N° d'ordre :

Université des Sciences et Technologies de Lille

Ecole doctorale des sciences pour l'ingénieur

THESE

pour obtenir le grade de

Docteur de l'Université des Sciences et Technologies de Lille

Discipline : Instrumentation et analyses avancées

Présentée par

Yves ROGGO

Détermination de la qualité de la betterave sucrière par spectroscopie proche infrarouge et chimiométrie

Soutenue publiquement le 17 juillet 2003 devant le jury :

Mr. DARDENNE, Directeur du département qualité des produits agricoles, Centre de recherches agronomiques, Gembloux (Belgique).
Mr. ULMSCHNEIDER, Responsable du département « NIR-chemometrics », Hoffmann - La Roche Ltd., Basel (Suisse).

Mr. BOUGEARD, Directeur de recherche, Université de Lille 1, Villeneuve d'Ascq.
Mr. BRUANDET, Directeur scientifique, Syndicat national des fabricants de sucre, Paris.
Mr. OPPENHEIM, Professeur, Université d'Orsay, Paris.
Mr. HUVENNE, Professeur, Université de Lille 1, Villeneuve d'Ascq.
Mr. DUPONCHEL, Maître de conférences, Université de Lille 1, Villeneuve d'Ascq.

Remerciements

Ce travail, financé par le Centre National de la Recherche Scientifique et le Syndicat National des Fabricants de Sucre (SNFS), a été réalisé au Laboratoire de Spectrochimie Infrarouge et Raman (LASIR - CNRS UMR 8516).

J'adresse mes remerciements à Monsieur Bougeard, Directeur du LASIR, et Monsieur Bruandet, Directeur scientifique du SNFS, pour m'avoir accepter dans leurs laboratoires respectifs et pour accepter de juger ce travail.

Je suis très sensible à l'honneur que me font Monsieur Dardenne, Directeur du département qualité du centre de recherches agronomiques de Gembloux, et Monsieur Ulmschneider, Responsable du département « NIR and chemometrics » de la société Hoffmann-La Roche, en acceptant d'être rapporteurs. De même, je remercie Monsieur Oppenheim, Professeur à l'université d'Orsay, d'accepter de faire partie de ce jury.

Cette recherche a été réalisée sous la direction de Monsieur HUVENNE, Professeur, responsable du « département IRTF » du LASIR et Monsieur DUPONCHEL, Maître de conférence. Je tiens à les remercier de m'avoir aider aux cours de mes recherches.

Je remercie également toutes les personnes du LASIR, du SNFS (Monsieur NOE en particulier), de la Confédération Générale des planteurs de Betteraves (CGB) et de la société FOSS pour la sympathie qu'ils m'ont témoignée.

Enfin, je n'oublie pas Emeline ROUX qui a aidé à relire et corriger ce manuscrit.

Sommaire

Liste des abréviations

ACP	Analyse en Composantes Principales
ANOVA	Analyse de la variance
CART	« Classification And Regression Trees »
CGB	Confédération Générale des planteurs de Betterave
CLHP	Chromatographie Liquide Haute Performance
CNRS	Centre National de la Recherche Scientifique
DPLS	« Discriminant Partial Least Squares »
DS	« Direct Standardisation »
el.	Elongation
F	Valeur du test Fisher
IR	Infrarouge
KNN	« K Nearest Neighbors »
LASIR	Laboratoire de Spectrochimie Infrarouge et Raman
LDA	« Linear Discriminant Analyse »
LSD	« Least Significant Difference »
LVQ	« Learning Vector Quantification »
MLR	« MultiLinear Regression »
mPLS	« Modified Partial Least Squares »
MSC	« Multiplicative Scatter Correction »
OG	Origine Géographique
PCR	« Principal Component Regression »
PDA	« Procustes Discriminant Analysis »
PDA	« Pulsed Amperometric Detection »
PDS	« Piecewise Direct Standardisation »
PIR	Proche InfraRouge
PLS	« Partial Least Squares »
PNN	« Probabilistic Neural Network »
PR	Période de Récolte
PRESS	« Predictive Residual Error Sum of Squares »
R^2	Coefficient de détermination
RER	Rapport gamme de concentrations / SEP
RPD	Rapport Ecart type des concentrations / SEP
RR	Résistance à la Rhizomanie

SCE	Somme au Carrée des Ecarts
SEC	« Standard Error of Calibation »
SECV	« Standard Error of Cross Validation »
SEP	« Standard Error of Prediction »
SEP(C)	SEP corrigé du biais
SIMCA	« Soft Independant Modelling of Class Analogy »
SNFS	Syndicat National des Fabricants de Sucres
SNV	« Standard Normal Variate »
SNVD	« Standard Normal Variate and Detrending »
SPIR	Spectroscopie Proche InfraRouge
SW	Méthode de transfert d'étalonnage développée par Shenk et Westerhaus
t	Valeur du test de Student

Introduction

L'intérêt porté à la spectroscopie proche infrarouge (SPIR) a été croissant grâce aux améliorations de l'instrumentation, au développement des fibres optiques permettant de délocaliser la mesure et aux progrès de l'informatique.

Il faut souligner l'importance des méthodes mathématiques et statistiques permettant de visualiser, d'extraire et de traiter l'information dans le développement de la SPIR. Les débuts de la chimiométrie datent de 1969 lorsque Jurs et ses collaborateurs[1] ont publié des articles concernant l'utilisation d'une méthode, « Linear Learning Machine », permettant de classer des spectres de masse.

La définition de la chimiométrie est la suivante : il s'agit de la discipline qui utilise les mathématiques et les méthodes statistiques pour sélectionner les procédures expérimentales optimales et pour extraire le maximum d'informations des données issues d'analyses chimiques[2]. La chimiométrie regroupe l'ensemble des méthodes de planification des expériences, des méthodes d'extraction de l'information (modélisation, classification et test d'hypothèses) et des techniques permettant de comprendre des mécanismes chimiques (relation structure / activité des molécules) ou de modéliser un procédé[3,4].

Ainsi la spectroscopie proche infrarouge et la chimiométrie ont prouvé leurs utilités dans des domaines variés tels que l'agriculture, les industries alimentaires, pharmaceutiques, chimiques et pétrolières[5]. Par sa rapidité et par son caractère non destructif, la SPIR est une méthode de choix pour l'analyse des procédés industriels en ligne.

Dans l'industrie sucrière, le dosage du saccharose s'effectue lors de l'achat de la betterave. Actuellement, les industriels cherchent à moderniser leurs centres de réception. Ils souhaitent à la fois automatiser cette étape mais également supprimer l'utilisation de l'acétate de plomb qui est un défécant des jus de betteraves. L'objectif principal de la thèse est de développer un modèle mathématique reliant la teneur en saccharose aux spectres proche infrarouge. Ensuite, il faut montrer que ce modèle est valable sur plusieurs instruments et stable d'une année sur l'autre pour pouvoir appliquer la méthode spectroscopique en usine. Enfin, l'objectif secondaire est d'obtenir une image de la qualité générale de la betterave grâce à la SPIR.

La mise en place d'une application analytique utilisant la SPIR nécessite un ensemble de connaissances pluridisciplinaires. C'est pourquoi les deux premiers chapitres de méthodologie sont importants.

Tout d'abord, le premier chapitre présente la SPIR, son instrumentation et la démarche permettant de développer une analyse quantitative. De même, ce chapitre détaille les méthodes chimiométriques utilisées au cours de ce travail pour extraire et modéliser l'information présente dans les spectres. Enfin, le deuxième chapitre décrit les méthodes chimiques classiquement utilisées pour doser les constituants de la betterave. Ces méthodes serviront de référence lors de l'établissement des modèles quantitatifs.

La partie consacrée à la présentation des résultats et à la discussion est divisée en trois chapitres : le chapitre 3 concerne la faisabilité du dosage du saccharose par la SPIR, le choix de l'instrumentation, du protocole d'analyse et de la modélisation. Le chapitre 4 traite des problèmes de validité du modèle au cours du temps et lors du transfert d'étalonnage entre plusieurs spectromètres. Le chapitre 5 montre que la SPIR peut être envisagée pour la quantification de plusieurs composés de la betterave et que des critères qualitatifs peuvent être prédits grâce aux méthodes de classification supervisées.

Chapitre 1

Applications analytiques de la spectroscopie proche infrarouge (SPIR)

1 Introduction

Les techniques de spectroscopie sont des méthodes physiques de caractérisation. La spectroscopie vibrationnelle peut être définie comme l'étude de l'interaction des ondes électromagnétiques et de la matière[6] sur le domaine des longueurs d'ondes ultraviolet, visible et infrarouge.

Le spectre électromagnétique est généralement divisé comme le montre la Figure 1 en diverses régions en fonction de la longueur d'onde des radiations : ainsi, on trouve les rayons γ qui sont les plus énergétiques, les rayons X, l'ultraviolet, le visible, l'infrarouge (IR), les micro-ondes et les ondes radio fréquences[7]. Pour le domaine de l'IR, les transitions d'énergie observées sont de type vibrationnel. Ce domaine est sous divisé en trois catégories selon la fréquence : le proche infrarouge (PIR) est compris entre 750 nm et 2500 nm, entre 2500 nm et 25000 nm se trouve le domaine de l'IR moyen et l'IR lointain à des longueurs d'onde supérieures à 25000 nm.

Les objectifs de ce chapitre sont de décrire les principes fondamentaux de la spectroscopie proche infrarouge, d'étudier les modalités de l'analyse quantitative par SPIR, de décrire des méthodes chimiométriques et de présenter l'instrumentation proche infrarouge.

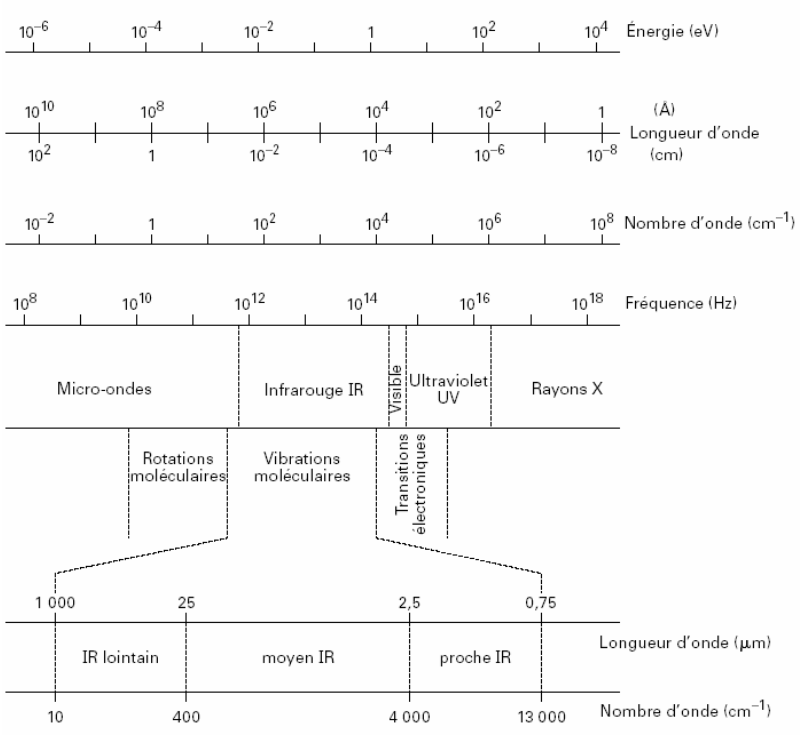


Figure 1 ■ Domaines spectraux du rayonnement électromagnétique (d'après Dalibart et Sevant[8]).

2 Principes fondamentaux de la spectroscopie proche infrarouge

2.1 Rayonnement électromagnétique

La radiation électromagnétique a une double nature ondulatoire et quantique. La principale caractéristique ondulatoire du rayonnement est sa fréquence de vibration ν, exprimée en Hertz. La longueur d'onde λ est la distance parcourue pendant un cycle complet. Elle est reliée à la fréquence par l'Équation 1 :

Équation 1 $\lambda = \frac{c}{\nu}$ avec c la célérité de la lumière (3.10^8 m.s^{-1})

L'approche quantique permet la description des interactions énergétiques avec la matière au niveau moléculaire. Une radiation lumineuse se comporte comme si elle était composée de corpuscules appelés photons. Ces photons possèdent la propriété de transporter une quantité d'énergie finie, liée à la fréquence de la radiation par la relation de Planck (ÉÉquation 2).

Équation 2 $E = h.\nu$ avec h la constante de Planck ($h = 6{,}626176.10^{-34}$ J.s)

2.2 Niveau d'énergie de vibration de la molécule

Les atomes d'une molécule ne restent pas en position fixe les uns par rapport aux autres, ils vibrent autour d'une position moyenne. Dans l'étude des spectres infrarouge, seuls les mouvements vibratoires sont considérés.

2.2.1 Molécule diatomique

- ***Modèle de l'oscillateur harmonique***

L'approche classique de la théorie vibrationnelle consiste dans un premier temps à étudier le cas de l'oscillateur harmonique pour une molécule diatomique (Figure 2). Dans une molécule diatomique, les atomes subissent deux phénomènes contradictoires. En effet, il y a une répulsion entre les nuages électroniques négatifs. De plus, les électrons d'un atome et le noyau de l'autre s'attirent mutuellement. Il en résulte une vibration continuelle de la liaison.

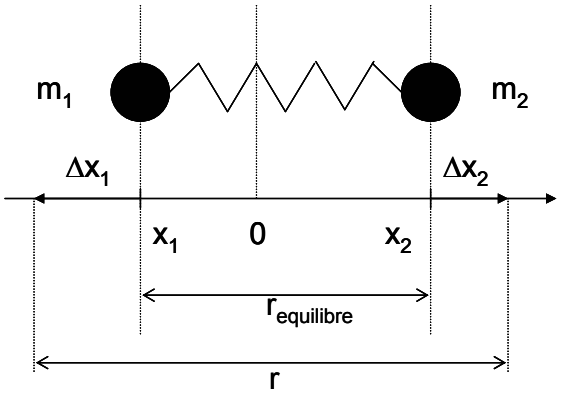


Figure 2 ■ Modèle de l'oscillateur

La force résultante de ces actions passe par un minimum à la distance d'équilibre (Figure 3). L'énergie potentielle (V) de ce système est décrite par l'équation suivante :

Équation 3 $V = 1/2\ k\ (r - r_{équilibre})^2 = 1/2\ k\ x^2$

Avec r = distance internucléaire, $r_{équilibre}$ = distance internucléaire à l'équilibre,

$x = r - r_{équilibre}$ = déplacement, k = constante de force (en $N.m^{-1}$).

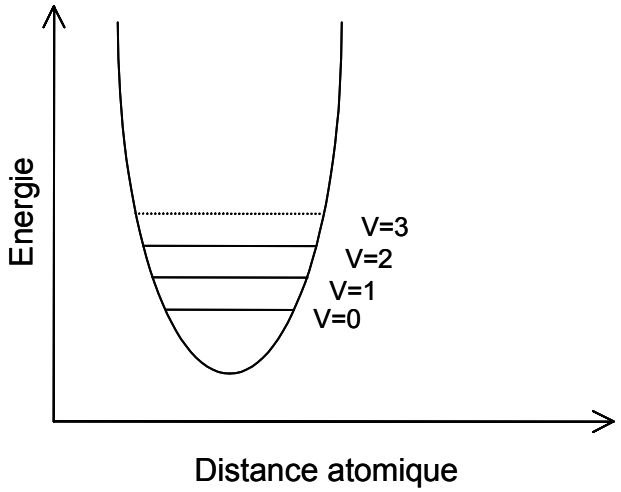


Figure 3 ■ Potentiel harmonique – Représentation de l'énergie potentielle de liaison en fonction de la distance interatomique.

Sans détailler les diverses étapes de calcul, largement explicitées[9], on obtient les résultats suivants. En mécanique classique, la fréquence de l'oscillateur harmonique est donnée par :

Équation 4 $\upsilon = \frac{1}{2\pi}\sqrt{\frac{k}{\mu}}$ avec $\mu = \frac{m_1 \times m_2}{m_1 + m_2}$ la masse réduite, k constante de force et m_1, m_2 les masses des 2 atomes.

Lorsqu'une molécule interagit avec une onde électromagnétique de fréquence ν l'augmentation de la quantité d'énergie contenue dans le système diatomique se traduit par une absorption lumineuse qui peut être enregistrée.

La mécanique quantique montre que toutes les valeurs de l'énergie totale ne sont pas possibles au niveau moléculaire. Plank proposa en 1900, l'hypothèse selon laquelle les changements énergétiques d'un système ne peuvent se faire que par sauts discontinus. L'énergie moléculaire ne prend que des valeurs discrètes. Les équations quantiques de Schrödinger permettent de calculer les énergies des vibrations possibles pour une molécule. Pour l'oscillateur harmonique simple, la solution de ces équations est :

Équation 5 $E_v = (v + \frac{1}{2}).h.\upsilon$ avec v le nombre quantique vibrationnel.

Le niveau énergétique exprimé en cm^{-1} est :

Équation 6 $$G(\upsilon) = E_v / h.c = (v + \frac{1}{2}).\bar{\upsilon}$$ avec $\bar{\upsilon} = 1 / \lambda$ le nombre d'onde en cm^{-1}

Ainsi seules certaines énergies sont permises et même pour v = 0, correspondant à la plus basse énergie, l'énergie de vibration n'est pas nulle. Les atomes ne peuvent pas être complètement immobiles.

L'équation de Schrödinger donne également les conditions de variation d'énergie. Le saut quantique ne peut correspondre qu'à une seule valeur énergétique. Il est possible par absorption ou émission d'une quantité E de passer d'un niveau énergétique à l'autre. La variation d'énergie pour passer d'un état à l'autre est :

Équation 7 $$\Delta E = E_{v+1} - E_v = h.\nu$$

La transition autorisée de v = 0 à v = 1 est appelée transition fondamentale. Les autres transitions autorisées (passage de v = 1 à v = 2 et v = 2 à v = 3…) correspondent à des bandes plus faibles en énergie. On les appelle bandes chaudes (« hot bands ») car la température augmente leur intensité. Dans le cas de l'oscillateur harmonique, ces « hot bands » ont la même fréquence que la transition fondamentale.

Le modèle harmonique est simple et son application est limitée aux très faibles déplacements des atomes. Certes, il peut donner une idée sur la position des bandes fondamentales ou sur la force des liaisons, mais il n'est pas représentatif des molécules réelles[7]. Dans ce modèle, il est admis que les liaisons sont parfaitement élastiques. Or, ces liaisons peuvent se briser quand l'amplitude des vibrations devient importante. La forme de la courbe d'énergie est donc plus complexe qu'une simple parabole.

- ***Modèle anharmonique***

Afin de remédier aux incertitudes générées par le modèle harmonique et d'obtenir une meilleure approximation des états vibrationnels des molécules réelles, le modèle de l'oscillateur anharmonique a été mis en place.

Deux observations montrent que les molécules ne sont pas des oscillateurs idéaux. Premièrement, les bandes chaudes n'ont pas la même fréquence que les bandes fondamentales. Deuxièmement, les transitions harmoniques sont autorisées de v = 0 à v = 2, 3, 4… Cette déviation du comportement harmonique peut s'expliquer par deux effets :

Le premier est l'anharmonicité mécanique dûe aux effets cubiques et de termes plus élevés dans l'expression de l'énergie potentielle (Équation 8).

Équation 8 $V = 1/2\ k.x^2 + k'.x^3 + \ldots$ avec $k' << k$

Le second effet est appelé anharmonicité électrique. Il est responsable des harmoniques qui correspondent aux transitions mettant en jeu 2v ou 3v. Comme le montre la figure précédente, les fréquences d'absorption des harmoniques ne sont pas exactement des multiples de la fréquence fondamentale.

De manière générale, on obtient une courbe d'énergie potentielle en fonction du déplacement comme le montre la Figure 4. En mécanique quantique, l'application du modèle de Morse dans les équations de Schrödinger donne l'équation ci-dessous.

Équation 9 $E_v = (v + \frac{1}{2}).h.\upsilon_a - (v + \frac{1}{2})^2.h.\upsilon_a.x_e$ avec υ_a la fréquence d'oscillation et x_e la constante d'anharmonicité

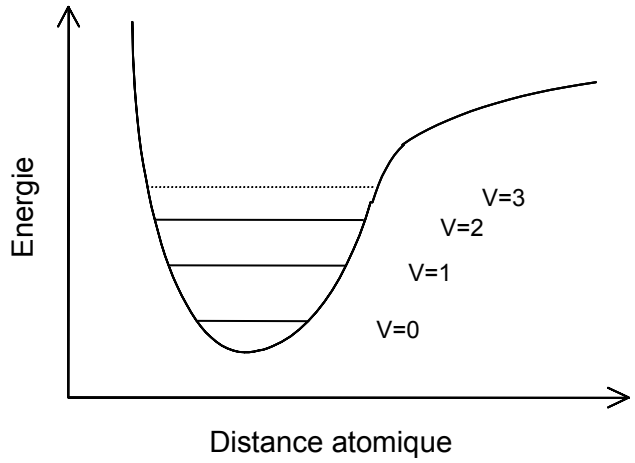


Figure 4 ■ Courbe d'énergie potentielle en fonction du déplacement dans le cadre du modèle anharmonique.

Dans l'hypothèse d'un équilibre thermodynamique, la loi de Boltzmann, qui permet de calculer le nombre de molécules se trouvant à un niveau d'énergie donné, montre que la plupart des molécules sont au niveau d'énergie le plus faible correspondant à v = 0. Les trois transitions observables dans les conditions habituelles mettent en jeu les variations Δv = +1, +2 ou +3. La constance d'anharmonicité étant très faible, les trois bandes spectrales sont à des fréquences voisines de $\nu, 2\nu, 3\nu$. La bande ν est appelée fondamentale, les bandes 2ν et 3ν sont appelées premières et secondes harmoniques.

2.2.2 Molécules polyatomiques

Ces propriétés peuvent être étendues à des molécules polyatomiques. Dans ce cas, la molécule est considérée comme un ensemble d'oscillateurs pouvant interagir. Une molécule contenant N atomes aura 3N-6 degrés de liberté de vibration (3N-5 pour une molécule linéaire). Ce nombre de degrés de liberté donne le nombre de fréquences de vibrations fondamentales. Un mode normal de vibrations correspond à un mouvement interatomique où tous les atomes vibrent à la même fréquence mais avec des amplitudes différentes.

- ***Approximation harmonique***

En mécanique classique, dans une première approximation, la molécule peut être considérée comme une superposition de 3N-6 mouvements harmoniques simples. Comme exemple, on peut considérer une molécule triatomique non linéaire telle que H_2O qui a trois vibrations fondamentales ν_1, ν_2 et ν_3.

Dans l'hypothèse de l'harmonicité, l'énergie de vibration de la molécule sera :

Équation 10 $E_v(v_1, v_2, v_3) = h.\nu_1 (v_1+1/2) + h.\nu_2 (v_2+1/2) + h.\nu_3 (v_3+1/2)$ avec v_1, v_2 et v_3 les nombres quantiques des trois vibrations normales.

Soit en nombre d'onde :

Équation 11 $G(v_1, v_2, v_3) = \bar{\upsilon}_1 (v_1+1/2) + \bar{\upsilon}_2 (v_2+1/2) + \bar{\upsilon}_3 (v_3+1/2)$

- ***Influence de l'anharmonicité***

Pour la molécule triatomique, en tenant compte de l'anharmonicité, l'énergie de vibration n'est plus une somme de termes indépendants mais contient des termes croisés produits de deux vibrations normales :

$$\begin{aligned} G(v_1, v_2, v_3) = {} & \bar{\upsilon}_1 (v_1+1/2) + \bar{\upsilon}_2 (v_2+1/2) + \bar{\upsilon}_3 (v_3+1/2) + X_{11}(v_1+1/2)^2 \\ & + X_{22} (v_2+1/2)^2 + X_{33} (v_3+1/2)^2 + X_{12}(v_1+1/2)(v_2+1/2) + \\ & + X_{23} (v_2+1/2)(v_3+1/2) + X_{13} (v_1+1/2)(v_3+1/2) \end{aligned}$$

Équation12

avec X_{ik} la constante d'anharmonicité correspondant à la molécule diatomique

En plus des harmoniques, des bandes de combinaisons peuvent être observées lorsque plusieurs vibrations interagissent pour donner des bandes dont la fréquence est la somme ou la différence de fréquences (Équation 13). Les bandes d'absorption seront plus faibles que les fondamentales et les combinaisons tertiaires seront plus faibles que les

secondaires. En pratique, seules les combinaisons de deux termes ont une intensité suffisante pour être observées[10].

Équation 13 $\nu = \Sigma\, n_i.\nu_i$

Les transitions apparaissent dans le spectre infrarouge seulement si elles mettent en jeu un changement du moment dipolaire de la molécule. Les changements du moment dipolaire et donc les modes normaux de vibration peuvent être prédits à partir de la symétrie de la molécule. Cependant d'autres phénomènes viennent compléter cette observation :

- Résonance de Fermi : une résonance conduit à une perturbation des niveaux d'énergie si deux niveaux de vibration ont des énergies similaires. Ce type de résonance peut se produire lorsque des bandes harmoniques ont la même symétrie et des fréquences voisines de la bande fondamentale. Les deux bandes sont observées à des fréquences plus élevées et moins élevées que les positions attendues de la bande fondamentale et de l'harmonique.

- Modèle des modes locaux : alors que l'interprétation des modes normaux de vibration est utilisée sur les régions de fréquences de la vibration fondamentale, le modèle des modes locaux est approprié pour décrire les hauts niveaux énergétiques. Les molécules contenant des liaisons symétriquement équivalentes comme H_2O forment un système de deux oscillateurs couplés vibrant soit en phase (ν_s) soit en opposition de phase (ν_a):

Équation 14 $\bar{\upsilon}_s = \bar{\upsilon}_M - C$

Équation 15 $\bar{\upsilon}_a = \bar{\upsilon}_M + C$

avec $\bar{\upsilon}_M$ le nombre d'onde associé à la liaison considérée comme un oscillateur indépendant et C la constante de couplage.

En fonction du rapport (constante de couplage C divisé par la constante d'anharmonicité X), on peut déterminer si la molécule a un comportement plutôt normal ou plutôt local. Par exemple, SO_2 a une constante de couplage très grande, dominant la constante d'anharmonicité ce qui conduit à un comportement « modes normaux ». Par contre, la molécule d'eau a une constante de couplage plus faible que la constance d'anharmonicité. Ainsi, les modes locaux sont utilisés pour rendre compte du comportement de la molécule d'eau[5].

2.3 Attribution des bandes spectrales

Les combinaisons et les harmoniques sont le cœur de la SPIR. L'anharmonicité détermine la fréquence et l'intensité des bandes. Les liaisons ayant l'anharmonicité la plus grande sont celles mettant un jeu l'hydrogène. Ces liaisons vibrent avec une énergie élevée et une large amplitude. C'est pourquoi l'interprétation du domaine PIR est dominée par l'absorption des liaisons du groupe fonctionnel XH_n.

De nombreux auteurs ont essayé d'attribuer les bandes spectrales du proche infrarouge[11,12]. L'identification des maxima d'absorption reste difficile car les bandes sont larges, souvent de faibles intensités. De plus les interactions entre les molécules peuvent entraîner des déplacements de longueurs d'onde[13]. Dans ce paragraphe, l'interprétation générale du spectre proche infrarouge sera présentée. Les bandes principales des glucides et de l'eau seront ensuite décrites à titre d'exemples.

2.3.1 Interprétation générale du spectre infrarouge

Tout d'abord, il est possible de délimiter les régions des bandes de combinaisons et des harmoniques sur la gamme [700 nm ; 2500 nm] (Figure 5).

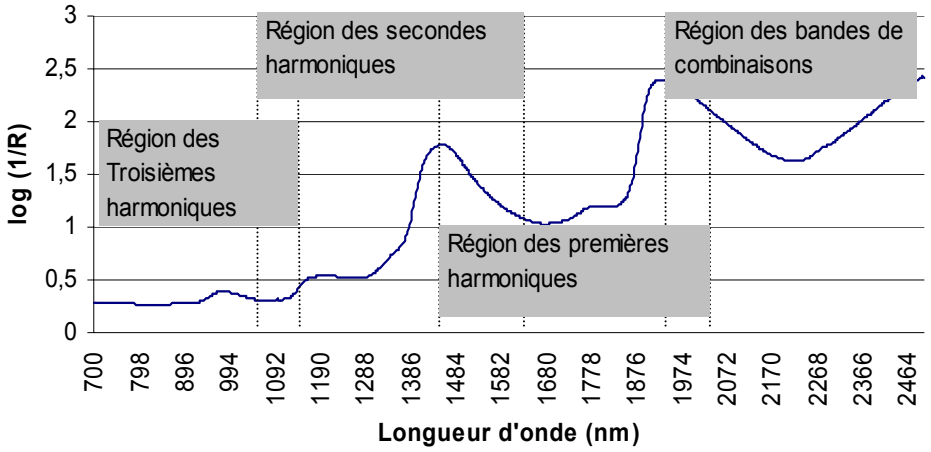


Figure 5 ■ Domaines spectraux des bandes de combinaisons et des harmoniques.

De plus, il est également possible de corréler le spectre à la structure chimique comme le montre la Figure 6. Cette figure est un résumé, elle ne présente que les bandes les plus importantes du domaine PIR. On constate la répétition de l'information spectrale dans les différentes régions (combinaisons et harmoniques). Ainsi, la construction de modèles quantitatifs sera facilitée.

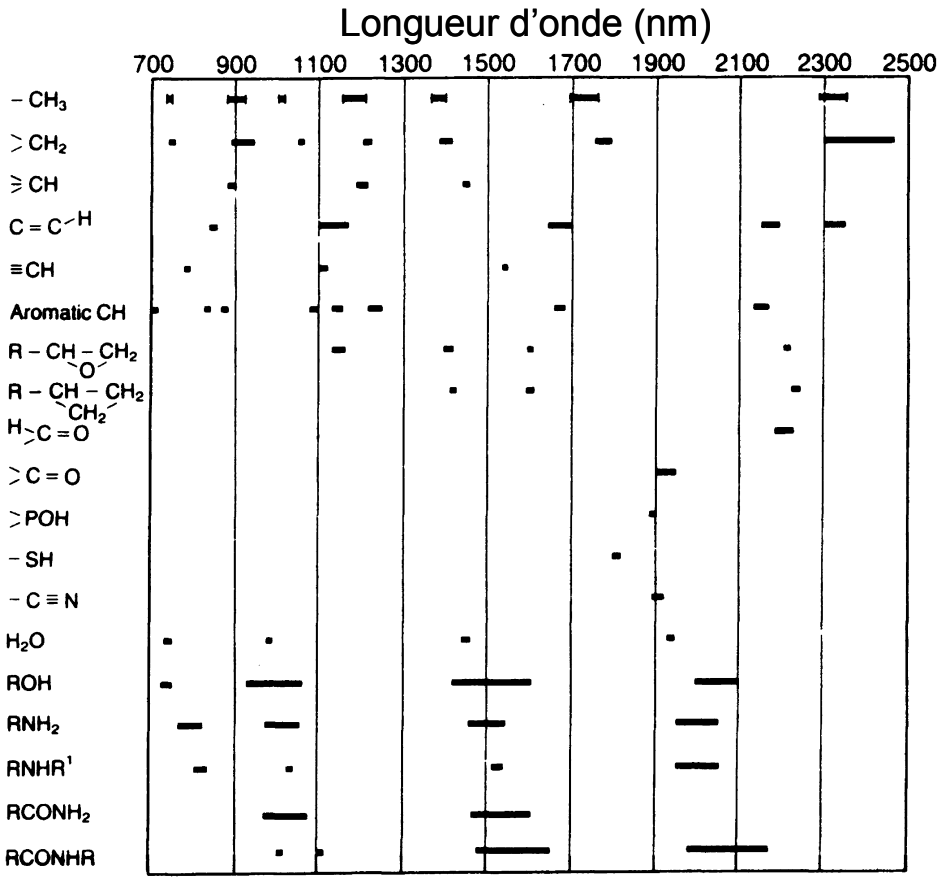


Figure 6 ■ Longueurs d'onde caractéristiques de quelques groupements chimiques (d'après Osborne et ses collaborateurs[14]).

2.3.2 Exemples d'attribution de bandes dans le proche infrarouge

Les deux composés majoritaires de la betterave sont l'eau et les glucides. C'est pourquoi, à titre d'exemples, il est apparu intéressant de développer l'attribution spectrale de ces composés. En ce qui concerne les autres composés tels que les lipides[15] ou les protéines[16], des ouvrages décrivent précisément l'attribution de leurs bandes.

- ***Attribution des bandes de l'eau***

La molécule d'eau présente trois modes fondamentaux de vibrations (Figure 7). Les fréquences fondamentales sont attribuées pour l'eau à l'état vapeur.

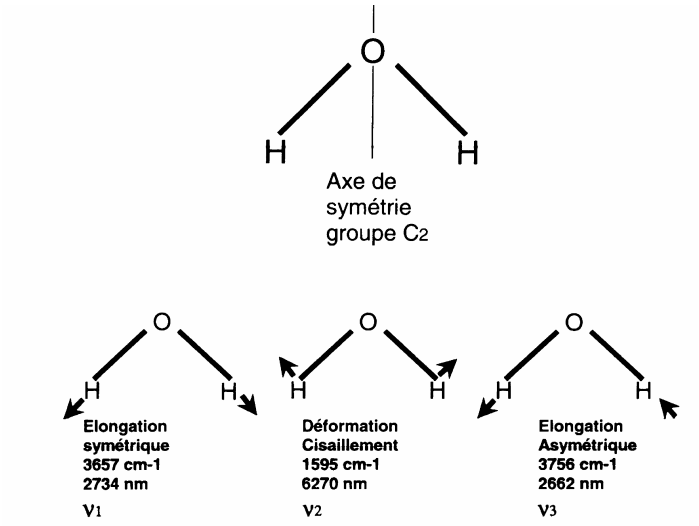


Figure 7 ■ Modes fondamentaux de vibration de la molécule d'eau (d'après D. Bertrand[18]).

Dans l'eau liquide, des liaisons hydrogènes apparaissent entre les molécules d'eau. Dans ces conditions, les fréquences fondamentales sont modifiées. D'après Iwamoto et ses collaborateurs[17], une attribution des bandes du proche infrarouge du spectre de l'eau à l'état liquide peut être effectuée :

- λ = 760 nm correspond à la troisième harmonique des bandes élongations $3\nu_3$.
- λ = 960 nm correspond à la combinaison $(2.\nu_1 + \nu_2)$.
- λ = 1150 nm correspond à la combinaison $(\nu_1 + \nu_2 + \nu_3)$.
- λ = 1440 nm correspond à la combinaison $(\nu_1 + \nu_3)$.
- λ = 1930 nm correspond à la combinaison $(\nu_2 + \nu_3)$.

Cependant, le spectre de l'eau pure est modifié par la présence de solutés et par des variations de la température[18]. Ainsi, Lin et Brown[19] ont mis en évidence un déplacement de - 34 nm du pic à 1440 nm quand l'eau passe de 5 °C à 65 °C. Ce décalage est attribué au fait que l'énergie des liaisons O-H de l'eau augmente en fonction de la température. Il y a alors augmentation de la fréquence de vibration. La présence de solutés a également un effet significatif sur le spectre proche infrarouge. Par exemple, le pic d'absorption à 1440 nm est déplacé vers les hautes fréquences par la présence de chlorure de sodium[20].

- ***Attribution des bandes des glucides***

Les spectres PIR des sucres à l'état solide présentent une allure générale similaire et des bandes faiblement résolues. De plus, en solution, les différences spectrales entre les glucides sont masquées par les bandes de l'eau. L'attribution des fréquences d'absorption est cependant possible et permet de différencier les groupements chimiques. D'après Cadet et ses collaborateurs[21], les spectres PIR des glucides présentent deux zones distinctes : la première, comprise entre 1100 et 1800 nm, correspond aux premières et deuxièmes harmoniques des groupements OH et CH et la seconde, allant de 1800 à 2500 nm, est représentative des bandes de combinaisons de ses mêmes groupements. Workman[22] propose les attributions figurant dans le tableau suivant pour les bandes de vibration CH, CC et OH des glucides dans le proche infrarouge.

Tableau 1 ■ Bandes de vibrations associées aux polysaccharides et monosaccharides dans le proche infrarouge. (avec él. = élongation) (d'après Cadet et collaborateurs[21])

Bandes associées	Longueurs d'onde
Combinaison C-H él./ C-C él. et C-O él.	2500 nm
Combinaison C-H él./ CH_2 déformation	2280 – 2330 nm
Combinaison O-H él./ ZOH déformation	2100 nm
O-H él. 1ère harmonique	1450 nm
O-H él. 2ème harmonique	1010 – 1030 nm
C-H él. 3ème harmonique	850 – 900 nm

Pour les monosaccharides, les bandes d'absorption sont centrées sur 1457 nm (première harmonique de l'élongation O-H) et sur 2062 nm (élongation O-H et déformation O-H). Des pics secondaires sont également identifiables à 2263 nm (élongation O-H et déformation O-H) et à 2440 nm (élongation C-H + élongation C-C). Pour les polysaccharides, on peut noter un déplacement de ces bandes, sur la gamme spectrale 1100 nm à 2170 nm, vers des valeurs plus faibles (1432 et 1931 nm) et vers des valeurs plus importantes sur la gamme spectrale 2170 nm à 2500 nm (2310 nm et 2477 nm). Le Tableau 2 montre les principales longueurs d'onde du glucose, fructose et du saccharose.

Tableau 2 ■ Attribution des bandes d'absorption du glucose, fructose et du saccharose entre 1100 nm et 2500 nm (d'après Cadet et ses collaborateurs[21]). él. = élongation.

Longueur d'onde (nm)	Référence	Attribution
Glucose		
1493 – 1598	Osborne et collaborateurs[23]	Première harmonique O-H
2085	Ghosh	él. O-H + déformation O-H
2275	Osborne et collaborateurs[23]	él. OH + él. C-C
2340	Ghosh	él. symétrique CH_2 + déformation CH_2
Fructose		
2073	Diffee[24]	él. O-H + déformation O-H
2282	Meurens et Alfaro[25]	él. symétrique CH_2 + déformation CH_2
Saccharose		
1480	Trott et collaborateurs[26]	él. O-H première harmonique
2080	Osborne et collaborateurs[14]	él. O-H + déformation O-H
2145	Diffee[24]	él. O-H + déformation O-H
2276	Law et Tkachuk[27]	él. O-H + él. C-C

2.4 Interaction entre le rayonnement et la matière

Après avoir décrit l'influence de la lumière au niveau moléculaire, l'objectif de ce paragraphe est de comprendre l'interaction entre le rayonnement électromagnétique et l'échantillon. Lorsque une radiation monochromatique éclaire un échantillon, elle peut être absorbée, transmise ou réfléchie (Figure 8). La loi de conservation de l'énergie permet d'écrire :

Équation 16 $I_0 = I_A + I_T + I_R$

Avec I_0 l'intensité du rayonnement incident,

I_A l'intensité absorbée,

I_T l'intensité transmise

et I_R l'intensité réfléchie.

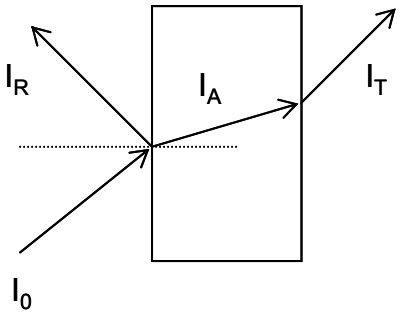


Figure 8 ■ Interaction de la radiation avec la matière.

L'expérimentation est réalisée pour que I_T (ou I_R) soit nulle ainsi I_A peut être déduite de la mesure de I_T (ou de I_R). Ainsi la mesure s'effectue soit en transmission soit en réflexion.

2.4.1 Transmission

La radiation qui traverse un échantillon, c'est à dire un milieu d'indice de réfraction différent, est soumise au phénomène d'absorption. L'atténuation de la radiation par absorption est décrite par la loi de Beer Lambert :

Équation 17 $\log (I_0 / I_T) = \varepsilon.c.L$ avec ε constante d'absorption, c la concentration et L l'épaisseur de l'échantillon.

Si c est exprimée en $mol.L^{-1}$ et L en cm, ε s'exprime en $L.mol^{-1}. cm^{-1}$ et s'appelle constante d'absorption molaire.

La loi de Beer Lambert n'est pas vérifiée si une partie du rayonnement est réfléchie, si des phénomènes de diffusion existent ou si des radiations parasites perturbent la mesure. De même, la linéarité de cette relation n'est plus vérifiée si les concentrations et les niveaux d'absorbances sont élevés[14].

2.4.2 Réflexion

- ***Réflexion spéculaire***

Pour un échantillon non absorbant et opaque, la radiation incidente est réfléchie totalement selon les principes d'optique : l'angle d'incidence est égal à l'angle de réflexion. La réflexion spéculaire à l'interface échantillon / air est décrite par les équations de Fresnel (Équation 18). Si l'échantillon est absorbant, l'indice de réfraction est un nombre complexe

n.(1-i.k) où n est la partie réelle du nombre complexe et k la constante d'absorption (Équation 19).

Équation 18 $$\frac{I_R}{I_0} = \frac{(n_2 - n_1)^2}{(n_2 + n_1)^2}$$

Équation 19 $$\frac{I_R}{I_0} = \frac{(n_2 - n_1)^2 + (n_2 k)^2}{(n_2 + n_1)^2 + (n_2 k)^2}$$ avec n_1 et n_2 les indices de réfraction des deux milieux.

Lorsque des échantillons partiellement transparents sont analysés, des cellules de mesure disposant d'une face réfléchissante (réflecteur) sont utilisées. Ainsi la réflexion et la transmission sont mises en jeu : on parle alors de transflexion.

- ***Réflexion diffuse***

Lorsqu'une radiation rencontre une particule de taille supérieure à la longueur d'onde la radiation se propage dans toutes les directions. Ce phénomène découvert par Tyndall en 1869 s'appelle la diffusion. Si la radiation est transmise à travers la première interface, une partie de la radiation est ensuite absorbée. La radiation sortant de l'interface est diffusée (Figure 9).

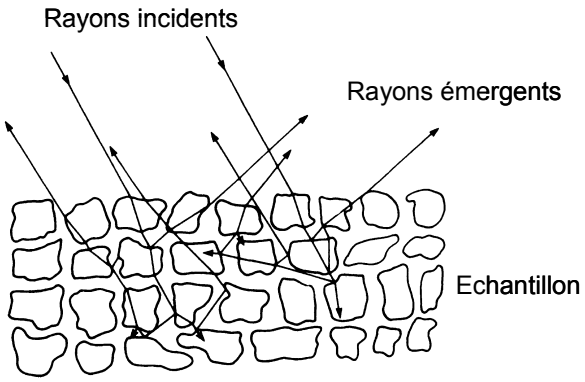


Figure 9 ■ Réflexion diffuse (d'après Osborne et collaborateurs[14])

Le chemin optique exact est extrêmement difficile à décrire de façon mathématique. En conséquence, aucune théorie rigoureuse n'a été proposée. Mais plusieurs théories décrivant le phénomène ont été développées, la plus connue étant celle de Kubelka et Munk[28,29]. Selon cette théorie, on obtient une relation simple entre la réflectance R et la concentration c. La théorie de Kubelka Munk a été vérifiée expérimentalement par Butler et Norris[30].

3 Méthodologie analytique

Après avoir expliqué les principes fondamentaux de la spectroscopie proche infrarouge (SPIR), cette partie décrit comment une application analytique utilisant la SPIR est développée.

Le développement d'une application analytique quantitative utilisant la SPIR s'inscrit dans le cadre général décrit par la Figure 10. Avant tout il faut déterminer le composé à analyser, choisir le principe de la méthode et ses conditions d'utilisation. Il y a ensuite une étape de développement au sein d'un laboratoire de référence. Enfin, il faut vérifier que la méthode est répétable et reproductible avant l'utilisation en routine. Il est également conseillé d'effectuer des tests de robustesse pour s'assurer que la méthode n'est pas influencée par de petites variations des conditions d'utilisation.

Cependant, l'étape de développement d'une solution analytique utilisant la SPIR est particulière. C'est pourquoi, nous allons la décrire.

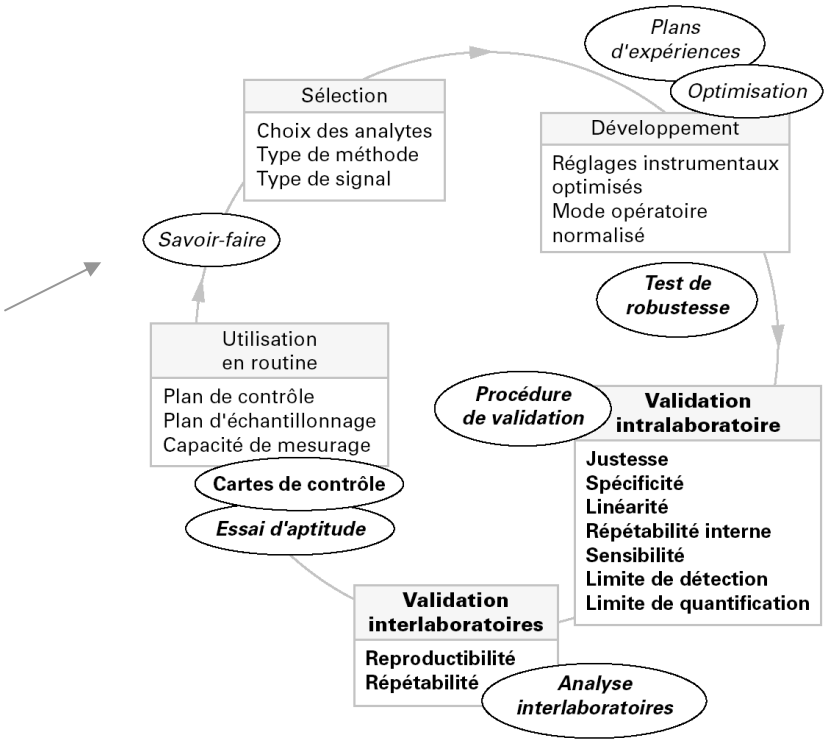


Figure 10 ■ Cycle de vie d'une méthode analytique (d'après M. Feinberg[31]).

Les notations suivantes seront utilisées :

- Les constantes sont notées par des lettres minuscules ou majuscules.
- Les variables scalaires sont notées par des lettres minuscules. $y_{j,k}$ représente une concentration pour l'échantillon j et un constituant k et $x_{j,i}$ l'absorbance à la longueur d'onde i pour l'échantillon j.
- Les vecteurs **x** sont notés en gras et en minuscule.
- Les matrices sont représentées par des lettres majuscules en gras. **Y** est la matrice des concentrations et **X** est la matrice des absorbances. **X**' est la transposée de la matrice **X** et $\mathbf{X}^{-1}$ sa matrice inverse.

3.1 Description générale de l'analyse quantitative par spectroscopie proche infrarouge

L'analyse quantitative[32] se divise en trois étapes :

- L'étape d'étalonnage permet de construire l'équation d'étalonnage, c'est à dire le modèle mathématique qui relie les absorbances aux différentes longueurs d'onde à la concentration recherchée.

- L'étape de validation vérifie si le modèle obtenu donne des résultats satisfaisants sur un lot d'échantillons différent de celui utilisé lors de la phase d'étalonnage.

- Enfin, l'étape de prédiction correspond à l'utilisation en routine de la technique spectroscopique en s'affranchissant de la méthode chimique de référence.

3.2 Critères statistiques pour l'évaluation de l'analyse quantitative

Différents critères statistiques permettent d'évaluer la qualité des étapes d'étalonnage et de validation. Tout d'abord, deux paramètres permettent de vérifier les performances d'étalonnage : il s'agit de l'erreur standard d'étalonnage (SEC – « Standard Error of Calibration »), du coefficient de détermination R^2 qui correspond au carré du coefficient de corrélation R (Équation 21). Le SEC (Équation 20) mesure les écarts entre la valeur prédite par SPIR et la valeur de référence. Le coefficient R^2 estime la variance expliquée par la régression. Il faut remarquer que faute de définitions normalisées les erreurs d'étalonnage sont parfois exprimées par d'autres paramètres tel que le PRESS (« Predictive Residual Error Sum of Squares» - Équation 24).

Les formules suivantes sont décrites par Naes[33].

Équation 20 $$SEC = \sqrt{\frac{\sum_{j=1}^{m}(y_j - y'_j)^2}{m-1-q}}$$

Équation 21 $R = \frac{cov(y',y)}{\sigma_{y'}.\sigma_y}$ avec cov(y',y) la covariance de y' et de y et σ_y, $\sigma_{y'}$ les écarts-types respectifs de y et y'.

Équation 22 $$cov(y',y) = \sqrt{\frac{\sum_{j=1}^{m}(y_j - \bar{y})(y'_j - \bar{y'})}{m-1}}$$

Équation 23 $$\sigma_y = \sqrt{\frac{\sum_{j=1}^{m}(y_j - \bar{y})^2}{m-1}}$$

Équation 24 $PRESS = \sum_{j=1}^{n}(y_j - y'_j)^2$ avec :

y_j: la concentration de référence pour l'échantillon j,

y'_j : la concentration prédite pour l'échantillon j,

m : nombre d'échantillons dans le lot d'étalonnage,

n : nombre d'échantillons dans le lot de validation,

q : nombre de termes de la régression.

Un SEC faible est une condition nécessaire mais non suffisante pour valider la méthode SPIR. L'étape de validation permet de contrôler l'équation d'étalonnage sur un lot d'échantillons indépendants. Au cours de cette étape, différents indicateurs statistiques sont calculés : le coefficient de détermination, l'erreur standard de prédiction SEP (« Standard Error of Prediction ») (Équation 25), le biais (Équation 26) et le SEP corrigé du biais (Équation 27). Le biais correspond à la moyenne des écarts entre la méthode de référence et la SPIR, il s'agit de l'erreur systématique entre les deux méthodes. Le SEP(C) est en fait l'écart-type des écarts entre la mesure de référence et la mesure SPIR. Deux autres indicateurs peuvent être également utilisé : le rapport écart-type des concentrations sur SEP

est noté RPD (Équation 28) et le rapport gamme de concentration sur SEP est appelé RER (Équation 29).

Équation 25 $$SEP = \sqrt{\frac{\sum_{j=1}^{n}(y_j - y'_j)^2}{n}}$$

Équation 26 $$biais = \frac{\sum_{j=1}^{n}(y_j - y'_j)}{n}$$

Équation 27 $$SEP(C) = \sqrt{\frac{\sum_{j=1}^{n}(y_j - y'_j - biais)^2}{n-1}}$$

Équation 28 $RPD = \sigma_{référence} / SEP$

Équation 29 RER = Gamme de concentration / SEP

Les paramètres suivants SEC, SEP, biais et SEP(C) sont exprimés dans l'unité de la méthode chimique de référence. Le coefficient de corrélation, le RER et le RPD sont sans unité. Si le SEC et le SEP sont faibles et les coefficients de corrélation proches de 1 (lors de la phase d'étalonnage et de validation), l'analyse quantitative peut être considérée comme satisfaisante. Pour développer des modèles quantitatifs, il est nécessaire d'utiliser des méthodes mathématiques qui seront présentées dans le paragraphe suivant.

4 Chimiométrie en spectroscopie proche infrarouge

L'objectif de ce paragraphe est de présenter les outils mathématiques utilisés pour visualiser les données, prétraiter les spectres proche infrarouge, modéliser une information quantitative et classer les échantillons à partir de leurs spectres. Enfin, les tests statistiques utilisés pour comparer les modèles quantitatifs et les méthodes de régression seront présentés.

4.1 Prétraitements mathématiques appliqués aux spectres

Le spectre proche infrarouge est affecté par la taille des particules de l'échantillon[34] et par des variations du chemin optique[35]. C'est pourquoi il est nécessaire d'avoir un protocole de préparation et d'analyse de l'échantillon bien défini[36]. Pour éliminer ou diminuer ces interférences, des prétraitements mathématiques sont appliqués aux spectres. Les traitements les plus couramment utilisés sont décrits dans ce paragraphe.

4.1.1 Dérivée

La dérivée a été historiquement le premier prétraitement utilisé. Elle permet de réduire la dérive de la ligne de base[37], de séparer plus clairement les bandes d'absorption[38] et de mettre en évidence certaines parties de l'information spectrale[39].

Il existe différentes méthodes pour calculer la dérivée :

- La dérivée par intervalle[40,41] (« gap derivative ») calcule la dérivée sur un intervalle de points fixés par l'utilisateur. Dans notre étude, seule la dérivée par intervalle est utilisée.

- La dérivée basée sur la technique de convolution de Savitsky et Golay[42]. Deux étapes sont nécessaires pour calculer la dérivée en un point i : tout d'abord, un polynôme de degré k est ajusté sur au moins k+1 points du spectre autour du point i. Ensuite, la dérivée du polynôme en ce point est calculée.

4.1.2 Lissage

Les données initiales contiennent un bruit, une erreur non systématique. Le lissage permet de diminuer cette erreur aléatoire. La méthode utilisée dans notre étude est celle de la moyenne mobile. La valeur $x_{j,i}$ est remplacée par la moyenne pondérée de $x_{j,i}$ et de ces voisins sur un intervalle compris entre i-D et i+D.

Équation 30 $$x_{j,i\ \text{lissé}} = \sum_{d=-D}^{D} x_{i,j+d}.u_d$$ avec u_d les poids définis par le lissage.

Les propriétés du lissage par la méthode de la moyenne mobile ont été étudiées par Anderson[43] et Rabiner et Gold[44]. Dans notre étude, le lissage est appliqué sur les spectres dérivés. La dérivée et le lissage sont codés par trois nombres : (X,Y,Z). Le premier correspond à l'ordre de la dérivée, le second à l'intervalle de la dérivée et le troisième à l'intervalle sur lequel est effectué le lissage par la méthode de la moyenne mobile.

4.1.3 Méthode « De-trending »

Vers les longueurs d'onde grandes, les valeurs de l'absorbance ont tendance à augmenter à cause des effets de diffusion. Quand les cellules d'analyses sont remplies avec des pressions différentes, cette tendance devient non linéaire. La méthode « De-trending »[45] est une méthode simple de calcul qui enlève la courbure du second degré de la ligne de base. Un polynôme du second degré est ajusté au spectre. Ensuite, le polynôme calculé est

soustrait du spectre pour donner le spectre corrigé. La méthode « De-trending » est utilisée le plus souvent après une normalisation de type « Standard Normal Variate ».

Équation 31 $x_{i,\text{de-trending}} = x_i - d_i$ avec $x_{i,\text{de-trending}}$ la valeur corrigée et d_i la valeur du polynôme à la longueur d'onde i.

4.1.4 Normalisation des spectres

- ***Normalisation « Multiplicative Scatter Correction »***

La correction de diffusion MSC (« Multiplicative Scatter Correction ») améliore la linéarité de la relation existant entre l'absorbance et la concentration[46]. Cette méthode est intéressante lorsque des techniques de régressions linéaires sont utilisées. Pour effectuer la correction MSC, il est nécessaire d'avoir un spectre de référence. Le spectre moyen des spectres du lot d'étalonnage est utilisé par défaut[47].

Un modèle linéaire est mis en place entre le spectre et le spectre moyen selon :

Équation 32 $x_i = a + b\,\bar{x}_i + e_i$ (avec a et b les coefficients du modèle et e_i l'erreur à la longueur d'onde i). La valeur corrigée est ensuite calculée comme suit :

Équation 33 $$x_{MSC,i} = (x_i - a) \,/\, b$$

- ***Normalisation « Standard Normal Variate »***

Le principal avantage de la méthode SNV[45] est qu'elle s'applique à chaque spectre pris séparément sans référence à l'ensemble des échantillons d'étalonnage[48].

Les données spectrales sont centrées et réduites selon l'équation suivante :

Équation 34 $$x_{SNV,i} = (x_i - \bar{x}) \,/\, \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{(w-1)}}$$

avec x_i la valeur du log (1/R) à la longueur d'onde i, w le nombre de longueurs d'onde, $\bar{x}$ est la valeur moyenne du log (1/R) et $x_{SNV,i}$ la valeur corrigée à la longueur d'onde i.

De nombreuses études ont montré que l'utilisation des normalisations MSC et SNV permettait d'améliorer les résultats de l'analyse quantitative[49,50,45].

4.2 Exploration des données

Avant de commencer une étude quantitative, il faut observer les données spectrales pour apprécier la structure des données et détecter la présence d'un spectre aberrant. Quand le nombre de données devient important, un examen direct est difficile. L'analyse multivariée permet de représenter ces données. Dans notre étude, l'analyse en composante principale a été utilisée.

4.2.1 Principe de l'analyse en composantes principales

L'analyse en composantes principales (ACP) permet de déterminer les caractéristiques principales des spectres, de les comparer entre eux et de mettre en évidence des liens entre les variables descriptives (les absorbances aux différentes longueurs d'onde)[51].

L'ACP projette le nuage de points dans un espace de représentation de faibles dimensions[52]. Elle calcule de nouvelles variables, appelées composantes principales qui sont des combinaisons linéaires des absorbances de départ. Puisque l'objectif de l'analyse est la simplification, il faut choisir la dimension de l'espace de représentation en effectuant un compromis entre deux objectifs contradictoires : prendre un espace de faibles dimensions et conserver une variance expliquée maximale.

Les différentes étapes de l'analyse en composantes principales sont les suivantes :

- Le centre de gravité du nuage de points est calculé. Il s'agit du point ayant pour coordonnées les moyennes des coordonnées des individus. Les données sont centrées, ce qui correspond à une translation du repère. Ainsi, l'origine du nouveau repère est le centre de gravité du nuage initial.

- Le premier axe principal est la droite passant par l'origine qui restitue le maximum d'inertie, c'est-à-dire le maximum de variance. Le deuxième axe est orthogonal au premier, passant par l'origine qui approxime le mieux les données, c'est-à-dire, qui exprime le mieux la variance résiduelle. Les composantes suivantes sont déterminées de la même façon.

- Ensuite, les coordonnées des individus dans le nouvel espace sont calculées.

4.2.2 Approche mathématique

- ***Centrage de la matrice initiale***

Le centrage le plus couramment utilisé consiste à soustraire la moyenne des variables selon l'équation :

Équation 35 $x_{corrigé\ i,j} = x_{i,j} - x_{.,i}$ avec $x_{corrigé\ i,j}$ l'absorbance transformée pour l'échantillon j à la longueur d'onde i et $x_{.,i}$ l'absorbance moyenne à la longueur d'onde i.

Dans le cas où les variables sont hétérogènes (unités différentes), les données sont centrées et réduites. Ainsi, toutes les variables ont le même poids dans le calcul des composantes principales. Par la suite, **Xt** représentera la matrice des absorbances corrigées (m échantillons et w longueurs d'onde).

- ***Recherche des composantes principales***

Le premier axe est la droite pour laquelle les carrés des écarts à la droite sont minima. Le critère des moindres carrés conduit à maximiser les valeurs des projections orthogonales des individus sur cette droite.

La coordonnée de projection d_i d'un vecteur $\mathbf{x}_i$ sur un axe est le produit scalaire de ce vecteur avec le vecteur unitaire $\mathbf{p}_i$ de cet axe : $d_i = \mathbf{x}_i.\mathbf{p}_i$ (Équation 36).

Pour l'ensemble des individus la relation précédente s'écrit :

$\mathbf{d} = \mathbf{Xt}.\mathbf{p}_i$ avec **d** : le vecteur des projections des m individus sur l'axe 1.

Soit s la somme des carrés des projections, $s = \mathbf{d}'.\mathbf{d} = \mathbf{p}_i'.\mathbf{Xt}'.\mathbf{Xt}.\mathbf{p}_i$

La première composante principale est telle que son vecteur unitaire $\mathbf{u}_1$ vérifie :

- $\mathbf{p}_1'.\mathbf{Xt}'.\mathbf{Xt}.\ \mathbf{p}_1$ est maximum
- $\mathbf{p}_1$ est un vecteur unitaire : $\mathbf{p}_1'.\ \mathbf{p}_1 = 1$

De même, le second axe est tel que son vecteur unitaire $\mathbf{p}_2$ vérifie les trois conditions suivantes :

- $\mathbf{p}_2'.\mathbf{Xt}'.\mathbf{Xt}.\ \mathbf{p}_2$ est maximum
- $\mathbf{p}_2$ est un vecteur unitaire : $\mathbf{p}_2'.\ \mathbf{p}_2 = 1$
- $\mathbf{p}_2$ est orthogonal à $\mathbf{p}_1$: $\mathbf{p}_2'.\ \mathbf{p}_1 = 0$

Les axes suivants sont définis de la même façon.

On montre que les vecteurs unitaires solutions du problème sont les vecteurs propres de la matrice de variance covariance initiale **V** (**V** = **Xt**'.**Xt**). Le calcul est réalisé par la diagonalisation de la matrice **V**. La diagonalisation de la matrice **V** donne deux types de résultats : la matrice des vecteurs propres (appelé « loading ») **P**(w,a) avec a nombre de composantes sélectionnées et la matrice diagonale des valeurs propres **L**(a,a). A chaque vecteur propre **p** est associé une valeur propre λ qui est la variance des individus sur l'axe correspondant.

- ***Calcul des coordonnées des individus***

Les coordonnées sont calculées en projetant les individus sur les nouveaux axes sélectionnés :

Équation 37 **T** = **Xt**.**P** avec **T**(m,a) : matrice des coordonnées factorielles (appelé également scores).

Il est possible de projeter dans le même espace, des individus n'ayant pas participé à la création des axes en utilisant l'équation précédente. Ces individus sont alors appelés individus supplémentaires.

- ***Calcul des coordonnées des variables***

Les coordonnées des variables sont données à un facteur près par les vecteurs propres **P**. Le facteur de proportionnalité est égal à la racine carrée de la valeur propre correspondante[53].

Équation 38 $\mathbf{C} = \mathbf{P}.\mathbf{L}^{1/2}$ avec **C** la matrice des coordonnées des variables.

L'algorithme NIPALS[33] (« Non linear Iterative Partial least Square ») est utilisé pour la détermination des vecteurs propres. Il ne calcule pas directement tous les vecteurs propres. Il procède de façon itérative : il calcule $\mathbf{t}_1$ et $\mathbf{p}_1$ à partir de **X**t. Ensuite le produit $\mathbf{t}_1.\mathbf{p}'_1$ est soustrait de la matrice **X**t et le résidu est utilisé pour calculer $\mathbf{t}_2$ et $\mathbf{p}_2$. Il a été montré que la méthode NIPALS donnait les mêmes solutions que les formules classiques de calculs des vecteurs propres et des valeurs propres.

La méthode NIPALS procède comme suit[54] :

1. initialise $\hat{\mathbf{t}}_a$: $\hat{\mathbf{t}}_a$ = le vecteur colonne de $\mathbf{X}_{a-1}$ (Pour a = 1, $\mathbf{X}_{a-1} = \mathbf{X}_0 = \mathbf{X}_t$)
2. calcule $\hat{\mathbf{p}}_a$ ' : $\hat{\mathbf{p}}_a$ ' = ($\hat{\mathbf{t}}_a$ '. $\hat{\mathbf{t}}_a$)$^{-1}$. $\hat{\mathbf{t}}_a$ '.$\mathbf{X}_{a-1}$
3. normalise $\hat{\mathbf{p}}_a$ ' à la longueur 1 : $\hat{\mathbf{p}}_a$ = $\hat{\mathbf{p}}_a$ ($\hat{\mathbf{p}}_a$ '. $\hat{\mathbf{p}}_a$)$^{-0.5}$
4. calcule $\hat{\mathbf{t}}_a$: $\hat{\mathbf{t}}_a$ = $\mathbf{X}_{a-1}$. $\hat{\mathbf{p}}_a$.($\hat{\mathbf{p}}_a$ '. $\hat{\mathbf{p}}_a$)$^{-1}$
5. estime la valeur propre τ_a = $\hat{\mathbf{t}}_a$ '. $\hat{\mathbf{t}}_a$
6. vérifie la convergence : si la différence entre la valeur propre τ_a et celle calulée à la précedente itération est plus petite qu'une constante fixée la méthode a convergée. Sinon, l'algorithme reprend à l'étape 2.
7. Calcule le facteur suivant avec $\mathbf{X}_a = \mathbf{X}_{a-1} - \hat{\mathbf{t}}_a . \hat{\mathbf{p}}_a$ '

L'ACP permet ainsi une décomposition spectrale. Les vecteurs propres **P**, appelés profils spectraux, ont la dimension d'un spectre et peuvent être observés en tant que tels. Ils permettent d'identifier les longueurs d'onde responsables de la plus grande variabilité. L'ACP permet également une condensation des données en diminuant le nombre de variables. En effet dans l'espace initial, les échantillons sont décrits par w variables et dans l'espace de ACP, l'utilisateur choisi un nombre de composantes principales qui est largement inférieur à w.

4.3 Méthodes de régression

Les méthodes de régression utilisées pour construire une équation d'étalonnage à partir de données de spectroscopie PIR sont nombreuses. L'objectif est de présenter les différentes familles de méthodes de régressions linéaires : régressions linéaires simples et les méthodes multivariées linéaires.

4.3.1 Régression linéaire simple et multiple

- ***Régression linéaire simple***

Le modèle mathématique le plus simple pour relier une absorbance à la longueur i à une concentration est du type :

Équation 39 $y_j = b_{0,i} + b_{1,i}.x_i + e_{i,j}$ avec $e_{i,j}$ le résidu pour la longueur i et l'échantillon j, $b_{0,i}$, $b_{1,i}$ l'intersection à l'origine et la pente de la régression pour la longueur i.

L'inconvénient de ce type de modèles réside dans le choix de la longueur d'onde i qui doit être la plus prédictive possible de la concentration et le plus souvent, une longueur d'onde n'est pas suffisante pour évaluer une concentration.

Cependant, l'équation de régression linéaire simple ne permet pas de tenir compte de l'interaction avec d'autres constituants chimiques susceptibles d'interférer à la longueur d'onde i. C'est pourquoi des méthodes de régressions multilinéaires ont été introduites.

- ***Régression linéaire multiple (MLR « MultiLinear Regression »)***

La méthode MLR[55,56,57] est également connue sous le nom de « Inverse Least Square ». L'équation de prédiction obtenue est de la forme :

Équation 40 $$y_j = b_0 + \sum_{i=1}^{k} b_i.x_{i,j} + e_{i,j}$$

avec k le nombre de variables prédictives (nombre de longueurs d'onde utilisées – k est inférieur ou égale à w) et $e_{i,j}$ le résidu (terme d'erreur aléatoire). La méthode des moindres carrés consiste à minimiser la somme des carrés des écarts $e_{i,j}$.

Dans la suite des calculs, la matrice **X** contient une colonne de 1 pour tenir compte de l'intersection à l'origine. En notation matricielle, le système d'équation devient : **Y** = **X**.**b** avec **b** le vecteurs des coefficients. Lors de l'étape d'étalonnage, la matrice $\hat{\mathbf{b}}$ est calculée :

Équation 41 $\mathbf{Y}_{étalonnage} = \mathbf{X}_{étalonnage} . \hat{\mathbf{b}}$

Soit $\mathbf{X}'_{étalonnage}.\mathbf{Y}_{étalonnage}. = \mathbf{X}'_{étalonnage}.\mathbf{X}_{étalonnage}. \hat{\mathbf{b}}$

Équation 42 $\hat{\mathbf{b}} = (\mathbf{X}'_{étalonnage}.\mathbf{X}_{étalonnage})^{-1}. \mathbf{X}'_{étalonnage}.\mathbf{Y}_{étalonnage}$

Lorsque l'étalonnage est terminé, la matrice **B** peut être utilisée pour déterminer les concentrations des échantillons inconnus : $\hat{\mathbf{Y}}_{validation} = \mathbf{X}_{validation} . \hat{\mathbf{b}}$

L'inconvénient majeur de cette méthode est la sélection des longueurs d'onde caractéristiques des éléments à doser. En effet, le problème majeur des spectres PIR est que les absorbances des longueurs d'onde proches sont fortement corrélées. La matrice $\mathbf{X}'_{étalonnage}.\mathbf{X}_{étalonnage}$ est quasi-singulière, c'est-à-dire que l'on peut écrire une ligne ou une colonne de cette matrice comme étant une combinaison linéaire des autres. Dans ces conditions, il n'est pas toujours possible d'inverser cette matrice et il existe alors une infinité de solutions équivalentes.

Il faut trouver un compromis entre sélectionner suffisamment de longueurs d'onde pour obtenir une prédiction satisfaisante mais pas trop pour ne pas sélectionner des variables colinéaires. Les longueurs d'onde peuvent être choisies en fonction de leur pouvoir prédictif par régression multilinéaire pas à pas ou grâce à des algorithmes de sélection. Les trois méthodes les plus couramment utilisées pour la sélection des longueurs d'onde sont les suivantes[58] :

- Lors de la sélection ascendante, les variables sont introduites dans le modèle une à une. Cette procédure procède de manière itérative en incluant à chaque étape la variable qui améliore le plus le modèle de prédiction. La première variable x_i utilisée pour la régression est celle qui est le plus fortement corrélée à la concentration y. La seconde variable est celle qui est le plus corrélée à y corrigée de l'effet de la première variable (résidu non modélisé). Les variables x_i sont introduites jusqu'à ce que le coefficient de régression de la dernière variable ne soit plus significatif d'après le test de Fisher ou que toutes les variables soient utilisées[59] (si $\mathbf{X}'.\mathbf{X}$ inversible).

- De manière analogue, il est possible d'envisager une procédure de régression descendante. Lors de la sélection descendante, le premier modèle construit utilise l'ensemble des variables x_i. Ensuite, on procède par itération et les variables les moins corrélées à la valeur y sont retirées une à une. Le processus de sélection s'arrête lorsque toutes les variables ont un coefficient de corrélation significatif ou lorsqu'il ne reste plus qu'une variable prédictive.

- Il est aussi possible de combiner les deux démarches (ascendante et descendante). L'intérêt est de tirer le meilleur parti des deux procédures, en envisageant à une étape donnée de la sélection ascendante, la possibilité de retirer une variable incluse au préalable. On parle alors de la méthode pas à pas ascendante « stepwise ». Ainsi, à chaque nouvelle introduction d'une variable, les variables introduites précédemment sont réévaluées en fonction du test de Fisher.

- Enfin toutes les combinaisons possibles peuvent être testées. D'après Martens et Naes[33], il s'agit de la meilleure méthode de sélection. Pour w longueurs d'onde, le nombre de modèle est $\sum_{i=1}^{w} C_w^i = \sum_{i=1}^{w} \frac{w!}{i!(w-i)!}$ avec C_w^i le nombre de combinaisons de i élements pris parmi les w longueurs d'onde.

4.3.2 Méthodes factorielles

- ***Régression en composantes principales (PCR)***

La régression en composantes principales (PCR) est constituée de deux étapes[60]. Tout d'abord les données spectrales sont traitées par ACP. Ensuite une régression MLR est appliquée aux données issues de l'ACP, avec les coordonnées factorielles comme variables prédictives.

Comme les données spectrales sont centrées, la constante de régression est nulle. L'équation de prédiction s'écrit de façon matricielle par : $\mathbf{Y}_{étalonnage} = \mathbf{T}_{étalonnage}.\mathbf{b}$ avec **T** les nouvelles coordonnées de dimensions (m,a) avec a le nombre de composantes principales sélectionnées et **b** le vecteur des coefficients.

Comme pour la méthode MLR, le vecteur **b** peut être calculé à partir des données du lot d'étalonnage (Équation 43).

Équation 43 $\hat{\mathbf{b}} = (\mathbf{T}'.\mathbf{T})^{-1}.\mathbf{T}'.\mathbf{Y}$

Ensuite les coefficients sont utilisés sur les données du lot de validation. La matrice spectrale $X_{validation}$ est projetée dans l'espace à a dimensions de l'ACP puis les concentrations sont calculées [61]:

Équation 44 $\mathbf{T}_{validation} = \mathbf{X}_{validation}.\mathbf{P}_{étalonnage}$

Équation 45 $\hat{\mathbf{Y}}_{validation} = \mathbf{T}_{validation}.\ \hat{\mathbf{b}}$

Les avantages de la méthode PCR sont nombreux :

- L'analyse en composantes principales supprime les colinéarités spectrales.
- Elle ne demande pas de sélection de longueurs d'onde *a priori*.

L'inconvénient majeur est qu'il n'y a pas de garantie que les composantes principales soient corrélées aux concentrations à prédire.

- ***Régression des moindres carrés partiels (PLS)***

L'algorithme Partial Least Squares (PLS) a été développé pour résoudre des problèmes de sciences économiques[33]. Ses premières applications à l'analyse quantitative remonte aux années 1980[62]. Comme la méthode PCR, la régression PLS réduit la matrice initiale. Le choix des nouveaux axes ne se fait plus en fonction de la variance maximale comme dans l'ACP mais selon les directions les plus pertinentes en termes de prédiction des concentrations[54], c'est-à-dire en fonction des covariances maximales spectres-concentrations.

Comme lors de l'analyse en composantes principales, la matrice des données spectrales est décomposée :

Équation 46 $\mathbf{X}(m,w) = \mathbf{T}(m,a).\mathbf{P}'(a,w)$.

De même la matrice des concentrations est décomposée :

Équation 47 $\mathbf{Y}(m,p) = \mathbf{U}(m,a).\mathbf{Q}'(a,p)$.

De plus, une relation lie les deux matrices **X** et **Y** :

Équation 48 $\mathbf{u_a} = b_a.\mathbf{t}_a$ avec b_a jouant le rôle de coefficient de régression.

Lorsque plusieurs constituants chimiques doivent être évalués, on peut effectuer une régression sur chacun des constituants pris séparément (méthode PLS1) ou sur toutes les concentrations en même temps (algorithme PLS2). Seule la méthode PLS1 sera présentée.

L'algorithme d'étalonnage (PLS1) proposé par Wold[63] est le suivant :

Etape 1 : La matrice spectrale **X** et le vecteur des concentrations **y** sont centrées pour obtenir les variables $\mathbf{X}_0$ et $\mathbf{y}_o$.

Pour chaque facteur a = 1, 2, …, A_{max}, les étapes suivantes (2.1 à 2.6) sont effectuées pour réaliser l'étalonnage.

Etape 2.1 : Déterminer le vecteur $\mathbf{w}_a$ respectant les conditions suivantes :

- Les $\mathbf{w}_a$ sont deux à deux orthogonaux. $\mathbf{w}_{a-1}'.\mathbf{w}_a = 0$
- $\mathbf{w}_a'.\mathbf{w}_a = 1$ ($\mathbf{w}_a$ est un vecteur unitaire)
- $\mathbf{X}_{a-1} = \mathbf{y}_{a-1}.\ \hat{\mathbf{W}}_a'$

d'où $\hat{\mathbf{W}}_a = c.\mathbf{X}_{a-1}'.\mathbf{y}_{a-1}$ avec c un facteur d'échelle permettant d'avoir $\mathbf{w}_a$ unitaire

$$c = (\mathbf{y}_{a-1}'.\mathbf{X}_{a-1}.\mathbf{X}_{a-1}'.\mathbf{y}_{a-1})^{-1/2}$$

Etape 2.2 : Construire les coordonnées factorielles $\hat{\mathbf{t}}_a$ par projection de $\mathbf{X}_{a-1}$ sur $\mathbf{w}_a$.

Il suffit de résoudre : $\mathbf{X}_{a-1} = \mathbf{t}_a . \hat{\mathbf{w}}_a'$

Soit $\hat{\mathbf{t}}_a = \mathbf{X}_{a-1} . \hat{\mathbf{w}}_a . (\hat{\mathbf{w}}_a' . \hat{\mathbf{w}}_a)^{-1} = \mathbf{X}_{a-1} . \hat{\mathbf{w}}_a$

Etape 2.3 : Effectuer une régression de $\mathbf{X}$ sur $\mathbf{t}_a$ pour obtenir le vecteur propre spectral $\mathbf{p}_a$

L'équation $\mathbf{X}_{a-1} = \hat{\mathbf{t}}_a . \mathbf{p}_a'$ a pour solution $\hat{\mathbf{p}}_a = \mathbf{X}_{a-1}' . \hat{\mathbf{t}}_a . (\hat{\mathbf{t}}_a' . \hat{\mathbf{t}}_a)^{-1}$.

Etape 2.4 : Calculer les composantes q_a en déterminant les solutions de

$$\mathbf{y}_{a-1} = \hat{\mathbf{t}}_a . q_a \text{ qui s'écrivent } \hat{q}_a = \mathbf{y}_{a-1}' . \hat{\mathbf{t}}_a . (\hat{\mathbf{t}}_a' . \hat{\mathbf{t}}_a)^{-1}$$

Etape 2.5 : Soustraire la contribution de la composante $\mathbf{p}_a$ de la matrice spectrale $\mathbf{X}$ pour obtenir une nouvelle matrice $\mathbf{X}_a$

$$\mathbf{X}_a = \mathbf{X}_{a-1} - \hat{\mathbf{t}}_a . \hat{\mathbf{p}}_a'$$

Etape 2.6 : Soustraire la contribution de la composante q_a de la matrice des concentrations y pour obtenir une nouvelle matrice $\mathbf{y}_{résiduelle}$

$$\mathbf{y}_a = \mathbf{y}_{a-1} - \hat{\mathbf{t}}_a . \hat{q}_a$$

Les étapes de 2.1 à 2.6 sont effectuées de façon récursive pour l'ensemble des $\mathbf{w}_a$.

Etape 3 : Déterminer le nombre de facteurs à conserver pour étalonner le modèle et créer la matrice $\mathbf{W}$ contenant les vecteurs $\mathbf{w}_a$.

Etape 4 : Calculer les coefficients pour les A termes conservés

$\mathbf{y} = \mathbf{X}.\mathbf{b}$ avec $\hat{\mathbf{b}} = \hat{\mathbf{W}} . (\hat{\mathbf{P}}' . \hat{\mathbf{W}})^{-1} . \hat{\mathbf{q}}$

L'équation d'étalonnage peut être utilisée sur des échantillons inconnus du modèle selon l'équation suivante :

$$\mathbf{y}_{validation} = \mathbf{X}_{validation} . \mathbf{b}$$

La régression PLS a les mêmes avantages que la méthode PCR. Elle permet de prendre en compte un grand nombre de variables observées sur un petit nombre d'échantillons[64]. Elle permet aussi d'interpréter les relations entre les variables prédictives et les concentrations à prédire. L'avantage de la PLS par rapport à la PCR réside dans l'extraction des facteurs qui s'effectue à partir de la matrice de la covariance spectre / concentration. Le seul point critique est le choix du nombre de facteurs.

- ***Choix du nombre de termes***

Le plus souvent lors de la régression PLS et PCR, le nombre de termes utilisés est déterminé par une validation croisée (« cross validation » en anglais). Le principe de base de la validation croisée est le suivant : tous les échantillons du lot d'étalonnage servent à la fois à l'élaboration et à la validation du modèle. On distingue la validation croisée totale (si les échantillons sont retirés un à un) ou partielle (si les échantillons sont retirés par groupe). Chaque échantillon ou groupe d'échantillons est écarté à tour de rôle et un modèle est élaboré avec les individus restants[65]. On calcule la somme des écarts au carré (PRESS) pour des modèles construits avec un nombre de termes variables de 1 à un nombre maximal choisi par l'utilisateur (Figure 11).

Le nombre de termes donnant le plus petit PRESS est identifié. On le note K_{min}. Ensuite, tous les modèles construits avec un nombre de termes inférieurs à K_{min} sont comparés avec le modèle construit avec K_{min} termes. Le modèle utilisant le plus petit nombre de termes donnant un PRESS non significativement différent de celui obtenu avec K_{min} termes est choisi. Haaland et Thomas[66] conseillent de comparer les PRESS à l'aide d'un test de Fisher selon l'Équation 49.

Équation 49 $$F = \frac{\text{PRESS(modèle k termes)}}{\text{PRESS(modèle k}_{min}\text{ termes)}}$$

Le F calculé est comparé au F des tables de Fisher avec un risque α = 5%, m degrés de liberté (m nombre d'échantillons dans le lot d'étalonnage).

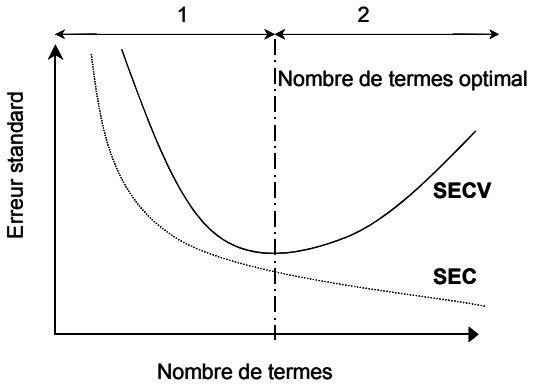


Figure 11 ■ Influence du nombre de terme de la régression sur les erreurs standard de validation croisée (SECV) et d'étalonnage (SEC)[33].

Le choix de la méthode de régression est fonction du nombre d'échantillons du lot d'étalonnage et de la complexité des modèles entre la variable à prédire et les données prédictives. Pour avoir un modèle stable, le SECV doit tendre vers le SEC. De même, la méthode MLR construit des modèles plus simples à interpréter que la PLS et la PCR. Enfin les trois méthodes MLR, PCR et PLS sont des méthodes linéaires. Si des relations non linéaires doivent être modélisées, d'autres méthodes de régression doivent être utilisées : par exemple des régressions PLS ou PCR non linéaires[67] et enfin des réseaux de neurones multicouches[68].

4.4 Méthodes de classification

Il existe deux grands types de classification[69]. Le premier est la classification dite non supervisée. On dispose d'un certain nombre de variables et il faut regrouper les échantillons similaires pour former des agrégats (« cluster » en anglais) puis les identifier. Les méthodes les plus couramment utilisées pour la classification non supervisée sont l'ACP et les méthodes de classification hiérachiques[69]. Le second type est la classification supervisée. Dans ce cas, on dispose d'un lot d'entraînement dont l'appartenance à chaque groupe est connue. Un modèle de classification est ainsi créé en utilisant cette information *a priori*. Il est ensuite validé sur de nouveaux échantillons. Dans notre étude, seules les méthodes de classification supervisées ont donné des résultats exploitables. C'est pourquoi seules ces méthodes sont décrites dans ce paragraphe.

Les méthodes de classification peuvent être regroupées suivant différentes caractéristiques. Ainsi on distingue les méthodes linéaires et les méthodes non linéaires[3]. On peut ensuite séparer les méthodes paramétriques et les méthodes non paramétriques. Les méthodes paramétriques supposent que les échantillons sont normalement distribués. Enfin, certaines méthodes recherchent les différences entre les échantillons alors que d'autres identifient les similarités au sein d'une même classe.

Dans notre étude, huit algorithmes[69] ont été utilisés :

- la méthode des K plus proches voisins[70,71] (« K-Nearest Neighbors » - KNN),
- l'analyse discriminante linéaire[72] (« Linear Discriminant Analysis » - LDA),
- la modélisation indépendante des analogies de classe (« Soft Independent Modelling of Class Analogy »[73] - SIMCA),
- la méthode discriminante des moindres carrés partiels[74] (« Discriminant Partial Least Squares » - DPLS),
- l'analyse procrustéenne discriminante[75] (« Procrustes Discriminant Analysis » - PDA),
- les arbres de décision[76] (« Classification And Regression Tree » - CART),
- et enfin les réseaux de neurones[77,78] utilisant des cartes de Kohonen (« Learning Vector Quantization neural network » - LVQ).
- les réseaux de neurones probabilistiques[79] (« Probabilistic Neural Network » - PNN)

Pour les méthodes KNN, LDA et SIMCA, les classes sont codées par un nombre entier variant de 1 au nombre total de classe. Pour les autres méthodes (PDA, DPLS, LVQ et PNN), une matrice de classe est utilisée : cette matrice contient autant de colonnes que de classes et autant de lignes que d'échantillons. L'appartenance à la classe J est codée par 1 dans la J^{ème} colonne et par des 0 dans les autres colonnes.

Remarque : une présentation exhaustive des méthodes, dont certaines sont relativement classiques mais d'autres plus originales, a été choisie avec le risque d'une lecture rapidement fastidieuse. Afin de diminuer ce risque, chaque paragraphe a été réduit au maximum néanmoins ils sont appuyés de références bibliographiques afin de compléter l'information.

4.4.1 KNN

La méthode KNN est non paramétrique[70,71]. Un échantillon inconnu du lot de validation est classé selon la classe des K échantillons du lot d'entraînement qui lui ressemblent le plus[80].

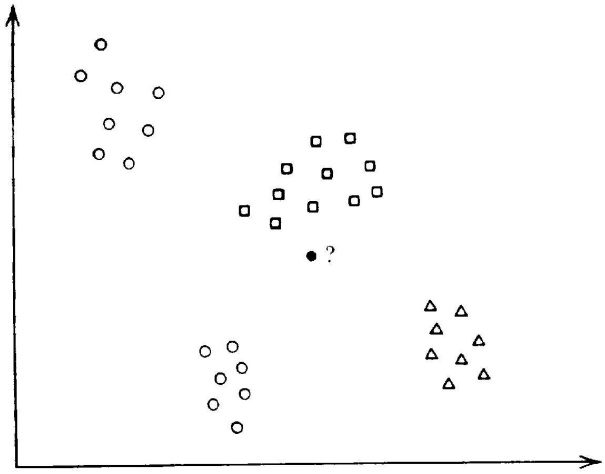


Figure 12 ■ Méthode des K plus proches voisins (d'après Massart *et al.*[81]) Le point noir inconnu est plus proche des échantillons représentés par un carré. Il est donc affecté à cette classe.

KNN est simple à mettre en place. Pour chaque échantillon du lot de validation, on calcule la distance euclidienne avec tous les échantillons du lot d'étalonnage. La distance Euclidienne ($d_{k,j}$) entre deux échantillons k et j est :

Équation 50 $$d_{k,j} = \sqrt{\sum_{i=1}^{w} (x_{j,i} - x_{k,i})^2}$$

avec w le nombre de variables (nombre de longueurs d'onde), $x_{k,i}$ la valeur de l'absorbance à la longueur d'onde i pour l'échantillon k.

Les K plus proches voisins d'un échantillon inconnu sont donc les échantillons du lot d'étalonnage pour lesquelles les distances euclidiennes sont les plus petites. La classe prédite est la classe qui est la plus représentée dans les K plus proches voisins (Figure 12).

Le nombre optimal K est déterminé par une procédure de validation croisée[82]. Chaque objet du lot d'étalonnage est sorti et considéré comme un échantillon de validation. Cette procédure est réalisée pour K = 1 à n-1 et le nombre K pour lequel le taux de prédiction est le plus important, est choisi.

4.4.2 Méthodes factorielles

- ***LDA***

« Linear Discriminant Analysis » (LDA) est une méthode linéaire de discrimination, paramétrique[70,81]. LDA, comme l'analyse en composantes principales (ACP), est une méthode qui factorise les données. Cependant, alors que l'ACP sélectionne les directions selon lesquelles la variance est la plus importante, LDA sectionne les directions qui permettent la meilleure séparation entre les classes[69]. Ainsi, LDA recherche les frontières optimales entre les classes (Figure 13). Nous utilisons la méthode LDA avec des distances euclidiennes.

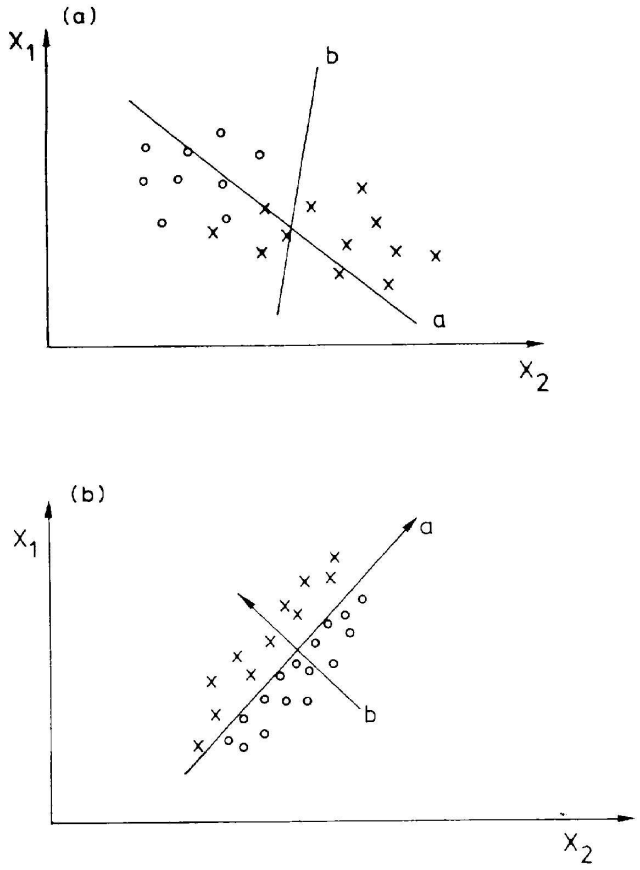


Figure 13 ■ Fonctions discriminantes (d'après Massart *et al.*[81]) (a) La droite a est une meilleure fonction discriminante pour séparer les échantillons x et o. (b) Cas où la première composante principale a et la fonction discriminante b sont différentes.

Des fonctions linéaires sont calculées pour chaque classe C :

Équation 51 $D_c = \sum_{i=1}^{k} v_i . \bar{x}_{c,i}$ avec v_j les poids donnés aux variables, $\bar{x}_{c,i}$ la moyenne de la variable j pour la classe C et m le nombre de variables.

Pour un échantillon inconnu u ayant les variables $x_{u,j}$, la même fonction est obtenue :

Équation 52 $D_u = \sum_{i=1}^{k} v_i . \bar{x}_{u,i}$

u est affecté à la classe C pour laquelle D_c est la plus proche de D_u. La seule condition pour utiliser LDA est que le nombre de variables k doit être inférieur au nombre d'échantillons m. Quand on utilise les données de SPIR, le nombre de longueurs d'onde est souvent supérieur au nombre d'échantillons. L'ACP est donc utilisée comme étape préliminaire pour réduire le nombre de variables.

- ***SIMCA***

SIMCA[73] utilise les propriétés de l'analyse en composantes principales (Figure 14). Il s'agit d'une méthode paramétrique qui considère chaque classe séparément. Pour chaque classe C, une décomposition en composantes principales est effectuée, ce qui conduit aux modèles suivants [81]:

Équation 53 $\mathbf{X}_C = \mathbf{T}_C . \mathbf{P}_C^{'} + \mathbf{E}_C$

avec $\mathbf{X}_C$ la matrice centrée des $\mathbf{X}$ pour la classe C, $\mathbf{E}_C$ l'erreur résiduelle, $\mathbf{P}_C$ et $\mathbf{T}_C$ la matrice des vecteurs propres et des coordonnées factorielles de la classe C.

Ainsi on obtient autant de modèles que de classes. On pose r le nombre de composantes principales retenues. Dans notre étude, les composantes principales représentant plus de 1 % de la variance totale sont conservées. Les limites de classe, c'est-à-dire l'intervalle de confiance, sont construites autour du modèle. Les résidus du lot d'entraînement par rapport à un modèle en composantes principales suivent une loi normale avec un écart type s :

Équation 54 $s = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{w} \frac{e_{i,j}^{\ 2}}{(w-r)(m-r-1)}}$

s est calculé à partir des coordonnées factorielles (scores) sur les vecteurs propres non conservés :

Équation 55 $$s = \sqrt{\sum_{i=1}^{m}\sum_{j=r+1}^{w}\frac{t_{i,j}^{\ 2}}{(w-r)(m-r-1)}}$$

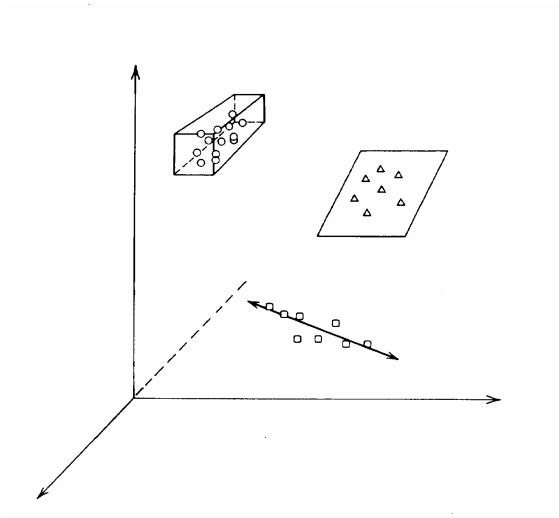


Figure 14 ■ Principe de la méthode SIMCA (d'après M. Sharaf *et al.*[69]). Des analyses en composantes principales et des enveloppes de confiance (hypervolumes) sont ajustées pour chaque classe : une droite quand une composante principale est utilisée, un plan pour un modèle à deux composantes et un hypercube pour celui à trois composantes.

s est la distance moyenne entre les objets de la classe K et son modèle. En utilisant un test de Fisher, on calcule une valeur critique de s :

Équation 56 $s_{critique} = s.\sqrt{F_{critique}}$ avec $F_{critique}$ le nombre de Fisher pour w et m-r-1 degrés de liberté et un risque α à 5 %.

Le lot de validation est projeté sur chacun des modèles. Pour un échantillon inconnu, le résidu $s_{nouveau}$ est calculé :

Équation 57 $$s_{nouveau} = \sqrt{\sum_{j=r+1}^{w}\frac{t_{i,j}^{\ 2}}{(w-r)}}$$

Si $s_{nouveau}$ est inférieur à $s_{critique}$ alors l'objet du lot de validation appartient à la classe K.

- ***DPLS***

DPLS[74] est une méthode paramétrique linéaire. La régression PLS est utilisée pour classer les échantillons[64]. On peut remarquer que les études utilisant cette méthode sont peu nombreuses[82].

La matrice des coefficients $\hat{\mathbf{B}}$ est calculée à partir du lot d'étalonnage :

Équation 58 $\mathbf{Y}_{étalonnage} = \mathbf{X}_{étalonnage}.\ \hat{\mathbf{B}}$

Y possède autant de colonnes qu'il y a de classes. L'algorithme PLS donne la matrice des vecteurs propres **T** et des coordonnées factorielles **P** tel que $\mathbf{T} = \mathbf{X}_{étalonnage}.\mathbf{P}$ [83]. L'algorithme PLS2 est utilisé.

Équation 59 $\hat{\mathbf{B}} = \mathbf{P}\ (\mathbf{T}'.\mathbf{T})^{-1}\ \mathbf{T}'.\mathbf{Y}$

La prédiction d'un nouveau lot est effectuée avec la même formule :

Équation 60 $\mathbf{Y}_{validation} = \mathbf{X}_{validation}.\ \hat{\mathbf{B}}$

- ***PDA***

PDA[75,84] est relativement proche de la méthode DPLS. Les vecteurs propres sont obtenus à partir de la matrice des covariances **Z**'**Z** (avec **Z** = **Y**'.**X**). L'étape fondamentale de cette méthode est la transformation des coordonnées factorielles pour chaque classe en une matrice cible Y par une combinaison de rotation, translation et homothétie. Comme pour DPLS, la matrice originale est décomposée : $\mathbf{T} = \mathbf{X}_{étalonnage}.\mathbf{P}$. La transformation de Procruste permet le passage de la matrice **T** à la matrice des classes **Y** (Équation 61).

Équation 61 $\mathbf{Y} = \mathbf{T}.\ \hat{\mathbf{W}}$ avec $\hat{\mathbf{W}} = (\mathbf{T}'.\mathbf{T})^{-1}.\mathbf{T}'.\mathbf{Y}$

La précision de la méthode est évaluée avec un lot de données indépendant. La matrice **X** de validation est décomposée de la même façon et les classes prédites sont calculées : $\mathbf{Y}_{validation} = \mathbf{T}_{validation}.\ \hat{\mathbf{W}}$

4.4.3 Arbre de décision CART

La méthode CART (« Classification And Regression Trees ») repose sur la construction d'un arbre de décision à choix binaire[85]. L'arbre de décision fournit une représentation hiérarchisée des variables de description des échantillons. Les arbres de décision de type CART sont des méthodes non paramétriques. Le lot d'étalonnage est divisé

de façon récursive en sous-groupes : à chaque nœud h, une seule variable x_h (une longueur d'onde) intervient et une valeur seuil s_h de cette variable est choisie. A ce nœud h, l'ensemble des échantillons est divisé en deux sous-ensembles selon le résultat observé pour la variable x_h (inférieur ou supérieur à s_h). Le critère de séparation est basé sur la « pureté » des sous ensembles formés : les nœuds descendants ainsi créés sont plus homogènes d'un point de vue de l'appartenance à une classe que le noeud parent. La procédure se déroule de façon itérative et le choix de l'arbre final est basé sur le taux d'erreur apparent[86].

4.4.4 Réseaux de neurones

Les réseaux de neurones artificiels sont des méthodes non linéaires et non paramétriques. Un neurone est une unité de calcul qui transforme l'information d'entrée en donnée de sortie grâce à une fonction d'activation. Les réseaux sont constitués de plusieurs couches de neurones : la couche d'entrée, la couche cachée et la couche de sortie. Dans notre étude, deux types de réseaux de neurones adaptés à la classification ont été utilisés : LVQ et PNN.

- ***LVQ***

Les réseaux de neurones de Kohonen[77] appartiennent à la famille des cartes auto-organisatrices. Ils sont généralement utilisés pour des problèmes non supervisées. Cependant, ils peuvent être étendus à des méthodes de classification supervisées.

LVQ[78] (« Learning Vector Quantification » ou quantification des vecteurs d'apprentissage) possède trois couches : une couche d'entrée, une couche compétitive et une couche de sortie linéaire[87]. La couche compétitive permet de classer les données d'entrée de façon non supervisée. Elle accepte des données d'entrée et renvoie en sortie des zéros pour tous les neurones sauf un, appelé neurone gagnant qui reçoit la valeur 1. Quand le réseau est entraîné, tous les spectres appartenant à la même classe ont le même neurone gagnant. La couche linéaire transforme les données issues de la couche compétitive en classe.

- ***PNN***

PNN[79] est un réseau de neurones multicouches sans rétropropagation de l'erreur[88] (Figure 15). La fonction d'activation est la suivante :

Équation 62 $f(\mathbf{x}) = e^{-\frac{\mathbf{x}.\mathbf{v}_i - 1}{\sigma^2}}$ avec x le vecteur d'entrée normalisé, v le vecteur poids, i le numéro de la classe et σ le facteur de lissage.

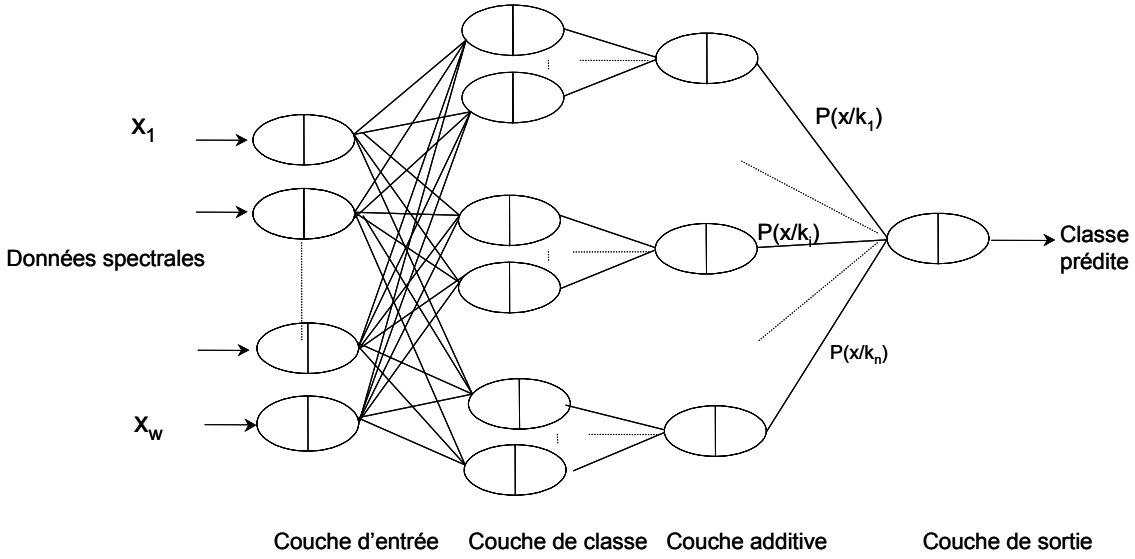


Figure 15 ■ Principe du réseau PNN

PNN est un réseau de neurones qui utilise des règles de décisions bayésiennes. La couche additive est constituée d'un neurone par classe qui somme les données sortant de la couche de classe.

Quand un nouveau spectre est présenté, la couche de classe calcule les distances entre ce nouveau spectre et ceux du lot d'entraînement. Ensuite la couche additive somme la contribution de chaque classe pour produire en sortie la probabilité d'appartenir à une classe donnée. Enfin, la fonction d'activation de la couche de sortie attribut une classe au nouvel échantillon en fonction des probabilités sortant de la couche additive.

4.5 Méthodologie pour la comparaison de modèles

Les méthodes chimiométriques étant nombreuses, il a été nécessaire de développer une méthodologie pour les comparer. Ainsi, les modèles quantitatifs sont comparés sur la base de l'erreur globale (SEP) ou en séparant l'erreur aléatoire de l'erreur

systématique. Quant aux méthodes de classification supervisées, les taux de bonne classification sur des échantillons de validations sont utilisés pour la comparaison de méthodes.

4.5.1 Comparaison des modèles quantitatifs

Pour comparer les modèles deux approches ont été utilisées. La première consiste à comparer leurs erreurs totales. La deuxième a été mise en place pour séparer la comparaison des erreurs systématiques et des erreurs aléatoires.

- Comparaison des erreurs standards de prédiction

Le SEP représente l'erreur totale de la méthode. Un test statistique, décrit par Haaland et Thomas[66], est réalisé pour comparer les SEP obtenus avec deux méthodes. L'hypothèse de départ H_0 est la suivante : les deux modèles ont des SEP qui ne sont pas significativement différents à α = 5 %. Le test de Fisher est calculé comme suit :

Équation 63 $$F_{calculé} = \frac{SEP_1^2}{SEP_2^2} \quad \text{avec } SEP_2 < SEP_1$$

On compare ensuite le $F_{calculé}$ et le $F_{critique}$ lu dans les tables pour n_1-1 et n_2-1 degrés de liberté avec n_1 le nombre d'échantillons pour valider le modèle 1 et n_2 le nombre d'échantillons pour valider le modèle 2. Si $F_{calculé}$ est supérieur au $F_{critique}$, H_0 est rejetée et les deux méthodes sont considérées comme étant différentes.

- Comparaison séparée de l'erreur aléatoire et de l'erreur systématique

- ***Comparaison des biais (erreur systématique)***

L'analyse de la variance (ANOVA) est utilisée pour vérifier si l'un des biais des modèles est significativement différents des autres[89]. L'ANOVA est une méthode univariée à but descriptif et décisionnel. Elle permet d'expliquer une variable quantitative par des effets qualitatifs. On compare la valeur du test de Fisher (F) à la valeur théorique donnée par les tables de Fisher (Tableau 3).

Dans notre étude, l'ANOVA est appliquée à un tableau de données constitué des erreurs de prédiction $y'_{j,k}$ - y_j des différents modèles sur le lot de validation. $y'_{j,k}$ est la

concentration prédite par le modèle k pour l'échantillon j et y_j est la valeur de la méthode chimique de référence pour l'échantillon j.

Tableau 3 ■ Principes de l'analyse de la variance.

n_i le nombre d'échantillons pour valider le modèle k, K le nombre de modèles, $N = \sum n_i$, $e_{j,k}$ l'erreur de prédiction ($y'_{j,k} - y_j$) pour l'échantillon j et le modèle k, $e_{.,.}$ la moyenne générale et $e_{.,k}$ la moyenne pour le modèle k.

Source de variations	Somme des carrés	Degrés de Liberté	Moyenne des carrés	Fisher
Facteurs	$SSF = \sum_i \sum_j (e_{.,k} - e_{.,.})^2$	K-1	MSf=SSF/(K-1)	MSf/MSr
Résidus	$SSR = \sum_i \sum_j (e_{j,k} - e_{.,.})^2 - \sum_i \sum_j (e_{.,k} - e_{.,.})^2$	(N-1)-(K-1)	MSr=SSR/(N-K)	
Total	$SST = \sum_i \sum_j (e_{j,k} - e_{.,.})^2$	(N-1)	MSt=SST/(N-1)	

L'hypothèse nulle de l'ANOVA (H_0) est : les biais de tous les modèles testés ne sont pas différents. Si la valeur calculée du Fisher est plus grande que la valeur du Fisher critique, alors H_0 est rejetée et l'hypothèse alternative H_1 est acceptée.

Si H_1 est acceptée, le test des plus petites différences significatives LSD (« Least Significant Difference ») est utilisé pour comparer les biais entre les modèles. Le niveau de significativité est le même pour l'ANOVA et le test LSD (5 %). La valeur LSD permet de regrouper les modèles ayant des biais qui ne sont pas significativement différents les uns des autres. Si la différence entre les biais de deux modèles est inférieure à la valeur LSD calculée alors ces deux modèles ne sont pas significativement différents avec α = 5 %.

Équation 64 $$\text{Valeur LSD} = t\sqrt{(2MSr/n)}$$

Avec t la valeur de Student lu dans les tables (bilatérale, 5 %, degré de liberté de l'ANOVA), n nombre d'échantillons du lot de validation et MSr la moyenne des carrés des résidus de l'ANOVA.

L'ANOVA est simple d'utilisation mais deux conditions doivent être vérifiées. Wold[90] propose deux règles pour son utilisation : la première concerne l'hypothèse de normalité et la deuxième concerne la variance intragroupe. "La première règle consiste à ne pas analyser des données qui ont un épaulement dans leurs distributions"[90]. La deuxième

concerne l'homoscedasticité : "Les variances intra-classes ne doivent pas différer de plus d'un facteur dix"[90].

- ***Comparaison des SEP(C) (erreur aléatoire)***

Pour comparer le SEP(C) qui correspond à l'écart-type des résidus, un test de Fisher[81] est utilisé comme précédemment pour la comparaison des SEP.

Afin de comparer plusieurs modèles, une valeur limite du SEP(C) est calculée avec la formule suivante :

Équation 65 $$\text{limite de confiance SEP(C)} = \text{SEP(C)}_{min} \cdot \sqrt{F_{critical}}$$

avec SEP(C)_{min} la plus petite valeur du SEP(C).

En conséquence, tous les modèles qui ont un SEP(C) compris entre SEP(C)_{min} et la limite de confiance du SEP(C) ne sont pas significativement différents. L'ANOVA, les tests LSD et les tests de Fisher sont effectués avec le logiciel Matlab® (The Mathworks, Natick, USA).

- Discussion sur les tests statistiques

Pour comparer les modèles, il est nécessaire d'avoir un critère objectif. C'est pourquoi des tests statistiques ont été utilisés. Cependant en toute rigueur, le test de Fisher et l'ANOVA ne s'appliquent que sur des données indépendantes. Ainsi, l'ANOVA appropriée est l'ANOVA à deux facteurs (Effet lié au modèle et effet lié aux échantillons). En utilisant l'ANOVA à un facteur (facteur liè au modèle), on estime que la précision de la méthode chimique de référence est suffisamment grande pour être négligée. De même, le F test aurait du être réalisé sur des résultats issus de modèles construits et validés avec des lots de données indépendantes.

Cependant dans la pratique, de nombreux auteurs utilisent ces tests sans vérifier cette hypothèse[66,91]. Quand l'hypothèse d'indépendance n'est pas vérifiée, les conséquences sont les suivantes : le test statistique perd en puissance et le risque α réel est supérieur au risque considéré. Des petites différences entre les modèles ne sont donc pas détectées. Mais dans notre étude, cette perte de puissance est compensée par le grand nombre d'échantillons analysés.

4.5.2 Tests statistiques pour la comparaison des méthodes de classification supervisées

Nous appellerons taux de prédiction le pourcentage d'échantillons convenablement classés. L'objectif est de pouvoir identifier des méthodes qui ont des taux de prédiction significativement différents. Dans notre étude, le test statistique de McNemar est utilisé pour comparer les taux de classifications correctes[92] obtenus avec deux méthodes différentes. Ce test est par exemple utilisé en médecine[93,94] pour comparer l'effet d'un principe actif par rapport au placebo lors d'essais cliniques en double aveugle.

D'autres tests peuvent être utilisés pour comparer des algorithmes de classification : F test, t test apparié[95]. Le test de McNemar a été choisi parce qu'il est le seul à avoir un niveau d'erreur de type I acceptable (α) et une faible erreur de type II (β)[95]. Le premier type d'erreur (α) correspond à la probabilité de rejeter un échantillon membre de la classe testée comme un non membre (faux négatif). Le second type d'erreur (β) est la probabilité de classer un échantillon non membre dans la classe étudiée (faux positif).

Deux méthodes A et B sont utilisées. L'hypothèse nulle H_0 est la suivante : les deux méthodes ont le même taux de prédiction. Les statistiques n_{01} et n_{10} suivent sous H_0 la même distribution binomiale de paramètre $n' = n_{01} + n_{10}$ et $p = ½$. Ce test est basé sur un test du χ^2 à un degré de liberté (si le nombre n' est supérieur à vingt).

Le Tableau 4 montre comment la valeur de McNemar est calculée. La valeur critique du χ^2 avec un niveau de significativité à 5 %, notée $\chi^2_{(1;0,95)}$, est 3,8414. Si l'hypothèse est vraie, la probabilité d'avoir une valeur de McNemar supérieure à $\chi^2_{(1;0,95)}$ est inférieure à 5 %. Si cette valeur est supérieure à $\chi^2_{(1;0,95)}$ alors l'hypothèse nulle est fausse et les deux méthodes sont significativement différentes. Dans la pratique, les tables de contingence sont construites à partir des résultats du lot de validation et les tests de McNemar effectués sous Matlab®.

Tableau 4 ■ Tableau de contingence de McNemar et calcul de sa valeur

n_{00} : Nombre d'échantillons mal classés par les deux méthodes A et B.	n_{01} : Nombre d'échantillons mal classés par A mais pas par B.
n_{10} : Nombre d'échantillons mal classés par B mais pas par A.	n_{11} : Nombre d'échantillons correctement classés par A et par B.
	$\text{Valeur de McNemar} = \frac{\left(\lvert n_{01} - n_{10} \rvert - 1\right)^2}{n_{01} + n_{10}}$

5 Mise en application : instrumentation, échantillonnage et logiciels

Les interactions entre le rayonnement électromagnétique et la molécule ont été décrites précédemment. De plus, l'interprétation générale du spectre PIR et la mise en place d'une analyse quantitative ont été présentées. Cette partie a pour but de présenter le matériel utilisé, c'est à dire l'instrumentation proche infrarouge et les logiciels de traitements des données.

Les instruments peuvent être classés selon leur principe de fonctionnement (Figure 16). Dans cette partie, nous détaillerons uniquement le principe des instruments utilisés dans cette étude : le spectromètre à réseau et à transformée de Fourier.

Les spectromètres comportent quatre parties essentielles[96] :

- Une source lumineuse,
- Un dispositif permettant de séparer les longueurs d'onde,
- Un système de présentation de l'échantillon,
- Un ou plusieurs capteurs photosensibles.

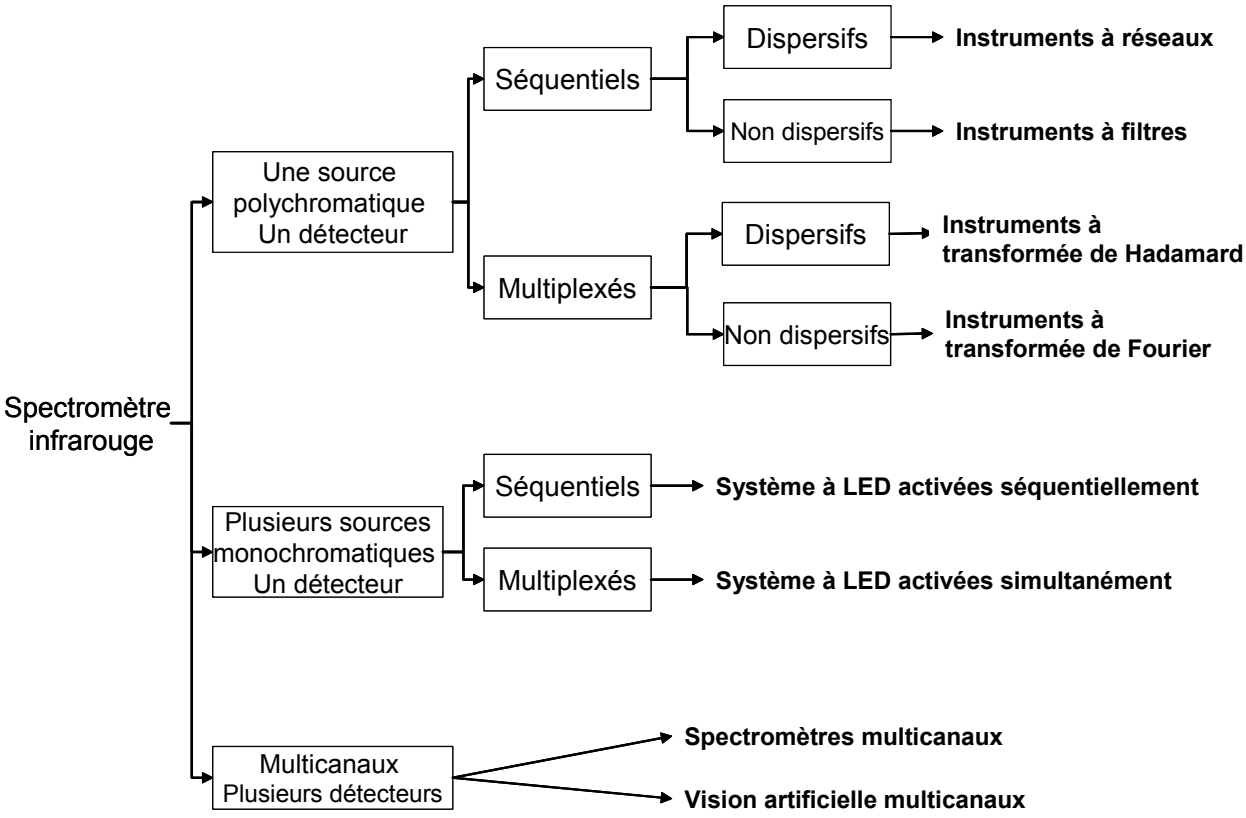


Figure 16 ■ Différents principes des spectromètres (d'après Smith[97]).

5.1 Sources

Les sources lumineuses présentant un intérêt pour la spectroscopie peuvent être divisées en plusieurs classes : les sources thermiques, les sources à décharge, les diodes électroluminescentes et les lasers.

Les sources thermiques sont les plus couramment utilisées en spectroscopie proche infrarouge. Les sources de rayonnement infrarouge usuelles sont constituées par des solides portés à haute température qui rayonnent par incandescence.

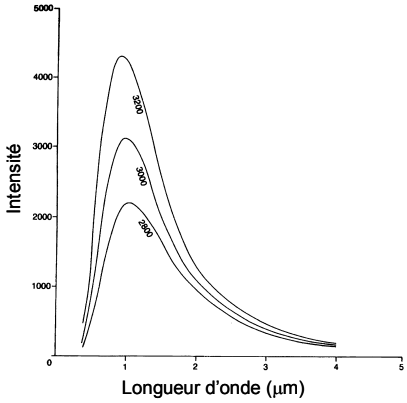


Figure 17 ■ Intensité du rayonnement des corps incandescents en fonction de la longueur d'onde et de la température en Kelvin (d'après Osborne et ses collaborateurs[14]).

Ces sources présentent de nombreux avantages. Elles émettent des radiations qui couvrent une large gamme spectrale et qui ont une forte intensité. De plus, elles restent stables pendant de longues périodes[18]. Dans la pratique, les sources lumineuses utilisées sont des ampoules à filament de tungstène maintenu à 2 400 K.

5.2 Dispositif d'analyse du rayonnement et principes des instruments

Les spectromètres peuvent être classés selon leur nature dispersive ou non. Il existe deux principes pour enregistrer un spectre. La première manière consiste à détecter une longueur d'onde de façon séquentielle et à déterminer l'énergie absorbée. La seconde solution consiste à détecter toutes les longueurs d'onde du domaine spectral de façon simultanée. On parle alors de méthodes multiplexées.

5.2.1 Instruments séquentiels

- **Systèmes à filtres optiques**

Un filtre interférentiel est basé sur le principe de l'interféromètre de Fabry Pérot, c'est-à-dire qu'il combine de multiples réflexions entre deux faces semi réfléchissantes. Ce dispositif ne permettant des mesures qu'à une seule longueur d'onde, son utilisation est donc limitée. C'est pourquoi ce dispositif ne sera pas détaillé.

- **Systèmes à monochromateurs**

Un réseau par réflexion est un système optique réfléchissant dont la surface est constituée d'une série de traits parallèles gravés. Ces rainures peuvent être obtenues de façons mécaniques ou par l'utilisation de lasers dans le cas des réseaux holographiques[98].

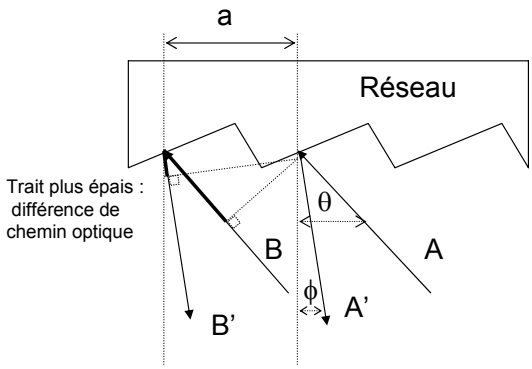


Figure 18 ■ Principe d'un réseau de diffraction

La Figure 18 présente le principe d'un réseau. Un rayonnement polychromatique incident sur un réseau sera diffracté dans plusieurs directions. Les rayons incidents A et B forment un angle θ avec la normale au réseau. La distance entre deux rainures du réseau est notée a. Les rayons réfléchis A' et B', de longueur d'onde λ, forment un angle ϕ avec la normale au réseau. La condition d'interférences constructives entre les rayons issus de deux miroirs voisins s'exprime en fonction de leur différence de marche par :

Équation 66 $a.[\sin \phi + \sin \theta] = k.\lambda$

avec k ordre de diffraction, a et ϕ sont définis sur la Figure 18.

Le réseau est intégré dans un ensemble optique appelé monochromateur. La lumière polychromatique est envoyée sur le système à travers une fente d'entrée et recueillie à travers une fente de sortie. Par rotation du réseau, différentes longueurs d'onde sont obtenues en sortie. La Figure 19 présente le fonctionnement d'un instrument à réseau. Dans notre étude, l'appareil utilisé est un instrument FOSS® NIRSystem modèle 6500. La mesure spectrale est faite en 1 minute : la référence et l'échantillon sont analysés respectivement 10 et 25 fois sur la gamme spectrale comprise entre 400 nm et 2498 avec un pas de 2 nm.

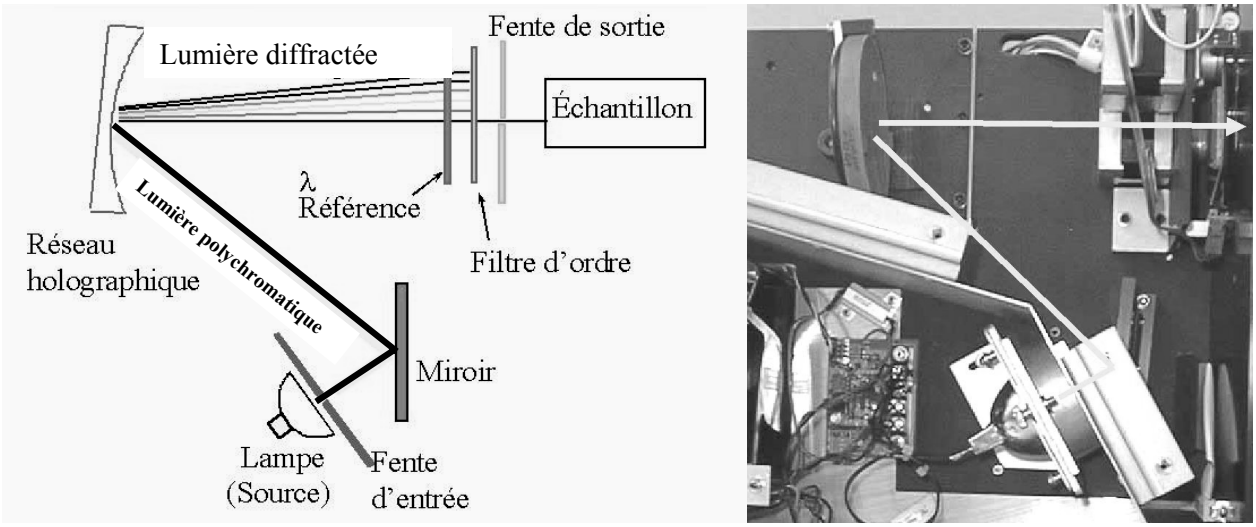


Figure 19 ■ Schéma de fonctionnement d'un instrument dispersif et photographie de l'instrument FOSS NIRsystem. Le trajet de la lumière est représenté par la flèche grise.

5.2.2 Instruments multiplexés à transformée de Fourier

Le dispositif central des spectromètres à transformée de Fourier est l'interféromètre de Michelson[99] (Figure 20). Un interféromètre de Michelson est formé de deux miroirs plans perpendiculaires dont l'un est fixe et l'autre est mobile. Une lame semi-réfléchissante appelée séparatrice, inclinée à 45° par rapport à la direction de propagation du rayonnement incident est placée au centre du montage. La séparatrice dédouble le faisceau incident en un faisceau transmis et un faisceau réfléchi. Ces deux faisceaux sont ensuite réfléchis par les miroirs. Si les deux miroirs sont à égale distance de la séparatrice, les chemins optiques suivis par les deux faisceaux sont identiques et ils émergent en phase. Si le miroir mobile est translaté de x, le chemin optique du premier trajet augmente de 2 x et les deux faisceaux sont déphasés[8]. Le signal enregistré par le détecteur dépend de x est suit une loi périodique (Équation 67).

Équation 67 $I(x) = B(\lambda).\cos^2(2\pi.x / \lambda)$ avec I(x) l'intensité enregistrée par le détecteur et $B(\lambda)$ l'énergie incidente à l'interféromètre.

L'intensité I(x) porte le nom d'interférogramme. L'interférogramme d'une source polychromatique est une somme de fonctions périodiques. Ensuite, par une transformation mathématique appelée transformée de Fourier[100], il est possible à partir de l'interférogramme de reconstruire le spectre infrarouge.

Par rapport aux méthodes séquentielles, pour un temps d'analyse donné, le spectre obtenu avec un instrument à transformée de Fourier sera moins bruité. Cependant dans le PIR, ces instruments ont de moins bonnes performances de sensibilité et de reproductibilité que les appareils à réseaux[98]. L'instrument à transformée de Fourier utilisé dans notre étude est un appareil Bruker®, modèle IFS 22 / N.

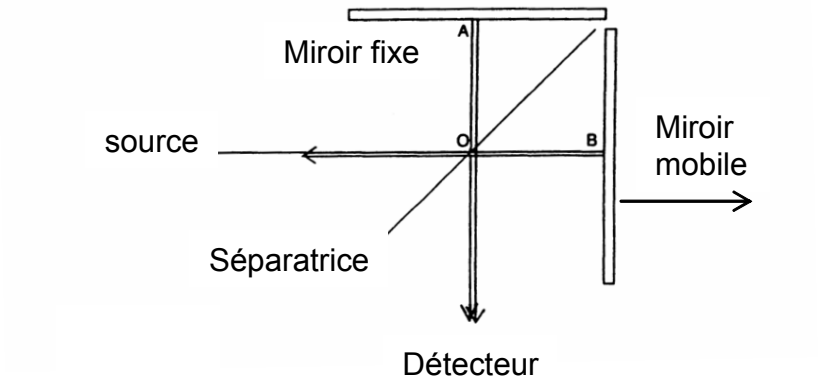


Figure 20 ■ Principe d'un spectromètre à transformée de Fourier (d'après Osborne et ses collaborateurs[14]).

5.3 Détecteurs

Un détecteur est caractérisé par sa limite de détection, son domaine spectral et son efficacité quantique[8]. L'efficacité quantique est le rapport du courant induit sur le flux incident.

En spectroscopie optique, il existe deux types de détecteurs :

- des détecteurs thermiques qui sont sensibles à la chaleur dégagée par l'absorption des photons (exemple détecteur DTGS Deuterated TriGlycide Sulfate).
- des détecteurs quantiques sensibles aux transitions électroniques. Ces détecteurs possèdent en général une réponse plus rapide que les détecteurs thermiques.

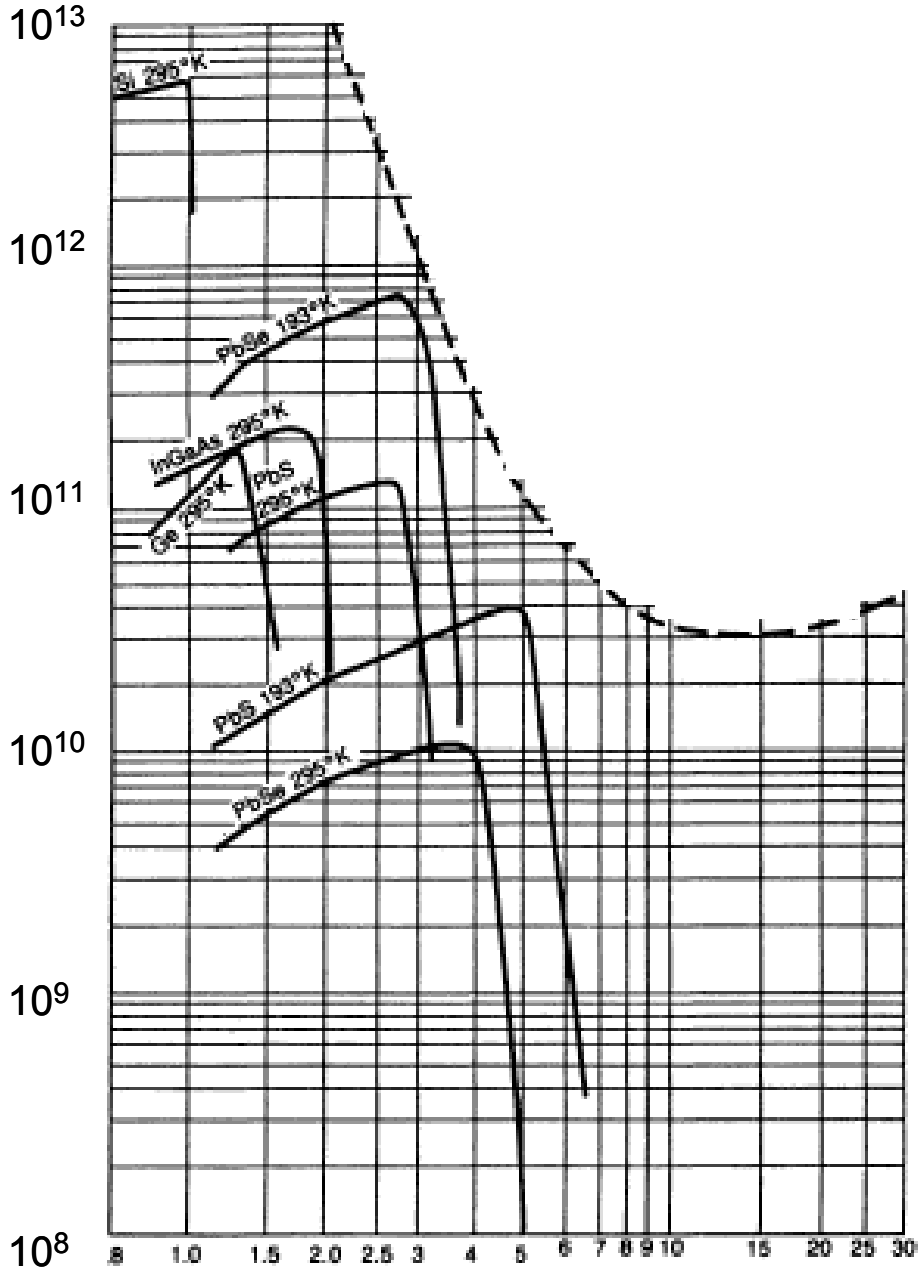


Figure 21 ■ Caractéristiques de quelques détecteurs dans le domaine du proche infrarouge. En abscisse, les longueurs d'onde en µm et en ordonnées, la détectitvité spécifique en $cm.Hz^{0.5}.W^{-1}$.

Depuis de nombreuses années, le détecteur au sulfure de plomb (PbS) tient la première place pour l'analyse du domaine proche infrarouge[101]. Le principal avantage du détecteur PbS est sa capacité à couvrir le domaine 900 nm-2500 nm avec un bon rapport signal sur bruit (Figure 22).

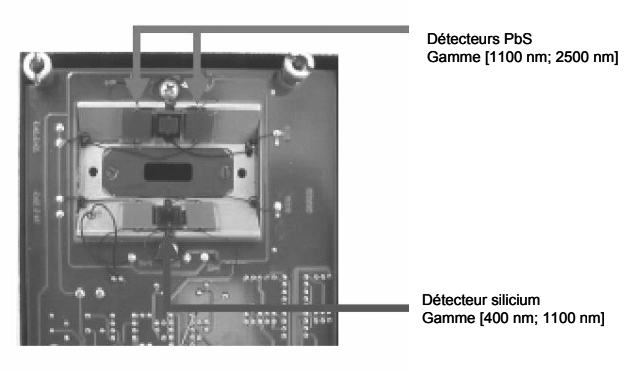


Figure 22 ■ Photographie des détecteurs de l'instrument FOSS NIRsystem 6500. Deux types de détecteurs sont utilisés : des détecteurs silicium et PbS qui recueillent respectivement l'information sur la gamme 400 nm - 1100 nm et 1100 nm - 2500 nm.

5.4 Echantillonnage

Les cellules sont adaptées à la mesure spectrale. On distingue ainsi des cellules de mesure en transmission, des cellules de mesure en réflexion diffuse ou des cellules en transflexion qui disposent d'une paroi réfléchissante. Dans notre étude, les deux instruments effectuent la mesure en réflexion.

L'utilisation de cellules de grandes dimensions, effectuant des mouvements de rotation ou de translation, permet de s'affranchir de l'hétérogénéité de l'échantillon. Dans notre étude, l'instrument FOSS utilise une cellule rectangulaire recouverte d'une fenêtre de quartz (Figure 24).

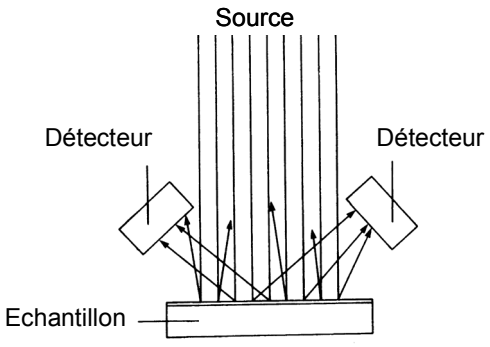


Figure 23 ■ Principe de l'échantillonnage. Les détecteurs sont placés à 45 °.

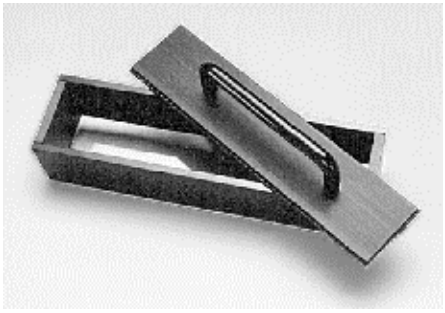


Figure 24 ■ Exemple de cellule de mesure utilisée avec l'instrument FOSS NIRsytem 6500 (Natural product sample cup® IH 0314P).

5.5 Logiciels utilisés

- ***Logiciels utilisés pour la mise en place de l'analyse quantitative***

Concernant la mise en place de l'analyse quantitative, la majorité des modèles ont été développés avec le logiciel Winisi (Infrasoft®, Port Matilda, USA) car il est adapté à l'instrument FOSS. Mais la cohérence des résultats entre ce logiciel et d'autres (Unscrambler, Matlab) a été vérifiée.

- ***Logiciels utilisés pour la classification des spectres de betterave***

Comme les tests statistiques, toutes les méthodes exceptées DPLS et PDA ont été développées sous Matlab® (v6.0 – The Math Works Inc., Natick, USA). SIMCA et LDA proviennent de la boîte à outils Matlab® PLS (v2.0, Eigenvector research,). PNN et LVQ sont disponibles avec la boîte à outils réseaux de neurones (neural network toolbox®v 4.0, The Math Works Inc., Natick, USA). DPLS est une méthode du logiciel Winisi (Infrasoft®, Port Matilda, USA) et PDA est réalisé avec le programme Holmes[84].

6 Bilan

La spectroscopie proche infrarouge possède de nombreux avantages. L'absorption des bandes harmoniques et des bandes de combinaisons sont moins intenses que celle des bandes fondamentales. Le chemin optique avant d'atteindre la saturation est donc plus long. Ainsi des produits plus concentrés pourront être analysés et l'étape de dilution des échantillons n'est plus nécessaire. De plus, les pics sont larges et arrondis[102] : les spectres seront peu sensibles aux décalages en longueurs d'onde. Enfin, les interférences liées à la taille des particules sont moins importantes dans le proche infrarouge que dans l'infrarouge conventionnel. Un autre avantage de la SPIR est lié à l'instrumentation. En effet, la source lumineuse PIR possède un bon rendement énergétique car le maximum d'émission se trouve dans le domaine étudié. Ce facteur combiné aux capacités des détecteurs, permet d'obtenir un rapport[13] signal sur bruit de plus de 10000. La spectroscopie proche infrarouge permet d'analyser in situ des échantillons naturels sans les détruire[103]. Les détecteurs ont un temps de réponse court et les fibres optiques[104] permettent de délocaliser l'analyse sur de grandes distances.

Le principal inconvénient est le suivant : les spectres PIR bruts sont difficiles à interpréter et pour utiliser la SPIR pour le dosage, il est nécessaire de mettre en place une analyse quantitative, c'est à dire, construire un modèle mathématique qui mette en relation les valeurs obtenues par la méthode d'analyse chimique de référence et les spectres proche infrarouge.

Dans ce chapitre, les outils mathématiques utilisés pour extraire l'information spectrale ont également été décrits. Grâce aux méthodes de régression, des concentrations en différents constituants tels que le saccharose pourront être calculées à partir des spectres proche infrarouge. De même, par l'utilisation des méthodes de classification, des critères qualitatifs pourront être déterminés.

Avant de présenter les principaux résultats, il faut décrire les caractéristiques de l'échantillon analysé et les analyses chimiques qui seront utilisées comme référence lors de la mise en place d'analyses quantitatives.

Chapitre 2

Betterave sucrière et technologie

1 Introduction

La première référence à la famille de la betterave se trouve dans la littérature grecque vers 420 avant J.C. Au 15ème siècle, la betterave est présente dans l'Europe entière. A l'origine, elle est cultivée pour ses feuilles qui sont consommées comme des épinards. En 1747, un scientifique allemand nommé Margraaf[105] démontre que les cristaux sucrés obtenus à partir de la betterave sont les mêmes que ceux de la canne à sucre. Malgré la découverte de Margraaf, la canne à sucre reste la principale source de sucre au début du 19ème siècle. Dès 1806, à cause du blocus continental, il y a une pénurie de sucre de canne. En 1811, des scientifiques français produisent des pains de sucre issus de la betterave. Napoléon décrète alors que 32 000 hectares de terre doivent immédiatement être dédiés à la betterave et il subventionne la construction de sucreries. A la fin du blocus continental, plusieurs pays stoppent leur production de betteraves. Au contraire, le gouvernement français encourage le développement de nouvelles variétés de betteraves et de nouvelles techniques d'extraction[106].

L'objectif de ce chapitre est de décrire les caractéristiques de la betterave, de présenter les analyses chimiques effectuées classiquement pour la détermination de sa qualité et enfin de faire le bilan de l'utilisation de la spectroscopie proche infrarouge dans l'industrie sucrière.

2 Caractéristiques de la betterave à sucre

2.1 Caractéristiques agronomiques

La betterave à sucre *(Beta vulgaris* - famille *Chenopodiaceae)* est une plante bisannuelle cultivée pour sa racine charnue (Figure 25) qui se forme la première année. Sa racine, riche en saccharose, est essentiellement utilisée pour la fabrication du sucre et dans une moindre mesure pour la distillation. Les collets et les feuilles sont consommés par le bétail ou utilisés comme engrais verts.

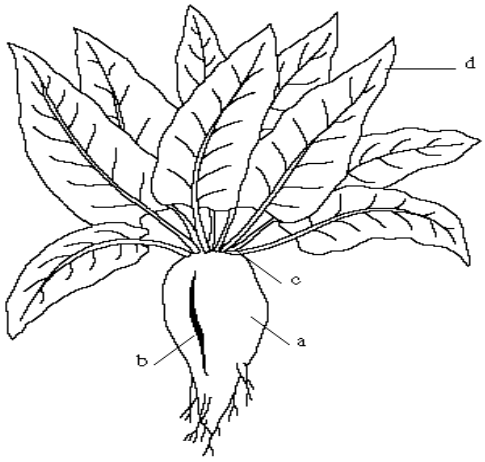


Figure 25 ■ Présentation de la betterave à sucre. a: Racine rugueuse à forme conique, b: sillon sacharifère, c: collet (tige), d: feuille (d'après Boiffin et Choppin de Janvry [107]).

Le semis a lieu tôt au printemps et la récolte s'effectue en automne (septembre - décembre), lorsque les feuilles jaunissent et prennent un port tombant. La culture des betteraves couvre une superficie de l'ordre de 2,6 % des terres labourables de France, soit 457 000 hectares en 1997. Même si la betterave est cultivée dans plus de 20 départements, on constate que douze départements situés au nord de la Loire représentent 90% des surfaces plantées en betterave (Tableau 5). Pour la construction d'une analyse quantitative, les variabilités géographiques, variétales et temporelles ont été prises en compte. De plus dans le chapitre 6, nous nous intéressons à la détermination par SPIR de critères qualitatifs tels que l'origine géographique, la période de récolte et la résistance à une maladie appelée rhizomanie.

Tableau 5 ■ Principaux départements français cultivant la betterave et superficie cultivée en 2000.

Départements	Surface en ha	en % de la surface totale
Aisne	71 840	16,30
Ardennes	12 543	2,85
Aube	24 368	5,53
Calvados	6 400	1,45
Côte-d'Or	3 688	0,84
Eure	10 855	2,46
Eure-et-Loir	6 296	1,43
Loiret	21 018	4,77
Marne	54 323	12,33
Nord	22 649	5,14
Oise	42 594	9,67
Pas-de-Calais	43 399	9,85
Puy-de-Dôme	2 826	0,64
Bas-Rhin	3 916	0,89
Haut-Rhin	1 654	0,38
Seine-Maritime	15 051	3,42
Seine-et-Marne	32 004	7,26
Somme	43 968	9,98
Autres	10388	4,81

2.2 Composition chimique

Comme le montre la Figure 26, la betterave a une composition chimique complexe. Elle contient un grand nombre de composés qui influenceront le spectre infrarouge. De plus, au sein de la betterave, il existe des distributions différentes de ces composés : ainsi par exemple, le collet est moins riche en sucre que le reste de la racine mais il a un taux de matières sèches plus élevé[108].

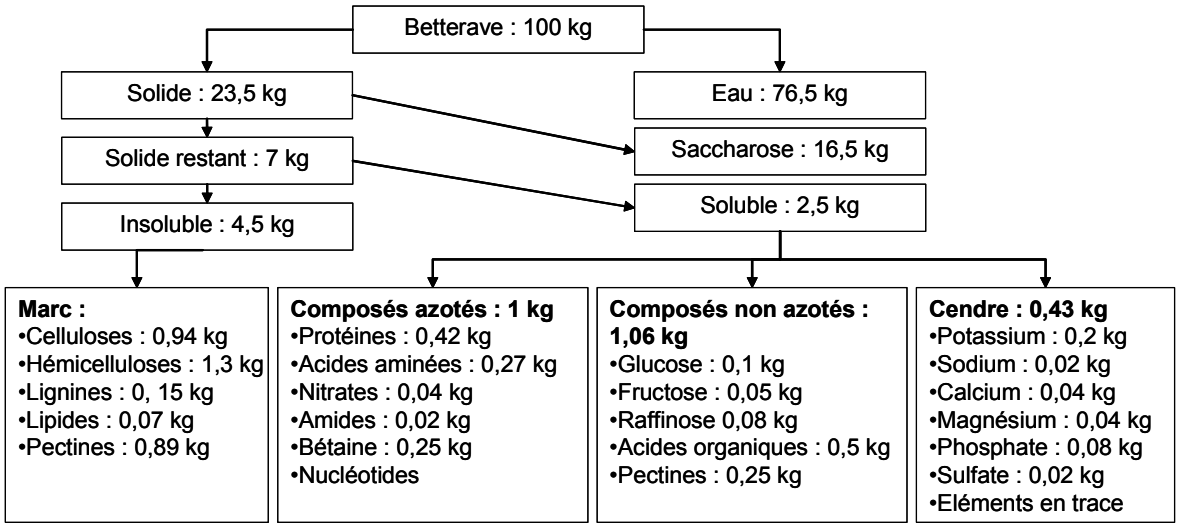


Figure 26 ■ Composition chimique moyenne de la betterave à sucre (D'après Schiweck et collaborateurs[109])

Pour évaluer la qualité globale de la betterave, les composés dosés en routine sont les suivants[110] : le saccharose, le glucose, l'azote total, le sodium et le potassium. De plus, le brix qui correspond au pourcentage de solide, le marc qui est le poids des composés insolubles dans l'eau et l'éthanol sont également mesurés. Les méthodes utilisées pour déterminer ces composés sont décrites car elles constituent les méthodes de référence pour la mise en place des analyses quantitatives par SPIR.

3 Analyses chimiques pour la détermination de la qualité de la betterave

3.1 Préparation de l'échantillon

La préparation de l'échantillon est la même pour le dosage du saccharose, du glucose, du sodium, du potassium et de l'azote.

- ***Broyage de l'échantillon***

Tous les échantillons ont été traités selon la méthode réglementaire de réception[111]: ils sont lavés en s'assurant qu'il ne reste pas de terre dans le sillon sacharifère puis le collet de la betterave est coupé. On procède alors à un sous échantillonnage. Cet échantillon est sous forme de râpure, c'est-à-dire de betterave broyée finement pour pouvoir extraire facilement les matières solubles. La râpe est normalisée pour les sucreries françaises. Il s'agit d'un instrument à scies circulaires (Parmentière®) qui permet d'obtenir un échantillon de râpure dont le poids est égal à 10 % du poids de départ. Enfin, il y a une étape d'homogénéisation pendant 9 secondes par un appareil homologué (modèle IUA®)[112]. Pour l'analyse par SPIR, la préparation de l'échantillon s'arrête à cette étape d'homogénéisation.

- ***Préparation d'une solution limpide***

Un prélèvement de 40 g ± 0,05 g de râpure homogénéisée soigneusement est placé sur un papier glacé. La râpure est ensuite transvasée dans un bécher de la chaîne automatique de digestion et on ajoute 165 mL ± 0,20 mL d'une solution diluée d'acétate de plomb hydraté à 2,5 %, appelé sous acétate de plomb (formule : $Pb(CH_3COO)_2 \cdot 3H_2O$). Par convention, le volume total est considéré égal à 200 mL. Le couvercle est placé sur le bécher et l'agitation est vigoureuse. Après environ 15 minutes, les échantillons sont filtrés sur papier filtre simple.

La température du laboratoire est maintenue à 20 °C et le degré d'humidité à 80 % minimum. La préparation de l'échantillon est réalisée grâce à un système automatisé comprenant une balance proportionneuse (Gallois® selon la réglementation de 1964) qui délivre le sous acétate de plomb, une chaîne d'homogénéisation et de filtration automatique.

3.2 Dosage du saccharose : méthode réglementaire

3.2.1 Structure du saccharose

Grâce au phénomène de photosynthèse, l'énergie lumineuse est convertie en énergie disponible pour les cellules végétales. Cette énergie est utilisée pour synthétiser des oses au niveau de la feuille. Le saccharose est la forme de transport qui permet le mouvement entre les différentes zones de synthèse, d'utilisation et de mise en réserve. Dans la plupart des espèces végétales, le saccharose est une forme de réserve transitoire et il est ensuite transformé en amidon. Mais dans la canne à sucre et la betterave, il constitue des réserves durables[113]. Le saccharose est un diholoside, constitué de glucose et de fructose réunis par une liaison osidique (Figure 27).

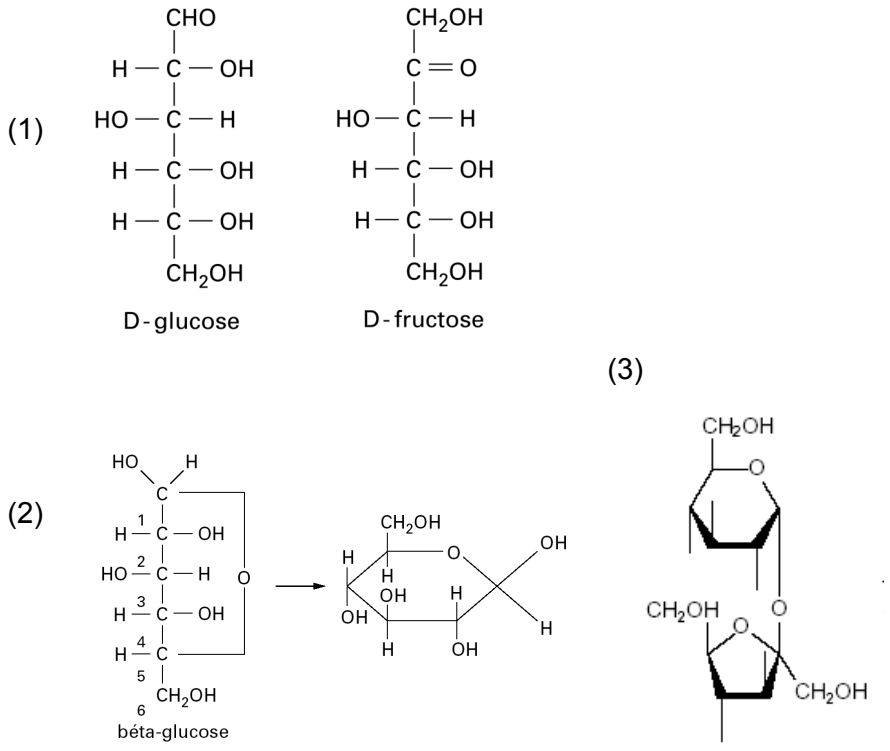


Figure 27 ■ Formules chimiques du glucose, fructose et saccharose. (1) Formules linéaires du glucose et du fructose ; (2) Formule cyclique du glucose ; (3) Formule du saccharose.

3.2.2 Principe physique de la méthode

Le pouvoir rotatoire est la propriété que possèdent certaines molécules de faire tourner le plan de polarisation d'un faisceau de lumière rectilignement polarisée. L'angle de rotation suit la loi de Biot (Équation 68). Ainsi, la rotation de la lumière due aux solutions de saccharose est très proche d'une fonction linéaire de leurs concentrations :

Équation 68 $\alpha_\lambda^t = [\alpha]_\lambda^t.c.l$ avec c : la concentration en saccharose en $g.mL^{-1}$, l : la longueur du tube polarimétrique en dm, α_λ^t : l'angle de rotation de la lumière en degrés à la température t et la longueur d'onde λ, $[\alpha]_\lambda^t$: le pouvoir rotatoire spécifique du saccharose en $°.mL.dm^{-1}.g^{-1}$ à la température t et la longueur d'onde λ.

La longueur d'onde de la mesure a été fixée à 589,3 nm[114]. A partir de la rotation produite par l'échantillon analysé, on peut retrouver la teneur en saccharose. Cette relation n'est réellement vérifiée que dans le cas théorique où le saccharose est la seule substance optiquement active de l'échantillon examiné[115]. En pratique, aucun échantillon de betterave, ne remplit ces conditions. L'avantage de la polarimétrie dans l'analyse des sucres est d'être simple et rapide.

3.2.3 Instrumentation

Le principe des polarimètres photométriques à lame de quartz est décrit par la Figure 28. Le polariseur et l'analyseur sont fixes en position croisée. La translation du coin de quartz permet de faire varier l'épaisseur de quartz traversée par le faisceau.

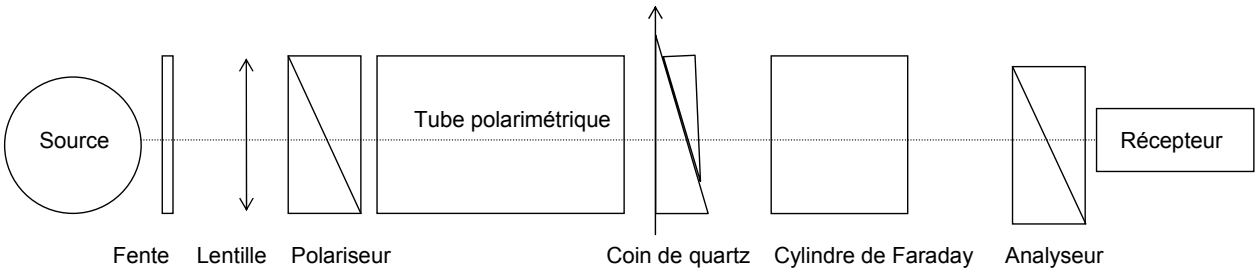


Figure 28 ■ Principe d'un saccharimètre.

En absence de substances actives, le coin de quartz étant hors du faisceau, il y a extinction. Quand une solution de saccharose est introduite, le plan de polarisation est dévié dans un sens. En déplaçant le coin de quartz selon l'axe vertical, une rotation en sens inverse est provoquée. Lorsque le quartz a compensé la rotation due à la substance, il y a à nouveau extinction[116]. La translation nécessaire du coin de quartz est proportionnelle à la teneur en saccharose. La lecture de la mesure se fait sur une échelle indiquant le déplacement du coin de quartz.

Les saccharimètres sont des polarimètres portant des graduations définissant le pourcentage de sucre et spécialement conçus pour l'industrie sucrière. L'échelle saccharimétrique internationale, exprimée en degré sucre, est par convention la suivante : 100°S correspond à une solution de 40 g de saccharose dans 100 cm^3. Pour la raie D du sodium, dans ces conditions standardisées, l'angle de rotation est de 34,626° et le pouvoir rotatoire spécifique[117] est de 66,59°.mL.dm^{-1}.g^{-1}. Dans notre étude, le polarimètre est de marque Lisabio® (modèle Pol F).

3.3 Dosage du saccharose : autres méthodes

3.3.1 Dosage des sucres par Chromatographie Liquide Haute Performance (CLHP)

- ***Principes***

La CLHP est la principale méthode de détermination des sucres. Elle a l'avantage d'être simple et rapide pour identifier et quantifier les sucres dans des mélanges complexes[118].

Les techniques chromatographiques mettent en jeu les coefficients de partage entre une phase mobile entraînant le mélange à analyser et une phase stationnaire. Le liquide éluant est injecté à l'aide d'une pompe à débit constant dans une colonne séparatrice. A la sortie le liquide passe par un système de détection. L'analyse des sucres, composés fortement polaires mais non ioniques sauf en milieu très basique, peut être réalisée selon différentes voies (Tableau 6).

Tableau 6 ■ Revue des différentes méthodes utilisables pour l'analyse des sucres par CLHP (d'après Mathouthi et Reiser[117])

Nature de la phase stationnaire	Principe de séparation	Eluant	Particularité
Silice greffée C18	Polarité	eau	Séparation par masses moléculaires croissantes.
Silice greffée NH_2	Différence de polarité	80 % acétonitrile / 20 % eau	
Résines cationiques	Echange de ligand	eau	Mauvaise séparation des osides
Résines anioniques	pKa	solution de soude	Très bonne séparation

L'utilisation des résines pelliculaires échangeuses d'ions à pH élevé, associée à la détection ampérométrique pulsée (« Pulsed Amperometric Detector » - PAD), a considérablement amélioré la séparation et l'analyse des mono- et oligosaccharides par CLHP[119,120,121].

Garcia-Jares et Médina[122] ont mis en place des modèles de prédiction des teneurs en glucose, fructose et sucres totaux dans le raisin à partir de spectres proche infrarouge et des analyses par CLHP. De même, Cho et Hong[123] mesurent les sucres du miel par spectroscopie proche infrarouge en utilisant la CLHP comme méthode de référence.

Le mécanisme de séparation de ces colonnes est lié au fait que les glucides sont des acides faibles qui sont ionisés à pH élevé. Les glucides ainsi séparés sont détectés en mesurant le courant électrique généré par leur oxydation à la surface d'une électrode en or à un pH optimal de 13[124].

- ***Protocole utilisé***

Un Système Dionex LC (Dionex Sunny Vale, CA, USA) connecté à un détecteur PAD (Dionex PAD-2) à électrode d'or avec une colonne (4,6 x 250 mm) de Carbo Pac PA-1 a été utilisé. La colonne est remplie avec un échangeur d'anions en amine quaternaire basique forte sur substrat pelliculaire de 10 μm en polystyrène-divinylbenzène. L'élution s'effectue à une vitesse de 1 $mL.min^{-1}$ (régime isocratique) par une solution de NaOH de concentration 0,2 $mol.L^{-1}$. La colonne et le détecteur sont thermostatés à 27 °C.

On prélève environ 5 g de râpure de betterave auxquels on ajoute 35 g d'acétate de plomb (à 36 %). Le lactose est utilisé comme un étalon interne. Avant la défécation au plomb, on introduit 4 mL de lactose à 100 $g.L^{-1}$ dans l'échantillon. L'échantillon est agité pendant 5 minutes. Ensuite, l'échantillon est filtré sur papier. On ajoute alors du carbonate de calcium (CO_3Ca) pour précipiter le plomb.

L'échantillon est filtré sur membrane de 0,45 μm puis injecté (volume 25 μL). L'identification et la concentration des glucides sont données par comparaison avec des solutions étalons analysées trois fois (Tableau 7). Une régression linéaire simple entre le rapport des masses (masse glucide/masse étalon interne) et le rapport des aires (aire glucide/aire étalon) permet d'obtenir la pente de la droite d'étalonnage (Annexe 2).

Tableau 7 ■ Solutions étalons utilisées. (1) Glucose, Fructose, Raffinose

	Solution 1	Solution 2	Solution 3	Solution 4
Masse de saccharose / masse de lactose	1	2	3	4
Masse autre glucide(1) / masse de lactose	0,1	0,2	0,3	0,4

Tous les échantillons de betterave sont analysés trois fois et les concentrations en saccharose, glucose, fructose et raffinose sont déterminées par la relation suivante :

Équation 69 $C = (A_{composé} / A_{lactose}) * (M_{lactose} / M_{échantillon})* (100 / \text{pente droite d'étalonnage})$

avec C la concentration en g / 100g de betterave et M la masse en g.

3.3.2 Dosage enzymatique

Le dosage enzymatique fait intervenir la β-fructosidase, une enzyme qui va hydrolyser le saccharose en glucose et en fructose. Ensuite le glucose est dosé par voie enzymatique[125] (méthode décrite en 3.4.3). Le dosage du glucose est double : avant et après hydrolyse du saccharose. Ainsi, on peut calculer la quantité de glucose produite par l'hydrolyse et donc la concentration en saccharose.

3.4 Dosage de composés caractéristiques de la qualité de la betterave

3.4.1 Ions

Le dosage du sodium et du potassium s'effectue par photométrie de flamme. Le principe de la méthode est le suivant : l'échantillon à analyser est vaporisé dans la flamme ce qui entraîne une excitation des atomes. Quand les électrons passent d'un niveau d'énergie supérieur à un niveau plus stable, il y a émission de photons[125]. La photométrie de flamme consiste à relier l'intensité de l'émission avec la concentration de l'élément à doser.

L'appareil est étalonné avec une solution contenant une quantité connue de sodium (5 et 10 $mg.L^{-1}$) et de potassium (10 et 20 $mg.L^{-1}$). Les émissions du sodium et du potassium sont mesurées respectivement à 589,6 nm et 766,5 nm.

Cette méthode présente des inconvénients. En effet, il peut y avoir des interférences d'origines diverses :

- Interactions chimiques entre plusieurs composés.
- Superpositions des émissions (quand plusieurs composés émettent à la même longueur d'onde).
- Fonds d'émissions dus à d'autres composés.

3.4.2 Azote

La méthode est celle de Carruthers et Oldfield[126] dérivée de la méthode colorimétrique de Moore et Stein[127]. Elle permet de mesurer l'azote alpha aminé dans le jus de betterave après défécation au plomb. Cette méthode est officiellement adoptée comme méthode de référence par l'ICUMSA (« International Commission for Uniform Methods of Sugar Analysis »).

L'échantillon subit l'action de l'hydrazine (N_2H_4) pour dénaturer les protéines. Les acides aminés libres réagissent avec la ninhydrine pour donner un complexe coloré bleu en présence de propanol. L'absorbance de cette solution est ensuite mesurée à 570 nm.

L'étalonnage est réalisé à partir de solutions contenant 1,3 ; 2,6 ; 3,9 ; 5,2 et 6,5 mg d'azote alpha aminé (acide glutamique) par litre. A chaque échantillon de 1 mL est ajouté 1 mL de réactif (20 g de ninhydrine et 3 g d'hydrazine dans 1 litre). L'échantillon est placé dans un bain marie bouillant pendant 15 minutes. Ensuite le volume est complété à 10 mL avec une solution de propanol (50 %). La teneur en azote alpha aminé est exprimée en mg pour 100 g de betterave fraîche.

Cette méthode présente un inconvénient majeur : la couleur et son intensité varient selon la nature chimique des acides aminés. Cependant ce dosage permet de déterminer de façon précise la teneur totale en acides aminés libres[125].

3.4.3 Glucose

Le dosage du glucose[125] est effectué par une méthode enzymatique connue sous le nom de méthode GOD / POD mise au point par Trinder[128]. Pour doser le glucose, une enzyme, la glucose oxydase (GOD) est utilisée. Cette méthode est spécifique au D-glucose.

Réaction 1 : $C_6H_{12}O_6 + O_2 + H_2O \rightarrow C_6H_{12}O_7 + H_2O_2$

Glucose Acide gluconique

Enfin, la mesure de H_2O_2 formé permet d'obtenir la concentration en glucose[129]. Cette réaction fait intervenir une deuxième enzyme la peroxydase (POD).

Réaction 2 : H_2O_2 + 4 aminophénazone + phénol $\rightarrow$ 4-phénylaminophénazone +H_2O

(4 amino phenazone = 4-diméthylamino-2,3-diméthyl-1-phényl-3-pyrazolin-5-one)

Le 4-phénylaminophénazone est un produit coloré rouge orangé. On effectue un dosage colorimétrique. L'absorbance du produit à analyser et d'une solution étalon est mesurée à 505 nm. Après étalonnage, la concentration en glucose est donnée par la formule suivante[130] :

Équation 70 $[glucose]_{échantillon} = (Absorbance_{échantillon} / Absorbance_{étalon})* [glucose]_{étalon}$

3.4.4 Brix

La préparation de l'échantillon est différente de celle des analyses précédentes. Environ 100 g de betterave sont prélevés et centrifugés pour récolter un jus trouble. Ce jus est ensuite filtré sur verre fritté. Le jus limpide ainsi obtenu est analysé par réfractométrie. Pour une température et une longueur d'onde incidente donnée, l'indice de réfraction mesuré est fonction de la teneur en matières sèches.

Équation 71 Brix = masse de matière sèche / masse totale * 100

Quand un rayon lumineux change de milieu, il est dévié. L'angle est le plus grand dans le milieu ayant l'indice de réfraction le moins élevé (avec une limite de 90°). Dans le cas limite, la relation de Snell devient :

Équation 72 $\sin(\alpha_{critique}) = n_{bas} / n_{haut}$ avec $\alpha_{critique}$ l'angle critique.

Tous les rayons atteignant l'interface à partir du milieu le plus réfringent avec un angle d'incidence plus grand que $\alpha_{critique}$ sont réfléchis totalement. Les réfractomètres commerciaux sont conçus pour détecter cet angle limite. Ils utilisent un prisme fabriqué avec une matière à haut indice de réfraction sur lequel l'échantillon à mesurer est déposé. En connaissant $\alpha_{critique}$ et n_{haut}, le calcul de n_{bas} de l'échantillon est simple.

Dans la profession sucrière, cette grandeur est traditionnellement mesurée en degré brix. 1° brix est équivalent à 1 % de matière sèche. Dans le but de déterminer la matière sèche, on fait l'hypothèse que la matière sèche autre que le saccharose a la même densité et le même indice de réfraction que le saccharose. Il existe des tables reliant l'indice de réfraction au brix. Ainsi, l'indice de réfraction des solutions de sucre est utilisé comme une méthode rapide pour la détermination approchée de la teneur en matière sèche[131].

3.5 Estimation de paramètres industriels

Les principaux produits issus de la transformation de la betterave et susceptibles d'être commercialisés, sont le sucre, la mélasse et les pulpes. Une tonne de betteraves à 16 % de sucre donne en théorie :

- 130 kg de sucres extraits,
- 18 kg de sucres perdus dans les mélasses, soit 37,5 kg de mélasses à 48 % de saccharose,
- 55 kg de matières sèches sous forme de pulpes.

La quantité de pulpe peut être estimée par le dosage des marcs de la betterave. De plus, à partir des analyses précédentes, la quantité de sucre non cristallisable qui se retrouvera dans la mélasse peut être estimée.

3.5.1 Marc

Le marc est la matière sèche insoluble dans l'eau à 50 °C et dans l'éthanol. Cette détermination permet de calculer les quantités de pulpes pressées obtenues pour 100 kg de betterave.

Le protocole d'analyse est le suivant. Une masse M comprise entre 10 et 20 g de râpure de betterave est prélevée. Après broyage dans de l'eau distillée, on filtre sur un verre fritté avec 100 mL d'eau à 50 °C quatre fois en remettant en suspension à chaque fois. Ensuite l'échantillon est rincé avec 100 mL d'éthanol (à 50 %). On laisse égoutter sous vide puis on sèche à l'étuve à 105 °C jusqu'à obtenir une masse constante M'. Enfin le marc est calculé selon l'Équation 73.

Équation 73 Marc en % = (M' / M) * 100

3.5.2 Sucre mélasse

En réalisant un bilan matière, la quantité de sucre produit par une usine m_{SP} est donnée par l'équation suivante[132] :

Équation 74 $m_{SP} = m_S – m_{SM} – m_{PP}$

avec m_S la teneur en saccharose de la betterave et m_{SM} la teneur en saccharose de la mélasse et m_{PP} pertes liées aux procédés en kg .

m_S est mesurée à la réception des matières premières et m_{PP} est calculable et constante quand le régime de l'usine est stationnaire. A part m_S, on remarque que m_{SM} est un paramètre intéressant pour estimer la production en sucre d'une usine.

A partir des dosages de l'azote, du glucose, de potassium et de sodium de la betterave, une équation empirique (Équation 75) permet d'estimer la quantité de sucre non cristallisable qui sera dans la mélasse[133] :

Équation 75 $m_{SM} = a_1 . m_{(K+Na)} + a_2.m_{azote} + a_3.m_{glucose} + a_4$ (% dans la betterave)

avec $a_1 = 0,14$ $a_2 = 0,25$ $a_3 = 3,3$ et $a_4 = 0,3$.

Les masses sont exprimées en g pour 100 g de betterave sauf pour le sodium et le potassium où elles sont exprimées en mmol par kg de betterave.

3.5.3 Pureté du jus

La pureté est le rapport masse de saccharose sur masse de la matière sèche. Il s'agit d'un indice de la qualité de la betterave. Comme pour le sucre mélasse, ce paramètre est évalué grâce à une formule empirique :

Équation 76 $Pj = 99,36 - (14,27 * (m_{(K+Na)} + a_2.m_{azote}) / m_{saccharose})$

4 Historique de l'utilisation de la spectroscopie proche infrarouge dans les industries agroalimentaires et sucrières

4.1 Analyse quantitative de produits agroalimentaires

Historiquement la SPIR a été développée pour le dosage de la teneur en eau d'aliments[134]. C'est pourquoi, uniquement des exemples tirés de l'application de la spectroscopie proche infrarouge pour l'analyse de produits agroalimentaires seront développés dans ce paragraphe.

Dans un premier temps, nous verrons que cette méthode est utilisée pour le contrôle de la qualité des produits agricoles puis dans un second temps, qu'elle est également la méthode d'analyse de référence pour l'achat des matières premières. La liste des exemples n'est pas exhaustive mais elle permet de se rendre compte de l'intérêt de la mesure infrarouge dans les industries agricoles et alimentaires.

- Analyse de la qualité de produits agricoles et alimentaires par SPIR

Les fabricants de tabac effectuent des mélanges de tabacs d'origines différentes pour avoir un goût et des qualités organoleptiques constantes. Ainsi la formulation du tabac nécessite un grand nombre d'analyses chimiques différentes. La méthode PIR a été choisie pour sa rapidité, son coût peu élevé et pour son caractère non destructeur de l'échantillon analysé[24].

Les glucides naturellement présents dans les feuilles de tabac sont le fructose, le glucose et le saccharose. De plus, les fabricants rajoutent du saccharose sous forme de sirop. Ainsi la concentration en sucres totaux dans le tabac varie de 0,2 % à 35 %. Les spectres sont acquis sur l'intervalle [1100 nm ; 2500 nm]. Ce modèle est ensuite testé sur 71 échantillons et l'erreur standard de prédiction (SEP) est de 1,36 % avec un coefficient de détermination de 0,986. Le dosage des glucides dans le tabac peut donc s'effectuer par spectroscopie PIR. Ainsi, la spectroscopie proche infrarouge permet de connaître rapidement les caractéristiques chimiques des tabacs.

Elle est également utilisée pour déterminer la qualité des fruits. La maturité de nombreux fruits est évaluée par spectroscopie PIR. On peut citer comme exemples : la pomme, la mandarine ou encore la pêche. Dans ces études, les auteurs s'intéressent à la

mesure du brix et Cho[123] mesure également les glucides (fructose, glucose, saccharose) (Tableau 8). Le but est de déterminer rapidement la maturité du fruit sans le détruire.

Tableau 8 ■ Analyse de fruits par SPIR. (- : non mesuré)

	Brix (en g / 100 g)	**Glucose** (g / 100 g)	**Fructose** (g / 100 g)	**Saccharose** (g / 100 g)
Mandarine	SEP = 0,339 Gamme de [9 ; 16]	-	-	-
Pomme	SEP = 0,6 Gamme de [12 ; 16]	SEP=0,28 Gamme de [2 ; 3,6]	SEP =0, 75 Gamme de [6 ; 9]	SEP=0,33 Gamme de [1,6 ; 4]
Pêche	SEP = 0,44	-	-	-

La spectroscopie PIR a de nombreuses applications pour le contrôle de la qualité des produits alimentaires. Elle permet le dosage de la teneur en glucides dans des produits très variés tels que les fruits secs[135], les jus de fruits[136], les moûts de fermentation, et le dosage de la teneur en eau, en protéines et en lipides dans des plats cuisinés[137].

- Achat de matières premières à partir de l'analyse PIR

La SPIR est la méthode utilisée pour doser des composés déterminant le prix d'achat de certains produits alimentaires tels que le lait et les céréales.

- ***Analyse du lait***

La composition du lait (Tableau 9) est très contrôlée à la fois sur le plan économique et sur le plan légal. Ainsi, la spectroscopie proche infrarouge permet :

- L'analyse du lait entrant dans l'usine pour le paiement des éleveurs.
- L'analyse du produit fini pour vérifier les spécifications légales.

Tableau 9 ■ Analyse du lait de chèvre par SPIR (d'après Frankhuizen[138]).

	Nombre d'échantillons	SEP (g / 100 g)	Gamme (g / 100 g)
Matières grasses	50	0,034	4-7
Protéines	50	0,038	2,1-2,8
Lactose	50	0,043	3,5-4,1

- ***Analyse à la réception dans les industries céréalières***

Depuis 20 ans, la spectroscopie PIR est appliquée en Australie, au Canada, en Europe et aux Etats-Unis pour contrôler les lots de blé et d'orge à la réception des silos. En 1980, la SPIR est utilisée comme la méthode officielle de l' « US Federal Grain Inspection Service » (FGIS). La réglementation américaine impose que les grains exportés soient analysés. L'utilisateur doit posséder un appareil d'analyse agréé et doit se servir des équations d'étalonnage établies par la FGIS. Les appareils sont standardisés selon des procédures agréées[139]. La spectroscopie permet de doser la teneur en eau et en protéines des grains mais elle permet également d'effectuer des classifications sur des critères qualitatifs. Ainsi, cette méthode permet de vérifier l'authenticité du riz ou des farines de blé.

4.2 Exemples d'application des méthodes de classification supervisées

Les applications des méthodes de classification supervisées sont nombreuses en chimie analytique[140,141,142] en biologie[143], en pharmacie[144] et en sciences alimentaires[145]. Ainsi, par exemple, ces méthodes sont utilisées dans les industries agricoles et alimentaires pour :

- L'authentification des huiles d'olive.
- La distinction entre le blé dur et le blé tendre[146].
- La détermination de la maturité, de la qualité organoleptique et de la capacité de stockage des fruits[147].
- La discrimination des huiles essentielles de citrus : mandarine, citron, orange et pamplemousse[148].

4.3 Utilisation de la SPIR dans l'industrie sucrière

La première étude (1983), utilisant la SPIR dans le milieu sucrier, a pour thème le dosage de l'azote dans les feuilles de canne à sucre[149,150]. Mais la SPIR s'applique à une large gamme de produits tout au long du procédé de fabrication, de la matière première au sucre.

Dans un premier temps, nous verrons les produits de sucreries analysés par des méthodes de spectroscopie infrarouge. Un bilan des principaux résultats sera établi dans un second temps.

4.3.1 Analyse de la matière première entrant dans l'usine

La sucrerie de canne doit connaître la composition chimique de la matière première pour déterminer le prix d'achat de celle-ci mais également pour estimer sa production de sucre et de co-produits. Ainsi, la spectroscopie proche infrarouge permet de mesurer sur la canne à sucre :

- La teneur en eau[151,152].
- Le brix[150, 153].
- La concentration en fructose[150] , en glucose[150] , en saccharose[151,150,] et en lignines[151, 150].

On remarque que les études concernant l'analyse de la betterave par la spectroscopie proche infrarouge sont plus récentes que celles étudiant la canne à sucre. En effet, l'échantillon (la râpure de betterave), moins homogène que le jus de canne, est plus difficile à analyser.

La plupart des composés analysés sur la canne à sucre sont également dosés dans la betterave par spectroscopie proche infrarouge. Ainsi, la SPIR a été utilisée pour évaluer :

- La teneur en eau[154].
- La teneur en saccharose[155,156,157 ,158, 159].
- La teneur en azote α-aminé[155,156,157].
- La teneur en sodium et en potassium[157].

4.3.2 Contrôle du procédé

La spectroscopie PIR permet également le contrôle en ligne du brix et de la polarisation. Elle s'applique tout au long du procédé (décrit en annexe 1) :

- pour l'analyse du jus de diffusion, ayant des teneurs en saccharose faibles, avant et après la carbonatation[160,161,162,163,164].
- et pour l'analyse des produits ayant des concentrations en saccharose élevées : sirops, liqueurs standards et masse cuite de l'atelier de cristallisation[165].

Marchetti[166,154] décrit les installations mises en place et les analyses réalisées sur 14 usines. De même, Vaccari et ses collaborateurs[167] ont montré comment l'analyse spectroscopique permet le contrôle du procédé sucrier.

4.3.3 Analyse des coproduits et des produits finis

Les coproduits de la sucrerie (mélasse, pulpes humides et pulpes séchées) ont également été dosés par spectroscopie proche infrarouge pour évaluer leur teneur en saccharose et leur brix[158].

Ames et ses collaborateurs[168] ont prédit la teneur en eau, la polarisation, les sucres réducteurs, les cendres et la couleur du sucre brut grâce à la spectroscopie infrarouge et visible. Ils ont également tenté d'obtenir des informations sur la taille et la forme des cristaux de sucre.

4.3.4 Synthèse des principaux résultats

Comme nous l'avons vu, un grand nombre d'études a été mené. Les résultats les plus récents sont présentés dans le Tableau 10 pour la canne à sucre et dans le Tableau 11 pour la betterave.

Tableau 10 ■ Analyse par SPIR dans la sucrerie de canne. (1) unité : g / 100 g. - : Donnée manquante dans l'article de référence.

Produit de sucrerie	Composé dosé	Référence	Gamme (1)	Nombre d'échantillons en étalonnage et en validation	Erreur standard d'étalonnage SEC(1)	Erreur standard de validation SEP(1)	R^2 en étalonnage et en validation
Jus de canne	Saccharose	150	0,7-13.	90 / 35	0,25	0,32	0,96 / 0,91
	Saccharose	153	12-19	350 / 350	0,06	0,19	0,998 / 0,98
	Brix	150	1,7-14	90 / 35	0,15	3,3	0,98 / 0,92
	Brix	153	12-22	500 / 500	0,15	0,16	0,988 / 0,97
Canne broyée	Brix	150	6-21	26 / 136	0,24	3,8	0,94 / 0,85
	Polarisation	150	5-19	26 / 136	0,42	3,9	0,93 / 0,88
	Matière sèche	150	16-38	26 / 136	0,86	4,7	0,88 / 0,82
	Fibre	150	7-14	26 / 136	0,41	6,0	0,85 / 0,8
Sucre	Teneur en eau	167	0,2-0,7	15 / 30	-	0,04	- / 0,961

On constate que les résultats obtenus sur le jus de canne sont meilleurs que ceux obtenus en analysant la canne broyée. En effet, la nature et l'homogénéité de l'échantillon influencent les erreurs de prédiction.

Tableau 11 ■ Analyse par SPIR dans la sucrerie de betterave. (1) unité : g / 100 g sauf pour le sodium et le potassium exprimés en 10^{-3}.mol.kg^{-1}. - : Donnée manquante dans l'article de référence.

Produit de sucrerie	Composé dosé	Référence	Gamme(1)	Nombre d'échantillons en étalonnage et en validation	Erreur standard d'étalonnage SEC(1)	Erreur standard de validation SEP(1)	R^2 en étalonnage et en validation
Râpure	Brix	158	15-18,5	146 / 36	-	0,19	- / 0,963
	Brix	156	-	75 / 75	0,24	0,27	- / 0,96
	Saccharose	158	15-18,5	146 / 36	-	0,1	-
	Saccharose	159	13-19,6	1000 / 4500	0,19	0,20	0,96 /
	Saccharose	156	-	75 / 75	0,37	0,4	- / 0,95
	Saccharose	157	-	175 / 75	-	0,25	-
	Azote α-aminé	158	-	146 / 36	-	1,7	- / 0,79
	Azote α-aminé	157	-	175 / 75	-	3,95	-
	Sodium	157	-	175 / 75	-	2,22	-
	Potassium	157	-	175 / 75	-	4,19	-
Cossette	Matière sèche	158	-	55 / 29	-	0,35	-
	Polarisation	158	-	55 / 29	-	0,30	-
Liquides	Polarisation	158	10-20	116 / 72	-	0,14	- / 0,999
Sirop et liqueur	Brix	158	13,5-20	116 / 72	-	0,12	- / 0,998
	Sucres	165	62-73	33	-	0,3	- / 0,995

La spectroscopie moyen infrarouge a été également appliquée en sucrerie. Elle a été utilisée pour détecter la présence de contaminations organiques (caoutchouc) et inorganiques (carbonates, sulfates) dans le sucre cristallisé[169], pour doser les jus de pressage de la canne à sucre[170] et pour contrôler les jus d'usine[171].

5 Bilan

Ce chapitre a montré la complexité chimique et la variabilité de l'échantillon. En effet, les variétés de betteraves comme les régions de culture sont nombreuses. De plus les différentes méthodes de dosage ont été présentées car elles constituent les mesures de référence pour la construction de modèles quantitatifs pour la spectroscopie.

Les applications de la spectroscopie proche infrarouge sont nombreuses en sucrerie de canne. Ainsi, la spectroscopie proche infrarouge sert de référence pour l'achat de la canne à sucre. La spectroscopie permet également d'effectuer des prédictions qualitatives du produit : par exemple, la prédiction de la résistance de la canne à sucre à certaines maladies[150]. Le but est de montrer que la SPIR et la chimiométrie sont également applicables à la sucrerie de betterave.

L'objectif de l'étude est triple. Tout d'abord, une méthode SPIR, complémentaire à la méthode réglementaire, va être définie. Ensuite, il faudra que cette méthode soit utilisable dans un contexte industriel multisite. Il faudra donc gérer un ensemble de spectromètres et une base de données spectrales. Enfin la qualité générale de la betterave sera évaluée par SPIR pour mettre en évidence le potentiel de la SPIR.

Chapitre 3

Détermination de la teneur en saccharose de la betterave sucrière par spectroscopie proche infrarouge

1 Objectifs

Les objectifs de ce chapitre sont de décrire les différentes étapes qui ont été nécessaire pour déterminer la teneur en saccharose de la betterave par SPIR : l'étude de faisabilité, l'optimisation de la méthode et la caractérisation du modèle mis en place au sein d'un laboratoire de référence. Enfin cette méthode spectrale sera évaluée en terme de répétabilité, reproductibilité et robustesse.

2 Etude de faisabilité

2.1 Choix de l'instrumentation

2.1.1 Objectifs

Cette pré-étude a été réalisée en 1999 par le LASIR avec le syndicat national des fabricants de sucre. L'objectif a été de choisir l'instrument le mieux adapté à l'analyse de la betterave à sucre. Quatre fabricants de spectromètres ont été contactés ce qui a permis le choix du type d'instrument et du mode d'échantillonnage.

2.1.2 Démarche

La démarche s'est effectuée en deux temps :

- Dans un premier temps, quatre instruments ont été évalués. Leurs caractéristiques techniques et la répétabilité spectrale ont été appréciées. De plus, leurs capacités à l'analyse quantitative ont été testées par la construction de modèles avec 50 échantillons (30 pour l'étalonnage et 20 pour la validation) conservés par congélation car l'étude s'est déroulée hors de la période de récolte.

- Dans un second temps, deux instruments ayant les meilleurs résultats ont été étudiés sur des échantillons frais. 1696 échantillons (1016 pour l'étalonnage et 680 pour la validation) ont été analysés sur les deux instruments et une application analytique a été développée.

2.1.3 Résultats et discussion

- Résultats et discussion concernant la première phase de l'étude de faisabilité

- ***Caractéristiques générales des instruments***

Tableau 12 ■ Caractéristiques générales des quatre instruments testés

Constructeur Instrument	**Foss NIRsystem** NIRS6500	**Bran & Luebbe** Infralyser 2000	**Bruker** Vector 22N-I	**Nicolet** Avatar 360
Type d'instrument	Dispersif à réseau holographique	Dispersif à filtres	Multiplexé à transformée de Fourier	Multiplexé à transformée de Fourier
Technique d'échantillonnage	Réflexion	Transflexion	Réflexion	Réflexion totale atténuée
Gamme spectrale	400 -2500 nm	19 filtres PIR	1100 - 2500 nm	4000 – 400 cm^{-1}
Source	Tungstène	Tungstène	Tungstène	Globar
Détecteur	PbS et Si	PbS	PbS	DTGS
Vitesse d'acquisition	1,8 $spectres.s^{-1}$	2 $spectres.min^{-1}$	1,7 $spectres.s^{-1}$	1 $spectre.s^{-1}$
Masse de râpure utilisée	100 g	20 g	50 g	- de 10 g
Système de moyennage de l'échantillon	Déplacement vertical	Aucun	Coupelle tournante	6 réflexions
Acquisition de la référence photométrique	Automatique	Automatique	Automatique	Manuelle

On constate que les fabricants ont proposé des instruments et des méthodes d'échantillonnages très différents. Les trois types d'instruments PIR (dispersif à filtres, dispersif à réseaux et multiplexé à transformée de Fourier) sont représentés ainsi qu'un spectromètre dans le domaine du moyen infrarouge. Le Tableau 12 met en évidence les limites pour chacun des instruments. Le spectromètre moyen infrarouge (Nicolet) a deux inconvénients. Le premier est l'acquisition manuelle de la référence photométrique. Le second est lié à l'échantillonnage. Il est apparu difficile d'analyser de la betterave par ATR car le positionnement de l'échantillon a été problématique. L'inconvénient de l'instrument Bran et Luebbe est lié aux filtres qui permettent d'analyser uniquement 19 longueurs d'onde : 2336, 2348, 2310, 2230, 2208, 2180, 2139, 2080, 1982, 1818, 1778, 2100, 1759, 1940, 1734, 1722, 1680, 1480 et 1445 nm. Il peut être intéressant d'avoir le spectre complet pour la détermination de plusieurs critères simultanément. De plus, cet instrument n'a pas de système de moyennage de l'échantillon ni de procédure de tests et le temps d'analyse est plus long. Enfin, les deux autres instruments Foss et Bruker n'ont pas d'inconvénients majeurs.

- ***Performances à l'analyse de la betterave***

Tableau 13 ■ Performances des instruments pour l'analyse d'échantillons de betterave râpée.

Constructeur	Foss Electric	Bran & Luebbe	Bruker	Nicolet
SEC (g / 100 g)	0,12	0,16	0,11	0,08
SEP (g / 100 g)	0,22	0,25	0,28	0,28
Ecart type de répétabilité (g / 100 g) (1)	0,05	0,04	0,04	0,09

(1) sur un échantillon analysé 20 fois

L'analyse quantitative a été possible avec tous les instruments. D'après le calcul du SEP limite qui vaut 0,32 (=0,22*2,12^{0,5}), les SEP obtenus sur les quatre instruments ne sont pas différents. On constate cependant que la répétabilité du spectromètre Nicolet est plus grande que celles des autres instruments. Cependant à cause des inconvénients précédemment décrits, l'instrument à filtre et le spectromètre moyen infrarouge ont été rejetés. Comme cette pré-étude rapide a été effectuée sur des échantillons congelés, il a été

décidé de conserver deux instruments pour une étude plus approfondie : l'instrument FOSS et l'instrument Bruker.

- Résultats et discussion de la deuxième phase de sélection

Tableau 14 ■ Comparaison de deux instruments sur 680 échantillons en validation.

Marque **Modèle**	**BRUKER** Vector 22N-I	**FOSS** 6500
Nombre de spectres	680	680
SEP (g / 100g)	0,16	0,10
Biais (g / 100g)	0,01	0,01
Pente	1,01	1,01

Au cours de la campagne sucrière de 1999, les deux instruments ont été utilisés. 1016 échantillons ont été analysés sur les deux instruments pour développer des modèles dont les résultats de la validation sur 680 échantillons sont présentés dans le Tableau 14. Les SEP obtenus sont plus petits que ceux obtenus lors de la pré-étude. L'étalonnage est développé sur un plus grand nombre d'échantillons et il semble donc plus précis.

Concernant la différence entre les deux instruments, on constate que les résultats de la pré-étude sont confirmés. Les résultats de l'instrument FOSS sont plus satisfaisants que ceux obtenus avec l'instrument Bruker comme le montre le résultat du test de Fisher ($F_{calculé}$ = 2,56 et $F_{critique}$ = 1,13). Nos résultats sont en accord avec la littérature qui considère que les systèmes dispersifs à réseaux donnent de meilleurs résultats que les appareils multiplexés à transformée de Fourier[98] dans les études quantitatives en SPIR.

2.2 Choix de la méthode chimique de référence

2.2.1 Objectifs

Il existe depuis 1964 une méthode légale pour le dosage du saccharose de la betterave. Avant de commencer la construction d'une base de données contenant les spectres et les concentrations en saccharose d'un grand nombre d'échantillons, il faut s'assurer que la méthode de référence est la mieux adaptée pour la construction d'une application quantitative. En effet, la qualité du modèle dépend non seulement de la qualité de l'instrument utilisé mais aussi de la méthode de référence.

2.2.2 Démarche

Trois méthodes de dosage du saccharose ont été utilisées : la polarimétrie, le dosage par CLHP sur résine échangeuse d'ions et le dosage enzymatique. 100 échantillons ont été analysés avec les trois méthodes. Pour des raisons pratiques, 100 échantillons ont d'abord été analysés par polarimétrie et par HPLC puis 100 autres échantillons ont été analysés par polarimétrie et par dosage enzymatique.

Dans un premier temps, 20 échantillons ont été analysés trois fois pour évaluer l'écart-type de répétabilité des trois méthodes de dosage. L'écart-type de répétabilité est calculé d'après les recommandations de Feinberg[31] selon les équations suivantes :

Équation 77 $$SCEr = \sum_i \sum_j (x_{i,j} - \bar{x}_i)^2$$

Avec $x_{i,j}$ valeur pour le $i^{ème}$ échantillon et la $j^{ème}$ répétition et $\bar{x}_i$ la moyenne pour l'échantillon i.

Équation 78 $$s_r = \sqrt{\frac{SCEr}{N-n}}$$

Avec N : nombre total de mesures, n : le nombre d'échantillons, SCEr : Somme des carrés des écarts, s_r écart-type de répétabilité.

Dans un second temps, des modèles quantitatifs ont été construits avec les spectres proche infrarouge et les données des différentes méthodes chimiques. L'objectif est de déterminer quel dosage chimique va permettre de prédire au mieux la teneur en saccharose de la betterave avec les mêmes données spectrales. 60 échantillons ont été utilisés pour l'étalonnage et 40 pour la validation.

2.2.3 Résultats et discussion

- ***Evaluation de répétabilité des méthodes de référence***

Tableau 15 ■ Comparaison de la répétabilité des trois méthodes chimiques de référence (unité g / 100g)

Méthode	Polarimétrie	CLHP	Dosage enzymatique
Répétabilité (s_r)	0,06	0,41	0,15

Comparée à la CLHP et au dosage enzymatique, la polarimétrie est la méthode la plus répétable. La répétabilité obtenue par CLHP est en accord avec celle trouvée par Cadet[10]. De même celle de la polarimétrie correspond aux exigences réglementaires[112]. Nous avons ainsi vérifiés que la polarimétrie est la méthode la plus répétable.

- ***Résultats des analyses quantitatives développées avec les différentes méthodes chimiques***

Tableau 16 ■ Comparaison des modèles construits avec les trois méthodes chimiques et les spectres proche infrarouge. (Unité g / 100g)

Méthode	Polarimétrie	CLHP
SEP	0,15	0,91
Pente	0,95	1,22
Biais	-0,03	0,04
SEP(C)	0,15	0,92
R^2	0,98	0,62

Méthode	Polarimétrie	Dosage enzymatique
SEP	0,16	0,46
Pente	0,99	0,68
Biais	-0,01	0,03
SEP(C)	0,16	0,47
R^2	0,98	0,32

On constate que la polarimétrie est la seule méthode permettant d'obtenir un SEP faible et un R^2 proche de 1 (Tableau 16).

D'un point de vue qualitatif, la CLHP a un avantage car elle permet d'identifier les différents sucres. Cependant, au niveau quantitatif, cette méthode, comme le dosage enzymatique, est moins répétable que la polarimétrie et donc l'erreur standard de prédiction est plus grande.

Comme la construction de modèles est satisfaisante avec les données de la polarimétrie, cette méthode chimique est conservée comme méthode de référence. Ainsi le modèle développé correspond bien à la méthode légale et il pourra avoir une équivalence entre les résultats de la SPIR et ceux de la méthode actuelle. On peut remarquer qu'il a été montré que les concentrations en glucose et fructose avaient des effets négligeables sur la valeur polarimétrique[172,10].

3 Optimisation de la modélisation

3.1 Construction de la base de données

3.1.1 Protocole de récolte des échantillons

Chaque échantillon constitué d'environ 15 kg de betteraves entières a été mis en sac convenablement fermé, étiqueté et stocké à température ambiante ou en chambre froide. Afin de couvrir la variabilité tant géographique que temporelle, les échantillons ont été prélevés dans diverses usines de la zone betteravière française tout au long de la campagne.

La variabilité entre les années est également intégrée par l'étude qui s'est déroulée de 1999 à 2001. Les dix-neuf usines concernées par l'étude ont été : Colleville, Etrepagny, Cagny, Arcis, Sillery, Guignicourt, Lillers, Eppeville, Roye, Dompierre, Origny, Sainte Emilie, Chevrière, Bucy, Vic sur Aisne, Villenoy, Pithiviers, Corbeilles, Artenay. De plus, une vingtaine de variétés de betteraves ont été analysées.

3.1.2 Données spectrales utilisées

- ***Spectres d'échantillons de betterave***

La figure suivante présente les spectres de 525 échantillons de betterave (lot de validation de la campagne 2001). On constate que les spectres sont très semblables. La variation importante sur le domaine du visible s'explique par le changement de couleur de l'échantillon de betterave lié à l'oxydation.

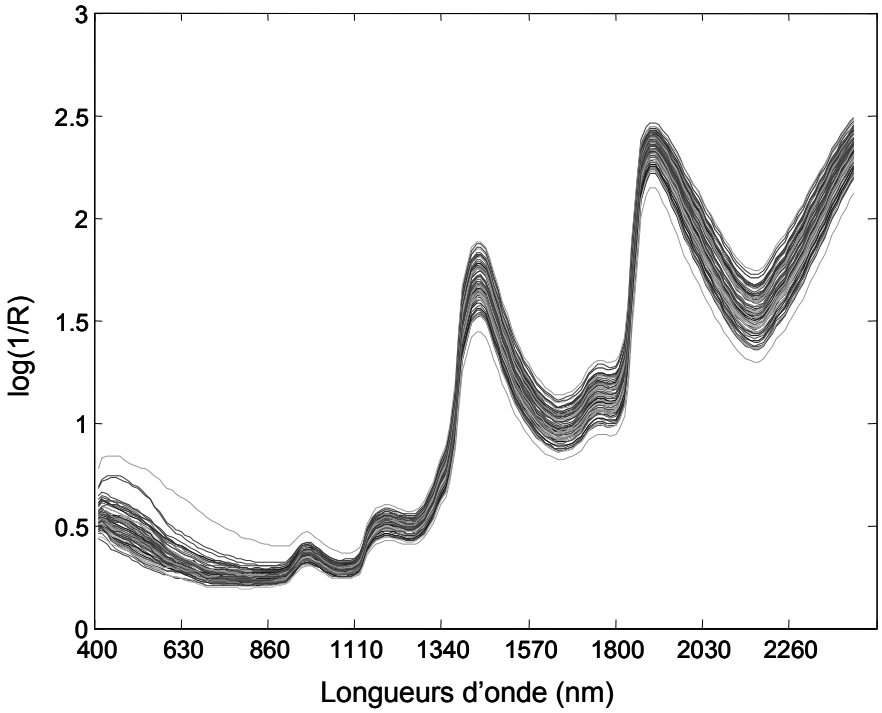


Figure 29 ■ Spectres proche infrarouge et visible des 525 échantillons du lot de validation de la campagne 2001.

- ***Visualisation des données***

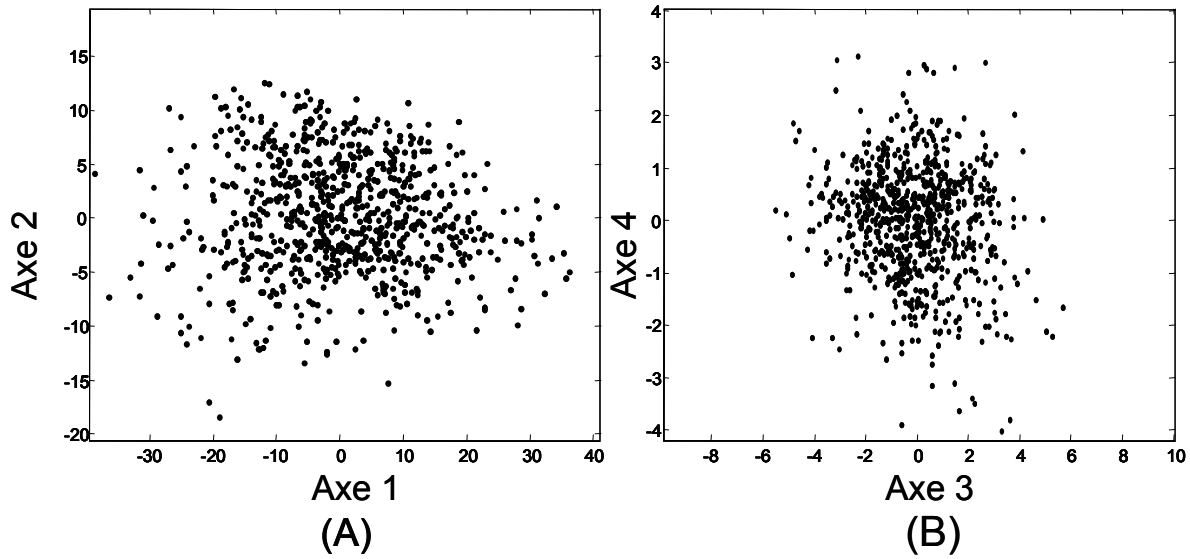


Figure 30 ■ Analyse en composantes principales des 2210 échantillons du lot d'étalonnage. (A) plan 1-2 représentant 96,9 % de la variance totale (B) plan 3-4 représentant 2,7 % de la variance totale.

On constate une répartition homogène des échantillons sur les résultats de l'ACP (Figure 30). Il n'y a pas de groupes distincts. Il n'existe donc pas de différences spectrales majeures entre les variétés ou entre les années de récolte. On peut donc penser qu'il sera possible de développer un modèle unique valable pour l'ensemble des échantillons analysés, c'est-à-dire applicable pour toutes les localisations géographiques françaises.

- ***Interprétation des vecteurs propres***

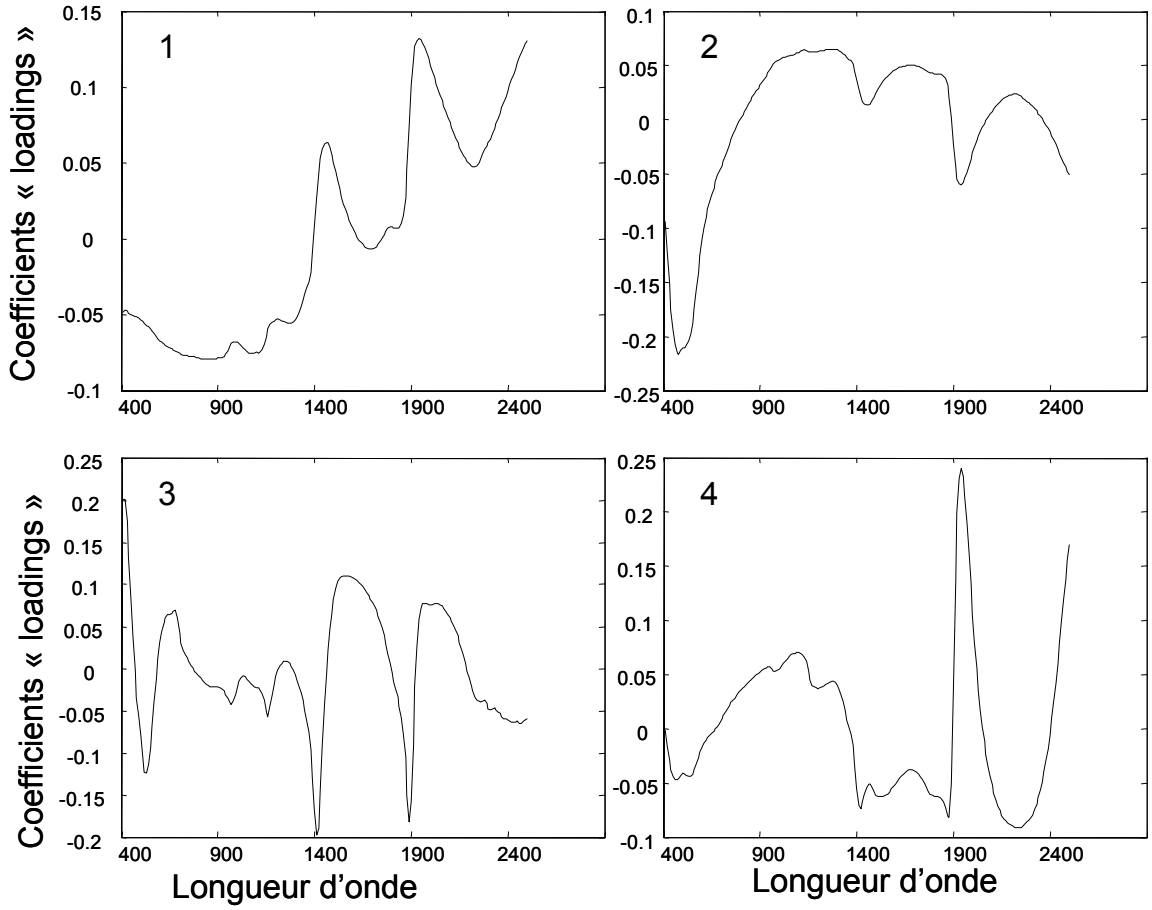


Figure 31 ■ Quatre premiers vecteurs propres obtenus sur un lot de 2210 spectres de betterave. (1), (2), (3) et (4) représentent les numéros des composantes principales.

Les vecteurs propres (Figure 31) permettent de mettre en évidence les longueurs d'onde les plus influentes. Le premier vecteur propre a la même allure que le spectre moyen. On remarque l'importance des pics de l'eau centrés sur 1450 et 1950 nm. Le deuxième s'explique par des longueurs du visible, c'est-à-dire la couleur de l'échantillon qui est caractéristique du degré d'oxydation. La troisième s'interprète par les deux bandes de l'eau

(et des glucides vers 1450 nm). Enfin la quatrième s'explique par la deuxième bande de l'eau et par les longueurs d'onde supérieures à 2200 nm (contribution des glucides).

3.1.3 Données chimiques utilisées

La Figure 32 montre la distribution des concentrations de référence. Le test d'aplatissement (Kurtosis[173]) a pour valeur 3,17 et le test d'épaulement (Skewness[173]) est de 0,01. On peut noter que pour une distribution normale, le kurtosis vaut 3 et le skewness 0. Un test statistique de Kolmogorov-Smirnov[173] permet de conclure que la distribution peut être considérée comme normale.

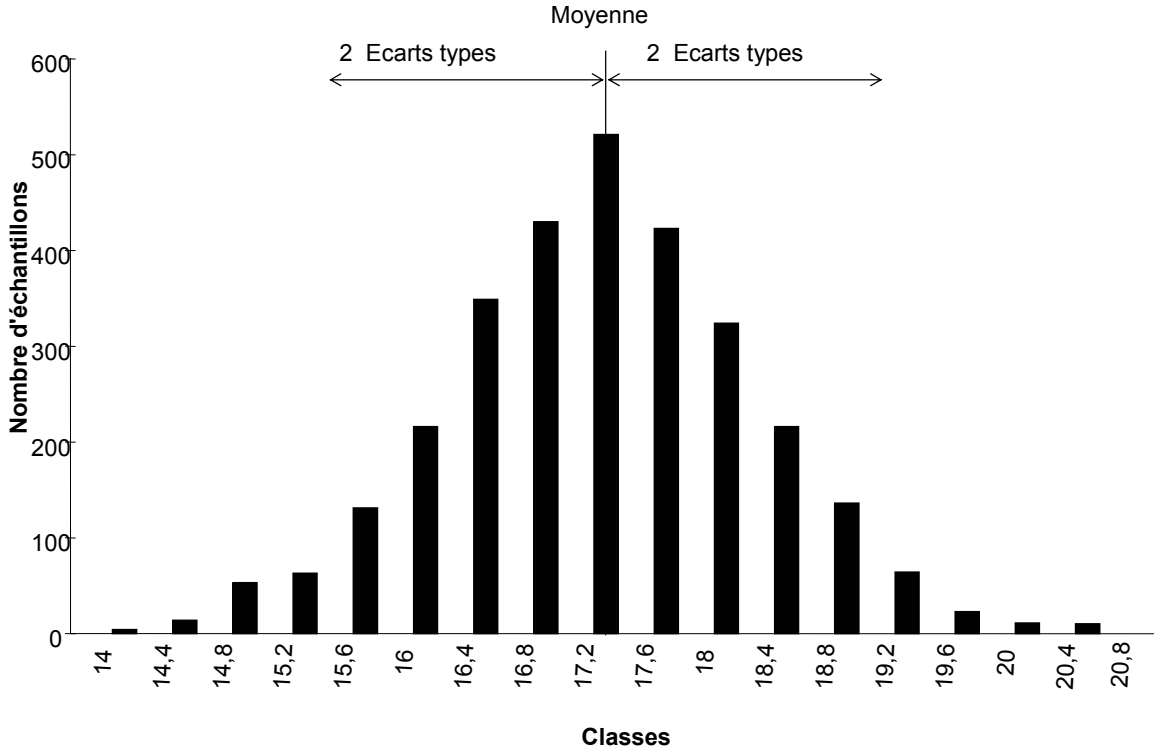


Figure 32 ■ Histogramme des valeurs de référence sur 2210 échantillons du lot d'étalonnage.

La moyenne des concentrations est de 17,54 g / 100 g, l'écart-type de 1,04 g / 100 g de betterave et la gamme des concentrations s'étend de 14,38 à 20,62 g de saccharose pour 100 g de betterave.

3.2 Influence des paramètres de la modélisation

3.2.1 Objectifs

L'objectif est la détermination de la meilleure méthode de modélisation. Il s'agit de trouver la méthode chimiométrique donnant l'erreur standard de prédiction la plus faible. Dans cette partie, l'influence du domaine spectral, des méthodes de prétraitement des spectres et des méthodes de régression sur l'erreur de prédiction est étudiée.

3.2.2 Protocole

- ***Spectres utilisés***

54 modèles différents sont construits avec le lot d'étalonnage constitué de 2210 échantillons analysés au cours des trois années de 1999 à 2001. Le SEP est calculé sur le lot de validation de 525 échantillons de 2001 indépendants de ceux utilisés pour l'étalonnage.

- ***Construction des modèles***

Tous les modèles sont construits avec WINISI. Lors de l'étalonnage, une validation croisée (10 groupes) est réalisée afin de déterminer le nombre de composantes à conserver pour créer le modèle.

Les paramètres variants sont :

- La méthode de régression : « Partial Least Squares » (PLS), PLS modifiée[174] (mPLS) ou « Principal Component Regression » (PCR).
- Les domaines spectraux : deux domaines sont évalués ([400, 2498 nm] ou [1100, 2498 nm]).
- Les prétraitements spectraux : neuf combinaisons entre les algorithmes (aucun, « Standard Normal Variate » (SNV) ou « Standard Normal Variate detrending » (SNVD)) et une dérivée (aucune, d'ordre 1 ou 2 notée D1 et D2) sont testées.

Remarque : La méthode MLR a été également utilisée. De même, MSC a été testée mais a donné des résultats similaires à ceux de SNV. Ces résultats ne sont pas présentés dans la thèse mais ils ont été publiés[177].

3.2.3 Résultats et discussions

- ***Effets des prétraitements sur les spectres***

La Figure 33 montre les effets des différents prétraitements sur les spectres proche infrarouge de betterave. On remarque les effets des prétraitements sur la ligne de base et sur la dispersion des spectres.

Il n'est pas possible de savoir a priori lequel des prétraitements donnera les meilleurs résultats. La solution est de tester de manière systématique les différents prétraitements et de sélectionner celui donnant l'erreur standard de prédiction la plus faible[175].

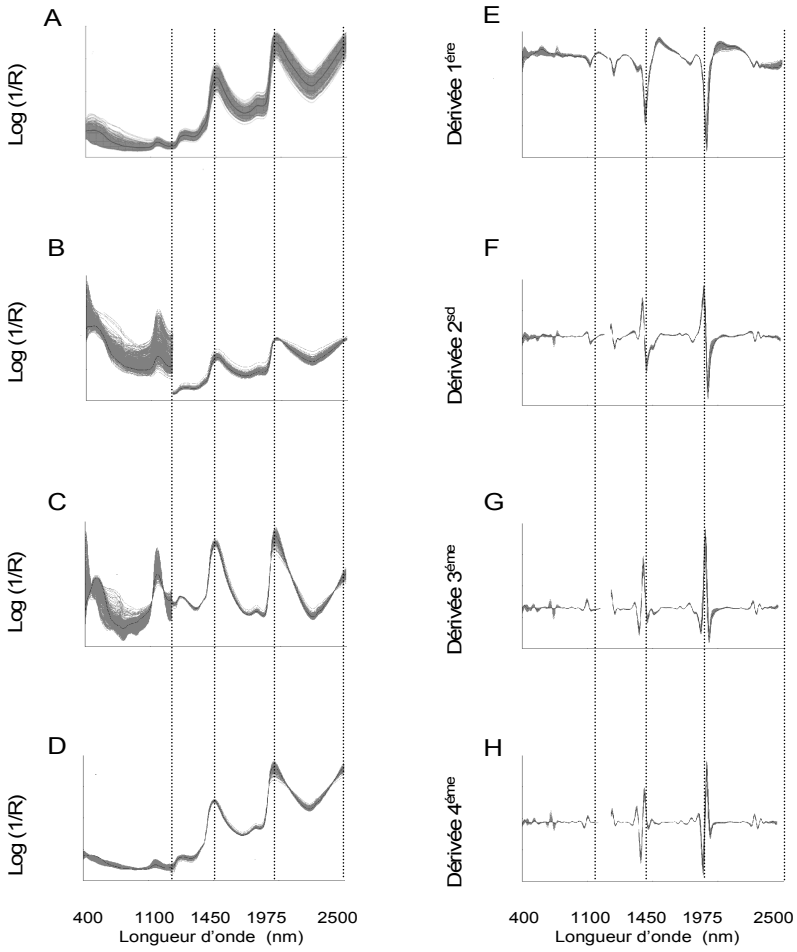


Figure 33 ■ Influence des prétraitements sur un lot de 525 spectres PIR de betterave A : Spectres bruts ; B : prétraitement SNV ; C : prétraitement SNV+D ; D : prétraitement MSC ; E : dérivée (1,5,5) ; F : dérivée (2,8,6) ; G : dérivée (3,10,10) ; H : dérivée (4,10,10)

- ***Comparaison des modèles***

Les principaux résultats sont présentés dans les deux tableaux suivants. On constate que les biais varient de -0,08 à 0,02 g / 100 g et que les SEP(C) sont compris entre 0,10 et 0,35 g / 100 g. Les biais significativement différents de zéro sont obtenus uniquement avec la méthode PCR. Seule la méthode mPLS permet d'obtenir un SEP(C) de 0,10 g / 100 g.

Tableau 17 ■ Comparaison des biais des 54 modèles. En gris, les biais significativement différents de zéro d'après l'ANOVA et le test LSD. (A) domaine spectral [400, 2498 nm] (B) domaine spectral [1100, 2498 nm].

Légende : D1 et D2 dérivées première et seconde, SNV « standard normal variate », SNVD « standard normal variate and detrend », PLS « partial least squares », mPLS « modified PLS » et PCR « principal component regression »

(A)

	Aucun	D1	D2	SNV	SNV+D1	SNV+D2	SNVD	SNVD+D1	SNVD+D2
MPLS	-0,01	0,00	0,00	-0,01	0,00	-0,01	0,00	0,00	-0,01
PLS	-0,02	-0,01	-0,00	-0,01	-0,00	0,00	-0,02	0,00	0,00
PCR	**-0,08**	**-0,05**	-0,01	**-0,06**	**-0,05**	-0,01	**-0,05**	**-0,05**	-0,01

(B)

	Aucun	D1	D2	SNV	SNV+D1	SNV+D2	SNVD	SNVD+D1	SNVD+D2
MPLS	-0,01	0,00	0,00	-0,01	0,00	0,00	-0,01	0,01	0,00
PLS	0,00	0,00	0,00	-0,01	0,00	0,00	0,00	0,00	0,00
PCR	-0,01	**-0,02**	0,01	0,00	0,01	**-0,02**	**0,02**	0,01	**-0,02**

Tableau 18 ■ Comparaison des SEP(C) des 54 modèles. En gris, les SEP(C) non significativement différents. (A) domaine spectral [400, 2498 nm] (B) domaine spectral [1100, 2498 nm]

(A)

	Aucun	D1	D2	SNV	SNV+D1	SNV+D2	SNVD	SNVD+D1	SNVD+D2
MPLS	0,12	**0,10**	**0,10**	0,12	**0,10**	**0,10**	0,13	**0,10**	**0,10**
PLS	0,15	0,13	0,12	0,17	0,15	0,16	0,20	0,15	0,15
PCR	0,21	0,20	0,15	0,35	0,27	0,32	0,35	0,27	0,32

(B)

	Aucun	D1	D2	SNV	SNV+D1	SNV+D2	SNVD	SNVD+D1	SNVD+D2
MPLS	0,12	**0,10**	**0,10**	0,12	**0,10**	**0,10**	0,12	**0,10**	**0,10**
PLS	0,15	0,14	0,16	0,14	0,14	0,16	0,15	0,14	0,16
PCR	0,15	0,17	0,16	0,16	0,18	0,16	0,16	0,18	0,16

- ***Gamme spectrale***

La PCR donne de meilleurs résultats en terme de SEP(C) et de biais quand le domaine visible n'est pas utilisé. Par contre, pour les méthodes mPLS et PLS, il n'y a pas de différence entre les modèles construits avec ou sans le domaine [400, 1100 nm]. Ce domaine ne contient apparemment aucune information relative à la détermination de la teneur en saccharose.

- ***Méthode de régression***

La PCR est la méthode donnant les SEP les plus grands. La régression par PLS donne des SEP plus grands que la méthode mPLS. La modification de PLS brevetée par Shenk et Westerhaus[174] améliore de façon significative la méthode. Elle est décrite par ses auteurs de la façon suivante : « Les résidus spectraux et les résidus des concentrations sont normalisés (centrés et réduits) à chaque itération de la PLS »[174].

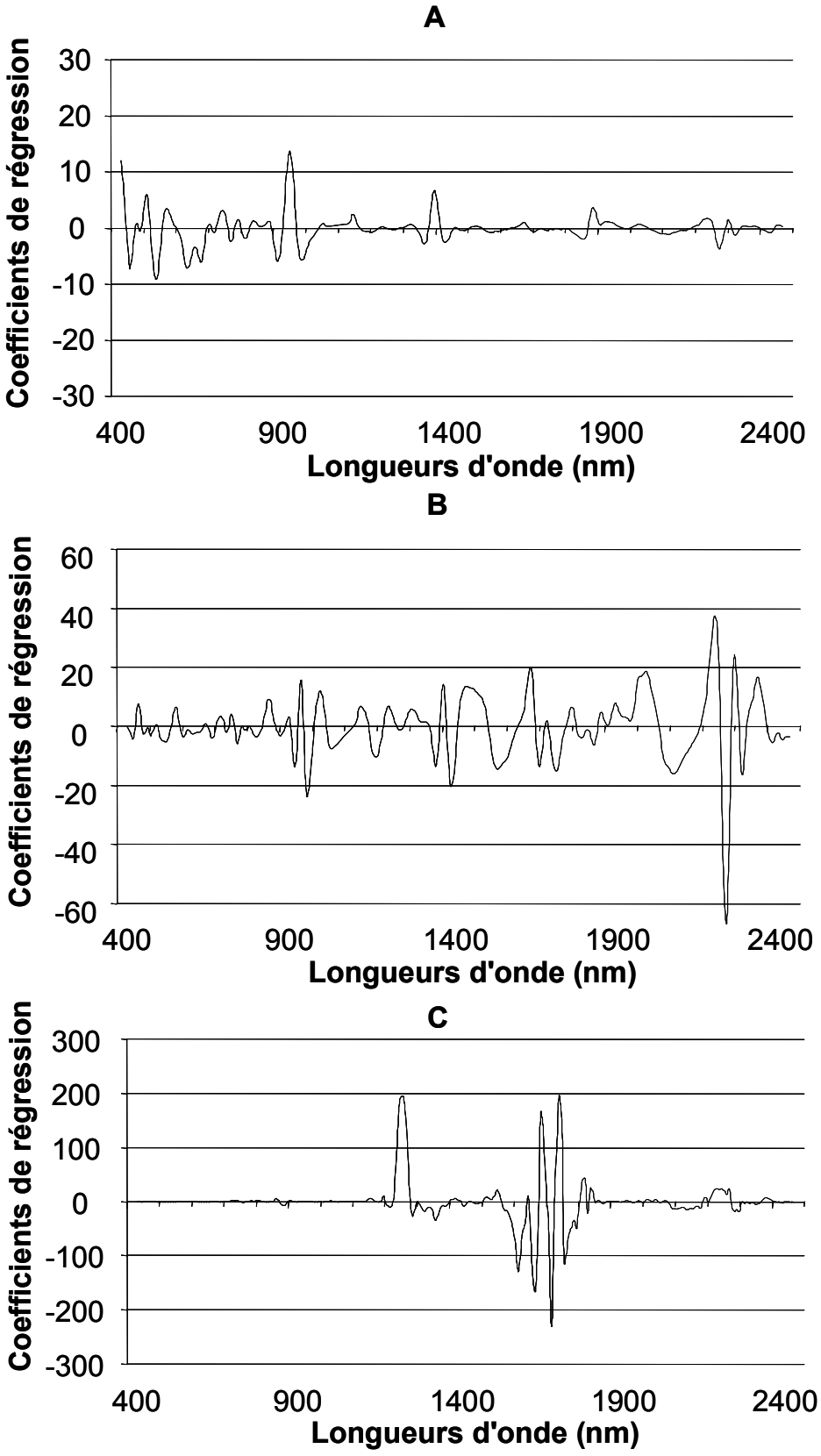


Figure 34 ■ Coefficients de régression obtenus avec les trois méthodes : (A) PCR, (B) PLS, (C) mPLS.

La différence entre les méthodes spectrales peut s'expliquer par l'analyse des coefficients de régression (Figure 34). La méthode PCR donne un poids important aux longueurs d'onde du visible car une grande partie de la variabilité spectrale provient de ce domaine. Ce domaine correspond à la couleur de l'échantillon, il témoigne du degré d'oxydation de l'échantillon mais il n'est pas directement relié à la teneur en saccharose. C'est pourquoi cette méthode donne les résultats les moins précis. Par contre, on constate que la méthode mPLS a des coefficients proches de zéro pour l'ensemble de la gamme [400,1100 nm]. De plus cette méthode a des coefficients très importants pour un nombre limité de longueurs d'onde.

- ***Prétraitements***

On remarque que pour la méthode PCR, l'utilisation des prétraitements n'améliore pas la prédiction (excepté la dérivée seconde).

L'utilisation de la dérivée et des prétraitements permet de réduire l'erreur de prédiction de façon significative pour la méthode mPLS. En effet, les prétraitements suppriment la dérive de la ligne de base et la dérivée exalte les informations spectrales. Pour la méthode mPLS, les combinaisons de prétraitements ne sont pas significativement différents : SNV+(1,4,4), SNV+(2,8,6), SNVD+(1,4,4) et SNVD+(2,8,6).

- ***Choix du meilleur modèle***

Plusieurs modèles donnent des SEP qui ne sont pas significativement différents. On constate qu'ils sont tous construits avec la méthode mPLS et qu'ils utilisent une dérivée (d'ordre 1 ou 2) et un algorithme SNV ou SNVD. Nous choisissons le modèle mPLS utilisant le domaine [1100, 2498 nm] et des spectres prétraités par SNVD et la dérivée seconde.

4 Caractéristiques du modèle optimal

Le modèle est développé à partir d'une base de données contenant 1016 échantillons analysés en 1999, 669 en 2000 et 525 en 2001. La gamme d'étalonnage, les résultats de l'étalonnage et de la validation sont décrits et comparés aux résultats des campagnes précédentes. Ensuite, la linéarité, la répartition des écarts entre la valeur PIR et celle de la polarimétrie seront observées. Enfin, les longueurs d'onde utilisées par le modèle seront présentées.

- ***Résultats de la phase d'étalonnage***

La gamme d'étalonnage est large (14 – 20 g / 100 g). Le SEC est faible, le R^2 est grand (Tableau 19).

Tableau 19 ■ Caractéristiques du modèle mis en place en 2001

Paramètres de modélisation	
Nombre d'échantillons	2210
Validation croisée	10 groupes
Résultat de l'étalonnage	
SEC (g / 100 g)	0,09
R^2	0,99

- ***Résultats de la phase de validation sur 525 échantillons de la campagne 2001***

Au cours de la campagne 2001, le SEP obtenu est de 0,10 g / 100 g (Tableau 20). L'erreur de prédiction est faible et correspond bien au cahier des charges industrielles (SEP inférieur à 0,15). La teneur en saccharose des betteraves peut être prédite par spectroscopie proche infrarouge de manière satisfaisante. De plus, le rapport écart-type des valeurs de référence divisé par le SEP, noté RPD (ratio prediction deviation) est de 10,72. On peut donc conclure que la méthode SPIR est comparable à la méthode de référence[176].

Tableau 20 ■ Résultats de validation

	campagne 2001
Nombre d'échantillons	525
SEP (g / 100 g)	0,10
Biais (g / 100 g)	-0,01
SEP (C) (g / 100 g)	0,10
Pente	1,00
R^2	0,99

- ***Linéarité du modèle***

La Figure 35 représente les valeurs prédites par le modèle 2001 en fonction des valeurs de référence (polarimétrie). La linéarité du modèle est bien vérifiée (Figure 35). On constate que le coefficient de détermination est proche de 1, que le biais est non significatif, c'est-à-dire qu'il n'y a donc pas d'erreur systématique entre les valeurs PIR et polarimétrique.

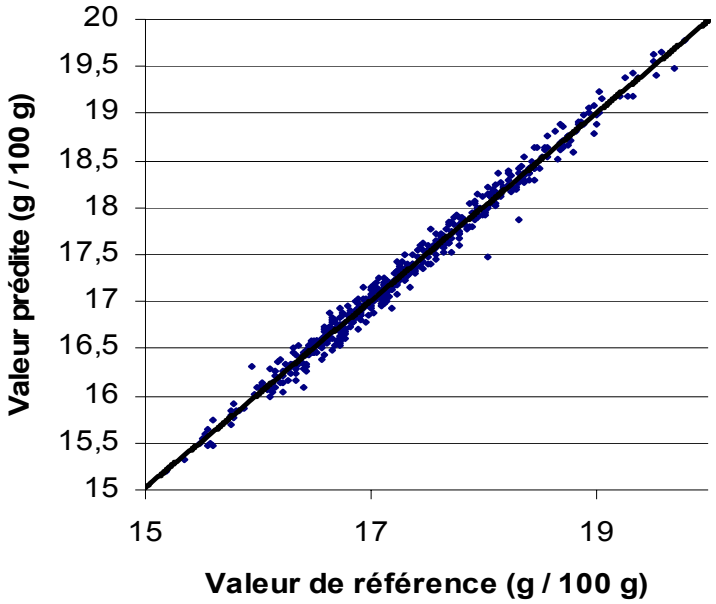


Figure 35 ■ Linéarité du modèle 2001 (Validation avec 525 échantillons inconnus du modèle).

- ***Répartition des résidus***

La Figure 36 représente les écarts entre la valeur SPIR et la polarimétrie pour l'ensemble des échantillons du lot de validation. La moyenne des écarts (valeur PIR – valeur polarimétrie) est proche de zéro. L'erreur systématique entre la SPIR et la polarimétrie est faible. La plupart des écarts se situent dans l'intervalle de confiance à 95 %, c'est-à-dire entre [- 2 SEP ; 2 SEP]. Seulement 14 échantillons se trouvent hors de cet intervalle (Figure 36).

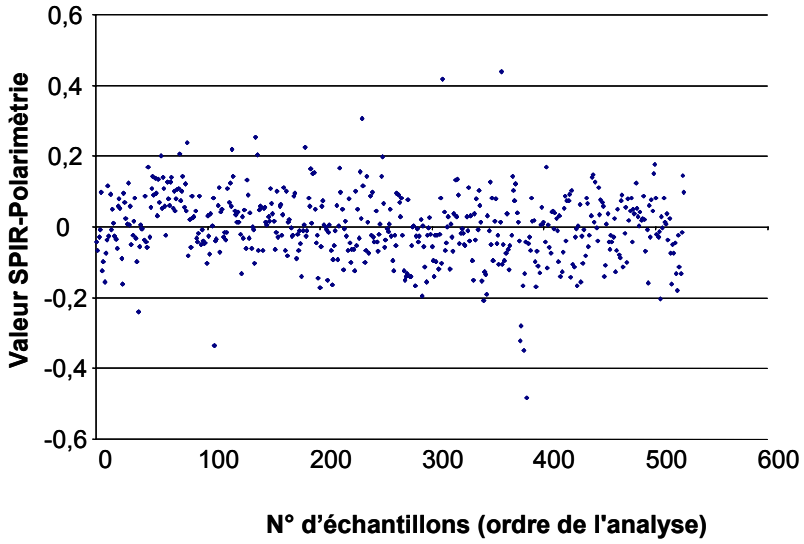


Figure 36 ■ Résidus du modèle 2001 sur un lot de validation de 525 échantillons.

- ***Comparaison avec les résultats des campagnes précédentes***

Le Tableau 21 montre les erreurs standard de prédiction obtenues au cours des trois années de l'étude. La Figure 37 compare les coefficients des modèles obtenus à partir des échantillons de 1999 ou à partir des échantillons des trois années. Dans cette partie, les échantillons du lot de validations sont sélectionnés au hasard parmi l'ensemble des analyses réalisées.

Tableau 21 ■ Résultats des campagnes précédentes.

Lot d'étalonnage	**1999**	**2000**	**2001**	**1999-2000**	**1999-2000-2001**
Lot de validation	1999	2000	2001	2000	2001
Nombre d'échantillons de l'étalonnage	1016	669	525	1016+669	1016+669+525
Nombre d'échantillons de la validation (inconnus du modèle)	1026	640	525	640	525
SEP (g / 100 g)	0,10	0,11	0,09	0,11	0,10
Biais	0,00	0,01	0,00	0,00	0,01
SEP(C)	0,10	0,11	0 ,09	0,11	0, 10

Les résultats des deux précédentes campagnes[177] se trouvent vérifiés (SEP = 0,1 g / 100 g au cours de la campagne 2001 (Tableau 21). Les résultats sont stables d'une campagne sur l'autre : le modèle semble pouvoir être pérennisé.

- ***Coefficients de régression du modèle***

Le modèle de 1999/2001 utilise les mêmes longueurs d'onde que le modèle de 1999 (Figure 37). Il y a un ajustement des coefficients entre ces deux modèles. Les principales longueurs d'onde sont 1236 nm, 1348 nm, 1593-1804 nm et 2228-2268 nm. Ces longueurs d'onde sont attribuées au saccharose par la littérature[178,179,180,181]. L'information spectrale utilisée pour prédire la teneur en saccharose reste inchangée d'une campagne à l'autre.

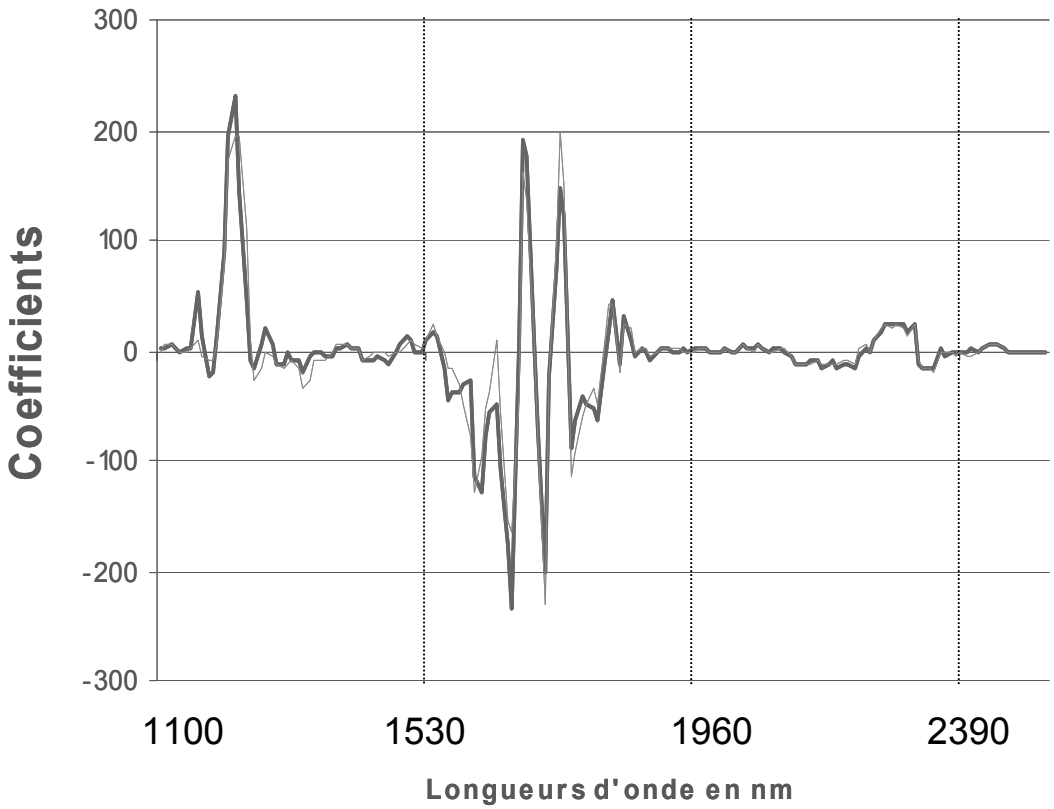


Figure 37 ■ Coefficients de régression des modèles construits avec les lots d'étalonnage de 1999 et de 2001. ___ modèle de 1999 et ___ modèle de 1999/2001.

Lors d'applications quantitatives, de nombreux auteurs se sont intéressés au dosage des glucides et grâce à des techniques d'analyses multivariées, ils ont identifié des bandes caractéristiques des glucides. Les principaux résultats sont présentés dans la Figure 38. On constate que les résultats de l'étude sont concordants avec ceux de la littérature.

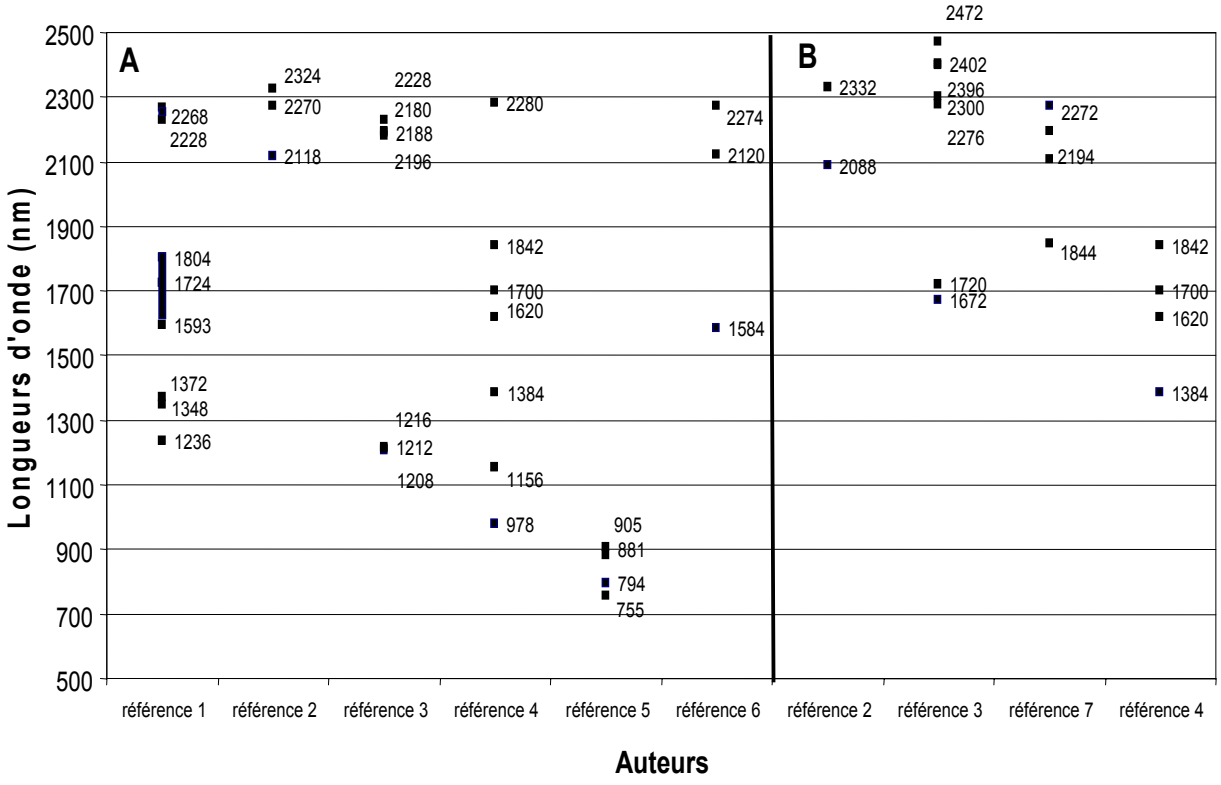


Figure 38 ■ Attribution des bandes des glucides par des méthodes multivariées. Cette figure présente les longueurs d'onde utilisées pour le dosage du saccharose (A) et le dosage des sucres totaux (B).

Référence 1 : Roggo *et al.*[177], betterave à sucre, réflexion.

Référence 2 : Robert *et al.*[178], solution aqueuse de saccharose, transmission.

Référence 3 : Cho *et al.*[180], pomme, réflexion.

Référence 4 : Salgo *et al.*[156], betterave à sucre, mesure en réflexion.

Référence 5 : Miyamoto et Kitano[179], mandarine, transmission.

Référence 6 : Rambla *et al.*[181], betterave à sucre, réflexion.

Référence 7 : Li *et al.*[182], jus d'orange, transmission.

La construction et la validation du modèle sont satisfaisantes (SEP = 0,097 g / 100 g). Le modèle est linéaire sur toute la gamme de concentration, il n'y a pas de biais. L'information utilisée par les modèles est stable sur les trois campagnes. Il a été possible de mettre en place une analyse quantitative sur le spectromètre du laboratoire de référence. Dans la partie suivante, des éléments pour la validation de la méthode SPIR seront abordés.

5 Evaluation de la méthode spectrale

5.1 Test du modèle après sa construction

5.1.1 Objectifs et protocole

Après avoir obtenu des résultats d'étalonnage et de validation satisfaisants, la méthode est validée par l'utilisation d'échantillons indépendants analysés un mois puis un an après la mise en place du modèle. Le modèle présenté précédemment est donc utilisé pour prédire des échantillons de la campagne 2001 analysés en janvier 2002 après un mois de conservation en chambre froide et des échantillons récoltés en septembre 2002. La polarimétrie et la spectroscopie proche infrarouge sont comparées grâce à un test de Fisher pour vérifier si les résultats sont statistiquement équivalents.

5.1.2 Résultats et discussion

- ***Résultats***

Tableau 22 ■ Utilisation du modèle 2001 avec 20 échantillons analysés un mois après la mise en place du modèle.

SEP (g / 100 g)	0,08
Biais (g / 100 g)	0,01
SEP(C) (g / 100 g)	0,08
R^2	0,96
F calculé	1,56
F table 5%	2,12

Tableau 23 ■ Utilisation du modèle 2001 avec 96 échantillons de 2002.

SEP (g / 100 g)	0,17
Biais (g / 100 g)	0,04
SEP(C) (g / 100 g)	0,16
R^2	0,98
F calculé	2,89
F table 5%	1,40

- ***Discussion***

Un mois après la construction du modèle, il n'y a pas différence entre les valeurs fournies par la spectroscopie proche infrarouge et celles de la polarimétrie. Par contre, on constate une augmentation de l'erreur de prédiction quand le modèle est utilisé un an après sa construction. Il semble que les échantillons sont différents d'une année sur l'autre ou que l'instrument évolue. Il faudra donc réfléchir à la gestion de la base de données et au contrôle du modèle (*cf.* chapitre 5).

5.2 Répétabilité et reproductibilité de la méthode spectrale

5.2.1 Définition

La fidélité est « l'étroitesse de l'accord entre des mesures effectuées sur des prises multiples d'un échantillon homogène » [31]. La répétabilité est « la mesure de la fidélité lorsque les mesures sont faites par un même opérateur, sur un même instrument avec une méthode unique et dans un délai court » [31]. La reproductibilité est « la mesure de la fidélité lorsque n'importe quelle condition change » [31].

La répétabilité de la mesure est évaluée. Ensuite, l'influence du changement d'instrument et celle de l'opérateur sur la mesure SPIR sont abordées et utilisées comme exemple de reproductibilité partielle. Enfin ces résultats seront comparés à la précision de la polarimétrie.

5.2.2 Répétabilité de la mesure SPIR

- Protocole

La mesure spectrale a été répétée trois fois sur cinquante échantillons. Une analyse de la variance (ANOVA) est réalisée pour déterminer si une répétition est significativement différente des deux autres.

- Résultats et discussion

Tableau 24 ■ ANOVA - Effet de la répétition sur la mesure PIR

Source de Variation	Somme des Carrés	Degrés de liberté	Moyenne des carrés
Facteur	0,0016	2	0,0008
Résidu	157,76	147	1,073
Total	157,76	149	

Fcalculé = 0,0007; Fcritique = 3,0576 ; Test non significatif

D'après l'ANOVA (Tableau 24), il n'y a pas d'effets significatifs de la répétition. Sur trois répétitions, l'oxydation de l'échantillon et la décantation de l'eau dans la cellule de mesure n'ont pas d'effets significatifs sur la valeur prédite par SPIR.

L'écart-type de répétabilité ($s_{r\text{-}SPIR}$) vaut 0,07 g / 100 g. Les répétabilités de la polarimétrie et de la SPIR sont comparables et non significativement différents ($F_{calculé}$ = 1,08 et $F_{tabulé}$ = 1,68).

5.2.3 Réproductibilité partielle de la mesure SPIR

- Influence de l'opérateur

  - ***Protocole***

Trois opérateurs ont effectué trois remplissages pour 25 échantillons. L'ordre des opérateurs est alterné pour s'affranchir de l'influence de l'oxydation. La moyenne des trois remplissages de chaque opérateur est utilisée pour réaliser une ANOVA.

- ***Résultats***

Tableau 25 ■ ANOVA - Effet de l'opérateur

Source de Variation	Somme des Carrés	Degrés de liberté	Moyenne des carrés
Facteur	0,006	2	0,003
Résidu	43,52	72	0,6044
Total	43,526	74	

Fcalculé = 0,00049; Fcritique = 3,1239 ; Test non significatif

- ***Conclusions***

D'après l'ANOVA, le changement d'opérateur n'a pas d'effet significatif sur la mesure PIR. Les trois opérateurs ont préparé l'échantillon de la même façon.

L'écart-type de reproductibilité (opérateur) est 0,09 g / 100 g. Il est plus grand que la répétabilité. En effet, l'écart-type calculé tient compte de l'influence de l'état de surface de l'échantillon (lié à l'opérateur) et de l'hétérogénéité de l'échantillon.

- Influence du changement d'instrument

- ***Protocole***

47 échantillons de betterave ont été analysés sur l'instrument de référence et sur un autre instrument en utilisant le modèle développé sur l'instrument de référence, appelé instrument maître.

- ***Résultats et discussion***

Tableau 26 ■ Influence du changement d'instrument (47 échantillons)

	Instrument de référence	Autre instrument
SEP	0,10	0,17
Biais	-0,01	0,12
SEP (C)	0,1	0,12
Pente	0,99	0,97
R^2	0,99	0,99

Il apparaît que les différences entre les deux instruments sont significatives ($F_{calculé}$ = 2,89 et $F_{critique}$ = 1,56). La reproductibilité est de 0,11 g / 100 g de betterave. Il semble donc nécessaire de mettre en place des solutions permettant de résoudre cette problématique (*cf.* chapitre 5).

5.3 Robustesse du modèle vis-à-vis de la nature de l'échantillon

5.3.1 Définition

La robustesse d'une procédure analytique est la mesure de sa capacité à ne pas être affectée par de petites variations des paramètres d'utilisation (choisies par l'expérimentateur) et elle fournit une indication sur la fiabilité de la méthode lors d'un usage normal. Seule l'influence de la nature de l'échantillon est étudiée.

5.3.2 Protocole

Lors de l'étalonnage, le protocole de préparation a été strict : la betterave est décolletée et l'échantillon est analysé par spectroscopie juste après sa préparation. De plus, seules des betteraves cultivées en France ont servi à développer le modèle.

La robustesse du modèle a été évaluée sur des échantillons ne correspondant pas à ce protocole. Les échantillons suivants ont été testés :

- Des betteraves entières
- Des betteraves récoltées avant maturité
- Des betteraves cultivées hors de France (4 origines : Belgique, Hongrie, Espagne, Angleterre)
- Des échantillons oxydés (20 minutes d'attente entre la préparation et l'analyse de l'échantillon)
- Des betteraves gelées (à – 18° C) puis dégelées

5.3.3 Résultats et discussion

Tableau 27 ■ Robustesse du modèle 2001

	Échantillons différents
Nombre d'échantillons	116
SEP	0,10
Biais	-0,01
SEP (C)	0,10
Pente	1,01
R^2	0,99

Des échantillons différents de ceux utilisés pour la construction du modèle (betterave gelée, entière, étrangère ou non mature) sont prédits avec une erreur faible (SEP = 0,10 g / 100 g) par rapport à la polarimétrie. Le modèle semble robuste vis-à-vis de certains changements de la nature de l'échantillon.

6 Bilan

- ***Points critiques de l'analyse quantitative***

Pour développer une analyse quantitative satisfaisante, il faut prendre certaines précautions. La précision de la méthode chimique de référence, le choix des échantillons de l'étalonnage, la mesure spectrale et l'étape de construction du modèle sont déterminantes dans la mise en place d'une analyse quantitative[183].

La méthode de référence doit être choisie avec soin. En effet, elle doit être spécifique du composé à analyser et la plus précise possible pour que l'analyse quantitative SPIR ait les mêmes qualités. Osborne et ses collaborateurs[14] montrent que le SEP tient compte de la précision de la méthode de référence (Équation 79).

Équation 79 $$SEP^2 = s_{r\text{-polarimétrie}}{}^2 + s_{r\text{-SPIR}}{}^2 + erreur^2$$

avec $s_{r\text{-polarimétrie}}$: la répétabilité de la mesure de référence

$s_{r\text{-SPIR}}$: la répétabilité de la mesure spectrale

erreur : un terme d'erreur lié au modèle.

Dans notre étude, l'erreur liée au modèle peut être estimée à $0,06 = (0,11^2-0,06^2-0,07^2)^{0,5}$. L'erreur apportée par le modèle mathématique est du même ordre de grandeur que la répétabilité de la polarimétrie ou de la SPIR.

Les échantillons utilisés lors de l'étalonnage doivent être représentatifs des échantillons à analyser. C'est à dire, pour des échantillons de produits agricoles, il faut avoir une diversité de variétés, d'origines et de périodes de récolte. Mais il faut également que les conditions de broyage et de stockage (température et humidité) soient représentatives des conditions rencontrées. Les échantillons doivent être en nombre suffisant. Leur nombre dépend de la méthode de régression utilisée. Ainsi, si une méthode de régression multilinéaire est utilisée, une règle empirique conseille d'avoir au moins dix échantillons par terme de la régression dans la base de données. Par contre, pour des méthodes utilisant des réseaux de neurones, le nombre d'échantillons doit être beaucoup plus important (de l'ordre de 1000). La plage de variation de la valeur de référence doit être suffisamment large pour couvrir l'ensemble des valeurs rencontrées. Quand les échantillons sont récoltés d'une façon aléatoire, les concentrations ont une distribution gaussienne. Une pratique couramment admise consiste à tenter d'avoir le même nombre d'échantillons pour toutes les

valeurs de concentrations. Fearn[184] montre que cette sélection n'améliore pas la prédiction. En effet, l'élimination d'échantillons de la base d'étalonnage conduit à réduire sa représentativité en supprimant des échantillons qui introduisaient une variabilité supplémentaire. Il semble préférable de sélectionner les échantillons sur des critères spectraux [185,186].

La mesure spectrale doit être contrôlée régulièrement pour réaliser un étalonnage correct. Deux types de contrôle existent. Il y a tout d'abord des tests instrumentaux qui sont pratiqués de façon journalière. Par exemple, sur l'instrument FOSS NIRsystem 6500, ceux-ci permettent de vérifier la répétabilité spectrale, l'alignement en longueurs d'onde sur la mesure d'un échantillon de référence (polystyrène) et contrôler le bon fonctionnement des différentes parties de l'instrument (lampe, détecteur, ventilateur)[187]. Ensuite, il est également possible d'analyser des échantillons témoins qui présentent l'avantage d'être chimiquement stables. En appliquant les mêmes équations d'étalonnage, les concentrations de différents composés de ces échantillons sont calculées et doivent être comprises entre des valeurs limites. Grâce à cet échantillon et à des cartes de contrôle, des dérives instrumentales peuvent être détectées. Dans notre étude, les dosages de la matière sèche, des protéines, des fibres et des lipides d'une cellule contenant du soja sont effectués.

- ***Bilan de la détermination du saccharose par spectroscopie proche infrarouge***

La mise en place d'une analyse quantitative a été possible pour le dosage du saccharose de la betterave. La méthode spectrale donne des résultats comparables à la méthode chimique de référence et sa répétabilité est acceptable. Cependant deux problèmes ont été soulevés. Il s'agit tout d'abord du contrôle de la validité du modèle au cours du temps et de la gestion de la base de données spectrales. Enfin, il faut également gérer les différences entre les instruments pour une application industrielle.

Chapitre 4

Application de la méthode spectroscopique pour le dosage du saccharose de la betterave sur sites industriels

1 Introduction

Le chapitre précédent a montré qu'il était possible de doser le saccharose de la betterave par SPIR sur un instrument de laboratoire. De plus ce modèle est robuste et la méthode a une répétabilité acceptable. Il faut donc maintenant s'assurer que le modèle reste valable lors d'une application industrielle. Les situations dans lesquelles le modèle peut devenir non valide sont de trois types :

- L'environnement de la mesure évolue. Les conditions de température et d'humidité influencent de façon importante la mesure spectroscopique[188].
- Les échantillons changent de nature chimique (évolution de leur composition) ou d'aspect physique (la taille des particules et la viscosité changent en fonction de la préparation de l'échantillon).
- La réponse instrumentale est différente. C'est le cas lorsqu'on change tout ou partie d'un instrument ou lorsqu'un instrument a des composants (lampe, détecteur) qui vieillissent[189].

L'environnement de la mesure, c'est à dire la température et humidité sont contrôlées : 20 °C et 80 % d'humidité relative. Seules les variations liées à l'échantillon et les différences instrumentales doivent être gérées.

Pour utiliser la méthode SPIR en milieu industriel, il faut résoudre les problèmes concernant la validité du modèle. Dans un premier temps, il faut donc savoir si un modèle développé à une année donnée, reste valable l'année suivante. En effet, la nature des échantillons est variable : des variétés nouvelles peuvent être utilisées et les conditions climatiques varient d'une année sur l'autre. Dans un second temps, l'utilisation de plusieurs instruments est abordée et des solutions sont proposées pour obtenir le même niveau d'erreur sur tous les instruments.

Enfin, il semble intéressant de développer un instrument industriel pour l'analyse de la betterave. Ainsi, la dernière partie de ce chapitre présente le développement d'un analyseur automatique adapté aux conditions et aux cadences industrielles.

2 Intégration des variabilités liées à l'échantillon et gestion de la base de données

L'objectif est de savoir si un modèle construit au cours d'une campagne peut être utilisé lors de la campagne suivante en conservant le même niveau d'erreur et s'il faut mettre à jour ce modèle en incorporant dans la base de données de nouveaux échantillons. Parallèlement à cette mise à jour, il se pose le problème de la gestion de la base de données. Chaque année, la taille de la base de données augmente par l'analyse d'échantillons supplémentaires. Il semble donc nécessaire de supprimer les échantillons qui n'apportent pas d'informations utiles.

2.1 Mise à jour annuelle du modèle

2.1.1 Protocole

Plusieurs modèles ont été construits à partir des lots d'étalonnages des campagnes 1999 et 2000 (1016 échantillons de 1999 et 669 échantillons de 2000) et un nombre variable d'échantillons de la campagne 2001 (20 à 525 échantillons). Les échantillons sont ajoutés dans l'ordre chronologique d'analyse pour simuler la situation réelle. Tous les modèles sont ensuite validés avec un même lot d'échantillons inconnus des modèles (525 échantillons).

2.1.2 Résultats

La Tableau 28 montre l'évolution du SEP en fonction du nombre d'échantillons de la campagne 2001 utilisés. On constate une dégradation du SEP lorsque le modèle 1999/2000 est utilisé sur les échantillons de 2001. Cependant, l'ajout d'échantillons de la

nouvelle campagne permet de diminuer de façon significative le SEP et d'obtenir le même niveau d'erreur que la campagne précédente. Le test de Fisher permet d'identifier les SEP non significativement différents. La limite de confiance du SEP vaut 0,11 g / 100 g, ce qui signifie que les valeurs comprises entre 0,10 et 0,11 ne sont pas différentes d'un point de vue statistique.

Tableau 28 ■ Mise à jour du modèle 1999/2000 au début de la campagne 2001. (Validation sur un lot de 525 échantillons de 2001 inconnus du modèle). En gras, les SEP non significativement différents.

Nombre d'échantillons de 2001	0[1]	20	50	100	150	200	300	400	525[2]
SEP	0,21	0,16	0,14	0,12	**0,11**	**0,11**	**0,10**	**0,10**	**0,10**
Biais	-0,17	-0,11	-0,09	-0,06	-0,03	-0,01	0,00	0,00	0,00
SEP (C)	0,12	0,11	0,11	0,11	0,11	0,10	0,10	0,10	0,10
Pente	0,99	0,99	0,99	0,99	0,99	1,00	1,00	1,00	1,00
R^2	0,98	0,98	0,98	0,98	0,99	0,99	0,99	0,99	0,99

1 : lot d'étalonnage = 1016 échantillons de 1999 et 669 échantillons de 2000.

2 : lot d'étalonnage = 1016 échantillons de 1999, 669 échantillons de 2000 et 525 échantillons de 2001.

2.1.3 Discussion

- ***Nombre d'échantillons pour la mise à jour***

La mise à jour du modèle peut se faire en début de campagne par l'analyse de 150 échantillons pour retrouver le niveau d'erreur de la campagne précédente (d'après le test de Fisher).

Il faut noter que dans notre étude, les 80 premiers échantillons ont la même origine. Si on effectue la mise à jour avec 21 échantillons provenant de 21 sites géographiques, on obtient un SEP de 0,13. Il semble donc préférable d'utiliser des échantillons ayant des origines différentes afin de tenir compte au maximum de la variabilité des échantillons.

Ce résultat confirme les résultats d'une étude similaire réalisée en 2000 : à savoir la mise à jour du modèle 1999 par l'ajout d'échantillons de la campagne 2000[177].

- ***Effet de la mise à jour***

La mise à jour diminue le biais, c'est-à-dire l'erreur systématique. Le fait de rajouter des échantillons de la nouvelle campagne dans la base de données permet de prendre en compte la variabilité inter campagne de la betterave mais également les variations instrumentales.

- ***Vérification de l'hypothèse***

Au cours de la campagne 2002, nous avons voulu vérifier que l'ajout de 150 échantillons de la nouvelle campagne permettait de diminuer de façon significative le SEP. Les résultats sont présentés dans le Tableau 29. La mise à jour du modèle 1999-2001 par ajout d'échantillons de 2002 permet donc de diminuer de façon significative le biais.

Tableau 29 ■ Mise à jour du modèle en 2002

Validation sur un lot de 70 échantillons de 2002 (analysés après la mise à jour du modèle)

Modèle	**Modèle 2001**	**Modèle après mise à jour (150 échantillons de 2002)**
SEP	0,15	0,10
Biais	0,06	0,03
SEP (C)	0,13	0,09
Pente	1,01	1,01
R^2	0,98	0,99

2.2 Gestion de la base de données spectrales

2.2.1 Principes

La gestion de la base de données a deux objectifs : le premier est la détection puis l'élimination des spectres hors normes et le second est l'élimination des informations redondantes.

- D'un point de vue statistique, les spectres hors normes peuvent être détectés par le calcul de la distance de Mahalanobis, notée GH, entre le spectre $\mathbf{x}_i$ et le spectre moyen $\mathbf{x}_m$. Avec le logiciel Winisi, le GH est calculé à partir des coordonnées factorielles issues de ACP. Une valeur limite du GH est fixée pour détecter les échantillons hors normes.

Équation 80 $GH_i = (\mathbf{x}_i-\mathbf{x}_m)\ (\mathbf{X}'\mathbf{X})^{-1}\ (\mathbf{x}_i-\mathbf{x}_m)'$

Dans notre étude avec des instruments de laboratoire, les spectres aberrants sont rares. En effet, un spectre est hors normes quand la cellule de mesure a été mal remplie, c'est à dire quand il y a présence de terre ou de bulles d'air à la surface du quartz ou quand il y a eu de la lumière parasite. Or avec un remplissage manuel, un contrôle visuel de l'état de surface du quartz est effectué avant l'analyse. C'est pourquoi, le critère GH n'est pas utilisé avec l'instrument de laboratoire. Mais il sera utilisé avec l'instrument automatisé pour une détection des erreurs de remplissage.

- Le second point est l'élimination des spectres redondants. Afin de ne pas avoir une base de données trop importante, on peut diminuer le nombre de spectres du lot d'étalonnage en supprimant des spectres qui sont similaires. Ainsi les distances de Mahalanobis entre tous les points sont calculées, la distance NH correspond à la distance entre le spectre et le spectre qui lui est le plus proche. Un seuil de proximité, noté NH_{limite}, est fixé. Un échantillon est sélectionné au hasard et les échantillons qui sont à une distance inférieure au NH_{limite} fixé par l'utilisateur sont éliminés. La procédure est effectuée de façon itérative. Dans cette partie, seuls les résultats concernant l'utilisation du critère NH seront présentés.

Équation 81 $NH_i = \min_{j\neq i} ((\mathbf{x}_i-\mathbf{x}_j)\ (\mathbf{X}'\mathbf{X})^{-1}\ (\mathbf{x}_i-\mathbf{x}_j)')$

2.2.2 Résultats et discussions

Tableau 30 ■ Influence de la taille de la base d'étalonnage sur la prédiction.

Validation sur un lot de 70 échantillons de 2002

Critère de sélection	Aucun Base mise à jour en 2002	NH = 0,3	NH = 0,5	NH = 0,8
Nombre d'échantillons de l'étalonnage	2360	2067	1075	357
Validation sur un lot de 70 échantillons de 2002				
SEP	0,10	0,10	0,11	0,10
Biais	0,03	0,04	0,04	0,03
SEP(C)	0,09	0,09	0,10	0,09
Validation sur 116 échantillons de 2001 (lot pour tester la robustesse – chapitre 4)				
SEP	0,10	0,10	0,10	0,12
Biais	-0,01	-0,01	-0,01	-0,02
SEP(C)	0,10	0,10	0,10	0,11

On constate qu'un modèle précis peut être développé avec un nombre beaucoup plus faible d'échantillons (357). Cependant, ce modèle est moins robuste que les autres. En effet, le SEP de 0,12 est significativement différent de 0,10 g / 100 g car le $F_{calculé}$ est de 1,44 et le $F_{critique}$ est de 1,36.

L'utilisation du NH permet de diminuer le nombre d'échantillons de la base d'étalonnage.

Le choix du nombre d'échantillons reste délicat. On peut envisager d'utiliser le NH lors de la mise à jour de la base d'étalonnage. Ainsi, on ne rajoute dans la base d'étalonnage que les spectres ayant des caractéristiques différentes.

3 Transfert d'étalonnage et utilisation d'un réseau d'instruments

3.1 Problématique liée à l'utilisation de plusieurs spectromètres

Au cours des trois années de l'étude, cinq autres instruments ont été testés. Le Tableau 31 montre les résultats obtenus en utilisant directement le modèle de 2002 développé sur l'instrument de référence appelé instrument maître sur les cinq autres instruments, appelés instruments esclaves. On constate une augmentation significative du SEP (d'après le test de Fisher) sur les instruments esclaves. En effet, le plus petit Fisher calculé est de 1,440 pour l'instrument 4 et le Fisher critique de 1,412. L'augmentation du SEP s'explique par l'amplification du biais. Il y a donc une erreur systématique entre les instruments esclaves et l'instrument maître.

Tableau 31 ■ Validation du modèle développé sur l'instrument maître avec cinq autres instruments.

	Maître	Instrument 1	Instrument 2	Instrument 3	Instrument 4	Instrument 5
Nombre d'échantillons du lot de validation	70	47	55	41	41	50
SEP (g / 100 g)	0,10	0,17	0,17	0,19	0,12	0,62
Biais (g / 100 g)	0,03	0,12	-0,14	-0,10	0,09	-0,59
SEP(C) (g / 100 g)	0,09	0,12	0,10	0,16	0,08	0,17

Une ACP est réalisée à partir du lot d'étalonnage de l'instrument maître. Ensuite les spectres des lots de validation des instruments esclaves sont projetés en temps qu'individus supplémentaires. Cette ACP permet de mettre en évidence les différences spectrales entre les instruments (Figure 39). Dans le plan 1-2 de l'ACP, les instruments 3 et 5 apparaissent différents de l'instrument maître. Comme l'axe 2 s'explique par des longueurs d'onde du visible comprises entre 400 nm et 600 nm, on peut penser que les détecteurs du domaine visible des instruments 3 et 5 ont des comportements différents de ceux de l'instrument maître. De même dans le plan 5-6, l'instrument 4 se distingue de l'instrument maître.

La Figure 40 montre la différence entre les spectres moyens de 40 échantillons identiques analysés sur l'instrument maître et sur trois des esclaves. On constate que les différences spectrales sont relativement faibles. Cependant elles expliquent l'augmentation des SEP sur les instruments esclaves.

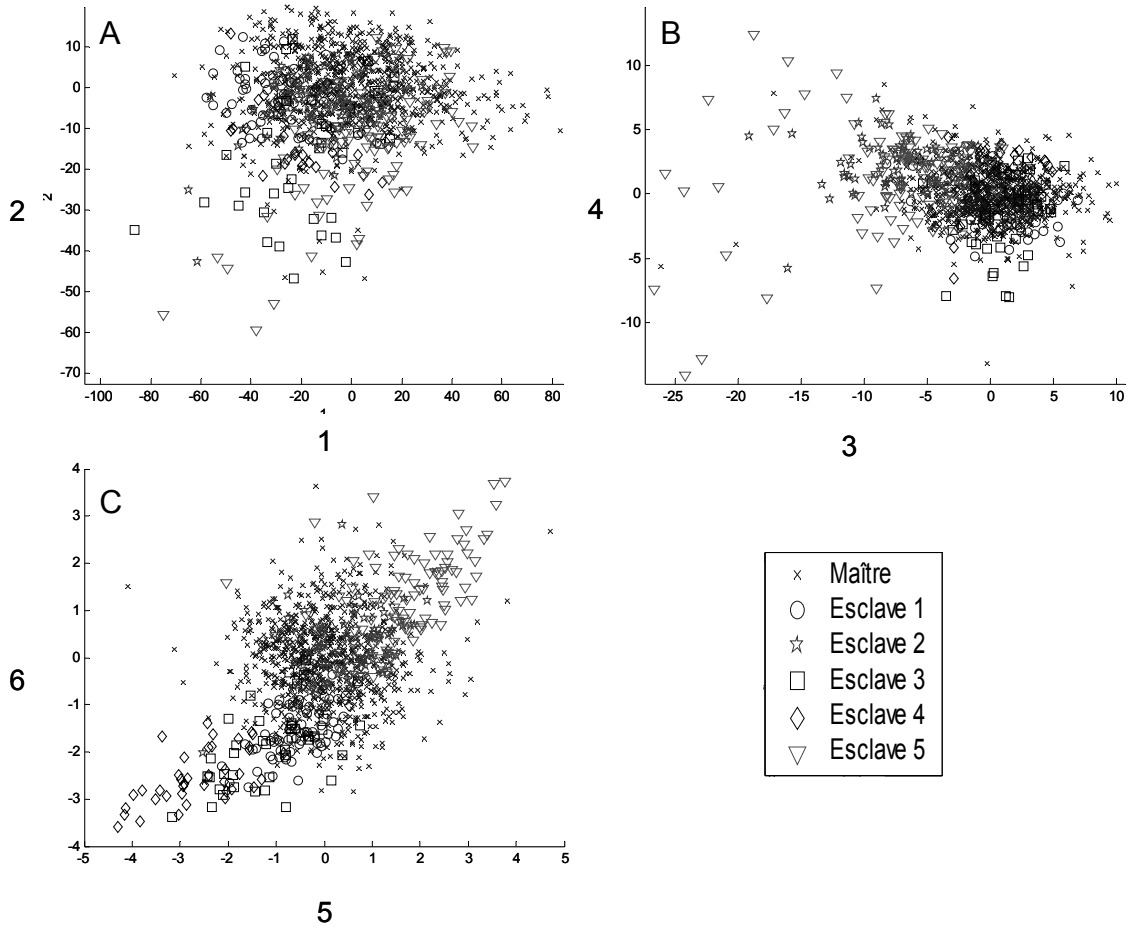


Figure 39 ■ Analyse en composantes principales sur la base de l'instrument maître et projection des spectres des cinq autres instruments en individus supplémentaires.

(A) plan 1-2 (96,9 % de la variance totale) (B) plan 3-4 (2,7 % de la variance totale) (C) plan 5-6 (0,3 % de la variance totale)

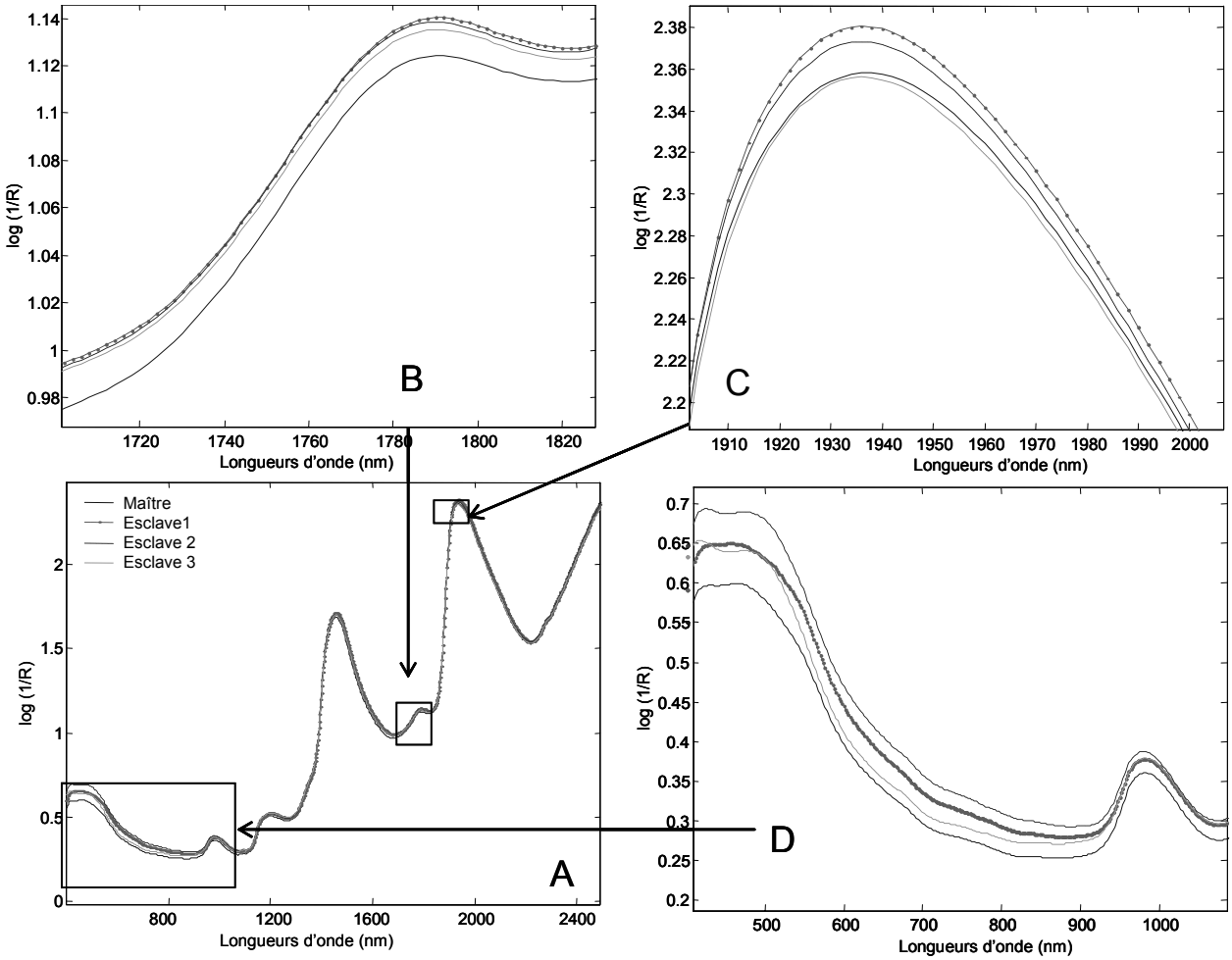


Figure 40 ■ Spectres moyens de l'instrument maître et des esclaves 2, 3 et 4 (moyenne sur 41 échantillons analysés sur tous les instruments). (A) spectre sur toute la gamme spectrale (B) (C) et (D) agrandissements de certains domaines spectraux.

L'objectif est d'obtenir sur les cinq instruments un SEP semblable à celui de l'instrument maître (d'après le test de Fisher) ou au moins un SEP inférieur à 0,15 (seuil susceptible d'être accepté par les industriels).

3.2 Données utilisées

L'étude de l'utilisation de plusieurs instruments a été menée pendant les trois années de la thèse. Seul le modèle 2002 est utilisé. Pour l'instrument maître, 70 spectres de 2002 sont utilisés en validation. Le Tableau 32 présente les données utilisées dans ce paragraphe. Les échantillons utilisés pour le transfert d'étalonnage sont choisis au hasard parmi l'ensemble des analyses.

Tableau 32 ■ Description des lots de données utilisés pour le transfert d'étalonnage

	Instrument 1	Instrument 2	Instrument 3	Instrument 4	Instrument 5
Date	Octobre 2001	Janvier 2001	Septembre 2002	Septembre 2002	Octobre 2000
Nombre d'échantillons analysés sur l'esclave et le maître	117	229	131	131	149

3.3 Méthodes

Trois approches ont été testées pour réduire l'influence des différences instrumentales sur le dosage du saccharose par SPIR. La première solution consiste à développer un modèle qui ne soit pas influencé par les différences instrumentales. La seconde approche est la correction des concentrations obtenues avec les instruments esclaves. Enfin, la dernière solution consiste à modifier les spectres des instruments esclaves pour qu'ils ressemblent aux spectres que l'on aurait pu acquérir sur l'instrument maître.

3.3.1 Modèles robustes

Tout d'abord, l'utilisation des prétraitements mathématiques tels que SNV et les dérivées, rend les modèles plus robustes au changement d'instrument[190]. En effet, les prétraitements normalisent les spectres, ajustent les lignes de base et donc réduisent les différences instrumentales. Mais les prétraitements seuls ne sont pas suffisants. Dans notre étude, une démarche supplémentaire a été introduite.

Afin de développer des modèles plus robustes, des bases contenant des spectres de plusieurs instruments sont utilisées. Les spectres des lots d'étalonnage des esclaves sont introduits dans la base de données de l'instrument maître. Différents modèles sont construits avec des données provenant du maître et d'un ou plusieurs instruments esclaves. Ensuite les lots de validation sont utilisés pour évaluer les écarts avec la mesure de référence.

3.3.2 Correction des concentrations

Les mêmes échantillons sont analysés à la fois sur l'instrument maître et sur un instrument esclave. La méthode de correction des concentrations est la méthode la plus utilisée[191].

- ***Corrections de la pente et de l'ordonnée à l'origine***

Une régression linéaire simple (Équation 82) est réalisée entre les concentrations obtenues sur l'instrument maître et sur l'instrument esclave notées respectivement, $y_{j,maître}$ et $y_{j,esclave}$ pour tous les échantillons (j allant de 1 à N).

Équation 82 $y_{j,maître} = a.y_{j,esclave} + b$

Dardenne[191] conseille de tester la significativité de la pente et de l'ordonnée à l'origine. Il faut calculer la variance de la pente σ_a^2 (Équation 83) et celle de l'ordonnée à l'origine σ_b^2 (Équation 84). On applique alors des tests de Student pour calculer des intervalles de confiance. La pente ne demande pas à être réajustée si la valeur 1 est incluse dans l'intervalle (Équation 85). De même, l'ordonnée à l'origine ne nécessite pas de réajustement si la valeur 0 est dans l'intervalle calculé (Équation 86).

Équation 83 $$\sigma_a^2 = \frac{\sigma_r^2}{\sum_{i=1}^{N} (y_{j,esclave} - \bar{y}_{esclave})^2}$$

Équation 84 $$\sigma_b^2 = \sigma_r^2.(\frac{1}{N} + \frac{\bar{y}_{esclave}^2}{\sum_{i=1}^{N}(y_{j,esclave} - \bar{y}_{esclave})^2})$$

Avec N : le nombre d'échantillons et t : la valeur du Student (α = 5 % et N degrés de liberté) $\bar{y}_{esclave}$: la moyenne des valeurs de l'instrument esclave et σ_r l'écart-type des résidus ($d_j = y_{j,maître} - y_{j,esclave}$) de cette régression.

Équation 85 $$a - t.\sigma_a \leq 1 \leq a + t.\sigma_a$$

Équation 86 $$b - t.\sigma_b \leq 0 \leq b + t.\sigma_b$$

- ***Correction du biais***

Une version simplifiée de cette méthode dispense de la correction de pente et corrige uniquement le biais (erreur systématique) entre les deux lots de données (Équation 87). Cette correction s'applique quand le biais est supérieur au biais limite[192] selon l'Équation 88. Dans notre étude, le logiciel Winisi® calcule le biais limite par approximation (n = 10) selon Équation 89.

Équation 87 $y_{maître} = y_{esclave} + biais$

Équation 88 Biais limite = $\pm$ (t.SECV) / $n^{0,5}$ avec n : le nombre d'échantillons du nouveau lot de données et t : la valeur du Student (bilatéral, degrés de liberté associés au SECV)

Équation 89 Biais limite = $\pm$ (0,6.SECV)

Ensuite les corrections des concentrations sont appliquées aux lots de validation. Les indicateurs statistiques SEP, R^2, biais et SEP(C) sont calculés après chaque correction.

3.3.3 Corrections spectrales

La dernière solution consiste à corriger les données spectrales. Les spectres des instruments esclaves sont transformés pour être semblables à ceux de l'instrument maître. Ainsi, le modèle de l'instrument maître est utilisé directement sur l'ensemble des instruments. Dans ce type de transfert, aucune concentration de référence n'est utilisée. Il est alors possible de le réaliser avec des échantillons génériques comme le polystyrène[193] ou des standards commerciaux.

Les principales méthodes pour effectuer cette correction spectrale sont les suivantes :

- ***la standardisation directe (DS « Direct Standardisation »)***

Il s'agit de la première méthode développée[194]. L'hypothèse de départ est qu'une longueur d'onde de l'instrument maître est corrélée à l'ensemble des longueurs d'onde de l'instrument esclave. La méthode PLS permet d'effectuer une régression entre la longueur d'onde i de la matrice $\mathbf{X}_{maître}$ et toutes les longueurs d'onde de la matrice $\mathbf{X}_{esclave}$ (avec i qui couvre l'ensemble du spectre). La standardisation DS est basée sur le calcul d'une matrice de passage **F** telle que

Équation 90 $\mathbf{X}_{maître,\ étalonnage} = \mathbf{X}_{esclave,\ étalonnage}.\mathbf{F}$

Ensuite la matrice est utilisée pour transformer les spectres de l'instrument esclave :

Équation 91 $\mathbf{X}_{\text{esclave, standardisé}} = \mathbf{X}_{\text{esclave, validation}}.\mathbf{F}$

- ***la standardisation pas à pas (PDS « Piecewise Direct Standardisation »)***

Cette méthode[194] s'inspire de la méthode précédente. La différence réside dans le calcul de la matrice de passage **F**. Cette méthode considère que l'absorbance à une longueur d'onde sur l'instrument maître est corrélée à un certain nombre de longueurs d'onde de l'instrument esclave sur une fenêtre déterminée.

Équation 92 $\mathbf{x}_{\text{maître,i}} = \mathbf{W}_i.\mathbf{b}_i$ avec $\mathbf{W}_i$ matrice d'absorbance de l'esclave pour les longueurs d'onde comprises entre i-k et i+k, $\mathbf{x}_{\text{maître,i}}$ vecteur d'absorbance à la longueur d'onde i et $\mathbf{b}_i$ vecteur des coefficients.

La fenêtre se déplace sur l'ensemble du spectre. Dans notre étude, la valeur de k est trois. L'ensemble des vecteurs $\mathbf{b}_i$ calculés permet d'obtenir la matrice de passage **F** qui sera utilisée sur les échantillons de validation.

- ***la méthode brevetée par Shenk et Westerhaus (SW)***

Une description complète de cet algorithme est réalisée par Bouveresse et ses collaborateurs[195]. Cette méthode est protégée par un brevet américain[196]. Elle procède en deux étapes. La première est un ajustement des longueurs d'onde entre les deux instruments. Comme la méthode PDS, on considère qu'une longueur d'onde de l'instrument maître est corrélée aux longueurs d'onde de l'instrument esclave sur une fenêtre spectrale. Les coefficients de corrélation entre la longueur d'onde i de l'instrument maître et les longueurs d'onde i' de l'instrument esclave sont calculés sur une fenêtre spectrale. Pour estimer la valeur maximale de cette corrélation, une régression quadratique est réalisée (Équation 93). La correction spectrale se termine par la modélisation de i' à l'aide d'une parabole (Équation 94) sur l'ensemble du domaine spectrale. Ainsi, à chaque longueur d'onde de l'instrument maître correspond une longueur d'onde de l'instrument esclave.

Équation 93 $\text{Corrélation} = t.i^2 + u.i + v$

Équation 94 $i' = t.i^2 + u.i + v$ avec i la longueur d'onde de l'instrument maître, i' celle de l'instrument esclave et t, u, v trois constantes.

La seconde étape consiste à une correction des absorbances longueur d'onde par longueur d'onde. Une régression linéaire simple entre l'absorbance $x_{i,maître}$ et $x_{i,eslcave}$ est effectuée pour l'ensemble des longueurs d'onde du spectre.

Équation 95 $x_{i,maître} = a_i.x_{i,esclave} + b_i$ avec a_i et b_i deux constantes

Dans notre étude, deux types d'échantillons ont été utilisés pour réaliser la standardisation des instruments : des échantillons de betterave qui ne sont pas stables dans le temps et un lot de 30 coupelles commerciales contenant des produits agricoles (Foss, IH-0328 Natural product® sample kit) qui présentent l'avantage d'être stables chimiquement et l'inconvénient d'avoir un spectre très différent des échantillons de betterave comme le montre la figure suivante.

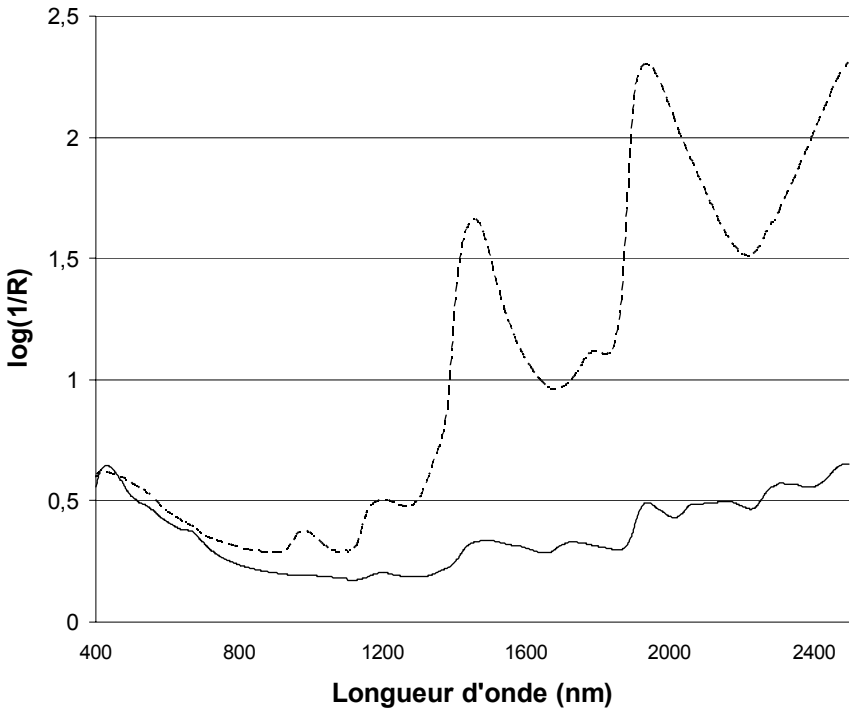


Figure 41 ■ Comparaison du spectre moyen des trente coupelles commerciales et du spectre moyen de 525 échantillons de betterave. ____ Spectre moyen des coupelles commerciales ------ Spectre moyen des échantillons de betterave.

3.4 Résultats et discussion

3.4.1 Modèles robustes

Les spectres des instruments esclaves sont rajoutés à la base d'étalonnage de l'instrument maître par étapes. Dans un premier temps, la base d'étalonnage de l'instrument maître, à laquelle on rajoute 70 spectres de l'instrument esclave 1, est utilisée pour développer un modèle. On constate que ce modèle (Tableau 33) permet d'obtenir un SEP significativement plus faible sur le lot de validation de l'instrument 1 et que l'erreur de prédiction sur l'instrument maître reste inchangée. Par contre, la prédiction des esclaves 2 et 3 n'est pas améliorée. Le SEP de l'instrument 5 diminue mais reste très élevé. Le SEP de l'instrument 4 diminue et devient non significativement différent de celui obtenu sur l'instrument maître.

Tableau 33 ■ Validation d'un modèle construit avec les spectres de l'instrument maître et 70 spectres de l'instrument 1. En gras les valeurs du SEP inférieur à 0,15 (en g / 100 g).

	Maître	Instrument 1	Instrument 2	Instrument 3	Instrument 4	Instrument 5
SEP	**0,10**	**0,11**	0,23	0,20	**0,10**	0,41
Biais	0,00	-0,01	-0,21	-0,10	0,02	-0,37
SEP(C)	0,10	0,11	0,09	0,18	0,10	0,17

Dans un second temps, un modèle (Tableau 34) est développé avec des spectres des instruments esclaves 1 et 2. De même, un troisième modèle (Tableau 35) est construit à partir des spectres de l'instrument maître et des esclaves 1, 2 et 5. On constate que pour un appareil donné, l'erreur de prédiction s'améliore si la base de données contient des spectres de cet instrument. Par contre, pour les instruments dont les spectres n'ont pas été rajoutés à la base de données, le SEP n'est pas suffisamment amélioré (excepté pour l'instrument 4 qui avant toute correction avait un SEP faible).

Tableau 34 ■ Validation d'un modèle développé avec les spectres de l'instrument maître, 70 et 164 spectres des instruments 1 et 2 respectivement. En gras les valeurs du SEP inférieur à 0,15 (en g / 100 g).

	Maître	Instrument 1	Instrument 2	Instrument 3	Instrument 4	Instrument 5
SEP	**0,10**	**0,10**	**0,11**	0,16	**0,10**	0,26
Biais	0,00	-0,01	-0,03	-0,04	0,01	-0,20
SEP(C)	0,10	0,10	0,09	0,16	0,10	0,16

Tableau 35 ■ Validation d'un modèle construit avec les spectres de l'instrument maître, 70, 164 et 99 spectres des instruments 1, 2 et 5. En gras les valeurs du SEP inférieur à 0,15 (en g / 100 g).

	Maître	Instrument 1	Instrument 2	Instrument 3	Instrument 4	Instrument 5
SEP	**0,10**	**0,10**	**0,11**	0,21	**0,11**	**0,14**
Biais	0,00	-0,02	0,04	-0,15	0,01	0,02
SEP(C)	0,10	0,10	0,09	0,15	0,10	0,14

Enfin un dernier modèle est développé à partir des spectres de l'instrument maître et des cinq lots d'étalonnage des esclaves. On remarque à nouveau que le SEP de l'instrument maître n'est pas dégradé. On constate (Tableau 36) alors que tous les instruments ont un SEP inférieur à 0,15. Les tests de Fisher montrent que les instruments esclaves 1, 2, 3 et 4 ont un SEP qui n'est pas significativement différent de l'instrument maître (pour l'instrument 4, le Fisher calculé est de 1,210 et le $F_{critique}$ de 1,412). L'instrument esclave 5 a un SEP significativement supérieur à celui de l'instrument maître ($F_{calculé}$ = 1,960 et $F_{critique}$ = 1,37). Cependant on remarque que l'erreur standard avant transfert était de 0,62 g / 100 g. La réduction de l'erreur a donc été très importante et la procédure est puissante même avec un instrument qui n'est pas directement opérationnel. En conclusion, l'utilisation d'une base de données contenant des spectres de plusieurs instruments permet de réduire les biais de prédiction sur l'ensemble des instruments.

Tableau 36 ■ Développement d'un modèle avec les spectres de l'instrument maître et 70, 164, 99, 90 et 90 spectres des cinq instruments respectivement. En gras les valeurs du SEP inférieur à 0,15 (en g / 100 g).

	Maître	Instrument 1	Instrument 2	Instrument 3	Instrument 4	Instrument 5
SEP	**0,10**	**0,11**	**0,11**	**0,09**	**0,09**	**0,14**
Biais	0,00	0,02	-0,03	0,00	0,02	0,01
SEP(C)	0,10	0,10	0,10	0,09	0,09	0,14

3.4.2 Correction des concentrations

- ***Correction de la pente et de l'ordonnée à l'origine***

Dans un premier temps, une étude préliminaire a permis de déterminer le nombre d'échantillons nécessaire au transfert. Quand 20 ou 10 échantillons sont utilisés, on constate une augmentation importante des écarts-types de la pente et de l'ordonnée à l'origine (Tableau 37). C'est pourquoi, nous choisissons d'utiliser 30 échantillons pour le transfert d'étalonnage. Ces résultats sont en accord avec ceux de la littérature qui conseillent d'utiliser un minimum de 10 standards[191] : par exemple, Osborne et Fearn[197] réalisent la correction avec 20 échantillons.

Tableau 37 ■ Influence du nombre d'échantillons sur la détermination de la pente et de l'ordonnée à l'origine. (Instrument 1)

Nombre d'échantillons	70	60	50	40	30	20	10
Ordonnée à l'origine	-0,34	0,71	0,49	0,38	0,51	0,26	-1,19
Pente	1,01	0,97	0,98	0,98	0,98	0,99	1,08
Ecart type de l'ordonnée à l'origine	0,33	0,32	0,29	0,34	0,36	**0,60**	**1,13**
Ecart type de la pente	0,02	0,02	0,02	0,02	0,02	**0,04**	**0,07**

Les résultats obtenus sont présentés dans le Tableau 38. Lors de l'étalonnage, il apparaît que la correction de pente n'est pas nécessaire. C'est pourquoi nous décidons d'ajuster uniquement les biais.

Tableau 38 ■ Influence de la correction pente et ordonnée à l'origine des concentrations sur les erreurs de prédiction (30 échantillons)

		Instrument 1	Instrument 2	Instrument 3	Instrument 4	Instrument 5
Etalonnage	Pente	0,98	1,04	0,99	1,01	1,04
	Ordonnée à l'origine	0,51	-0,97	-0,082	-0,13	-0,27
	Ecart type ordonnée origine	0,36	0,55	0,54	0,22	0,49
	Ecart type pente	0,02	0,03	0,03	0,01	0,03
Validation	SEP (g / 100 g)	**0,12**	0,15	0,15	**0,09**	0,16
	Biais (g / 100 g)	0,05	0,10	0,02	-0,04	-0,05
	SEP(C) (g / 100 g)	0,11	0,11	0,15	0,08	0,15

- ***Ajustement des biais***

Comme précédemment, trente échantillons sont utilisés pour le calcul du biais. Lors de la phase d'étalonnage (Tableau 39), nous constatons que les biais calculés sur l'ensemble des instruments esclaves sont supérieurs au biais limite qui vaut 0,06 en valeur absolue (calculé par le logiciel Winisi selon l'Équation 89). Il apparaît donc nécessaire de corriger l'erreur systématique existante entre les instruments esclaves et l'instrument maître.

La validation montre que cette méthode permet de diminuer le SEP sur l'ensemble des instruments esclaves. Après correction, quatre des cinq instruments ont un SEP inférieur à 0,15. La Figure 42 montre l'effet de la correction de biais sur l'instrument 2.

Tableau 39 ■ Influence de la correction des concentrations sur les erreurs de prédiction. Unité g / 100 g. En gras, les SEP inférieurs à 0,15 g / 100 g.

		Instrument 1	Instrument 2	Instrument 3	Instrument 4	Instrument 5
Etalonnage	Biais	0,14	-0,17	-0,10	0,09	-0,61
Validation	SEP	**0,12**	**0 ,10**	**0,14**	**0,08**	0,18
	Biais	0,01	0,03	0,01	0,01	0,10
	SEP(C)	0,11	0,09	0,14	0,08	016

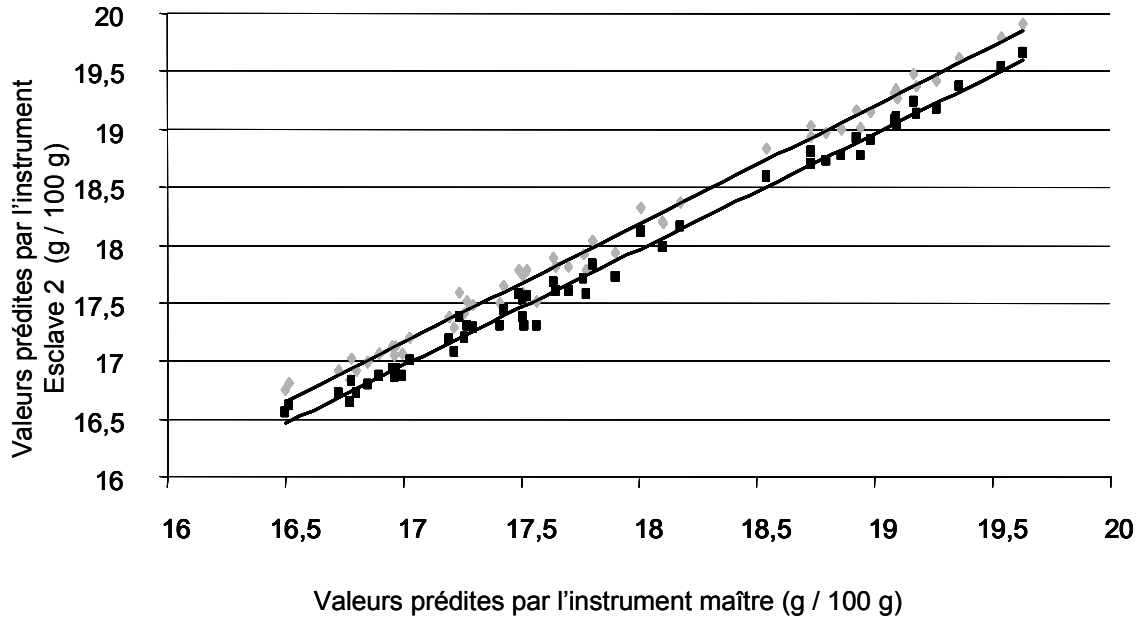


Figure 42 ■ Effet de la correction des biais sur le lot de validation de l'instrument 2. En gris, avant la correction de biais et en noir, après la correction de biais.

- ***Discussion***

Sur nos données, seule la correction de biais est apparue nécessaire. Cette méthode donne des résultats satisfaisants quand elle est appliquée sur des instruments de même type et quand les ajustements sont de faibles amplitudes[198].

L'avantage de cette méthode est son aspect univarié. Ainsi, cette méthode est simple et rapide. Cependant l'inconvénient majeur est que cette méthode nécessite l'analyse d'échantillons de betterave qui ne sont pas stables chimiquement. De plus, cette correction est valable uniquement pour un modèle donné. Ainsi, si le modèle doit être mis à jour, la correction doit être de nouveau effectuée.

3.4.3 Corrections spectrales

- Choix du protocole

Dans un premier temps, le protocole a été mis en place : la nature des échantillons, leur nombre et l'algorithme à utiliser ont été déterminés.

- ***Choix des échantillons***

Tout d'abord, deux types d'échantillons ont été étudiés : des échantillons de betterave et des coupelles commerciales. L'algorithme de Shenk et Westerhaus est utilisé pour construire les équations de standardisation. Pour l'instrument 4, les coupelles commerciales n'ont pas permis de réduire le biais (Tableau 40). On retrouve ainsi les résultats de Bouveresse et ses collaborateurs[195] : pour effectuer une modification spectrale, il semble nécessaire d'avoir des échantillons qui ont des spectres semblables à ceux des échantillons analysés, c'est à dire avec le même niveau d'absorbance. Même si les échantillons de betterave ne sont pas chimiquement stables, ils sont les mieux adaptés pour le transfert et ils ont donc été utilisés pour cette opération de transfert.

Tableau 40 ■ Standardisation de l'instrument 4 à l'aide d'échantillons de betterave ou de coupelles commerciales. Choix de la nature des échantillons. Validation de la standardisation sur 41 échantillons indépendants (unité en g / 100 g).

Echantillon pour la standardisation	30 échantillons de betterave	30 coupelles agricoles
SEP	0,11	0,16
Biais	0,02	0,09
SEP(C)	0,11	0,13

Le choix de la nature des échantillons et de leur nombre est également important. Pour caractériser de façon précise les différences entre les réponses de deux instruments, il faut que les échantillons analysés soient stables chimiquement et représentatifs des échantillons[199]. En effet, les échantillons standards doivent avoir des propriétés physiques et chimiques stables afin que les différences entre les spectres soient uniquement dues aux différences instrumentales. De même, il est recommandé d'utiliser des échantillons qui soient représentatifs de ceux analysés.

- ***Choix du nombre d'échantillons***

Des équations de standardisation sont mises en place à partir d'un nombre variable d'échantillons de betterave. Pour choisir le nombre, on prend la valeur minimum du SEP dans le Tableau 41 (0,13 pour 40 échantillons). Le nombre d'échantillons inférieur à 40 tel que le SEP ne soit pas significativement différent de 0,13 par le test de Fisher est alors utilisé. Quand on compare les SEP après standardisation en utilisant les équations développées avec 30 ou 40 échantillons, on obtient un Fisher non significatif de 1,159

($F_{critique}$ = 1,744). Par contre, en comparant l'équation de transfert développée avec 20 échantillons à celle utilisant 40 échantillons, le Fisher est significatif et vaut 1,917 ($F_{critique}$ = 1,837). C'est pourquoi, nous décidons d'utiliser 30 échantillons pour le calcul des équations de transfert. Ces résultats sont en accord avec la littérature : Shenk et Westerhaus utilisent entre 15 et 30 échantillons pour le développement des équations de transfert.

Tableau 41 ■ Choix du nombre d'échantillons pour la modification spectrale (en g / 100 g)

Instrument 3	20	30	40	50	60	70	80	90
SEP	0,18	**0,14**	**0,13**	**0,13**	**0,13**	**0,13**	**0,13**	**0,14**
Biais	-0,01	0,01	0,02	0,02	0,01	0,02	0,03	0,03
SEP(C)	0,15	0,15	0,12	0,12	0,13	0,13	0,12	0,14

- ***Comparaison des algorithmes de transfert***

Enfin, trois algorithmes ont été testés. Les équations sont développées en analysant 30 échantillons de betterave. On constate que les méthodes SW et PDS donnent les meilleurs résultats. Par contre, la méthode DS ne permet pas de standardiser les instruments dans notre étude.

Tableau 42 ■ Choix de l'algorithme de transfert (en g / 100 g)

Instrument1	SW	PDS	DS
SEP	**0,11**	**0,11**	0,18
Biais	0,02	-0,01	-0,04
SEP(C)	0,10	0,11	0,18

La PDS donne des meilleurs résultats que la méthode DS. En effet, PDS construit des modèles multivariés locaux. Ainsi le risque de surentraînement est réduit[199] et les non linéarités sont mieux modélisées par plusieurs modèles locaux que par un modèle général.

La méthode SW a comme inconvénient majeur son caractère univarié. Quand les différences instrumentales deviennent plus complexes, la correction SW ne peut pas gérer de telles différences[199] et des résultats non satisfaisants peuvent être obtenus. Cependant dans notre étude, les instruments utilisés sont tous de même type et donc les différences

entre instruments sont faibles. Un autre avantage de cette méthode est son intégration au logiciel qui pilote l'instrument. Ainsi dans le cadre d'une application industrielle, la gestion des équations de standardisation est facilitée. Nous avons donc choisi d'utiliser cette méthode.

- Bilan sur les cinq instruments

Le Tableau 43 montre l'influence de la standardisation à l'aide de trente échantillons de betterave et de la méthode SW. On constate que quatre instruments ont un SEP inférieur au seuil fixé de 0,15 g / 100 g. Cependant d'après le test de Fisher, la différence entre l'instrument 3 et le maître est significative. En effet, le $F_{calculé}$ est 1,960 et le $F_{critique}$ de 1,412. Il en est de même pour l'instrument 2. On constate également que le SEP de l'instrument 5 reste important.

Tableau 43 ■ Effet de la modification spectrale par la méthode SW avec 30 échantillons de betterave.

	Maître	Instrument 1	Instrument 2	Instrument 3	Instrument 4	Instrument 5
SEP	**0,10**	**0,11**	**0,12**	**0,14**	**0,11**	0,16
Biais	0,01	0,02	0,01	0,03	0,02	0,04
SEP(C)	0,10	0,11	0,10	0,13	0,10	0,15

3.4.4 Conclusions sur la gestion de plusieurs instruments

Trois approches ont été proposées pour résoudre le problème du transfert d'étalonnage (Tableau 44). Les trois méthodes permettent de réduire l'amplification du SEP quand un instrument esclave est utilisé. Cependant la première solution qui consiste à développer un modèle robuste au changement d'instruments semble être la plus satisfaisante.

L'inconvénient de cette méthode est qu'*a priori* on ne peut pas encore savoir si un nouvel instrument sera adapté au modèle. Pour le moment, il semble nécessaire d'analyser des échantillons sur les nouveaux instruments. Il faut espérer qu'à terme, lorsque le modèle sera suffisamment robuste, il ne soit plus nécessaire de reconstruire un modèle lors de l'intégration d'un nouvel instrument.

Cette méthode permet d'utiliser un ensemble d'instruments sans analyser d'échantillons supplémentaires. De plus, si cette méthode n'est pas suffisante, il est possible de corriger les biais des instruments esclaves.

Tableau 44 ■ Comparaison des trois approches pour la gestion d'un réseau de cinq instruments. En gras, les SEP inférieurs à 0,15 g / 100 g.

	Instrument 1	Instrument 2	Instrument 3	Instrument 4	Instrument 5
SEP avant standardisation	0,17	0,17	0,19	**0,12**	0,62
Ajustement des biais	**0,12**	**0 ,10**	**0,14**	**0,08**	0,18
Modèle robuste	**0,11**	**0,11**	**0,09**	**0,09**	**0,14**
Correction spectrale	**0,11**	**0,12**	**0,14**	**0,11**	0,16

4 Automatisation du spectromètre de laboratoire

4.1 Principe de l'automatisation

Comme il a été possible de doser la teneur en saccharose de la betterave par SPIR sur plusieurs instruments en conservant une erreur acceptable, un analyseur automatique a été développé. La Figure 43 présente l'instrument automatique utilisé. Cet appareil est conçu sur la base d'un spectromètre traditionnel équipé d'un module d'échantillonnage automatique. L'instrument automatique permet de réaliser le cycle : présentation de l'échantillon, analyse SPIR, vidange du module automatique et archivage des spectres. Pour l'instant, un technicien doit introduire de façon manuelle l'échantillon dans le bol automatique. Mais à terme, cette étape sera également automatisée.

Le système de remplissage automatique est un bol cylindrique en acier inoxydable. Au fond de ce bol se trouve une fenêtre en quartz. Un bras motorisé avec une raclette en caoutchouc étale l'échantillon de betterave sur cette fenêtre. Après la phase d'étalement, le module de remplissage automatique va se déplacer au dessus du spectromètre au cours de l'analyse. Enfin, l'échantillon est expulsé par une trappe amovible située au fond du bol.

Par rapport à l'instrument de laboratoire où la mesure se fait en configuration verticale, sur l'instrument automatique, elle est horizontale. La seconde modification par rapport à l'instrument de laboratoire concerne l'optique. Sur l'instrument automatique, la distance entre les détecteurs et l'échantillon est plus grande de 5 mm par rapport à celle sur l'instrument de laboratoire. Ces deux modifications peuvent entraîner des modifications du spectre et donc des écarts de prédiction plus importants.

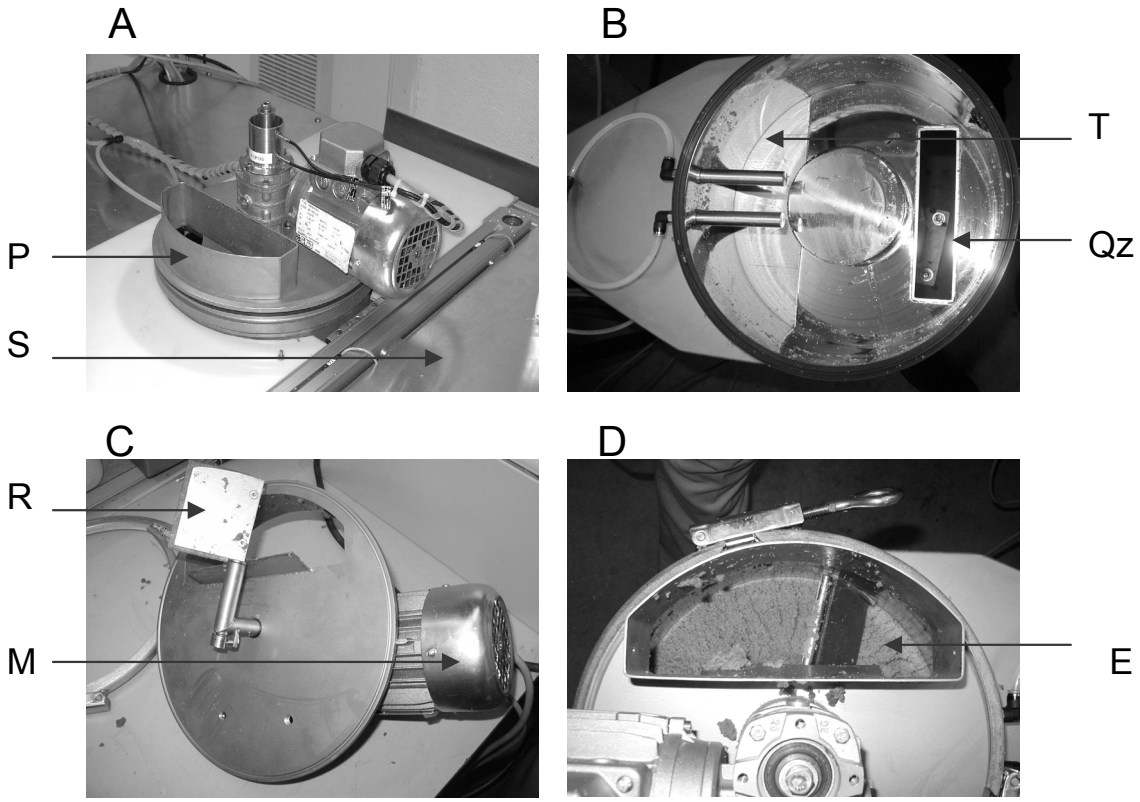


Figure 43 ■ Principe de l'automatisation. (A) vue externe de l'instrument automatique – P : unité d'étalement (passeur d'échantillons), S : spectromètre (B) vue interne – Qz : fenêtre de quartz, T : trappe de vidange (C) bras motorisé permettant d'étaler l'échantillon – R : raclette, M : moteur (D) bras en mouvement – E : échantillon

4.2 Résultats des essais en conditions de réception

Des essais ont été réalisés en usines au cours de la campagne 2002. L'objectif est de montrer la possibilité de l'automatisation et de développer un étalonnage prenant en compte la spécificité de l'instrument automatique.

La démarche est la suivante : tout d'abord, le modèle développé sur l'instrument de laboratoire est utilisé avec le nouvel instrument, ensuite un modèle spécifique à l'instrument automatique est mis en place. Enfin un modèle construit sur une base de données contenant à la fois des spectres de l'instrument automatique et de l'instrument de laboratoire est testé. Tous les modèles sont développés avec la méthode mPLS et des prétraitements SNVD + D2 sur la gamme spectrale comprise entre 1100 nm et 2498 nm.

Les principaux résultats sont présentés dans le Tableau 45. On constate que le modèle de l'instrument du laboratoire ne donne pas des résultats satisfaisants. Les différences entre les deux instruments sont donc suffisamment importantes pour influencer la

prédiction. On constate qu'il y a une erreur systématique entre l'instrument de laboratoire et l'instrument automatique. Cependant un modèle spécifique à l'instrument automatique peut être développé. De plus, il semble intéressant de conserver la base de données de l'instrument de laboratoire en la fusionnant avec celle de l'instrument automatique. Il est donc possible d'automatiser la mesure spectrale en maintenant une erreur de prédiction faible. Cependant environ 4 % des échantillons ne sont pas bien étalés à la surface du quartz. Ces spectres hors normes sont détectés en utilisant un GH limite et éliminés de la base de données.

Tableau 45 ■ Validation sur un lot de spectres de l'instrument automatique (636 échantillons indépendants (une semaine complète de mesure) analysés avec le module automatique).

	SEP	Biais	SEP(C)
Modèle 2002 spécifique à l'instrument de laboratoire (2360 échantillons dans le lot d'étalonnage)	0,19	-0,14	0,12
Modèle spécifique à l'instrument automatique (1718 échantillons dans le lot d'étalonnage)	**0,10**	0,01	0,10
Modèle mixte (fusion des deux bases spectrales précédentes)	**0,11**	0,02	0,10

5 Bilan

La détermination de la teneur en saccharose de la betterave peut se faire par spectroscopie proche infrarouge de façon précise. Les deux problèmes majeurs de cette méthode sont la validité du modèle lorsque les échantillons changent ou lorsqu'un nouvel instrument est utilisé. C'est pourquoi, une réflexion sur la gestion de la base de données spectrales et d'un réseau d'instruments a été menée. La gestion du réseau d'instruments pourrait être réalisée de la façon suivante.

Nous supposons que lors de l'année n, il existe une base de données spectrales contenant les spectres de l'instrument maître réalisés de façon manuelle ou automatique et environ 100 spectres par instrument esclave. Cette base est donc représentative de la variabilité temporelle, géographique, variétale de l'échantillon et de la variabilité instrumentale. De plus, nous supposons que des valeurs limites du biais et du SEP(C) ont été fixées par les utilisateurs de la méthode.

L'objectif est de développer avant la nouvelle campagne (année n+1) un modèle valide tenant compte des nouvelles variabilités liées à l'échantillon et à l'utilisation d'un spectromètre supplémentaire. Le protocole à réaliser avant la campagne de mesure pourrait être le suivant :

- Dans un premier temps, 150 échantillons d'origines géographiques diverses sont analysés. Le modèle de l'année n est utilisé pour valider ces échantillons. Si les valeurs du SEP et du biais sont supérieures aux valeurs limites, ces échantillons sont rajoutés dans la base de données et un nouveau modèle est développé.

- Le second temps concerne l'utilisation d'un nouvel instrument. Tout d'abord, une centaine d'échantillons sont analysés sur cet instrument et incorporés à la base d'étalonnage afin de construire un nouveau modèle. Ensuite, 50 échantillons sont analysés pour valider ce modèle. Si le SEP et le biais de la validation sont inférieurs aux valeurs limites, le nouvel instrument est validé. Dans le cas contraire, la solution consiste à effectuer une correction de biais ou une correction des spectres de cet instrument. L'inconvénient majeur de cette seconde solution est la nécessité d'analyser une trentaine d'échantillons de betterave sur l'instrument maître et sur le nouvel esclave.

Enfin, il semble nécessaire de contrôler régulièrement l'ensemble des instruments du réseau au cours de la campagne de mesure. Pour détecter des problèmes instrumentaux, les diagnostics de l'instrument (répétabilité spectrale, précision des longueurs d'onde et test des différentes composantes de l'instrument) seront réalisés de façon journalière. De plus, l'analyse des cellules de contrôle telles que des cellules de soja ou des kits commerciaux doit permettre de détecter des dérives instrumentales pouvant perturber la détermination de la teneur en saccharose. Dans le cas où les biais des différents instruments sont corrigés, il faudrait également vérifier en début de campagne à l'aide du nouveau modèle la valeur du biais de chacun des instruments du réseau. Ce contrôle des biais nécessite l'analyse d'échantillons de betterave avec les instruments esclaves et avec la méthode chimique de référence.

Chapitre 5

Valorisation de la méthode spectrale par la caractérisation de la qualité globale de l'échantillon

1 Objectifs

Ce chapitre a pour but de montrer le potentiel de la méthode SPIR pour la caractérisation de la qualité globale de la betterave et donc de valoriser la méthode spectroscopique. Dans une première partie, des critères quantitatifs tels que la teneur en sodium ou en azote seront déterminés par SPIR. Puis dans une seconde partie, des critères qualitatifs tels que l'origine géographique de l'échantillon seront étudiés grâce à des méthodes de classification supervisées.

2 Détermination de critères quantitatifs

2.1 Protocole

Le protocole de l'analyse quantitative décrit dans le chapitre 1 a été utilisé pour développer des modèles déterminant les paramètres suivants : le brix, le marc, le sucre mélasse, la pureté du jus, la teneur en azote, en potassium, en sodium et en glucose. Les échantillons ont été analysés à la fois par SPIR et par les méthodes chimiques de référence présentées dans le chapitre 3. Les modèles sont ensuite validés sur un lot d'échantillons indépendants. Le nombre d'échantillons analysés se trouve dans le Tableau 46.

Comme pour la détermination du saccharose par SPIR, les modèles ont été optimisés. Différents prétraitements et différentes méthodes de régression ont été utilisés.

Les résultats sont similaires à ceux obtenus avec le saccharose. Ils ne seront pas détaillés. Pour chacun des constituants, la méthode mPLS et une combinaison de prétraitements (SNVD et une dérivée seconde) ont donné les résultats les plus satisfaisants. Tous les modèles sont développés sur le domaine [1100 ; 2500 nm].

Pour juger de l'efficacité des modèles, des indicateurs statistiques ont été utilisés : SEP, biais, SEP(C) et R^2. De plus, deux autres paramètres sont également calculés[200] : le RPD et RER. Quand un modèle a un RPD supérieur à 3, c'est-à-dire que le SEP est petit comparé à l'écart-type des valeurs de référence, il est considéré comme satisfaisant pour la quantification. Un modèle ayant un RPD compris entre deux et trois permet de discriminer les échantillons ayant une faible concentration de ceux ayant une concentration élevée. Quand les écarts-types des valeurs de référence sont faibles, le RPD et le R^2 ne sont pas satisfaisants. C'est pourquoi, le RER est utilisé. Un RER supérieur à dix peut être considéré comme satisfaisant.

2.2 Résultats et discussion

- ***Détermination par SPIR des neuf composés***

Le Tableau 46 montre les principaux résultats. Il décrit les données utilisées et les principaux résultats obtenus lors de l'étalonnage et de la validation. De plus, la Figure 44 représente les valeurs prédites en fonction des valeurs de référence.

Tableau 46 ■ Détermination des composés de la betterave.

	Brix	Marc	Pureté du jus	Sucre Mélasse	Azote	Potassium	Sodium	Glucose
Données								
Unité	g / 100 g	g / 100 g	g / 100 g	g / 100 g	g / 100 g	$mmol.kg^{-1}$	$mmol.kg^{-1}$	g / 100 g
Minimum	17,00	3,48	91,31	0,87	0,21	2,15	0,02	0,01
Maximum	26,24	5,33	96,98	2,45	2,33	6,25	1,48	0,34
Moyenne	20,85	4,36	95,55	1,21	0,62	3,60	0,33	0,05
Ecart type	1,36	0,32	0,68	0,19	0,24	0,66	0,23	0,03
Etalonnage								
Nombre d'échantillons	1025	218	994	994	994	994	994	994
SEC	0,17	0,11	0,29	0,07	0,10	0,38	0,12	0,01
R^2	0,98	0,91	0,79	0,75	0,74	0,58	0,42	0,49
Validation								
Nombre d'échantillons	955	198	1066	1066	1066	1066	1066	1066
SEP	0,19	0,13	0,31	0,08	0,11	0,41	0,14	0,01
Biais	0,00	0,00	0,03	-0,01	0,00	-0,01	-0,02	0,00
SEP(C)	0,19	0,13	0,31	0,08	0,11	0,41	0,14	0,01
Pente	1,01	1,02	0,95	0,97	0,94	0,93	0,81	0,71
R^2	0,98	0,83	0,74	0,71	0,64	0,48	0,32	0,31
RPD	7,16	2,42	2,20	2,38	2,24	1,61	1,63	2,47
RER	48,63	14,02	18,23	19,75	20,19	10,05	10,43	27,33

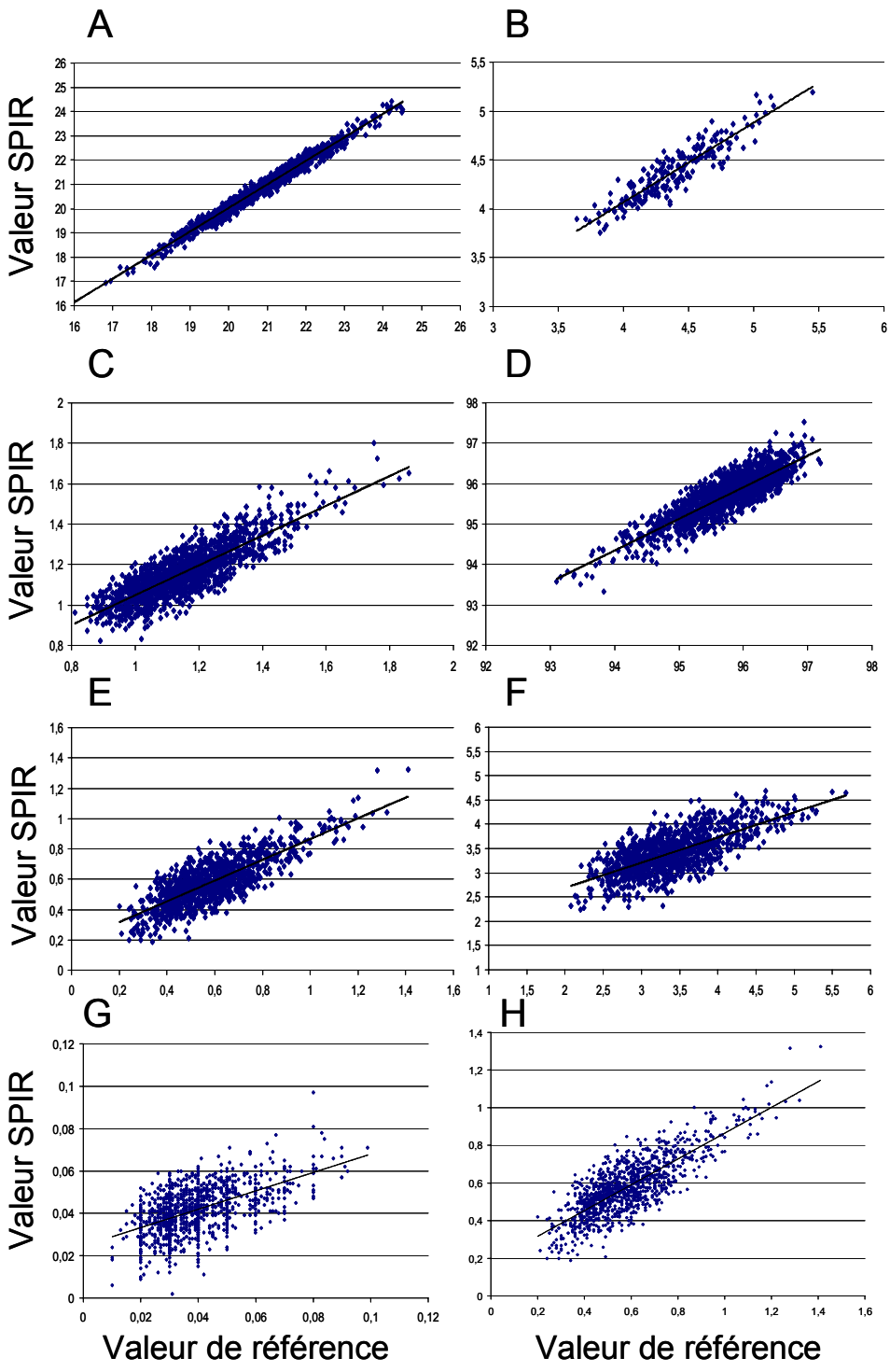


Figure 44 ■ Résultats de la validation pour les différents composés de la betterave. En abscisses les valeurs de référence et en ordonnées les valeurs fournies par la SPIR. (A) Brix, (B) Marc, (C) Pureté du jus, (D) Sucre mélasse, (E) Azote alpha aminé, (F) Potassium, (G) Glucose et (H) Sodium. Unité : g / 100 g sauf pour le sodium et le potassium exprimés en 10^{-3}.mol.kg^{-1}.

On constate que le brix est bien prédit par la SPIR. Le SEP est faible, le R^2 est grand et le RPD est proche de 7. Ceci peut s'expliquer par le fait que le saccharose représente 80 % de la matière sèche de la betterave. Comme le saccharose peut être déterminé précisément par SPIR, le brix peut l'être également.

En ce qui concerne le marc, le sucre mélasse, la pureté du jus, l'azote et le glucose, les modèles développés sont moins satisfaisants. Cependant les RPD sont compris entre deux et trois. Ces modèles peuvent donc être utilisés pour le criblage ou le tri des échantillons. Le marc, la pureté et le sucre mélasse sont mieux prédits que l'azote et le glucose.

Pour le potassium et le sodium le RPD est faible, cependant le RER est supérieur à dix. Ces modèles peuvent être utiles pour discriminer les échantillons ayant des concentrations élevées d'échantillons ayant des concentrations faibles.

Pour un modèle, la précision recherchée dépend de son utilisation. Ainsi, pour la détermination du saccharose la meilleure précision a été recherchée car le résultat obtenu détermine une valeur marchande. Par contre pour d'autres paramètres comme la teneur en sodium, la précision requise est moins grande car ce critère pourra être utilisé pour un tri entre les échantillons.

Les causes du manque de précision de certains modèles peuvent être les suivantes :

- La précision des méthodes de référence n'est pas suffisante.
- Des erreurs sont dues aux modèles mathématiques.
- L'information n'est que faiblement identifiée dans le spectre proche infrarouge (cas des ions).
- Les concentrations sont trop faibles pour être bien déterminées par SPIR.

La première hypothèse a été contrôlée. Trente échantillons ont été analysés trois fois sur la chaîne analytique et par SPIR. Le Tableau 47 regroupe les principaux résultats. On constate que les méthodes de référence sont précises et que la première hypothèse est à rejeter.

Tableau 47 ■ Précision des analyses chimiques de référence

	Brix g / 100 g	Marc g / 100 g	Pureté du jus %	Sucre mélasse %	Azote alpha-aminé g / 100 g	Potassium $mmol.kg^{-1}$	Sodium $mmol.\ kg^{-1}$	Glucose g / 100 g
Moyenne	20,85	4,36	95,55	1,21	0,62	3,60	0,33	0,05
Ecart type	1,36	0,32	0,68	0,19	0,24	0,66	0,23	0,03
Ecart type de répétabilité de la méthode chimique	0,05	0,08	0,04	0,01	0,03	0,04	0,01	0,01
Ecart type de répétabilité de la méthode SPIR	0,09	0,11	0,10	0,05	0,06	0,11	0,03	0,01

- ***Coefficients de régression***

Comme le montre la figure suivante, les zones où les coefficients de régression sont les plus importants sont :

- la zone entre 1100 et 1300 nm qui correspond aux secondes harmoniques des liaisons C-H.
- la zone entre 1600 et 1800 nm qui correspond à la première harmonique de la liaison C-H.
- la zone entre 2100 et 2300 nm qui correspond aux bandes de combinaison C-H+N-H et C-H+C-C.

Par contre, on constate que les bandes de l'eau centrées sur 1450 nm et sur 1950 nm ont toujours des coefficients de régression faibles. L'information de ces zones n'est pas utilisée pour déterminer les teneurs des différents constituants.

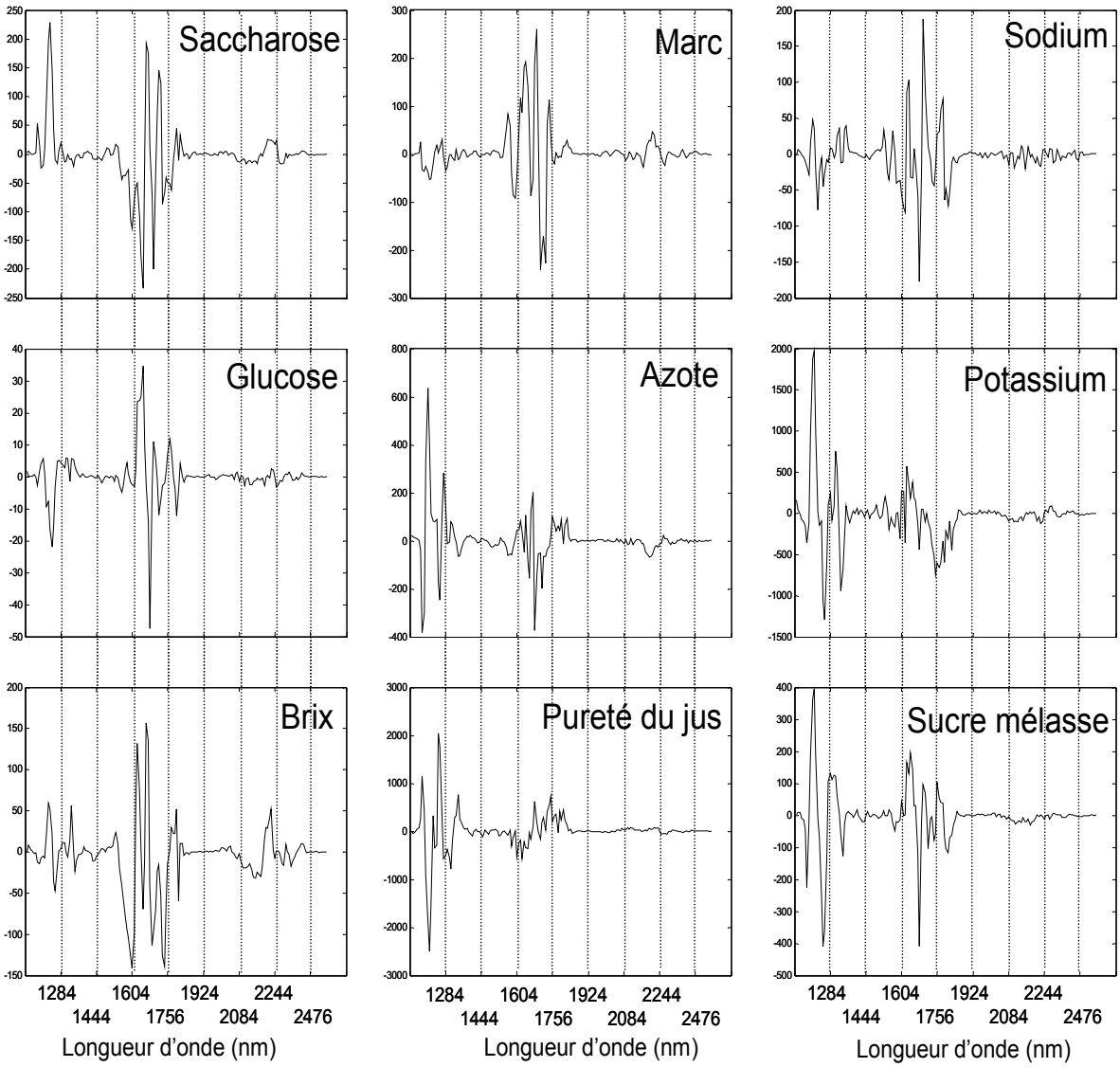


Figure 45 ■ Coefficients de régression des différents modèles

Nous avons étudié les corrélations entre les valeurs prédites des différents composés (Tableau 48). On remarque que les coefficients sont très élevés entre la pureté du jus et l'azote ou le potassium. Ceci est dû à la façon de calculer la valeur de la pureté.

La corrélation entre le sucre mélasse et la pureté du jus est compréhensible. En effet, plus le jus avant cristallisation est pur, moins il y a de pertes de saccharose dans la mélasse. C'est pourquoi quand la pureté augmente, la valeur du sucre mélasse diminue. Des constituants minoritaires peuvent être prédits par spectroscopie proche infrarouge car ils sont en fait très corrélés à d'autres constituants.

Tableau 48 ■ Corrélation entre les valeurs prédites (525 échantillons).

En gras les valeurs du R^2 supérieures à 0,70

	Saccharose	Brix	Marc	Pureté du jus	Sucre Mélasse	Azote	Potassium	Sodium	Glucose
Saccharose	1,00	**0,99**	0,57	0,59	-0,22	-0,27	-0,61	0,07	-0,10
Brix		1,00	0,59	0,52	-0,14	-0,21	-0,58	0,15	-0,07
Marc			1,00	0,54	-0,22	-0,26	-0,63	-0,20	0,10
Pureté du jus				1,00	**-0,73**	**-0,77**	**-0,73**	-0,62	-0,08
Sucre Mélasse					1,00	0,48	0,49	0,60	0,42
Azote						1,00	0,55	0,52	0,01
Potassium							1,00	0,24	0,05
Sodium								1,00	0,10
Glucose									1,00

3 Détermination de critères qualitatifs

Les trois critères étudiés sont les suivants : la résistance à la rhizomanie, la région de culture et la période de récolte. L'objectif est d'optimiser la classification des échantillons de betterave en fonction de ces trois critères qualitatifs. Ainsi, différents algorithmes de classification seront évalués. De même, les classifications effectuées, soit avec les données chimiques de référence, soit avec les données spectrales seront comparées. Pour effectuer ces comparaisons, il a été nécessaire d'utiliser un test statistique : le test de McNemar.

3.1 Critères étudiés

3.1.1 Résistance à une maladie la Rhizomanie (RR)

Les betteraves atteintes de rhizomanie ont une prolifération massive de leurs racines secondaires et une atrophie de la racine principale. Le lot de données d'étalonnage

est constitué de 83 échantillons dont 42 sont résistants à cette maladie (notés R) et 41 qui ne le sont pas (notés NR). Le lot de validation est lui aussi composé de 83 échantillons dont 42 échantillons sont NR et 41 sont R. Les échantillons NR et R ont été répartis de façon aléatoire entre le lot de validation et le lot d'étalonnage.

Les échantillons résistants à la rhizomanie ont des variétés diverses (et des fabricants différents) : Rhist (Deleplanque), Shérif (SES), Rosana (KWS). Il en est de même pour les échantillons non résistants : Sonate (SES), Lynx (Deleplanque), Ariana et Roberta (KWS).

3.1.2 Origines géographiques (OG)

Les différentes origines sont représentatives de la production française. Elles sont codées par des chiffres de 0 à 7 (Tableau 49). Pour chaque origine, les échantillons sont répartis de façon aléatoire entre le lot d'étalonnage et le lot de validation. Le lot d'étalonnage est composé de 320 échantillons (40 échantillons par origine). Il en est de même du lot de validation qui contient 40 échantillons pour chacune des 8 origines.

Tableau 49 ■ Origines géographiques des betteraves

Code	Région
0	Marne
1	Calvados
2	Somme
3	Seine et Marne
4	Aisne
5	Eure
6	Pas de calais
7	Loiret

3.1.3 Période de récolte (PR)

La période de récolte est divisée en quatre mois de septembre à décembre (codée par 0, 1, 2 et 3). Pour chacune des périodes, les échantillons sont répartis en deux lots : étalonnage et validation. Le lot de données d'étalonnage contient 262 échantillons : 37

de septembre, 75 d'octobre, 75 de novembre et 75 de décembre. Le lot de validation est également composé de 262 échantillons ayant la même répartition dans chaque classe.

3.2 Résultats et discussion

3.2.1 Analyse des données par ACP

Dans une première étape, l'ACP est utilisée pour l'exploration des données. Pour chacun des trois lots d'étalonnage, les plans 1-2 et 3-4 de l'ACP sont représentés dans la Figure 46 et Annexe 3. Il n'est pas possible de trouver des séparations simples entre les différents groupes des lots OG et PR. Par contre sur le lot RR, il semble possible de séparer les échantillons résistants des non résistants en fonction des axes 3 et 4. Cependant la frontière n'est pas clairement définie et les groupes se chevauchent. On remarque ainsi que l'ACP et en général les méthodes de classification non supervisées ne donnent pas forcément les groupes recherchés. C'est pourquoi il a été nécessaire d'utiliser des méthodes de classification supervisées.

3.2.2 Résultats et discussion concernant les classifications supervisées

- ***Optimisation des méthodes***

Avant de comparer les méthodes, il faut s'assurer que les modèles qualitatifs sont correctement construits. Chaque méthode est optimisée séparément et on compare ensuite le meilleur résultat de chaque méthode. Ainsi pour les méthodes PDA et DPLS, le nombre de composantes principales est déterminé par la méthode de validation croisée. Ces deux méthodes utilisent 16 composantes sur les trois lots. Ce nombre influence le taux de bonnes classifications en validation. Ainsi, par exemple sur le lot RR, un modèle DPLS construit avec 12 composantes principales a un taux de prédiction correct de 63,9 % alors qu'un modèle construit avec 16 composantes donne 80,2 % de bonnes réponses. Il est vrai que le nombre de composantes est grand, mais il faut remarquer que pour prédire la teneur en saccharose ce nombre est compris en 10 et 14.

Concernant la méthode CART, le nombre de nœuds est déterminé par une validation croisée. Pour les lots RR, GO et PR, on utilise respectivement 6,12 et 14 nœuds.

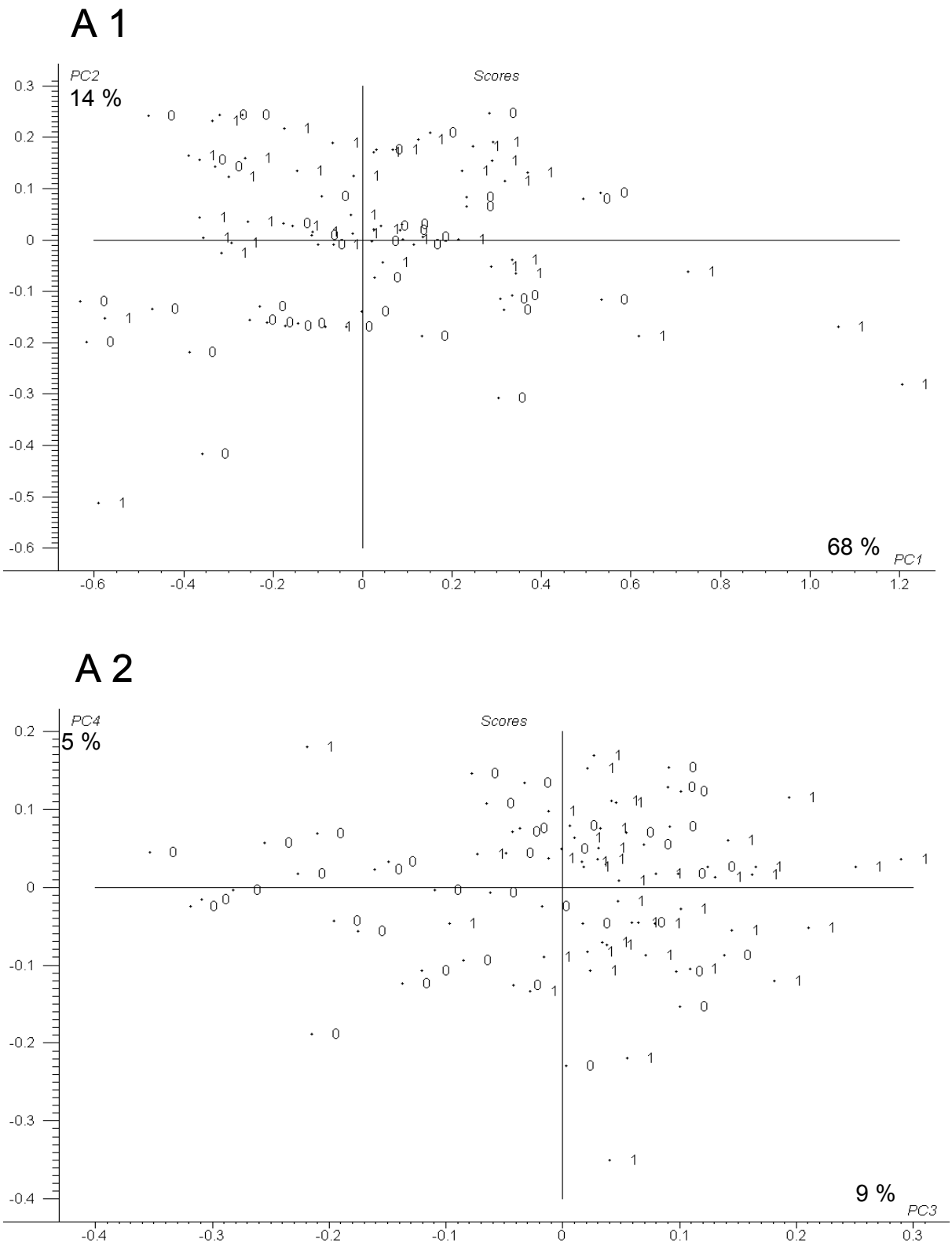


Figure 46 ■ Analyse en composantes principales du lot concernant la résistance à la rhizomanie (Pour les autres lots d'étalonnage utilisés pour la classification *cf.* Annexe 3) (A1) Plan 1-2 (A 2) Plan 3-4.

Le nombre des plus proches voisins K est important pour la méthode KNN. Les valeurs K sont testées au cours d'une validation croisée. Le taux de bonne prédiction de la validation croisée varie entre 61 % pour K = 11 et 77 % pour K = 3 sur le lot RR, entre 49 % pour K=13 et 58 % pour K = 5 pour les données OG et enfin entre 61 % (K = 9) et 67 % (K = 3) pour les données PR. Les valeurs de K utilisées sont donc 3, 5 et 3 pour les données RR, OG et PR respectivement.

Concernant la méthode LDA, vingt composantes principales sont extraites et les coordonnées factorielles résultantes sont utilisées comme données d'entrée de LDA. Ces 20 composantes représentent 99,99 % de la variance totale pour les trois lots de données.

L'algorithme PNN est largement influencé par le facteur de lissage σ. Si σ est proche de zéro le réseau fonctionne comme la méthode KNN avec K = 1. Le taux de prédiction est de 48,8 % si σ vaut 0,1 pour le lot RR. L'optimal est à 10 (Le taux de prédiction est 75,6 %).

Le nombre de neurones cachés ainsi que le nombre de cycles de l'apprentissage influencent les résultats de la méthode LVQ. Le nombre minimum de cycles est 100. Le nombre optimum de neurones cachés est 9 pour le lot DR, 20 pour OG et 9 pour PR. Ce nombre influence de façon significative le taux de prédiction. Par exemple, le taux de prédiction est de 14 % avec 2 neurones, 54 % avec 4, 62 % avec 9 et 59 % avec 10 sur le lot de validation de PR.

- ***Comparaison des méthodes***

Les méthodes de classification sont comparées à partir des résultats des lots de validation (c'est-à-dire respectivement 83, 320 et 262 échantillons pour les lots de données RR, OG et PR). La moyenne des taux de prédiction des huit méthodes est : 74,4 % pour RR, 53 % pour OG et 68,4 % pour PR. Chacun des trois paramètres a donc pu être prédit à partir des données spectrales.

Les trois tableaux suivants comparent les huit méthodes de classification utilisées sur les trois lots de données. Les taux de prédiction sont présentés dans ces tableaux. De plus, toutes les méthodes sont comparées par paire avec le test de McNemar. Les méthodes donnant les meilleurs résultats sur le lot RR sont les suivantes : PDA, SIMCA, LDA et KNN. Le test de McNemar montre que ces trois méthodes ne sont pas significativement différentes (Tableau 50). Sur les données OG, les méthodes les plus précises sont SIMCA et DPLS. Elles ont des taux de prédiction non différents (Tableau 51). Enfin concernant le lot PR, DPLS est significativement la méthode la plus précise (Tableau 52).

Tableau 50 ■ Comparaison des méthodes de classification supervisées pour le lot de données RR. En gras, les valeurs significatives du test de McNemar.

Méthodes (Taux de prédiction)	PDA (85,5 %)	SIMCA (81,9 %)	LDA (80,7 %)	KNN (75,9 %)	PNN (69,8 %)	LVQ (69,8 %)	DPLS (66,2 %)	CART (65,1 %)
PDA	-	0,16	0,41	2,22	**5,33**	**7,58**	**8,65**	**7,75**
SIMCA		-	0	0,76	3,38	**4,05**	3,69	**5,63**
LDA			-	0,37	3,22	2,78	**5,28**	**9,33**
KNN				-	1,56	0,64	2,7	3,7
PNN					-	0,03	0,13	0,26
LVQ						-	0,45	0,45
DPLS							-	0
CART								-

Tableau 51 ■ Comparaison des méthodes de classification supervisées pour le lot de données OG. En gras, les valeurs significatives du test de McNemar.

Méthodes (Taux de prédiction)	SIMCA (69,0 %)	DPLS (66,5 %)	PDA (61,5 %)	LDA (59,3 %)	CART (53,4 %)	KNN (52,1 %)	PNN (48,1 %)	LVQ (12,5 %)
SIMCA	-	0,43	**5,38**	**8,37**	**20,8**	**23,02**	**33,25**	**155,02**
DPLS		-	2,83	**3,94**	**11,83**	**15,34**	**26,27**	**137,6**
PDA			-	0,28	**5,48**	**7,37**	**14,34**	**121,04**
LDA				-	2,79	**4,21**	**10,37**	**120,65**
CART					-	0,07	2,4	**94,41**
KNN						-	2,25	**95,06**
PNN							-	**80,18**
LVQ								-

Tableau 52 ■ Comparaison des méthodes de classification supervisées pour le lot de données PR. En gras, les valeurs significatives du test de McNemar.

Méthodes (Taux de prédiction)	DPLS (94,2 %)	SIMCA (81,3 %)	PDA (77,1 %)	CART (73,2 %)	LDA (69,8 %)	KNN (65,2 %)	LVQ (62,5 %)	PNN (28,6 %)
DPLS	-	**21,875**	**30,68**	**37,87**	**49,61**	**62,5**	**71,8**	**168,14**
SIMCA		-	1,42	**3,81**	**19,11**	**14,24**	**21,63**	**132,17**
PDA			-	0,9	3,02	**10,58**	**141,01**	**85,81**
CART				-	0,57	**4,21**	**10,22**	**78,66**
LDA					-	1,14	**4,13**	**102,42**
KNN						-	1,23	**58,8**
LVQ							-	**38,32**
PNN								-

- ***Précision des méthodes***

Nous pouvons considérer que les trois méthodes les plus précises sur les données utilisées sont SIMCA, PDA et DPLS. Elles ont des points communs. En effet, elles sont linéaires, paramétriques et utilisent des méthodes de régression. L'avantage de ces méthodes est qu'elles peuvent gérer des données hautement colinéaires, des données manquantes, des variables bruitées et des classes qui se chevauchent[201]. C'est pourquoi elles sont le mieux adaptées à nos données spectrales.

LDA, KNN et CART sont moins précises que les méthodes précédentes. KNN ne donne pas de bons résultats car les spectres proche infrarouge contiennent beaucoup d'informations redondantes[201]. Le désavantage de LDA est qu'elle ne peut pas être appliquée à des données ayant plus de variables que d'échantillons. De plus, LDA s'applique de façon optimale quand la dispersion des classes est égale et de même direction[69]. Les méthodes non linéaires n'ont pas donné de résultats satisfaisants sur les données de notre étude.

- ***Autres critères de comparaison***

Le premier critère pour choisir la méthode est la précision. Si plusieurs méthodes ont la même précision d'autres critères peuvent être appliqués. Trois autres critères de comparaison peuvent être utilisés : la détection des échantillons hors normes, la mémoire vive nécessaire aux calculs et l'interprétation des résultats.

Les méthodes de classification supervisées attribuent à chaque échantillon une classe. Les méthodes utilisant des distances (toutes sauf CART) peuvent détecter des échantillons hors normes. En général, si la distance entre l'échantillon inconnu et le centre de gravité des échantillons du lot d'étalonnage est grande alors l'échantillon testé est considéré comme un échantillon aberrant. Mais le choix de la valeur critique est difficile. Dans notre étude, les lots de données ont été construits, il n'y a donc pas d'échantillons aberrants en validation c'est-à-dire des échantillons appartenant à une classe qui ne se trouvait pas dans le lot d'étalonnage.

La mémoire vive de l'ordinateur est également un critère important, surtout quand on utilise une large base de données. KNN et PNN ont besoin de plus de mémoire que les autres méthodes. En effet, les deux méthodes stockent l'information du lot d'étalonnage pour réaliser la classification. Sur des lots d'étalonnage plus grands que ceux de notre étude, ce stockage peut être un inconvénient majeur pour ces deux méthodes.

- ***Test de McNemar***

Le test de McNemar est un test conservatif sur des lots de données de petite taille (100 échantillons). L'écart significatif entre deux méthodes est élevé pour le lot RR : la différence entre deux taux de prédiction doit être supérieure à 10 % pour être significative. Avec des lots de données plus grands, la différence significative entre deux méthodes est plus faible : 5 % pour OG et 4 % pour PR.

La valeur de McNemar calculée est comparée à une valeur critique. Certains résultats sont difficiles à interpréter car la valeur critique est un seuil. Par exemple, sur le lot de données RR, la méthode SIMCA a 81,9 % de chances de donner la bonne classe et elle est significativement différente de LVQ qui a un taux de prédiction de 69,8 %. Mais elle n'est pas différente de la méthode DPLS qui a un taux de bonnes classifications de 66,2 %. En effet, la valeur de McNemar est de 3,69 quand DPLS est comparé à SIMCA. Cette valeur est juste en dessous de la valeur critique $\chi^2_{(1\ ;\ 0,95)}$.

Le test de McNemar est cependant un outil utile pour déterminer si deux résultats issus de deux classifications sont significativement différents. Si deux méthodes ont le même taux de prédiction, le choix de l'algorithme peut être ensuite basé sur des critères pratiques.

Tableau 53 ■ Bilan : comparaison des méthodes de classification supervisées.

	Méthode	SIMCA	DPLS	PDA	LDA	CART	KNN	PNN	LVQ
Caractéristiques des méthodes	Linéaire	oui	oui	oui	oui	oui	oui	non	non
	Paramétrique	oui	oui	oui	oui	non	non	non	non
	Discriminant (D) ou Modélisation des classes (MC)	MC	MC	MC	D	D	D	D	D
Critères qualitatifs	Rapidité d'entraînement	+	+ +	+ +	++	+ +	+ +	+	+
	Mémoire nécessaire	faible	faible	faible	faible	faible	élevée	élevée	faible
Précision	Lot RR	+ +	+	+ +	+	+	+	-	-
	Lot PR et OG	+ +	++	+ +	+	+	+	-	-

- ***Détermination des critères qualitatifs de la betterave***

Les trois critères qualitatifs de la betterave sont prédits avec un pourcentage élevé de bonnes réponses : 85,5 % pour la résistance à la rhizomanie avec la méthode PDA, 69 % pour l'origine géographique (méthode SIMCA) et 94,2 % pour la période de récolte (classification DPLS).

Tableau 54 ■ Table de contingence pour les méthodes donnant les meilleurs résultats sur les trois lots de données en validation.

α = 1 - <n_i> / n_i avec n_i : le nombre d'échantillons $\in$ i, <n_i> nombre d'échantillons $\in$ i et classés dans la classe i

et β = 1 - <$\overline{n_i}$> / $\overline{n_i}$ avec $\overline{n_i}$: nombre d'échantillons $\notin$ i et <$\overline{n_i}$> nombre d'échantillons $\notin$ i et classés comme non-i.

(I) Résistance à la rhizomanie – Méthode PDA

Classe prédite	R	NR	α	β
Classe R	**30**	11	0,27	0,02
Classe NR	1	**41**	0,02	0,27

(II) Origine géographique – Méthode SIMCA

Classe prédite	0	1	2	3	4	5	6	7	α	β
Classe 0	**24**	0	9	0	4	1	2	0	0,40	0,01
Classe 1	1	**26**	6	0	1	0	3	3	0,35	0,04
Classe 2	0	1	**33**	0	3	1	0	2	0,17	0,17
Classe 3	1	2	6	**28**	0	3	0	0	0,30	0,01
Classe 4	0	5	6	0	**26**	1	1	1	0,35	0,08
Classe 5	0	0	5	2	9	**20**	0	4	0,50	0,02
Classe 6	1	1	7	0	1	0	**28**	2	0,30	0,02
Classe 7	0	0	2	0	2	0	0	**36**	0,1	0,05

(III) Période de récolte – Méthode DPLS

Classe prédite	0	1	2	3	α	β
Classe 0	**33**	4	1	0	0,05	0,01
Classe 1	2	**72**	0	0	0,04	0,03
Classe 2	0	1	**70**	4	0,07	0,02
Classe 3	0	0	3	**72**	0,04	0,02

Pour le lot RR, on remarque que les échantillons non résistants sont bien classés. La majorité des erreurs est due aux échantillons résistants qui sont mal classés (Tableau 54).

Pour le lot OG, les régions 7 et 2 sont globalement mieux prédites que les autres. Comme peu d'échantillons mal classés sont attribués à la classe 3 ou 0, on peut penser que ces deux classes ont des caractéristiques spécifiques. Au contraire, beaucoup d'échantillons

sont classés à tort dans la classe 2, cette classe ne semble pas avoir de caractéristiques spécifiques et doit recouvrir les autres.

Concernant PR, le taux de bonne classification est élevé cependant on constate que les confusions s'effectuent de façon majoritaire entre les classes 0 et 1 ou entre les classes 2 et 3. Il semble que la différence entre les échantillons analysés au début de la campagne (classes 0 et 1) et la fin (classes 2 et 3) soit plus grande que celle existante entre les classes 0 et 1 ou entre les classes 2 et 3.

3.2.3 Comparaison des classifications utilisant les données spectrales ou les données des méthodes chimiques de référence

- ***Protocole***

Dans cette seconde partie concernant la classification, seule la méthode SIMCA est utilisée. On effectue des classifications à partir des données chimiques de référence (teneur en saccharose, glucose, azote, sodium et potassium) et des données spectrales. Des lots de données spectrales (RAN) où les classes sont attribuées au hasard sont également utilisés. Les lots de données RAN contiennent les mêmes spectres et ont le même nombre de classes.

- ***Résultats***

Pour les données spectrales, les taux de prédiction sont les suivants : 81,9 % avec le lot RR, 69,0 % avec OG et 81,3 % pour le lot PR. En ce qui concerne la classification effectuée à partir des données chimiques de référence, les pourcentages de bonnes classifications sont : 80,3 % avec RR, 30,1 % avec OG et 60,0 % avec PR. Les classifications obtenues avec les lots RAN sont : 45,8 % pour RR, 12,0 % pour OG et 22,0 % avec le lot PR. Les modèles SIMCA utilisent 16 composantes principales par classe pour les données spectrales et 3 composantes principales pour les données chimiques de référence.

Le Tableau 55 compare les classifications effectuées avec les données spectrales ou les données de chimie. On constate que pour deux des lots testés (OG et PR), les données spectrales améliorent la prédiction. Par contre, concernant la détermination de la résistance à la rhizomanie, les résultats obtenus à partir des données spectrales ou des données chimiques de référence sont similaires.

Tableau 55 ■ Comparaison des classifications effectuées à partir des données chimiques ou spectrales : table de contingence de McNemar pour les trois lots de données.

Méthode A : SPIR - Méthode B : Polarimétrie (D'après les formules de la table de McNemar – *cf.* Tableau 4) n_{01} : Nombre d'échantillons mal classés par A mais pas par B. n_{10} : Nombre d'échantillons mal classés par B mais pas par A.

(I) Résistance à la rhizomanie

n_{00} = 2	n_{01} = 16
n_{10} = 15	n_{11} = 50
	Valeur de McNemar = 0 (NS)

(II) Origine Géographique

n_{00} = 63	n_{01} = 28
n_{10} = 148	n_{11} = 81
	Valeur de McNemar = 80.46 (S)

(III) Périodes de récoltes

n_{00} = 41	n_{01} = 8
n_{10} = 114	n_{11} = 99
	Valeur de McNemar = 90.37 (S)

- ***Discussion***

Globalement les classifications obtenues à partir des données spectrales sont meilleures que celles obtenues à partir des données chimiques de référence. De plus, les taux de classification obtenus à partir des données RAN sont celles d'une attribution au hasard. Donc, on constate que l'information modélisée est bien réelle.

Pour le critère RR, les deux lots de données (SPIR et données de référence) donnent des résultats similaires. L'hypothèse pour expliquer ce résultat est la suivante : les betteraves qui sont résistantes à la rhizomanie sont globalement moins riches en sucre. Or, comme l'information teneur en sucre est contenue dans les deux lots de données, la résistance à la rhizomanie est déterminée avec la même précision à partir des données

spectrales ou des données de chimie. Pour vérifier cette hypothèse, des classifications ont été effectuées en supprimant l'un des composants des données chimiques. On constate que si le glucose, le potassium, le sodium ou l'azote alpha aminé sont supprimés un à un, on obtient respectivement 78,3 %, 75,9 % 75,9 % et 77,1 %. Par contre, lorsque le saccharose est supprimé le taux de classifications correctes est de 69,1 %. Mais la concentration en saccharose seule ne permet pas de classer les échantillons selon le critère de résistance.

Pour les deux autres lots de données (OG et PR), on constate que les classifications effectuées à partir des données chimiques ont des taux de prédiction corrects inférieurs à ceux obtenus avec les données spectrales. Les données chimiques ne contiennent pas l'information nécessaire à la classification. On peut supposer que la teneur en eau aurait pu être une information utile pour classer les échantillons selon leur origine et selon la période de récolte. Les données spectrales contiennent cette information au travers des bandes de l'eau qui sont centrées sur 1450 nm et sur 1950 nm. Les spectres proche infrarouge contiennent en effet des informations sur la majorité des composants de la betterave alors que les données chimiques contiennent uniquement les concentrations de 5 constituants.

L'utilisation des prétraitements spectraux améliore la classification. Dans notre étude, les prétraitements ont été testés avec la méthode DPLS. L'utilisation de la normalisation SNVD et de la dérivée seconde donne les résultats les plus précis. Les taux de prédiction obtenus sur les trois lots de données RR, OG, PR sont respectivement 73,9 %, 62 % et 85 % sans prétraitement et 85,5 %, 69 % et 94,2 % avec prétraitements. Le test de McNemar montre que les prétraitements améliorent de façon significative les résultats de la classification. Nos résultats sont en accord avec les études précédentes[202] qui montraient l'influence des prétraitements sur les résultats de la classification supervisée. Ces transformations réduisent en effet la variation spectrale due à la taille des particules.

4 Bilan

Ce chapitre a montré le potentiel de la méthode spectroscopique. En effet, la SPIR permet de doser rapidement le saccharose de la betterave mais elle permet également d'estimer d'autres composés (sodium, potassium, glucose et azote) et de déterminer d'autres paramètres (sucre mélasse, pureté du jus) qui sont utiles au fonctionnement de l'usine.

La deuxième partie de ce chapitre a montré que des critères qualitatifs peuvent également être prédits à partir des spectres. La détermination des critères de résistance à la rhizomanie, de l'origine géographique de la betterave et de la période de récolte a pu être réalisée à partir de SPIR avec un taux de prédiction important pour les trois critères respectivement 85 %, 69 % et 95 %. D'un point de vue plus théorique, l'utilisation du test de McNemar a permis de comparer la précision de différentes méthodes de classification supervisées et a montré l'avantage de la SPIR par rapport aux données chimiques de référence. Parmi les huit méthodes utilisées, SIMCA, DPLS et PDA semblent être les mieux adaptées à nos données spectrales. Elles sont toutes les trois linéaires, paramétriques et utilisent des méthodes de régression.

Conclusion

Les objectifs de ce sujet de thèse ont été multiples. Tout d'abord, le premier objectif a été de valider une méthode spectroscopique pour le dosage du saccharose de la betterave. Ensuite, les problèmes de transfert d'étalonnage et de validité du modèle ont été abordés pour démonter la possibilité d'utiliser la spectroscopie proche infrarouge dans un contexte industriel. Enfin le dernier objectif a été de déterminer de façon globale la qualité de la betterave pour valoriser la méthode SPIR.

L'originalité de l'étude se situe à plusieurs niveaux. Tout d'abord, l'ensemble du projet industriel a été abordé. Ainsi, la recherche s'est déroulée depuis les essais de faisabilité en laboratoire jusqu'aux tests en usine avec un prototype automatisé. De plus, le caractère pluridisciplinaire de la thèse est également important. Les thématiques abordés sont variées : chimie analytique, spectroscopie, instrumentation, chimiométrie et statistiques.

Dans un premier temps, le dosage du saccharose par SPIR est abordé. Le modèle, reliant la concentration au spectre, a été optimisé d'un point de vue chimiométrique. Dans cette étude, l'influence du domaine spectral, des méthodes de prétraitement et des méthodes de régression sur l'erreur de prédiction a été étudiée. Le meilleur étalonnage a été réalisé en utilisant le domaine spectral [1100 nm, 2500 nm], les prétraitements « Standard Normal Variate » et « detrending » associés à la dérivée seconde et la méthode de régression « Modified Partial Least Squares ». Ainsi, l'erreur standard de prédiction (SEP) obtenue est de 0,10 g de saccharose pour 100 g de betterave. Le modèle est linéaire sur toute la gamme de concentration qui s'étend de 14 à 20 g / 100 g, il n'y a pas de biais c'est-à-dire pas d'erreur systématique entre la SPIR et la méthode polarimétrique. L'erreur de prédiction est faible compte tenu du fait que le SEP tient compte de la répétabilité de la mesure de référence. La teneur en saccharose de la betterave est donc déterminée par SPIR de manière satisfaisante.

La seconde partie aborde les problèmes relatifs à la validité du modèle et à l'utilisation de SPIR dans un contexte industriel. Il est important de savoir si le modèle utilisé est valable d'une campagne à l'autre. Dans notre étude, l'ajout d'échantillons de la nouvelle campagne à la base de données spectrales permet d'améliorer la précision du modèle. *A priori*, l'analyse de 150 échantillons en début de campagne permet de vérifier si la mise à jour annuelle du modèle est nécessaire et de l'effectuer. De plus, le modèle développé sur un instrument de référence, appelé « instrument maître » peut être utilisé sur plusieurs instruments, appelés « instruments esclaves » tout en conservant un SEP faible. Enfin, l'automatisation de la mesure spectrale est réalisable. Les conditions nécessaires pour une application industrielle de la SPIR sont donc réunies.

La dernière partie montre que la SPIR offre la possibilité d'évaluer la qualité globale de l'échantillon et des critères qualitatifs tels que l'origine géographique ou la résistance à la rhizomanie. On remarque ainsi le potentiel analytique offert par la SPIR et les méthodes chimiométriques.

D'un point de vue théorique, des tests statistiques facilement applicables ont été proposés pour comparer des modèles quantitatifs et des modèles de classification. D'un point de vue applicatif, la spectroscopie proche infrarouge est capable de remplacer la polarimétrie pour le dosage du saccharose de la betterave lors de la réception en usine.

Certes l'utilisation de la SPIR dans un contexte industriel doit être contrôlée, il faut vérifier si le modèle prédisant la teneur en saccharose de la betterave doit être mis à jour et pour l'utilisation d'un réseau d'instruments, un protocole de transfert d'étalonnage doit être mis en place. Mais les avantages de la SPIR sont nombreux. Cette méthode est précise, rapide, peu coûteuse et non polluante.

ANNEXES

Annexe 1 : Procédé sucrier et SPIR

Au centre de réception :

Des betteraves sont râpées et échantillonnées pour déterminer leur propreté (tare) et leur richesse en sucre pour calculer la recette du planteur*.

Dans l'usine :

1. Le lavage.

Des betteraves sont brassées dans des tambours laveurs et débarrassées de la terre, des cailloux et des herbes.

2. La diffusion

Les betteraves sont découpées en fines lamelles appelé « cossettes », puis placées dans un courant d'eau chaude pour en extraire le sucre, par diffusion. A la sortie de l'atelier de diffusion, on obtient un jus de diffusion*.*

3. Epuration

Le jus de diffusion est purifié par un traitement à la chaux et au gaz carbonique qui se termine par une filtration. On obtient alors un jus épuré.*

4. Evaporation

Le jus épuré contenant 85 % d'eau est concentré en sirop par évaporation (dans des cuves chauffées à la vapeur). On obtient alors un sirop*.*

5. Cristallisation

Le sirop contient 65 % de sucre. Il est chauffé sous vide et concentré pour déclencher la cristallisation.*

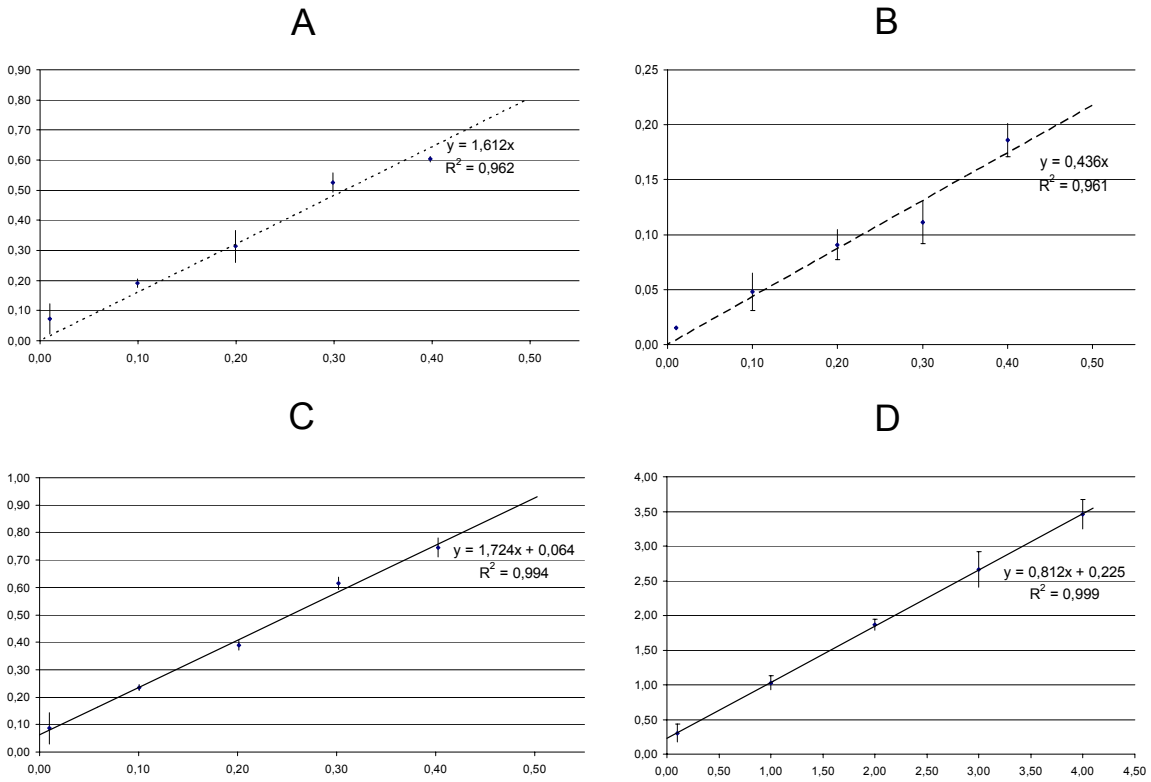
6. Centrifugation

La masse cuite produite est envoyée dans des turbines qui évacuent la phase liquide par centrifugation et retiennent le sucre blanc* cristallisé. Il est séché, puis stocké avant sa commercialisation.*

(*) Produit potentiellement analysable par SPIR.

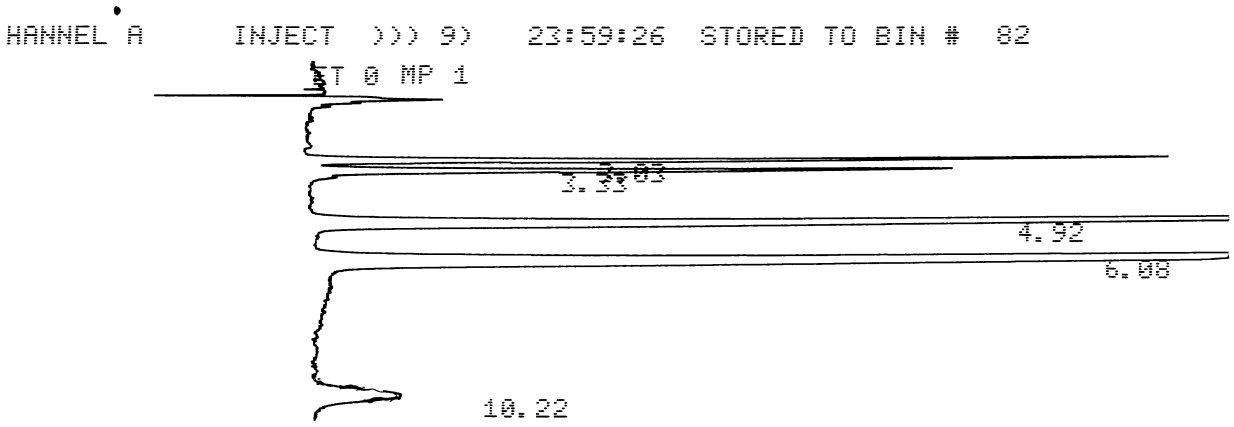
Annexe 2 : Dosage des sucres de la betterave par CLHP

- ***Etalonnage de la méthode***



Annexe 2.a : Courbe d'étalonnage du dosage des sucres par HPLC : (A) fructose, (B) raffinose, (C) glucose et (D) saccharose. En abscisse, rapport des masses : masse du glucide dosé sur la masse de l'étalon interne (lactose). En ordonnée, rapport des aires : aire du glucide dosé sur l'aire de l'étalon interne (lactose).

- ***Chromatogrammes***

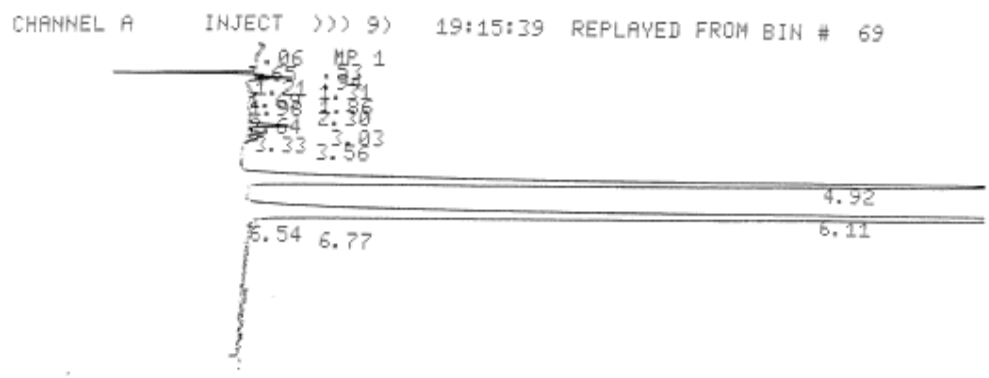


))) 9) 23:59:26 CH= "A" PS= 1.

ILE 1. METHOD 0. RUN 92 INDEX 1 CALIB BIN 82

EAK#	AREA%	RT	AREA	BC
1	10.285	3.03	32044	01
2	8.452	3.33	26334	01
3	30.219	4.92	94149	01
4	46.558	6.08	145054	01
5	4.424	10.22	13783	03
OTAL	100.		311557	

Annexe 2.b : Chromatogramme de l'étalon 2 (rapport de masse saccharose/lactose = 2 et 0,2 pour les autres sucres). RT = Temps de rétention. Pic 1 : glucose, Pic 2 : fructose, Pic 3 : lactose, Pic 4 : saccharose, Pic 5 : raffinose.



))) 9) 19:15:39 CH= "A" PS= 1.

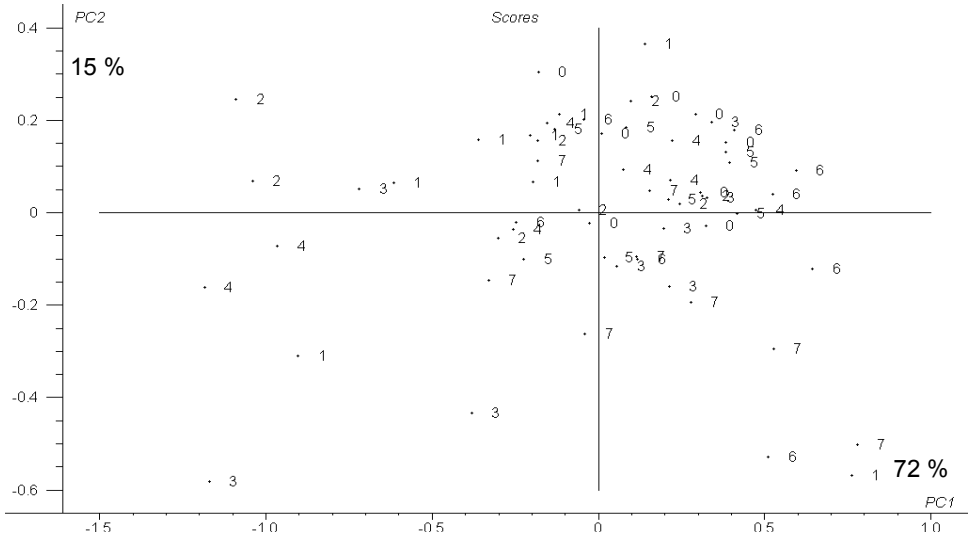
FILE 1. METHOD 0. RUN 79 INDEX 1 CALIB BIN 69

PEAK#	AREA%	RT	AREA BC
1	2.748	0.06	8252 02
2	1.371	0.53	4117 02
3	4.544	0.65	13644 02
4	3.521	0.94	10572 02
5	4.12	1.21	12371 02
6	3.483	1.31	10458 02
7	3.19	1.69	9578 02
8	1.143	1.86	3433 02
9	0.577	1.98	1733 03
10	0.067	2.3	202 01
11	0.076	2.64	228 02
12	0.818	3.03	2457 02
13	0.403	3.33	1209 02
14	0.266	3.56	799 03
15	28.663	4.92	86069 01
16	44.332	6.11	133117 02
17	0.122	6.54	365 02
18	0.26	6.77	780 03
19	0.042	9.83	125 02
20	0.04	10.28	120 02
21	0.072	10.38	217 02
22	0.064	10.71	192 02
23	0.037	10.8	112 02
24	0.041	10.89	124 03
TOTAL	100.		300274

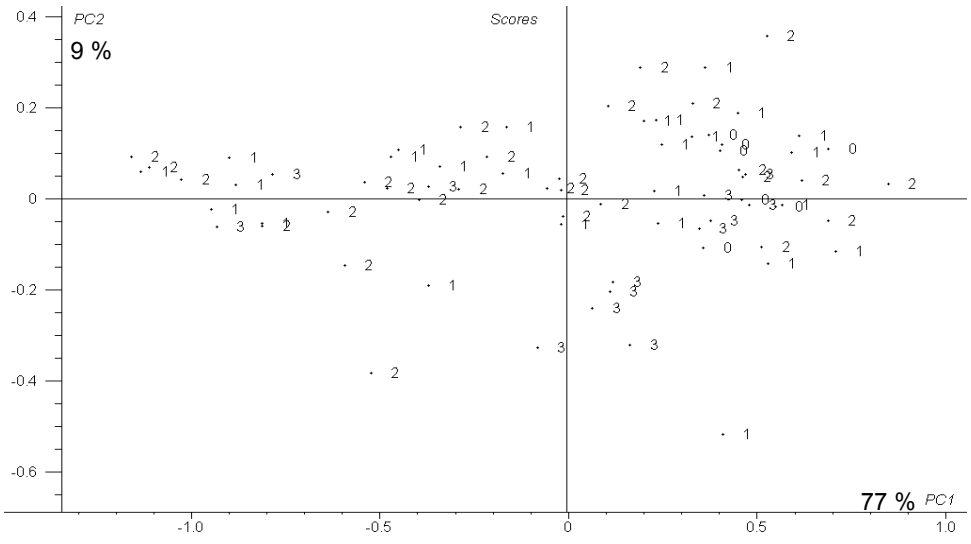
Annexe 2.c : Chromatogramme d'un échantillon de betterave (N°324). RT = Temps de rétention. Pic 12 : glucose, Pic 13 : fructose, Pic 15 : lactose, Pic 16 : saccharose.

Annexe 3 : Analyse en composantes principales sur les données de classification.

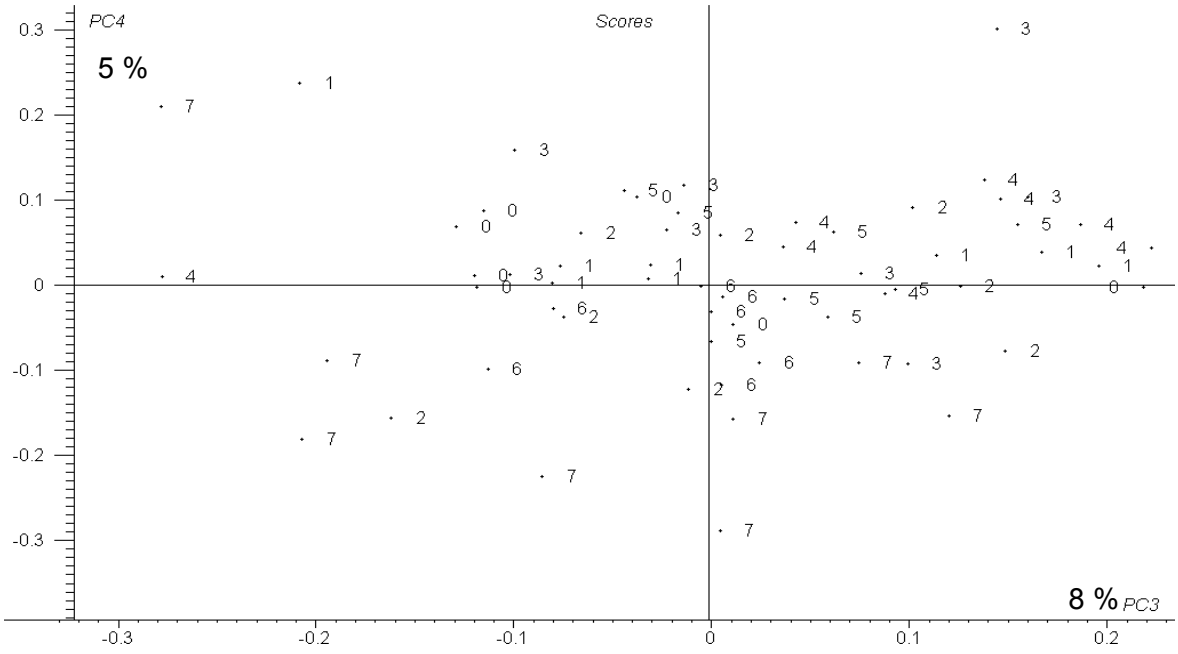
B 1



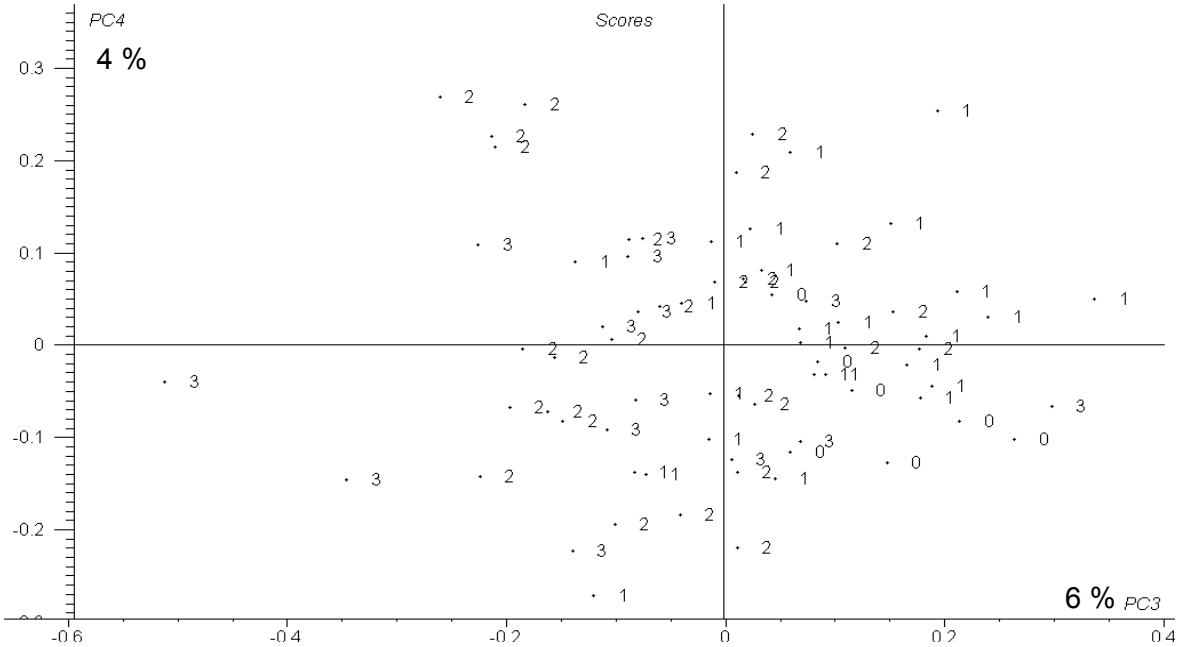
C 1



B 2



C 2



Légende :

B : lot d'étalonnage concernant l'origine géographique (OG)

C : lot d'étalonnage concernant la période de récolte (PR)

1 : Plan 1 – 2

2 : Plan 3 - 4

Annexe 4 : Communications scientifiques

PUBLICATIONS INTERNATIONALES A COMITE DE LECTURE

1. Sucrose content determination of sugar beets by near infrared reflectance spectroscopy. Comparison of calibration methods and calibration transfer.

Y. Roggo, L. Duponchel, B. Noé, J. P. Huvenne.

Journal of Near Infrared Spectroscopy, **10**, 137-150, 2002.

2. Comparison of supervised pattern recognition methods with McNemar's statistical test. Application to qualitative analysis of sugar beet by near-infrared spectroscopy.

Y. Roggo, L. Duponchel, J. P. Huvenne.

Analytica Chimica Acta, **477**, 187-200, 2003.

3. Statistical tests for comparison of quantitative and qualitative models developed with near infrared spectral data.

Y. Roggo, L. Duponchel, C. Ruckebusch, J. P. Huvenne.

Journal of Molecular Structure, **654**, 253-262, 2003.

COMMUNICATIONS ORALES

4. Applications des outils de la chimiométrie aux spectres proche infrarouge de betteraves sucrières. Dosage du saccharose et détermination de critères qualitatifs.

Y. Roggo, L. Duponchel, J. P. Huvenne.

Chimiométrie 2001.

Paris (Décembre 2001).

Prix de la communication junior attribué par la société française de chimiométrie

5. Le calcul distribué pour l'optimisation stochastique : une nouvelle voie pour la chimiométrie.

L. Duponchel, Y. Roggo, E. G. Talbi, J. P. Huvenne.

Chimiométrie 2001.

Paris (Décembre 2001).

6. Détermination de la teneur en saccharose de la betterave sucrière par spectroscopie proche infrarouge.

Y. Roggo, L. Duponchel, J. P. Huvenne.

6èmes Journées Jeunes Chercheurs.

Villeneuve d'Ascq (Décembre 2001).

7. La spectroscopie proche infrarouge, une méthode de dosage rapide pour les industries agricoles et alimentaires. Notions d'étalonnage, de validation et de transfert d'étalonnage.

Y. Roggo, L. Duponchel, J. P. Huvenne.

Colloque sciences alimentaires de l'ENSIA.

Massy (Mai 2002).

8. Apport de la chimiométrie au transfert d'étalonnage entre spectromètres proche infrarouge. Application au dosage du saccharose de la betterave.

Y. Roggo, L. Duponchel, J. P. Huvenne.

Chimiométrie 2002.

Paris (Décembre 2002).

9. Sucrose content determination of sugar beet by near infrared spectroscopy: calibration, calibration transfert and automation of NIRS measurement.

Y. Roggo, L. Duponchel, J. P. Huvenne.

11th International Conference on Near Infrared Spectroscopy.

Cordoba (Avril 2003).

COMMUNICATIONS PAR AFFICHES

10. Optimisation des méthodes chimiométriques par les algorithmes génétiques.

L. Duponchel, Y. Roggo, E. G. Talbi, J. P. Huvenne.

Chimiométrie 2001.

Paris (Décembre 2001). ***Prix du meilleur poster attribué par la société française de chimiométrie***

11. La spectroscopie proche infrarouge pour la détermination des critères qualitatifs de la betterave sucrière. Comparaison de méthodes de classification supervisées.

Y. Roggo, B. Noé, L. Duponchel, J. P. Huvenne.

9ième Journées Thématiques du Groupe Français de Spectroscopie Vibrationnelle.

Nantes (Juin 2002).

12. Statistical tests for comparison of quantitative and qualitative models developed with near infrared spectral data.

Y. Roggo, L. Duponchel, C. Ruckbusch, J. P. Huvenne.

XXVI European Congress on Molecular Spectroscopy 2002.

Villeneuve d'Ascq (Septembre 2002).

13. Quantitative and qualitative evaluation of sugar beet with near infrared spectroscopy.

Y. Roggo, L. Duponchel, C. Ruckbusch, J. P. Huvenne.

Carrefour Européen des Biotechnologies (6ème édition).

Lille (Octobre 2002).

14. Mise en place d'un réseau de spectromètres proche infrarouge pour le dosage du saccharose de la betterave.

Y. Roggo, L. Duponchel, J. P. Huvenne.

7èmes Journées Jeunes Chercheurs.

Villeneuve d'Ascq (Février 2003).

15. Sucrose content determination of sugar beet in industrial context: optimisation and automation of NIRS measurement.

Y. Roggo, L. Duponchel, J. P. Huvenne.

11th International Conference on Near Infrared Spectroscopy.

Cordoba (Avril 2003).

PROCEEDINGS ET AUTRES COMMUNICATIONS ECRITES

16. Utilisation de la spectroscopie infrarouge dans les industries agricoles et alimentaires.

Y. Roggo, L. Duponchel.

Bulletin d'Information du Syndicat National des Fabricants de Sucre, 3, 5 mars 2001.

17. Sucrose content determination of sugar beet in industrial context: method development, instrument network management and NIRS automation.

Y. Roggo, L. Duponchel, J. P. Huvenne.

Proceeding of 11th International Conference on Near Infrared Spectroscopy (soumis avril 2003).

Index

P

R

S

T

Table des illustrations

FIGURES

TABLEAUX

Bibliographie

[1] P. Jurs, B. Kowalski, T. Isenhour et C. Reilly, Anal. Chem., 41 (1969) 690.

[2] B. Kowalski, Chem. Ind., 22 (1978) 882.

[3] G. Vandeginste, D. Massart, L. Buydens, S. De Jong, P. Lewi et J. Smeyers-Verbeke, Handbook of chemometrics and qualimetrics part A, Elsevier, New York, 1988, p 207.

[4] S. Wold et M. Sjöström, Chemometr. Int. Lab. Syst., 44 (1998) 3.

[5] L. Bokobza, J. Near Infrared Spectrosc., 6 (1998) 3.

[6] C. Banwell, Fundamentals of molecular spectroscopy. McGraw Hill, Londres, 1983, p 338.

[7] G. Lachenal, Introduction à la spectroscopie infrarouge. *In* : La spectroscopie infrarouge et ses applications analytiques, D. Bertrand et E. Dufour (Ed.), Tec&Doc, Paris, 2000, p 32.

[8] M. Dalibart et S. Servant, Spectroscopie dans l'infrarouge, techniques de l'ingénieur, PA, 2000, p 2845.

[9] C. Cohen-Tannoudji, B. Diu et F. Laloe, Mécanique quantique, tome1, Hermann, Paris, 1996, p 480.

[10] F. Cadet, Analyse multidimensionnelle de données et spectroscopie moyen infrarouge, deuxième partie, thèse pour obtenir l'habilatation à diriger des recherches, Université de la Réunion, France, 1992.

[11] R. Lauer et E. Rosenbaum, Appl. Spectrosc., 6 (1954) 29.

[12] K. Whetsel, Appl. Spectrosc. Rev., 2 (1968) 1.

[13] L. Duponchel, Validation par les méthodes chimiométriques d'un analyseur spectrométrique de lait travaillant dans le proche infrarouge. Transfert de calibration en vue de développements industriels, thèse de doctorat (N°2085), Université des Sciences et Techniques de Lille, France, 1997.

[14] B. Osborne, T. Fearn et P. Hindle, Pratical NIR spectroscopy with application in food and beverage analysis. Prentice Hall, Harlow, 1993, p 227.

[15] A. Riaublanc, D. Bertrand et E. Dufour, Lipides. *In* : La spectroscopie infrarouge et ses applications analytiques, D. Bertrand et E. Dufour (Ed.), Tec&Doc, Paris, 2000, p 139.

[16] E. Dufour et P. Robert. Protéines. *In* : La spectroscopie infrarouge et ses applications analytiques, D. Bertrand et E. Dufour (Ed.), Tec&Doc, Paris, 2000, p 107.

[17] M. Iwamoto, Proceedings International Near Infrared Diffuse Reflectance Spectroscopy Conference, (1986) 1.

[18] D. Bertrand, Spectroscopie de l'eau. *In* : La spectroscopie infrarouge et ses applications analytiques, D. Bertrand et E. Dufour (Ed.), Tec&Doc, Paris, 2000, p 93.

[19] J. Lin et C. Brown, Appl. Spectrosc., 47 (1993) 62.

[20] H. Begley, E. Lanza, K. Norris et W. Hruschka, J. Agr. Food Chem., 32 (1984) 984.

[21] F. Cadet, M. Safar et E. Dufour, Glucides. *In* : La spectroscopie infrarouge et ses applications analytiques, D. Bertrand et E. Dufour (Ed.), Tec&Doc, Paris, 2000, p 171.

[22] J. Workman, Appl. Spectrosc. Rev., 31 (1996) 251.

[23] B. Osborne, T. Fearn, A. Miller et S. Douglas, J. Sci. Food Agr., 35 (1984) 99.

[24] J. Diffee, Tobacco analysis by NIR spectroscopy. *In* : Handbook of Near-infrared Analysis, D. Burns, W. Ciurczak, Dekker Inc., New York,1992 p 433.

[25] M. Meurens et G. Alfaro, Belg. J. Food, 45 (1990) 63.

[26] G. Trott, E. Woodside, K. Taylor et J. Deck, Carbohyd. Res., 51 (1973) 133.

[27] D. Law et R. Tkachuk, Cereal Chem., 54 (1977) 65.

[28] P. Kubelka, J. Opt. Soc. Am., 38 (1948) 448.

[29] P. Kubelka et F. Munk, Z. Technische Phys., 12 (1931) 593.

[30] W. Butler et K. Norris, Arch. Biochem. Biophys., 87 (1960) 31.

[31] M. Feinberg, L'assurance qualité dans les laboratories agroalimentaires et pharmaceutiques, 2eme éditonTec&Doc, Paris, 2001, p 355.

[32] J. Workman, NIR spectroscopy calibration basics, *In* : Handbook of Near-Infrared Analysis, D. Burns et E. Ciurczak, Dekker, New York, 1992, p 247.

[33] H. Martens et T. Naes, Multivariate calibration, Wiley&Sons, Chichester, 1989, p 419.

[34] M. Dhanoa, S. Lister et R Sanderson, J. Near Infrared Spectrosc., 2 (1994) 43.

[35] J. Olinger et P. Griffiths, Appl. Spectrosc., 47 (1993) 163.

[36] T. Isksson et B Kowalski, Appl. Spectrosc., 47 (1993) 6.

[37] L. Arakaki et D. Burns, Appl. Spectrosc., 46 (1992) 1919.

[38] Y. Ozaki, T. Miura, K. Sakurai et T. Matsunaga, Appl. Spectrosc., 46 (1992) 875.

[39] P. Levillain et D. Pompeydie, Analusis, 14 (1986) 1.

[40] K. Norris et P. Williams, Cereal Chem., 661 (1984) 158.

[41] J. Rodriguez-Oterio, M. Hermida et A. Cepeda, J. Assoc. Off. Am. Chem., 78 (1995) 802.

[42] A. Savitsky et M. Golay, Anal. Chem., 36 (1964) 1627.

[43] T. Anderson, The statistical analysis of time series. John Wiley&Sons, New York, 1971, p 288.

[44] L. Rabiner et B. Gold, Theory and application of digital signal processing, 2sd edition, Prentice-hall, Englewood Cliffs, 1975, p 332.

[45] R. Barnes, M. Dhanoa et S. Lister, Appl. Spectrosc., 43 (1989) 772.

[46] P. Geladi, D. McDougall et H. Martens, Appl. Spec., 39 (1985) 491.

[47] T. Naes et T. Isaksoon, Anal. Chem. , 62 (1990) 664.

[48] E. Bertan, H. Iturriaga, S. Maspoch et I. Montoliu, Anal. Chim. Acta, 431 (2001) 303.

[49] M. Manley, A. McGill et B. Osborne, J. Near Infrared Spectrosc., 2 (1994) 93.

[50] M. Dhanoa, S. Lister, R. Sanderson et R. Barnes, J. near Infrared Spectrosc., 2 (1994) 43.

[51] L. Lebart, A. Morineau et M. Piro, L'Analyse en Composantes Principales. *In* : Statistiques exploratoires multidimensionnelles, Dunod, Bordas, Paris, 1997, p 32.

[52] M. Danzart, Statistique descriptive. *In* : SSHA&ISHA, Analyse sensorielle. Manuel méthodologique, Tec&Doc, Paris, 1990, p 209.

[53] L. Lebart, A. Morineau et N. Tabard, Techniques de la description statistique. Méthodes et logiciels pour l'analyse des grands tableaux. Dunod, Bordas, Paris, 1987, p 351.

[54] P. Geladi et B. Kowalski, Anal. Chim. Acta, 185 (1986) 1.

[55] C. Brown, P. Lynch, R. Obremski et D. Lavery, Anal. Chem., 54 (1982) 1472.

[56] D. Haaland, Appl. Spectrosc., 2 (1987) 56.

[57] H. Mark, Anal. Chem., 58 (1986) 2814.

[58] S. Chatterjee et B. Price, Regression analysis by example, 2nd edition, Wiley & sons, New York, 1997, p 304.

[59] L. Xu et W. Zhang, Anal. Chim. Acta., 446 (2001) 477.

[60] K. Mardia, J. Kent et J. Bibby, Multivariate Analysis, Wiley & sons, New York, 1981, p 430.

[61] K. Beebe et B. Kowalski, Anal. Chem., 59 (1987) 1007A.

[62] W. Linberg, J. Pearson et S. Wold, Anal. Chem., 55 (1983) 643.

[63] H. Wold, Multivariate Analysis, Academic Press, New York, 1975, p 307.

[64] M. Tenenhaus, La régression PLS : théorie et pratique, Edition Technip, Paris, 1998, p 254.

[65] E. Vigneau, E. Qannari et M-F Devaux, Méthodes prédictives. *In* : La spectroscopie infrarouge et ses applications analytiques, Bertrand D., Dufour E., Tec&Doc Lavoisier, Paris, 2000, p 295.

[66] D. Haaland et E. Thomas, Anal. Chem., 60 (1988) 1193.

[67] G. Baffi, E. Martin et A. Morris, Comput. Chem. Eng., 23 (1999) 395.

[68] F. Despagne et D. Massart, Analyst, 123 (1998) 157R.

[69] M. Sharaf, D. Illman et B. Kowalski, Chemometrics, Wiley-interscience publication, New York, 1986, p 228.

[70] T. Cover, IEEE T. Infor. Theory, 13 (1967), 21.

[71] M. Derde, L. Buydens, D. Massart et P. Hopke, Anal. Chem., 59 (1987) 1868.

[72] D. Coomans, M. Jonckheer, D. Masssart, I. Broeckaert et P. Block, Anal. Chim. Acta, 103 (1978) 409.

[73] S. Wold, Pattern Recogn., 8 (1976) 127.

[74] L. Stahle et S. Wold, J. Chemometr., 1 (1987) 185.

[75] D. Gonzalez-Arjona, G. Lopez-Perez et A. Gonzalez, Chemometr. Intell. Lab. Syst., 57 (2001) 133.

[76] L. Breiman, R. Friedman, R. Olsen et C. Stone, Classification and regression trees, Wadsworth, Pacific Grove, 1984.

[77] T. Kohonen, Self-organization and associative memory, Springer Verlag, Berlin, 1989, p 236.

[78] T. Kohonen, Proc. IEEE, 78 (1990) 1464.

[79] D. Specht, Neural Networks, 3 (1990) 109.

[80] D. Coomans et D. Massart, Anal. Chim. Acta, 138 (1982) 153.

[81] D. Massart, B. Vandeginste, S. Deming, Y. Michotte et L. Kaufman, Chemometrics: a textbook, Elsevier, New York, 1988, p 488.

[82] B. Alsberg, R. Goodacre. J. Rowland et D. Kell, Anal. Chim. Acta, 348 (1997) 389.

[83] A. Burnham, R. Viveros et J. McGregor, J. Chemometr., 10 (1996) 31.

[84] D. Gonzalez-Arjona, G. Loperez-Perez et A. Gonzalez, Talanta, 49 (1999) 189.

[85] C. Cappelli, F. Mola et R. Siciliano, Comput. Stat. Data An., 38 (2002) 285.

[86] A. Pallara, Statisca Applicata, 4 (1992) 255.

[87] H. Demuth et M. Beale, Neural Network Toolbox: User's Guide version 4, The Math Works Inc., Natick, 2001.

[88] L. Simon et M. Karim, Biochem. Eng. J., 7 (2001) 41.

[89] T. Fearn, NIR News, 5 (1996) 5.

[90] L. Stahle et S. Wold, Chemometr. Intell. Lab. Syst., 6 (1989) 259.

[91] P. Dagnelie, Statistique théorique et appliquée. 2, Inférence statistique à une et à deux dimensions, De Boeck Université, Bruxelles, 1998, p 480.

[92] B. Everitt, The analysis of contingency tables. 2sd edition Chapman et Hall, London, 1992, p 164.

[93] W. Chapman, M. Fizman, B. Chapman et P. Hary, J. Biomed. Inform., 34 (2001) 4.

[94] C. Tan, Y. Wang et C. Lee, Inform. Process. Manag., 38 (2002) 329.

[95] T. Dietterich, Neural Comput., 10 (1998) 1895.

[96] V. Bellon-Maurel. Application de la spectroscopie proche infrarouge au contrôle en ligne de la qualité des fruits et des légumes. Thèse de doctorat, Institut polytechnique, Toulouse, France (1992).

[97] A. Smith, Applied Infrared Spectroscopy Fundamentals, Techniques, and Analytical Problem-Solving. John Wiley&Sons, New York.1979, p 322.

[98] D. Bertrand, Instrumentation. *In* : La spectroscopie infrarouge et ses applications analytiques, D. Bertrand et E. Dufour (Ed.), Tec&Doc, Paris, 2000, p 213.

[99] A. Michelson, Philosophical Magazine, 5 (1891) 31.

[100] P. Griffiths et J. De Haseth, Fourier transform infrared spectroscopy, Wiley-interscience, New York, 1986, p 425.

[101] J. Chatard, Nouvelle revue d'aéronautique et d'astronomie, 4 (1995) 33.

[102] D. Bertrand, Utilisation des analyses multidimensionnelles en spectroscopie de reflexion dans le proche infrarouge : application à la caractérisation de la qualité technologique du blé et de ses produits de mouture, Thèse de doctorat, Université de Dijon, France, 1988.

[103] R. Miller et H. Willis, J. Appl. Chem., 6 (1956) 385.

[104] B. Buchanan et D. Honigs, Appl. Spectrosc., 41 (1987)1388.

[105] T. Ky, F. Drouart et Y-M Lucot, La betterave, plante miracle. Edition La vague verte, Woignarue, 1997, p 205.

[106] D. Cooke et R. Scott, The beet sugar crop, Chapman & Hall, London, 1993, p 412.

[107] J. Boiffin et E.Choppin de Janvry, L'implantation de la betterave industrielle, Publications INRA, Paris, 1994, p 170.

[108] K. Mahn, C. Hoffmann et B. Märländer, Eur. J. Agron., 17 (2002) 29.

[109] H. Schiweck, G. Kozianowski, J. Anderlei et M. Burba, Zuckerindustrie, 1198 (1994) 268.

[110] J. Hobbis, J. Kysilka et M. Holle, La sucrerie belge, 101 (1982) 49.

[111] Bulletin Officiel de la République Française, Décret du 24 avril (1964).

[112] ICUMSA, Méthode GS 6-1 (1994).

[113] J.-L. Guignard, Biochimie végétale, Dunod, Paris, 2000, p 274.

[114] ICUMSA, ICUMSA Proceedings19 th session, ICUMSA publication, Peterborough, UK, 1986, p 66.

[115] ICUMSA, Spécification et normes SPS -1,1994.

[116] C. Audigié, G. Dupont et F. Zonszain, Principes des méthodes d'analyse biochimique, tome 1, Doin, Paris, 1982, p 190.

[117] M. Mathlouthi et P. Reiser, Le saccharose : propriétés et applications, Polytechnica, Paris, 1995, p 315.

[118] J. Ceirwyn, Analytical chemistry of foods. Blackie academic & professional, London 1995, p 178.

[119] J. Montreuil et A. Verbert, Techniques de l'ingénieur, PA (1997) 3320.

[120] S. Bichsel, ICUMSA Proceedings 20th session. ICUMSA publication, Peterborough, UK, 1990, p 352.

[121] J. Peschet et A. Giacalone, Industries Agricoles et Alimentaires, 108 (1991) 583.

[122] C. Garcia-Jares et B. Médina, Fresen. J. Anal. Chem., 357 (1997) 86.

[123] H. Cho et D. Hong, J. Near Infrared Spectrosc., 6 (1998) A329.

[124] ICUMSA, Méthode GS1/2/3-4 (1998).

[125] Y. Pomeranz et C. Meloan, Food analysis. Theory and practice. Chapman & Hall, New York, 1994, p 778.

[126] A. Carruthers et J. Oldfield, Proc. 11th Session CITS, Elsevier, Amsterdam, 1962, p 224.

[127] S. Moore et W. Stein, J. Biol. Chem., 211 (1954) 907.

[128] P. Trinder et D. Webster, Ann. Clin. Biochem., 21 (1984) 430.

[129] P. Devillers, R. Detavernier et L. Roger, Sucrerie Française, 116 (1975) 299.

[130] D. Barham et P. Trinder, Analyst, 97 (1972) 142.

[131] ICUMSA, Réfractométrie et table officielle SPS-3, 2000.

[132] U. Beiss, Zuckerindustrie,106 (1981) 820.

[133] P. Devillers, Sucrerie Française, 129 (1988) 1190.

[134] K. Norris et J. Hart, Principles and methods of measuring moisture in liquids and solids. Vol.4, Reinhold, New York, 1965, p 19.

[135] C. Burks, F. Dowell et F. Xie, J. Stored Produ. Res., 36 (2000) 289.

[136] M. Meurens, W. Li, M. Foulon et V. Acha, Cerevisa, 20 (1995) 33.

[137] U. Wählby et C. Skjoöldebrand, J. Food Eng., 47 (2001) 303.

[138] R. Frankhuizen, NIR analysis of dairy products. *In* : Handbook of Near-infrared Analysis, Burns D., Ciurczak W., Dekker Inc., New York, 1992, p 609.

[139] B. Osborne, Utilisation de la spectroscopie proche infrarouge dans les industries céréalières. *In* : La spectroscopie infrarouge et ses applications analytiques, Bertrand D., Dufour E., Tec&Doc Lavoisier, Paris, 2000, p 423.

[140] M. Martin, F. Pablo et A. Gonzalez, Anal. Chim. Acta, 350 (1996) 191.

[141] D. Gonzalez-Arjona, V. Gonzalez-Gallero, F. Pablo et A. Gonzalez, Anal. Chim. Acta, 381 (1999) 257.

[142] C. Armanino, R. De Acutis et M. Festa, Anal. Chim. Acta, 454 (2002) 315.

[143] L. Simon and M. Karim, Biochem. Engineering J., 7 (2001) 41.

[144] A. Candolfi, W. Wu, D. Massart et S. Heuerding, J. Pharmaceut. Biomed., 16 (1998) 1229.

[145] R. Bucci, A. Magri, A. Magri, D. Marini et F. Marini, J. Agr. Food Chem., 50 (2002) 413.

[146] C. Armanino, R. De Acutis et M.-R. Festa, Anal. Chimi. Acta, 454 (2002) 315.

[147] J. Kim, A. Mowat, P. Poole et N. Kasabov, Chemometr. Intell. Lab. Syst., 51 (2000) 201.

[148] B. Steur, H. Shulz et E. Lager, Food Chem., 72 (2001) 113.

[149] H. Meyer, Proceedings South Africa Sugar Technologists Association, 1988, p 9.

[150] H. Meyer, Int. Sugar J., 100 (1998) 279.

[151] C. Sverzut, L. Verma et A. French, Am. Soc. Agr. Eng., 30 (1987) 258.

[152] G. Vaccari et G.Mantovani, Zuckerindustrie, 114 (1989) 75.

[153] M. Clark, B. Legendre, L. Edye et C. Scott, Sugar J., 61 (1997) 22.

[154] G. Vaccari, G. Mantovani et G. Sguldino, Sugar J., 54 (1990) 4.

[155] L. Edye et M. Clarke, Proceedings of the Conference on Sugar Processing Reseach, 1996, p 350.

[156] A. Salgo, J. Nagy et E. Miko, J. Near Infrared Spectrosc., 6 (1998) A101.

[157] S. Heppner, K. Thielecke, K. Buchholz et D. Wullbrandt, Zuckerindustrie, 125 (2000) 325.

[158] J. De Bruijn, Int. Sugar J., 97 (1995) 147.

[159] M. Steegmans et H. Hoeberg, La Sucrerie Belge, 116 (1998) 30.

[160] G. Vaccari, G. Mantovani, G. Sguldino et P. Goberti, Zuckerindustrie, 112 (1987) 800.

[161] G. Vaccari, G. Mantovani et G. Sguldino, Sugar J., 51 (1988) 4.

[162] J. De Bruijn, Zuckerindustrie, 122 (1997) 878.

[163] M. Clark, L. Edeye, X. Miranda et C. Scott, Sugar Ind. Technol., 7 (1994) 81.

[164] L. Edeye et M. Clarke, Zuckerindustrie, 120 (1995) 284.

[165] E. Burzawa et M. Melle, Industries Agricoles Alimentaires, 105 (1988) 629.

[166] G. Marchetti, L'industria Saccarifera Italiana, 82 (1989) 221.

[167] G. Vaccari., G. Mantovani et G. Sguldino, Sugar J., 53 (1990) 4.

[168] R. Ames, S. Norton et H. Nguyen, Sugar J., 52 (1989) 7.

[169] S. Rearick, Indian Sugar, 40 (1990) 403.

[170] F. Cadet, Analyse du jus de canne à sucre par spectroscopie moyen infrarouge. *In* : La spectroscopie infrarouge et ses applications analytiques, D. Bertrand, E. Dufour, Tec&Doc Lavoisier, Paris, 2000, p 397.

[171] J. Moravcova, Z. Vanclova et M. Novotna, Potrav Vedy, 11 (1993) 493.

[172] S. Brokensha, R. Niemeyer et R. Schaffler, Proc. 52nd Congr. S. African Sugar Tech. Assoc., 1978, p 54.

[173] I. Chakravarti, R. Laha et J. Roy, Handbook of methods of applied statistics, volume1, Wiley&sons, 1967, p 160.

[174] J. Shenk et M. Westerhaus, Crop Sci., 31 (1991) 1548.

[175] G. Sinnaeve, P. Dardenne et R. Agneessens, J. Near Infrared Spectrosc., 2 (1994) 163.

[176] T. Kemper et S. Sommer, Environ. Sci. Technol., 36 (2002) 2742.

[177] Y. Roggo, L. Duponchel, B. Noé et J.-P. Huvenne, J. Near Infrared Spectrosc., 10 (2002) 137.

[178] P. Robert, M. Devaux, A. Qannari et M. Safar, J. Near Infrared Spectrosc., 1 (1993) 99.

[179] K. Miyamoto et Y. Kitano, J. Near Infrared Spectrosc., 3 (1995) 227.

[180] R. Cho, M. Sohn et Y. Kwon, J. Near Infrared Spectrosc., 6 (1998) A75.

[181] F. Rambla, S. Garrigues et M. de la Guardia, Anal. Chim. Acta, 344 (1997) 41.

[182] W. Li, P. Goovaerts et M. Meurens, J. Agric. Food Chem., 44 (1996) 2252.

[183] D. Bertrand, Une démarche générale pour l'établissement d'applications analytiques. *In* : La spectroscopie infrarouge et ses applications analytiques, D. Bertrand, E. Dufour, Tec&Doc Lavoisier, Paris, 2000, p 381.

[184] T. Fearn, Flat or natural ? A note on choice of calibration samples. *In :* Near infrared spectroscopy bridging the gap between data analysis et NIR applications, K. Hildrum, T. Isaksson, T. Naes et A. Tandbeg. Ellis Hordwood, New York, 1992, p 61.

[185] T. Isaksson et T. Naes, Appl. Spectrosc., 44 (1990) 1152.

[186] T. Naes, Chemometr. Intell. Lab. Syst., 5 (1987) 155.

[187] WINISI manual version 1.02A, The complete software solution for routine analysis, robust calibration and networking, FOSS NIRsystems, 1999.

[188] F. Koehler, G. Small, R. Combs, B. Knapp et R. Kroutil, Anal. Chem., 72 (2000) 1690.

[189] J. Lin, S. Lo et C. Brown, Anal. Chim. Acta, 349 (1997) 263.

[190] R. Feudale, N. Woody, H. Tan, A. Myles, S. Brown et J. Ferré, Chemometr. Intell. Lab. Syst., 64 (2002) 181.

[191] P. Dardenne, Transfert d'équation d'étalonnage et mise en réseau d'instruments, D. Bertrand, E. Dufour, Tec&Doc Lavoisier, Paris, 2000, p 371.

[192] G. Martin, J. Shenk et F. Barton, Near infrared reflectance spectroscopy (NIRS): analysis of forage quality. United States Department of Agriculture, Agricultural Research Service. Agricultural Handbook, No. 643, 1989.

[193] Y. Wang et B. Kowalski, Appl. Spectrosc., 46 (1992) 764.

[194] Y. Wang, D. Veltakamp et B. Kowalski, Anal. Chem., 63 (1991) 2750.

[195] E. Bouversse, D. Massart et P. Dardenne, Anal. Chim. Acta, 297 (1994) 405.

[196] J. Shenk et M. Westerhaus, US Pat. N° 4866644, September 1991.

[197] B. Osborn et T. Fearn, J. Food Technol.,18 (1983) 453.

[198] T. Fearn, J. Near Infrared Spectrosc., 9 (2001) 229.

[199] E. Bouveresse et D. Massart, Vibrational Spectrosc., 11 (1996) 3.

[200] P. Williams et D. Sobering. Near Infrared Spectroscopy: The Future Waves. A. Davies et P. Williams. NIR Publications, Chicheter, 1996, p 185.

[201] W. Wu et D. Massart, Anal. Chim. Acta, 349 (1997) 253.

[202] A. Candolfi, R. De Maesschalck, D. Jouan-Rimbaud, P. Hailey, et D. Massart, J. Pharmaceut. Biomed., 21 (1999) 115.

DETERMINATION DE LA QUALITE DE LA BETTERAVE SUCRIERE PAR SPECTROSCOPIE PROCHE INFRAROUGE ET CHIMIOMETRIE

Résumé Actuellement, l'évaluation de la qualité de la betterave sucrière (*Beta vulgaris*) est réalisée par analyse d'un jus limpide obtenu après une défécation à l'acétate de plomb. Cependant les métaux lourds sont polluants et leur utilisation pourrait être interdite. C'est pourquoi la spectroscopie proche infrarouge (SPIR) est envisagée comme méthode alternative.

La première partie de l'étude concerne la faisabilité du dosage du saccharose par SPIR en utilisant la polarimétrie comme méthode de référence. Afin d'obtenir l'erreur standard de prédiction (SEP) la plus faible possible, différents prétraitements spectraux et différentes méthodes de régression sont évalués. Une approche statistique permet de choisir le modèle utilisé. Ainsi un SEP de 0,1 g de saccharose pour 100 g de betteraves est obtenu sur une gamme de concentration allant de 14 à 21 g / 100 g.

La seconde partie développe les problèmes de transfert d'étalonnage et de l'utilisation de la SPIR dans un contexte industriel. Plusieurs approches sont comparées : correction spectrale, correction des valeurs prédites et développement d'un modèle robuste. La dernière solution apparaît être la plus adaptée à notre étude. Il semble donc possible de déterminer la teneur en saccharose de la betterave sur plusieurs instruments en conservant la même précision. Enfin, la faisabilité de l'automatisation de la mesure spectrale est également abordée pour répondre aux cadences industrielles.

La troisième partie concerne la détermination simultanée de plusieurs constituants de la betterave afin d'estimer sa qualité. Ainsi, le brix, la teneur en azote et d'autres paramètres sont évalués en appliquant la même démarche que pour le dosage du saccharose. De plus, des paramètres qualitatifs tels que l'origine géographique, la résistance à une maladie ou la période de récolte sont évalués grâce à des méthodes de classification supervisées.

Mot-clefs : spectroscopie proche infrarouge, betterave sucrière, saccharose, régression multivariée, classification supervisée, transfert d'étalonnage.

SUGAR BEET QUALITY DETERMINATION BY NEAR INFRARED SPECTROSCOPY AND CHEMOMETRIC

Abstract Since 1964, the official method for the determination of beet sucrose content, i.e. polarimetry, uses lead acetate. However, because heavy metals pollute the environment, the use of lead acetate is likely to be banned in the near future. Near infrared spectroscopy (NIRS) is a suitable replacement for this method.

The first part is a feasibility study to determine beet sucrose content by NIRS. Several spectral pre-processing and regression methods were tested in order to produce an accurate prediction of sugar content. Analyse of variance and Fisher's tests were used to compare models (bias and Standard Error of Prediction corrected for bias) in terms of statistical significance. A model, developed with spectra pre-treated by SNV and second derivative, gave the most accurate results on a validation set of 525 samples. The standard error of prediction was 0.10 g of sucrose / 100 g of fresh beet over a large concentration range (14 – 21 g / 100 g).

The transfer of this calibration to an industrial application was studied in a second part. Different alternative solutions were evaluated: among these were bias correction of predicted values, spectral correction and development of robust models. The solution adopted was a robust model developed using a calibration set containing spectra from several instruments. The conclusion reached was that prediction of the polarimetric measurement using several NIRS instruments is feasible in an industrial context. Finally, the automation of NIRS measurements using an automatic filling system coupled to a spectrometer was studied.

In the last part, the determination of several components contents, which determine sugar beet quality, is studied. Brix, Nitrogen content and others parameters are evaluated by NIRS. Supervised pattern recognition methods are used to determine qualitative parameters like disease resistance, geographical origin and crop period.

Keywords : near infrared spectroscopy, sugar beet, sucrose, multivariate regression, supervised pattern recognition method, calibration transfer.

Auteur : Yves ROGGO
Ecole doctorale : Sciences pour l'ingénieur
Discipline : Instrumentation et Analyses Avancées, USTL, 59655 Villeneuve d'Ascq, France.
Laboratoire : Laboratoire de Spectrochimie Infrarouge et Raman, LASIR, CNRS UMR 8516, USTL, Bât C5, 59655 Villeneuve d'Ascq, France.