ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

COMME EXIGENCE PARTIELLE
À L'OBTENTION DE LA
MAÎTRISE EN GÉNIE DE LA PRODUCTION AUTOMATISÉE

M.Ing.

PAR
GIROD, DENIS

DÉTERMINATION DE LA MATURITÉ DES AVOCATS HASS PAR IMAGERIE HYPERSPECTRALE

MONTRÉAL, LE 22 AOÛT 2008

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE :

M. Jacques-André Landry Ph. D., directeur de mémoire
Département de génie de la production automatisée à l'École de technologie supérieure

M. Robert Hausler Ph. D., président du jury
Département de génie de la construction à l'École de technologie supérieure

M. Gilles Doyon, Agr. Ph. D., membre du jury
Agriculture et Agroalimentaire Canada

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 4 AOÛT 2008

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Nous tenons en premier lieu à remercier M. Jacques-André Landry, pour son chaleureux soutient tout au long de ce projet. C'est grâce à M. Landry que nous avons eu la chance de poursuivre nos études au Québec, et ainsi approfondir nos connaissances dans ce domaine qui nous passionne. Que ce travail soit le témoin de notre gratitude et de notre reconnaissance.

Nous remercions notre famille, qui d'une part a partagé son temps et ses connaissances de la langue française pour ce mémoire, mais surtout nous a soutenus moralement et matériellement pendant toutes nos années d'études.

Nous voulons également remercier M. Gilles Doyon, agronome et professeur associé à l'ETS, pour ses recommandations précieuses et avisées concernant les besoins de l'industrie agroalimentaire.

Ce projet n'aurait pas pu être mené à bien sans la disponibilité, et les remarquables conseils techniques de M. Wes Procino de la société Autovision.

Enfin, nous adressons nos remerciements à toutes les personnes qui ont concouru, de prés ou de loin à l'accomplissement de ce travail.

DÉTERMINATION DE LA MATURIITÉ DES AVOCATS HASS PAR IMAGERIE HYPERSPECTRALE

GIROD DENIS

RÉSUMÉ

La maturité de l'avocat est établie habituellement en mesurant son contenu en matière sèche, un processus long et destructif. Le but de cette étude est d'introduire une technique rapide et non destructrice pour estimer le taux de matière sèche de ce fruit tropical.

Des avocats de la variété « Hass » à différents stades de maturité et différentes couleurs de peau ont été analysés par imagerie hyperspectrale en mode réflectance et absorbance. La plage des taux de matière sèche s'étend de 19.8% à 42.5%. Les données hyperspectrales consistent en des spectres moyennés d'une zone du fruit acquis dans le visible et proche infrarouge (de 400nm à 1000nm), pour un total de 163 bandes spectrales distinctes.

La relation entre les spectres et les taux de matière sèche a été établie grâce à des techniques d'analyse chimiométrique, notamment la régression des moindres carrés partiels (PLS). Les statistiques de calibration et validation comme le coefficient de détermination (R^2), et l'erreur quadratique moyenne de prédiction (RMSEP) ont été utilisés afin de comparer les capacités de prédiction des différents modèles. Les résultats des modélisations PLS portant sur plusieurs randomisations de la base de données, en utilisant le spectre dans son ensemble, donnent un R^2 de 0.86 avec une erreur moyenne RMSEP de 2.45 en mode réflectance, ainsi qu'un R^2 moyen de 0.94 avec une erreur RMSEP moyenne de 1.59 pour le mode absorbance. Cela indique que des modèles raisonnablement précis (R^2>0.8) peuvent être obtenus pour l'évaluation du taux de matière sèche avec le spectre dans son entier.

Cette étude montre également que les concepts de réduction de bandes spectrales peuvent être appliqués à ce sujet. Partant de 163 bandes spectrales, le taux de matière sèche a pu être prédit avec les mêmes performances en utilisant 10% des bandes initiales (16 bandes).

Par conséquent cette étude démontre la faisabilité d'utiliser l'imagerie hyperspectrale dans le domaine du visible et proche infrarouge en mode absorbance, dans le but de déterminer des propriétés physicochimiques, le taux de matière sèche dans notre étude, des avocats Hass de manière non destructive. En outre, cette étude donne des indices permettant de déterminer quelles bandes spectrales semblent être pertinentes à cette fin.

PREDICTING FRUIT MATURITY OF HASS AVOCADO USING HYPERSPECTRAL IMAGERY

GIROD Denis

ABSTRACT

The maturity of avocado fruit is usually assessed by measuring its dry matter content (DM), a destructive and time consuming process. The aim of this study is to introduce a quick and non-destructive technique that can estimate the dry matter content of an avocado fruit.

'Hass' avocado fruits at different maturity stages and varying skin fruit color were content analyzed by hyperspectral imaging in reflectance and absorbance modes. The dry matter ranged from 19.8% to 42.5%. The hyperspectral data consist of mean spectra of avocados in the visible and near infrared regions, from 400nm to 1000nm, for a total of 163 different spectral bands.

Relationship between spectral wavelengths and dry matter content were carried out using a chemometric partial least squares (PLS) regression technique. Calibration and validation statistics, such as correlation coefficient (R^2) and prediction error (RMSEP) were used as means of comparing the predictive accuracies of the different models. The results of PLS modeling, over several different randomizations of the database, with full cross validation methods using the entire spectral range, resulted in a mean R^2 of 0.86 with a mean RMSEP of 2.45 in reflectance mode, and a mean R^2 of 0.94 with a mean RMSEP of 1.59 for the absorbance mode. This indicates that reasonably accurate models ($R^2>0.8$) could be obtained for DM content with the entire spectral range.

Also this study shows that wavelengths reduction can be applied to the problem. Starting with 163 spectral bands, the dry matter could be predicted with identical performances using 10% of the initial wavelengths (16 spectral bands).

Thus the study demonstrates the feasibility of using visible, near infrared region hyperspectral imaging in absorbance mode in order to determine a physicochemical property, namely dry matter content, of 'Hass' avocados in a non-destructive way. Furthermore it gives some clues about which spectral bands could be useful to this end.

TABLE DES MATIÈRES

Page

INTRODUCTION 1

CHAPITRE 1 REVUE DE LITTERATURE 3
1.1 Aperçu général 3
1.2 Le spectre lumineux 3
1.2.1 Spectroscopie 5
1.2.2 Imagerie hyperspectrale 7
1.2.2.1 Avantages et limitations 9
1.3 Outils d'analyse et de traitement des données hyperspectrales 10
1.3.1 La chimiométrie 10
1.3.2 La régression 11
1.3.2.1 Régression linéaire multiple 14
1.3.2.2 Régression en composantes principales 16
1.3.2.3 Régression des moindres carrés partiels 16
1.3.3 Analyse statistique de modèles prédictifs 17
1.3.4 Sélection de variables spectrales 19
1.3.4.1 Élagage PLS 21
1.3.4.2 Algorithme de colonie de fourmis 26
1.4 Utilisation de la vision par ordinateur comme méthode non destructrice d'évaluation de paramètres dans l'industrie agro-alimentaire 28
1.4.1 Mise en contexte 28
1.4.2 Spectroscopie 32
1.4.2.1 Évaluation des paramètres 33
1.4.2.2 Plage spectrale 34
1.4.2.3 Modèle statistique 36
1.4.2.4 Prétraitements 36
1.4.2.5 Étalonnage et performances du modèle. 37
1.4.3 Imagerie hyperspectrale 38
1.5 L'avocat 39
1.5.1 Description 39
1.5.2 Qualité et maturité des avocats 41

CHAPITRE 2 PROTOCOLE EXPERIMENTAL 42
2.1 Aperçu général 42
2.2 Analyse physique des avocats 42
2.2.1 Extraction de la matière sèche 42
2.2.2 Préparation des échantillons 45
2.2.3 Couleur et maturité des avocats 46
2.2.4 Base de données de matière sèche 46
2.3 Imagerie hyperspectrale 50
2.3.1 Mise en contexte 50

2.3.2 Caractérisation de la lumière 51
2.3.2.1 Problématique et types d'éclairage 51
2.3.2.2 Réflexion spéculaire et montage final 54
2.3.3 Étalonnage des données spectrales 57
2.3.4 Résolution spectrale 64
2.3.5 Dimension spatiale 68
2.3.6 Logiciel d'acquisition 70
2.3.7 Prétraitements 71
2.3.7.1 Étalonnage du cube 71
2.3.7.2 Segmentation de l'avocat 72
2.3.7.3 Région d'intérêt 75
2.4 La régression PLS 76
2.4.1 Évaluation de la complexité du modèle 79
2.4.1.1 Validation croisée 79
2.4.1.2 Diagrammes de Pareto 83
2.4.2 Mesure de performance 84
2.5 Sélection de variables 86
2.5.1 Élagage PLS 87
2.5.2 Algorithme de colonie de fourmis 92

CHAPITRE 3 ETUDE QUALITATIVE TOUTES BANDES 100
3.1 Aperçu général 100
3.2 Présentation des résultats 100
3.3 Analyse du spectre des avocats 102
3.3.1 Pigments présents dans la peau des fruits. 103
3.3.2 Différences entre les spectres en mode réflectance et absorbance 104
3.4 Choix de la méthode d'évaluation de la complexité des modèles 107
3.5 Configuration retenue 110

CHAPITRE 4 SÉLECTION DE BANDES SPECTRALES 112
4.1 Aperçu général 112
4.2 Présentation des résultats 112
4.2.1 Élagage PLS 113
4.2.2 Algorithmes de colonie de fourmis 120
4.3 Comparaison des deux approches 122
4.4 Bénéfices de la sélection de bandes spectrales 122

CONCLUSION 124

RECOMMANDATIONS 126

ANNEXE I EXTRACTION DE LA MATIÈRE SÈCHE DES AVOCATS 127

ANNEXE II BASE DE DONNEES D'AVOCATS 132

ANNEXE III CALCUL DES COMPOSANTES RGB A PARTIR DU SPECTRE 136

ANNEXE IV PERFORMANCES OBTENUES POUR DES MODELES TOUTES BANDES.....140

BIBLIOGRAPHIE.....152

LISTE DES TABLEAUX

Page

Tableau 1.1 Comparaison des résultats prédictifs entre 3 méthodes de sélection de variables 28

Tableau 1.2 Avantages et inconvénients d'un système de vision 31

Tableau 1.3 Performance sur l'ensemble de validation des modèles MLR 39

Tableau 1.4 Principaux pays producteurs d'avocats 40

Tableau 2.1 Avantage/désavantages des méthodes de chauffage de l'avocat 43

Tableau 3.1 Moyennes des erreurs d'étalonnage et de prédiction des modèles 101

Tableau 3.2 Moyennes des coefficients de détermination et des écarts type des résidus des modèles 101

Tableau 3.3 Nombres moyens de variables latentes retenues pour les modèles 102

Tableau 3.4 Moyennes des performances en mode réflectance sur les 46 modèles non aberrants 107

Tableau 3.5 Moyennes des performances en mode réflectance sur les 46 modèles non aberrants 109

Tableau 3.6 Moyennes des performances en mode absorbance sur les 49 modèles non aberrants 109

Tableau 3.7 Moyennes des performances en mode absorbance sur les 49 modèles non aberrants 110

Tableau 4.1 Moyennes des performances des deux fonctions de sélection de variables 113

Tableau 4.2 Bandes sélectionnées et leurs longueurs d'ondes correspondantes 114

Tableau 4.3 Moyennes des performances de deux configurations de sélection de variables 115

Tableau 4.4 Moyennes des performances données par la sélection des bandes 116

Tableau 4.5 Moyennes des performances des modèles issus de la sélection de certaines bandes 118

Tableau 4.6 Moyennes des meilleures performances obtenues pour chaque combinaison de bandes....119

Tableau 4.7 Bandes sélectionnées par l'ACF et leurs longueurs d'ondes correspondantes...120

Tableau 4.8 Moyennes des performances de deux configurations de sélection de variables....121

LISTE DES FIGURES

Page

Figure 1.1 *Rayon lumineux décomposé par un prisme.* 4

Figure 1.2 *Le spectre électromagnétique.* 4

Figure 1.3 *Le spectromètre.* 5

Figure 1.4 *Comparaison des réponses de différents capteurs.* 6

Figure 1.5 *Schémas d'un imageur hyperspectral.* 7

Figure 1.6 *Représentation d'un cube hyperspectral, et relation entre les dimensions spectrales et spatiales.* 8

Figure 1.7 *Composantes d'un système d'imagerie hyperspectrale.* 9

Figure 1.8 *Graphe des salaires théoriques versus salaires prédits avec la régression.* 12

Figure 1.9 *Spectres de sirop de sucre dans la région proche infrarouge.* 23

Figure 1.10 *Erreur de validation en fonction la réduction de variable.* 24

Figure 1.11 *Sélection de bandes spectrales par l'approche PLS pruning.* 25

Figure 1.12 *Diagramme de chromaticité CIE 1931 incluant l'espace de couleur RGB.* 32

Figure 1.13 *Mesure spectrale sur le concombre en mode interactance.* 35

Figure 2.1 *Étalonnage du temps de chauffage de différents avocats.* 44

Figure 2.2 *Intervalles de confiance des échantillons à 95%.* 48

Figure 2.3 *Énergie relative de 4 illuminants en fonction de leur longueur d'onde.* 52

Figure 2.4 *Réponse spectrale de la caméra.* 53

Figure 2.5 *Aspect spéculaire de la peau des avocats.* 54

Figure 2.6 *Boite de diffusion.* 55

Figure 2.7 *Intérieur de la boite de diffusion.* 56

Figure 2.8 *Spectre de la Référence blanche ainsi que des courants sombres.* 58

Figure 2.9 *Réflectance de la référence blanche.* 60

Figure 2.10 *Tuile de calibration Spectralon.* 62

Figure 2.11 *Données brutes provenant de l'acquisition des spectres pour l'étalonnage.*....63

Figure 2.12 *Valeurs de réflectance obtenues après étalonnage.*....63

Figure 2.13 *Spectres bruts obtenus après la combinaison de pixels aux extrémités.*....66

Figure 2.14 *Comparaison de valeurs de réflectance pour deux combinaisons de pixels.* 67

Figure 2.15 *Bruit présent dans les spectres étalonnés.* 68

Figure 2.16 *Répartition de l'intensité lumineuse d'une ligne.*....69

Figure 2.17 *Intensité lumineuse de la bande 684nm.* 69

Figure 2.18 *Interface du logiciel SpectralCube.* 70

Figure 2.19 *Image RGB recomposée à partir du spectre d'un avocat.*....72

Figure 2.20 *Image d'un avocat isolé du fond.* 73

Figure 2.21 *Cicatrices présentes dans 2 bandes spectrales.* 74

Figure 2.22 *Avocat segmenté.*....74

Figure 2.23 *Région d'extraction du spectre sur l'avocat.* 75

Figure 2.24 *Spectre extrait d'un avocat en mode réflectance et absorbance.* 76

Figure 2.25 *Étapes du processus de la régression.* 77

Figure 2.26 *Erreur cumulative PRESS en fonction du nombre de variables latentes.*....81

Figure 2.27 *Diagramme de Pareto.* 84

Figure 2.28 *Erreur RMSECV en fonction du nombre de variables.*....91

Figure 2.29 *Évolution de la moyenne et écart type des solutions pour* α *=0.8,* ρ *=1.* 98

Figure 2.30 *Évolution de la moyenne et écart type des solutions pour* α *=0.3,* ρ *=0.3.* 98

Figure 2.31 *Évolution de la moyenne et écart type des solutions pour* α *=0.55,* ρ *=0.8.* 99

Figure 3.1 *Spectre typique d'un avocat en mode réflectance et absorbance.* 103

Figure 3.2 *Spectres d'absorption de trois principaux pigments.* 104

Figure 3.3 *Groupage des échantillons en fonction des taux de matière sèche.* 105

Figure 3.4 *Spectres moyens des échantillons.* 106

Figure 4.1 *Résultats des votes pour l'élagage PLS.* 114

Figure 4.2 *Décompte des bandes sélectionnées dans les modèles performants.* 117

Figure 4.3 *Résultats des votes pour l'ACF.* 120

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

ACF	Algorithme de Colonie de Fourmis
ACP	Analyse en Composantes Principales
CCD	Charged Coupled Device
CMOS	Complementary Metal Semi-Conductor
CV %	Coefficient de variation
MLR	Régression linéaire multiple (Multiple Linear Regression)
PCR	Régression en Composantes Principales (Principal Component Régression)
PLS	Régression aux moindres carrés partiels (Partial Least Square)
PRESS	Predicted Residual Error Sum of Square
RMSE	Erreur quadratique moyenne, (Root Mean Squared Error)
RMSEC	Erreur quadratique moyenne d'étalonnage, (Root Mean Squared Error of Calibration)
RMSECV	Erreur quadratique moyenne de validation croisée, (Root Mean Squared Error of Cross Validation)
RMSEP	Erreur quadratique moyenne de prédiction, (Root Mean Squared Error of Prediction)
SDR	Écart-type des résidus, (Standard Deviation Ratio)
R^2	Coefficient de détermination
RVB	Rouge Vert Bleu (Red Green Blue)
TSS	Solides solubles totaux (Total Soluble Solids)

INTRODUCTION

La vision par ordinateur est un domaine en pleine expansion. Les développements continus en matière de capteurs ainsi que l'augmentation exponentielle des capacités de traitement de l'information élargissent constamment ses horizons.

L'agroalimentaire est un champ d'applications où la vision par ordinateur est en plein essor. En effet, être en mesure d'analyser précisément, de manière presque instantanée, et surtout de façon non destructrice la qualité d'un aliment transformé ou non, est un atout considérable. Non seulement les coûts sont réduits (il n'y a aucune perte ni transformation du produit), mais cela permet de faire une analyse en continu, sur tous les produits, et non une analyse sur seulement une partie « représentative » du lot. Cette précision combinée à la rapidité de traitement permet d'atteindre des standards de production toujours plus hauts et ainsi une qualité toujours meilleure des aliments.

L'évolution des technologies permet actuellement de décomposer très précisément la lumière dans le domaine du visible, au-delà et même en-deçà (proche infrarouge et ultra-violet). Alors qu'il n'y a que trois canaux principaux (rouge vert et bleu avec les capteurs de caméras conventionnelles), des capteurs toujours plus performants sont en mesure de décomposer la lumière en plusieurs centaines de bandes spectrales, correspondant à des partitions toujours plus petites du spectre lumineux. Cette décomposition apporte une information considérable et très précise de la lumière captée. Les récentes percées technologiques en vision par ordinateur ont permis de concevoir des outils capables de faire l'acquisition d'une scène en deux dimensions un peu à l'instar d'une caméra numérique. Mais au lieu des 3 canaux (rouge, vert et bleu), il y en a plusieurs centaines : ce sont les imageurs hyperspectraux.

Notre travail porte sur l'étude de la maturité des avocats Hass par imagerie hyperspectrale. Ce mémoire est divisé en quatre chapitres, puis nous présentons notre conclusion, ainsi que nos recommandations et enfin les annexes.

Le premier chapitre propose une revue d'ensemble des connaissances actuelles en matière de vision par ordinateur, et plus particulièrement de l'imagerie hyperspectrale utilisée dans le cadre de l'analyse de la qualité de produits agroalimentaires. Nous rapporterons les principaux sujets d'études de ces dernières années, et les outils utilisés dans le cadre de ces études. Nous décrivons enfin le sujet de notre recherche qui est l'avocat à l'état frais.

Nous présentons notre protocole expérimental au chapitre 2. Nous détaillons le déroulement des deux types d'analyses sur le fruit faites au cours de notre étude : Les analyses physiques et destructrices de l'évaluation de la maturité du fruit, ainsi que les acquisitions des données faites de manière non destructrice au moyen de notre système d'imagerie. Au delà, nous allons présenter les outils que nous avons choisis pour traiter nos données, et comment nous les avons paramétrés pour mener rigoureusement notre analyse.

Les chapitres 3 et 4 présentent les résultats de nos expérimentations avec leur analyse. Le chapitre 3 présente une analyse des résultats obtenus lors de l'évaluation de la maturité du fruit en utilisant toute l'information spectrale disponible. Cependant, dans le domaine de la l'imagerie hyperspectrale, le nombre de bandes spectrales peut atteindre plusieurs centaines. Nous allons voir au chapitre 4 que toutes ces bandes spectrales ne sont pas, à priori, nécessaires dans leur ensemble pour bien caractériser notre sujet d'étude. Nous allons donc procéder à la sélection des quelques bandes spectrales qui sont les plus pertinentes, et analyser les résultats ainsi obtenus.

Nous présentons la synthèse de notre étude dans la conclusion, et proposons quelques voies pour la poursuite éventuelle de la recherche de cette étude.

CHAPITRE 1

REVUE DE LITTERATURE

1.1 Aperçu général

Dans ce chapitre, introduirons les concepts fondamentaux liés à notre problématique. Tout d'abord, nous étudierons le spectre lumineux : les moyens de le décomposer, et la description des principaux outils de traitement des données spectrales. Ensuite, Nous passerons en revue différentes applications de vision par ordinateur et plus précisément d'analyse spectrale au domaine de l'industrie agroalimentaire. Quelles sont ces applications et quels sont les bénéfices d'une telle démarche. Enfin, nous décrirons notre sujet d'étude qu'est l'avocat.

1.2 Le spectre lumineux

La lumière est composée de la superposition d'ondes électromagnétiques. Newton, au XVII^{ème} siècle démontra que la lumière du soleil pouvait être décomposée en une multitude de composantes monochromatiques. Il en vint à cette conclusion en faisant passer des rayons lumineux au travers de 2 prismes consécutifs. Si la lumière blanche est décomposée au passage du premier prisme, les rayons, au passage du deuxième prisme étaient diffractés, mais non décomposés.

Chaque « couleur » correspond en fait à une longueur d'onde bien précise du spectre du visible. C'est cette décomposition que l'on appelle spectre lumineux. Le prisme n'est pas le seul moyen de décomposer la lumière. Une autre technique pour la décomposition de la lumière est l'emploi d'un réseau de diffraction. Un réseau est un dispositif optique composé de fentes (réseau en transmission), ou de rayures réfléchissantes (réseau en réflexion), espacées de manière régulière. Si l'espacement entre les fentes ou les rayures est de l'ordre de grandeur de la longueur d'onde, la lumière transmise (ou réfléchie), va être décomposée à la manière du prisme.

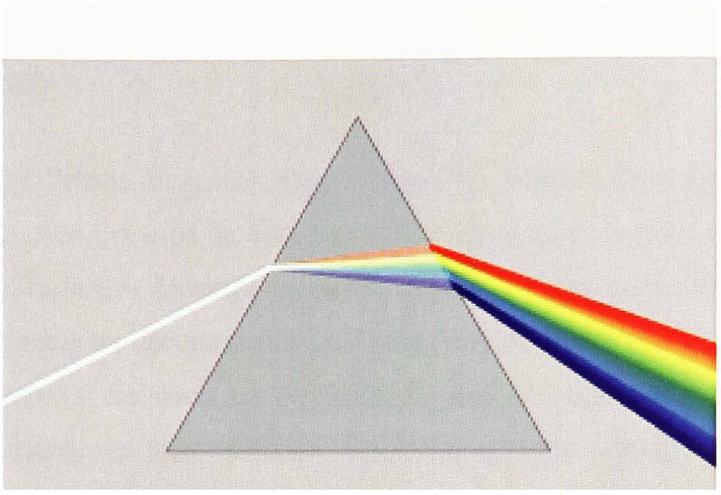

Figure 1.1 ***Rayon lumineux décomposé par un prisme.***

(Tiré de Wikipédia, 2008b)

Nous pouvons observer un exemple de décomposition, en arc-en-ciel, de la lumière par un prisme à la Figure 1.1. Le spectre lumineux, par extension, contient les ultraviolets, les longueurs d'ondes visibles, ainsi que les infrarouges. Concernant le domaine du visible, le spectre s'étend de 380nm à 780nm, au delà, ce sont les infrarouges, et en deçà, les ultra-violets (*Voir* Figure 1.2).

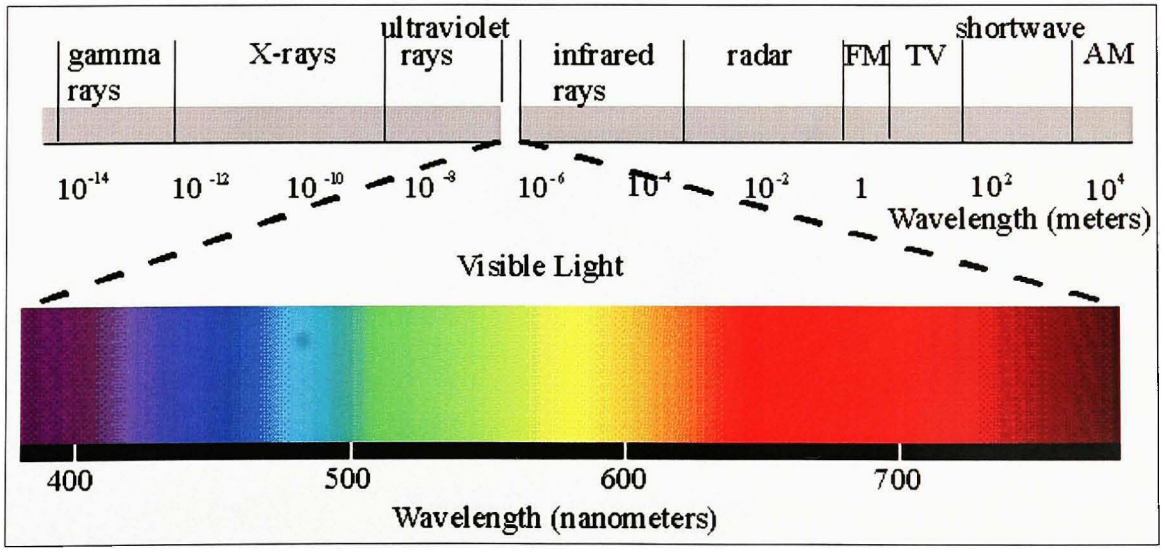

Figure 1.2 ***Le spectre électromagnétique.***

(Tiré de Kaiser, 2005)

1.2.1 Spectroscopie

La spectroscopie est l'étude du spectre émis ou absorbé. L'instrument de mesure du spectre est le spectroscope. Son principe de fonctionnement est simple, la lumière passe au travers d'une fente pour ensuite être décomposée par un dispositif prévu à cet effet (par exemple un prisme). En observant cette décomposition, on peut donc étudier le spectre, en examinant les bandes d'absorption ou d'émission. La spectroscopie dans le visible ou le proche infrarouge a rapidement été adoptée comme méthode d'acquisition sans contact, pour l'analyse de différents paramètres dans des domaines très variés. La spectroscopie proche infrarouge mesure la lumière réfléchie ou transmise selon une plage de longueur d'onde allant de l'UV à l'IR. Cette mesure est acquise pour une petite partie de l'objet étudié. Le matériel type nécessaire est composé d'un spectromètre doté d'un dispositif d'éclairage contrôlé (le plus souvent un dispositif à fibre optique intégré à l'appareil), et d'un ordinateur muni de logiciels adéquats pour l'acquisition et le traitement des données. Le spectromètre est généralement composé d'un prisme et d'une cellule photoélectrique linéaire (*Voir* Figure 1.3) dont la sensibilité et la résolution sont adaptées en fonction des spécifications voulues de l'appareil.

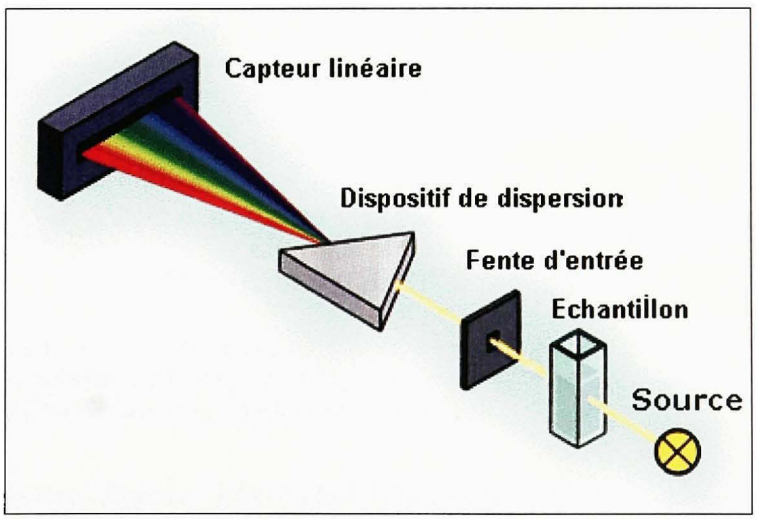

Figure 1.3 ***Le spectromètre.***
(Tiré de GMI, 2006 [Notre traduction])

Suivant la plage spectrale désirée, le type de cellule photosensible utilisé ne sera pas le même pour observer la décomposition de la lumière visible et proche infrarouge. Si les cellules CCD (Charged Coupled Device) ou CMOS (Complementary Metal Semi-Conductor) sont efficaces dans le domaine visible, elles sont peu sensibles dans le proche infrarouge. On leur préfère alors des cellules dites InGaAs (Indium Gallium Arsenide) beaucoup plus adaptées, mais aussi plus onéreuses.

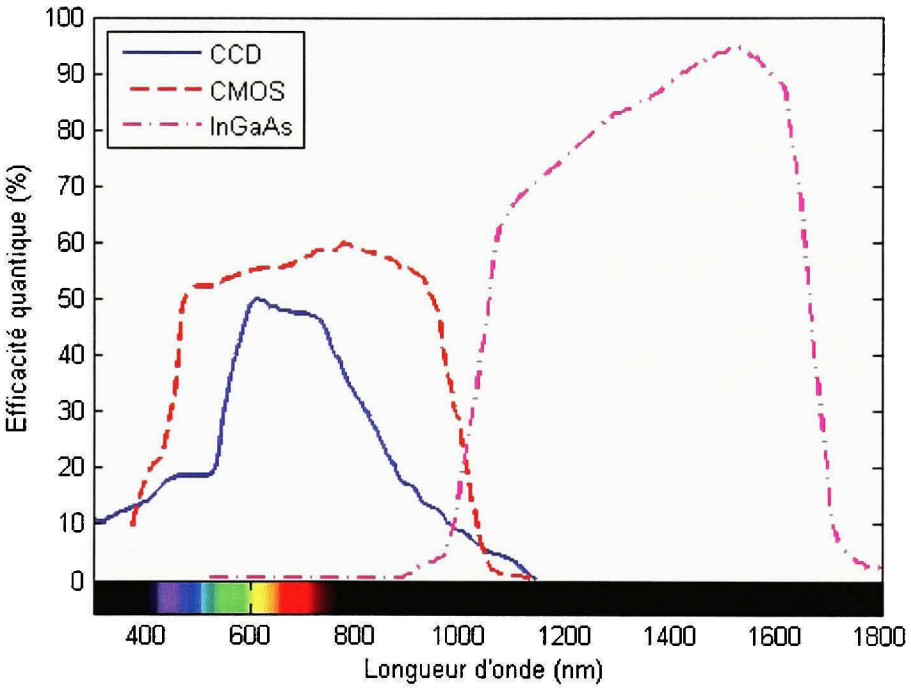

Figure 1.4 ***Comparaison des réponses de différents capteurs.***

(Tiré de Colvin, 2005; Newport, 2008)

Source : Cette figure sous forme de graphique est tirée de la réunion de deux sources. La première est la fiche technique de capteurs CCD/CMOS de marque Newport, et la deuxième d'un document en ligne de M. James Barry Colvin, *Microscopy of electonic devices*.

Comme nous pouvons observer à la Figure 1.4, les réponses d'un capteur CCD/CMOS et InGaAs sont complémentaires. L'appareil mesure la réflexion ou transmission par l'objet, du spectre lumineux incident. Le spectre est discret et l'acquisition donne une mesure unique pour la zone concernée. Généralement un spectromètre se caractérise par deux facteurs qui sont : la plage de sensibilité (domaine dans lequel la réponse est optimale : visible, proche

infrarouge, ultra-violet...), et la résolution (caractérise la précision de l'appareil, c'est la partition minimale du spectre que l'appareil peut mesurer).

1.2.2 Imagerie hyperspectrale

L'imagerie hyperspectrale, est une nouvelle technique d'imagerie, qui mesure le spectre transmis ou plus généralement réfléchi sur une plage de longueur d'ondes allant de l'UV à l'IR. À la différence du spectromètre, qui fournit une mesure de spectre par acquisition, l'imagerie hyperspectrale intègre une dimension spatiale à la mesure. C'est en quelque sorte l'union de techniques de spectroscopie et de techniques avancées d'acquisition d'images. De cette manière on réalise l'acquisition des données à la fois en termes de spectre, mais aussi en termes de surface. Nous pouvons observer une vue détaillée d'un tel système à la Figure 1.5.

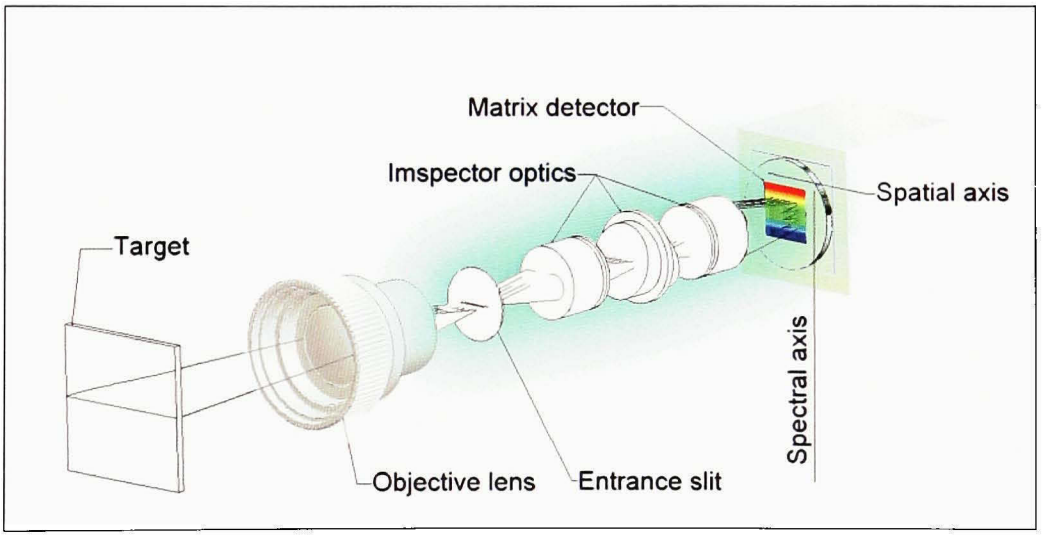

Figure 1.5 ***Schémas d'un imageur hyperspectral.***

(Tiré de Specim, 2003)

Source : Cette figure est issue de la documentation technique de l'imageur hyperspectral « Imspectror V10E » fabriqué par la société Specim.

Le principe de base est le même que pour le spectroscope, cependant le capteur n'est plus linéaire mais est un capteur deux dimensions. Généralement c'est une caméra monochrome dont la cellule, à la manière du spectromètre, est adaptée à la plage spectrale et à la résolution

de l'appareil. Dans ce cas, les pixels du capteur parallèles à la fente détermineront la dimension spatiale, et les pixels perpendiculaires à la fente, la dimension spectrale.

A chaque acquisition, une ligne de la scène est numérisée, et chaque pixel de cette ligne est décomposé selon son spectre. On effectue l'acquisition séquentielle de lignes successives pour former une image complète. Ces lignes mises côte-à-côte forment un cube hyperspectral, et donne des informations sur le spectre de toute la scène acquise. Nous pouvons observer à la Figure 1.6 les concepts de dimension spatiale et spectrale d'une scène acquise présentes dans le cube hyperspectral.

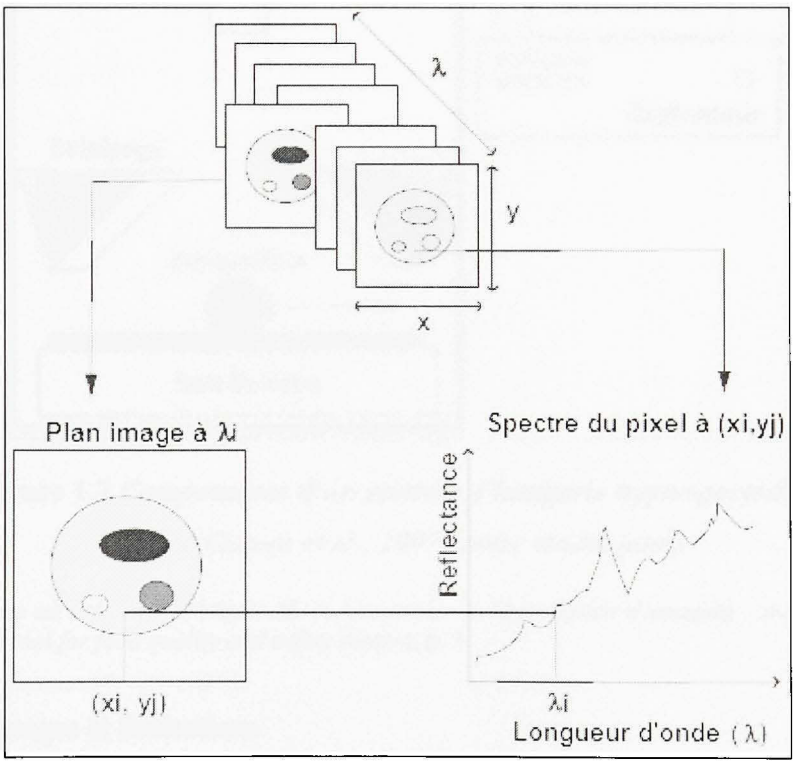

Figure 1.6 ***Représentation d'un cube hyperspectral, et relation entre les dimensions spectrales et spatiales.***

(Tiré de Gowen et al., 2007 [notre traduction])

Source : Cette figure est tirée de l'article de M. A. Gowen et al., *Hyperspectral imaging - an emerging process analytical tool for food quality and safety control*, p. 2.

Un système d'imagerie hyperspectrale se compose d'un imageur, d'un système d'éclairage, et d'un ordinateur doté des logiciels adéquats pour l'acquisition et le traitement des données, ainsi que d'un système de déplacement de l'imageur ou de la scène à acquérir, comme la montre la Figure 1.7.

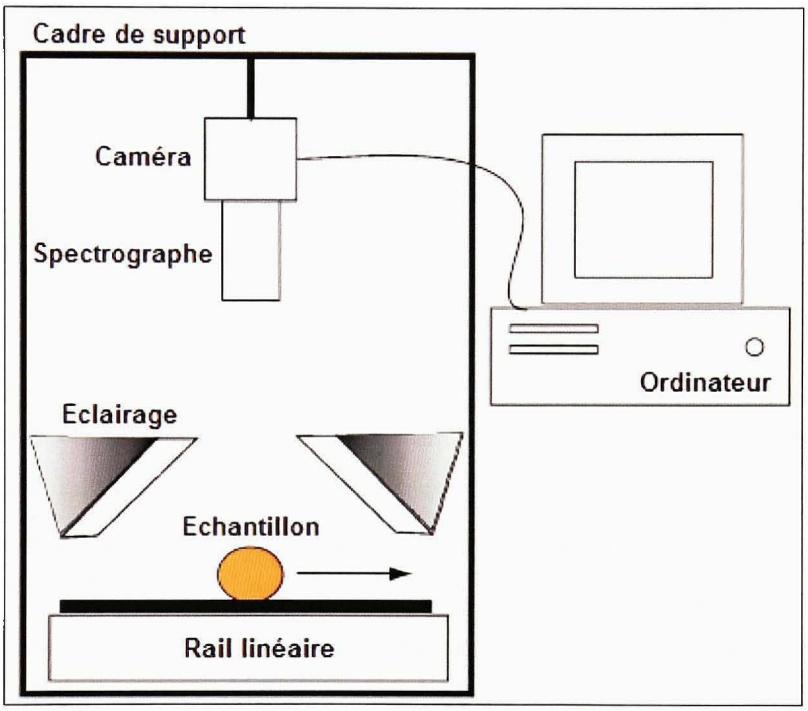

Figure 1.7 ***Composantes d'un système d'imagerie hyperspectrale.***

(Tiré de Gowen et al., 2007 [notre traduction])

Source : Cette figure est tirée de l'article de M. A. Gowen et al., *Hyperspectral imaging - an emerging process analytical tool for food quality and safety control*, p. 3.

1.2.2.1 Avantages et limitations

L'imagerie hyperspectrale est très avantageuse car le nombre de données acquises est important. La perte d'information est minime, ce qui est primordial lorsque l'on fait de l'extraction d'informations. L'aspect spatial ajouté donne beaucoup de renseignements sur la répartition de tel ou tel phénomène (par exemple un composé chimique spécifique). Cela peut

aussi être utile pour faire de l'analyse de texture, ou pour trouver les régions d'intérêt de la scène observée.

Cependant, un tel système d'acquisition est plutôt onéreux et peu disponible du fait de sa nouveauté. De plus, le temps de capture est plus long, puisqu'il faut déplacer la scène (ou la caméra) pour effectuer l'acquisition du cube. Enfin, les données générées sont colossales, il faut une forte puissance de calcul et beaucoup de mémoire pour effectuer l'acquisition et surtout le traitement des données issues des cubes.

1.3 Outils d'analyse et de traitement des données hyperspectrales

1.3.1 La chimiométrie

La chimiométrie est un outil utilisé pour extraire des informations pertinentes à partir de données physicochimiques mesurées, (Lantéri et R., 1998). Pour ce faire, il faut tenter de construire un modèle reliant nos données mesurées aux données observées. Le terme chimiométrie vient de l'anglais *chemometrics*, discipline associant initialement l'analyse de données et la chimie analytique. De manière moins restrictive on appelle aussi cela analyse multivariée (*multivariate analysis*). On peut, grâce à ces méthodes, traiter des systèmes complexes mettant en jeu des centaines de variables à corréler entre elles.

Aujourd'hui, ce domaine inclus maintes disciplines. En effet, de nombreux problèmes peuvent être décrits sous la forme d'un modèle d'entrée-sortie (Tenenhaus, 1998). Dans l'industrie, ce type d'analyse se résume à essayer de décrire des variables de sortie Y, issues d'analyses physico-chimiques (analyses longues et potentiellement destructrices), à partir de variables d'entrée X mesurables plus aisément (à l'aide de capteurs par exemple). Le concept est de construire un modèle décrivant les relations complexes entre les variables d'entrée et de sortie sans avoir à notre disposition de modèles théoriques à appliquer. Ainsi à l'aide d'un modèle robuste, on peut déduire les variables de sortie de notre système à partir des variables d'entrée sans avoir recours aux analyses physico-chimiques. Dans notre cas, nous désirons prédire la maturité des avocats de manière non destructrice à l'aide de données

hyperspectrales. Tel que nous le verrons plus loin, l'analyse de la maturité passe par le calcul du taux de matière sèche de l'avocat, étape longue (environ 5h par échantillon), et surtout destructrice. Grace à l'analyse hyperspectrale, le processus d'acquisition est court (environ 30 secondes par échantillon), et principalement non destructeur. Ainsi, en déterminant le lien entre la maturité des fruits et leurs spectres, il sera possible de déduire la maturité des avocats de manière presque immédiate sans altérer le fruit.

Les outils utilisés sont très nombreux et assez différents allant des réseaux de neurones à la régression, en passant pas l'analyse discriminante (Lantéri et R., 1998). De la littérature, nous le verrons plus tard, il ressort que l'approche utilisant la régression est largement utilisée pour le type de problème que l'on cherche à résoudre. En effet, on cherche à lier un taux de maturité au spectre, deux variables numériques. La régression est une méthode d'analyse des données spécifiquement construite pour l'étude de ce type de problème.

1.3.2 La régression

La régression est une approche aidant à établir une relation entre un certain nombre de variables indépendantes (ou prédicteurs, en fait les variables d'entrée), et des variables dépendantes (les variables de sortie).

Prenons un exemple. Dans une entreprise, le département des ressources humaines désire savoir si le personnel est payé en concordance avec les autres entreprises du marché. Pour ce faire, une enquête est réalisée. Il est établi que le salaire des employés dépend essentiellement de deux facteurs qui sont la quantité de responsabilité, soit un entier variant de 0 (pas de responsabilité), à 5 (beaucoup de responsabilités), que l'on nommera variable *resp*, et le nombre de personnes sous sa supervision, variable *nb_super*. L'enquête est menée au sein d'autres entreprises de taille et d'activité comparable.

L'information obtenue dans l'enquête peut être utilisée avec une analyse de régression multiple (multiple car on a plusieurs variables descriptives). L'hypothèse est que les

variables de sortie (ici les salaires) sont la combinaison linéaire des variables d'entrée (*resp* et *nb_super*). A l'aide de cette hypothèse et des résultats de l'enquête, nous déterminons une équation de régression de la forme :

$$salaires = 0.5 \times resp + 0.8 \times nb\_super \tag{1.1}$$

Une fois que cette équation a été déterminée, on peut alors construire un graphe des salaires théoriques (salaires prédits avec l'équation) en fonction des salaires réels des employés de l'entreprise.

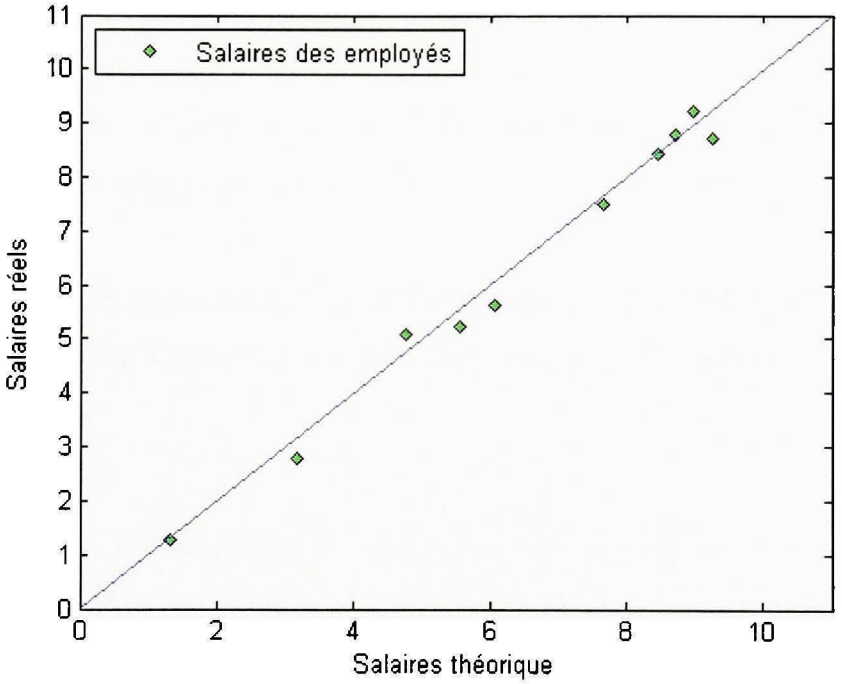

Figure 1.8 ***Graphe des salaires théoriques versus salaires prédits avec la régression.***

A partir de ce graphe, nous pouvons alors analyser les salaires donnés aux employés. Un point en dessous de la ligne indique que l'on sous-paye l'employé, à l'inverse, un point au dessus de la ligne indique le sur-paiement de celui-ci. Un point proche de la ligne ou sur celle-ci, implique le paiement équitable de l'employé.

La régression peut se mettre mathématiquement sous la forme :

Soit $\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}$ le vecteur colonne contenant les variables de sortie de notre système,

$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ \vdots & \ddots & & \vdots \\ x_{i,1} & \cdots & \cdots & x_{i,m} \\ \vdots & & \ddots & \vdots \\ x_{n,m} & \cdots & \cdots & x_{n,m} \end{bmatrix}$, la matrice contenant les variables d'entrée.

Pour l'observation i, $i \in [\![1,n]\!]$, y_i est la variable dépendante, et $x_{i,j}, j \in [\![1,m]\!]$ les variables indépendantes ou explicatives liées à y_i.

L'objectif de la régression est de trouver la relation entre **X** et **Y**. Dans le cas linéaire, on cherche à relier **X** et **Y** par une relation linéaire du type (pour le couple de variable entrée/ sortie $\mathbf{x}_i / y_i$):

$$y_i = b_1 x_{i,1} + b_2 x_{i,2} + \cdots + b_m x_{i,m} + e_i \Leftrightarrow y_i = e_i + \sum_{j=1}^{m} b_j x_{i,j} \tag{1.2}$$

e_i est l'erreur commise par le modèle régressif (elle explique ou résume l'information manquante lors de l'évaluation de y_i à partir de $\mathbf{x}_i$, et $b_i, i \in [\![1,n]\!]$ sont les coefficients de régression.

L'hypothèse de linéarité ne peut évidemment pas être confirmée de manière théorique, mais heureusement, de petites déviations de cette hypothèse sont généralement bien tolérées par

les modèles. Si effectivement les relations liant les variables d'entrées/sorties ne sont pas tout à fait linéaires, le modèle calculé avec l'hypothèse de linéarité est quand même adéquat.

De manière plus générale, en utilisant le produit matriciel, on a :

$$\mathbf{Y} = \mathbf{X} \times \hat{\mathbf{b}} + \mathbf{e} \tag{1.3}$$

Cela équivaut à :

$$\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ \vdots & \ddots & & \vdots \\ x_{i,1} & \cdots & \cdots & x_{i,m} \\ \vdots & & \ddots & \vdots \\ x_{n,m} & \cdots & \cdots & x_{n,m} \end{bmatrix} \times \begin{bmatrix} b_1 \\ \vdots \\ b_i \\ \vdots \\ b_m \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{bmatrix} \tag{1.4}$$

Le processus de régression consiste à trouver les coefficients du vecteur de régression $\hat{\mathbf{b}}$, en réalisant l'approximation

$$\hat{\mathbf{b}} = \mathbf{X}^{-1}\mathbf{Y} \tag{1.5}$$

Différentes méthode de modélisation peuvent être utilisées pour calculer $\hat{\mathbf{b}}$. C'est en fait l'inversion de $\mathbf{X}$ qui pose généralement problème. Les méthodes les plus utilisées sont : la régression linéaire multiple (*Multiple Linear Régression*, MLR), la régression en composantes principales (*Principal Component Regression*, PCR), et la régression des moindres carrés partiels (*Partial Least Square*, PLS).

1.3.2.1 Régression linéaire multiple

La régression linéaire multiple est une généralisation à n variables de la régression linéaire simple. En rappel, la régression linéaire simple ne prend en compte qu'une seule variable

explicative pour une seule variable expliquée (Geladi, 1986). Dans le cas de n échantillons pour m variables indépendantes, nous pouvons distinguer 3 cas :

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

$$\hat{\mathbf{b}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \tag{1.6}$$

Cependant, l'inverse de $\mathbf{X}'\mathbf{X}$ n'existe pas forcement : colinéarité des éléments, singularité de la matrice, déterminant nul, sont des termes synonymes décrivant ce même problème (Helge, 2006).

La MLR est la plus simple des régressions pour relier un ensemble de prédicteurs à une variable dépendante. Malgré sa simplicité, la MLR a quelques défauts. Tout d'abord, la détermination du modèle est très instable lorsque le nombre d'observations est inférieur au nombre de variables : c'est une situation pourtant rencontrée fréquemment. Ensuite, on se heurte à l'impossibilité de prendre en compte les données manquantes, et cela oblige souvent le praticien à rejeter certaines observations qui peuvent néanmoins se révéler utiles pour le modèle, des informations importantes peuvent être contenues dans les variables disponibles. Enfin, la MLR présente une grande sensibilité aux données colinéaires au sein des variables explicatives.

1.3.2.2 Régression en composantes principales

Pour palier au problème de la colinéarité des variables prédictives, on a jumelé l'analyse en composantes principales (ACP) avec la régression. En effet, en effectuant une ACP avant la régression, on projette les variables initiales dans un nouvel espace où elles sont deux à deux orthogonales. Ne sont retenues que les composantes apportant le plus de variance. Ce nombre de composantes, encore appelé nombre de facteur, détermine la complexité du modèle. C'est une méthode très efficace, mais elle ne tient pas compte, dans la projection, des variables dépendantes. Il peut aussi arriver que certaines variations non corrélées avec les variables de sortie, soient retenues dans les composantes principales, et inversement, que de l'information pertinente soit rejetée. Les bases de projection ne seront alors pas optimales. On lui préfère donc une méthode un peu plus avancée qui est la régression des moindres carrés partiels (*Partial Least Square*, PLS).

1.3.2.3 Régression des moindres carrés partiels

La régression PLS est une méthode de détermination d'un modèle prédictif permettant de passer outre les limitations rencontrées précédemment. En fait on peut voir la régression PLS comme une généralisation de la régression linéaire multiple et de la régression en composantes principales. On pallie au problème des données manquantes, ainsi qu'à la colinéarité éventuelle des prédicteurs, et le fait d'avoir beaucoup plus de prédicteurs que d'observations n'est plus un problème.

La régression PLS a été proposée par Herman Wold dans les années 1960 pour résoudre des problèmes issus des sciences économiques. Il développe l'algorithme NIPALS (Wold, 1966) (*Non Linear Interactive Partial Least Squares*). Svante Wold, son fils, généralise cette méthode qui devient très populaire dans le milieu de la chimie. Cette méthode est adaptée pour traiter les problèmes de régression avec beaucoup de prédicteurs, cette approche est alors appelée *Partial Least Squares*, PLS (Wold, Martens et Wold, 1983).

La technique PLS est basée sur la transformation linéaire d'un grand nombre de descripteurs vers un nouvel espace de variables orthogonales le plus petit possible (ce sont les variables latentes). Ces facteurs sont mutuellement indépendants (orthogonaux) et aussi combinaison linéaire des descripteurs initiaux. À la différence de la régression en composantes principales, les variables latentes sont déterminées de telle manière que la corrélation entre elles et les variables dépendantes soit maximum.

Mathématiquement, pour une matrice de variables dépendantes **Y**, et une matrice de prédicteurs **X**, on a :

- Les composantes principales de la régression PCR sont les vecteurs propres de la matrice $\mathbf{X'X}$

- Les facteurs de la régression PLS sont les vecteurs propres du produit matriciel $\mathbf{Y'XX'Y}$. Les variables d'entrées et de sorties sont prises en compte simultanément, ce qui généralement induit un nombre de facteurs nécessaires inférieurs à l'approche PCR. Cela signigie une complexité moindre du modèle. Notons que le nombre de facteurs maximum du modèle est inférieur ou égal au rang de la matrice **X**, (Vancolen, 2004).

Étant donné la souplesse d'utilisation de la méthode PLS, et sa facilité d'implémentation, celle-ci est devenue très populaire en chimie, particulièrement en spectroscopie, mais aussi dans nombre d'autres domaines.

1.3.3 Analyse statistique de modèles prédictifs

L'objectif de la régression est d'estimer ou d'exprimer des variables de réponses avec un ensemble de variables indépendantes. En d'autres termes, c'est une méthode avec laquelle on fait de la prédiction. Quelle que soit le type d'approche choisie pour effectuer de la prédiction, il faut évaluer les performances et la viabilité de notre modèle. De cette manière,

nous pouvons l'apprécier à sa juste valeur, et le comparer à d'autres modèles, ou d'autres approches. Nous allons présenter succinctement les indices généralement utilisés pour mesurer la performance de modèles. Dans le protocole expérimental nous reviendrons plus en détail sur le formalisme de ces différents indices.

La régression est un processus en deux étapes : la première consiste à construire le modèle, c'est-à-dire déterminer le vecteur de régression. C'est l'étape d'étalonnage. La deuxième consiste à vérifier que notre modèle réagit bien face à de nouvelles données. C'est l'étape dite de test.

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

Un autre critère pour juger le modèle est le calcul de son coefficient de détermination communément noté R^2. « R^2 mesure la proportion de variation totale sur la moyenne des échantillons expliquée par la régression » (Draper et Smith, 1981). R^2 peut être vu comme la corrélation qui existe entre les variables prédites et les variables effectivement mesurées. C'est une valeur comprise entre 0 (peu de corrélation), et 1 (beaucoup de corrélation).

Un bon modèle prédictif est un modèle ayant une erreur d'étalonnage et de prédiction peu élevée et proches l'une de l'autre, ainsi qu'un coefficient de détermination voisin de 1 (Tian et al., 2007).

En jumelant ces deux dernières mesures de performance, un indice appelé écart type des résidus (SDR, *Standard Deviation of Residuals*) est obtenu. C'est le ratio de l'écart type des données de référence et de l'erreur RMSEP. Cet indice peut aiguiller sur les réelles performances de notre modèle quand à l'habilité à la classification (McGlone, Jordan et Martinsen, 2002; McGlone et Kawano, 1998). Une valeur de 3 est généralement considérée comme un minimum pour un but visant une classification acceptable.

Au delà de la technique utilisée pour relier les variables d'entrée et de sortie, il peut être intéressant d'évaluer quelles variables sont effectivement pertinentes dans ce processus. De cette manière, les variables les moins appropriées sont éliminées. Cela nous permettra d'alléger le modèle. Cette sélection de variables va évidement avoir des conséquences sur les performances de notre système. Dans beaucoup de cas, la sélection de variables aboutissant à des modèles plus simples augmente les performances; mais dans d'autre cas, les performances en sont amoindries. Il faut alors faire un compromis entre sélection et performances, pour aboutir à un modèle adapté.

1.3.4 Sélection de variables spectrales

Une particularité majeure des techniques de régression multivariée comme PLS est la génération d'un modèle qui minimise l'influence des variables non contributives de manière significative, et maximise les variables qui contiennent le plus d'information. L'identification de telles variables est très importante puisque le nombre souvent très important de variables inutiles dans les données spectrales contribuera largement à la composante d'erreur, et donc influer sur les capacités prédictives du modèle.

En analyse hyperspectrale, on dispose d'un nombre conséquent de variables pour décrire un phénomène particulier. Généralement la totalité des variables ne sont pas toute pertinentes à la description de nos observations. Il faut donc, à partir de notre ensemble de variables descriptives, effectuer une sélection. Cette sélection a un double avantage : non seulement elle permet de réduire la complexité du modèle, mais aussi, dans la majeure partie des cas,

elle améliore ses performances, et contribue à la détermination d'un modèle plus robuste (Haswell et Walmsley, 1999) .

Une approche possible, et peut-être la plus simple, pour remédier au problème de sélection de variables, est de retirer manuellement les variables qui ne semblent contenir que peu d'informations. Néanmoins, cette approche souffre de deux défauts majeurs : Il n'y a aucune certitude qu'exactement la même zone du spectre soit enlevée à chaque fois. Les zones retirées ne sont peut être pas optimales du point de vue du modèle construit : des parties du spectre peuvent ne pas sembler importantes à l'œil nu, mais pour la construction du modèle, elles contiennent de l'information pertinente. La tendance, lorsque l'on retire manuellement les longueurs d'ondes est de soustraire les zones les plus bruitées, ainsi que les zones où la réponse du détecteur est faible. Cette approche peut être dans certains cas contre-performante lorsqu'on construit un modèle. En effet, l'information contenue dans le bruit des spectres peut se révéler extrêmement utile pour la détermination d'un modèle robuste.

Dans la littérature, la sélection de variables spectrales est un sujet largement débattu. Nombre de méthodes ont vu le jour, toutes avec des approches et des aspects différents. La première approche peut être d'examiner et modifier les coefficients de régression pour effectuer une sélection des variables correspondantes suivant l'amplitude des coefficients obtenus (Garrido Frenich et al., 1995). Cependant cela implique un traitement préalable, étape, comme nous le verrons plus tard, que nous avons choisi de ne pas effectuer. Par la suite, on trouve des méthodes itératives partant de m variables et retirant une à une certaines variables suivant un critère prédéfini. Inversement, partant d'une variable, et les ajoutant une à une itérativement, (Sutter et Kalivas, 1993). D'autres méthodes sélectionnent un ou plusieurs intervalles de longueur d'ondes consécutives (*PLS moving windows* (Du et al., 2004), IPLS (Garrido Frenich et al., 1995; Osborne, Jordan et Rainer, 1997). De même, on trouve des méthodes qui, à partir d'un sous-ensemble, cherchent la meilleure combinaison possible dans le voisinage du sous-ensemble, telle la méthode « tabu » (Hageman et al., 2003). Enfin, des méthodes évolutionnistes comme les algorithmes génétiques ont été appliquées au problème de la sélection de variables (Leardi et Lupianez Gonzalez, 1998) et (Leardi, 2000).

Dans notre cas, nous cherchons à déterminer spécifiquement les longueurs d'ondes les plus pertinentes caractérisant nos observations, et ce avec la meilleure performance. Nous ne pouvons pas directement utiliser les méthodes fournissant un intervalle de variables consécutives, parce que nous cherchons à identifier les longueurs d'ondes de manière spécifique.

Nous avons sélectionné deux approches différentes afin de les comparer. Une méthode itérative, et une approche basée sur le concept de la modélisation d'une colonie de fourmis, c'est une technique probabiliste de recherche de solution, adaptée pour la sélection de caractéristiques.

1.3.4.1 Élagage PLS

L'idée de la méthode d'élagage PLS (du terme anglais *pruning*), introduite par (Lima, Mello et Poppi, 2005), est d'observer la variation de l'erreur du modèle lorsqu'on élimine une variable spécifique. Partant d'un modèle possédant toutes les variables (*m* variables) et, à chaque itération, l'effet sur l'erreur de prédiction du retrait une à la fois de chacune des variables est évalué. Ensuite la variable dont l'erreur évaluée est minimale est retirée, et la procédure est recommencée avec *m-1* variables. L'évaluation de la minimisation est effectuée à partir d'un calcul de saillance vu plus en détail dans le protocole expérimental. Issu du vocabulaire visuel, saillance est un néologisme venant de saillant, signifiant apparaître nettement par contraste.

À chaque étape, un modèle PLS différent est calculé. Il faut donc veiller à optimiser la complexité du modèle (soit le nombre de variables latentes de notre modèle). Cela est réalisé par validation croisée et détermination du minimum de variables latentes en fonction du minimum de la racine carrée de l'erreur quadratique de validation croisée (RMSECV).

Avec cette méthode, est donc calculé *m* modèles différents, le premier avec *m* coefficients, le deuxième avec *m-1*, et ainsi de suite jusqu'à obtenir un modèle à deux, puis un unique

coefficient. Pour chaque modèle une erreur associée est calculée, et le meilleur modèle sera un compromis entre la minimisation de l'erreur, et le plus petit nombre de variables.

Les auteurs (Lima, Mello et Poppi, 2005) ont mis à l'épreuve leur méthode de manière expérimentale. 300 spectres de sirop de sucre en boite ont été acquis entre 1200 et 1600nm à une résolution de 2nm, soit 626 longueurs d'ondes au total. À ces 300 spectres ont été associées 300 valeurs de BRIX (l'échelle servant à mesurer la fraction de sucre présente dans un liquide).

L'évaluation de la performance des modèles construits se fait en calculant l'erreur RMSE, sur un ensemble de test. Ce calcul d'erreur est donc utilisé pour comparer le modèle issu de la méthode d'élagage, et le modèle issu de la régression sur le spectre entier. À cette comparaison de performance, un *F-test* à 95% de confiance à été ajouté afin de vérifier s'il y a des différences significatives entre les modèles (Miller et Miller, 2005).

De manière à réduire le temps de calcul, des 626 longueurs d'ondes initiales, seulement 278 ont été conservées, tout d'abord en supprimant 71 variables dans la zone autour de 2000nm ou le spectre est saturé (*voir* Figure 1.9), ensuite en ne gardant que les longueurs d'ondes impaires.

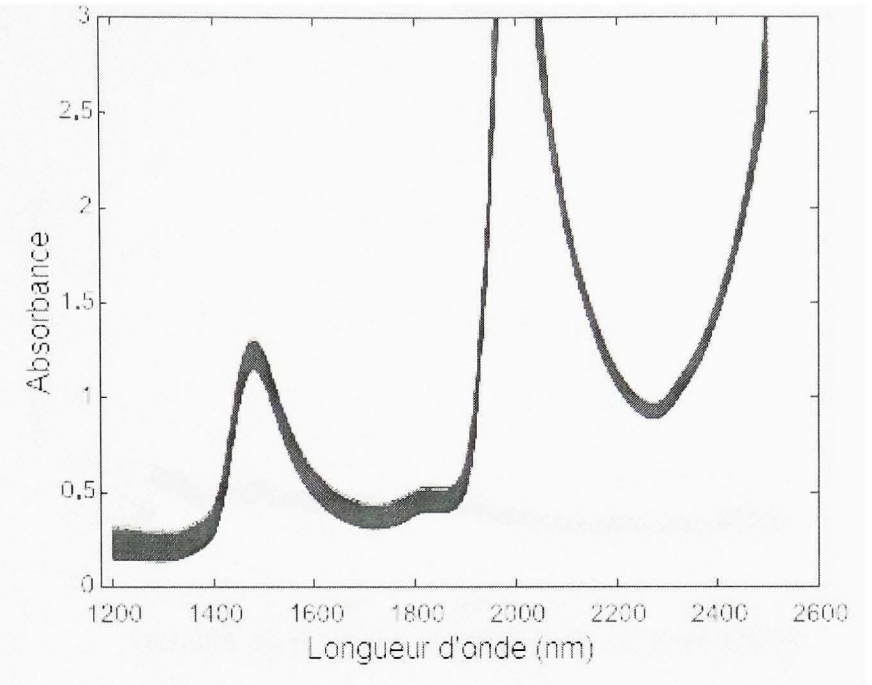

Figure 1.9 ***Spectres de sirop de sucre dans la région proche infrarouge.***

(Tiré de Lima, Mello et Poppi, 2005 [notre traduction])

Source : Cette figure sous forme de graphique est tirée de l'article de M. Silvio Lima et al., *PLS pruning: a new approach to variable selection for multivariate calibration based on Hessian matrix of errors*, p. 4.

Les 300 échantillons ont été divisés en 3 groupes. La répartition s'est faite de manière à ce que dans chaque groupe la variation de la valeur de BRIX soit représentative de la variation au sein de tous les échantillons. 150 sont regroupés pour l'étalonnage du modèle, 75 pour la validation du modèle préalablement construit, et les 75 derniers sont utilisés pour comparer les différents modèles entre eux.

La Figure 1.10 montre l'erreur faite sur l'ensemble de validation au fur et à mesure que l'on diminue le nombre de caractéristiques suivant la méthode d'élagage.

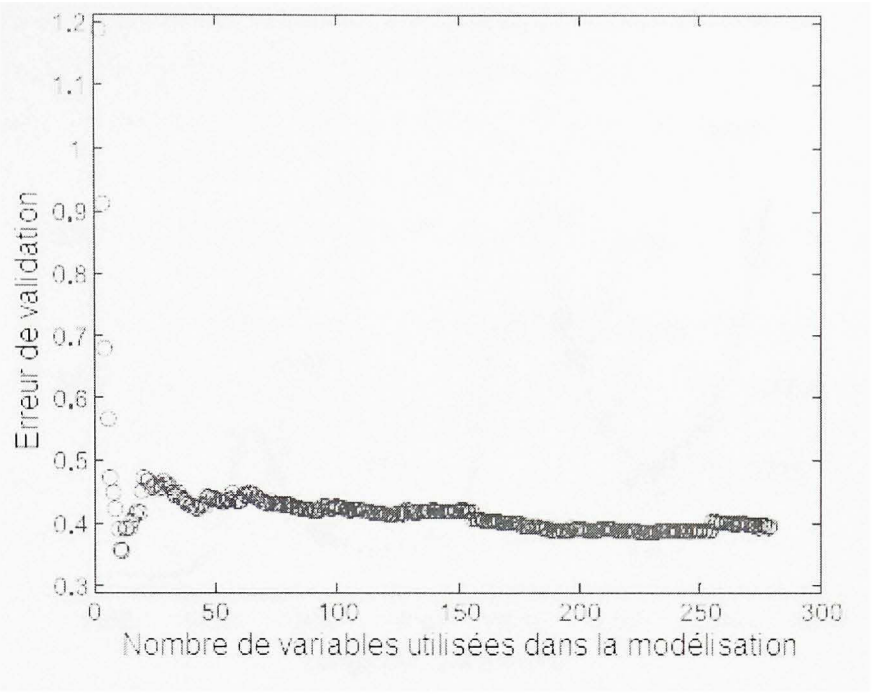

Figure 1.10 ***Erreur de validation en fonction la réduction de variable.***

(Tiré de Lima, Mello et Poppi, 2005 [notre traduction])

Source : Cette figure sous forme de graphique est tirée de l'article de M. Silvio Lima et al., *PLS pruning: a new approach to variable selection for multivariate calibration based on Hessian matrix of errors*, p. 5.

La forme du graphe (*voir* Figure 1.10) est typiquement représentative de l'évolution de cette méthode. Une trop grande réduction des variables (lorsque inférieures à 11) induit une forte augmentation de l'erreur. C'est-à-dire que le modèle n'est plus capable d'établir la relation entre le spectre et les valeurs de BRIX. D'après ces résultats, 11 variables semblent être ici le meilleur compromis entre un minimum de variables et un minimum d'erreurs sur l'ensemble de validation.

Ces 11 variables ont été identifiées au long de la procédure. Nous pouvons voir leur répartition à la Figure 1.11.

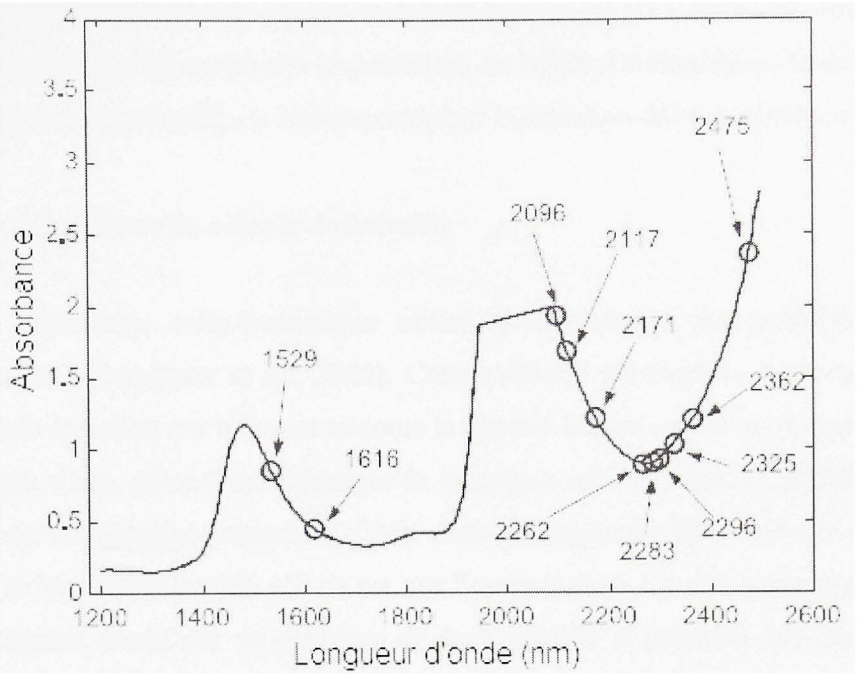

Figure 1.11 ***Sélection de bandes spectrales par l'approche PLS pruning.***

(Tiré de Lima, Mello et Poppi, 2005 [notre traduction])

Source : Cette figure sous forme de graphique est tirée de l'article de M. Silvio Lima et al., *PLS pruning: a new approach to variable selection for multivariate calibration based on Hessian matrix of errors*, p. 5.

Dans le cadre de cette étude, la majeure partie des bandes spectrales retenues sont des bandes impliquées dans les liens chimiques de glucides. Ces bandes sont donc corrélées avec les sucres (fructose, glucose et saccharose) présents dans les échantillons.

La comparaison des modèles issus de l'élagage et de la régression sur l'ensemble des longueurs d'ondes (régression PLS1) se fait en calculant l'erreur commise sur l'ensemble de prédiction. Cette erreur est évaluée à 0.38 pour *pruning*, et 0.40 pour la régression PLS1. Les deux modèles induisent des erreurs très proches. La comparaison de ces valeurs à l'aide du test de Fisher (Haaland et Thomas, 1988) amène à conclure que ces modèles sont statistiquement équivalents. C'est-à-dire que le modèle à 11 variables va produire des résultats équivalents à un modèle construit à partir des 278 bandes originales.

Avec cette méthode et dans le cas particulier de leur étude, les auteurs ont réussi à diminuer de 96% les variables nécessaires à la prédiction du BRIX d'échantillons de sirop de sucre. Il semble que se soit une méthode intéressante pour la sélection de caractéristiques.

1.3.4.2 Algorithme de colonie de fourmis

C'est un algorithme méta-heuristique utilisé pour résoudre des problèmes complexes d'optimisation (Shamsipur et al., 2006). Cette méthode est inspirée du comportement des fourmis dans la nature qui trouvent toujours le chemin le plus rapide entre leur fourmilière et la nourriture. Cela, grâce à des échanges de messages sous la forme de phéromones (odeur) déposées sur les chemins empruntés. C'est donc une approche basée sur des populations de solutions et leur retour positif, s'il en est, sur l'optimisation. Un des remarquables avantages des algorithmes basés sur ce principe est leur habilité à produire une solution presque optimale en peu de temps.

Les fourmis artificielles sont des agents qui imitent le comportement de vraies fourmis, c'est-à-dire qu'elles coopèrent entre elles définissant un système flexible unifié et capable de trouver des solutions performantes aux problèmes où l'espace des solutions possibles est très vaste.

Pour adapter le concept à la sélection de caractéristiques, les fourmis artificielles sont considérées comme des solutions construites de manière probabiliste en choisissant un sous ensemble de variables. Afin que les solutions puissent avoir un retour sur l'optimisation du problème, un vecteur dit de phéromone est utilisé. La taille de ce vecteur est égale au nombre de variables initiales de notre problème, dont chaque élément représente une variable du système. Chaque solution retenue suivant un certain critère de performance va « déposer » sur chacune des variables correspondantes du vecteur de phéromone une certaine quantité proportionnelle à sa performance. Le vecteur de phéromone est utilisé dans la construction probabiliste des fourmis. Cette construction probabiliste va prendre en compte la quantité de phéromone déposée sur chaque variable. C'est-à-dire qu'une variable se trouvant dans

beaucoup de solutions jugées bonnes aura plus de chance d'être choisie qu'une variable étant dans les solutions peu retenues. Une fourmi artificielle a donc une préférence pour les solutions ayant une quantité supérieure de phéromone.

Un tel algorithme réalise itérativement une boucle contenant deux procédures basiques : l'une spécifiant comment construire/modifier les solutions au problème, et l'autre mettant à jour le vecteur de phéromone. Le détail cet algorithme et son implémentation seront vus plus en détail dans le protocole expérimental.

L'algorithme de colonie de fourmis (ACF) conventionnelle est un processus d'optimisation comme le problème du voyageur de commerce (*Traveling Salesman Problem,* TSP) (Colorni, Dorigo et Maniezzo, 1991), où, l'on cherche à optimiser le chemin le plus court pour rejoindre certaines villes. Cependant, la sélection de variables est un problème de sélection d'un sous-ensemble, et est différente d'un problème d'ordonnancement. Ici, il n'y a pas véritablement de notion de chemin, une adaptation a donc été faite pour appliquer le concept à la sélection de variables.

Pour tester les habilités de sélection de cet algorithme les auteurs (Shamsipur et al., 2006), ont expérimenté trois modèles de régression différents *Classical Least Squares (CLS), Inverse Least Squares (ILS), et Partial Least Square (PLS*), sur quatre bases de données spectroscopiques différentes (visible et proche infrarouge).

Afin de mesurer l'efficacité de l'algorithme, les performances données par les modèles construits à partir des bandes sélectionnées par cette méthode sont comparées aux bandes sélectionnées par deux autres algorithmes, qui sont l'élimination descendante (en partant du nombre total de variables, on tente de le diminuer), et la sélection directe (méthode inverse, on part d'une seule longueur d'onde, et on en ajoute à chaque itération). Sur les quatre bases de données, il ressort que la sélection de variables obtenue par la méthode stochastique est identique en nombre, sinon moindre, et cela pour des performances toujours améliorées. Pour

exemple, la quatrième base de données est composée d'échantillons de blé associés à des teneurs en humidité et protéines, dont les résultats sont présentés au Tableau 1.1.

Tableau 1.1

Comparaison des résultats prédictifs entre 3 méthodes de sélection de variables (Tiré de Shamsipur et al., 2006)

Méthode de sélection des variables	Humidité		Protéines	
	MSEP	**Nb. bandes**	**MSEP**	**Nb. bandes**
ACF	0.0627	5	0.2422	4
Sélection directe	0.0715	5	0.8027	9
Elimination descendante	0.0631	170	0.4830	423

Source : Ce tableau est tiré de l'article de M. Mojtaba Shamsipur et al., *Ant colony optimisation: a powerful tool for wavelength selection*, p.9.

La méthode rapportée ici est une méthode intéressante comme alternative aux méthodes plus conventionnelles. Testée dans le cadre de quatre bases de données différentes (couvrant les ultraviolets, le domaine du visible, et le proche infrarouge), elle a démontré sa capacité à sélectionner les longueurs d'ondes pertinentes dans le cadre de ces bases de données. Cependant, cette méthode requiert certains ajustements de paramètres qui seront différents suivant les conditions d'application de l'algorithme. Le détail de ces paramètres, ainsi que leur optimisation pour aboutir à des solutions viables, sera vu dans le protocole expérimental.

1.4 Utilisation de la vision par ordinateur comme méthode non destructrice d'évaluation de paramètres dans l'industrie agro-alimentaire

1.4.1 Mise en contexte

Le domaine de la vision par ordinateur est une technologie récente servant à l'acquisition et au traitement d'images provenant de scènes réelles. A l'aide de ces acquisitions, on peut extraire l'information pertinente, ou encore effectuer le contrôle de divers procédés. La base

de la vision par ordinateur repose sur l'analyse et le traitement des informations. C'est à partir de cette analyse/traitement que l'on va quantifier, et éventuellement classifier les images, ou certains objets pertinents contenus dans les images (Sun, 2004) .

La vision par ordinateur est un domaine assez jeune qui remonte aux années 1960. Créant un fort intérêt dans le monde industriel, la vision par ordinateur s'est fortement développée autant au niveau théorique qu'au niveau pratique. De nos jours, la vision est présente dans de domaines nombreux et très variés, comme par exemple la médecine, la surveillance, ou encore le contrôle à distance de véhicules ou de robots. Avec ces développements, la vision par ordinateur tend à remplacer l'œil et les méthodes humaines dans l'évaluation qualitative et l'automatisation des procédés jugés selon leur aspect visuel.

Conjointement au développement du matériel et des logiciels, la vision par ordinateur a été étendue au contrôle de qualité appliqué à l'agriculture. On peut citer par exemple : la télédétection, l'agriculture de précision, l'évaluation de la qualité et salubrité de la production avant et après la récolte, aussi bien que le tri des produits selon des critères adéquats.

La vision par ordinateur a démontré de gros avantages dans l'évaluation rapide et sans contact de la nourriture. En effet, c'est un moyen rapide d'obtenir non seulement de nombreuses informations telles que la couleur, la forme, la taille, ou encore des attributs de texture, mais aussi, de recueillir des attributs numériques concernant certains objets spécifiques de la scène étudiée (Chen, Chao et Kim, 2002) .

L'estimation de la qualité par des méthodes de vision a un éventail d'applications très large dans l'agriculture. Détection de défauts, de maladies, de la maturité, et autres attributs de qualité des fruits et légumes.

L'utilisation de la vision par ordinateur donne généralement des résultats assez précis et cohérents, et ce de manière non destructrice. Les applications de la vision par ordinateur améliorent la production de l'industrie, en conséquence de quoi, les coûts sont réduits, les

opérations agricoles sur le terrain sont minimisées, les chaines de production sont ainsi réduites et plus sécurisées. L'automatisation et l'optimisation données par les systèmes de vision permettent d'établir et d'appliquer des standards de production élevés. Ce qui, au final, donne des produits de bonne qualité pour le consommateur, et moins de rejets.

Un système de vision par ordinateur se compose d'une caméra, d'un ordinateur muni d'une carte d'acquisition et de logiciels, et d'un système d'éclairage. C'est avec ce matériel qu'est réalisée l'acquisition de la réflectance, transmitance, ou fluorescence sous une illumination ultraviolette (200-380nm), visible (380-780nm) et proche infrarouge (780-2500nm).

Selon ses propriétés optiques, lorsqu'un objet est éclairé, la lumière est réfléchie, transmise, ou absorbée (lorsque la lumière absorbée est réémise on appelle cela fluorescence). Ces propriétés dépendent de sa composition physico-chimique, ainsi que de la lumière utilisée (type de lumière et angle d'incidence).

La vision par ordinateur a un gros potentiel dans le secteur agricole. Dans une étude réalisée sur le sujet, les auteurs ont passé en revue les dernières techniques d'analyse dans le domaine de la vision par ordinateur appliquée à l'industrie agroalimentaire (Brosnan et Sun, 2004). Ils ont basé leur étude sur différentes applications de systèmes de vision à des produits issus de l'industrie agroalimentaire. Leur étude s'appuie sur le travail effectué par d'autres chercheurs allant des produits à l'état brut (fruits, légumes), aux produits finis (nourriture préparée prête à être consommée), en passant par l'inspection des contenants. De cette étude ils tirent les avantages et inconvénients de l'application de la vision par ordinateur à l'industrie agroalimentaire (*Voir* Tableau 1.2).

Tableau 1.2

Avantages et inconvénients d'un système de vision
(Tiré de Brosnan et Sun, 2004)

Avantages	Inconvénients
Génération de données descriptives précises Technique non destructrice et non perturbante Réduction de la contribution parfois fastidieuse de l'humain Méthode efficace, cohérente, et rentable Automatisation de processus d'acquisition parfois intensif Technique de perception robuste et compétitive monétairement Données enregistrées de manière permanente, permettant une analyse décalée dans le temps	Identification d'objets difficile parfois (dans une scène peu ou pas structurée). Besoin de lumière artificielle lorsque la lumière présente n'est pas satisfaisante (intensité, et/ou type)

Source : Ce tableau est tiré de l'article de Tadhg Brosnan et al., *Improving quality inspection of food products by computer vision--a review*, p. 2.

La vision traditionnelle en couleur consiste en la perception par les appareils de 3 principales composantes qui sont le rouge, le vert et le bleu. Munis de ces 3 composantes, toutes les couleurs visibles ne peuvent être recomposées. Lors de l'acquisition il y a une perte d'information concernant les informations réellement contenues dans le spectre (à la fois dans le visible, mais surtout dans le non-visible). Nous pouvons observer l'ensemble des couleurs du spectre défini par la Commission Internationale de l'Eclairage (CIE), ainsi que l'espace de couleur défini par les trois couleurs rouge, vert et bleu à la Figure 1.12.

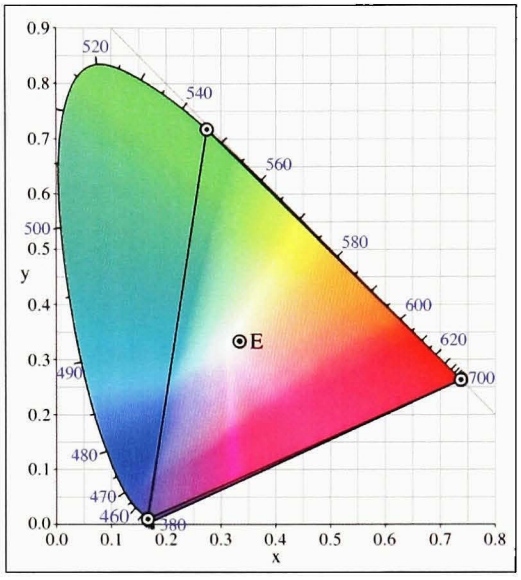

Figure 1.12 ***Diagramme de chromaticité CIE 1931 incluant l'espace de couleur RGB.***
(Tiré de Wikipédia, 2008a)

Nous remarquons (*Voir* Figure 1.12) qu'avec les 3 composantes RGB, toutes les couleurs du spectre ne peuvent pas être recomposées : seulement les couleurs présentes à l'intérieur du triangle le peuvent. Afin d'augmenter le nombre de couleurs représentables par notre espace de couleur, il faut augmenter l'aire couverte par le triangle initial. C'est réalisable en maximisant l'aire couverte par le triangle, cependant, cette aire à une limite bien inférieure à l'aire totale. Des dimensions peuvent aussi être ajoutées à notre espace, en adjoignant des points de référence. De cette manière l'aire couverte par ce polygone ainsi défini peut augmenter significativement. Ainsi sa limite tend vers l'aire maximum. La collecte simultanée de différentes couleurs successives, appelée spectre, peut être réalisée au moyen d'appareils d'acquisition spécifiques : les spectromètres.

1.4.2 Spectroscopie

La spectroscopie est largement utilisée dans l'industrie agro-alimentaire afin de déterminer divers paramètres qualitatifs des produits. Ces produits sont souvent bruts (fruits ou légumes), mais peuvent aussi être des produits manufacturés. L'inspection du fromage manufacturé a été réalisée par spectroscopie (Curda et Kukackova, 2004).

L'approche de l'analyse spectroscopique consiste à relier les données spectrales et les données physico-chimiques du produit étudié. Le cheminement consiste donc à trouver la relation qui existe entre les variations présentes dans le spectre et les variations des données d'analyses physico-chimiques.

L'analyse spectroscopique passe par plusieurs étapes qui commencent par la sélection du ou des paramètres à étudier (paramètres qualitatifs). C'est, en d'autres termes, les variables dépendantes de notre système. Ensuite est définie la plage spectrale utilisée pour tenter d'évaluer ce paramètre. Une fois cette plage spectrale choisie, il faut sélectionner les modèles statistiques à utiliser pour étudier et lier le paramètre qualitatif aux données spectrales. Après cela, c'est l'étape de prétraitement, soit les traitements à appliquer aux données spectrales de manière à éventuellement améliorer les résultats prédictifs. Enfin viennent les étapes d'étalonnage et test du modèle. L'étalonnage consiste à calculer les différents paramètres du modèle à partir d'un sous-ensemble de la base de données. Le test consiste à utiliser le modèle calculé lors de l'étalonnage sur un autre sous-ensemble de la base de donnée et d'évaluer certains paramètres de performance afin d'appréhender la viabilité du modèle.

1.4.2.1 Évaluation des paramètres

Les paramètres étudiés sont assez divers, mais toujours bien appropriés au produit étudié. Les paramètres peuvent être très subjectifs, par exemple l'évaluation du goût des pommes de terre; en fait le but est de déceler les goûts éventuellement inhabituels (Van de Laer et al., 2001). Outre les paramètres intrinsèques aux produits, la spectroscopie peut aider à classer différentes variétés d'un même produit. La distinction entre différentes variétés, ou différents types de productions, de pommes (Bennedsen et Abu-Khalaf, 2004) , ou encore différentes cultures pour les carottes (Abu-Khalaf, Bennedsen et Bjorn, 2004). Plus généralement, concernant les fruits et légumes, les paramètres les plus étudiés sont : le degré BRIX (ou taux de sucre), la fermeté, et le taux de matière sèche. Concernant la mangue, l'évolution du degré BRIX contenu dans la mangue pour caractériser leur goût (Jha, Chopra et Kingsly, 2005),

ainsi que le taux de matière sèche et la fermeté de la mangue (Subedi, Walsh et Owens, 2007), sont deux études qui ont de bons résultats. On retrouve aussi des paramètres de densité, d'acidité, et d'autres vraiment spécifiques à une sorte de produits. Parfois, la combinaison de plusieurs paramètres peut donner une évaluation d'un paramètre plus subjectif à priori, par exemple des chercheurs ont combiné le degré BRIX, la fermeté et l'acidité pour caractériser le goût des pommes (Bennedsen et Abu-Khalaf, 2004). Nous allons étudier la maturité de l'avocat. Cette maturité est représentée par la quantité de matière sèche présente dans l'avocat.

1.4.2.2 Plage spectrale

Les différentes plages spectrales utilisées dans l'industrie agro-alimentaire vont des ultra-violets au proches infrarouges, soit de 200 à 2500 nm. Suivant la plage spectrale, et selon la résolution spectrale, cela peut représenter un très grand nombre de données. Il est généralement préférable de définir une plage éventuellement plus petite dans laquelle cibler la recherche. Soit parce que le cahier des charges l'impose, soit parce que l'on sait à priori que cette zone contient les informations discriminantes. Dans leur étude de la qualité des concombres, les auteurs utilisent une plage spectrale en interactance de 400 à 700nm à une résolution de 2 nm soit 220 données spectrales par échantillon (Kavdir et al., 2007). L'interactance est une technique pour évaluer la composition d'un corps. On émet vers ce corps un rayonnement électromagnétique, et l'on mesure l'énergie que l'on récupère. Nous pouvons observer le fonctionnement de l'acquisition du spectre en interactance sur Figure 1.13.

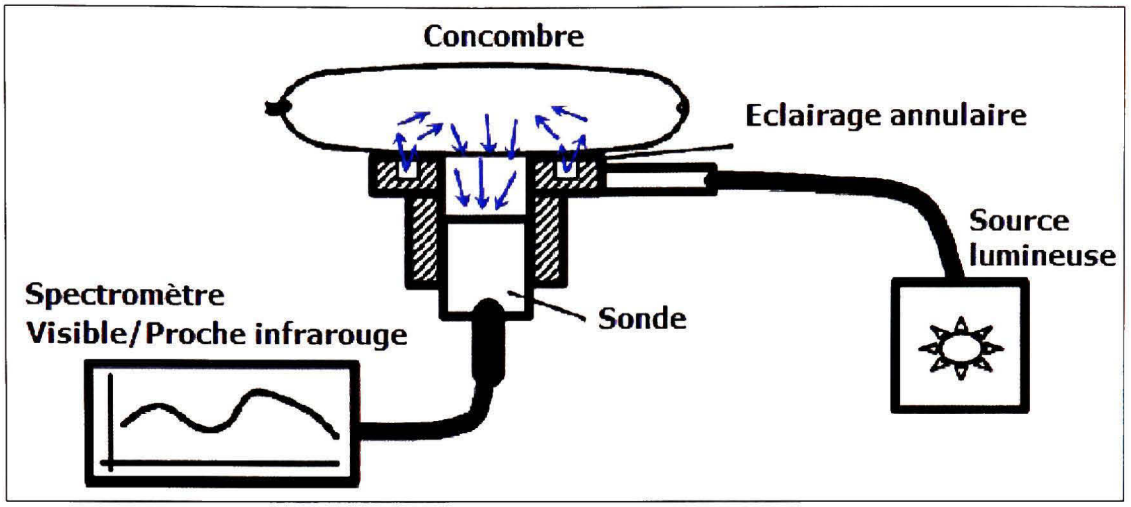

Figure 1.13 ***Mesure spectrale sur le concombre en mode interactance.***

(Tiré de Kavdir et al., 2007 [notre traduction])

Source : Cette figure est tirée de l'article de M. I. Kavdir et al., *Visible and near-infrared spectroscopy for non-destructive quality assessment of pickling cucumbers*, p. 3.

Lorsque la plage spectrale est vraiment large, il est souvent nécessaire de diviser celle-ci afin de voir dans quelle sous plage est présente l'information discriminante. L'étude de la qualité de la tomate a été faite en utilisant une plage spectrale en réflectance de 350 à 2500nm (He et al., 2005). Les acquisitions ont été réalisées à 3 points équidistants autour de l'équateur de la tomate puis moyennés. Au vu de la taille de la plage spectrale, il a été décidé de couper celle-ci en 2 : de 350 à 1000nm, et de 1000 à 2500nm. Le choix des plages spectrales peut aussi résulter en des plages se chevauchant. Dans leur étude de la maturité des avocats (Clark et al., 2003), les auteurs ont une plage spectrale de 300 à 1140nm pour une résolution de 3.3 nm, soit 256 valeurs. Quatre fenêtres ont été définies : de 500 à 1050nm, de 500 à 750nm, de 750 à 1050nm, et de 800 à 1000nm. Cela permet de cibler mieux les zones. Les modes d'acquisition choisis ont été l'interactance, et la réflectance. Il s'est avéré que l'interactance a donné de meilleurs résultats. L'étude de la maturité des kiwis (McGlone et Kawano, 1998) a été effectuée à partir d'une plage spectrale de 400 à 1000nm pour une résolution de 10nm. Les données ont été acquises en mode réflectance, mais le traitement des données a été réalisé à partir de l'absorbance (l'absorbance est le logarithme de l'inverse de la réflectance). Dans notre étude, le spectre dans son ensemble à la fois en absorbance et en réflectance va

être considéré. Sans chercher une zone spectrale intéressante, nous voulons trouver quelles bandes sont discriminantes pour la prédiction, et ainsi pouvoir déterminer quel mode (réflectance ou absorbance) est le meilleur.

1.4.2.3 Modèle statistique

Les modèles statistiques les plus souvent utilisés sont les modèles PCR, PLS, et MLR. De ces 3 modèles, le PLS est plus largement utilisé. Dans l'étude pour la prédiction de l'acidité des pommes (Liu, Ying et Fu, 2004), les 2 modèles prédictifs ont été le PCR et le PLS. C'est le PLS qui a donné les meilleurs résultats. De manière générale, le modèle MLR est souvent utilisé, car il est plus simple, mais plus limité, que le model PLS. Dans notre étude, nous allons étudier le modèle PLS, qui semble être le plus abouti de ces modèles de régression. Le modèle PCR est, comme nous l'avons vu, à mi-chemin entre le PLS et MLR, nous n'allons donc pas nous attarder à évaluer ses performances, puisque la régression PLS est plus complète, et la régression MLR plus simple.

1.4.2.4 Prétraitements

Le prétraitement (ou *preprocessing*) est une étape précédant la construction du modèle. Cette phase est utile, non seulement pour enlever le bruit éventuel dû aux instruments, mais aussi pour effectuer des traitements sur le spectre visant à augmenter les performances du modèle.

L'élimination du bruit peut se faire par lissage : moyenne, polynômes de Savinsky-Golay, ou par analyse de Fourier. La méthode de Savinsky-Golay calcule une série de polynôme en adéquation avec les valeurs du spectre. On utilise alors les données issues de cette interpolation. La méthode fréquentielle supprime les bruits haute fréquence en calculant la transformée de Fourier du spectre et en éliminant une bonne partie des coefficients présents dans les hautes fréquences. La transformée inverse est ensuite calculée pour obtenir les données spectrales dont le spectre a ainsi été filtré passe-bas (Fearn et Davies, 2002).

Les prétraitements cherchant à améliorer les performances, sont des méthodes visant à retirer du spectre l'information sans lien avec l'analyse effectuée. Parmi ces prétraitements, sont retrouvées les dérivées premières et secondes (Davies, 2007), et des méthodes un peu plus avancées comme « Multiplicative Scatter Correction » (MSC) (Fearn et Davies, 2007), « Standard Normal Variate » (SNV) ou encore « Orthogonal Signal Correction » (OSC) (Fearn et Davies, 2002).

Dans la littérature, l'utilisation de prétraitement est largement répandue, mais les résultats sont assez mitigés. Dans leur étude sur la mangue (Jha, Chopra et Kingsly, 2005), aucune des méthodes de prétraitement : lissage, seconde dérivée, ou MSC ne produit de meilleurs modèles.

Dans notre étude, nous avons choisi de ne pas utiliser de prétraitement. En effet, nous voulons développer des modèles utilisant au mieux les données brutes afin de trouver quelles sont les longueurs d'ondes les plus pertinentes répondant à notre problème.

1.4.2.5 Étalonnage et performances du modèle.

Une fois que toutes les étapes précédentes ont été accomplies, le modèle est prêt à être étalonné, pour ensuite tester ses performances. De manière générale, les travaux présentés dans la littérature donnent de bons résultats concernant la prédiction de leurs paramètres à l'aide de la spectroscopie. Dans leur étude de la matière sèche contenue dans les grains de maïs (Montes et al., 2006), les auteurs arrivent à de bons résultats prédictifs : un coefficient de détermination (R^2) de 0.95, et une erreur de prédiction (RMSEP) de 1.2. De la même manière, l'étude de la matière sèche contenue dans des fromages manufacturés (Curda et Kukackova, 2004), prouve que la spectroscopie est un moyen efficace d'inspection sans contact. En effet, les auteurs obtiennent un coefficient de corrélation de 0.99, et une erreur de prédiction de 0.43. Une autre étude sur la mangue (Subedi, Walsh et Owens, 2007), prouve que la matière sèche est bien corrélée avec le spectre, avec un coefficient de corrélation de 0.96, et une erreur de prédiction de 0.41. De l'étude de la maturité des avocats (Clark et al.,

2003), il ressort que le mode interactance (R^2=0.81) donne de meilleurs performances que le mode réflectance (R^2=0.64) lors de l'analyse utilisant le spectre dans son entier.

La spectroscopie semble être un moyen sans contact assez efficace pour la prédiction de divers paramètres de produits. Cependant, la spectroscopie, si elle fournit l'information spectrale, n'offre aucune information spatiale de la scène.

1.4.3 Imagerie hyperspectrale

Comme nous l'avons vu précédemment, l'imagerie hyperspectrale est une généralisation de la spectroscopie, puisque en plus de la dimension spectrale, on ajoute la dimension spatiale à l'instrument. Gowen et al., (2007) ont fait une étude de l'imagerie hyperspectrale et son utilité dans l'industrie agro-alimentaire. C'est une discipline émergente et un puissant outil pour l'identification des bandes pertinentes ainsi que pour l'automatisation de l'inspection. Comme la spectroscopie, l'imagerie hyperspectrale peut être réalisée en réflectance, transmitance ou fluorescence. Cependant, il ressort de récents travaux que la réflectance est le mode le plus communément utilisé. Cela dans le visible (400-700nm), et dans le proche infrarouge (100-1700nm). Son utilisation est assez variée : caractérisation des défauts, maladies, contaminants, et attributs des fruits, légumes et viandes.

Une équipe de recherche a utilisé l'imagerie hyperspectrale pour la caractérisation de la maturité des fraises (ElMasry et al., 2007). Les bandes spectrales optimales ont été obtenues à l'aide de la régression PLS, et l'analyse MLR a ensuite été conduite pour déterminer l'humidité (MC, Moisture Content), le pH et le BRIX. Les données spectrales consistaient en des spectres (826 variables spectrales) moyennés suivant une région d'intérêt du fruit. L'analyse PLS a permis de déterminer les longueurs d'ondes optimales en ne sélectionnant que les bandes correspondantes aux plus forts coefficients du vecteur de régression. À la suite de quoi une régression MLR est conduite et aboutie aux performances présentées dans le Tableau 1.3.

Tableau 1.3

Performance sur l'ensemble de validation des modèles MLR
(Tiré de El Masry et al., 2007)

Attribut	Nb. Bandes retenues	Validation	
		RMSEP	R^2
MC	8	5.786	0.91
Brix	6	0.211	0.80
pH	8	0.091	0.94

Source : Ce tableau est tiré de l'article de M. Gamal El Masry, *Hyperspectral imaging for nondestructive determination of some quality attributes for strawberry*, p. 7.

L'inspection de la tomate a montré que dans la région 396-736nm, l'imagerie hyperspectrale est plus performante que l'analyse RGB pour évaluer l'état de maturité indépendamment des conditions d'illuminations (Polder, Heijden et Young, 2002).

Nous voulons étudier la maturité des avocats à l'aide de l'imagerie hyperspectrale. Au vu des travaux qui ont été menés, l'imagerie hyperspectrale peut être un bon outil pour caractériser la maturité des fruits et légumes.

1.5 L'avocat

1.5.1 Description

L'avocat (*persea americana*) est un fruit tropical originaire du Mexique où il est cultivé depuis plus de 8000 ans. De nos jours, l'avocat est produit dans 66 pays de par le monde. La superficie mondiale utilisée pour cette culture s'élève à 388 000 hectares, pour un total de 3 330 168 tonnes (FAO, 2006). Le Tableau 1.4 présente les principaux pays producteurs et leur part de la production mondiale.

Tableau 1.4

Principaux pays producteurs d'avocats

(Tiré de FAO, 2006)

Pays	**% De la production mondiale**
Mexique	34.28
Etats Unis D'Amérique	7.45
Indonésie	6.86
Colombie	5.60
Brésil	5.11
Chili	4.92
République Dominicaine	3.43
Pérou	3.12
Autres 56 pays	29.23

Source : Ce tableau est tiré des statistiques fournies par la FAO (Food and Agriculture Organization of the United Nations) pour l'année 2006.

L'avocat frais est retrouvé à peu près partout dans le monde. L'avocat est essentiellement consommé comme fruit, ou sous forme de purée.

Pour la majorité des avocats, les constituants principaux sont : 73% d'eau, 15% de lipides, 7% de fibres, 9% de glucides, 2% de protéines. Les avocats sont de bonnes sources de potassium, et de vitamines A, K, et C.

L'avocat est normalement récolté après avoir atteint sa maturité. Pour le marché de l'exportation, il est récolté à maturité légale, 20.8% (Koo-Lee, 2003), mais lorsqu'il a atteint sa taille adulte. Les avocats sont récoltés à la main, ou à l'aide de bras motorisés actionnés par un operateur pour les fruits hors de portée. C'est un fruit climatérique, c'est-à-dire qu'il continue à mûrir une fois récolté de l'arbre. Au fur et à mesure que le fruit mûrit, sa peau passe du vert clair au marron foncé.

Un avocat de la variété Hass, qui est notre sujet d'étude, met environ 5 à 10 jours pour mûrir complètement lorsqu'il est stocké à température ambiante (22°C-23°C). On peut par contre le garder de 7 à 9 semaines lorsqu'il est entreposé à une température de 4°C-6°C.

1.5.2 Qualité et maturité des avocats

La qualité d'un avocat se reconnait à sa forme, taille, couleur de la peau, fermeté de la chair, couleur de la chair, dégâts éventuels dus à des maladies, ou des chocs.

Les avocats ont des caractéristiques qui peuvent aiguiller sur la maturité du fruit. Juger la maturité simplement sur ces caractéristiques n'est pas une méthode fiable. Parmi ces caractéristiques, nous pouvons remarquer que (McCarthy, 2005) :

- La queue du fruit est plus large, plutôt gonflée, et jaunâtre plus que verte;
- L'enveloppe du noyau est sèche, foncée;
- La peau est terne et sans éclat;

De manière standard, la maturité de l'avocat est évaluée en calculant la teneur en huile du fruit. L'extraction de la teneur en huile du fruit est une méthode qui, pour être assez précise, utilise la RMN (Résonnance Magnétique Nucléaire), ce qui nécessite un équipement très onéreux. Cependant il à été démontré (Lee et al., 1983) que cette teneur en huile est hautement corrélée avec le taux de matière sèche, qui est un paramètre plus facile à extraire. C'est pourquoi la détermination du taux de matière sèche est utilisé comme standard mondial pour évaluer la maturité des avocats à récolter, ou fraichement récoltés (Woolf et al., 2003).

CHAPITRE 2

PROTOCOLE EXPERIMENTAL

2.1 Aperçu général

Au travers de ce chapitre nous expliquerons en détail les différentes facettes de nos expérimentations. Premièrement, nous verrons comment est extraite la matière sèche de l'avocat. Ensuite, nous étudierons le matériel d'acquisition à notre disposition : quelles sont ses caractéristiques, et comment nous avons conçu le montage pour tirer le meilleur parti de l'équipement compte tenu de ses limitations et des spécificités de notre sujet. Les données brutes issues du système d'acquisition ne sont pas utilisables directement. Nous verrons comment étalonner ces données afin de pouvoir les utiliser correctement. Enfin nous nous pencherons en détail sur les outils que nous avons choisis pour étudier la maturité des avocats par imagerie hyperspectrale; de quelle manière fonctionnent ces outils, comment nous les avons paramétrés et utilisés dans le cadre de notre étude.

2.2 Analyse physique des avocats

2.2.1 Extraction de la matière sèche

La maturité des avocats est évaluée en théorie par le taux d'huile contenu dans l'avocat, (Charles, 1978). Extraire l'huile contenue dans un avocat est un processus plutôt compliqué et onéreux, requérant un équipement spécifique et certaines compétences. Cependant, il a été démontré que le taux d'huile contenu dans le fruit est intimement lié au taux de matière sèche de celui-ci (Lee et al., 1983), en effet le taux d'huile varie proportionnellement au taux de matière sèche du fruit.

Le taux de matière sèche, exprimé généralement en pourcentage, représente la quantité de matière privée de son eau. Cette quantité est obtenue en faisant chauffer un échantillon du fruit de manière à faire évaporer l'eau. L'extraction de la matière sèche, à l'inverse de

l'extraction de l'huile est un processus plus facile à mettre en place et plusieurs techniques sont adéquates. Afin de faire chauffer les échantillons, deux méthodes distinctes sont possibles : l'utilisation d'un four à micro-ondes ou l'utilisation d'une étuve (four à chaleur contrôlée). Le four à micro-ondes est une méthode où un seul échantillon peut être traité à la fois. Le procédé ne peut pas vraiment être calibré car le temps de chauffage varie de 5 minutes à 30 minutes, suivant la maturité de l'échantillon et la puissance utilisée, la chaleur étant très forte et peu contrôlable. Il faut systématiquement vérifier que l'échantillon ne brûle pas car même à basse puissance, la température est très haute (Woolf et al., 2003). Avec une étuve, la température utilisée est constante, et réglée de telle manière à ce que l'eau s'évapore : 105°C, c'est-à-dire une température juste supérieure à celle de l'ébullition de l'eau (100°C). Le temps de chauffage, pour un échantillon de 10g, est à priori de 5 heures (Hickson, 2006). Généralement, les étuves permettent le chauffage de plusieurs échantillons simultanément. Le Tableau 2.1 présente un résumé des avantages et limitations des deux méthodes.

Tableau 2.1

Avantage/désavantages des méthodes de chauffage de l'avocat
(Tiré de Hickson, 2006)

Four micro-ondes	**Four conventionnel**
Rapide (<30 minutes)	Lent (5 heures)
Très haute température (varie selon le four)	Température constante (105°C)
Ne peut pas être calibré	Peut être calibré
Un seul échantillon chauffé simultanément	Plusieurs échantillons chauffés simultanément

Source : Ce tableau est tiré d'une présentation de M. Brett Hickson, *Quality assessment of avocados by means of dry matter content*, p. 9.

Le four à micro-ondes peut être à priori une solution viable ne requérant que peu d'équipement. Cependant, le fait de ne pas pouvoir calibrer le temps de chauffage, ainsi que la limitation à un échantillon chauffé à la fois, mène à la conclusion que cette approche n'est

pas forcement adaptée pour réaliser une étude dans laquelle plusieurs centaines d'échantillons devront être chauffés. Pour ces raisons, nous nous sommes tournés vers l'option intéressante que peut représenter l'étuve. Une telle méthode est aisément étalonnable, et nous pouvons faire chauffer plusieurs échantillons simultanément. Le laboratoire d'environnement STEPPE à l'ETS possède une étuve. Mr R. Hausler, directeur du laboratoire, a eu la gentillesse de nous laisser l'utiliser pour faire nos chauffages. Au moyen de cette étuve, nous pouvons faire chauffer 24 échantillons d'avocats, soit les données de 6 avocats.

Nous avons surveillé l'évolution du taux de matière sèche de quatre avocats à différents stades de maturité. Les résultats sont présentés à la Figure 2.1. Nous pouvons remarquer qu'effectivement 5h de chauffage semble être un temps adéquat pour assécher complètement les échantillons d'avocats.

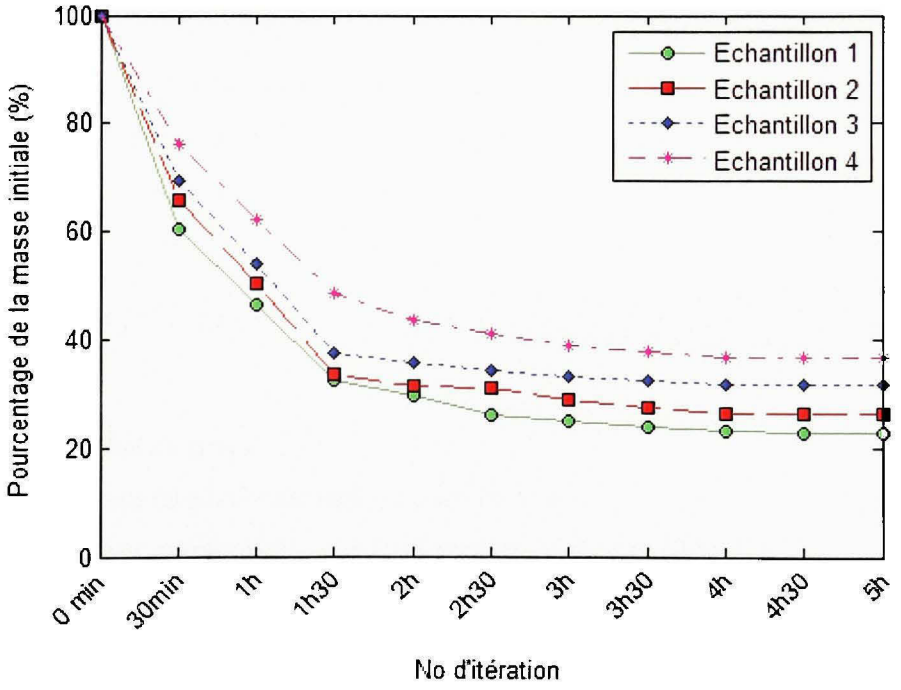

Figure 2.1 ***Étalonnage du temps de chauffage de différents avocats.***

2.2.2 Préparation des échantillons

La préparation des échantillons requiert le matériel suivant :

- Étuve à 105°C
- Balance avec une précision de 0.01g
- Couteau et planche à découper
- Plats allant au four

Quatre échantillons de chair sont prélevés suivant les quatre points cardinaux de l'avocat par rapport au pédoncule du fruit. De fines tranches, entre un et deux millimètres d'épaisseur, sont coupées suivant la longueur du fruit. Ces tranches sont pelées et le noyau est enlevé pour former une masse de chair de 10g±1g. Les échantillons sont disposés dans des plats adéquats préalablement pesés. La pesée de la chair fraiche des échantillons avec le plat est réalisée, ensuite les échantillons sont séchés. Une fois le processus de chauffage terminé, le plat est à nouveau pesé. Le taux de matière sèche, exprimé en pourcent, est le rapport des masses avant et après chauffage suivant la formule (2.1). Une description plus complète du procédé est disponible à l'Annexe I.

$$\text{matière sèche } (\%) = \frac{C-A}{B-A} \times 100 = \frac{\text{masse de matière sechée}}{\text{masse de matière fraiche}} \times 100 \qquad (2.1)$$

Avec : A : masse du plat à vide.

B : masse du plat contenant l'échantillon frais.

C : masse du plat contenant l'échantillon après chauffage.

La matière sèche est, comme nous l'avons vu, le standard pratique pour évaluer la maturité des avocats. D'après les conventions, un avocat peut être mis sur le marché lorsque son taux de matière sèche est supérieur à 20.8% pour la variété Hass (Koo-Lee, 2003).

2.2.3 Couleur et maturité des avocats

Lorsque le fruit est laissé à température ambiante, sa fermeté varie : il devient plus mou, et sa couleur change. Cependant, nous avons remarqué que ces changements n'influaient pas sur le taux de matière sèche du fruit. Le taux de matière sèche n'est donc pas, à priori, corrélé au changement de couleur et au changement de fermeté. Deux études sur l'évolution de la maturité de ces fruits, semblent confirmer ces résultats (Cox et al., 2004; Ozdemir et Topuz, 2004).

2.2.4 Base de données de matière sèche

Initialement, nous avons mesuré la matière sèche et le spectre des échantillons issus de 42 avocats. Parmi ces 168 mesures réalisées, certaines ont été faites le jour même de l'acquisition du lot, et d'autre ont été réalisées de un à trois jours suivant l'acquisition des lots (pendant ce laps de temps, les avocats étaient laissés à la température ambiante de la pièce soit environ 22°C). Sachant que, lorsqu'un avocat est laissé à température ambiante, sa couleur et sa fermeté varient, à l'inverse de son taux de matière sèche qui reste constant, il faut mettre de coté ces échantillons, de manière à ne pas handicaper notre modèle. Des 168 échantillons initiaux, nous ne gardons donc que ceux dont les mesures ont été réalisées dans la journée suivant leur acquisition. Cela correspond à 21 avocats (nous avons donc mis 21 avocats de coté). Ayant retiré les 84 échantillons, nous allons, à partir des 84 échantillons sélectionnés, effectuer quelques calculs statistiques, de manière à voir si ceux-ci sont homogènes. Il est à noter que les avocats viennent de différents marchés locaux et ont été achetés à différentes périodes de l'été. Nous n'avons aucune information précise sur la provenance, conditions de transports, ni saison de cueillette de ces avocats.

Pour chaque groupe de quatre échantillons appartenant au même avocat, nous allons calculer le minimum, le maximum, ce qui va nous donner l'étendue (=*min-max*). Nous allons déterminer la moyenne, équation (2.2), l'écart type, équation (2.3), l'erreur type, équation (2.4), le coefficient de variation, équation (2.5), et l'intervalle de confiance à 95%, équation (2.6). Cet intervalle à 95% est calculé à partir de la distribution du *t* de Student car nous

avons moins de 100 échantillons au total (Crow, Davis et Maxfield, 1960). Toutes ces valeurs sont regroupées dans un tableau (*Voir* Annexe II, tableau 2.1).

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.2}$$

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} \tag{2.3}$$

$$ErrType = \frac{\sigma}{\sqrt{n}} \tag{2.4}$$

$$CV = \frac{\sigma}{\overline{x}} \times 100 \ (\%) \tag{2.5}$$

$$CI_{\alpha} = \overline{x} \pm \left(ErrType \times t_{\frac{\alpha}{2},n-1} \right) \tag{2.6}$$

L'écart type est une mesure de dispersion d'une variable. L'erreur type est l'estimation de l'écart type de la différence entre les valeurs mesurées et les valeurs réelles. Le coefficient de variation est une mesure de dispersion des observations. Ce coefficient est sans unité, et généralement exprimé en pourcentage, il permet de comparer facilement la dispersion de groupes de variables. L'intervalle de confiance est un intervalle autour de la moyenne des échantillons qui contient la vraie valeur du paramètre. Cet intervalle est estimé à l'aide d'un coefficient α qui représente le risque d'erreur. 5% (soit 95% de confiance) d'erreur correspond à $\alpha = 0.05$. Dans notre cas, nous avons 4 échantillons par avocat, nous obtenons le coefficient correspondant issu de la table *t* de Student :

$$\boxed{t_{\frac{0.05}{2},3} = 3.1824}$$

Nous pouvons alors calculer et observer (*Voir* Figure 2.2), les intervalles de confiance à 95% de nos échantillons. Les échantillons sont regroupés par avocat, et sont identifiés par les

étoiles. Les ronds représentent les moyennes, et les intervalles dont délimités par les segments de droite.

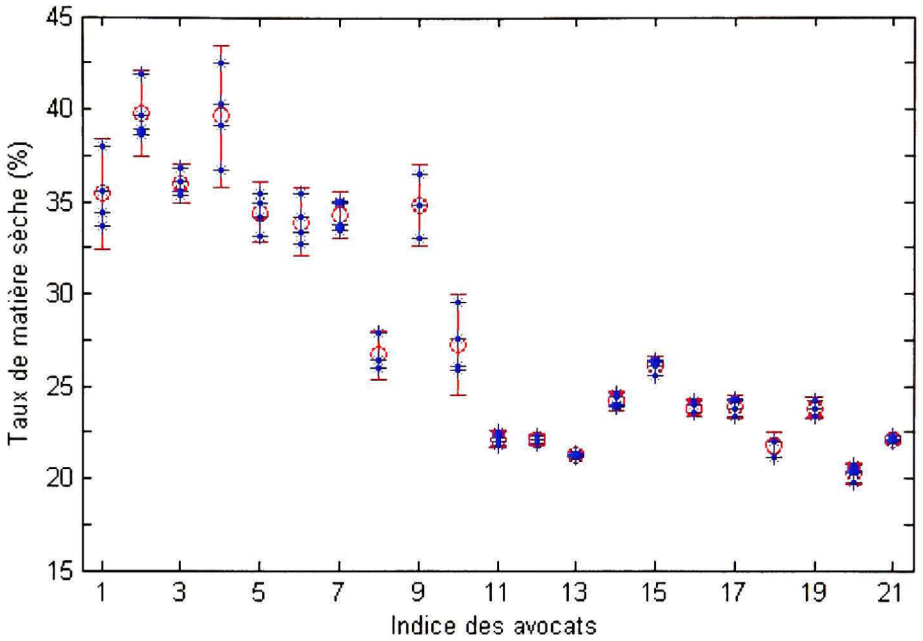

Figure 2.2 ***Intervalles de confiance des échantillons à 95%.***

Nous avons trois avocats dont les échantillons ont une erreur type supérieure à 5% : l'avocat numéro 1, 4, et 10. Nous pouvons voir que ce sont les avocats avec la plus grande étendue (respectivement 4.31, 5.79, et 3.74). Cela donne les trois plus grands intervalles de confiance. Mis à part ces trois fruits, les coefficients de variations ne dépassent pas 4%. A l'aide de ces outils statistiques, nous pouvons faire une analyse critique de notre échantillonnage (nombre de mesures intra-fruits, ici 4), et de nos répétitions (nombre de mesures inter-fruits, ici 21).

Suivant le calcul de nos coefficients de variation, nous pouvons établir si ceux-ci sont acceptables, ou s'ils ne le sont pas. À l'aide d'un expert nous avons établi que concernant l'échantillonnage, un coefficient de variation viable est de 5% ou moins; et concernant les répétitions, 10% ou moins serait acceptable. C'est ce que nous voulons idéalement pour nos données. Avec notre base de données, nous avons un coefficient de variation de 2.55% pour

l'échantillonnage, et 23.36% pour les répétitions. Alors que ce coefficient est adéquat pour l'échantillonnage, nous avons une valeur plus de deux fois supérieure pour nos répétitions.

Avec la formule du coefficient de variation, si nous connaissons la valeur que nous voulons, et en utilisant la moyenne actuelle de notre base de données, nous pouvons calculer l'écart type voulu :

$$\sigma_{voulu} = \frac{\overline{x}_{obtenu} \times CV_{voulu}}{100} \tag{2.7}$$

À partir de cet écart type voulu et de l'écart type obtenu actuellement avec notre base de données, nous sommes en mesure de déterminer combien d'échantillons ou de répétitions nous devons mesurer lors d'une prochaine expérience, et suivant une certaine valeur de confiance α :

$$N = \left(t_{\frac{\alpha}{2}, N_{actuel}-1} \times \frac{\sigma_{obtenu}}{\sigma_{voulu}} \right)^2 \tag{2.8}$$

Dans notre cas, la moyenne est de 28.29%, pour l'échantillonnage, nous avons :

$$\boxed{\sigma_{voulu}^{ech} = \frac{28.28 \times 5}{100} = 1.41}$$

Donc pour 4 échantillons par avocats, à 95% de confiance :

$$N^{ech} = \left(t_{\frac{0.05}{2}, 3} \times \frac{\sigma_{obtenu}}{\sigma_{voulu}} \right)^2 = \left(3.1824 \times \frac{0.79}{1.41} \right)^2 \tag{2.9}$$

$$\boxed{N^{ech} = 3.19 \Leftrightarrow N^{ech} = 4}$$

Concernant le nombre de répétitions, nous avons, la moyenne est toujours égale à 28.28%, et l'écart type est égal à 6.61 :

$$\boxed{\sigma_{voulu}^{rep} = \frac{28.28 \times 10}{100} = 2.83}$$

Donc pour 21 répétitions, à 95% de confiance :

$$N^{rep} = \left(t_{\frac{0.05}{2},20} \times \frac{\sigma_{obtenu}}{\sigma_{voulu}} \right)^2 = \left(2.09 \times \frac{6.61}{2.83} \right)^2 \tag{2.10}$$

$$\boxed{N^{rep} = 23.84 \Leftrightarrow N^{rep} = 24}$$

Nous constatons, que le nombre d'échantillonnage (intra-fruits) est correct, alors qu'il faudrait quelques fruits en plus (nombre de répétition) lors d'une prochaine expérience.

Maintenant que nous avons notre base de données de matière sèche, nous allons voir plus en détail comment nous avons obtenu notre base de données des spectres correspondants.

2.3 Imagerie hyperspectrale

2.3.1 Mise en contexte

Un système d'imagerie hyperspectrale se compose d'un système optique décomposant la lumière (spectrographe), d'une caméra, d'un système d'éclairage, ainsi que d'un ordinateur doté d'une carte d'acquisition et des logiciels adéquats. Le spectrographe et la caméra montés ensembles, forment ce que l'on appelle un imageur hyperspectral. Dans cette section, nous aborderons les différents aspects du système d'imagerie hyperspectrale : comment nous avons conçu notre montage. Quel type d'éclairage nous avons choisi, et pourquoi. Ensuite,

quelle configuration du matériel nous avons retenue. Enfin, nous aborderons l'interprétation et la transformation (autrement dit étalonnage) des données brutes collectées.

2.3.2 Caractérisation de la lumière

2.3.2.1 Problématique et types d'éclairage

Dans tout système de vision, l'éclairage est un aspect à prendre en compte convenablement. C'est en partie grâce à l'éclairage que nous mettrons en évidence les éléments pertinents d'une scène. En analyse hyperspectrale l'éclairage est à considérer de manière très sérieuse. Le fait d'avoir un équipement permettant de décomposer la lumière très précisément implique qu'un effort particulier doit être effectué sur l'éclairage, cela afin d'avoir une réponse maximale du matériel.

Une source lumineuse est une source émettrice de radiations (par exemple une bougie, une lampe, ou le soleil). Cette source peut être caractérisée numériquement par une distribution spectrale en fonction de la puissance relative qui lui est propre. Pour chaque source lumineuse, les longueurs d'ondes nous informent où l'énergie lumineuse est présente, et la puissance relative nous indique la quantité d'énergie présente. La Commission Internationale de l'Énergie (CIE) a codifié la distribution spectrale de différents types de sources lumineuses et les a appelés « illuminant ». Un illuminant est une représentation (longueur d'onde en fonction de l'énergie) de la qualité d'un type de source lumineuse (HunterLab, 2008). Les trois principaux illuminants sont le D65 : lumière du jour (6504°K), le A : source lumineuse incandescente tungstène (2856°K), et le F2 : lampe fluorescente commune (4100°K). La température donnée entre parenthèses est un indicatif de la couleur de la lumière. Cette température, donnée en degrés K, varie de 1700°K (couleur orangée issue de la flamme d'une bougie), à 25000°K (couleur du ciel très bleu). Généralement, en dessous de 4000°K, la couleur est qualifiée de chaude, et au dessus de 4000°K, la couleur est dite froide (Cornwell, 2005). Nous pouvons observer la répartition de l'énergie fournie par les 3 principaux illuminants, ainsi que celle d'une diode électroluminescente blanche (DEL, environ 7000°K), à la Figure 2.3.

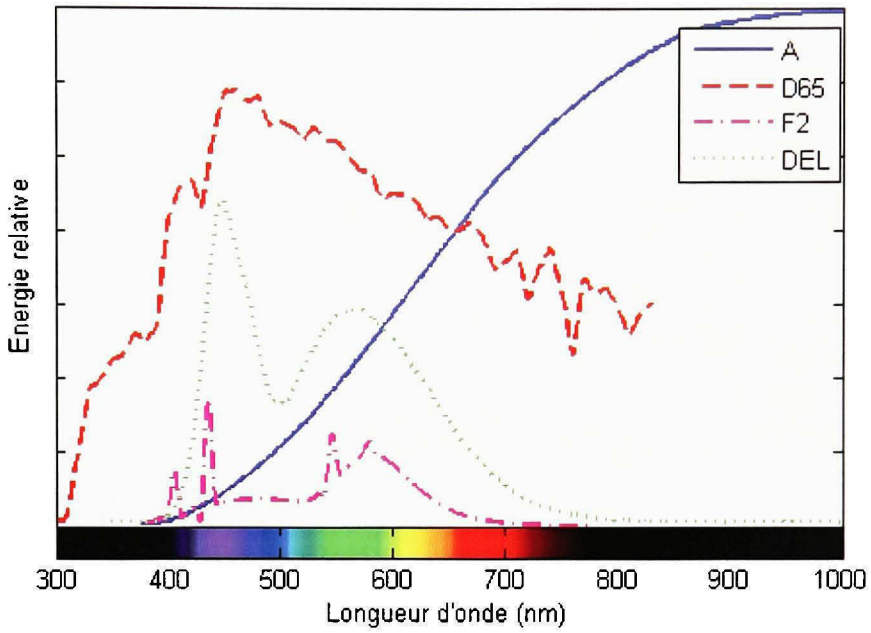

Figure 2.3 ***Énergie relative de 4 illuminants en fonction de leur longueur d'onde.***

(Tiré de Alessi et al., 2004)

Source : Cette figure sous forme de graphique est tirée des données disponibles en ligne de tables colorimétriques provenant du rapport de M. P. J. Alessi et al., *CIE Technical Report. Colorimetry*.

Si le spectrographe en lui même décompose le spectre lumineux entre 400 et 1000nm, la caméra a cependant une réponse non linéaire sur cette plage. Nous pouvons observer la réponse spectrale de la caméra sur la plage spectrale de sensibilité du spectrographe à la Figure 2.4. Cette réponse est évaluée en efficacité quantique, soit le nombre d'électrons qui parviennent à la cellule photosensible.

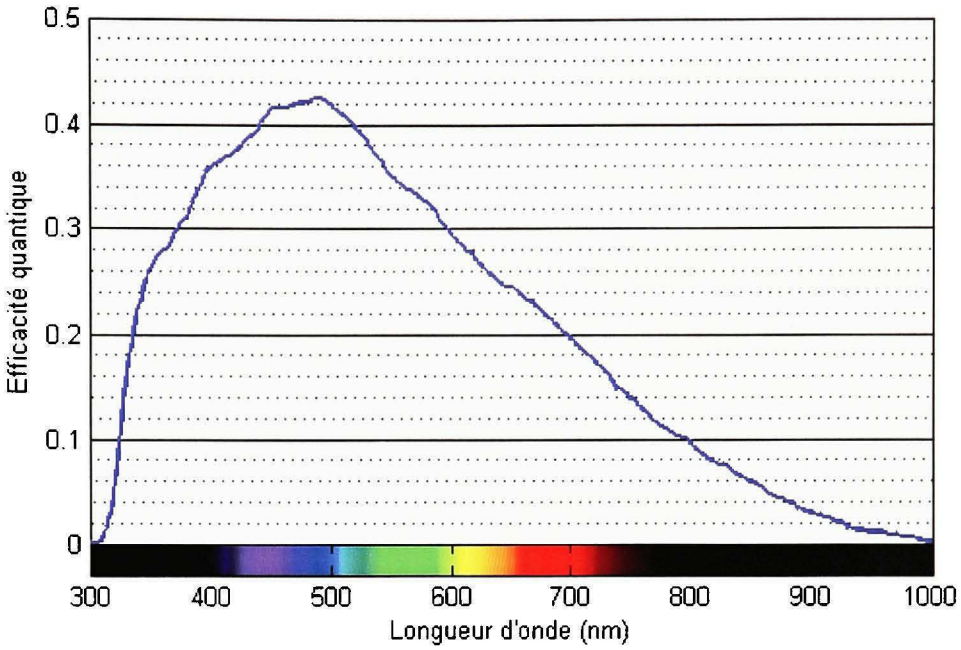

Figure 2.4 ***Réponse spectrale de la caméra.***

(Tiré de Imperx, 2006)

Source : Cette figure sous forme de graphique est issue de la documentation technique de notre caméra (modèle IPX-2M30), fabriquée par la société Imperx.

Nous pouvons remarquer qu'à partir de 800nm, la réponse commence à être faible (inférieure à 10%). Cela implique qu'il nous faut un type d'éclairage dont la puissance dans le proche infrarouge est importante de manière à compenser cette réponse.

À la vue de la Figure 2.3 et de l'analyse de la Figure 2.4, l'illuminant A (source incandescente Tungstène-halogène) semble être parmi ces courbes d'éclairage, celui qui répond le mieux à nos spécifications. En effet, sa répartition est plutôt monotone, avec sa puissance répartie de 400nm à plus de 1000nm. Par opposition, les éclairages fluorescents (F2) et à base de diodes électroluminescentes (DEL), présentent quant à eux une répartition peu uniforme où nous pouvons dénoter des discontinuités. De plus, à partir de 650nm leur puissance devient négligeable : nous en déduisons que ces types d'éclairage ne répondent pas à nos spécifications. Nous pouvons aussi observer la répartition de la lumière du jour (D65),

en notant que cette répartition semble être aussi adaptée que l'illuminant A à notre problème : répartition uniforme, et presque plus constante. Il n'est cependant pas concevable de retenir cela comme éclairage : nos expériences ne peuvent pas être conduites en plein air, les jours de beau temps uniquement... En notant que les éclairages dits « lumière du jour », n'ont rien à voir avec la répartition spectrale du D65 vue précédemment.

Les lumières de type Tungstène Halogènes sont habituellement choisies pour des études dans le visible et le proche infrarouge pour les mesures de spectres. Les sources de type halogène fournissent un éclairage très stable, et ont une durée de vie importante. L'avantage de ce type d'éclairage est que le spectre émis est régulier, monotone et sans discontinuité.

2.3.2.2 Réflexion spéculaire et montage final

Les avocats ont une peau plutôt lisse et sujette aux réflexions spéculaires lorsqu'ils sont éclairés directement par une source lumineuse.

Figure 2.5 ***Aspect spéculaire de la peau des avocats.***

Nous pouvons observer la spécularité inhérente à la peau des avocats à la Figure 2.5. L'éclairage utilisé est une source tungstène-halogène (Schott-Fostec DCR-III EKE) munie d'un module annulaire. Cette spécularité se répercute par des zones blanches sur l'image (observées au sommet des fruits)

Les réflexions spéculaires sont problématiques en vision par ordinateur, car leurs réponses ne sont pas utilisables. En spectroscopie, une réflexion spéculaire se traduit souvent par une réponse saturant le capteur. Afin de palier ce problème, nous devons éclairer les avocats de manière indirecte. Nous avons donc construit une boite de diffusion autour de la zone d'acquisition. Le matériel utilisé pour diffuser la lumière est un filtre de diffusion en papier (*Voir* Figure 2.6, les numéros de légende de cette figure sont décrits plus bas, à la suite de la Figure 2.7).

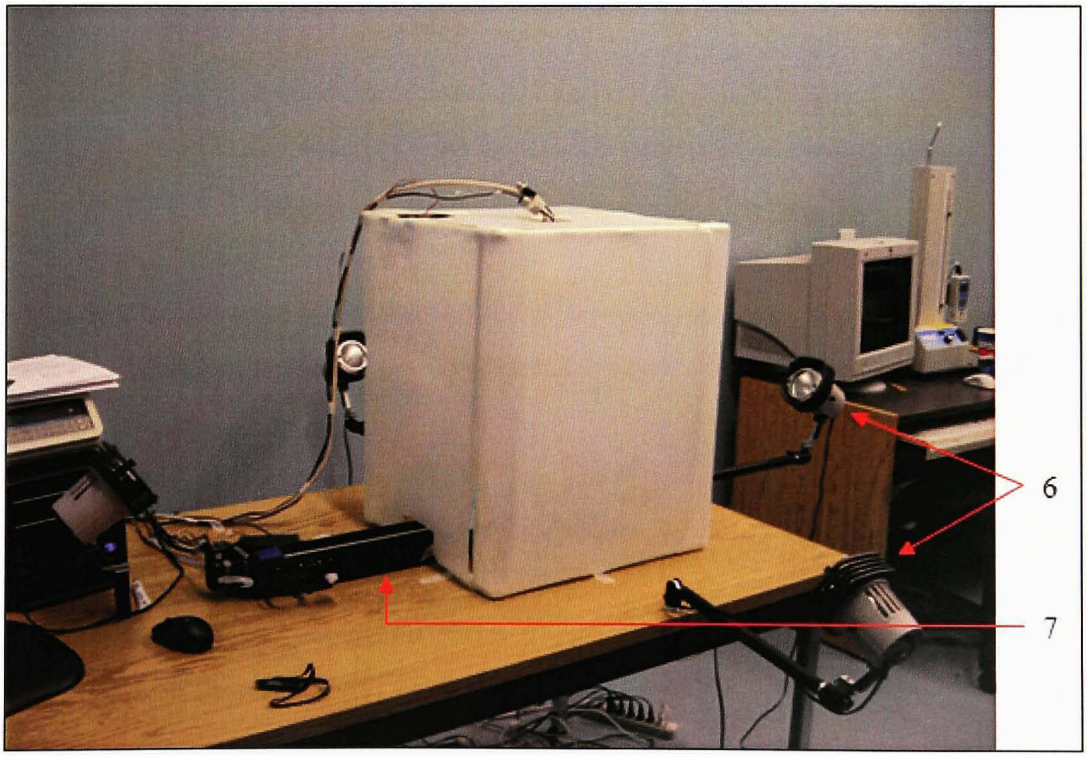

Figure 2.6 ***Boite de diffusion.***

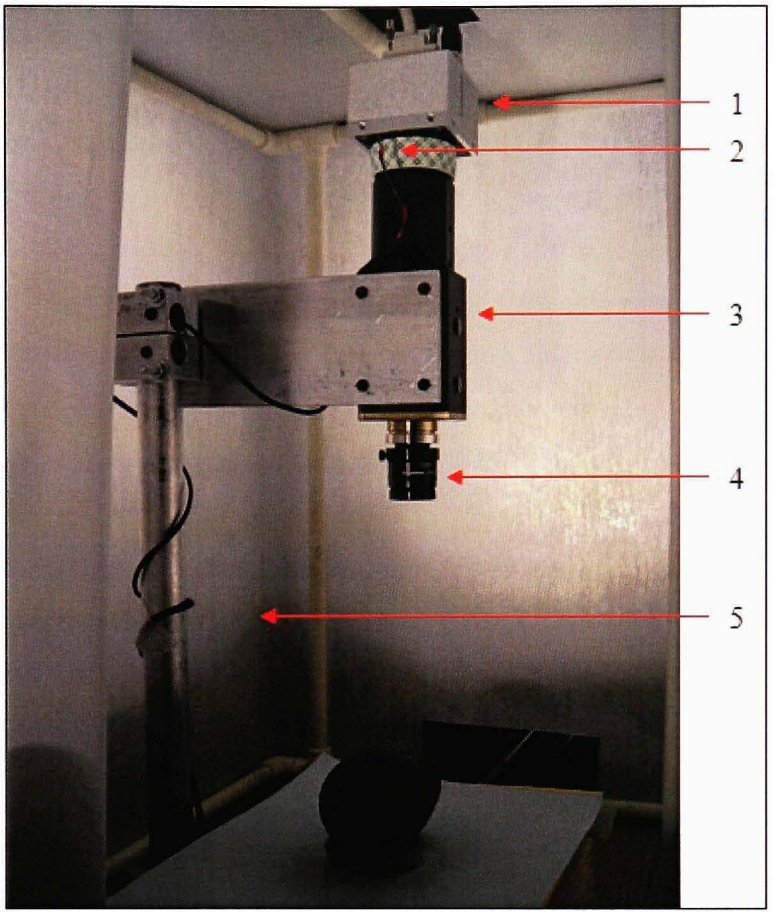

Figure 2.7 ***Intérieur de la boite de diffusion.***

Une vue de l'intérieur de la boite de diffusion est présentée à la Figure 2.7. Nous pouvons résumer les différentes parties de notre système d'imagerie hyperspectrale comme suit :

1. Caméra CCD (Imperx IPX-2M30);
2. Sonde de température (thermocouple + afficheur);
3. Spectrographe (Imspector V10E);
4. Lentille (Schneider Kreuznah 23mm achromatique);
5. Filtre de diffusion Lee N°261;
6. Source lumineuse Ambico V-100, ampoules 300W, (4 spots);
7. Rail linéaire;

Nous surveillons la température à l'intérieur de la boite de diffusion. En effet, les 4 projecteurs halogènes étant assez puissants (300W), cela fait augmenter la chaleur dans la boite. Aussi bien le spectrographe que la caméra possèdent des plages de températures de fonctionnement spécifiques. De 5°C à 40°C pour le spectrographe, et de -5°C à 50°C concernant la caméra. Il convient que la température à l'intérieure de la boite ne dépasse donc pas 40°C. Nous avons placé un ventilateur en extraction au sommet de la boite afin d'avoir une circulation d'air et ainsi maintenir une température presque constante et inférieure à 40°C. La température moyenne observée pendant les acquisitions est de 30°C±3°C. Nous devons attendre un certain temps entre la mise sous tension du système (imageur et spots), et la première acquisition. En effet, à la mise en marche de ceux-ci, la température des équipements se situe autour de 23°C, pour atteindre 30°C au bout d'une heure. Durant ce laps de temps, la précision et la linéarité de l'imageur ne sont pas constantes. Nous sommes donc obligés d'attendre que la température se stabilise et que l'imageur atteigne une température stable.

2.3.3 Étalonnage des données spectrales

L'imageur fonctionne en mode réflectance : il capte la lumière réfléchie par les objets de la scène observée. Cependant, les données brutes issues d'un capteur CCD ne sont pas directement les valeurs de réflectance. Ces données représentent le compte de l'intensité lumineuse reçue par le capteur. Nous pouvons voir à la Figure 2.8 les données brutes typiques acquises pour une référence blanche et une acquisition où nous avons bloqué l'accès à la lumière de l'imageur. Théoriquement, la référence blanche est constituée d'un matériel dont la réflectance à la lumière est connue et proche de 100% sur la plage spectrale de notre instrument. La structure et la variation du signal reçu, qui n'est donc pas constant (courbe rouge), sont dues aux caractéristiques spectrales du type d'éclairage ainsi qu'à la sensibilité spectrale du capteur utilisé. L'étalonnage de nos données brutes nous permettra de transformer les décomptes d'intensité bruts, en considérant la structure des données issues de nos références (l'absence de lumière et la référence blanche), en valeurs de réflectance ou d'absorbance ($\log(1/\textit{réflectance})$). Ces deux références nous aident ainsi à compenser pour

les non-uniformités de l'imageur : chaque pixel ayant sa propre réponse aux photons incidents.

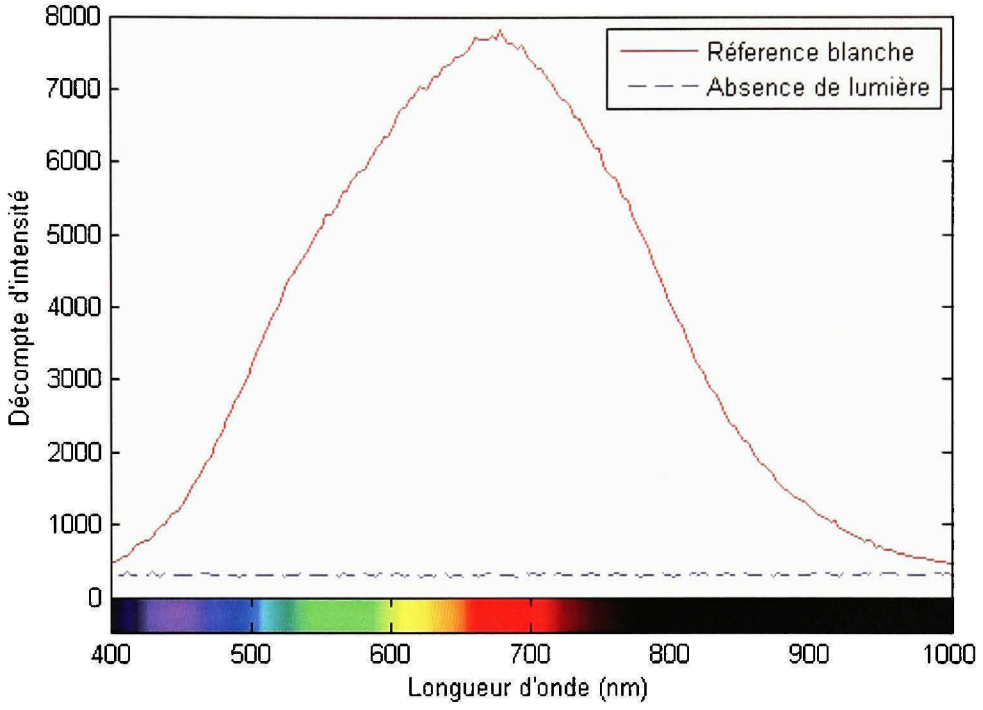

Figure 2.8 ***Spectre de la Référence blanche ainsi que des courants sombres.***

Les données brutes acquises par un imageur lorsqu'il n'y a pas d'éclairage ne sont en fait pas nulles (courbe bleue). Théoriquement elles devraient l'être, mais en fait ces décomptes d'intensité, appelés courants sombres (*dark currents*), sont dus à l'énergie thermique de la structure composant le silicium du capteur. Des électrons sont créés au fils du temps indépendamment de la lumière incidente. Ces électrons sont alors capturés, ce qui induit une réponse non nulle du capteur. Ces courants sont fonction du temps d'exposition ainsi que de la température du capteur : plus ce dernier est chaud, plus ces courants sont importants. Pour palier ce problème, il faut refroidir la caméra. Ainsi les fluctuations dues à l'agitation thermique sont diminuées.

Les données brutes des références vont nous informer sur l'interprétation du blanc et du noir par le système. Cela va nous permettre de corriger la colorisation issue de l'illumination, de l'optique, du spectrographe, et enfin, de la caméra.

Un cube hyperspectral présente les données en 3 dimensions : une dimension spectrale (λ) et deux dimensions spatiales (x, y). La dimension spatiale x est définie par les données issues de la lumière passant par la fente de l'imageur. La dimension spectrale est le spectre associé à chaque pixel de la dimension spatiale. Ces deux premières dimensions forment une image correspondant à l'acquisition d'une ligne de la scène. La troisième dimension y, est la mise côte-à-côte de ces images (lorsque l'on déplace l'objet sur le rail linéaire). Ainsi chaque valeur brute du cube est repéré par $m(x, y, \lambda)$.

L'étalonnage des données, afin d'obtenir des valeurs de réflectance (ou absorbance), est réalisé en corrigeant les données de notre cube à l'aide des données de réflectance théoriques de la référence blanche (R_{blanc}), des données brutes de la référence blanche (*blanc*), ainsi que des données brutes issues l'absence de lumière (*noir*). L'étalonnage doit se faire ligne par ligne, donc suivant la dimension spatiale y. Pour cela nous devons avoir à notre disposition une image spatiale/spectrale (x, λ), contenant les données brutes pour chaque référence. Nous réalisons l'acquisition de 30 images successives de la référence blanche, et de 30 images successives en absence de lumière. Ces deux cubes sont moyennés suivant la dimension spatiale y, de manière à avoir deux références chacune à deux dimensions (une spatiale, et une spectrale). C'est avec ces deux images que nous calculons le gain de notre système:

$$Gain(x, \lambda) = \frac{R_{blanc}(\lambda)}{blanc(x, \lambda) - noir(x, \lambda)} \tag{2.11}$$

À partir de ce gain, nous effectuerons l'étalonnage à proprement parler des données brutes issues de notre échantillon, nous rajoutons ici la deuxième dimension spatiale, y :

$$R(x,y,\lambda)=\left[m(x,y,\lambda)-noir(x,\lambda)\right]\times Gain(x,\lambda) \tag{2.12}$$

Si la réflectance de la référence blanche, R_{blanc}, n'est pas connue, nous pouvons assumer qu'elle est égale à un. Nous pouvons mettre (2.12) sous la forme plus commune suivante :

$$R(x,y,\lambda)=\frac{m(x,y,\lambda)-noir(x,\lambda)}{blanc(x,\lambda)-noir(x,\lambda)} \tag{2.13}$$

La référence est alors considérée comme idéale. Nous pouvons observer l'allure de la réflectance de la référence blanche (que nous décrirons en détail un peu plus loin) sur la Figure 2.9. Nous constatons que les valeurs ne sont pas tout à fait égales à un et ne sont pas non plus constantes. Il est donc important des les prendre en compte d'autant plus que l'instrument que nous possédons est de haute précision.

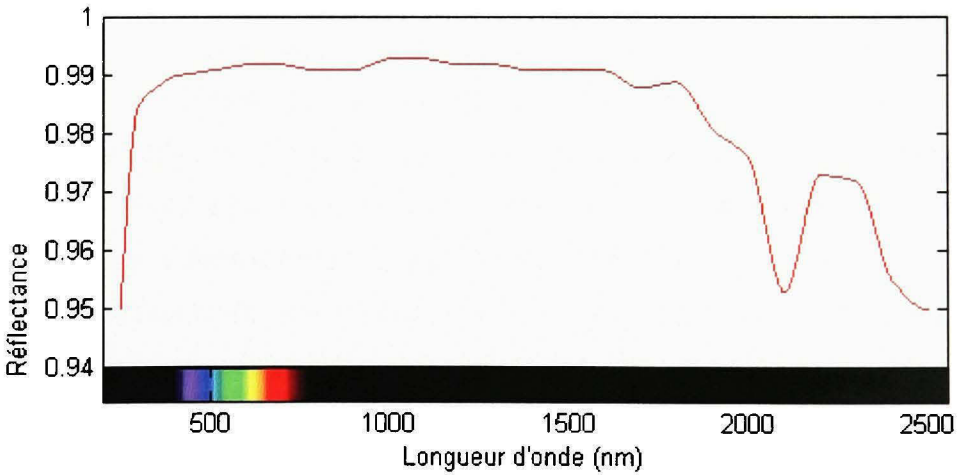

Figure 2.9 ***Réflectance de la référence blanche.***

(Tiré de Channel-Systems, 2007)

Source : Cette figure sous forme de graphique est issue des données de réflectance de la référence blanche que nous utilisons. Ces données ont été fournies par Channel Systems, la société qui nous a fourni le produit.

D'un point de vu plus mathématique, les équations (2.12) et (2.13) transforment les mesures effectuées par l'instrument en valeur de réflectance. Ces valeurs sont sans unité et comprises entre zéro et un. Cette transformation est aussi appelée étalonnage à un point. Nous pouvons mettre (2.12) sous la forme :

$$R(x,y,t) = -\frac{noir(x,\lambda)\times Gain(x,\lambda)}{blanc(x,\lambda)-noir(x,\lambda)} + \frac{Gain(x,\lambda)}{blanc(x,\lambda)-noir(x,\lambda)}\times m(x,y,\lambda) \qquad (2.14)$$

Si nous faisons abstraction des dimensions (pour une lecture plus aisée), (2.14) équivaut à :

$$R = b_0 + b_1 \times m \qquad (2.15)$$

$$b_0 = \frac{noir \times Gain}{blanc - noir}$$

$$b_1 = \frac{Gain}{blanc - noir}$$

L'équation (2.15) à une forme linéaire nécessitant seulement deux spectres bruts pour l'étalonnage : un reflétant la réflexion maximale du système (référence blanche), et un représentant la réflexion minimale du système (absence de lumière). Les modèles calculés ainsi couvrent l'entière plage de valeurs de réflectance (proche de zéro à presque un), mais ne donne aucune information sur la linéarité du système entre ces deux valeurs extrêmes. Pour ce faire, il faut modifier l'équation (2.15), et utiliser des cibles d'étalonnage blanc dont les réflectances sont graduellement comprises entre zéro et un (Burger et Geladi, 2005). Ainsi, nous pouvons avoir une meilleure évaluation de la linéarité de l'instrument suivant la réflectance. Cependant, nous ne possédons pas de telles cibles. Nous nous contenterons de l'étalonnage à un point vu à l'équation (2.14).

La tuile de calibration pour la référence blanche que nous avons utilisée est une tuile de 30cmx2.5cmx0.5cm faite de Polytétrafluoroéthylène (PFTE). Ce polymère fluoré est un

matériau doux, facilement déformable, blanc mat dont la réflectance est proche de un. Autrement connu sous le nom de Spectralon (*Voir* Figure 2.10), cette référence est considérée comme très bonne et offre une réflexion diffuse adéquate pour servir de standard. Ce matériau est ainsi utilisé comme référence dans le monde entier (Springsteen, 1999).

Figure 2.10 ***Tuile de calibration Spectralon.***

Comme nous l'avons vu précédemment, nous échantillonnons six avocats par session. Six avocats représentent 24 spectres à acquérir et étalonner. Quand l'instrument a atteint une température constante (une heure), nous réalisons l'acquisition du spectre brut de la référence blanche. Ce sont ces données qui nous serviront pour étalonner tous les échantillons acquis par la suite. L'acquisition du spectre brut en absence de lumière se fait quant à elle avant chaque avocat. En effet, les données brutes issues de cette acquisition sont beaucoup plus sujettes aux fluctuations. Étant donné que nous ne refroidissons pas notre caméra, il est préférable de prendre plusieurs spectres bruts en absence de lumière tout au long de la session.

Nous pouvons observer (*Voir* Figure 2.11) les données brutes recueillies lors de l'acquisition du spectre d'un avocat. Nous montrons ici les spectres provenant du pixel central de l'image.

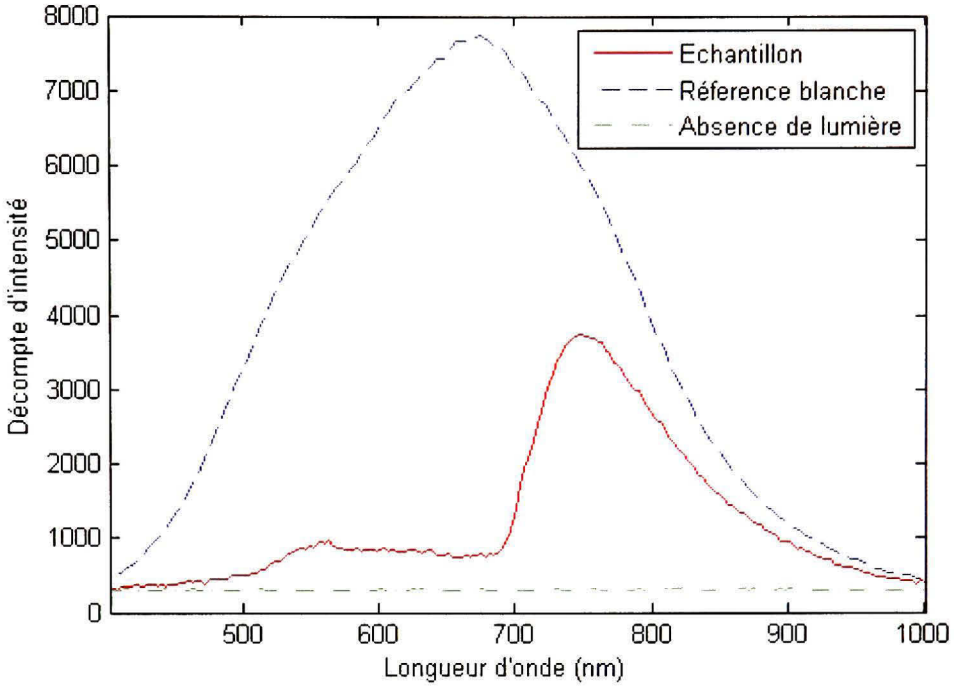

Figure 2.11 ***Données brutes provenant de l'acquisition des spectres pour l'étalonnage.***

C'est à partir de ces données que nous étalonnerons le spectre afin d'obtenir des images de réflectance, en appliquant l'équation (2.12). Nous obtenons le spectre de la Figure 2.12.

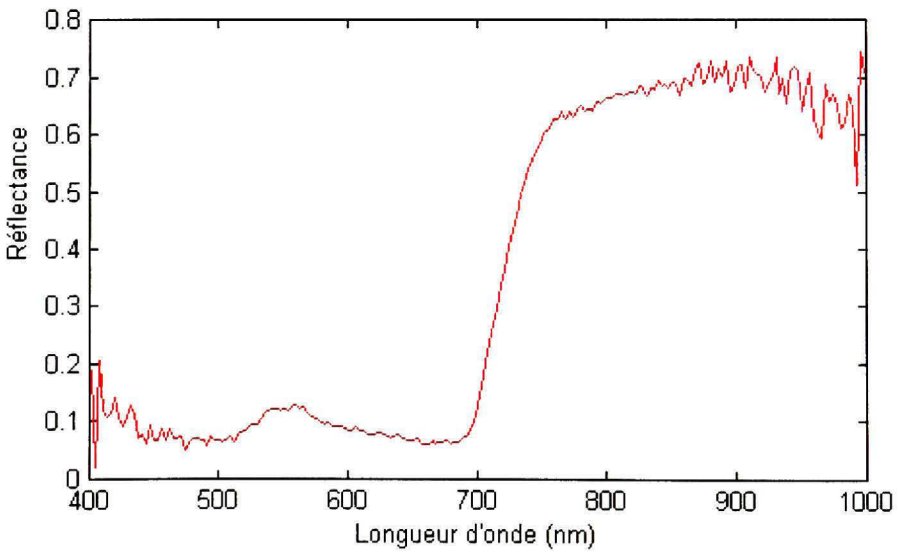

Figure 2.12 ***Valeurs de réflectance obtenues après étalonnage.***

Les spectres ainsi enregistrés sont à la résolution maximale de l'imageur, soit 2.94nm. Cela équivaut à 204 bandes spectrales. Nous pouvons immédiatement remarquer que du bruit non négligeable est présent au début du spectre (400nm-450nm) et à la fin (900nm-1000nm). Cela est dû notamment à la faible réponse de la caméra dans ces zones. Nous avons tenté de palier ce problème en utilisant un éclairage adapté. Ce n'est peut-être pas suffisant. Nous pouvons alors influer sur la résolution spectrale, éventuellement diminuer celle-ci, de manière à diminuer le bruit présent dans le spectre.

2.3.4 Résolution spectrale

Le spectrographe possède une résolution spectrale de 2.8nm au maximum, cela correspond à la partition minimale du spectre que l'optique du V10E peut distinguer. Afin de comprendre le concept de résolution spectrale, il nous faut saisir comment est captée la lumière par la caméra.

La caméra possède une résolution native de 1600x1200 pixels dont nous combinons les pixels quatre à quatre, pour obtenir une résolution de 800x600. Nous avons donc 800 pixels suivant la dimension spatiale, et 600 suivant la dimension spectrale : L'optique du V10E offre une résolution spectrale de 2.8nm, cela correspond à 215 bandes spectrales théoriques. Nous réduisons de manière matérielle la résolution de la caméra dans le but d'avoir en sortie une valeur d'intensité correspondant à une bande spectrale. Après cette combinaison matérielle, nous en effectuerons une autre au travers du logiciel d'acquisition. Lorsque la lumière est décomposée et arrive au capteur, les pixels suivant la dimension spectrale ne reçoivent pas tous la lumière incidente. En fait, tel qu'est assemblé le système, les 600 pixels correspondent à une plage spectrale de 275nm à 1195nm. Bien évidement, ni l'optique du spectrographe (400nm à 1000nm), ni le capteur ne peuvent décomposer la lumière provenant des extrémités de cette plage spectrale. 400nm correspond au pixel numéro 92 suivant la dimension spectrale, et 1000nm correspond au pixel 500. Cela donne une plage de 408 pixels, ce qui, pour une plage spectrale de 600nm donne 1.47nm par pixel. Le spectrographe

ayant une résolution native de 2.8nm, il nous faut donc combiner 2 pixels successifs, afin d'obtenir une résolution de 2.94nm par pixel. Il en résulte 204 bandes spectrales distinctes.

De la même manière, suivant la dimension spatiale, nous ne recueillons pas tout à fait l'information suivant les 800 pixels. Nous avons une plage active de 784 pixels. Lorsque nous utilisons l'instrument pour avoir la plus fine résolution, nous réalisons l'acquisition d'image de 784x204. A cela nous ajoutons la dimension spatiale y. Pour un avocat, il faut environ 350 images mises cote à cote pour couvrir le fruit. Cela représente des fichiers de 150Mo à 170Mo.

Nous pouvons influer sur la manière dont nous combinons les pixels successifs de la dimension spectrale. La combinaison de plusieurs pixels successifs fonctionne comme un accumulateur, c'est-à-dire que le résultat de cette combinaison est la somme pure et simple des valeurs brutes de compte d'intensité donnée par les pixels du capteur. Initialement, les pixels sont combinés deux par deux. Pour diminuer la résolution spectrale, nous pouvons les combiner trois par trois ou plus. Nous pouvons aussi décider de les combiner non linéairement, c'est-à-dire par groupes non homogènes, tout au long de la plage spectrale.

Si nous considérons le problème de bruit présent aux extrémités du spectre en réflectance (*Voir* Figure 2.12), nous pouvons attribuer le bruit au fait qu'à ces endroits (*Voir* Figure 2.11), les données brutes de l'échantillon sont proches des données brutes de la référence blanche et des données brutes en absence d'éclairage. En effet, réécrivons l'équation d'étalonnage (2.12) comme suit :

$$R(x,y,t)=\frac{m(x,y,t)-noir(x,y)}{blanc(x,y)-noir(x,y)}\times G(x,y) \tag{2.16}$$

Aux extrémités, le dénominateur et le numérateur tendent vers de petites valeurs. Les spectres bruts sont proches les uns des autres. L'influence du bruit induit par les valeurs des références et de l'échantillon tend donc à être amplifiée. Afin d'amenuiser le bruit il faudrait

être en mesure d'augmenter l'écart existant entre le spectre brute de l'échantillon et les deux spectres brutes des références. C'est une chose que nous pouvons faire en combinant plusieurs pixels successifs. En effet, cette combinaison agissant comme un accumulateur, en accumulant les bandes nous accumulons aussi les différences entre les courbes.

Ainsi, nous avons combiné les pixels huit par huit en début et fin de spectre, donnant une résolution spectrale de 11.76nm, tout en laissant intacte la combinaison deux par deux des autres pixels. Dans la région 400nm-450nm nous avons maintenant 5 bandes spectrales là ou nous en avions 22 auparavant, et concernant la plage 900nm-1000nm, nous avons 7 bandes spectrales contre 31 initiales. Nous pouvons voir les données brutes des spectres ainsi acquis à la Figure 2.13.

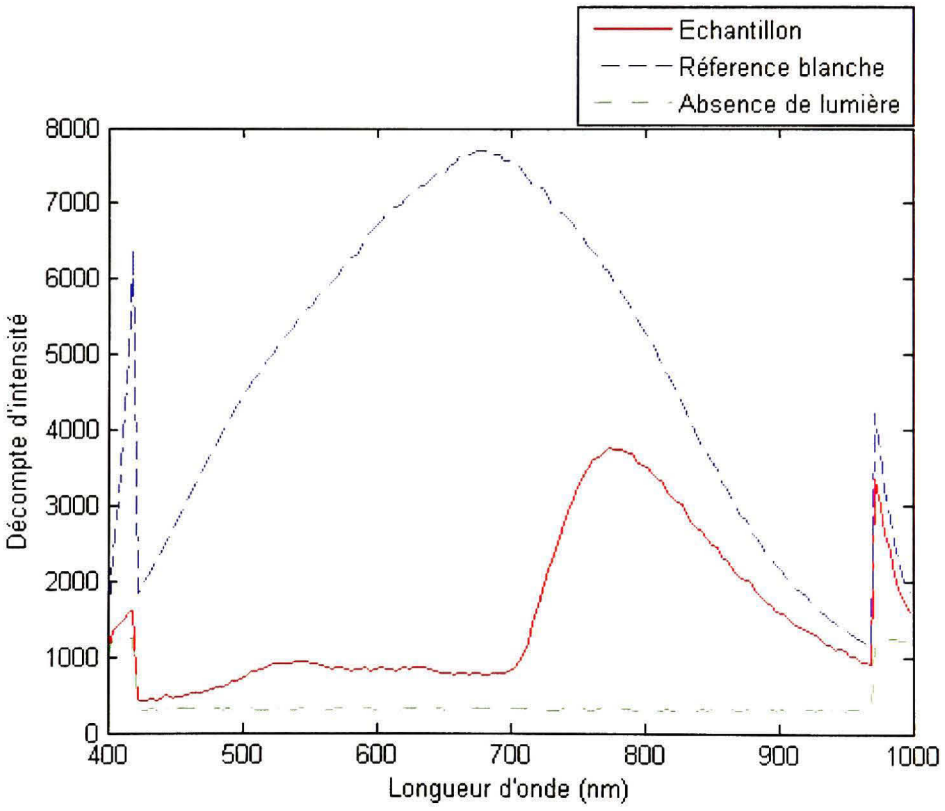

Figure 2.13 ***Spectres bruts obtenus après la combinaison de pixels aux extrémités.***

Les discontinuités observées à 450nm et 900nm marquent le passage de l'accumulation de huit pixels à l'accumulation de deux pixels, et inversement. Nous remarquons aussi que cette accumulation a comme prévu permis d'augmenter les différences entre les spectres bruts de l'échantillon et des références. Par définition, le fait de les regrouper non linéairement n'influe pas sur le calcul d'étalonnage, en notant que tous les spectres doivent être acquis avec la même configuration d'accumulation.

À la Figure 2.14, nous pouvons observer deux spectres en réflectance de deux combinaisons distinctes des pixels du capteur. La combinaison initiale deux à deux (à gauche), et la combinaison non linéaire citée précédemment (à droite).

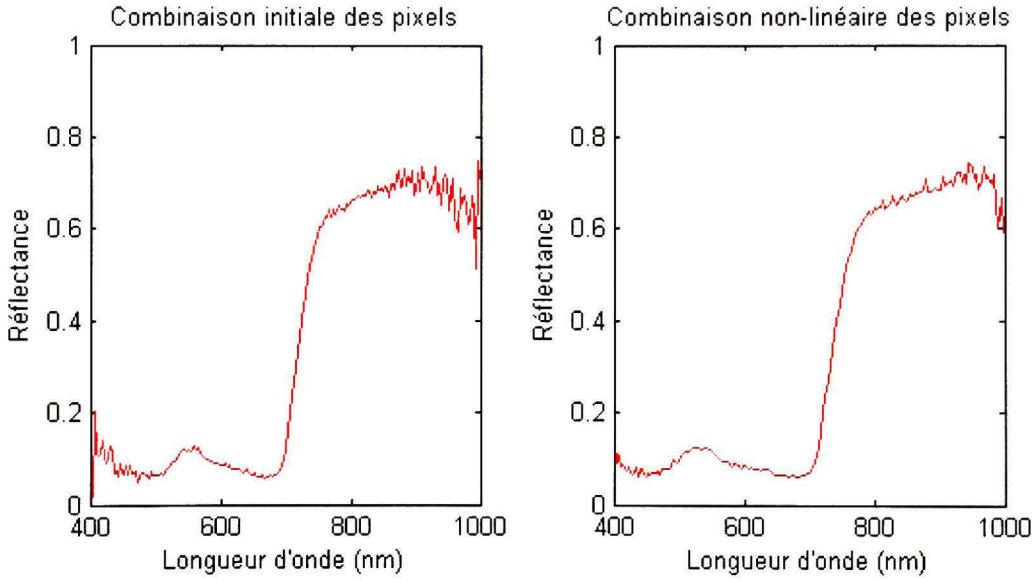

Figure 2.14 ***Comparaison de valeurs de réflectance pour deux combinaisons de pixels.***

Nous pouvons observer l'influence du bruit sur les données spectrales étalonnées à la Figure 2.15. Le bruit est plus présent aux extrémités (204 bandes spectrales), et se manifeste par des fluctuations locales importantes des valeurs de réflectance.

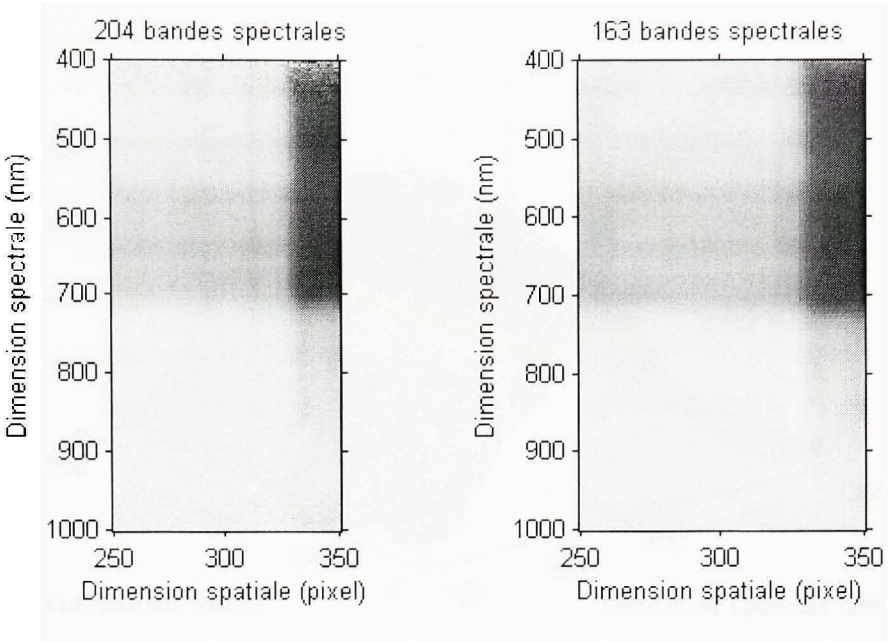

Figure 2.15 ***Bruit présent dans les spectres étalonnés.***

Nous remarquons que la combinaison non-linéaire nous a permis de réduire fortement le bruit présent aux extrémités du spectre. Son inconvénient est de perdre en résolution spectrale : de 204 bandes initiales, nous avons maintenant 163 bandes.

2.3.5 Dimension spatiale

En raison de l'optique de la lentille, il s'avère que la lumière n'arrive pas uniformément suivant la ligne de la dimension spatiale.

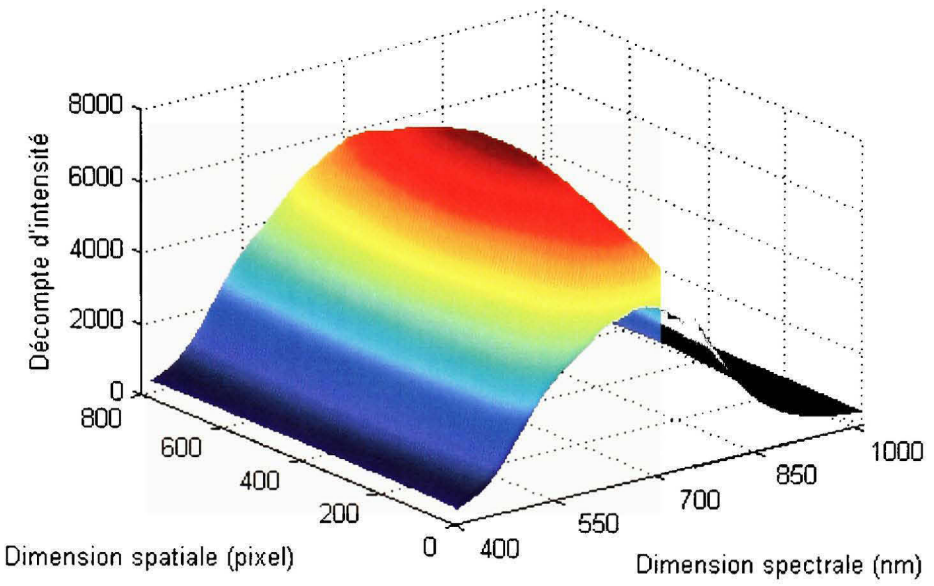

Figure 2.16 ***Répartition de l'intensité lumineuse d'une ligne.***

Comme nous pouvons l'observer à la Figure 2.16, qui présente les données brutes de la référence blanche, l'intensité lumineuse reçue n'est pas également répartie, celle-ci est plus importante au centre que sur les extrémités.

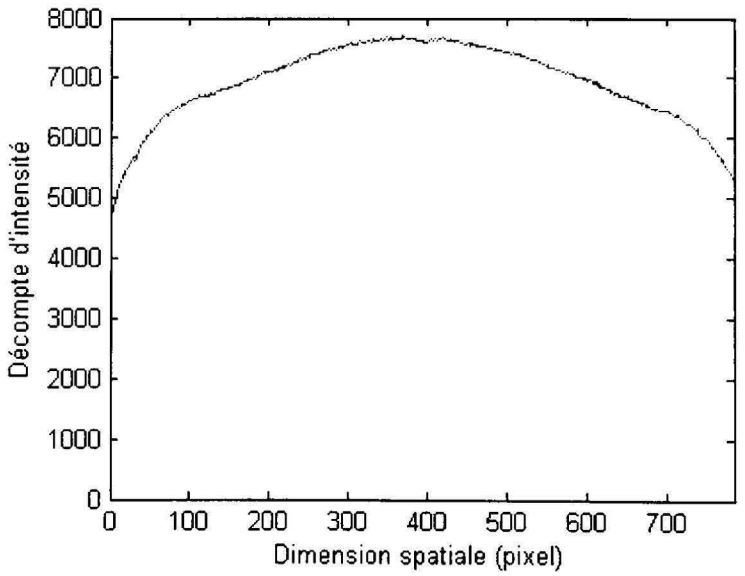

Figure 2.17 ***Intensité lumineuse de la bande 684nm.***

À la Figure 2.17, nous pouvons voir la répartition spatiale de l'intensité lumineuse, correspondant à la bande 684nm (bande correspondant au maximum de compte d'intensité), suivant la dimension spatiale de l'imageur. Nous pouvons remarquer que nous avons des décomptes d'intensité variant de 4700 à 7700, cela équivaut à une variation de 3000 entre le maximum et le minimum soit près de 43%, cela n'est pas négligeable. Pour palier cela, nous garderons l'avocat au centre de l'image, au pixel 400±150 pixels. Nous aurons donc une variation des décomptes de 400, soit une variation inférieure à 6%, ce qui est nettement plus acceptable.

2.3.6 Logiciel d'acquisition

Le logiciel dont nous nous servons pour l'acquisition des cubes hyperspectraux est SpectralCube v2.72, fournit avec le matériel par la société Autovision, dont nous pouvons voir une capture à la Figure 2.18.

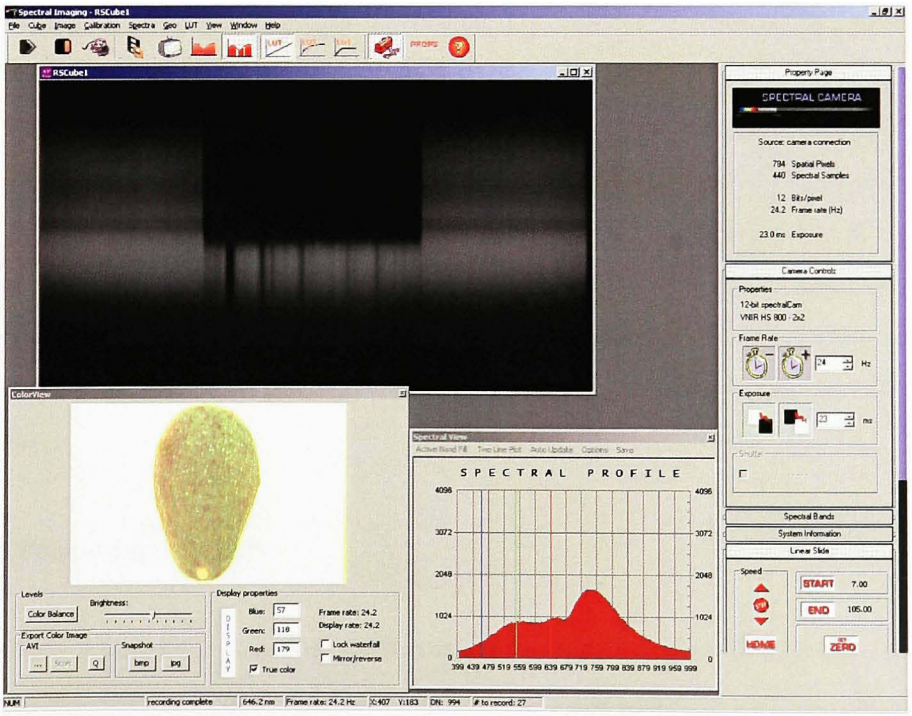

Figure 2.18 ***Interface du logiciel SpectralCube.***

Le logiciel nous permet de voir le profil spectral brute d'un pixel de la ligne (fenêtre en bas à droite), une image d'intensité (spatiale/spectrale) correspondant à une ligne de la scène (fenêtre en tons de gris en haut à gauche), et une image RGB (spatiale/spatiale) de la scène acquise (fenêtre en bas à gauche ou nous voyons l'avocat). Sur la droite nous avons les commandes du rail linéaire ainsi que les différentes options de configuration du logiciel.

2.3.7 Prétraitements

Dans cette section, nous verrons comment nous avons obtenu les spectres qui nous servirons pour la détermination du taux de matière sèche des avocats. Ces étapes consistent à calibrer les données hyperspectrales, segmenter l'avocat dans le cube, éventuellement segmenter les défauts présents sur la peau du fruit, et isoler notre région d'intérêt. Ces étapes sont réalisées avec le logiciel Matlab R2007b, et ENVI 4.2.

Premièrement, comme nous venons de le voir, l'avocat n'occupe pas toute la place du cube de données recueillies. Afin d'alléger le volume des données traitées, nous procéderons à la suppression de la zone inutilisée autour du fruit. Ce découpage est fait avec une macro sous le programme ENVI. Ce logiciel permet d'ouvrir un cube spectral sans pour autant charger en mémoire toutes les bandes spectrales, mais seulement 3, que nous définissons comme les 3 canaux RGB à des fins de visualisation. De cette manière, le chargement de seulement 3 bandes est nettement plus rapide que celui de 163 bandes. De plus, nous découpons les mêmes zones sur les fichiers de calibration, pour ne retenir des cubes initiaux, seulement la zone couvrant l'avocat. Nous passons ainsi de fichiers de 160Mo à environ 60Mo occupés sur le disque.

2.3.7.1 Étalonnage du cube

Nous pouvons maintenant étalonner notre cube hyperspectral, avec la référence blanche et la référence en absence de lumière correspondante, suivant la méthode vue précédemment (*Voir* 2.3.3). Nous avons donc à notre disposition le cube des valeurs de réflectance de l'avocat.

Nous pouvons observer à la Figure 2.20 une image RGB recomposée à partir des valeurs de réflectance du spectre d'un avocat. Le détail de la méthode pour passer du spectre aux composantes RGB est décrit à l'Annexe III.

Figure 2.19 ***Image RGB recomposée à partir du spectre d'un avocat.***

2.3.7.2 Segmentation de l'avocat

Afin de ne garder que les données spectrales de l'avocat, il faut segmenter celui-ci du fond de l'image. Nous réalisons un seuillage sur la bande 20 (504nm), soit celle, après inspection, qui présentait une grande différence entre le fond et le fruit. Nous réalisons l'extraction d'une ligne verticale au centre de l'image. En calculant la dérivée première de cette tranche, nous sommes en mesure de voir les discontinuités (passage de la zone du fond à l'avocat, et inversement), sachant que l'espace entre les deux discontinuités représente l'avocat. À partir de ces valeurs de réflectance de l'avocat, et du calcul de la valeur moyenne, nous pouvons seuiller cette bande pour avoir une image binaire avec des un sous l'avocat, et des zéros partout ailleurs. À partir de cette image binaire, nous pouvons ajuster la découpe des bords de l'avocat, ainsi que remplacer les valeurs de réflectance du fond par des zéros. Nous pouvons voir le résultat à la Figure 2.20.

Figure 2.20 ***Image d'un avocat isolé du fond.***

Comme nous pouvons le voir, la peau de l'avocat est généralement imparfaitement conservée : il arrive fréquemment que pendant le transport certains chocs se produisent créant plaies et cicatrices. Cela se traduit par de petites surfaces foncées. Il peut être intéressant d'éliminer ces surfaces, afin que les spectres de celles-ci n'interférent pas dans notre étude. Nous avons pu remarquer, après inspection des différentes bandes spectrales, que de telles cicatrices pouvaient être facilement segmentales sur les bandes 74 (663nm) et 94 (722nm). Nous remarquons aisément les taches claires (bande 74) et les taches foncées (bande 94) présentes dans la Figure 2.21. Le seuillage sur chaque bande est réalisé en regardant la moyenne des valeurs de réflectance de l'avocat. Ne sont gardés que les pixels étant inférieurs (respectivement supérieurs) à la moyenne plus 1.5 fois l'écart type (respectivement la moyenne moins 1.5 fois l'écart type). Nous réalisons un ET logique entre les deux masques ainsi calculés pour obtenir un masque que nous appliquons successivement à chaque bande de notre cube hyperspectral afin d'éliminer les cicatrices.

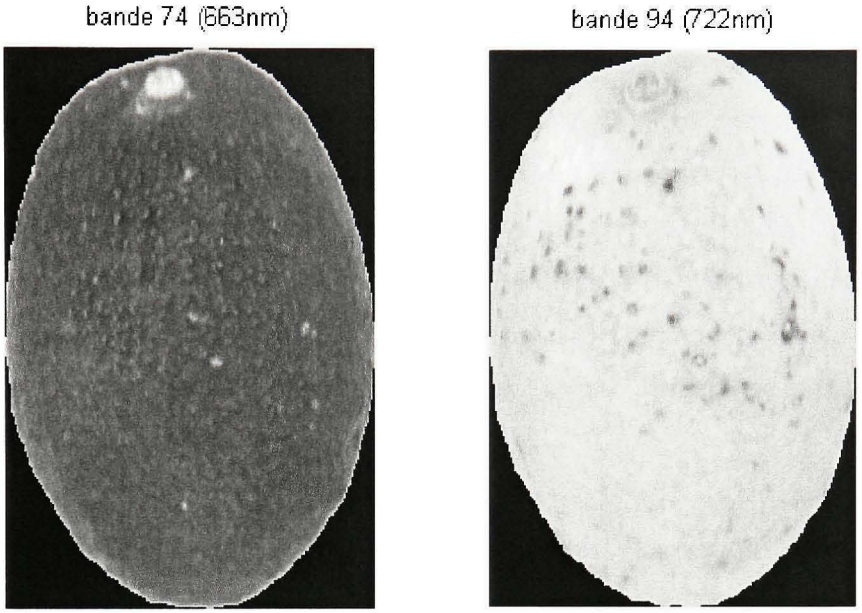

Figure 2.21 ***Cicatrices présentes dans 2 bandes spectrales.***

Nous pouvons observer à la Figure 2.22 le résultat d'un tel traitement. Nous pouvons voir que nous avons segmenté la majeure partie des cicatrices de l'avocat.

Figure 2.22 ***Avocat segmenté.***

2.3.7.3 Région d'intérêt

Pour chaque avocat, nous avons prélevé 4 échantillons de chair pour faire l'analyse de matière sèche. La répartition de la matière sèche au sein de l'avocat n'est pas constante (*Voir* Annexe I, Figure 1.1). Il nous faut donc déterminer les régions d'intérêt, dans les cubes, correspondantes aux zones d'extraction de la matière sèche. Les cubes hyperspectraux, tout comme les zones d'extraction de matière sèche ont été acquis aux quatre points cardinaux de l'avocat, selon la longueur du fruit. Sur le cube, notre région d'intérêt se situe donc suivant une zone verticale au centre du fruit.

Nous avons une valeur de matière sèche par échantillon. Nous avons donc besoin d'un spectre correspondant. Tout comme la valeur de matière sèche a été prise suivant une zone verticale au centre du fruit, nous réalisons la moyenne suivant les bandes spectrales des valeurs de réflectance d'une zone verticale située au même endroit que l'extraction de matière sèche. De cette manière, nous avons un spectre composé de 163 bandes spectrales, représentant toute la zone. Ainsi, à ce spectre, correspond une valeur de matière sèche. Nous pouvons observer à la Figure 2.23 la région d'intérêt de l'avocat.

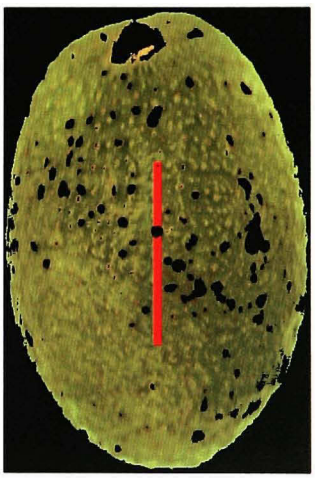

Figure 2.23 ***Région d'extraction du spectre sur l'avocat.***

La zone rouge correspond à la région d'extraction et de moyenne du spectre. Nous pouvons observer à la Figure 2.24 les spectres en mode réflectance et en mode absorbance ($\log(1/reflectance)$), ainsi extraits d'un avocat.

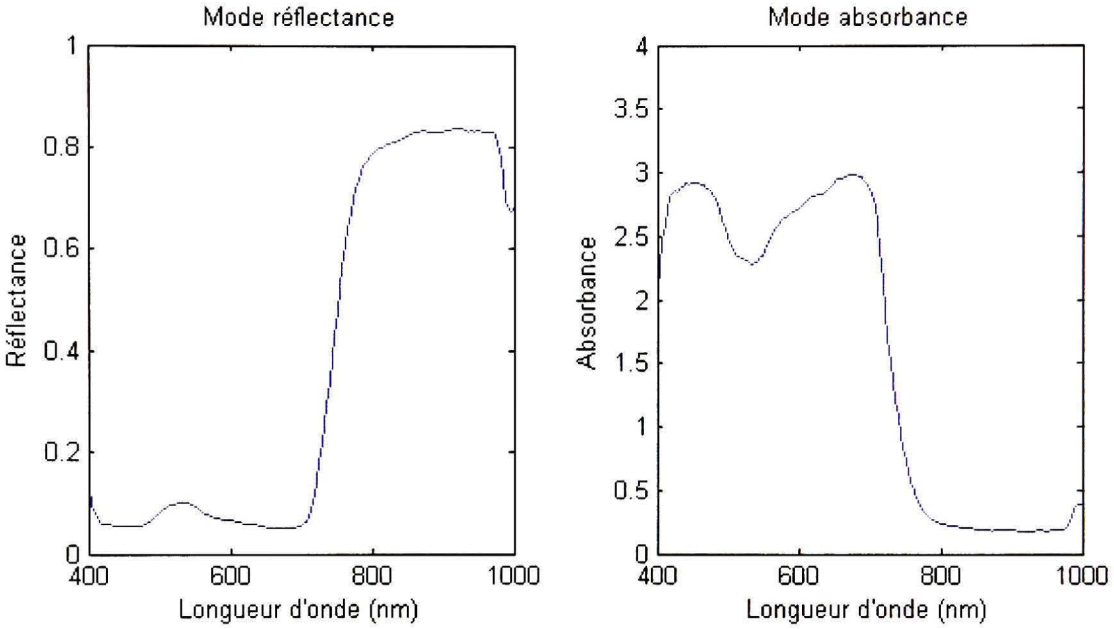

Figure 2.24 ***Spectre extrait d'un avocat en mode réflectance et absorbance.***

Muni du spectre et de la valeur de matière sèche correspondante, nous tenterons de découvrir la corrélation éventuelle qui existe entre ces deux mesures. S'il existe effectivement une corrélation, il serait judicieux de déterminer quelles bandes spectrales sont les plus pertinentes lors de la détermination de la matière sèche des avocats par imagerie hyperspectrale. Nous utiliserons un outil mathématique créé pour cela : la régression, aidée de diverses approches pour la sélection de variables.

2.4 La régression PLS

La régression, est l'outil que nous utiliserons pour tenter de trouver le lien existant entre les données spectrales et les mesures de matière sèche de l'avocat. Nous faisons l'hypothèse que

ce lien est de type linéaire : nous supposons que la matière sèche est la combinaison linéaire des n valeurs de réflectance ou d'absorbance, λ_i, du spectre. Dans notre cas, $n = 163$.

$$\%DM = \sum_{i=1}^{163} \beta_i \lambda_i \tag{2.17}$$

L'objectif est donc de déterminer les coefficients β_i du vecteur de régression.

Pour la construction d'un modèle prédictif, nous distinguons deux opérations distinctes :

1. L'étalonnage (ou modélisation) : à l'aide des mesures faites sur les variables explicatives et expliquées, nous déterminerons le modèle, c'est-à-dire que nous évaluerons les paramètres de celui-ci, et déterminerons le vecteur de régression.

2. La prédiction/test : nous déduisons les variables de sortie à l'aide du modèle et des variables explicatives.

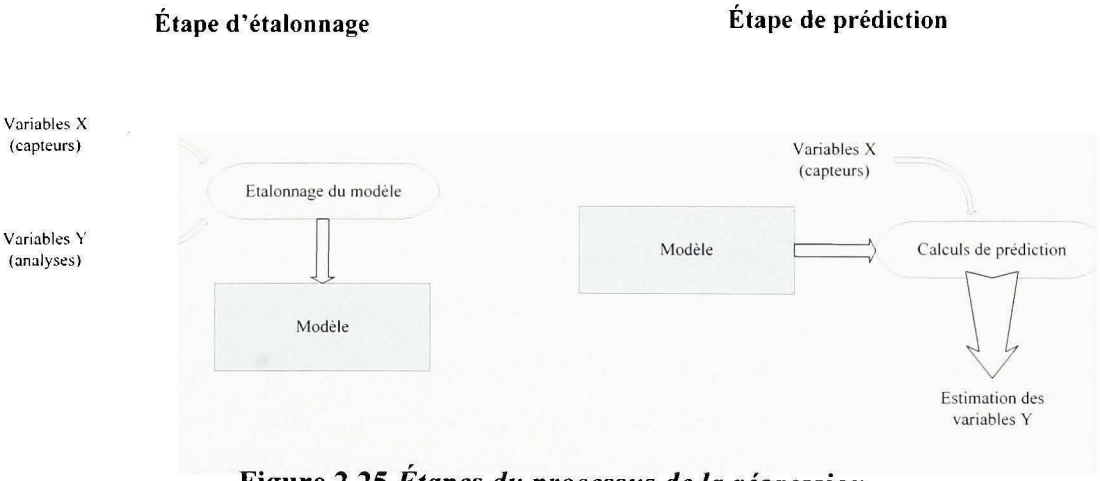

Figure 2.25 ***Étapes du processus de la régression.***

Pour satisfaire ces étapes, nous devons avoir à notre disposition un ensemble de variables de réponse, ou variables dépendantes ainsi que les variables explicatives correspondantes. Cela

forme notre base de données à partir de laquelle nous allons construire puis tester notre modèle. Évidemment nous ne pouvons pas utiliser toute la base de données pour faire simultanément la construction et le test du modèle. Dans le cas idéal, il faudrait diviser la base de données en trois parties contenant à peu près le même nombre d'échantillons (Davies, 2006). Une première partie pour la construction du modèle (ensemble d'étalonnage), une deuxième pour évaluer les paramètres du modèle (ensemble de validation), enfin une troisième partie pour tester, évaluer la performance du modèle (ensemble de test). L'ensemble de test nous permettra de comparer la performance obtenue par différents modèles prédictifs. Cet ensemble ne doit en aucun cas intervenir dans la construction des modèles. Il doit donc être indépendant. Il se peut, et nous sommes dans ce cas, que le nombre d'échantillons disponibles ne nous permette pas de séparer correctement notre base de données en trois parties également distribuées. Nous devons alors scinder la base de données en deux parties, un tiers pour la validation, deux tiers pour l'étalonnage, et utiliser l'ensemble d'étalonnage pour à la fois construire et évaluer les paramètres du modèle.

Nous possédons 84 échantillons de matière sèche couplés à 84 mesures spectrales. Lors de la division de la base de données en une base d'étalonnage (deux tiers des échantillons) et une base de test (un tiers des échantillons), nous avons 57 échantillons pour l'étalonnage du modèle et 27 pour évaluer ses performances. Nous avons pu observer à la Figure 2.2 que la répartition du taux de matière sèche n'était pas très uniforme. Nous avons 2 noyaux principaux autour de 25% et 35%. De manière à avoir des performances représentant bien nos données, et pour éliminer le facteur chance, nous ne pouvons pas faire de modélisation sur une seule et unique base de données. Nous avons donc créé 50 bases de données (étalonnage/test) distinctes ayant chacune une répartition des taux de matière sèche équivalente à la base de données contenant tous les échantillons. Nous allons donc réaliser la modélisation avec et sans la sélection de variables sur ces 50 bases de données et compiler les résultats obtenus (moyenne et écarts types des performances obtenues par les 50 modélisations). Ce sont ces moyennes et écarts types que nous analyserons.

Nous avons décidé d'utiliser le modèle PLS dans la suite de nos travaux. Lors de la détermination d'un modèle PLS, il y a un paramètre qu'il nous faut fixer. Communément appelé complexité du modèle, il s'agit du nombre de variables latentes, ou nombre de projecteurs lors de l'analyse en composante principale.

2.4.1 Évaluation de la complexité du modèle

La complexité du modèle PLS s'évalue au nombre total de variables latentes dans le modèle, dont le maximum est égal au rang de la matrice des variables explicatives (**X**). Nous voulons avoir le modèle ayant la meilleure performance pour un nombre de variables latentes minimal. Dans le cas idéal, en ayant deux ensembles d'étalonnage indépendants, nous construisons autant de modèles que l'on peut avoir de variables latentes, puis nous testons la performance de chaque modèle avec le deuxième ensemble. Nous déterminons le modèle ayant la meilleure performance, et nous en déduisons le nombre de variables latentes que doit avoir le futur modèle. Nous construisons alors un modèle PLS avec les deux ensembles d'étalonnage réunis et le nombre de variables latentes déterminé auparavant.

Avec notre base de données, il nous est impossible d'avoir 3 ensembles distincts. Nous avons retenu deux approches différentes pour la détermination du nombre de variables latentes du modèle PLS.

2.4.1.1 Validation croisée

La validation croisée est une technique pour évaluer les paramètres d'un modèle sans avoir à notre disposition un ensemble de validation à proprement parler (Davies, 1998). Comme nous n'avons pas assez d'échantillons pour avoir à la fois un ensemble pour la construction de notre modèle et un ensemble distinct pour évaluer la complexité, nous avons donc recours à la validation croisée sur notre unique ensemble d'étalonnage pour déterminer la complexité du modèle.

La validation croisée est un processus itératif reposant sur le principe suivant : nous scindons en deux parties notre base de données d'étalonnage : une partie pour la détermination du modèle (appelé sous-ensemble d'étalonnage), et une autre partie pour le test (appelé sous-ensemble de validation). Nous construisons un modèle suivant une valeur de complexité avec le sous-ensemble d'étalonnage, et effectuons un test de performance avec le sous-ensemble de validation. Nous répétons alors ces deux étapes de sorte que tous les échantillons soient utilisés une fois pour évaluer la performance du modèle. Lorsque ces deux étapes sont terminées, nous recommençons pour une autre valeur de complexité. Une fois que toutes les valeurs de complexité ont été testées, nous pouvons tracer un graphique de la performance de notre système en fonction de la complexité, et ainsi repérer la valeur pour laquelle l'erreur est minimale.

Il existe plusieurs choix à notre disposition concernant la sélection de la partition à retirer de l'ensemble d'étalonnage pour le test de performance. Nous pouvons tout d'abord retirer un seul élément à la fois et ainsi réaliser le calcul du modèle avec les *N-1* échantillons restants, puis répéter la procédure *N* fois. On appelle cette technique *Leave One Out* (LOO). Nous pouvons aussi diviser l'ensemble en υ blocs égaux, ces blocs peuvent être composés d'éléments consécutifs, on parle alors de validation croisée continue υ-blocs. Ils peuvent aussi être sélectionnés à un tout les υ échantillons, on emploie le terme anglais de validation croisée *venetian blinds* υ-blocs. Enfin les échantillons à mettre de coté peuvent aussi être sélectionnés aléatoirement. Il est à remarquer que dans le cas particulier où $\upsilon = N$, la validation croisée υ-blocs est identique à la validation croisée LOO.

Pour évaluer la performance du modèle, nous étalonnons le modèle avec le sous-ensemble détalonnage selon une valeur de complexité, soit selon un certain nombre de variables latentes. Ensuite nous déterminons l'erreur de prédiction du modèle sur les échantillons du sous-ensemble de validation. Ce calcul d'erreur est appelé la somme des erreurs de prédiction au carré (*PRediction Error Sum of Square*, PRESS). Pour chaque modèle k, $k \in [\![1, \upsilon]\!]$, l'erreur est calculée suivant:

$$PRESS_k = \sum_{i=1}^{n_2} \left(\hat{y}_{i,k} - y_{i,k} \right)^2 \tag{2.18}$$

y_i, la i^{ième} valeur mesurée.

$\hat{y}_i$, la i^{ième} valeur prédite.

n_2, le nombre d'échantillons du sous-ensemble de validation.

L'indice de performance consiste à effectuer la somme cumulative des erreurs PRESS obtenues pour les v modèles déterminés successivement :

$$cumPRESS = \sum_{k=1}^{v} PRESS_k \tag{2.19}$$

En répétant cette procédure pour plusieurs nombres de variables latentes, nous pouvons tracer un graphique de l'erreur cumPRESS en fonction du nombre de variables latentes (*Voir* Figure 2.26). Nous cherchons alors le nombre de variables latentes qui minimise cette erreur.

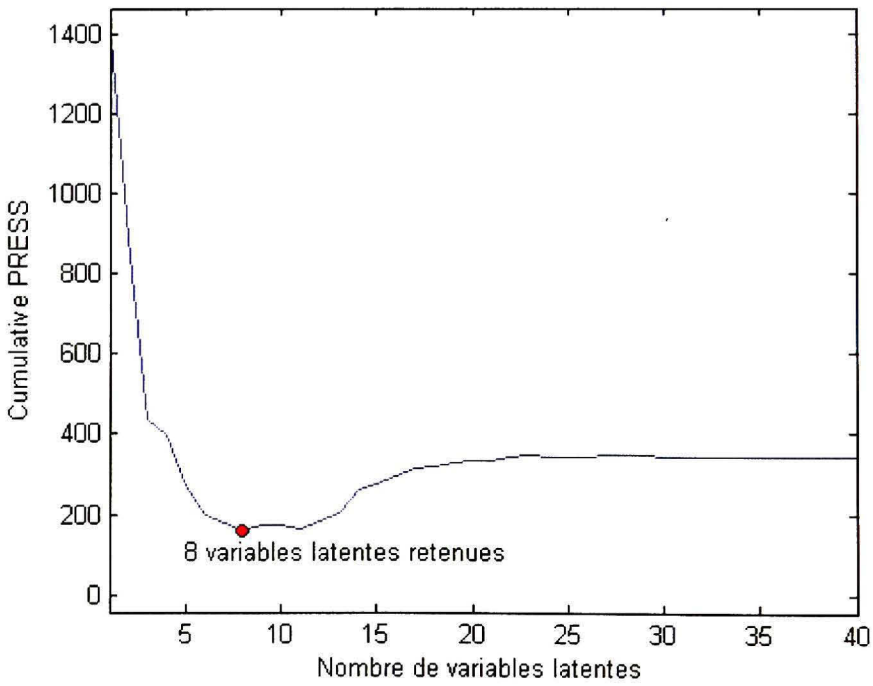

Figure 2.26 ***Erreur cumulative PRESS en fonction du nombre de variables latentes.***

Un autre calcul d'erreur peut aussi être utilisé : la racine carrée de l'erreur quadratique moyenne de validation croisée (*Root Mean Square Error of Cross Validation*, RMSECV), pour chaque modèle k, $k \in [\![1, \upsilon]\!]$, l'erreur est calculée suivant:

$$RMSECV_k = \sqrt{\frac{1}{n_2}\sum_{i=1}^{n_2}\left(\hat{y}_{i,k} - y_{i,k}\right)^2} = \sqrt{\frac{PRESS_k}{n_2}} \tag{2.20}$$

y_i, la i^{ième} valeur mesurée.
$\hat{y}_i$, la i^{ième} valeur prédite.
n_2, le nombre d'échantillons du sous-ensemble de validation.

Afin d'évaluer la performance des modèles pour une variable latente donnée, nous réalisons la moyenne des erreurs RMSECV des υ modèles.

$$RMSECV = \frac{1}{\upsilon}\sum_{k=1}^{\upsilon} RMSECV_k \tag{2.21}$$

Nous traçons alors la courbe de l'erreur RMSECV en fonction du nombre de variables latentes, et déterminons pour quel nombre de variables latentes cette erreur est minimisée.

Ces deux calculs, PRESS et RMSECV sont en fait liés. En réunissant les équations (2.20) et (2.21), nous obtenons:

$$RMSECV = \frac{1}{\upsilon\sqrt{n_2}}\sum_{k=1}^{\upsilon}\sqrt{PRESS_k} \tag{2.22}$$

L'allure de la courbe de l'erreur RMSECV est donc la même que celle de l'erreur cumPRESS. Dans notre exemple (*Voir* Figure 2.26), le modèle avec le minimum de variables latentes donnant la performance la meilleure est le modèle composé de 8 variables latentes. Nous pouvons remarquer, et c'est un comportement typique, qu'avoir des modèles plus

complexes ne semble pas être un atout, mais plutôt un inconvénient : l'erreur est plus importante. C'est ce que l'on appelle la malédiction de la dimensionnalité, menant à un sur-apprentissage, et donc engendrant une diminution des performances. De même, construire des modèles très simples (avec un nombre de variables latentes inférieur à 5) ne permet pas non plus d'avoir de bonnes performances.

2.4.1.2 Diagrammes de Pareto

Outre l'observation l'erreur commise par le modèle, il peut être intéressant d'ajouter un indice de variance. Il a été démontré (Green et Kalivas, 2002; Kalivas, 2004) qu'un indice de variance du modèle est un diagnostic important dans la régression et peut être utilisé en conjonction avec une mesure d'erreur afin d'évaluer au mieux les paramètres du modèle.

Un indicateur de variance peut être la norme euclidienne du vecteur de régression. La norme du vecteur de régression est reliée à la variance des coefficients de régression, et aux résultats de prédiction (Faber et Kowalski, 1997). Des modèles ayant une grande variance sont reconnus pour avoir de grands coefficients de régression, alors que des modèles qui ont une petite variance sont préférés et ont de petits coefficients de régression.

Afin de déterminer les paramètres d'un modèle régressif, la norme du vecteur de régression peut être évaluée en fonction d'une mesure de résidus, pour former ce que l'on appelle des diagrammes de Pareto. Nous traçons un graphique de variance en fonction des résidus pour différents paramètres du modèle. Sachant que nous voulons un modèle ayant peu de variance et de petits résidus, l'optimisation consiste donc à trouver les paramètres du modèle qui minimisent ces deux attributs.

Dans le cas de la régression PLS, le paramètre à optimiser est le nombre de variables latentes. De la même manière que pour la validation croisée, nous déterminerons et tracerons les valeurs de résidus versus la norme du vecteur de régression, cela pour plusieurs valeurs de variables latentes. La mesure de résidus utilisée dans ce cas est la racine carrée de l'erreur

quadratique sur l'ensemble de calibration (RMSEC). Nous utiliserons donc l'ensemble d'étalonnage dans son intégralité, et sans partition. Nous pouvons observer à la Figure 2.27, un exemple de diagramme de Pareto. Chaque point de ce graphique correspond à un modèle PLS construit suivant une valeur de complexité, et à partir duquel nous avons calculé l'erreur RMSEP, et la norme de son vecteur de régression. Le neuvième point en partant de la droite correspond au modèle construit avec 14 variables latentes, et est le modèle qui minimise à la fois les résidus et la variance.

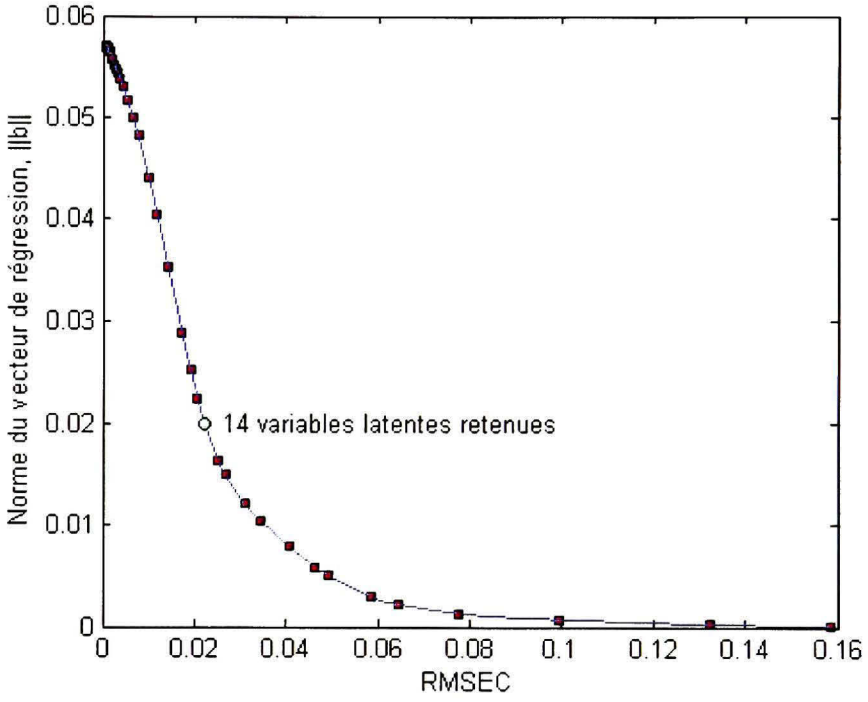

Figure 2.27 ***Diagramme de Pareto.***

2.4.2 Mesure de performance

La première mesure de performance est l'erreur (e_i) que commet notre modèle : pour le couple $\hat{y}_i$ (variable prédite par le modèle), y_i (variable de référence) :

$$e_i = \hat{y}_i - y_i \tag{2.23}$$

La moyenne des ces erreurs sur leur ensemble est appelée biais (*bias)* :

$$bias = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i) \tag{2.24}$$

A partir de cette mesure de résidus, nous pouvons définir l'erreur quadratique moyenne (*Mean Squared Error*), et sa racine carrée (*Root Mean Squared Error*) :

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{2.25}$$

$$RMSE = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{2.26}$$

Au delà de la mesure d'erreur, nous pouvons utiliser des outils statistiques nous permettant de savoir si nos données explicatives et expliquées sont bien reliées entre elles. La plus largement utilisée est le coefficient de corrélation R^2. C'est une valeur comprise entre 0.0 et 1.0 sans unité. Une valeur égale à 0.0 indique que, connaissant nos variables explicatives, nous n'arrivons pas du tout à déduire les variables expliquées. Il n'y a donc pas de relation linéaire existante entre elles. À l'inverse, une valeur de 1.0 indique que connaissant nos variables explicatives, nous en déduisons parfaitement nos variables expliquées. Afin de calculer le coefficient de détermination pour le couple de variables mesurée/prédite, $(y_i, \hat{y}_i)$, nous définissons la somme des erreurs au carré (*Sum of Squared Error*), équation (2.27), et la somme totale des carrés (*Total Sum of Squares*), équation (2.28), pour aboutir à la définition du coefficient, équation (2.29) :

$$SSE = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2.27}$$

$$SST = \sum_{i=1}^{N} \left(y_i - \overline{\mathbf{y}} \right)^2 \tag{2.28}$$

$$R^2 = 1 - \frac{SSE}{SST} \tag{2.29}$$

Afin d'évaluer les capacités de notre modèle régressif, nous avons un dernier indice de performance, l'écart type des résidus (*Standard Deviation of Residuals*), qui correspond à l'écart type des données explicatives mesurées, sur la racine carrée de l'erreur quadratique moyenne :

$$SDR = \frac{EcartType(\mathbf{y})}{RMSE} \tag{2.30}$$

Pour analyser les résultats de nos modèles prédictifs, nous retenons la racine carrée de l'erreur quadratique moyenne, le coefficient de détermination, et l'écart type des résidus. Ce sont des paramètres qui sont largement utilisés dans la littérature lors de l'analyse des performances de modèles régressifs.

2.5 Sélection de variables

La sélection de variables est une étape importante dans un processus de prédiction. En effet, toutes les variables indépendantes n'étant pas forcement pertinentes à l'explication du phénomène, il peut être judicieux d'essayer de déterminer le sous ensemble de celles qui ont plus d'impact. En plus, il peut apparaitre que diminuer le nombre de variables permet d'augmenter les performances du modèle. La sélection de variables est un problème d'optimisation, c'est-à-dire que nous essaierons différentes solutions (différentes configurations de variables sélectionnées) à notre problème, puis choisirons parmi les solutions proposées, celles qui minimisent un paramètre donné de l'optimisation. Généralement ce paramètre est une mesure de performance : l'erreur que commet le modèle.

Nous aborderons deux approches différentes de sélection de variables. Une première méthode itérative, évalue la variation de l'erreur du modèle suivant l'élimination successive des variables une à une. La deuxième méthode est stochastique, basée sur des essais de population de solutions, ainsi que leur retour plus ou moins positif sur l'optimisation.

2.5.1 Élagage PLS

L'élagage PLS est une technique itérative d'élimination de variables (Lima, Mello et Poppi, 2005). Partant de la totalité des variables, nous tenterons de diminuer ce nombre en minimisant l'augmentation d'erreur due au retrait de telle ou telle variable. C'est une technique adaptée à la régression PLS provenant de la minimisation de l'erreur de généralisation d'un réseau de neurones en optimisant ses poids. Pour un niveau de performance donné, les modèles avec le moins de paramètres sont supposés avoir de meilleures performances (Hassibi et al., 1994).

Le modèle de régression peut être exprimé sous la forme :

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \tag{2.31}$$

b, le vecteur contenant les coefficients de régression.
e, l'erreur.
X, la matrice des variables explicatives.
y, le vecteur des variables expliquées.

Nous tenterons, avec ce procédé itératif, d'éliminer les variables (les colonnes de **X**) ayant le moins de signification dans l'explication du phénomène observé. Cela concorde avec la minimisation de l'erreur commise par le modèle.

Nous pouvons considérer la prédiction $\hat{y}_i$ définie par la fonction :

$$\hat{y}_i = \mathbf{x}_i\mathbf{b} \Leftrightarrow \hat{y}_i = f(\mathbf{x}_i, \mathbf{b}) \tag{2.32}$$

$\mathbf{x}_i$, le vecteur contenant les variables explicatives de l'observation y_i.

Il est possible d'écrire la fonction d'erreur du système comme :

$$E(\mathbf{b}) = \sum \left(y - f(\mathbf{x}, \mathbf{b}) \right)^2 \Leftrightarrow E(\mathbf{b}) = \sum \left(y - \hat{y} \right)^2 \tag{2.33}$$

Avec cette reformulation, nous voyons que l'erreur est fonction des coefficients du vecteur de régression et nous allons essayer de mesurer l'impact du retrait d'une longueur d'onde (d'un coefficient du vecteur de régression), sur cette erreur. Pour ce faire, nous effectuons un calcul de saillance. Afin de minimiser cette erreur, nous pouvons la développer en une série de Taylor. Le développement en série de Taylor d'une fonction f autour du point a, est défini par :

$$f(a+h) = \sum_{n=0}^{+\infty} h^n \frac{f^{(n)}(a)}{n!} \tag{2.34}$$

Appliqué à notre problème, nous cherchons à évaluer la variation de l'erreur du modèle, $E(\mathbf{b})$, pour une petite variation des paramètres (le vecteur de régression). Nous faisons les deux hypothèses suivantes :

1. Les termes d'ordre trois et supérieurs peuvent être négligés.

2. Pour un vecteur de régression donné, le modèle calculé est optimal : nombre de variables latentes minimum pour une erreur minimum. Nous sommes donc à un minimum local d'erreur, ce qui implique que le gradient de la fonction d'erreur est nul.

L'équation (2.34) s'écrit donc sous la forme :

$$E(\mathbf{b}+\delta\mathbf{b}) = E(\mathbf{b}) + \underbrace{\frac{\partial E(\mathbf{b})}{\partial \mathbf{b}}}_{\approx 0}\delta\mathbf{b} + \frac{1}{2!}\delta\mathbf{b}^T \underbrace{\frac{\partial^2 E(\mathbf{b})}{\partial \mathbf{b}^2}}_{=\mathbf{H}}\delta\mathbf{b} \underbrace{+\cdots}_{\approx 0} \tag{2.35}$$

H est la matrice hessienne définie par:

$$\mathbf{H} = \frac{\partial^2 E(\mathbf{b})}{\partial \mathbf{b}^2} = \begin{bmatrix} \frac{\partial^2 E(\mathbf{b})}{\partial b_1^2} & \cdots & \frac{\partial^2 E(\mathbf{b})}{\partial b_1 \partial b_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 E(\mathbf{b})}{\partial b_n \partial b_1} & \cdots & \frac{\partial^2 E(\mathbf{b})}{\partial b_n^2} \end{bmatrix}$$

Posons $\Delta E(\mathbf{b}) = E(\mathbf{b}+\delta\mathbf{b}) - E(\mathbf{b})$, l'équation (2.35) se réécrit :

$$\Delta E(\mathbf{b}) = \frac{1}{2}\delta\mathbf{b}^T\mathbf{H}\delta\mathbf{b} \tag{2.36}$$

Nous regarderons la variation de la fonction d'erreur lorsque nous mettrons un des coefficients b_i du vecteur de régression à zéro. Mettre un coefficient du vecteur de régression à zéro revient à éliminer ce coefficient du modèle. Cela s'exprime sous la forme :

$$b_i + \delta b_i = 0 \Leftrightarrow b_i + \mathbf{e}_i^T \cdot \delta\mathbf{b} = 0 \tag{2.37}$$

$\mathbf{e}_i^T$ est un vecteur avec un 1 à la position *i* et des 0 partout ailleurs tel que $\mathbf{e}_i^T \cdot \delta\mathbf{b} = \delta b_i$

Nous devons résoudre :

$$\min_i \left\{ \min_{\delta\mathbf{b}} \left\{ \frac{1}{2}\delta\mathbf{b}^T\mathbf{H}\delta\mathbf{b} \mid b_i + \mathbf{e}_i^T \cdot \delta\mathbf{b} = 0 \right\} \right\} \tag{2.38}$$

Ce genre de problème de minimisation d'une fonction peut être résolu par la méthode des multiplicateurs de Lagrange. Nous réécrivons l'équation (2.38) avec λ le multiplicateur de Lagrange associé à la condition de minimisation de l'équation (2.37) :

$$L = \frac{1}{2}\delta\mathbf{b}^T\mathbf{H}\delta\mathbf{b} + \lambda\left(b_i + \mathbf{e}_i^T \cdot \delta\mathbf{b}\right)_i = 0 \tag{2.39}$$

En dérivant la fonction de Lagrange, et appliquant la contrainte de l'équation (2.37), nous trouvons que le changement de paramètre ainsi que le changement dans l'erreur du modèle sont :

$$\delta\mathbf{b} = -\frac{b_i}{\left[\mathbf{H}^{-1}\right]_{i,i}}\mathbf{H}^{-1} \cdot \mathbf{e}_i \tag{2.40}$$

$$\boxed{L_i = \frac{1}{2}\frac{b_i^2}{\left[\mathbf{H}^{-1}\right]_{i,i}}} \tag{2.41}$$

L_i est appelée saillance du coefficient b_i. C'est-à-dire l'augmentation de l'erreur due au retrait de ce coefficient.

Voici les étapes de la procédure permettant de réaliser l'élagage :

(1) Calcul du modèle PLS avec *m* (initialement le nombre maximal) variables;

(2) Détermination du coefficient b_i à éliminer en fonction du calcul de saillance, et élimination de la variable *i* de **X**;

(3) Calcul d'un nouveau modèle PLS avec *m-1* variables, cela implique de nouveaux coefficients de régression, et une nouvelle erreur;

(4) Exécution des étapes 2 et 3 jusqu'à ce qu'il ne reste plus qu'une variable (soit une colonne) dans **X**;

À chaque itération, la complexité du modèle calculé est évaluée par validation croisée, et le calcul de la racine carrée de l'erreur quadratique correspondante (RMSECV) est effectué. Dans l'article (Lima, Mello et Poppi, 2005), les auteurs utilisent un ensemble distinct leur permettant, à chaque itération, de calculer l'erreur que commet le modèle en fonction du nombre de variables. Nous n'avons pas assez de données pour posséder un ensemble de validation distinct. Nous utiliserons donc l'erreur de validation croisée pour évaluer les performances de notre modèle en fonction du nombre de variables latentes retenues. Nous relevons la racine carrée de l'erreur quadratique de validation croisée. Nous pouvons tracer le graphique de cette erreur en fonction du nombre de variables utilisées dans le modèle. Nous choisissons alors le nombre de variables qui correspond au minimum d'erreur. Cependant, il peut arriver que le minimum global de notre graphique ne soit pas la solution optimale à notre problème. Les solutions correspondantes à une erreur légèrement supérieure, à l'erreur minimum globale peuvent être de meilleures candidates à notre problème. Nous considérons une augmentation de l'erreur inférieure 5% (Osten, 1988). En effet, il peut arriver que les solutions ainsi sélectionnées aient jusqu'à plus de trois fois moins de variables que la solution correspondant au minimum global (*Voir* Figure 1.1).

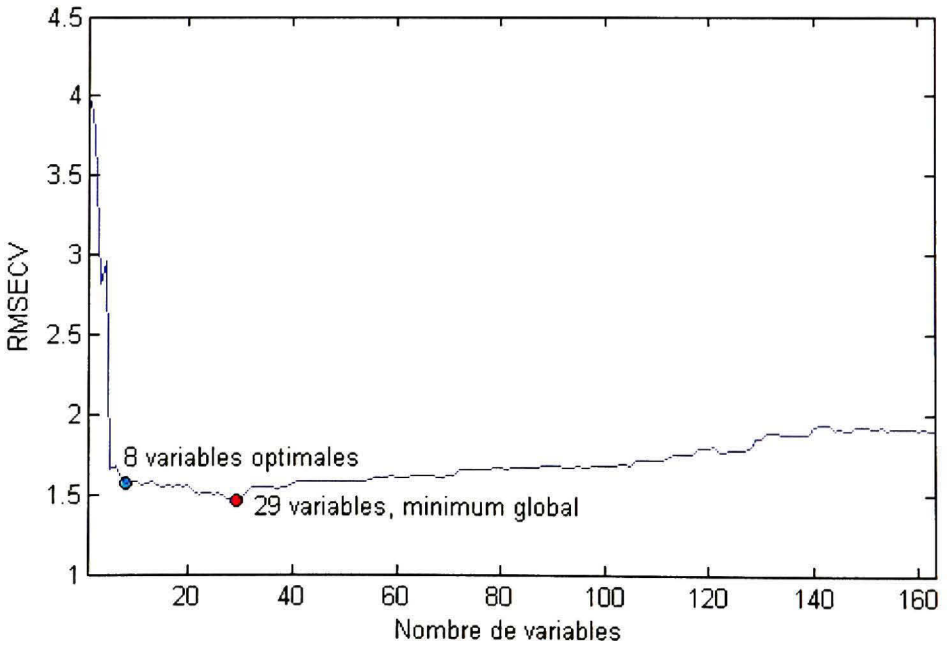

Figure 2.28 ***Erreur RMSECV en fonction du nombre de variables.***

Au cours des itérations, en plus de relever l'erreur RMSECV, nous relevons aussi les bandes spectrales mises en jeu dans l'étalonnage des modèles. De cette manière, nous savons, pour une erreur donnée, quelles bandes spectrales ont été utilisées lors de la modélisation. En sortie de notre procédure d'élagage PLS, nous avons les indices des bandes sélectionnées.

2.5.2 Algorithme de colonie de fourmis

Comme nous l'avons vu dans le chapitre précédent, l'ACF est une méthode utilisée dans le cadre de la sélection de variables spectrales. Nous revenons ici plus en détails sur les mécanismes de cette approche.

Cet algorithme repose sur des essais de solutions construites de manière probabiliste. Une solution est une sélection d'un sous-ensemble des variables initiales. La construction probabiliste des solutions se fait à l'aide d'un vecteur dit de phéromone, $\boldsymbol{\tau}$, dont la taille est égale au nombre de variables initiales de notre système, et dont chaque élément représente une des variable initiales. La valeur de chaque élément de ce vecteur influera sur la sélection ou non de la variable correspondante lors de la génération de la solution. Plus cette valeur sera élevée, plus la chance de sélection de cette variable sera élevée.

Initialement toutes les m variables doivent avoir une probabilité égale de sélection lors de la génération des solutions. Cela implique que tous les éléments du vecteur de phéromone doivent être égaux à 1, (2.42). Ce sont ces valeurs que nous ferons varier au sein du processus itératif de manière à identifier les variables pertinentes, et influer sur leur sélection par les solutions. On définit un vecteur de probabilité, $\boldsymbol{P}$ (2.43), et un vecteur cumulatif des probabilités, $\boldsymbol{CP}$ (2.44).

$$\tau_k = 1,\ k \in [\![1, m]\!] \tag{2.42}$$

$$P_k = \frac{\tau_k}{\sum_{k=1}^{m} \tau_k} \tag{2.43}$$

$$CP_k = \sum_{i=1}^{k} P_i \tag{2.44}$$

Un ensemble ϕ de solutions est construit en se basant sur le vecteur ***CP***. Nous pouvons considérer les solutions comme un vecteur binaire de taille m (nombre de variables totales) avec des uns là où une variable est retenue, et des zéros là où la variable n'est pas sélectionnée. Nous considérons la taille d'une solution, l, comme étant le nombre de variables sélectionnées par cette solution (c'est-à-dire la somme des éléments du vecteur binaire). La génération d'une solution est réalisée à l'aide du vecteur ***CP*** de probabilités cumulatives, et de nombres générés aléatoirement. Les éléments du vecteur ***CP***, sont compris entre zéro et un et sont classés par ordre croissant. Nous pouvons les voir comme des bornes, délimitant des sous-ensembles plus ou moins grands suivant la valeur plus ou moins élevée correspondante dans le vecteur $\boldsymbol{\tau}$. En effet, pour une longueur d'onde k, plus la valeur de τ_k est élevée, plus P_k est élevé et ainsi plus la différence entre CP_{k-1} et CP_k est importante. Le principe sera de générer des nombres aléatoires compris entre zéro et un, et identifier le sous-ensemble dans lequel il se trouve pour sélectionner la longueur d'onde correspondante dans le vecteur de solution. Le sous ensemble de valeurs pour la longueur d'onde k est défini par $]CP_{k-1}, CP_k]$. Pour la première longueur d'onde, le sous-ensemble est défini par $]0, CP_1]$. Comme une longueur d'onde ne peut pas être sélectionnée plus d'une fois, si un nombre aléatoirement généré tombe dans le sous-ensemble d'une longueur d'onde déjà sélectionnée, rien ne se passe, et il faut générer un nouveau nombre. La sélection des longueurs d'ondes présentes dans une solution s'arrête lorsque le nombre total de longueurs d'ondes sélectionnées est égal à sa taille, l, générée initialement aussi de manière aléatoire. Voici le pseudo-code de la fonction de génération d'une solution :

Génération d'une solution

```
Génération aléatoire de taille_solution
taille_coutrante = 0

Tant que taille_courante ≠ taille_solution faire
        Génération d'un nombre aléatoire compris entre 0 et 1
        Identifier dans quel sous-ensemble se situe le nombre
        Si la longueur d'onde correspondante n'est pas déjà sélectionnée
                Sélection de la longueur d'onde
                taille_courante = taille_courante + 1
        Fin si
Fin tant que
```

Une fois l'ensemble des solutions généré, nous calculons autant de modèles qu'il y a de solutions proposées. Pour simplifier les calculs et gagner du temps, le nombre de variables latentes des modèles est défini par le nombre de variables latentes déterminé par la modélisation utilisant toutes les bandes spectrales. La performance brute de chaque solution est évaluée par validation croisée v-blocs. Ainsi, une valeur de PRESS est associée à chaque solution. Cette valeur de PRESS est utile dans l'évaluation d'une fonction d'adéquation qui nous permettra de classer la performance de chaque solution i les une vis-à-vis des autres. Nous évaluons cette fonction suivant :

$$G_i = \frac{1}{PRESS_i} \tag{2.45}$$

Ainsi, G est inversement proportionnel à l'erreur PRESS, et c'est ce que nous cherchons, car les meilleurs modèles seront ceux ayant la valeur de PRESS la plus faible.

Afin de pouvoir comparer sur un pied d'égalité chaque solution proposée, nous normalisons et multiplions par une constante α chaque valeur G_i. Nous appelons Gn_i cette nouvelle valeur, et elle montre la qualité de la solution. Plus Gn_i est élevé meilleure est jugée la solution.

$$Gn_i = \alpha \cdot \left[\frac{G_i}{\sum_{i=1}^{\phi} G_i} \right] \tag{2.46}$$

Gn_i est comprise entre zéro et un. Il est important de garder Gn_i inférieur à un car de larges valeurs d'adéquations peuvent amplifier de mauvais choix de longueur d'ondes dus au caractère aléatoire de la génération des solutions dans les premières itérations de l'algorithme. C'est une valeur comprise entre 0.3 et 0.8

L'étape suivante consiste à mettre à jour le vecteur de sélection de variables. Nous retenons un ratio des meilleures solutions proposées et mettons à jour la probabilité de sélection de chaque variable suivant l'équation (2.47). Nous ne retenons que 50% des meilleures solutions, en effet, retenir trop ou trop peu de solutions empêche la méthode de converger. Nous appelons ϕ_{best} cet ensemble. Nous ne mettons à jour que les éléments de $\boldsymbol{\tau}$ qui correspondent aux longueurs d'ondes sélectionnées par la solution. Notons β_i le sous-ensemble de longueurs d'ondes sélectionnées par la solution i, $\tau_k(t)$ est la valeur mise à jour, et $\tau_k(t-1)$ est la valeur du vecteur à l'itération précédente.

$$\tau_k(t) = \tau_k(t-1) \times (1 + Gn_i),\ i \in \phi_{best},\ k \in \beta_i \tag{2.47}$$

Ainsi, les variables présentes dans les solutions retenues auront plus de chance d'être sélectionnées à l'itération suivante. Dans les algorithmes de colonie de fourmis adaptés à la sélection de variables, une fourmi est assimilée à une solution de notre problème, dont la construction sera influencée par le vecteur de phéromone $\boldsymbol{\tau}$. De manière à reproduire le

phénomène naturel d'évaporation des phéromones, τ est multiplié par une constante ρ comprise entre 0.3 et 1.

Pour pallier la perte d'informations, et ajouter éventuellement des variables non forcement présentes dans les populations de solutions, nous ajoutons une étape de mutation au processus. Cette étape permet d'augmenter la diversité des solutions candidates au problème. De cette manière nous pouvons introduire des variables éventuellement importantes qui n'ont pas eu la chance d'être sélectionnées dans des solutions jusqu'alors viables. La mutation change de manière aléatoire la sélection ou non d'une variable. Une nouvelle solution est donc proposée, et avec elle une nouvelle valeur de PRESS. Dans le cas d'une désélection de variables, si la nouvelle valeur de PRESS est inférieure à la valeur de PRESS initiale, la mutation est acceptée. Dans le cas contraire (sélection de variables initialement désélectionnées), un test de Fisher (Haaland et Thomas, 1988) est appliqué pour déterminer si nous acceptons ou non la solution. Nous calculons le ratio des erreurs PRESS avant mutation sur erreur PRESS après mutation, et calculons la p-valeur correspondante avec les degrés de liberté égaux aux nombre de variables sélectionnées par la solution moins un avant et après mutation. Nous retenons la mutation que si $p < 0.3$. p est le seuil à partir duquel nous considérons que la différence observée est statistiquement significative.

$$F = \frac{PRESS_{avant\ mutation}}{PRESS_{après\ mutation}} \tag{2.48}$$

Au final, plusieurs paramètres sont donc à prendre en compte dans cette méthode. Le nombre d'itération maximum de la procédure, le nombre de solutions candidates à générer, et certaines constantes utilisées lors de calculs de mise à jour. La meilleure combinaison des paramètres du modèle : $[\alpha, \rho, \phi]$ est établie grâce à un design factoriel. Nous retenons 3 valeurs pour chaque paramètre : les deux extrêmes, et une valeur centrale. Toutes les combinaisons des variations sont testées. Pour chaque combinaison de paramètres, nous notons l'évolution des performances à chaque itération de l'algorithme. Nous relevons la moyenne, ainsi que l'écart type des erreurs PRESS des solutions retenues.

Nous observerons l'influence des paramètres α et ρ. Nous étudierons la moyenne et l'écart type des PRESS des solutions retenues à chaque itération pour différentes valeurs de α et ρ. La moyenne nous montrera le comportement de l'algorithme. Nous nous attendons à une diminution et une stabilisation de celle-ci à un minimum. Il en va de même pour l'écart type, celui-ci doit diminuer puis se stabiliser, indiquant que les solutions semblent converger vers une même valeur.

α et ρ influeront directement sur la manière dont va converger l'algorithme. Lorsque trop importants, $\alpha = 0.8$, $\rho = 1$, nous remarquons qu'après une phase rapide de diminution de l'erreur, celle-ci augmente de manière conséquente (il en va de même pour l'écart type). Le choix des longueurs d'ondes dans les itérations initiales de l'algorithme repose sur un phénomène aléatoire. Beaucoup des longueurs d'ondes initialement choisies ne sont pas forcement pertinentes pour notre problème, mais sont quand même retenues dans les meilleures solutions, et plus α sera élevé, plus leur score sera important. Durant la phase de diminution de l'erreur PRESS, les longueurs d'ondes pertinentes ont encore la possibilité d'être choisies avec les non-pertinentes, mais le fait d'avoir α élevé donne l'avantage aux longueurs d'ondes qui sont choisies depuis le début, et l'algorithme tend alors au fur et à mesure de son avancement à ne sélectionner que celles-ci. Donc si ces longueurs d'ondes initiales ne sont pas parmi les plus pertinentes, alors les performances tendent à diminuer. C'est ce que nous observons à la Figure 2.29.

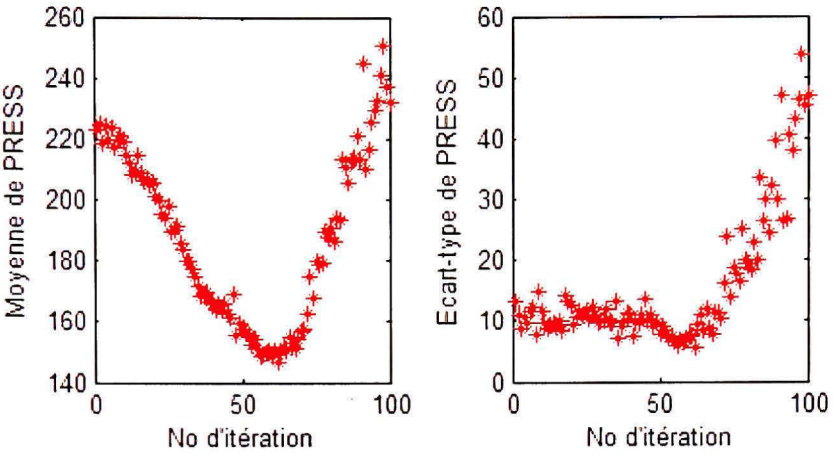

Figure 2.29 ***Évolution de la moyenne et écart type des solutions pour α =0.8, ρ =1.***

Inversement, lorsque α et ρ sont faibles, $\alpha = 0.3$, $\rho = 0.3$, la diminution de l'erreur est plus lente, et l'écart type semble indiquer que les solution retenues ont des performances assez différentes, indiquant que la méthode ne converge pas (*Voir* Figure 2.30).

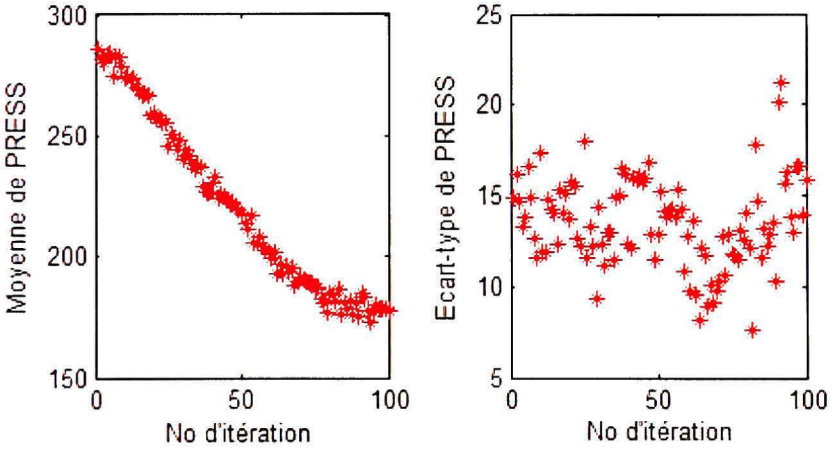

Figure 2.30 ***Évolution de la moyenne et écart type des solutions pour α =0.3, ρ =0.3.***

Des valeurs convenables sont $\alpha = 0.5$, $\rho = 0.8$, l'erreur diminue puis se stabilise vers un minimum, de même que l'écart type des meilleurs solutions retenues diminue, impliquant que les solutions semblent converger (*Voir* Figure 2.31).

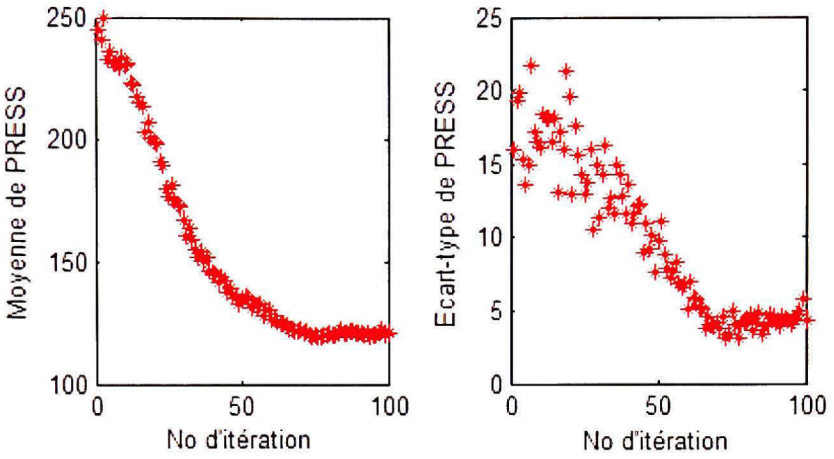

Figure 2.31 ***Évolution de la moyenne et écart type des solutions pour*** $\alpha = 0.55$, $\rho = 0.8$.

De la même manière, nous avons choisi un nombre ϕ de solutions totales candidates à notre problème, égale à 100. Trop peu de solutions (25-75) empêchent l'algorithme de converger, et trop de solutions (125-150) impliquent un temps de traitement supplémentaire pour des résultats équivalents. À chaque itération, nous ne retenons donc que les 50 meilleures solutions. Le nombre maximal d'itérations a été fixé à 100. D'après la Figure 2.31, 100 itérations semblent suffisantes pour permettre à la méthode de converger. Lorsque toutes les solutions générées sont identiques, alors nous avons convergence de l'algorithme, et celui-ci s'arrête, sinon celui-ci continue jusqu'à atteindre 100 itérations. Lorsque les itérations sont finies, nous calculons le vecteur des probabilités $\boldsymbol{P}$ (2.43).Nous ne retenons que les indices des longueurs d'ondes dont la probabilité de sélection est supérieure à 0.1. Cette valeur à été obtenue par essais et erreurs.

CHAPITRE 3

ETUDE QUALITATIVE TOUTES BANDES

3.1 Aperçu général

Au travers de ce chapitre, nous comparerons les performances de modèles de régression PLS déterminées suivant l'application de différentes configurations de paramètres de base. En utilisant l'ensemble des bandes spectrales disponibles, nous considérerons le spectre en mode réflectance et en mode absorbance, de même, nous étudierons l'influence de trois méthodes d'évaluation de la complexité des modèle PLS : validation croisée *leave-one-out* (LOO), validation croisée 5-blocs, et méthode utilisant les diagrammes de Pareto. Nous présenterons donc les résultats obtenus par les modèles déterminés à partir des différentes configurations. Les résultats présentés sont les moyennes et les écarts types des indices de performance, obtenus en fonction des 50 différentes bases de données aléatoires.

La comparaison de ces résultats nous permettra de faire ressortir quelle configuration semble être la meilleure en utilisant toutes les bandes spectrales pour prédire la matière sèche des avocats. C'est cette configuration que nous allons utiliser lors de l'étape suivante qui consiste en la sélection des bandes spectrales les plus pertinentes, pour la prédiction du taux de matière sèche.

3.2 Présentation des résultats

Les résultats que nous avons obtenus seront présentés sous forme de tableaux. Pour ce faire, nous avons déterminé à partir des 50 bases de données aléatoires les 50 modèles correspondant en utilisant chacune des trois méthodes d'évaluation de la complexité des modèles PLS : validation croisée *Leave-One-Out*, validation croisée 5-blocs, diagrammes de Pareto, et ce, suivant chacun des deux modes : réflectance et absorbance. Nous présentons ici les moyennes de chaque indice de performance obtenues pour les 50 modèles ainsi

déterminés. Cela nous donne six configurations à analyser. Chaque moyenne est présentée avec entre parenthèses l'écart type correspondant.

Nous pouvons observer au Tableau 3.1 les racines carrées des erreurs quadratiques obtenues sur les échantillons de l'ensemble d'étalonnage (RMSEC) et sur l'ensemble de prédiction (RMSEP).

Tableau 3.1

Moyennes des erreurs d'étalonnage et de prédiction des modèles

Évaluation de la complexité des modèles	RMSEC		RMSEP	
	Réflectance	**Absorbance**	**Réflectance**	**Absorbance**
Pareto	1.03 (0.07)	0.73 (0.05)	2.46 (0.28)	1.70 (0.16)
CV 5-blocs	1.10 (0.74)	0.80 (0.17)	2.41 (0.42)	1.65 (0.17)
CV LOO	6.04 (23.71)	1.87 (7.58)	8.73 (26.06)	3.08 (9.66)

Note : Le nombre entre parenthèses représente l'écart type correspondant.

Les moyennes des coefficients de détermination et les écarts types des résidus (SDR) obtenus sur les 50 modèles suivant chaque configuration sont donnés au Tableau 3.2.

Tableau 3.2

Moyennes des coefficients de détermination et des écarts type des résidus des modèles

Évaluation de la complexité des modèles	R^2		SDR	
	Réflectance	**Absorbance**	**Réflectance**	**Absorbance**
Pareto	0.86 (0.03)	0.93 (0.01)	2.75 (0.44)	4.02 (0.41)
CV 5-blocs	0.86 (0.05)	0.94 (0.01)	2.82 (0.59)	4.10 (0.45)
CV LOO	0.79 (0.24)	0.91 (0.13)	2.62 (0.74)	3.92 (0.75)

Note : Le nombre entre parenthèses représente l'écart type correspondant.

Enfin les nombres moyens de variables latentes retenus pour chaque configuration sont présenté au Tableau 1.3.

Tableau 3.3

Nombres moyens de variables latentes retenues pour les modèles

Évaluation de la complexité des modèles	Nb. variables latentes	
	Réflectance	**Absorbance**
Pareto	12.90 (0.30)	13.44 (0.50)
CV 5-blocs	16.60 (11.12)	12.84 (1.43)
CV LOO	21.86 (9.92)	14.30 (7.28)

Note : Le nombre entre parenthèses représente l'écart type correspondant.

De manière générale les modèles en mode absorbance performent mieux que les modèles issus du mode réflectance. De même, à première vue, la méthode d'évaluation du nombre de variables latentes des modèles qui à les meilleures performances semble être la validation croisée 5-blocs. Nous analyserons ces résultats, en commençant par la comparaison des spectres des deux modes pour expliquer cette différence dans les performances; ensuite nous verrons plus en détail les performances données par les différentes méthodes pour le choix du nombre de variables latentes à retenir pour nos modèles.

3.3 Analyse du spectre des avocats

Nous pouvons observer à la Figure 3.1 le spectre typique en mode réflectance et absorbance d'un avocat. Nous remarquons le peu de réflectance entre 400nm et 700nm, avec un maximum local autour de 550nm, ainsi qu'un accroissement rapide de la réflectance à partir de 700nm, appelé la chute du rouge, (*red edge*) (Blackburn, 2007), menant à une zone de haute réflectance, ou les pigments n'absorbent que peu les radiations incidentes.

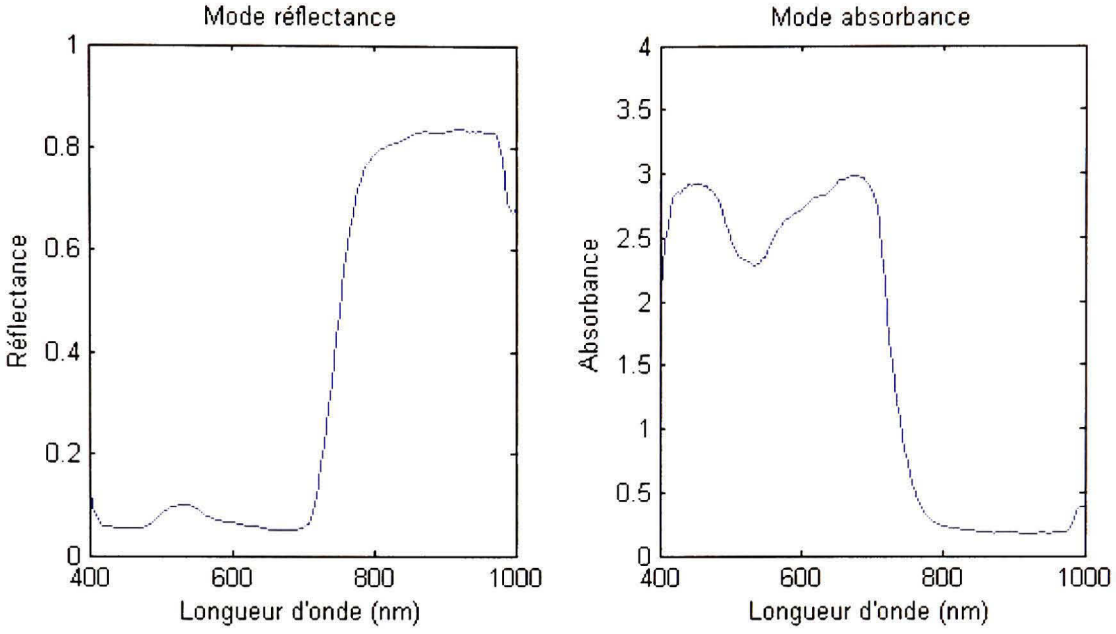

Figure 3.1 ***Spectre typique d'un avocat en mode réflectance et absorbance.***

3.3.1 Pigments présents dans la peau des fruits.

L'allure des ces spectres peut être expliquée par les pigments présents dans la peau des fruits. Les principaux pigments responsables des couleurs des végétaux (peau et chair) sont la chlorophylle (α et β), l'anthocyane, et les caroténoïdes. Dans la peau des avocats, sont présents les pigment chlorophylliens, ceux issus des caroténoïdes (lutéine, β-carotène), et ceux issus de l'anthocyane (Ashton et al., 2006). Nous pouvons observer à la Figure 3.2 l'allure des spectres d'absorption de trois principaux pigments : la chlorophylle α, la chlorophylle β, et le β-carotène. Nous pouvons remarquer les pics d'absorption présents entre 400nm et 475nm, et entre 500nm, et 600nm. Cela explique la faible réflectance du spectre observée à ces intervalles de longueurs d'ondes sur la Figure 3.1.

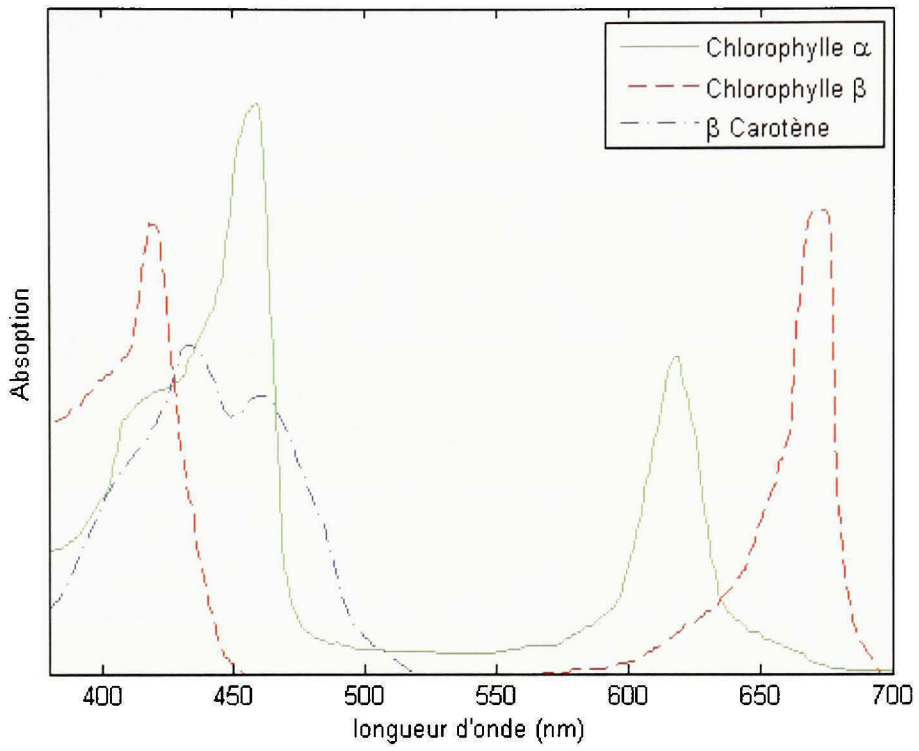

Figure 3.2 ***Spectres d'absorption de trois principaux pigments.***

(Tiré de Hemraj et al., 1997; Isler, Gutmann et Solms, 1971)

Source : Cette figure sous forme de graphique est tirée de la réunion de deux sources. Les courbes des pigments chlorophylliens sont issues du travail de Indradeo Hemraj *et al.*, *Absoption Spectum of chlorophyll*, p. 5. La courbe du pigment β-carotène est tiré du livre de Otto Isler *et al.*, *Carotenoids*, p. 194.

3.3.2 Différences entre les spectres en mode réflectance et absorbance

Afin d'analyser la différence des résultats obtenus entre les modes réflectance et absorbance, nous étudierons les spectres moyens de quatre groupes d'avocats. Ces quatre groupes ont été obtenus en regardant la répartition des taux de matière sèche de nos échantillons. Nous avons un groupe dont la matière sèche est supérieure à 37.5%, un deuxième groupe dont la matière sèche est comprise entre 31% et 37.5%, le troisième groupe est formé par les avocats dont le taux de matière sèche est supérieur à 25% mais inférieur à 31%, enfin le quatrième et dernier groupe représente les avocats dont le taux de matière sèche est inférieur à 25%. Nous

pouvons observer les quatre groupes à la Figure 3.3. Suivant l'axe des abscisses, nous avons les numéros des avocats étudiés avec les quatre mesures de matière sèche correspondantes. Suivant l'axe des ordonnées, nous avons le taux de matière sèche.

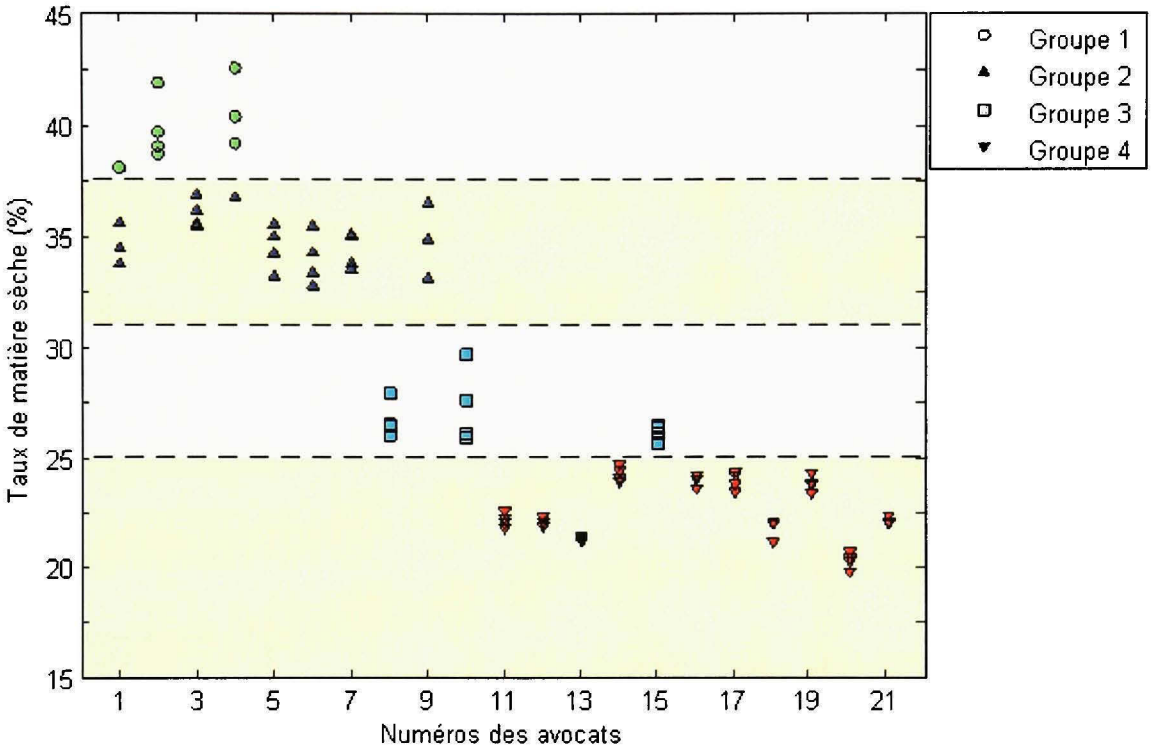

Figure 3.3 ***Groupage des échantillons en fonction des taux de matière sèche.***

À partir de ces quatre groupes, nous pouvons calculer le spectre moyen fourni par les échantillons des groupes. Nous pouvons observer les quatre spectres moyens à la Figure 3.4. Le passage du mode réflectance au mode absorbance nous a permis d'augmenter les différences entre les spectres dans la zone de 400nm à 700nm. À l'inverse les spectres moyens entre 700nm et 1000nm semblent plus rapprochés en mode absorbance qu'en mode réflectance.

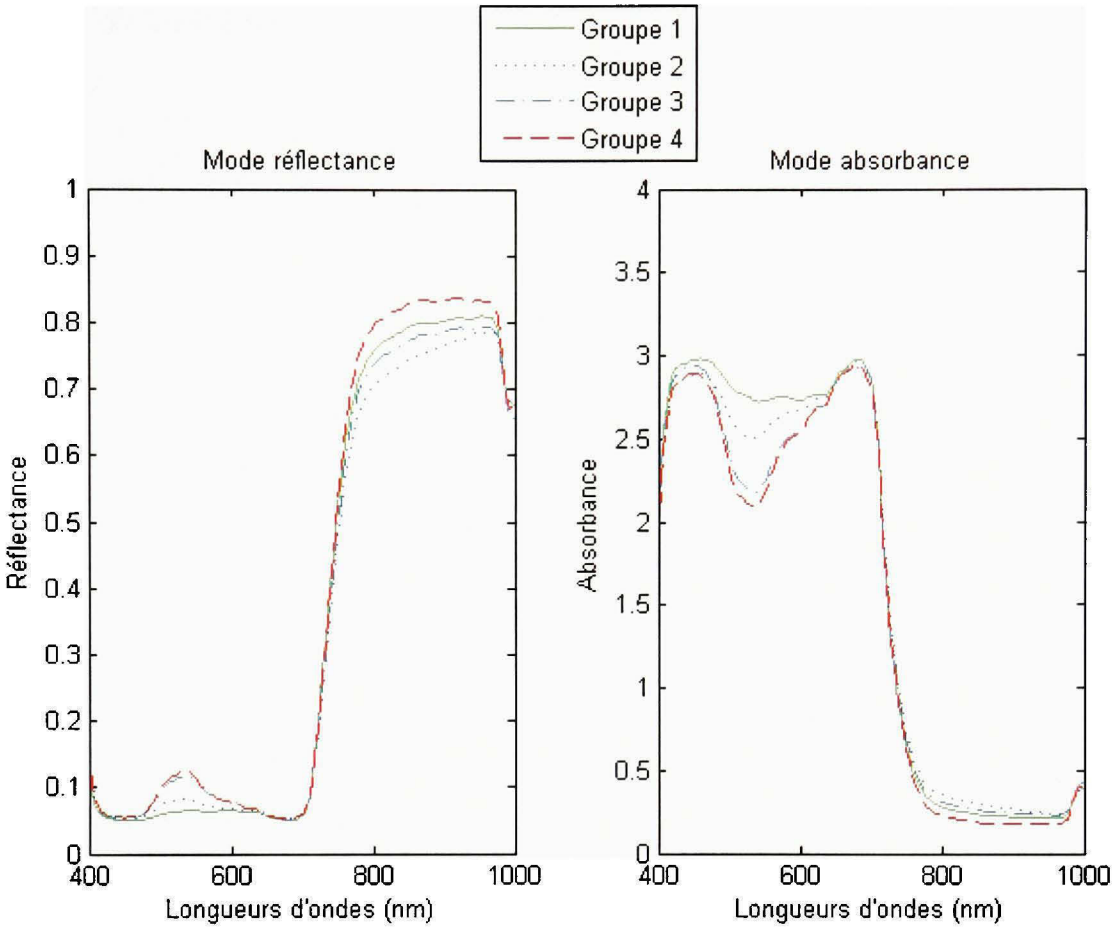

Figure 3.4 ***Spectres moyens des échantillons.***

Nous avons pu remarquer que l'utilisation conjointe du mode réflectance et absorbance, de manière à avoir la première partie du spectre en mode réflectance, et la deuxième en mode absorbance (la délimitation étant avant ou après la forte pente), ne donnait pas de modèles plus performants que l'utilisation du seul mode absorbance. De même, l'utilisation de toutes les valeurs d'absorbance, et toutes les valeurs de réflectance, soit 326 variables donne des résultats proches mais légèrement inférieurs à ce ceux donnés par les 163 bandes spectrales en mode absorbance.

3.4 Choix de la méthode d'évaluation de la complexité des modèles

Nous analyserons maintenant les résultats obtenus en fonction des méthodes servant à définir la complexité des modèles PLS. Nous avons 2 méthodes de validation croisée, et une méthode utilisant les graphes de Pareto (minimisation variance/résidus).

Tout d'abord, concernant les deux méthodes de validation croisée, nous pouvons remarquer que la méthode 5–blocs performe mieux que la méthode LOO (*Leave-One-Out*), qui est un cas particulier de 5–blocs (lorsque v est égal au nombre d'échantillons de notre base d'étalonnage). D'une manière générale la validation croisée LOO retiendra plus de variables latentes que nécessaire, ce qui induit une faible erreur de calibration, mais au détriment de l'erreur de prédiction (Qing-Song Xu, 2004). Tels quels, les résultats moyens des erreurs de calibration et de prédiction ne reflètent pas cela. En examinant mieux les résultats fournis par les 50 modèles, nous pouvons remarquer que quatre d'entre eux semblent être aberrants (*Voir* Annexe IV, tableau 3.1). Ce sont les modèles 7, 13, 22, et 31, dont les erreurs RMSEV sont plus conséquentes (respectivement 80.91, 54.15, 26.11, et 163.25). Cela explique les fortes moyennes et les hauts écarts types. Si nous retirons ces modèles, et calculons à nouveau les moyennes des performances des 46 autres modèles, nous obtenons les résultats présentés au Tableau 3.4.

Tableau 3.4

Moyennes des performances en mode réflectance sur les 46 modèles non aberrants

Évaluation de la complexité des modèles	RMSEC	RMSEP	Nb. Variables Latentes
CV-LOO	0.48 (0.36)	2.45 (0.30)	19.6 (6.34)
CV 5-blocs	0.97 (0.74)	2.44 (0.39)	17.17 (10.82)
Pareto	1.03 (0.07)	2.45 (0.28)	12.89 (0.31)

Note : Le nombre entre parenthèses représente l'écart type correspondant.

Nous pouvons remarquer que l'erreur de calibration (RMSEC) est faible, et même inférieure (jusqu'à 2 fois) aux erreurs des autres modèles. Les erreurs de prédiction (RMSEP) sont équivalentes, mais le nombre de variables latentes données par la validation croisée LOO est supérieur aux deux autres. Un grand nombre de variables latentes retenues par un modèle induit souvent du sur-apprentissage (Baumann, 2003). Sur les 46 modèles, que ce soit avec la validation croisée LOO ou la validation croisée 5–blocs, le nombre de variables latentes retenu est élevé, et peu constant suivant les modèles, en effet l'écart type est lui aussi très élevé, cela induit de fortes fluctuations : tantôt le nombre de variables latentes est raisonnable, et le modèle a de bonnes performances (validation croisée 5–blocs : modèle 36, nombre de variables latentes = 7, $R^2 = 0.93$), tantôt le modèle comporte beaucoup de variables latentes et les performances sont mauvaises (validation croisée 5–blocs : modèle 38, nombre de variables latentes = 43, $R^2 = 0.68$). Les modèles déterminés au moyen des graphes de Pareto ne semblent pas affectés par le sur-apprentissage, en effet le nombre de variables latentes retenues est d'environ 13 avec un faible écart type de 0.31. C'est cette fonction d'évaluation de la complexité qui, dans le mode réflectance, semble donner les meilleures performances.

Même si les performances de modèles ainsi obtenus semblent être acceptables, $R^2 > 0.8$ (*Voir* Tableau 3.5), en considérant l'écart type des résidus (SDR), nous remarquons que celui-ci est inférieur à 3. L'écart type des résidus nous permet d'appréhender mieux que le coefficient de détermination et la racine carrée de l'erreur quadratique moyenne séparément ne le peuvent, les réelles aptitudes de prédiction de nos modèles (McGlone et Kawano, 1998). Généralement une valeur de 3 est un minimum pour de bonnes aptitudes de prédiction.

Tableau 3.5

Moyennes des performances en mode réflectance sur les 46 modèles non aberrants

Évaluation de la complexité du modèle	R^2	SDR
CV-LOO	0.86 (0.04)	2.79 (0.46)
CV 5-blocs	0.86 (0.05)	2.78 (0.54)
Pareto	0.86 (0.04)	2.73 (0.44)

Note : Le nombre entre parenthèses représente l'écart type correspondant.

De la même manière que le nombre moyen de variables latentes des modèles possède un écart type élevé, les écarts types des coefficients de détermination indiquent une forte variation de ceux-ci. Cela nous indique que les performances des modèles sont assez variables, et donc, que les modèles ainsi déterminés ne sont pas très stables au niveau des performances. Le mode réflectance ne semble pas être adapté à notre problème.

Concernant le mode absorbance, pour la validation croisée LOO, nous pouvons remarquer un modèle dont les résultats sont aberrants, le modèle 20 (*Voir* Annexe IV, tableau 3.3). De la même manière, si nous retirons ce modèle et recalculons les moyennes des performances, nous obtenons les résultats présentés au Tableau 3.6.

Tableau 3.6

Moyennes des performances en mode absorbance sur les 49 modèles non aberrants

Évaluation de la complexité du modèle	RMSEC	RMSEP	Nb. Variables Latentes
CV-LOO	0.79 (0.20)	1.72 (0.27)	13.53 (4.89)
CV 5-blocs	0.80 (0.17)	1.65 (0.17)	12.88 (1.42)
Pareto	0.73 (0.05)	1.69 (0.17)	13.44 (0.50)

Note : Le nombre entre parenthèses représente l'écart type correspondant.

En mode absorbance, les performances des modèles sont meilleures et plus homogènes. Notons toutefois que l'écart type du nombre de variables latentes retenues par la validation croisée LOO est presque de 5, cela indique une forte variation du nombre pour les différents modèles déterminés.

Tableau 3.7

Moyennes des performances en mode absorbance sur les 49 modèles non aberrants

Évaluation de la complexité du modèle	R^2	SDR
CV-LOO	0.93 (0.02)	3.99 (0.56)
CV 5-blocs	0.94 (0.01)	4.10 (0.48)
Pareto	0.93 (0.01)	4.02 (0.41)

Note : Le nombre entre parenthèses représente l'écart type correspondant.

De même, les coefficients de détermination ainsi que les écarts types sont meilleurs que ceux obtenus en mode réflectance, et nous avons dépassé la barre de 3 pour les écarts types des résidus (SDR).

3.5 Configuration retenue

Au vu de l'analyse précédente, Le mode réflectance ne semble pas être un mode où nous pouvons conduire notre étude plus avant. Les faibles performances obtenues ainsi que les fortes fluctuations de celles-ci pour la validation croisée ne nous permettent pas de choisir ces fonctions d'évaluation pour ce mode. L'utilisation des diagrammes de Pareto est un peu plus stable au niveau des performances. Les écarts types des résidus (SDR) moyens étant inférieurs à une valeur de 3 pour les trois fonctions d'évaluations de la complexité, nous ne pouvons retenir ce mode.

À l'inverse, en mode absorbance, les performances sont meilleures, et équivalentes pour les différentes fonctions d'évaluation de la complexité des modèles. La validation croisée LOO retient plus de variables latentes que la validation croisée 5-blocs, c'est pourquoi nous avons choisi de la mettre de coté. Pareto donne des résultats de même niveau que la validation croisée 5-blocs; cela peut représenter une alternative intéressante. Nous avons néanmoins choisi d'utiliser la validation croisée 5-blocs, en effet, c'est cette méthode qui est largement utilisée dans la littérature pour évaluer la complexité d'un modèle PLS.

Nous avons analysé les performances de 50 modèles PLS en utilisant le spectre dans sa globalité de manière à déterminer quelle configuration des données nous devrions utiliser, et comment évaluer le nombre de variables latentes des modèles PLS. Il est intéressant, maintenant que nous avons identifié la configuration optimale, de nous pencher sur le problème de la sélection des bandes spectrales.

CHAPITRE 4

SÉLECTION DE BANDES SPECTRALES

4.1 Aperçu général

Au travers de ce chapitre, nous présenterons et analyserons les résultats obtenus lorsque nous avons appliqué nos deux méthodes de sélection de bandes spectrales. La configuration utilisée pour la détermination des différents modèles PLS est celle obtenue au chapitre précédent : mode absorbance, et validation croisée 5-blocs. Cette analyse nous permettra de déterminer si nous pouvons appliquer des concepts de réduction de variables à notre problème. Si c'est le cas, nous présenterons quelles ont été les bandes pertinentes déterminées par les deux méthodes et discuter ce choix.

4.2 Présentation des résultats

Nous avons appliqué les méthodes de sélection de variables aux 50 bases de données aléatoires et obtenu à chaque itération les bandes sélectionnées. Nous avons alors déterminé le modèle et les performances correspondants. De même, nous avons relevé les bandes sélectionnées pour les 50 bases de données. Ainsi, nous pouvons mettre en place un système de vote. Chaque bande lorsque sélectionnée donnant une voie, nous pouvons alors établir un diagramme en bâton où nous pouvons observer quelles sont les bandes majoritairement choisies.

Le Tableau 4.1 présente les moyennes ainsi que les écarts types des performances obtenues à chaque itération des 50 bases de données aléatoires. Nous avons relevé l'erreur sur l'ensemble de test (RMSEP), le coefficient de détermination (R^2), l'écart type des résidus (SDR), et le nombre de variables proposé par la méthode de sélection.

Tableau 4.1

Moyennes des performances des deux fonctions de sélection de variables

Méthode de sélection de variables	RMSEP	R^2	SDR	Nombre de variables
Élagage PLS	1.60 (0.23)	0.94 (0.02)	4.31 (0.58)	14.76 (5.60)
ACF	2.18 (0.58)	0.88 (0.07)	3.28 (0.75)	17.92 (8.29)

Note : Le nombre entre parenthèses représente l'écart type correspondant.

D'après ces résultats, l'élagage PLS semble donner de meilleurs résultats que l'algorithme de colonie de fourmis. En effet les performances sont toujours supérieures et les écarts types sont systématiquement plus faibles. Cela indique que sur les 50 bases de données aléatoires les résultats sont meilleurs, et aussi plus uniformes en utilisant l'élagage PLS.

Nous allons maintenant présenter les résultats des votes des deux méthodes de sélection de variables, et interpréter les choix donnés par ces algorithmes.

4.2.1 Élagage PLS

La Figure 4.1 présente les résultats des votes relevés suivant les 50 bases de données aléatoires et leurs bandes sélectionnées correspondantes. D'une manière générale, nous pouvons observer que certaines bandes (en rouge) sont beaucoup plus souvent sélectionnées que d'autres. Après analyse de la Figure 4.1, nous avons choisi ces bandes de telle manière que leur nombre de vote soit supérieur à 18. Cela donne un nombre de bandes égal à 16. Ces 16 bandes sont détaillées au Tableau 4.2.

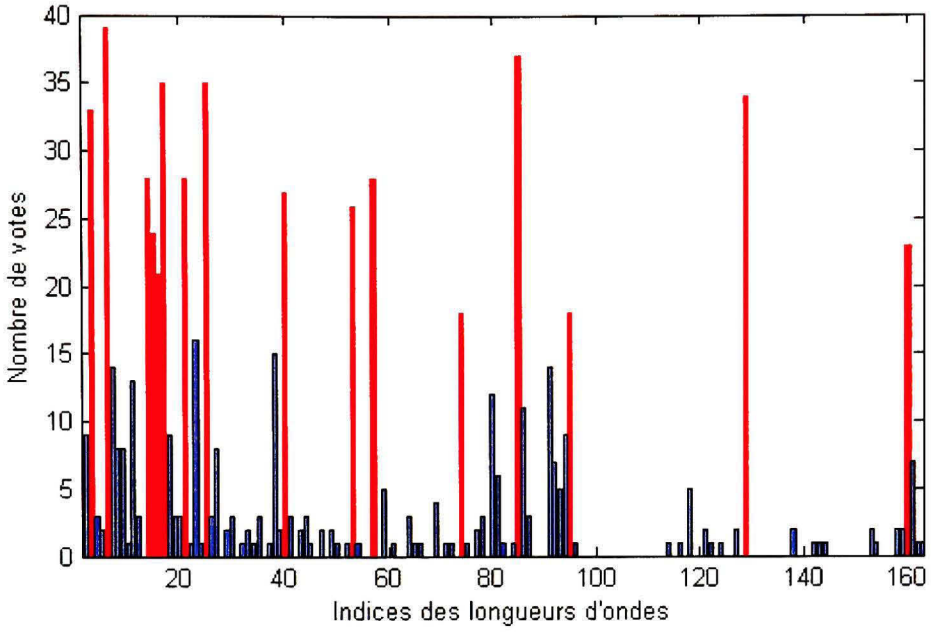

Figure 4.1 ***Résultats des votes pour l'élagage PLS.***

Tableau 4.2

Bandes sélectionnées et leurs longueurs d'ondes correspondantes

Indices des bandes	3	6	14	15	16	17	21	25
Longueurs d'ondes (nm)	425.58	458.97	486.35	489.27	492.19	495.110	506.88	510.69
Indices des bandes	40	53	57	74	85	95	129	160
Longueurs d'ondes (nm)	563.08	601.40	613.16	663.09	695.41	724.83	825.74	959.81

Nous pouvons remarquer que sur ces 16 bandes, 14 sont dans le domaine du visible, et deux dans le domaine du proche infrarouge. Cela semble confirmer l'analyse des différences entre les spectres en mode absorbance vue au chapitre précédent. Si nous utilisons ces 16 bandes et relevons les performances moyennes obtenues pour les 50 bases de données, nous obtenons

les résultats présentés au Tableau 4.3. Dans un but de comparaison, nous avons aussi inscrit les résultats obtenus lors de l'analyse toutes bandes.

Tableau 4.3

Moyennes des performances de deux configurations de sélection de variables

Nombre de bandes	RMSEP	R^2	SDR
16	1.35 (0.17)	0.96 (0.01)	5.03 (0.66)
163	1.65 (0.17)	0.94 (0.01)	4.10 (0.45)

Note : Le nombre entre parenthèses représente l'écart type correspondant.

Nous pouvons remarquer qu'avec 10% des bandes initiales, nous obtenons des résultats meilleurs qu'en utilisant le spectre dans son entier. Nous avons gagné 0.02 pour le coefficient de détermination, et quasiment 1 sur l'écart type des résidus.

Si nous nous intéressons aux bandes sélectionnées, nous remarquons que quatre d'entres elles sont adjacentes : ce sont les bandes 14, 15, 16, et 17. Sur cette plage de longueur d'onde (486.35nm-495.110nm), la résolution de l'appareil est de 2.94nm, ce qui est relativement petit. Il est possible que l'information véhiculée par ces quatre bandes soit en fait la même. Pour évaluer cela, nous pouvons analyser l'effet, sur les performances, des différentes combinaisons possibles de leur sélection ajoutées aux 12 autres bandes. De la même manière, les bandes 74 et 95 obtiennent le nombre de votes minimum pour être retenues. Il peut être judicieux de voir quel est l'apport de leur ajout/combinaison aux 14 autres bandes.

Nous avons six bandes et nous voulons vérifier l'influence de différentes combinaisons possible de leurs sélections sur les performances. Cela représente $2^6 = 64$ modélisations. Nous avons 10 bandes de bases : 3, 6, 21, 25, 40, 53, 57, 85, 129, 160, et six bandes dont nous voulons évaluer l'impact de leur sélection : 14, 15, 16, 17, 74, 95. Le Tableau 4.4 présente les moyennes des indices de performance données par les combinaisons de

longueurs d'ondes issus de la sélection de ces six longueurs d'ondes. La dernière ligne de ce tableau présente les moyennes des performances obtenues par la modélisation excluant ces six bandes, c'est-à-dire en utilisant seulement les 10 bandes de base.

Tableau 4.4

Moyennes des performances données par la sélection des bandes

Indice de la bande	RMSEP	R^2	SDR
14	1.71	0.93	4.05
15	1.73	0.93	4.01
16	1.56	0.94	4.38
17	1.73	0.93	4.05
74	1.87	0.92	3.70
95	1.50	0.95	4.52
Aucune	2.43	0.86	2.79

La modélisation issue uniquement à partir des 10 bandes de base donne des performances en retrait face à celles obtenues avec les combinaisons de sélections de ces six bandes étudiées, et l'écart type des résidus est inférieur à 3. Il semble qu'au moins une des bandes de chaque groupe soit nécessaire pour une bonne modélisation. Nous pouvons observer que les modèles issus de la sélection de la bande 95 ainsi que ceux issus de la sélection de la $16^{ième}$ bande sont les plus performants de chaque groupe.

Des résultats identiques sont obtenus en étudiant le nombre de fois où l'une de ces six longueurs d'ondes est sélectionnée dans un modèle ayant un coefficient de détermination supérieur à 0.94 et un écart type des résidus supérieur à 4.10 (performances obtenues lors de l'analyse toutes bandes). Nous pouvons observer un tel décompte à la Figure 4.2. Nous observons alors que parmi les 4 bandes adjacentes, c'est la $16^{ième}$ qui est le plus souvent

sélectionnée Parmi les deux autres bandes, la 95^{ième} qui apporte les modèles les plus performants.

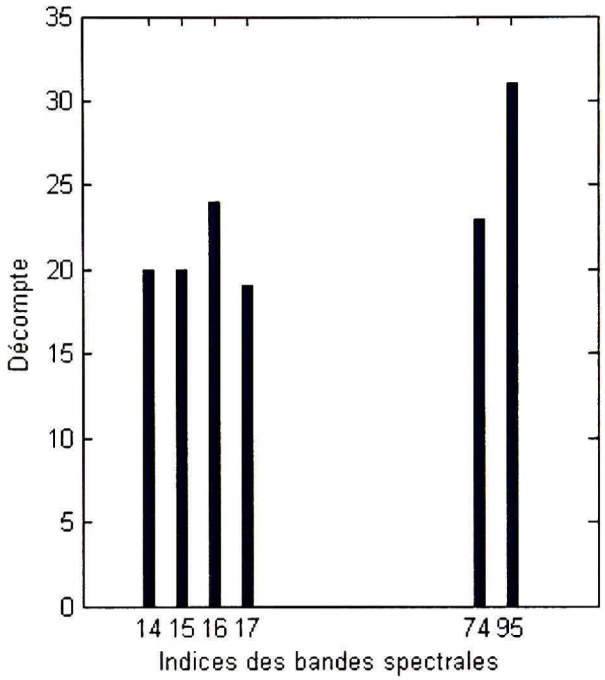

Figure 4.2 ***Décompte des bandes sélectionnées dans les modèles performants.***

Le Tableau 4.5 présente la comparaison des performances données par les différentes configurations de sélection de bandes : les 10 bandes de bases plus la 16^{ième}, soit 11 bandes; les 10 bandes de base plus la 95^{ième}, soit 11 bandes, les 10 bandes de base plus la 16^{ième} et la 95^{ième}, soit 12 bandes, et enfin les 16 bandes d'origine. Puis nous avons joint les performances des modèles utilisant toutes les bandes spectrales.

Tableau 4.5

Moyennes des performances des modèles issus de la sélection de certaines bandes

Choix des bandes	RMSEP	R^2	SDR
Base + 16^{ième} (11 bandes)	1.82 (0.20)	0.92 (0.02)	3.70 (0.37)
Base + 95^{ième} (11 bandes)	1.62 (0.19)	0.94 (0.02)	4.18 (0.53)
Base + 16^{ième} + 95^{ième} (12 bandes)	1.46 (0.17)	0.95 (0.01)	4.64 (0.55)
Base + 6 étudiées (16 bandes)	1.35 (0.17)	0.96 (0.01)	5.03 (0.66)
Toute (163 bandes)	1.65 (0.17)	0.94 (0.01)	4.10 (0.45)

Note : Le nombre entre parenthèses représente l'écart type correspondant.

Nous pouvons remarquer que les performances données par les modèles utilisant 11 bandes et les modèles utilisant toutes les bandes sont identiques. De la même manière, les performances des modèles à 12 et 16 bandes sont à peu près équivalentes.

Pour améliorer encore la sélection de variables, nous pouvons déterminer quel est le nombre minimal de ces 12 bandes spectrales nécessaire pour obtenir les performances minimales acceptables dans notre étude, soit un écart type des résidus (SDR) juste supérieur à 3. Pour cela, nous allons déterminer les modèles correspondants à toutes les combinaisons possibles des 12 bandes, cela correspond à $2^{12}-1=4095$ modèles. Nous ne comptons pas le premier modèle où aucune bande n'est sélectionnée.

Nous présentons les résultats obtenus au Tableau 4.6. Nous avons déterminé les 4095 modèles, et pour chaque nombre de bandes sélectionnées, nous avons déterminé quelle était la combinaison donnant les meilleures performances. Nous avons ainsi relevé son erreur moyenne (RMSEP), son coefficient de détermination moyen (R^2), et son écart type des résidus moyen (SDR).

Tableau 4.6

Moyennes des meilleures performances obtenues pour chaque combinaison de bandes

Nombre de bandes	RMSEP	R^2	SDR	Indices des bandes sélectionnées
1	4.32	0.56	1.55	40
2	3.03	0.79	2.23	25 57
3	2.54	0.85	2.66	21 40 57
4	2.33	0.87	2.89	16 21 40 57
5	2.04	0.90	3.31	40 57 85 95 129
6	1.79	0.92	3.77	16 21 25 40 57 85
7	1.77	0.93	3.84	21 25 40 57 85 95 129
8	1.60	0.94	4.21	16 21 25 40 57 85 95 129
9	1.60	0.94	4.23	3 6 25 40 53 85 95 129 160
10	1.52	0.95	4.44	3 6 16 21 25 40 57 85 95 129
11	1.47	0.95	4.60	3 6 16 21 25 40 53 85 95 129 160
12	1.46	0.95	4.64	3 6 16 21 25 40 53 57 85 95 129 160

D'après ce tableau, nous pouvons remarquer que le nombre minimum de bandes donnant des performances acceptables est de 5 : le coefficient de détermination est de 0.90, et l'écart type des résidus est égal à 3.31. Aussi, nous observons que le nombre de bandes offrant des performances identiques à l'analyse toutes bandes est de 8 : coefficient de détermination égale à 0.94, et écart type des résidus de 4.21.

D'après cette analyse, l'élagage PLS semble être une méthode qui fonctionne bien dans notre étude. Partant de 163 bandes spectrales, nous avons réduit ce nombre à 12 bandes (7.4% des 163 bandes initiales), cela avec de meilleures performances, 8 bandes (4.9% des 163 bandes initiales) avec des performances équivalentes, et 5 bandes (3.01% des 163 bandes initiales) avec les performances minimales acceptables.

4.2.2 Algorithmes de colonie de fourmis

La Figure 4.3 présente les résultats des votes relevés suivant les 50 bases de données aléatoires et leurs bandes sélectionnées correspondantes. Nous pouvons observer que certaines bandes (en rouge) sont plus souvent sélectionnées que d'autres.

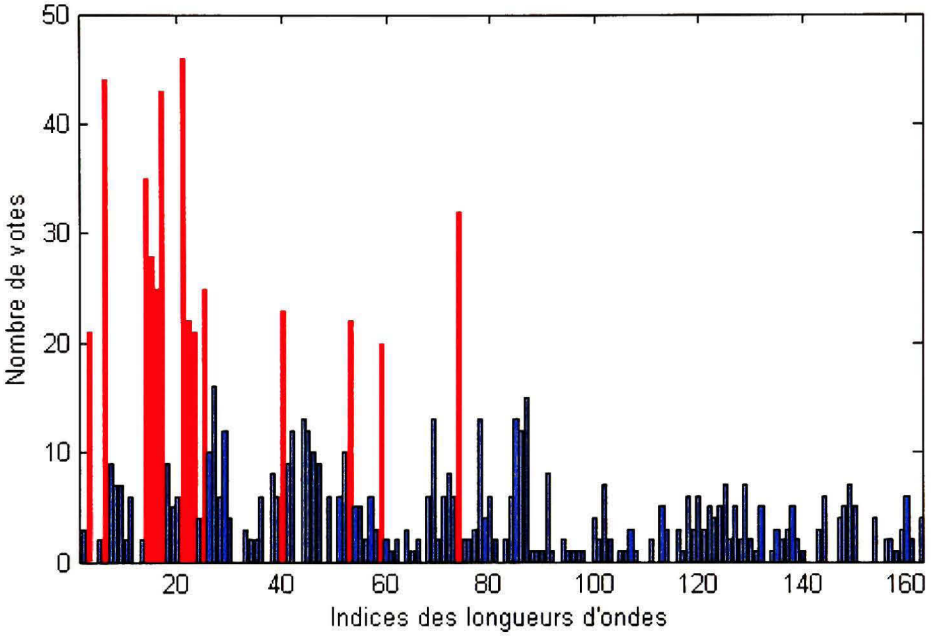

Figure 4.3 ***Résultats des votes pour l'ACF.***

Tableau 4.7

Bandes sélectionnées par l'ACF et leurs longueurs d'ondes correspondantes

Indices des bandes	3	6	14	15	16	17	21
Longueurs d'ondes (nm)	425.58	458.97	486.35	489.27	492.19	495.110	506.88
Indices des bandes	22	23	25	40	53	59	74
Longueurs d'ondes (nm)	509.83	512.78	510.69	563.08	601.40	619.04	663.09

À l'observation de la Figure 4.3, nous avons choisi ces bandes de telle manière que leur nombre de votes soit supérieur à 18. Cela donne un nombre de bandes égal à 15. Ces 15 bandes sont détaillées au Tableau 4.7 Nous pouvons d'emblée remarquer que les longueurs d'ondes proposées par l'ACF sont toutes dans le visible, et très peu de votes peuvent être observés dans le proche infrarouge.

Les moyennes des performances sont présentées au Tableau 4.8. Les performances, bien qu'acceptables (l'écart type des résidus étant supérieur à 3, et le coefficient de détermination supérieur à 0.90), sont en retrait par rapport aux performances toutes bandes.

Tableau 4.8

Moyennes des performances de deux configurations de sélection de variables

Nombre de bandes	RMSEP	R^2	SDR
15	1.96 (0.27)	0.91 (0.02)	3.47 (0.48)
163	1.65 (0.17)	0.94 (0.01)	4.10 (0.45)

Note : Le nombre entre parenthèses représente l'écart type correspondant.

Parmi ces 15 bandes spectrales, nous remarquons que certaines sont encore adjacentes, les $14^{ième}$, $15^{ième}$, $16^{ième}$, et $17^{ième}$, de même que les $21^{ième}$, $22^{ième}$, et $23^{ième}$. Si nous tentons de sélectionner parmi ces deux groupes les bandes qui ont le plus d'influence ($16^{ième}$ et $21^{ième}$), nous obtenons toujours des performances largement en retrait. Coefficient de détermination de 0.89, et un écart type des résidus de 3.17.

En analysant ces résultats, l'ACF ne fonctionne pas très bien dans notre étude. Dans sa description, l'algorithme semble intéressant, mais l'optimisation de ses paramètres intrinsèques est peut-être à considérer de manière plus approfondie, ce qui requiert beaucoup

de temps machine. Nous n'en avons fait qu'une analyse rapide, et cette méthode nécessite un temps de calcul relativement important.

4.3 Comparaison des deux approches

Les deux approches donnent des résultats quelques peu différents, avec des performances en retrait pour l'algorithme de colonie de fourmis. En effet, les bandes sélectionnées par ce dernier se situent toutes dans le visible et sont à peu près similaires au choix fourni par l'élagage, cependant, il manque deux bandes dans le proche infrarouge, bandes qui semblent avoir leur importance.

Nous avons vu que l'ACF requiert certains ajustements de paramètres : deux d'entre eux influencent directement la manière dont l'algorithme convergera. L'étude de ce type d'algorithme n'étant pas l'objectif de notre travail, nous n'avons que sommairement regardé cela de manière à avoir des résultats à priori acceptables. Il semble que certains ajustements doivent être réalisés.

Sur une même machine, le temps de calcul nécessaire pour une ACF est de sept à huit heures, alors que celui de l'élagage PLS dure moins de 10 minutes. Il y a d'autre part des paramètres à ajuster, alors qu'aucun ajustement n'est nécessaire pour l'élagage. Nous en concluons que la méthode utilisant l'élagage PLS est préférable dans notre étude en termes de complexité, temps de calcul, et performance.

4.4 Bénéfices de la sélection de bandes spectrales

Un système de vision tel que celui utilisé dans notre étude, étant capable de décomposer la lumière en plusieurs centaines de bandes spectrales avec une grande précision, est très dispendieux. De plus, il est peu mobile car il requiert des conditions stables d'utilisation, comme par exemple la température, ainsi qu'une certaine puissance électrique pour son fonctionnement, et un ordinateur pour le contrôler et traiter efficacement les données.

Nous venons de le voir, Il n'est pas forcement nécessaire d'utiliser toute l'information contenue dans les 163 bandes spectrales afin d'obtenir des modèles adéquats : huit bandes nous permettent d'avoir des performances identiques à la modélisation toutes bandes, et cinq bandes pour les performances minimales acceptables. Non seulement cette réduction de bandes mène à une modélisation plus simple, mais aussi, elle offre la possibilité de développer un instrument spécifique.

En effet, un appareil utilisant cinq à huit bandes spectrales est plus aisé à construire, et moins onéreux. Plusieurs options peuvent être envisagées : plusieurs capteurs spécifiques (cinq à huit dans notre cas), permettant une acquisition simultanée de l'information, ou bien un seul capteur muni de filtres interférentiels (cinq à huit filtres), permettant l'acquisition en série de l'information. Cet appareil peut alors aisément être portatif, ne requérant que peu de puissance, et donnant presque instantanément le résultat de l'acquisition.

CONCLUSION

Dans ce mémoire nous avons étudié la faisabilité de l'utilisation de l'imagerie hyperspectrale pour prédire la maturité des avocats de manière non destructrice. Nous parlons de faisabilité, car comme nous l'avons vu, nous n'avons aucune information concernant l'origine, les conditions de transport et d'entreposage de nos avocats entre leur récolte et l'étude en laboratoire. Nous sommes seulement certains de leur variété : Hass. À noter que la répartition du taux de matière sèche n'est pas uniforme sur la plage de maturité du fruit, et que nous avons un manque d'échantillons aux environs de 30%. Ce manque de données ainsi que le flou concernant leurs origines ne nous permet donc pas tirer des conclusions très précises pour notre étude. Nous nous sommes donc donné pour objectif de prouver la faisabilité d'une telle approche, et proposer quelques indices pour la poursuite éventuelle de la recherche.

Au delà de cette preuve de concept, nous nous sommes familiarisés avec ce domaine de pointe qu'est l'imagerie hyperspectrale. Nous avons réalisé un montage fonctionnel en tenant compte des spécificités du sujet de notre d'étude, afin de tirer le meilleur parti de ce matériel. Nous avons ainsi établi un protocole expérimental qui pourrait être réutilisé et appliqué à d'autres sujets d'étude. Ce protocole inclut l'utilisation adéquate du système d'imagerie pour l'acquisition d'une scène, ainsi que le traitement des données brutes directement issues de l'imageur.

Nous avons utilisé un outil éprouvé dans le domaine de la vision par ordinateur appliquée à l'agroalimentaire pour la prédiction de nos taux de matière sèche. C'est la technique de la régression linéaire. Les résultats obtenus semblent indiquer qu'il y a un avenir dans l'étude de la maturité de ce fruit tropical par imagerie hyperspectrale. Non seulement le spectre, considéré dans son ensemble, induit de bons résultats en mode absorbance, mais encore il semble possible d'appliquer des concepts de réduction de bandes, menant à une modélisation plus simple et plus efficace. Des deux algorithmes que nous nous sommes proposés d'utiliser, l'élagage PLS semble être le plus intéressant : en effet, son temps d'exécution est nettement plus faible que celui de l'ACF, et il ne nécessite aucuns ajustements de paramètres.

Nous avions 163 bandes spectrales initiales, et nous avons réduit ce nombre à 12 bandes (7.4% bandes initiales) avec de meilleures performances, 8 bandes (4.9% bandes initiales) pour des performances équivalentes, et 5 bandes (3.01% bandes initiales) avec les performances minimales acceptables.

La réduction du nombre de bandes peut permettre le développement d'outils portables efficaces et peu onéreux, permettant une analyse aisée sur le terrain (dans l'arbre ou dans les centres de tri par exemple).

RECOMMANDATIONS

Comme nous l'avons souligné dans la conclusion, les résultats prometteurs donnés par cette étude ne nous permettent pas de tirer de conclusions définitives. À la suite de l'expérience que nous avons acquise, nous pouvons conseiller ces quelques voies de recherche :

- Avoir le maximum de variabilité ainsi qu'une répartition adéquate des données issues des mesures de taux de matière sèche.

- Disposer d'une source fiable d'avocats afin de garantir des conditions identiques de production, d'entreposage et de transport entre la récolte et l'analyse des fruits.

- D'après les résultats statistiques que nous avons obtenus sur notre base de données, une étude portant sur 24 fruits permet d'avoir un coefficient de variabilité de 10%, ce qui, dans le domaine agroalimentaire, est une valeur acceptable. Si l'objectif est un coefficient de 5%, il faudra 95 fruits.

- Effectuer 4 mesures de matière sèche aux quatre points cardinaux du fruit par rapport au pédoncule, et réaliser l'acquisition des quatre mesures hyperspectrales correspondantes.

- Étudier d'autres outils permettant la prédiction du taux de matière sèche en fonction du spectre : par exemple les algorithmes évolutionnaires comme la programmation génétique. Là où nous ne réalisons qu'une combinaison linéaire des bandes en les considérant indépendamment les unes des autres, il peut être intéressant d'étudier les combinaisons des bandes spectrales entre elles de manière linéaire ou non.

ANNEXE I

EXTRACTION DE LA MATIÈRE SÈCHE DES AVOCATS

Nous allons décrire en détail la méthode que nous avons utilisée afin d'extraire la matière sèche des avocats avec l'utilisation d'une étuve chauffée à 105°C. Sachant que la matière sèche au sein d'un avocat n'est pas constante (*Voir* Figure 1.1), il est important de faire un prélèvement de chair dans le fruit qui soit représentatif de cette variation. C'est pourquoi lors de nos manipulations, nous extrayons la matière sèche des avocats à quatre endroits, correspondant aux quatre points cardinaux de l'avocat par rapport à son pédoncule, suivant la longueur du fruit.

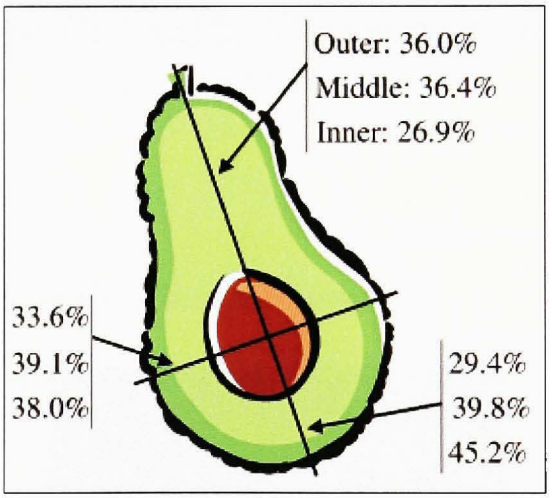

Figure 1.1 ***Répartition non uniforme du taux de matière sèche au sein d'un avocat.***
(Tiré de Woolf et al., 2003)

Source : Cette figure sous forme de graphique est tirée de l'article de M. Allan Woolf et al., *Measuring Avocado maturity; ongoing developments*, p. 1.

Tout d'abord, nous avons besoin du matériel suivant :

- Étuve à 105°C
- Balance avec une précision de 0.01g

- Couteau et planche à découper
- Plats allant au four

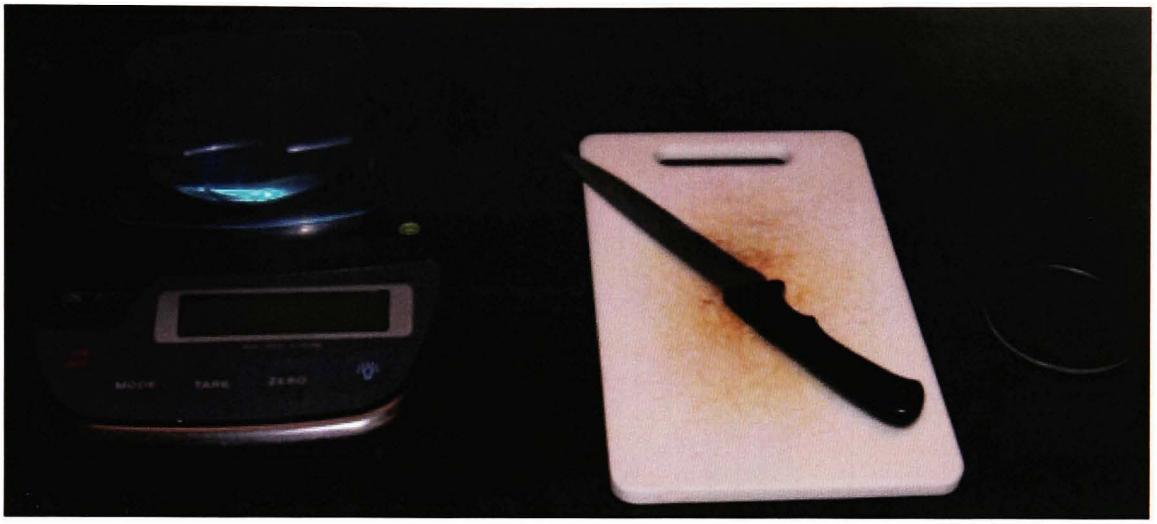

Figure 1.2 ***Matériel nécessaire.***

Voici les étapes à suivre :

1. Peser le plat à vide, et relever la masse obtenue :

Figure 1.3 ***Pesée à vide.***

2. A l'aide du couteau et de la planche à découper, découper l'avocat en quatre quarts :

Figure 1.4 ***Découpe des quarts d'avocat.***

3. Une fois ces quarts obtenus, découper de fines tranches selon la longueur du fruit, et retirer la peau ainsi que le noyau à l'aide du couteau :

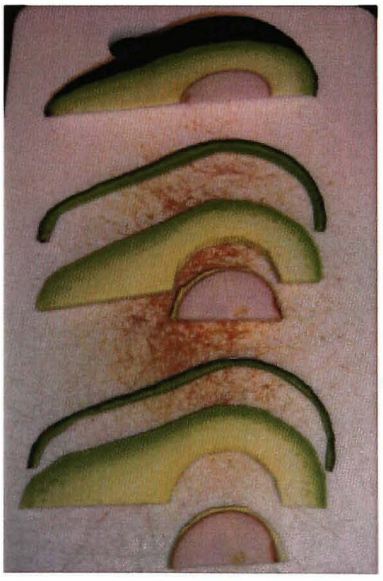

Figure 1.5 ***Découpe des tranches d'avocat.***

4. Découper les tranches en petits morceaux, et les placer dans le plat de manière à avoir une masse de $10g \pm 1g$, et relever la masse ainsi obtenue :

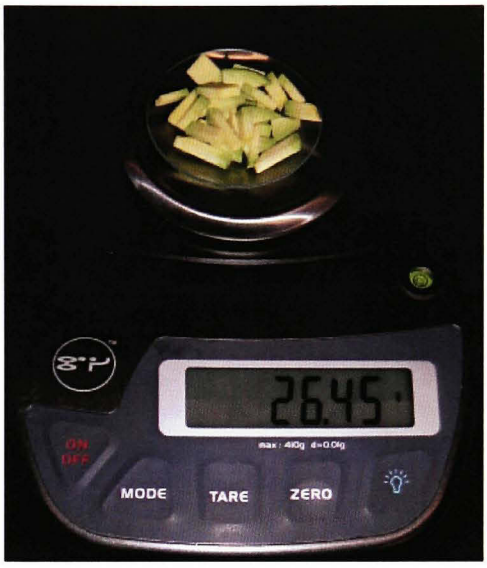

Figure 1.6 ***Pesée de la chaire fraiche.***

5. Effectuer ces quatre premières étapes pour les trois échantillons restants du fruit et éventuellement pour les autres fruits à étudier.

6. Une fois que tous les échantillons sont préparés, les mettre au four (noter l'heure) :

Figure 1.7 ***Échantillons frais dans le four.***

7. Vérifier de temps à autre que le chauffage s'effectue correctement (constance de la température du four). Sa durée est de cinq heures. Sortir les échantillons du four, et effectuer leur pesée :

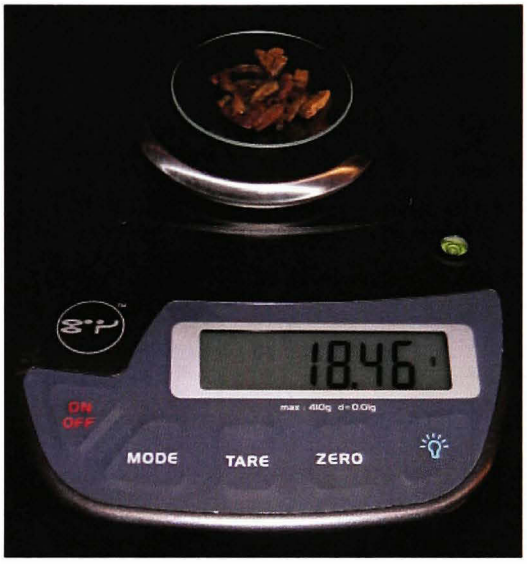

Figure 1.8 ***Pesée de la chaire chauffée.***

8. Calculer les taux de matière sèche ainsi obtenus avec la formule :

$$\text{matière sèche } (\%) = \frac{C-A}{B-A} \times 100 = \frac{\text{masse de matière sechée}}{\text{masse de matière fraiche}} \times 100$$

A : masse du plat à vide.

B : masse du plat contenant l'échantillon frais.

C : masse du plat contenant l'échantillon après chauffage.

Dans notre exemple, nous avons :

$$\text{matière sèche } (\%) = \frac{18.46-16.43}{26.45-16.43} \times 100 = \frac{2.03}{10.02} \times 100 = 20.26\%$$

ANNEXE II

BASE DE DONNEES D'AVOCATS

Tableau 2.1

Base de données

	Taux de matière sèche (%)	Moyenne	Minimum Maximum	Étendue	Écart type	Erreur type	Coefficient de variation (%)	Intervalle de confiance à 95% (±)
hass_1-1	38.04	35.44	33.73 38.04	4.31	1.89	0.95	5.34	3.01
hass_1-2	34.42							
hass_1-3	33.73							
hass_1-4	35.57							
hass_2-1	41.85	39.78	38.64 41.85	3.21	1.44	0.72	3.63	2.30
hass_2-2	38.98							
hass_2-3	39.64							
hass_2-4	38.64							
hass_3-1	36.15	35.99	35.40 36.86	1.46	0.67	0.33	1.86	1.06
hass_3-2	35.40							
hass_3-3	35.53							
hass_3-4	36.86							
hass_4-1	42.51	39.68	36.72 42.51	5.79	2.42	1.21	6.09	3.84
hass_4-2	40.34							
hass_4-3	36.72							
hass_4-4	39.13							

	Taux de matière sèche (%)	Moyenne	Minimum Maximum	Étendue	Écart type	Erreur type	Coefficient de variation (CV%)	Intervalle de confiance à 95% (±)
hass_5-1	35.51	34.47	33.18 35.51	2.33	1.01	0.51	2.93	1.61
hass_5-2	34.97							
hass_5-3	33.18							
hass_5-4	34.21							
hass_6-1	35.43	33.94	32.75 35.43	2.68	1.17	0.58	3.45	1.86
hass_6-2	33.33							
hass_6-3	32.75							
hass_6-4	34.25							
hass_7-1	35.05	34.32	33.50 35.05	1.55	0.80	0.40	2.34	1.28
hass_7-2	33.76							
hass_7-3	34.97							
hass_7-4	33.5							
hass_8-1	27.88	26.69	26.01 27.88	1.87	0.82	0.41	3.06	1.30
hass_8-2	26.42							
hass_8-3	26.46							
hass_8-4	26.01							
hass_9-1	36.48	34.81	33.09 36.48	3.39	1.38	0.69	3.98	2.20
hass_9-2	34.84							
hass_9-3	34.82							
hass_9-4	33.09							
hass_10-1	29.61	27.29	25.87 29.61	3.74	1.73	0.86	6.33	2.75
hass_10-2	25.87							
hass_10-3	26.08							
hass_10-4	27.59							
hass_11-1	22.55	22.16	21.82 22.55	0.73	0.31	0.16	1.40	0.49
hass_11-2	22.23							
hass_11-3	21.82							
hass_11-4	22.03							

	Taux de matière sèche (%)	Moyenne	Minimum Maximum	Étendue	Écart type	Erreur type	Coefficient de variation (CV%)	Intervalle de confiance à 95% (±)
hass_12-1	22.32	22.13	21.86 22.32	0.46	0.21	0.11	0.95	0.34
hass_12-2	22.06							
hass_12-3	22.27							
hass_12-4	21.86							
hass_13-1	21.2	21.30	21.20 21.41	0.21	0.09	0.04	0.41	0.14
hass_13-2	21.28							
hass_13-3	21.31							
hass_13-4	21.41							
hass_14-1	24.66	24.24	23.89 24.66	0.77	0.34	0.17	1.42	0.55
hass_14-2	23.89							
hass_14-3	24.04							
hass_14-4	24.37							
hass_15-1	26.39	26.10	25.62 26.39	0.77	0.35	0.18	1.35	0.56
hass_15-2	25.62							
hass_15-3	26.06							
hass_15-4	26.34							
hass_16-1	23.99	23.82	23.57 24.16	0.59	0.30	0.15	1.26	0.48
hass_16-2	23.57							
hass_16-3	23.57							
hass_16-4	24.16							
hass_17-1	24.33	23.92	23.41 24.33	0.92	0.42	0.21	1.74	0.66
hass_17-2	24.17							
hass_17-3	23.41							
hass_17-4	23.75							
hass_18-1	22.01	21.80	21.16 22.03	0.87	0.43	0.21	1.95	0.68
hass_18-2	22.03							
hass_18-3	21.16							
hass_18-4	21.99							

	Taux de matière sèche (%)	Moyenne	Minimum Maximum	Étendue	Écart type	Erreur type	Coefficient de variation (CV%)	Intervalle de confiance à 95% (±)
hass_19-1	24.25	23.80	23.36 24.25	0.89	0.37	0.18	1.54	0.58
hass_19-2	23.83							
hass_19-3	23.36							
hass_19-4	23.74							
hass_20_1	20.44	20.30	19.81 20.69	0.88	0.37	0.19	1.83	0.59
hass_20_2	20.69							
hass_20_3	19.81							
hass_20_4	20.27							
hass_21-1	22.06	22.11	21.98 22.34	0.36	0.16	0.08	0.72	0.25
hass_21-2	21.98							
hass_21-3	22.05							
hass_21-4	22.34							

ANNEXE III

CALCUL DES COMPOSANTES RGB A PARTIR DU SPECTRE

De nos jours, la plupart des applications numériques modélisent la couleur comme la combinaison de trois composantes (RGB, HSI). Alors que le fonctionnement de l'œil humain est assez bien modélisé par cette approche, l'interaction entre la lumière et la matière, ainsi que les véritables mécanismes de perception de la vision humaine, sont beaucoup plus compliqués. Cependant, lorsque le spectre issu de la lumière provenant d'un objet est connu de manière précise, la perception de ce spectre par l'œil aux moyens de nos modèles doit être déterminée. Il nous faut donc calculer les composantes (RGB, HSI) issues de ce spectre.

Un spectre est une fonction $P(\lambda)$ qui, sur une plage de longueurs d'ondes donnée, nous informe sur la répartition de la puissance des radiations lumineuses, en watt/nm. Nous déterminerons les valeurs du tri-stimulus $[X,Y,Z]$ à partir de cette répartition. Ce tri-stimulus caractérise la perception standard de la couleur par un observateur humain. Enfin, à partir de ce tri-stimulus, nous calculerons les valeurs des trois composantes RGB correspondantes.

Le système colorimétrique $[X,Y,Z]$ a été défini par la Commission Internationale de l'Éclairage (CIE) en 1931, afin de contourner les limitations du modèle RGB qui ne peut pas reproduire toutes les couleurs du spectre. L'ensemble des couleurs du spectre visible peut être recomposé par ces trois primaires, $[X,Y,Z]$, pondérées par des coefficients appelés fonctions colorimétriques. Ce sont les fonctions $\overline{x}$, $\overline{y}$, $\overline{z}$ (aussi appelées CMF, Color Matching Functions). Elles donnent la relative contribution de la lumière, selon les longueurs d'ondes, aux tri-stimulus $[X,Y,Z]$ définis par la C.I.E. (Walker, 1996).

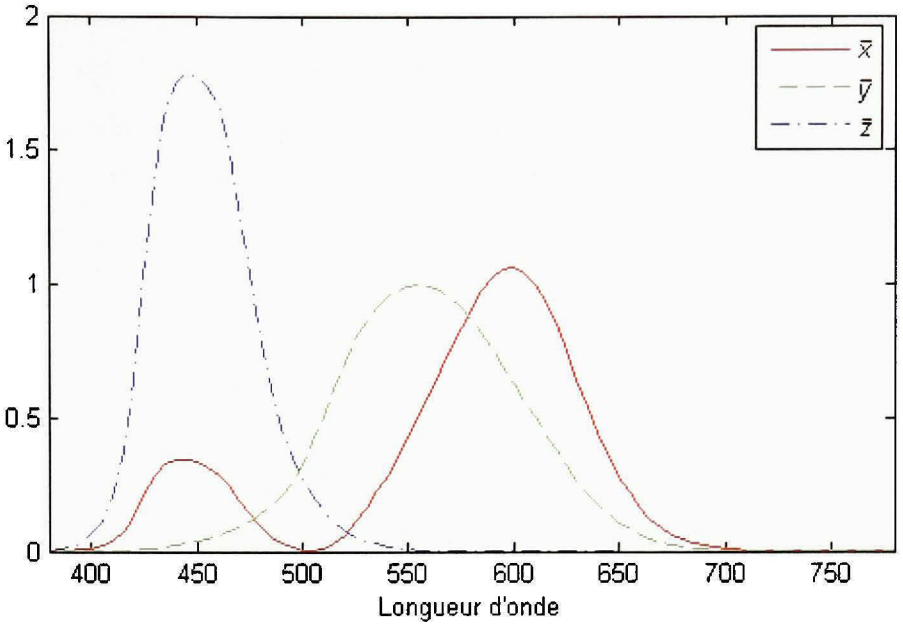

Figure 3.1 ***Fonctions colorimétriques CIE 1931.***

(Tiré de Smith et Guild, 1931)

Source : Cette figure sous forme de graphique est tirée de des données présentées sous forme de tableau dans l'article de M. M. A. Smith et al., *The C.I.E. colorimetric standards and their use* , p. 103-105.

Ces fonctions colorimétriques que nous pouvons observer à la Figure 3.1, ont été déterminées en mesurant la perception moyenne de la couleur dans le domaine du visible (de 380nm à 780nm), par un groupe d'observateurs humains. Afin de calculer de manière théorique les valeurs du tri-stimulus $[X,Y,Z]$ à partir d'un spectre émissif $P(\lambda)$, nous calculons l'intégrale, dans le domaine du visible, du produit des CMF par le spectre :

$$\begin{cases} X = \int \bar{x}(\lambda) P(\lambda) d\lambda \\ Y = \int \bar{y}(\lambda) P(\lambda) d\lambda \\ Z = \int \bar{z}(\lambda) P(\lambda) d\lambda \end{cases}$$

Ces équations sont valables dans le cas d'un spectre émissif, c'est-à-dire un spectre provenant d'une source lumineuse. Dans notre cas, le spectre est issu de la réflexion de la lumière sur

un objet. Nous remplaçons donc la fonction $P(\lambda)$, par des valeurs de réflectance $R(\lambda)$ multipliées par la fonction de répartition de référence de l'illuminant utilisé $I(\lambda)$ (Lindbloom, 2003). Nous incluons l'illuminant car celui-ci influe sur la couleur perçue de l'objet éclairé. Dans notre cas, l'illuminant utilisé est le A (*Voir* 2.3.2.1, et Figure 2.3). Les équations deviennent :

$$\begin{cases} X = \frac{1}{N}\int \bar{x}(\lambda)R(\lambda)I(\lambda)d\lambda \\ Y = \frac{1}{N}\int \bar{y}(\lambda)R(\lambda)I(\lambda)d\lambda \\ Z = \frac{1}{N}\int \bar{z}(\lambda)R(\lambda)I(\lambda)d\lambda \\ N = \int \bar{y}(\lambda)I(\lambda)d\lambda \end{cases}$$

Nous connaissons les spectres de manière empirique, c'est-à-dire mesurés expérimentalement. Nous ne connaissons pas les équations mathématiques qui représentent des spectres, mais seulement leurs valeurs discrètes, ce qui implique que le signe intégral est remplacé par un signe de sommation, Nous limitons les bornes à 380nm et 780nm, car à l'extérieur les fonctions colorimétriques sont nulles. Dans notre cas :

$$\begin{cases} X = \frac{1}{N}\sum_{\lambda=380}^{780} \bar{x}(\lambda)P(\lambda)I(\lambda) \\ Y = \frac{1}{N}\sum_{\lambda=380}^{780} \bar{y}(\lambda)P(\lambda)I(\lambda) \\ Z = \frac{1}{N}\sum_{\lambda=380}^{780} \bar{z}(\lambda)P(\lambda)I(\lambda) \\ N = \sum_{\lambda=380}^{780} \bar{y}(\lambda)I(\lambda) \end{cases}$$

Pour passer du tri-stimulus $[X,Y,Z]$ aux trois composantes RGB, nous utilisons les équations de conversion suivantes :

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 2.37067 & -0.513885 & 0.00529818 \\ -0.900040 & 1.42530 & -0.0146949 \\ -0.470634 & 0.0885814 & 1.00940 \end{bmatrix} \times \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

La plage de valeur pour le système RGB est $[0,1]$. Noter que certaines valeurs sont négatives dans la matrice de transformation, induisant des valeurs négatives ou supérieures à un pour le triplet RGB : la totalité des couleurs visibles n'est pas reproductible dans le système RGB.

ANNEXE IV

PERFORMANCES OBTENUES POUR DES MODELES TOUTES BANDES

Tableau 3.1

Performances des modèles en mode réflectance et validation croisée *Leave-One-Out*

N°Modèle	Nb. Variables Latentes	RMSEC	RMSEP	R^2	SDR
1	15	0.8572	2.1944	0.8894	3.3141
2	12	1.1101	2.5075	0.8576	2.8860
3	17	0.5348	2.6648	0.8368	2.9432
4	17	0.5025	2.7311	0.8299	2.5444
5	28	0.0507	2.9047	0.8071	2.4809
6	19	0.4408	2.1998	0.8907	3.2732
7	49	54.4097	80.9074	0.0000	0.6192
8	36	0.0225	2.1851	0.8825	2.9405
9	36	0.0028	2.2572	0.8847	3.4105
10	16	0.5411	2.6356	0.8421	2.5968
11	17	0.5949	1.8612	0.9184	3.6476
12	16	0.6566	2.2300	0.8935	2.8941
13	50	58.7008	54.1576	0.0000	0.6367
14	26	0.0474	2.7213	0.8187	2.2590
15	19	0.3477	2.4201	0.8558	2.7670
16	18	0.4275	3.2128	0.7758	2.1598
17	29	0.0342	2.2481	0.8846	3.3632
18	19	0.3817	2.4327	0.8590	2.5279
19	19	0.3719	2.5751	0.8378	1.9739
20	19	0.3717	2.5293	0.8427	2.7520
21	22	0.1320	2.8392	0.8196	2.4499
22	51	18.2308	26.1167	0.0000	1.0878
23	20	0.3706	1.9592	0.9076	3.2430
24	16	0.6876	2.1758	0.8931	3.6042
25	15	0.8164	2.0529	0.9035	3.5741

N°Modèle	Nb. Variables Latentes	RMSEC	RMSEP	R^2	SDR
26	8	1.5912	2.6571	0.8329	2.3717
27	21	0.2180	2.6085	0.8461	2.6757
28	15	0.7597	2.3765	0.8650	3.0125
29	20	0.2883	2.6869	0.8293	2.3706
30	19	0.3628	2.5099	0.8544	3.0698
31	52	150.4038	163.2507	0.0000	0.3170
32	15	0.8042	2.4577	0.8586	2.6368
33	18	0.4461	2.2045	0.8883	3.3490
34	15	0.6408	2.6377	0.8444	2.5542
35	28	0.0472	2.3704	0.8702	2.8117
36	15	0.8233	2.3770	0.8654	2.8960
37	19	0.3210	2.0206	0.9054	3.3992
38	18	0.3783	3.0148	0.7765	2.0387
39	19	0.3379	2.2131	0.8850	3.0248
40	19	0.3347	2.6732	0.8315	2.0487
41	17	0.5138	2.3251	0.8765	2.9821
42	21	0.2336	2.3123	0.8818	2.4533
43	20	0.3197	2.0197	0.9009	3.5037
44	21	0.2346	2.2989	0.8690	2.3706
45	14	0.9076	2.1465	0.8910	2.6112
46	19	0.2545	2.7910	0.8235	2.7161
47	22	0.1804	3.1903	0.7505	1.9175
48	19	0.3689	2.2853	0.8858	2.5334
49	20	0.3221	2.0699	0.8970	3.0798
50	18	0.3866	2.4286	0.8593	2.3975

Moyenne	21.86	6.0425	8.7329	0.7904	2.6218
Écart type	9.924	23.7107	26.0599	0.2381	0.7361

Tableau 3.2

Performances des modèles en mode réflectance et validation croisée 5-blocs

N°Modèle	Nb. Variables Latentes	RMSEC	RMSEP	R^2	SDR
1	42	0.1279	2.4142	0.8661	3.0055
2	14	0.7771	2.8218	0.8197	2.5289
3	43	0.9164	2.7767	0.8228	2.9503
4	17	0.5025	2.7311	0.8299	2.5444
5	18	0.4704	2.5448	0.8519	2.6463
6	16	0.7079	1.9752	0.9119	3.4806
7	17	0.5322	2.2835	0.8813	3.0158
8	20	0.2754	2.2606	0.8742	2.9534
9	43	0.0298	2.3009	0.8802	3.3581
10	21	0.2607	2.7007	0.8342	2.5688
11	20	0.3580	2.0558	0.9004	3.3208
12	8	1.6879	2.2298	0.8935	2.9083
13	8	1.9745	1.7539	0.9280	3.8484
14	18	0.3453	2.5509	0.8407	2.2866
15	12	1.1708	2.7152	0.8185	2.4810
16	8	1.6910	2.5809	0.8553	2.4126
17	29	0.0342	2.2481	0.8846	3.3632
18	7	2.0014	1.8700	0.9167	3.3418
19	13	0.9959	2.9413	0.7884	1.6033
20	9	1.6155	2.3767	0.8611	2.5915
21	22	0.1320	2.8392	0.8196	2.4499
22	42	0.0294	2.4458	0.8575	2.8700
23	7	1.8135	2.1080	0.8931	2.9733
24	8	1.8317	2.2595	0.8847	3.3896
25	6	2.1084	2.3890	0.8693	2.9251
26	7	1.7382	3.1071	0.7716	2.0081
27	7	1.9173	2.3399	0.8762	2.6262
28	8	1.8973	1.9257	0.9114	3.8147
29	8	1.6678	2.4371	0.8595	2.3365
30	43	0.4149	2.8954	0.8062	2.7983

N°Modèle	Nb. Variables Latentes	RMSEC	RMSEP	R^2	SDR
31	8	1.9172	1.9916	0.9060	3.3980
32	6	2.4138	2.2862	0.8776	3.2242
33	17	0.5217	2.2471	0.8839	3.2757
34	8	1.5866	2.5117	0.8589	2.4918
35	17	0.5488	2.2608	0.8819	3.1972
36	7	2.0137	1.7117	0.9302	3.7477
37	43	0.9045	3.7278	0.6780	1.8974
38	22	0.1717	2.9302	0.7889	2.0955
39	22	0.1641	2.3153	0.8741	2.9194
40	21	0.2450	2.7029	0.8278	1.9957
41	8	1.8591	1.9758	0.9108	3.7198
42	19	0.3484	2.2707	0.8860	2.5253
43	20	0.3197	2.0197	0.9009	3.5037
44	21	0.2346	2.2989	0.8690	2.3706
45	16	0.7042	2.1528	0.8903	2.6292
46	8	1.8628	2.0309	0.9066	3.5284
47	22	0.1804	3.1903	0.7505	1.9175
48	18	0.4403	2.3364	0.8806	2.4876
49	7	1.8938	2.3568	0.8665	2.6187
50	14	0.9253	2.5196	0.8486	2.2306
Moyenne	17.3000	0.9856	2.4143	0.8605	2.8235
Écart type	11.1140	0.7483	0.3884	0.0479	0.5493

Tableau 3.3

Performances des modèles en mode réflectance et diagrammes de Pareto

N° Modèle	Nb. Variables Latentes	RMSEC	RMSEP	R^2	SDR
1	13	1.0805	2.0852	0.9001	3.3592
2	13	0.9303	2.7383	0.8302	2.5877
3	13	0.9741	2.7440	0.8270	2.9933
4	13	0.9051	2.5543	0.8512	2.5981
5	13	1.0459	2.6903	0.8345	2.5306
6	13	1.0590	2.1747	0.8932	3.1329
7	13	1.0883	2.4572	0.8625	2.9257
8	13	1.0571	2.4324	0.8543	2.8859
9	13	1.1367	2.2541	0.8851	3.1713
10	12	1.1190	2.5615	0.8509	2.4635
11	13	1.0678	1.8649	0.9180	3.8818
12	13	1.0119	2.1986	0.8965	3.0351
13	13	1.1203	2.2225	0.8845	3.2347
14	12	1.0991	2.4171	0.8570	2.3872
15	13	1.0548	2.6844	0.8226	2.5337
16	13	0.9721	3.2063	0.7767	2.0545
17	12	1.0250	2.5509	0.8514	2.9846
18	13	1.0362	2.3732	0.8659	2.6300
19	13	0.9959	2.9413	0.7884	1.6033
20	13	0.9777	2.7865	0.8091	2.2560
21	12	1.0858	2.4583	0.8648	2.6289
22	13	0.9276	2.8501	0.8064	2.4274
23	13	0.9834	2.3481	0.8673	2.7193
24	13	1.0183	2.3979	0.8701	3.2377
25	13	1.1108	2.2444	0.8847	3.2340
26	13	0.9564	2.5283	0.8487	2.4791
27	13	1.0221	2.5262	0.8557	2.6652
28	13	1.0401	2.4757	0.8535	2.9659
29	13	0.9772	2.5701	0.8438	2.2750
30	13	1.1721	2.4286	0.8636	3.2177

N°Modèle	Nb. Variables Latentes	RMSEC	RMSEP	R^2	SDR
31	13	0.9583	2.2241	0.8828	3.2646
32	13	1.0911	2.2149	0.8852	2.9220
33	13	0.9610	2.6685	0.8363	2.8106
34	12	0.9757	2.5577	0.8537	2.6223
35	13	0.9608	2.0845	0.8996	3.5653
36	13	1.0682	2.4057	0.8621	2.8531
37	13	1.1625	2.3039	0.8770	2.8121
38	13	1.0125	3.0105	0.7772	1.8227
39	13	1.0994	2.0984	0.8966	3.1304
40	13	1.0644	2.4990	0.8528	2.3096
41	13	1.0413	2.5506	0.8514	2.8744
42	13	0.9881	2.4153	0.8710	2.3701
43	13	1.0433	2.1447	0.8882	3.1780
44	13	1.0518	2.3711	0.8607	2.4054
45	13	1.0123	2.2044	0.8850	2.6370
46	13	0.9377	2.8008	0.8223	2.7330
47	13	0.8570	2.8769	0.7971	2.2076
48	13	1.0812	1.9147	0.9198	3.1325
49	13	1.0073	2.2059	0.8831	2.8555
50	13	1.0791	2.5760	0.8417	2.1253
Moyenne	12.9000	1.0301	2.4579	0.8572	2.7546
Écart type	0.3030	0.0674	0.2783	0.0334	0.4413

Tableau 3.4

Performances des modèles en mode absorbance et validation croisée *Leave-One-Out*

N°Modèle	Nb. Variables Latentes	RMSEC	RMSEP	R^2	SDR
1	12	0.8543	1.6285	0.9391	4.2372
2	12	0.9190	1.4616	0.9516	4.5779
3	11	1.0149	1.6645	0.9363	4.4690
4	12	0.8809	1.8063	0.9256	3.3123
5	14	0.6135	1.4759	0.9502	4.6110
6	16	0.5707	1.5799	0.9436	4.3824
7	16	0.6286	1.6844	0.9354	4.0495
8	11	1.0383	1.7206	0.9271	3.9977
9	12	0.8429	1.7602	0.9299	4.1455
10	14	0.6179	1.8322	0.9237	3.8647
11	22	0.1792	2.7923	0.8163	2.3672
12	14	0.5767	1.6803	0.9395	3.8823
13	11	1.0723	1.4935	0.9478	4.2352
14	11	1.0101	1.8414	0.9170	3.6709
15	15	0.5878	1.8951	0.9116	3.9773
16	16	0.4486	1.8952	0.9220	3.5926
17	13	0.7239	1.8492	0.9219	3.8203
18	12	0.9184	1.2672	0.9618	4.8415
19	15	0.5317	1.7795	0.9225	3.1695
20	52	54.3869	70.0159	0.0000	0.5759
21	11	0.9493	1.9145	0.9180	3.3164
22	11	0.9407	1.7488	0.9271	3.8830
23	12	0.9245	1.4553	0.9490	4.3559
24	12	0.8718	1.7264	0.9327	4.0201
25	15	0.5722	1.5489	0.9451	4.6475
26	11	1.0838	1.3278	0.9583	4.8264
27	12	0.8639	1.6490	0.9385	4.1084
28	13	0.7379	1.5936	0.9393	3.9595
29	14	0.7214	1.5419	0.9438	4.1272
30	15	0.5755	1.7662	0.9279	4.2489

N°Modèle	Nb. Variables Latentes	RMSEC	RMSEP	R^2	SDR
31	13	0.8113	1.3339	0.9578	5.1387
32	12	0.9671	1.3348	0.9583	4.6619
33	11	0.9649	1.8469	0.9216	4.0470
34	44	0.3330	2.1874	0.8930	2.9078
35	12	0.9066	1.3710	0.9566	5.1807
36	12	0.8257	2.0050	0.9042	3.5426
37	12	0.9228	1.5023	0.9477	4.6200
38	12	0.8898	1.7803	0.9221	3.4851
39	15	0.5651	1.9205	0.9134	3.4258
40	15	0.6720	1.4910	0.9476	4.3530
41	11	1.0089	1.8315	0.9234	3.6619
42	13	0.7521	1.7059	0.9357	3.6536
43	11	1.0576	1.5253	0.9435	4.3333
44	12	0.9801	1.4251	0.9497	4.6556
45	12	0.8438	1.9043	0.9142	3.4108
46	13	0.7743	2.1774	0.8926	3.3326
47	12	0.8005	2.0641	0.8955	3.3398
48	14	0.6271	1.6078	0.9435	4.0339
49	11	0.9494	1.8179	0.9206	3.6527
50	11	0.9775	1.8424	0.9190	3.3115
Moyenne	14.3000	1.8657	3.0814	0.9112	3.9204
Écart type	7.2822	7.5818	9.6628	0.1336	0.7496

Tableau 3.5

Performances des modèles en mode absorbance et validation croisée 5-blocs

N°Modèle	Nb. Variables Latentes	RMSEC	RMSEP	R^2	SDR
1	12	0.8543	1.6285	0.9391	4.2372
2	13	0.8295	1.6431	0.9389	4.1074
3	11	1.0149	1.6645	0.9363	4.4690
4	12	0.8809	1.8063	0.9256	3.3123
5	13	0.7175	1.5883	0.9423	4.1980
6	11	1.1648	1.2846	0.9627	5.1834
7	14	0.7021	1.5040	0.9485	4.4978
8	11	1.0383	1.7206	0.9271	3.9977
9	15	0.6190	1.5726	0.9441	4.5488
10	13	0.7158	1.7939	0.9269	3.7849
11	13	0.7989	1.9115	0.9139	3.4999
12	13	0.6580	1.7262	0.9362	3.8701
13	12	0.8960	1.5132	0.9464	4.3083
14	12	0.9298	1.6128	0.9363	4.1533
15	12	0.8209	1.7966	0.9206	4.1913
16	17	0.4032	1.9072	0.9210	3.5806
17	13	0.7239	1.8492	0.9219	3.8203
18	11	1.1505	1.3500	0.9566	4.4713
19	13	0.7225	1.7923	0.9214	3.2139
20	11	0.9738	1.6386	0.9340	4.0149
21	15	0.5320	1.5711	0.9448	4.1210
22	11	0.9407	1.7488	0.9271	3.8830
23	12	0.9245	1.4553	0.9490	4.3559
24	12	0.8718	1.7264	0.9327	4.0201
25	13	0.7957	1.7674	0.9285	4.1362
26	12	0.9122	1.2948	0.9603	4.9033
27	13	0.7853	1.6825	0.9360	4.0366
28	13	0.7379	1.5936	0.9393	3.9595
29	12	0.8844	1.4718	0.9488	4.3152
30	13	0.7441	1.6365	0.9381	4.6138

N°Modèle	Nb. Variables Latentes	RMSEC	RMSEP	R^2	SDR
31	14	0.7193	1.4078	0.9530	4.8563
32	12	0.9671	1.3348	0.9583	4.6619
33	12	0.8270	1.8145	0.9243	4.1531
34	16	0.4280	1.8529	0.9232	3.7317
35	12	0.9066	1.3710	0.9566	5.1807
36	15	0.5924	1.7088	0.9304	3.9013
37	12	0.9228	1.5023	0.9477	4.6200
38	14	0.7449	1.5971	0.9373	3.8206
39	13	0.7024	1.8149	0.9226	3.5236
40	13	0.7982	1.5522	0.9432	4.1710
41	16	0.5696	1.7147	0.9328	3.9123
42	13	0.7521	1.7059	0.9357	3.6536
43	11	1.0576	1.5253	0.9435	4.3333
44	12	0.9801	1.4251	0.9497	4.6556
45	15	0.6090	1.8150	0.9220	3.6525
46	12	0.8636	1.8981	0.9184	3.7742
47	14	0.6803	1.9330	0.9084	3.5232
48	14	0.6271	1.6078	0.9435	4.0339
49	13	0.6966	1.7534	0.9261	3.8193
50	11	0.9775	1.8424	0.9190	3.3115
Moyenne	12.8400	0.8033	1.6486	0.9360	4.1019
Écart type	1.4337	0.1658	0.1717	0.0129	0.4548

Tableau 3.6

Performances des modèles en mode absorbance et diagrammes de Pareto

N°Modèle	Nb. Variables Latentes	RMSEC	RMSEP	R^2	SDR
1	13	0.7533	1.7685	0.9281	4.1356
2	14	0.7617	1.6685	0.9370	4.0922
3	14	0.7601	1.6789	0.9352	4.5580
4	13	0.7526	1.8077	0.9255	3.3675
5	13	0.7175	1.5883	0.9423	4.1980
6	14	0.6784	1.4724	0.9510	4.5486
7	14	0.7021	1.5040	0.9485	4.4978
8	14	0.7537	1.8177	0.9187	3.7559
9	13	0.7628	1.6619	0.9375	4.3732
10	13	0.7158	1.7939	0.9269	3.7849
11	13	0.7989	1.9115	0.9139	3.4999
12	13	0.6580	1.7262	0.9362	3.8701
13	14	0.7190	1.5218	0.9458	4.4620
14	14	0.7786	1.7417	0.9257	3.8951
15	13	0.7506	1.9528	0.9061	3.9060
16	13	0.6663	2.0564	0.9081	3.3411
17	13	0.7239	1.8492	0.9219	3.8203
18	14	0.7603	1.4949	0.9468	4.2352
19	13	0.7225	1.7923	0.9214	3.2139
20	13	0.7239	1.7659	0.9233	3.7956
21	13	0.6378	1.7337	0.9327	3.6956
22	13	0.7403	1.5395	0.9435	4.4539
23	14	0.7273	1.5494	0.9422	4.2037
24	13	0.8149	1.6351	0.9396	4.3028
25	13	0.7957	1.7674	0.9285	4.1362
26	14	0.7037	1.3020	0.9599	4.9274
27	13	0.7853	1.6825	0.9360	4.0366
28	13	0.7379	1.5936	0.9393	3.9595
29	14	0.7214	1.5419	0.9438	4.1272
30	13	0.7441	1.6365	0.9381	4.6138

N°Modèle	Nb. Variables Latentes	RMSEC	RMSEP	R^2	SDR
31	14	0.7193	1.4078	0.9530	4.8563
32	14	0.7716	1.4456	0.9511	4.4744
33	14	0.6445	1.8923	0.9177	3.9974
34	13	0.6512	1.9594	0.9141	3.5915
35	14	0.6819	1.6525	0.9369	4.4368
36	13	0.7228	1.8022	0.9226	3.7544
37	14	0.7463	1.6294	0.9385	4.4250
38	14	0.7449	1.5971	0.9373	3.8206
39	13	0.7024	1.8149	0.9226	3.5236
40	13	0.7982	1.5522	0.9432	4.1710
41	14	0.6988	1.8149	0.9248	3.7173
42	13	0.7521	1.7059	0.9357	3.6536
43	13	0.8320	1.6643	0.9327	4.1444
44	13	0.8142	1.4389	0.9487	4.6278
45	13	0.7377	1.8446	0.9195	3.5519
46	14	0.7215	1.9684	0.9122	3.5821
47	14	0.6803	1.9330	0.9084	3.5232
48	14	0.6271	1.6078	0.9435	4.0339
49	13	0.6966	1.7534	0.9261	3.8193
50	14	0.7626	1.7443	0.9274	3.4472
Moyenne	13.4400	0.7315	1.6957	0.9324	4.0192
Écart type	0.5014	0.0470	0.1634	0.0128	0.4126

BIBLIOGRAPHIE

Abu-Khalaf, N., B. S. Bennedsen et G. K. Bjorn. 2004. « Distinguishing carrot's characteristics by Near Infrared (NIR) reflectance and multivariate data analysis ». *Agricultural Engineering International,* vol. 6.

Alessi, P.J., E.C. Carter, M.D. Fairchild et R.W.G. Hunt. 2004. *CIE Technical Report. Colorimetry*. En Ligne. Commission Internationale de L'Eclairage. <http://www.cie.co.at/publ/abst/datatables15_2004/CIE_sel_colorimetric_tables.xls>. Consulté le 21 avril 2008.

Ashton, O. B. O., M. Wong, T. K. McGhie, R. Vather, Y. Wang, C. Requejo-Jackman, P. Ramankutty et A. B. Woolf. 2006. « Pigments in Avocado Tissue and Oil ». *J. Agric. Food Chem.,* vol. 54, n° 26, p. 10151-10158.

Baumann, Knut. 2003. « Cross-validation as the objective function for variable-selection techniques ». *TrAC Trends in Analytical Chemistry,* vol. 22, n° 6, p. 395-406.

Bennedsen, B. S., et N. Abu-Khalaf. 2004. « Near infrared (NIR) technology and multivariate data analysis for sensing taste attributes of apples. ». *International Agrophysics,* vol. 18, n° 3, p. 9.

Blackburn, George Alan. 2007. « Hyperspectral remote sensing of plant pigments ». *Journal of Experimental Botany,* vol. 58, p. 855-867.

Brosnan, Tadhg, et Da-Wen Sun. 2004. « Improving quality inspection of food products by computer vision--a review ». *Journal of Food Engineering,* vol. 61, n° 1, p. 3-16.

Burger, James, et Paul Geladi. 2005. « Hyperspectral NIR image regression part I: calibration and correction ». *Journal of Chemometrics,* vol. 19, n° 5-7, p. 355-363.

Charles, E. Lewis. 1978. « The maturity of avocados - a general review ». *Journal of the Science of Food and Agriculture,* vol. 29, n° 10, p. 857-866.

Chen, Yud-Ren, Kuanglin Chao et Moon S. Kim. 2002. « Machine vision technology for agricultural applications ». *Computers and Electronics in Agriculture,* vol. 36, n° 2-3, p. 173-191.

Clark, C. J., V. A. McGlone, C. Requejo, A. White et A. B. Woolf. 2003. « Dry matter determination in `Hass' avocado by NIR spectroscopy ». *Postharvest Biology and Technology,* vol. 29, n° 3, p. 301-308.

Colorni, A., M. Dorigo et V. Maniezzo. 1991. « Distributed Optimization by Ant Colonies ». In *Actes de la première conférence européenne sur la vie artificielle*. p. 134-142. Paris: Elsevier.

Colvin, James Barry. 2005. « Microscopy of Electronic Devices ». *EDFAS*. En Ligne. Vol. 17, n° 2. <http://www.fainstruments.com/FAQ.htm>. Consulté le 25 mars 2008.

Cornwell, M. Gerry 2005. *Guide de référence de l'éclairage*. Ressources Naturelles Canada, 84 p. <http://www.oee.nrcan.gc.ca/publications/equipement/eclairage/index.cfm?attr=24>.

Cox, Katy A., Tony K. McGhie, Anne White et Allan B. Woolf. 2004. « Skin colour and pigment changes during ripening of `Hass' avocado fruit ». *Postharvest Biology and Technology,* vol. 31, n° 3, p. 287-294.

Crow, Edwin L., Frances A. Davis et Margaret W. Maxfield. 1960. *Statistics Manual* Dover Publications 288 p.

Curda, L., et O. Kukackova. 2004. « NIR spectroscopy: a useful tool for rapid monitoring of processed cheeses manufacture ». *Journal of Food Engineering,* vol. 61, n° 4, p. 557-560.

Davies, A. M. C. 1998. « Cross Validation, do we love it too much? ». *Spectroscopy Europe,* vol. 18, n° 23, p. 2.

Davies, A. M. C. 2006. « Preparing for PLS calibration ». *Spectroscopy Europe,* vol. 18, n° 23, p. 2.

Davies, A. M. C. 2007. « Spectral pretreatments - derivatives ». *Spectroscopy Europe,* vol. 19, n° 32, p. 2.

Draper, N. R., et H. Smith. 1981. *Applied Regression Analysis (2nd Ed.)*, 1. USA: John Wiley & Sons, 709 p.

Du, Y. P., Y. Z. Liang, J. H. Jiang, R. J. Berry et Y. Ozaki. 2004. « Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares ». *Analytica Chimica Acta,* vol. 501, n° 2, p. 183-191.

El Masry, Gamal, Ning Wang, Adel El Sayed et Michael Ngadi. 2007. « Hyperspectral imaging for nondestructive determination of some quality attributes for strawberry ». *Journal of Food Engineering,* vol. 81, n° 1, p. 98-107.

ElMasry, Gamal, Ning Wang, Adel ElSayed et Michael Ngadi. 2007. « Hyperspectral imaging for nondestructive determination of some quality attributes for strawberry ». *Journal of Food Engineering,* vol. 81, n° 1, p. 98-107.

Faber, Klaas, et Bruce R. Kowalski. 1997. « Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares ». *Journal of Chemometrics,* vol. 11, n° 3, p. 181-238.

FAO, (FAOSTAT) 2006. « Principaux pays producteurs d'avocats ». En ligne. <http://faostat.fao.org/site/567/DesktopDefault.aspx?PageID=567>. Consulté le 25 janvier 2008.

Fearn, Tom, et A. M. C. Davies. 2002. « Sorting the wheat from the chaff ». *Spectroscopy Europe,* vol. 14, n° 16, p. 2.

Fearn, Tom, et A. M. C. Davies. 2007. « Removing Multiplicative effects ». *Spectroscopy Europe,* vol. 19, n° 32, p. 5.

Garrido Frenich, A., D Jouan-Rimbaud, D. L. Massart, S. Kuttatharmmakul, M. Martínez Galera et J. L. Martínez Vidal. 1995. « Wavelength selection method for multicomponent spectrophotometric determinations using partial least squares ». *Analyst,* vol. 120, p. 2787 - 2792.

Geladi, P. 1986. « Partial least-squares regression: a tutorial ». *Analytica Chimica Acta,* vol. 185, p. 1.

GMI, inc. 2006. « Spectrophotometry ». In *GMI.* En ligne. <http://www.gmi-inc.com/Genlab/spec_fig2%20diode%20array.gif>. Consulté le 10 avril 2008.

Gowen, A. A., C. P. O'Donnell, P. J. Cullen, G. Downey et J. M. Frias. 2007. « Hyperspectral imaging - an emerging process analytical tool for food quality and safety control ». *Trends in Food Science & Technology,* vol. In Press, Corrected Proof, p. 590-598.

Green, Robert L., et John H. Kalivas. 2002. « Graphical diagnostics for regression model determinations with consideration of the bias/variance trade-off ». *Chemometrics and Intelligent Laboratory Systems,* vol. 60, n° 1-2, p. 173-188.

Haaland, David M., et Edward V. Thomas. 1988. « Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information ». *Anal. Chem.,* vol. 60, n° 11, p. 1193-1202.

Hageman, J. A., M. Streppel, R. Wehrens et L. M. C. Buydens. 2003. « Wavelength selection with Tabu Search ». *Journal of Chemometrics,* vol. 17, n° 8-9, p. 427-437.

Hassibi, B. , D.G. Stork, G Wolf et T. Watanabe. 1994. « Optimal brain surgeon: Extensions, streamlining and performance comparisons ». *Advances in Neural Information Processing Systems,* vol. 6, p. 263-271.

Haswell, S. J., et A. D. Walmsley. 1999. « Chemometrics: the issues of measurement and modelling ». *Analytica Chimica Acta,* vol. 400, n° 1-3, p. 399-412.

He, Y., Y. Zhang, A. G. Pereira, A. H. Gómez et J Wang. 2005. « Nondestructive Determination of Tomato Fruit Quality Characteristics Using Vis/NIR Spectroscopy Technique ». *International Journal of Information Technology,* vol. 11, n° 11, p. 12.

Helge, Toutenburg. 2006. « Book Review: Regression Diagnostics - Identifying influential data and sources of collinearity. By D. A. Belsley, E. Kuh and R. E. Welsch ». *Biometrical Journal,* vol. 48, n° 6, p. 1044.

Hemraj, Indradeo, Craig Rineer, Sushmitha Kurapati et Mariza Clement. 1997. « Absorption Spectrum of Chlorophyll ». <http://www.seas.upenn.edu/courses/belab/LabProjects/1997/BE210S97R1R01.doc>.

Hickson, Brett. 2006. « Quality assessment of avocados by means of dry matter content ». <http://www.unece.org/trade/agr/meetings/capacity-building/2006_mojmirovce-SK/Dry%20matter%20content%20of%20avocados%20(June%202006).pdf>.

HunterLab. 2008. « Equivalent white light sources and CIE illuminants ». *Insight on color.* Application Note. Vol. 17, n° 5, p. 1-5.

Isler, Otto, Hugo Gutmann et Ulrich Solms. 1971. *Carotenoids.* Birkhauser-Verlag <http://www.chm.bris.ac.uk/motm/carotene/beta-carotene_colourings.html>.

Jha, S. N., S. Chopra et A. R. P. Kingsly. 2005. « Determination of Sweetness of Intact Mango using Visual Spectral Analysis ». *Biosystems Engineering,* vol. 91, n° 2, p. 157-161.

Kaiser, Peter K. 2005. « Electromagnetic Spectrum ». In *The Joy of Visual Perception.* En ligne. <http://www.yorku.ca/eye/spectru.htm>. Consulté le 10 avril 2008.

Kalivas, John H. 2004. « Pareto calibration with built-in wavelength selection ». *Analytica Chimica Acta,* vol. 505, n° 1, p. 9-14.

Kavdir, I., R. Lu, D. Ariana et M. Ngouajio. 2007. « Visible and near-infrared spectroscopy for nondestructive quality assessment of pickling cucumbers ». *Postharvest Biology and Technology,* vol. 44, n° 2, p. 165-174.

Koo-Lee, Seung. 2003. « Harvesting Avocados ». <http://www.avocadosource.com/books/avocadohandbook/harvesting_files/harvesting_avo.pdf>.

Lantéri, P., et Longeray R. 1998. « Chimiométrie : outils du XXème siècle, méthode du XXIème siècle ? ». *ANALUSIS MAGAZINE*. p. 4.

Leardi, Riccardo. 2000. « Application of genetic algorithm-PLS for feature selection in spectral data sets ». *Journal of Chemometrics,* vol. 14, n° 5-6, p. 643-655.

Leardi, Riccardo, et A. Lupianez Gonzalez. 1998. « Genetic algorithms applied to feature selection in PLS regression: how and when to use them ». *Chemometrics and Intelligent Laboratory Systems,* vol. 41, p. 195-207.

Lee, S. K. , R. E. Young, P. M. Schiffman et C. W. Coggins. 1983. « Maturity Studies of Avocado Fruit Based on Picking Dates and Dry Weight ». *Journal of the American Society for Horticultural Science,* vol. 108, n° 3, p. 390-393.

Lima, Silvio L. T., Cesar Mello et Ronei J. Poppi. 2005. « PLS pruning: a new approach to variable selection for multivariate calibration based on Hessian matrix of errors ». *Chemometrics and Intelligent Laboratory Systems,* vol. 76, n° 1, p. 73-78.

Lindbloom, Bruce Justin. 2003. « Computing XYZ From Spectral Data ». In *Useful Color Equations*. En Ligne. <http://www.brucelindbloom.com/>.

Liu, Y., Y. Ying et X. Fu. 2004. « Prediction of valid acidity in intact apples with Fourier transform near infrared spectroscopy ». *Journal of Zhejiang University* vol. 6, n° 3, p. 7.

McCarthy, Alec. 2005. *Avocado maturity testing*. Western Austrialia: Bunbury District Office, 2 p.

McGlone, V. Andrew, Robert B. Jordan et Paul J. Martinsen. 2002. « Vis/NIR estimation at harvest of pre- and post-storage quality indices for `Royal Gala' apple ». *Postharvest Biology and Technology,* vol. 25, n° 2, p. 135-144.

McGlone, V. Andrew, et Sumio Kawano. 1998. « Firmness, dry-matter and soluble-solids assessment of postharvest kiwifruit by NIR spectroscopy ». *Postharvest Biology and Technology,* vol. 13, n° 2, p. 131-141.

Miller, James, et Jane Miller. 2005. *Statistics and Chemometrics for Analytical Chemistry (5th Ed.)*. Pearson, 268 p.

Montes, J. M., H. F. Utz, W. Schipprack, B. Kusterer, J. Muminovic, C. Paul et A. E. Melchinger. 2006. « Near-infrared spectroscopy on combine harvesters to measure

maize grain dry matter content and quality parameters ». *Plant Breeding,* vol. 125, n° 6, p. 591-595.

Newport. 2008. *CMOS vs CCD Detectors*. En Ligne. <http://www.newport.com/store/genproduct.aspx?id=642094&lang=1033&Section=Detail>. Consulté le 25 mars 2008.

Osborne, Scott D., Robert B. Jordan et Kunnemeyera Rainer. 1997. « Method of Wavelength Selection for Partial Least Squares ». *Analyst,* vol. 122.

Osten, David, W. . 1988. « Selection of optimal regression models via cross-validation ». *Journal of Chemometrics,* vol. 2, n° 1, p. 39-48.

Ozdemir, Feramuz, et Ayhan Topuz. 2004. « Changes in dry matter, oil content and fatty acids composition of avocado during harvesting time and post-harvesting ripening period ». *Food Chemistry,* vol. 86, n° 1, p. 79-83.

Qing-Song Xu, Yi-Zeng Liang Yi-Ping Du. 2004. « Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration ». *Journal of Chemometrics,* vol. 18, n° 2, p. 112-120.

Shamsipur, Mojtaba, Vali Zare-Shahabadi, Bahram Hemmateenejad et Morteza Akhond 2006. « Ant colony optimisation: a powerful tool for wavelength selection ». *Journal of Chemometrics,* vol. 20, n° 3-4, p. 146-157.

Smith, T., et J. Guild. 1931. « The C.I.E. colorimetric standards and their use ». *Transactions of the Optical Society,* vol. 33, n° 3, p. 73-134.

Specim. 2003. *ImSpector user manual*, 1. Specim, 31 p.

Springsteen, Art. 1999. « Standards for the measurement of diffuse reflectance - an overview of available materials and measurement laboratories ». *Analytica Chimica Acta,* vol. 380, n° 2-3, p. 379-390.

Subedi, P. P., K. B. Walsh et G. Owens. 2007. « Prediction of mango eating quality at harvest using short-wave near infrared spectrometry ». *Postharvest Biology and Technology,* vol. 43, n° 3, p. 326-334.

Sun, Da-Wen. 2004. « Computer vision--an objective, rapid and non-contact quality evaluation tool for the food industry ». *Journal of Food Engineering,* vol. 61, n° 1, p. 1-2.

Sutter, J. M., et J. H. Kalivas. 1993. « Comparison of Forward Selection, Backward Elimination, and Generalized Simulated Annealing for Variable Selection ». *Microchemical Journal,* vol. 47, n° 1-2, p. 60-66.

Tenenhaus, Michel. 1998. *La régression PLS: Théorie et pratique* 1. 1. Technip, 254 p.

Tian, Hai-qing, Yi-bin Ying, Hui-shan Lu, Xia-ping Fu et Hai-yan Yu. 2007. « Measurement of soluble solids content in watermelon by Vis/NIR diffuse transmittance technique ». *Journal of Zhejiang University - Science B,* vol. 8, n° 2, p. 105-110.

Van de Laer, G., J. L. Rolot, Dardenne P. et R. Agneessens. 2001. « Qualité des pommes de terre : nouvelles méthodes d'évaluation calibrées sur l'analyse sensorielle ». *Biotechnologie, Agronomie, Société et Environnement* vol. 5, n° 3, p. 166-170.

Vancolen, Séverine. 2004. « La régression PLS ». Neuchâtel, 28 p. <http://doc.rero.ch/lm.php?url=1000,41,4,20070716085523-YM/mem_VancolenS.pdf>.

Walker, John. 1996. « CIE X,Y,Z Components ». In *Colour Rendering of Spectra.* En ligne. <http://www.fourmilab.ch/documents/specrend/>.

Wikipédia. 2008a. « Image:CIExy1931 CIERGB.png ». In *Wikipédia : L'encyclopédie libre.* En Ligne. <http://commons.wikimedia.org/wiki/Image:CIExy1931_CIERGB.png>. Consulté le 4 avril 2008.

Wikipédia. 2008b. « Image:Prism-rainbow.svg ». In *Wikipédia : L'encyclopédie libre.* En ligne. <http://commons.wikimedia.org/wiki/Image:Prism-rainbow.svg>. Consulté le 4 février 2008.

Wold, H. 1966. « Estimation of principal components and related models by iterative least squares ». *Multivariate Analysis* (New York). p. 391-420.

Wold, S., H. Martens et H. Wold. 1983. « The multivariate calibration method in chemistry solved by the PLS method ». In *Lecture Notes in Mathematics.* Vol. 973, p. 286-293. Berlin: Springer.

Woolf, Allan, Chris Clark, Emma Terander, Vong Phetsomphou, Reuben Hofshi, Mary Lu Arpaia, Donella Boreham, Marie Wong et Anne Whit. 2003. « Measuring avocado maturity; ongoing developments ». *The Orchadist*, p. 40-45.