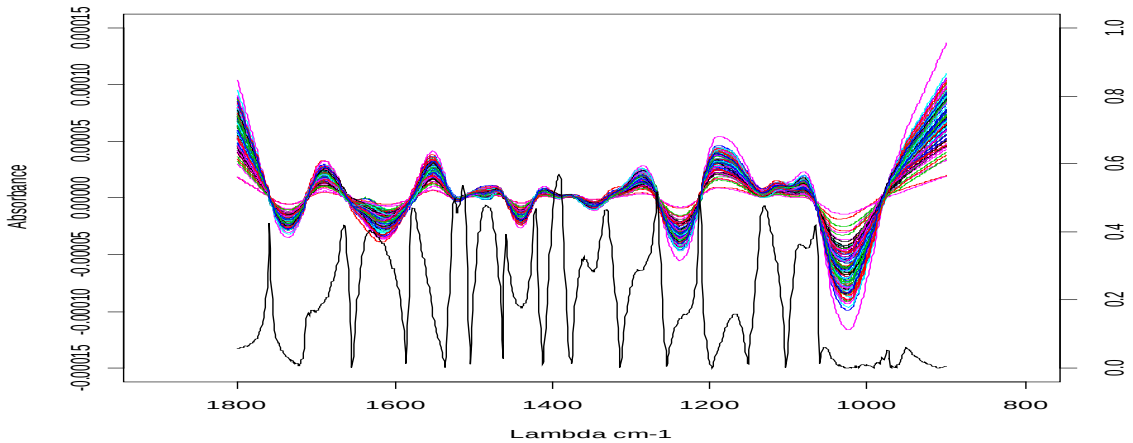
Gluc_tot et données spectrales normalisées



Gluc_tot et données spectrales dérivées 1 fois:

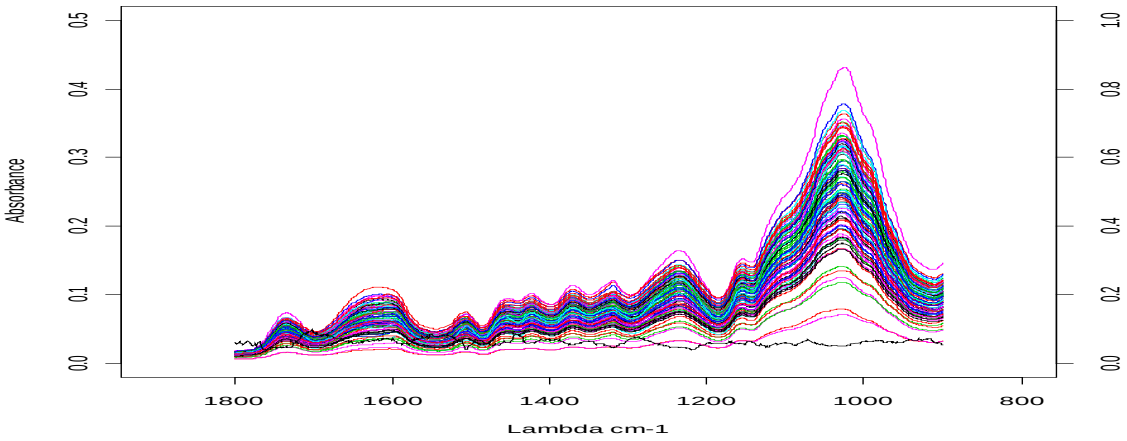


Gluc_tot et données spectrales dérivées 2 fois:

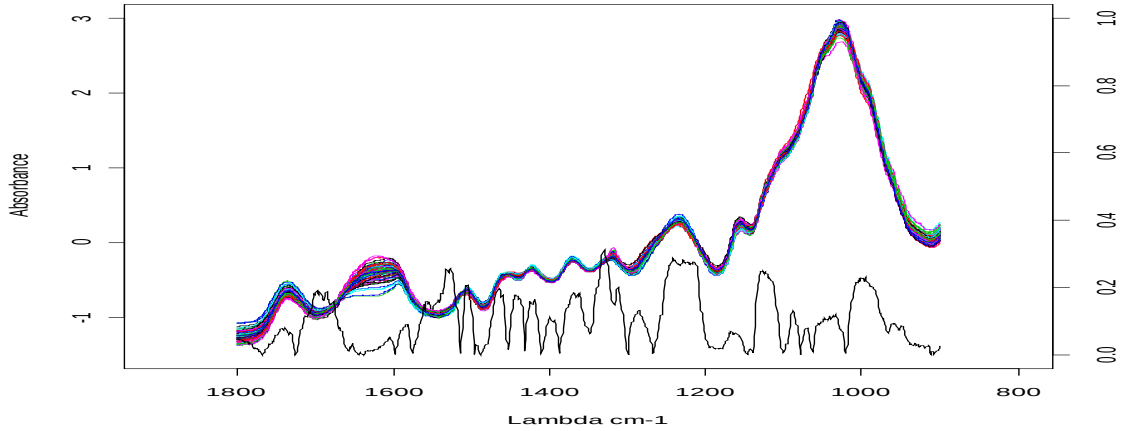


Non glucose hydrolysé :

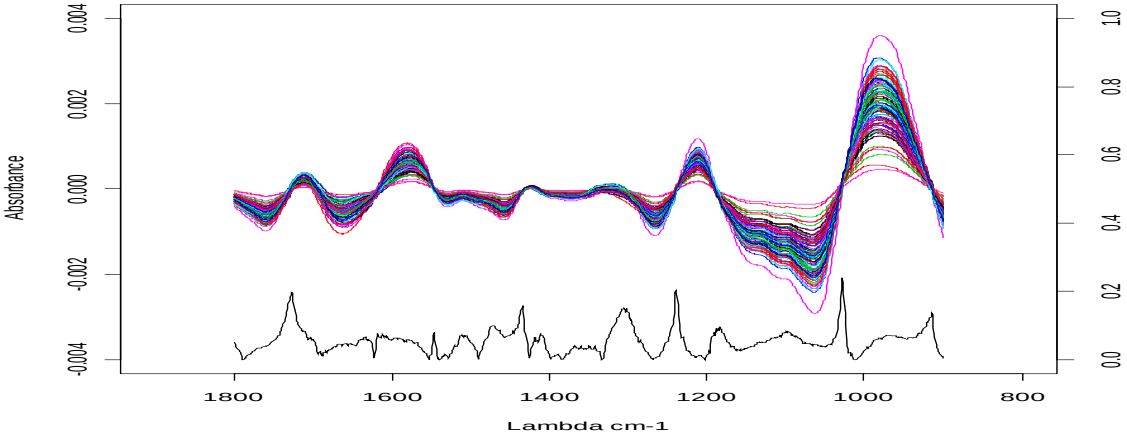
NonGluc_hydrol et données spectrales brutes:



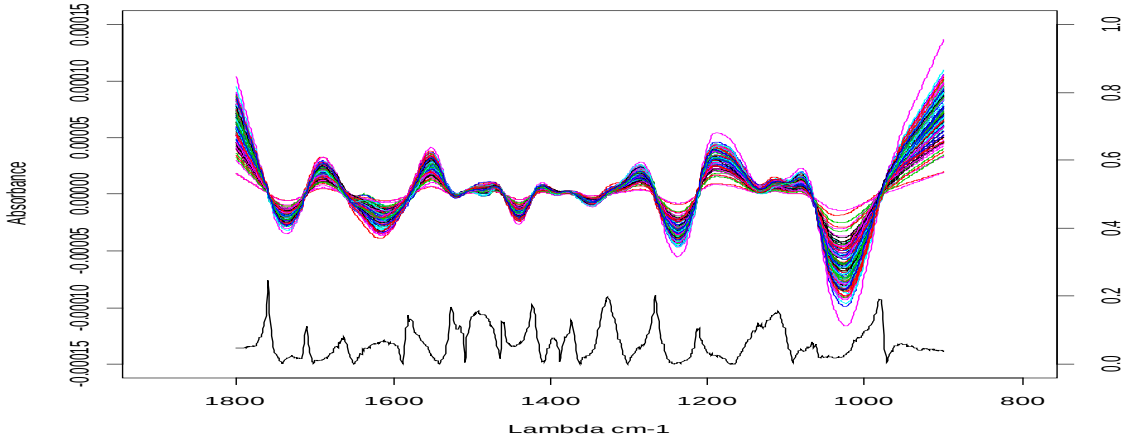
NonGluc_hydrol et données spectrales normalisées:



NonGluc_hydrol et données spectrales dérivées 1 fois:



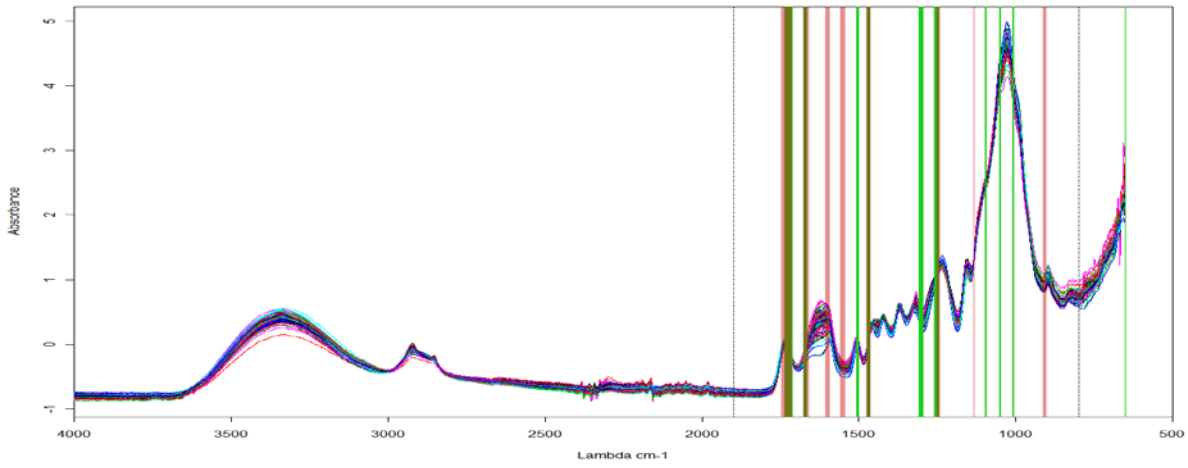
NonGluc_hydrol et données spectrales dérivées 2 fois:



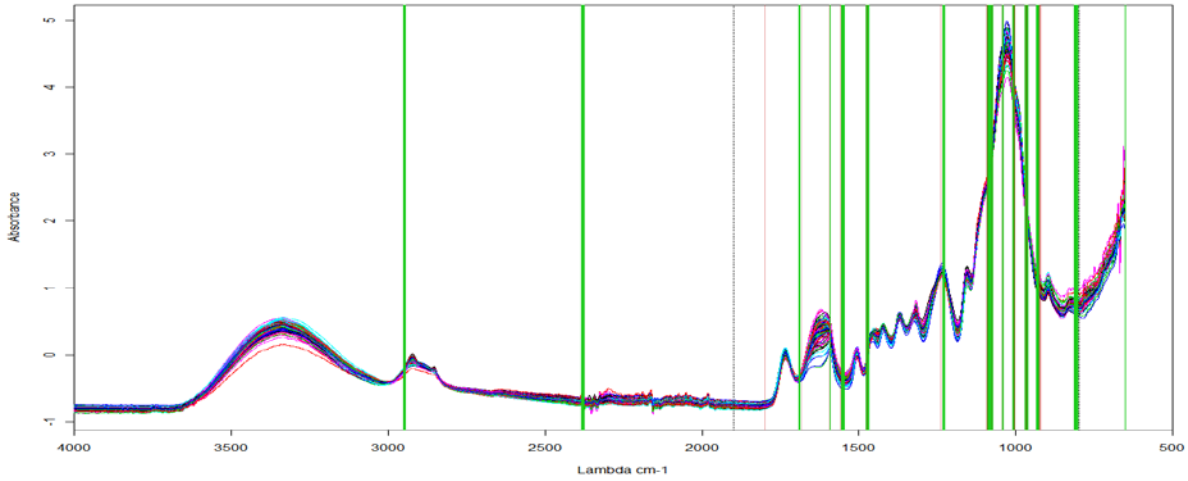
Annexe 6 : Graphes des nombres d'onde sélectionnés.

Légende : en vert nombres d'onde correspondant au spectre entier et en rouge au spectre découpé.

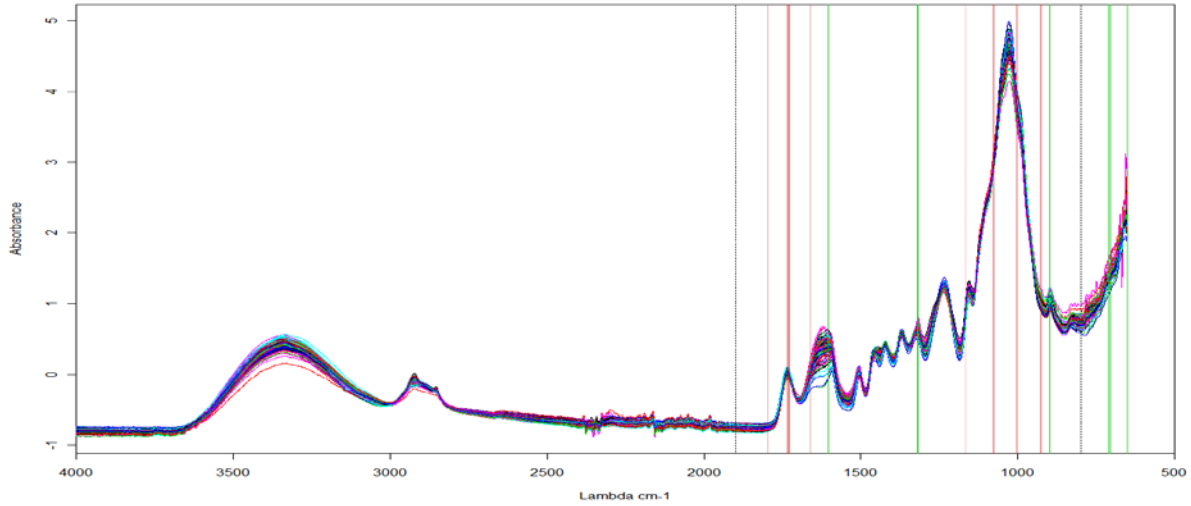
Extract :



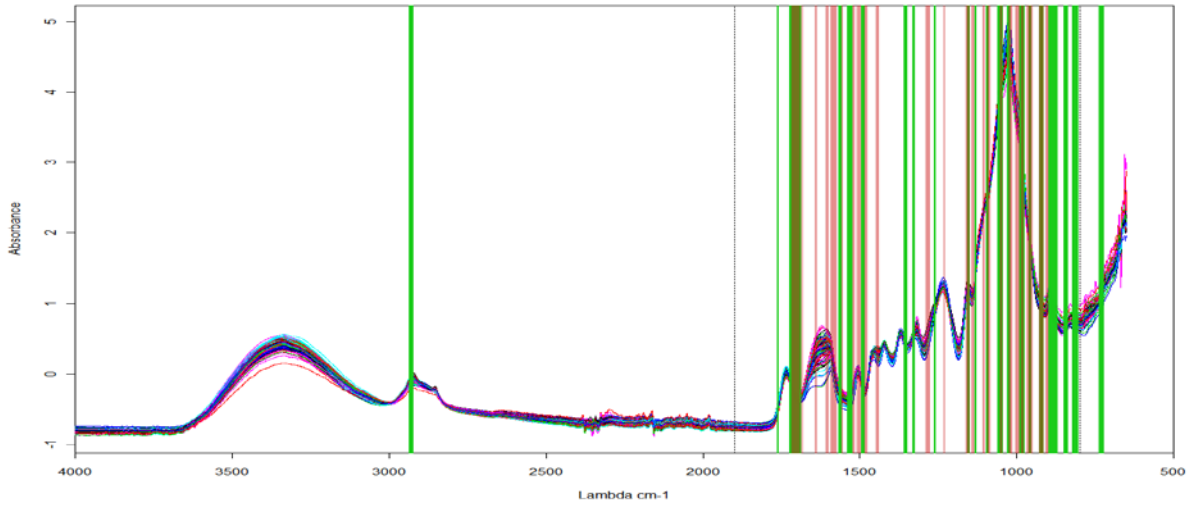
Lignine :



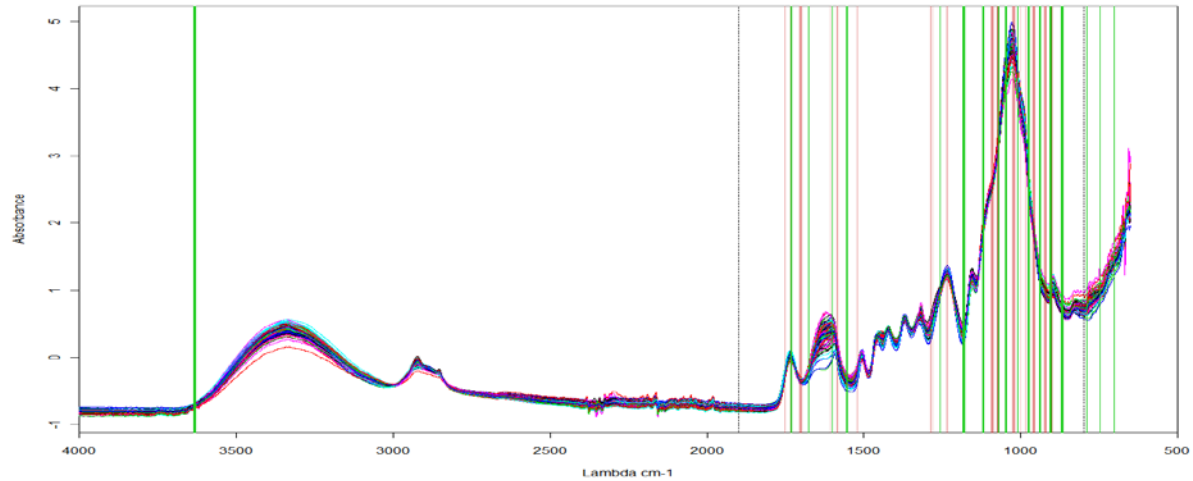
Holo :



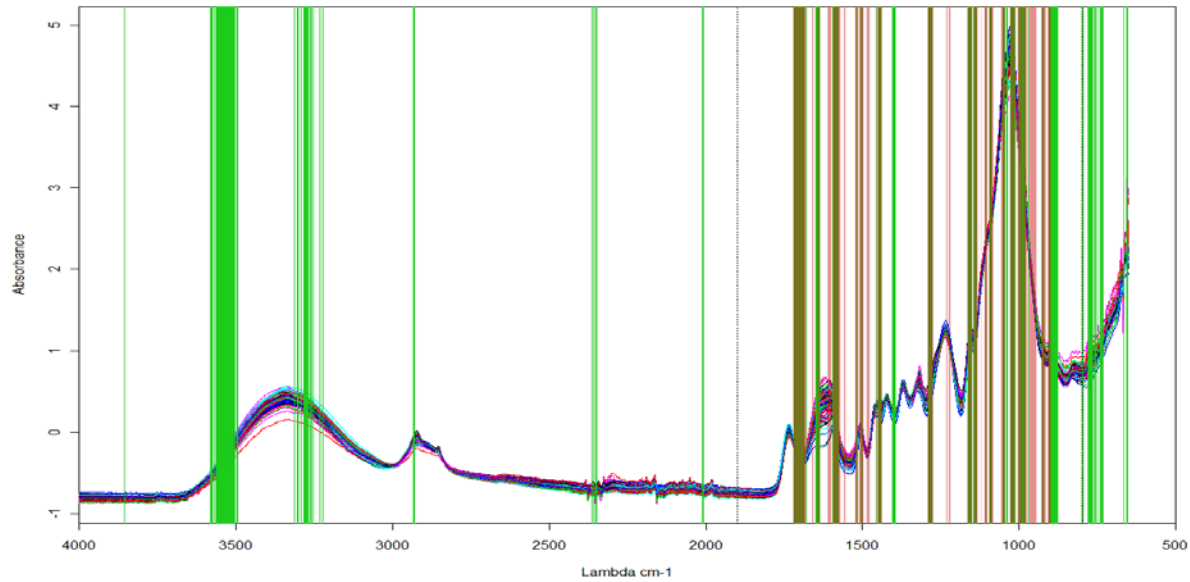
Sucres_sol :



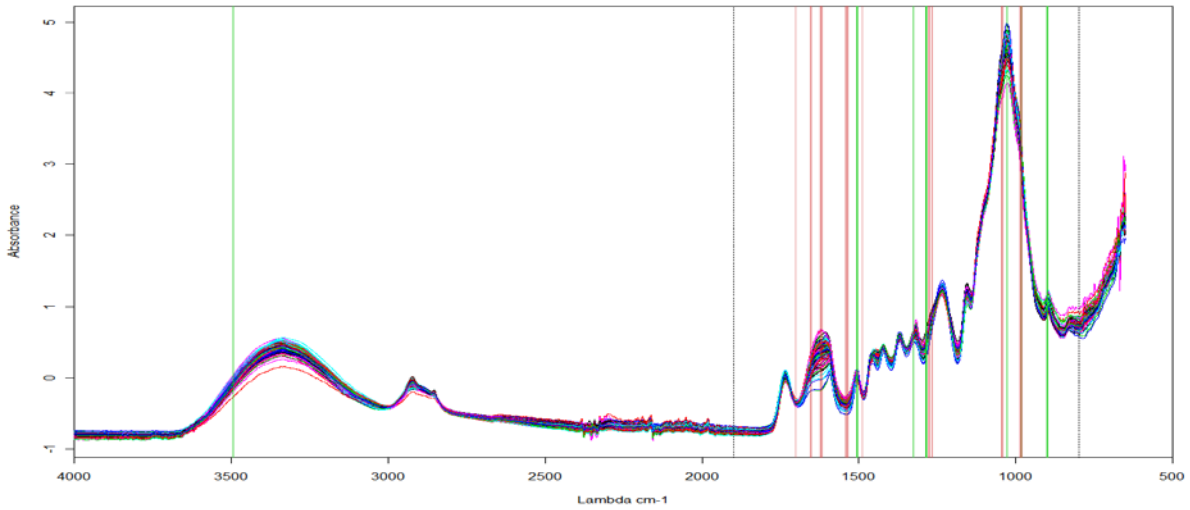
Gluc_sol :



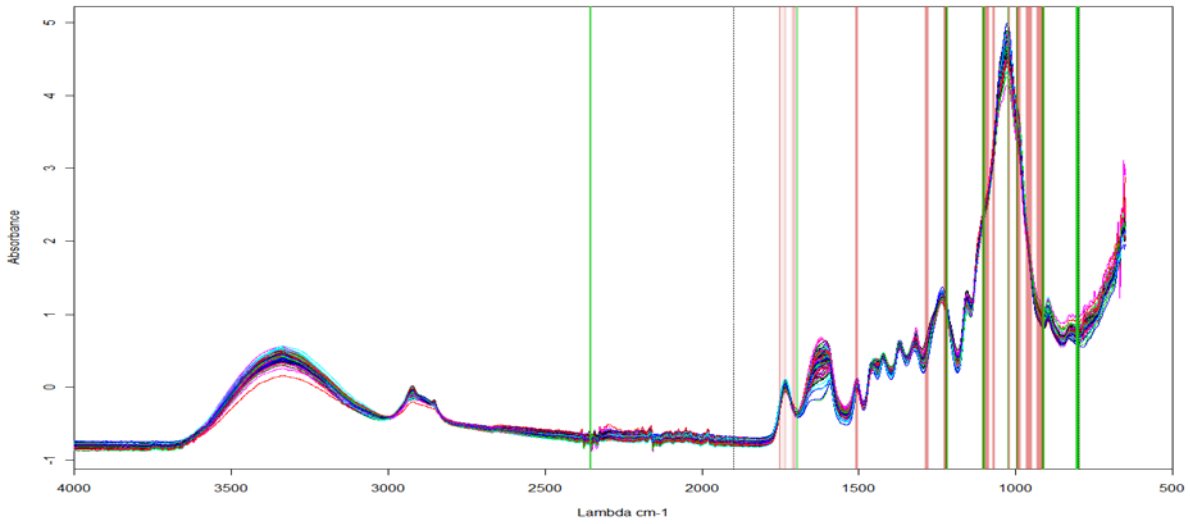
NonGluc_sol :



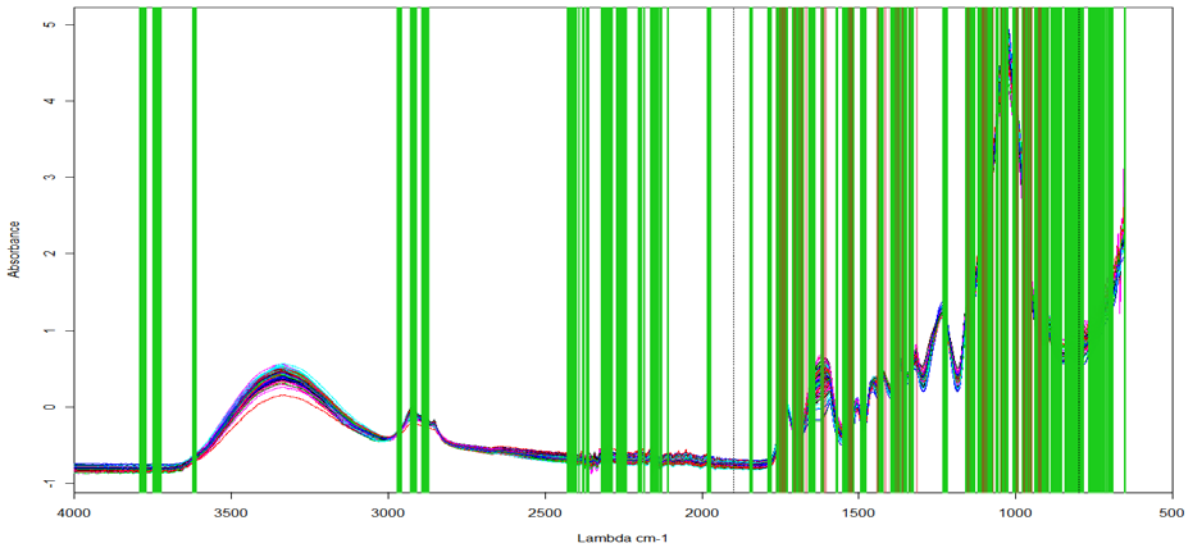
Gluc_sol_prop :



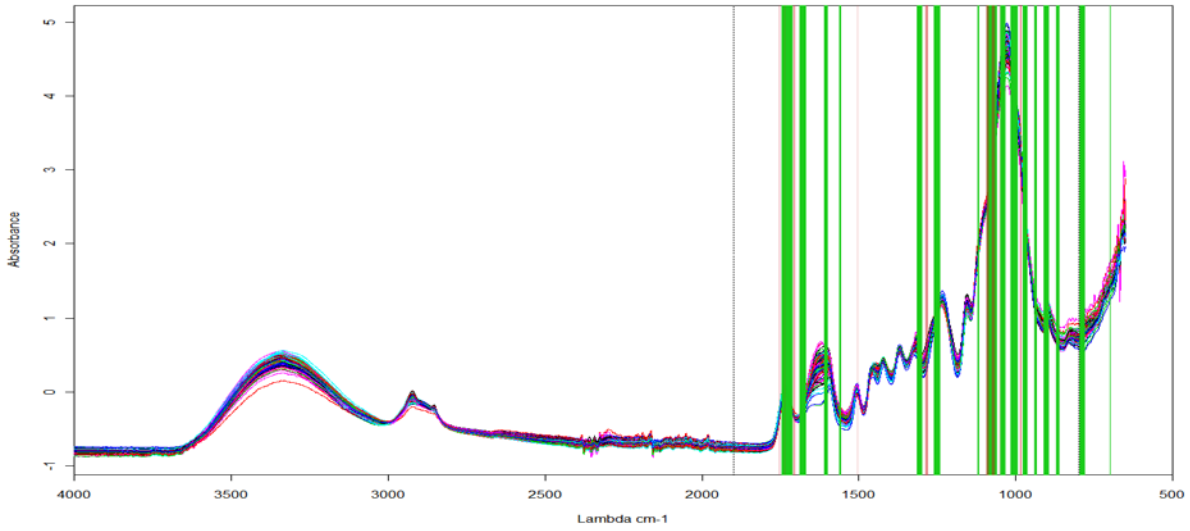
Sucres_tot :



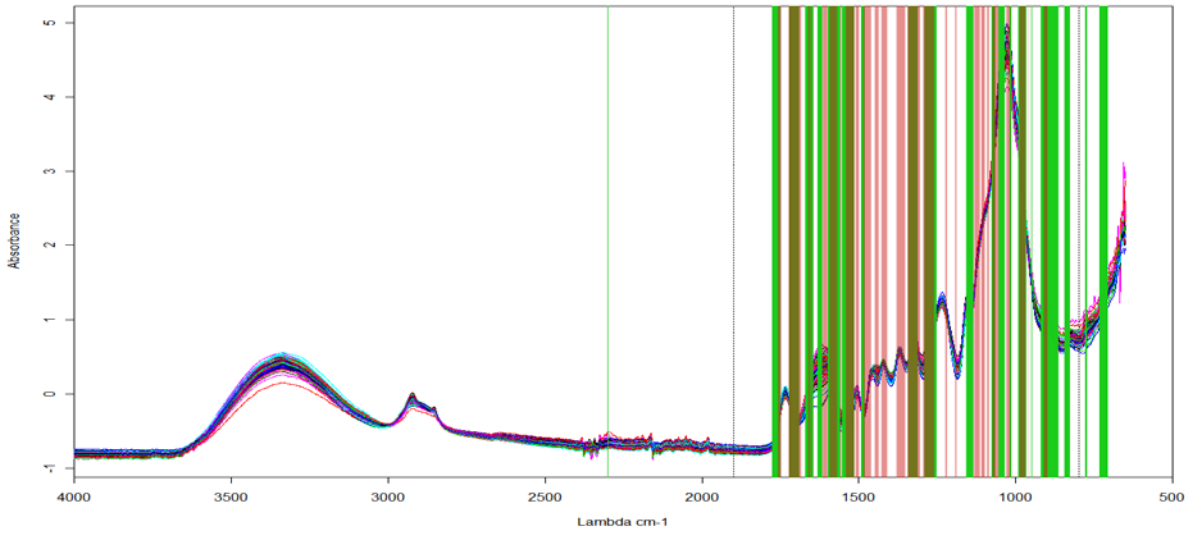
Gluc_tot :



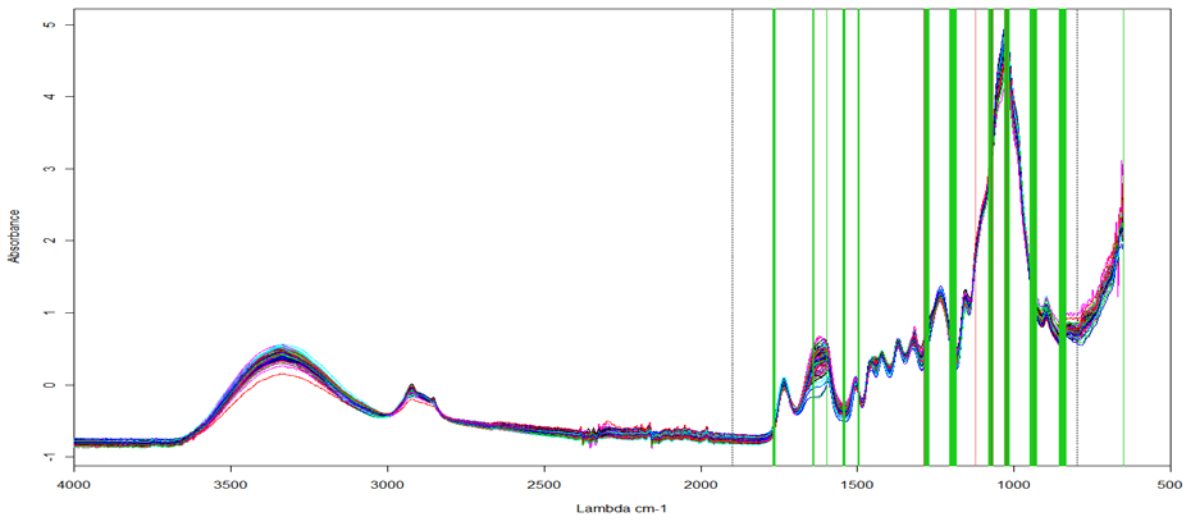
NonGluc_tot :



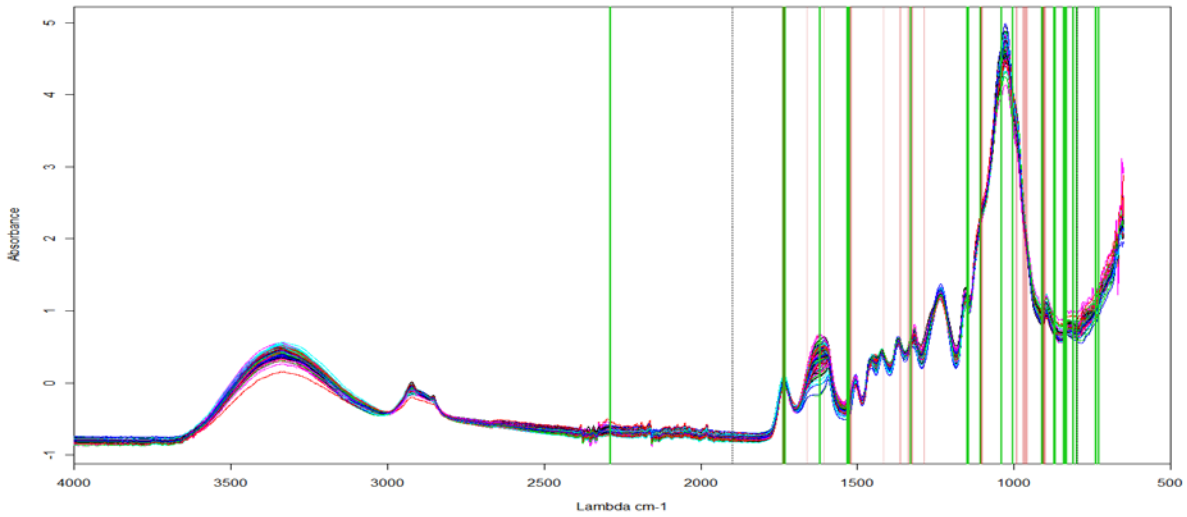
Gluc_tot_prop :



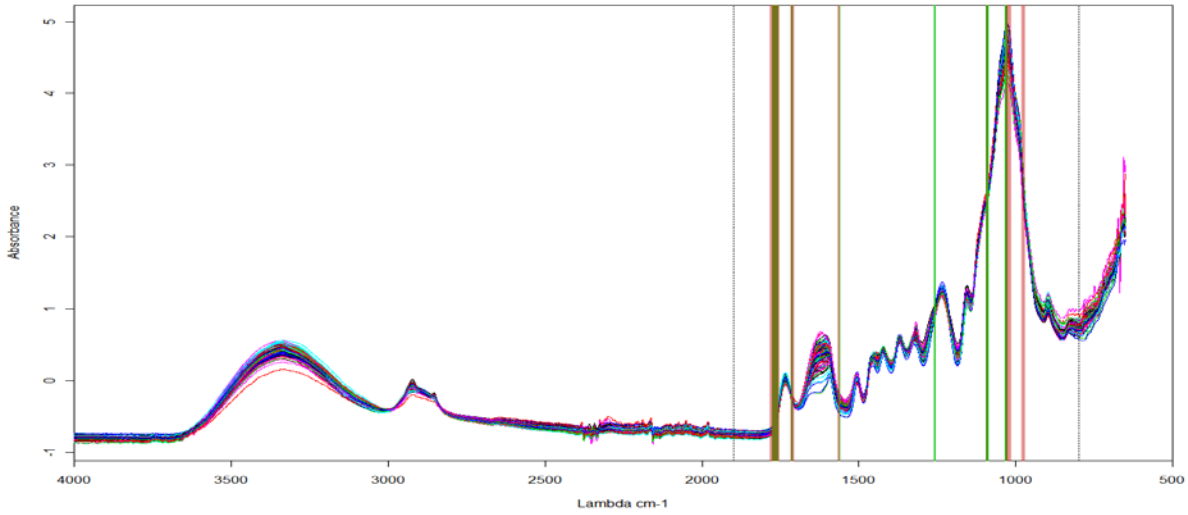
Sucres_hydrol :



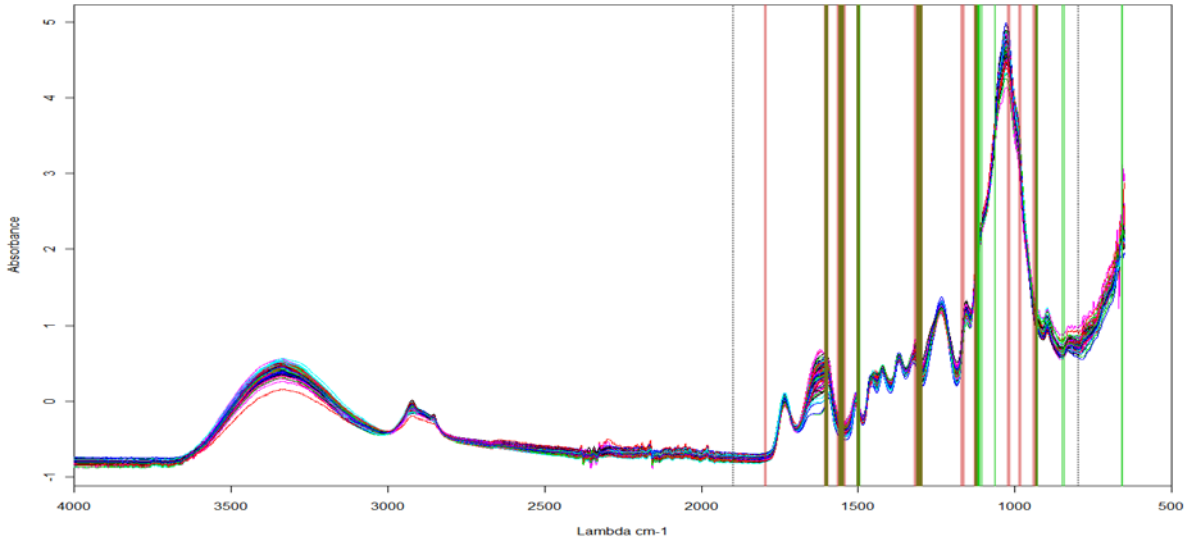
Gluc_hydrol :



NonGluc_hydrol :



Gluc_hydrol_prop :



Résumé

Le travail réalisé durant ce stage se situe dans un contexte de réduction des émissions des gaz à effets de serre, de production de biomasse lignocellulosique qui constitue une ressource d'intérêt pour la production d'énergie et plus particulièrement de biocarburants. L'objectif était d'effectuer des calibrations pour les propriétés chimiques du bois de peuplier en utilisant des données spectrales acquises en moyen infrarouge et de les comparer avec d'autres calibrations déjà obtenues au sein de l'INRA d'Orléans pour des spectres acquis en proche infrarouge.

Tout d'abord une phase d'exploration des données spectrales, chimiques (données acquises par des tests en laboratoires) et combinaison des 2 a été mise en place avec des analyses statistiques descriptives de base, prétraitements, ACP.

Afin de choisir la meilleure méthode à appliquer pour la calibration des modèles, plusieurs méthodes ont été testées comme les régressions PC et PLS avec diverses validations croisées (3, 4, 5 et 100 segments) et MCCV (Monte Carlo Cross Validation), un filtrage des données pour les valeurs aberrantes ainsi que les nombres d'onde sélectionnés avec la méthode CARS. Une fois la meilleure méthode choisie, elle a été appliquée aux différentes variables afin de sélectionner pour chacune des variables chimiques les meilleurs modèles établis avec tout le spectre moyen infrarouge mais aussi avec un sous-échantillon du spectre préalablement découpé.

Enfin, une fois que les meilleurs modèles ont été sélectionnés, ils ont été comparés avec ceux obtenus avec les spectres en proche infrarouge. Ces comparaisons ont été faites à l'aide des statistiques obtenues dans les validations croisées des régressions, comme le R^2, le RMSE et le RPD, mais aussi en prenant en compte le nombre de composantes et d'« outliers ».

Ainsi j'ai pu établir des modèles de calibration de bonne qualité pour presque toutes les variables chimiques. En effet pour la variable NonGluc_hydrol je ne suis pas parvenue à avoir un modèle de calibration correct. Certains des modèles que j'ai obtenus sont meilleurs que ceux en proche infrarouge. C'est le cas des modèles pour les variables chimiques Sucres_sol, Gluc_sol, NonGluc_sol, Gluc_sol_prop, Sucres_tot, NonGluc_tot, Gluc_tot_prop, Gluc_hydrol et NonGluc_hydrol.